Technische Universität München
Zentrum Mathematik
Lehrstuhl für Optimalsteuerung

On sparse sensor placement for parameter identification problems with partial differential equations

Daniel Walter

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Michael Ulbrich
Prüfer der Dissertation: 1. Prof. Dr. Boris Vexler
2. Prof. Dr. Karl Kunisch
Karl-Franzens-Universität Graz
(nur schriftliche Beurteilung)
3. Prof. Dr. Kristian Bredies
Karl-Franzens-Universität Graz

Die Dissertation wurde am 11.12.2018 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 23.5.2019 angenommen.

Abstract

This thesis is concerned with the formulation and analysis of a sparse optimization framework for the optimal placement of measurement sensors in inverse problems. At the focus of attention are settings in which an unknown parameter entering a partial differential equation is estimated from finitely many observations of the corresponding state. To mitigate the influence of measurement noise we propose to determine the optimal number of sensors and their positions based on the solution of a mathematical optimization problem. Therefore we minimize a suitable optimality criterion for the distribution of the sensors which is modeled as a measure on the set of possible candidate locations. The proposed sensor placement framework is applied for two model problems. Suitable approximation approaches based on a finite element discretization as well as efficient solution algorithms are discussed. The last part of the thesis introduces a first order solution algorithm for composite minimization problems in a general setting. Convergence of the method is addressed and worst case convergence rates are derived. In the case of measure-valued optimization variables the method is augmented by additional acceleration steps leading to improved convergence results.

Zusammenfassung

Diese Arbeit befasst sich mit der Formulierung und Analysis eines "sparsen" Optimierungsansatzes für die optimale Platzierung von Messsensoren in inversen Problemen. Im Mittelpunkt stehen Problemformulierungen bei denen unbekannte Parameter in partiellen Differentialgleichungen aus endlich vielen Beobachtungen des zugehörigen Zustands geschätzt werden sollen. Um den Einfluss von Messfehlern zu verringern wird vorgeschlagen die optimale Anzahl von Sensoren und deren Positionen basierend auf der Lösung eines mathematischen Optimierungsproblems zu bestimmen. Dahingehend minimieren wir ein geeignets Optimalitätskriterium bezüglich der Verteilung der Messensoren. Diese wird als Maß auf der Menge der möglichen Positionen modelliert. Wir wenden den vorgeschlagenen Ansatz zur optimalen Sensorplatzierung auf zwei Modelbeispiele an. Geeignete Approximationsverfahren basierend auf Finite Elemente Diskretisierungen sowie effiziente Lösungsverfahren werden diskutiert. Im letzten Teil der Arbeit wird ein allgemeines Optimierungsverfahren erster Ordnung für Kompositminimierungsprobleme eingeführt. Konvergenz des Verfahrens wird behandelt und Worst-Case Konvergenzraten werden hergeleitet. Für den Fall maßwertiger Optimierungsvariablen wird die Methode mit zusätzlichen Beschleunigungsschritten versehen. Dies führt zu verbesserten Konvergenzresultaten.

Contents

1 Introduction

This thesis focusses on the description and analysis of an optimization based approach to the placement of measurement sensors for the identification of unknown parameters in processes described by partial differential equations (PDEs). With the advent of constantly rising computational capacities and steadily improving numerical methods, mathematical models as surrogates for complex real-life processes have become a cornerstone and indispensable tool of modern day science. Applications range from simulating the smallest of particles in chemistry or physics to the characterization of global phenomena such as changes in the weather or the ocean current. In many cases, such processes are well-described by a state variable whose dynamics are governed by a system of partial differential equations. In most cases, a full description of such mathematical surrogates requires knowledge on the value of additional parameters entering in the equation. These may arise in the modeling process or correspond to unknown physical quantities such as material constants. Thus, rather than one particular partial differential equation to describe the modeled process, we have given a parametrized family of possible ones.

A properly chosen mathematical model may enable to predict on the behavior of the underlying process based on simulations. For this purpose, it is however indispensable to calibrate the unknown parameters i.e. to select them such that the associated equation and its solution describe the modeled process most faithfully. A direct measurement of the parameters often requires disproportional effort or is not possible at all. Inference on their value is only possible indirectly by e.g. measuring the quantity resembled by the state variable. From a practical point of view, this process of measuring observable quantities corresponds to conducting an experiment in which data is collected by measurement devices or sensors. A sophisticated mathematical approach to the problem of parameter identification is then constituted by solving a so-called *inverse problem*: Here, we also describe the measurement process mathematically e.g. by an operator mapping the state variable into the space of measurements. For each possible value of the unknown parameter we can now solve the associated partial differential equation and plug the obtained state into the measurement model. Subsequently, we identify a particular parameter such that the measurements predicted by the associated mathematical model match those obtained in the experiment.

In practice, this task is aggravated by several factors. First, mathematical modeling usually involves simplifying assumptions to yield equations that describe the modeled process sufficiently good and that are still tractable with numerical methods. In particular, this implies that modeling errors are present and there might be no parameter such that the response of the measurement model matches the observed experimental data exactly. Moreover, the experimental data which is used to infer on the unknown parameter is perturbed by measurement errors. These stem back to the imperfectness of the involved measurement devices in the experiment. Last, in many cases, the experimental data is given by a possibly small number of scalars where each one corresponds to the measurement taken by a particular sensor. In contrast, the unknown parameter may be a high-dimensional vector or even a distributed function. If there is such a discrepancy between the amount of provided data and the dimension of the parameter space, the inverse problem is underdetermined and an exact identification of the unknown parameter is impossible without further assumptions on its structure. The consequences on the inverse problem induced by the described defects, are summarized under the notion of *ill-posedness*. A characteristic feature of most ill-posed inverse problems is the discontinuity of their solutions with respect to the collected measurements. Thus, slight changes in the experimental data due to measurement errors may lead to the wrong conclusions on the unknown parameter if the problem is solved directly. As a consequence, ill-posed inverse problems call for appropriate *regularization strategies* which allow to compute stable approximations of their solutions. We refer, e.g., to the famous concept of *Tikhonov regularization* or the *Bayesian approach* to inverse problems.

After obtaining an approximate solution to the inverse problem for a given set of experimental data, we have to assess its reliability due to the presence of measurement errors. For this purpose at least slight assumptions on the nature of these perturbations have to be made. In practical experiments, measurements are not reproducible i.e. taking the same measurement twice leads to slightly different outcomes. These inaccuracies stem back to the inability of an experimenter or of the used sensor to repeat the measurement in the exact same way. This observation suggests to adopt a probabilistic model for the measurement error and assume prior knowledge on its distribution. As a consequence, since the approximate solution to the inverse problem depends on the measurements, it should also be interpreted as realization of a random variable taking values in the parameter space. In particular, we should drop the notion of *identifying* the unknown parameter and replace it by the more appropriate term of *parameter estimation* to stress the randomness in the problem. The distribution of the random parameter estimator depends on properties of the measurement error model. For this reason, it allows to study the influence of perturbations in the experimental data and assess the reliability of approximate solutions to the inverse problem in a probabilistic sense. More generally, the results of the parameter estimation process rely on the conditions of the measurement experiment such as the placement of available measurement sensors and the amount of provided measurements. Poorly conducted measurements may yield uninformative experimental data i.e. no conclusions on the value of the unknown parameter can be made based upon them. In contrast, a well-planned experiment allows for an, in some suitable sense, optimal estimation of the unknown parameters while simultaneously minimizing the overall cost of the measurement process. This leads to the task of optimally designing experiments *before* any measurements are taken in practice.

In the context of the present thesis, our focus lies on the inverse problem of estimating an unknown parameter q in some Hilbert space Q entering into a partial differential equation described by an operator $A(q,\cdot)$. Inference on its value is possible based on a finite number of $N \in \mathbb{N}$ scalar measurements $\mathbf{y}_d^i$, $i = 1, \dots N$, taken of the quantity resembled by the state variable $y \in Y$. We assume that the dependence between one of these measurements and the state is linear. Moreover, each measurement $\mathbf{y}_d^i$ is associated to a point x_i in a compact set $\Omega_o \subset \mathbb{R}^d$, $d \in \mathbb{N}$. For example, x_i may describe the position of the applied measurement sensor. Now, we model the action of the sensor at $x_i \in \Omega_o$ on the state variable by an element $\mathcal{O}(x_i) \in Y^*$, $i = 1, \dots, N$, in the topological dual of the state space. As an example, if the elements of Y are continuous functions on Ω_o, then we may consider a pointwise measurement taken at a point x_i. The corresponding element in the dual space Y^* is given by the associated Dirac delta function $\mathcal{O}(x_i) = \delta_{x_i}$. The resulting inverse problem now reads as

$$\text{find } q \in Q_{ad},\ y \in Y\colon \quad \langle y, O(x_i)\rangle_{Y,Y^*} = \mathbf{y}_d^i,\ i = 1, \dots, N, \quad A(q,y) = 0,$$

where $Q_{ad} \subset Q$ denotes a set of admissible parameters and $\langle \cdot, \cdot \rangle_{Y,Y^*}$ denotes the duality paring between the state space Y and its dual Y^*. The measurement at x_i is subject to additive perturbation by normally distributed noise $\varepsilon_i \sim \mathcal{N}(0, 1/\mathbf{u}_i)$, $i = 1, \dots, N$. The scalar quantity $\mathbf{u}_i > 0$ should be interpreted as diligence factor giving information on how carefully the data should be collected at the corresponding measurement point. Measurement errors at distinct locations are assumed to be uncorrelated. In order to mitigate the influence of perturbations in the measurements on (approximate) solutions of the inverse problem, we fix a parameter $\hat{q} \in Q$ and consider the associated Fisher information operator $X^* \varSigma^{-1} X$ which acts on the parameter space Q as

$$(\delta q_1, X^* \varSigma^{-1} X \delta q_2)_Q = \sum_{i=1}^{N} \mathbf{u}_i \langle \partial S[\hat{q}] \delta q_1, O(x_i) \rangle_{Y,Y^*} \langle \partial S[\hat{q}] \delta q_2, O(x_i) \rangle_{Y,Y^*} \quad \forall \delta q_1, \delta q_2 \in Q. \tag{1.1}$$

Here, the sensitivity $\partial S[\hat{q}] \colon Q \to Y$ describes the effect of perturbations in $\hat{q}$ on the associated solution $y = S[\hat{q}]$ to the partial differential equation. The parameter $\hat{q}$ represents e.g. an a priori guess to the solution of the inverse problem given the noise-free measurements. For optimal inference on the unknown parameter entering into the partial differential equation, we propose to optimize or design the measurement experiment in which the data $\mathbf{y}_d$ is obtained, according to the solution of a mathematical optimization problem based on properties of the Fisher information operator. More in detail, we parametrize $X^* \varSigma^{-1} X$ as a function of the number of measurements $N \in \mathbb{N}$, their positions $\{x_i\}_{i=1}^N$ in the admissible set $\varOmega_o$ as well as the diligence factors $\{\mathbf{u}_i\}_{i=1}^N \subset \mathbb{R}_+$ and solve the optimal sensor placement problem

$$\min_{x_i \in \varOmega_o, \mathbf{u}_i \in \mathbb{R}_+, N \in \mathbb{N}} [\varPsi(X^* \varSigma^{-1} X) + G(\|\mathbf{u}\|_{l_1})] \quad \text{where} \quad X^* \varSigma^{-1} X \text{ fulfills (1.1).}$$

Here, $\varPsi$ denotes a scalar-valued smooth and convex design criterion. To capture the overall cost of the measurement experiment, we add an additional convex term $G(\|\mathbf{u}\|_{l_1})$ to the problem which involves the l_1 norm of the measurement weight vector $\|\mathbf{u}\|_{l_1} = \sum_{i=1}^N \mathbf{u}_i$. For example, we may choose $G(\|\mathbf{u}\|_{l_1}) = \beta \|\mathbf{u}\|_{l_1}$ where $\beta > 0$ denotes the cost associated to a single measurement.

The optimal selection of measurement points and weights based on the Fisher information first came up in the context of polynomial regression, [245]. Nowadays, such formulations form the basis for the vast field of model-based optimal design of experiments which is concerned with the optimization-aided selection of experimental conditions. If the dependence of the state variable y on the unknown parameter is nonlinear, the obtained optimal solutions depend on the parameter $\hat{q}$ leading to the notion of *locally optimal designs*, see e.g [217]. Approaches to cope with this dependence include the consideration of robust/worst-case or averaged design criteria, [37, 171, 217] as well as sequential design approaches, [170], where one alternates between estimating the unknown parameter and obtaining a new measurement setup based on the Fisher information of the current estimate. Optimal design of experiments has been frequently studied and successfully applied for processes described by ordinary differential equations, [71, 244], and differential-algebraic equations, [20, 38, 169]. More recently, extensions of this concept to models given by partial differential equations have been considered in e.g. [4, 15, 103, 138, 181]. We also point out to the thesis [57] and the early work [157].

We emphasize that the optimal sensor placement problem has a combinatorial aspect due to the unknown optimal number of measurements. This aggravates discussions on the well-posedness of the problem, i.e. the existence of solutions, as well as the derivation of necessary optimality conditions. Clearly, it also poses a serious difficulty for the practical computation of optimal measurement setups. For this reason, the maximum number N of measurements is often fixed a priori and the design criterion is only minimized with respect to the positions $\{x_i\}_{i=1}^N$ and the measurement weights $\{\mathbf{u}_i\}_{i=1}^N$. In this case, optimal sensor placement is a nonlinear finite-dimensional optimization problem. We point out that, while the design criterion as well as the regularization term are assumed to be convex, the possibly complicated dependence of the Fisher information operator on the positions of the sensors renders the sensor placement problem nonconvex in general. In particular, first-order optimality conditions, if they can be derived, are only necessary but not sufficient. Thus, the problem may admit a large number of stationary points which are not necessarily (local) minimizers. As a consequence, the computation of a global minimizer to the problem is not feasible in most cases. Moreover, we also mention that, if the problem is smooth, the application of first-order optimization methods in order to compute a stationary point requires derivatives of the Fisher information with respect to the positions of the sensors. This can be a challenging problem in itself, see e.g. [116]. For this reason, the admissible set Ω_o is often chosen as a finite collection of points which correspond e.g. to nodal points of a triangulation. This additional simplification reduces optimal sensor placement to a finite-dimensional convex minimization problem for the measurement weights $\{\mathbf{u}_i\}_{i=1}^N$. For sensor placement problems with l_1 regularization term in this setting we refer e.g. to [4,71,127]. It is, by now, a well-known fact that penalizing the l_1 norm of the optimization variable favors optimal measurement weight vectors that are sparse i.e. they will only contain few nonzero entries.

In the context of this thesis we will neither prescribe an a priori upper bound on the number of used sensors nor will we, at least for most of the derived results, impose any restrictions, beyond compactness, on the admissible set of possible sensor positions Ω_o. In particular, we stress that Ω_o is not necessarily given by a finite collection of points. The main novelty of the present work is to bypass the aforementioned difficulties, i.e. the non-convexity and combinatorial nature of the problem, by embedding optimal sensor placement into a more abstract framework: Associated to a vector of sensor positions $\mathbf{x} = (x_1, \dots, x_N)^\top \in \Omega_o^N$ and a vector of measurement weights $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_N)^\top \in \mathbb{R}_+^N$ we define the design measure $u = \sum_{i=1}^N \mathbf{u}_i \delta_{x_i}$. We point out that the total variation norm of this conic combination of Dirac delta functions equals the l_1 norm of the measurement weight vector $\mathbf{u}$:

$$\|u\|_{\mathcal{M}} = \int_{\Omega_o} \mathrm{d}u(x) = \sum_{i=1}^N \mathbf{u}_i = \|\mathbf{u}\|_{l_1}.$$

Moreover, for an arbitrary spatial point $x \in \Omega_o$ consider the operator $\mathcal{O}(x) \otimes \mathcal{O}(x)$ acting on the parameter space Q as

$$(\delta q_1, [\mathcal{O}(x) \otimes \mathcal{O}(x)]\delta q_2)_Q = \langle \partial S[\hat{q}]\delta q_1, O(x)\rangle_{Y,Y^*} \langle \partial S[\hat{q}]\delta q_2, O(x)\rangle_{Y,Y^*} \quad \forall \delta q_1, \delta q_2 \in Q.$$

Now, we make the important observation that the Fisher information operator can be equivalently rewritten as an integral with respect to the design measure u:

$$X^* \Sigma^{-1} X = \sum_{i=1}^N \mathbf{u}_i [\mathcal{O}(x_i) \otimes \mathcal{O}(x_i)] = \int_{\Omega_o} [\mathcal{O}(x) \otimes \mathcal{O}(x)] \,\mathrm{d}u(x).$$

Here, integration has to be understood in the sense of Bochner. This reasoning leads to the *sparse sensor placement* problem

$$\min_{u \in \mathcal{M}^+(\Omega_o)} [\Psi(\mathcal{I}(u)) + G(\|u\|_{\mathcal{M}})] \quad s.t. \quad \mathcal{I}(u) = \int_{\Omega_o} [\mathcal{O}(x) \otimes \mathcal{O}(x)] \,\mathrm{d}u(x),$$

where we minimize with respect to the design measure u in the set of positive Borel measure $\mathcal{M}^+(\Omega_o)$ on the admissible set. Loosely speaking, this reformulation can be interpreted as minimization problem for the distribution of the measurement sensors on Ω_o instead of minimizing for the position of each sensor, the associated measurement weight as well as the overall number of measurements separately. The crucial advantage of the new formulation is the linear dependence of the Bochner integral on the measure u. As a consequence, in contrast to the original problem, the resulting sparse sensor placement problem is convex. Moreover, the combinatorial nature of the problem vanishes. This allows to treat sparse sensor placement as nonsmooth but convex minimization problem. In particular, necessary *and* sufficient first-order optimality conditions for optimal design measures can be derived.

Nevertheless, this comes at the price of having to deal with minimization problems on the space of Borel measures $\mathcal{M}(\Omega_o)$ which, in some sense, shifts the difficulties in the problem to the considered function space. For example, the Banach space of Borel measures on Ω_o lacks desirable properties such as reflexivity or smoothness which complicates the design and analysis of efficient numerical solution algorithms. Moreover, we point out that we minimize over the whole set $\mathcal{M}^+(\Omega_o)$ rather than only considering measures of the form $u = \sum_{i=1}^{N} \mathbf{u}_i \delta_{x_i}$. This is an, a priori, necessary extension of the problem to discuss its well-posedness and to derive optimality conditions since the cone of Dirac delta functions on Ω_o is not closed with respect to a suitable topology. We will however discuss conditions which ensure the existence of a minimizer comprising finitely many Dirac Delta functions. In particular, this is the case if the unknown parameter is finite-dimensional. The number of Diracs in such a solution, their positions and the associated coefficients then provide a solution of the nonconvex and combinatorial problem. This makes both problems essentially equivalent with the crucial difference that the sparse sensor placement problem is convex.

A rigorous analysis of sparse sensor placement problems and their efficient algorithmic solution are at the heart of this thesis. Moreover, for the practical computation of optimal measurement designs, we present a discretization framework based on a finite element discretization of the partial differential equation and, if Q is infinite-dimensional, on a sophisticated discretization of the parameter space. All arguments are backed up by accompanying a priori error estimates. For finite-dimensional parameter spaces, in contrast to prior approaches on robust optimal design, we cope with the dependence of optimal solutions on the linearization point by providing stability results and sensitivities for optimal design measures with respect to perturbations in the data of the sensor placement problem. While the presentation of these results is restricted to the important case of pointwise measurements and norm regularization i.e. $G(\|u\|_{\mathcal{M}}) = \beta\|u\|_{\mathcal{M}}$, we are confident that an extension to more general measurement models and different regularization terms is possible.

Before proceeding to a more detailed outline of the presented results, we give a brief overview on similar approaches to optimal sensor placement and the rapidly developing area of optimization with sparsity enhancing regularizers. This allows to put the derived results in the bigger picture and highlights their novelty. As already mentioned at an earlier point of these introductory remarks, optimal design of experiments based on the Fisher information of the parameter estimates first came up in the context of linear regression in statistics. Here, the interest lies in a sophisticated choice of sampling points in a set Ω_o in order to guarantee optimal inference on the unknown regression coefficients from the obtained samples. Systematic approaches to this problem are often based on the notion of approximate or continuous design theory stemming back to the works of Kiefer and Wolfowitz [163, 165]. This approach models potential measurement experiments as probability measures over the design space Ω_o. The mass associated to a Dirac delta function in such a measure describes the fraction of available measurements that should be conducted at the corresponding point in Ω_o. This idea has been further developed in numerous publications. For further reference we point out to the monographs [9, 205, 222] and [197, 198]. An extension of this method to nonlinear models is based on linearization, see e.g. [107, 217]. A fundamental pillar of continuous design theory is constituted by the famous Kiefer-Wolfowitz Theorem, [164, 166], which allows to check if a given design measure is an optimal one. In general, optimal designs cannot be given in closed form. If the design space Ω_o consists of finitely many points, the algorithmic solution of the continuous design problem is usually based on the multiplicative algorithm due to Silvey and Titterington, [241]. For general sets Ω_o, algorithms of the Fedorov-Wynn type, [105, 271–273], are applied to compute optimal designs. These methods compute optimal designs by the sequential selection of new sensors based on the gradient of the design criterion and an update of the associated measurement weights. For an adaptation of continuous designs to the estimation of finite dimensional parameters in a partial differential equation we refer to [17, 256] and the references therein. In the context of the present thesis, the continuous design problem can be recovered by an appropriate choice of the regularization term $G(\|u\|_{\mathcal{M}})$ if the design criterion Ψ fulfils mild monotonicity assumptions which is the case for all prominent examples. At this point, we emphasize that our results should not be seen as a simple adaptation of this well-established approach. Quite the contrary, we also contribute to the theory of continuous designs in a substantial way. First and foremost, we stress that, to our best knowledge, all previous works in this direction were restricted to finite dimensional parameter spaces while we also deal with the infinite-dimensional case. We provide a set of equivalent necessary and sufficient first-order optimality conditions for the sparse sensor placement problem which reduce to the result of the classic Kiefer-Wolfowitz Theorem if Q is finite-dimensional. We refer to Theorem 3.17 and Example 3.5. Moreover, in Section 4.4.5 we identify the Fedorov-Wynn algorithm as special instance of a conditional gradient method. Based on the results obtained in Chapter 6, we derive worst-case convergence rates for a general class of optimal design criteria. Most important, we provide an accelerated version of the method reminiscent to those proposed by Wu, [269, 270] and, more recently, by Biedermann, [275], and Boyd, [44]. In contrast to these previous works, we obtain a provable improved convergence behavior of the method if additional structural requirements are met. To the best of our knowledge, comparable results are only available if Ω_o consists of finitely many points, [1, 2]. Last, while the monograph [256] extends the idea of continuous designs to parameter estimation problems with partial differential equations, it does neither touch the topic of discretizing the problem nor does it study the influence of perturbations in the problem on the obtained designs. To sum up, while our primary interest lies in parameter identification problems with partial differential equations, the derived results may also have a considerable impact on continuous design theory which is, traditionally, a topic studied in statistics. These considerations highlight the strong interdisciplinary component of the present thesis.

For a complete overview, the sparse sensor placement problem

$$\min_{u \in \mathcal{M}^+(\Omega_o)} [\Psi(\mathcal{I}(u)) + G(\|u\|_{\mathcal{M}})]$$

should also be discussed in the broader context of nonsmooth composite minimization problems. By now, it is a well-known fact that a penalization of the total variation norm favours optimal solutions that are sparse i.e. they are supported on sets of Lebesgue measure zero. In particular, minimizers may be only supported on finitely many points. This observation makes measure-valued optimization variables appealing for inverse problems. For example, we mention acoustic and seismic inversion, [178, 209], as well as super-resolution, [55, 95]. In optimal control, sparsity provides a suitable framework for e.g. the optimal placement of actuators, [54,74,113]. The overall aim of the present thesis is to showcase the applicability of sparse minimization to the problem of optimal sensor placement. As outlined in [73], sparse minimization problems are closely related to the well-studied subject of state-constraint optimization.

The results contained in this thesis benefit from the advanced level of research in these fields but we also contribute to them in several ways. For example, we point out to the sensitivity results of Section 4.5 which are based on generalizing techniques from the recent work [95]. The finite element discretization framework in Sections 4.6 and 5.2, respectively, is inspired by the variational approaches for sparse optimal control problems in [59, 210]. A characteristic trait of sparse minimization problems is that sequences of perturbed optimal solutions obtained by e.g. discretizing the problem, do not converge in the total variation norm but only with respect to weaker topologies. Therefore, it is not obvious how to quantify the convergence of such sequences. For finite-dimensional parameter spaces Q, we will prove that the sparse sensor placement problem admits solutions consisting of finitely many Dirac delta functions. Considering a sequence of such sparse measures, we study and quantify the convergence of the position and coefficient associated to each Dirac delta. This can be achieved by extending results from semi-infinite optimization cf. [191]. These considerations require additional structural assumptions on the problem which we obtain by adapting the recent concept of non-degenerate source conditions, [94], from super-resolution theory to the problem at hand. On a more abstract level, we will observe that these convergence results imply convergence rates for sequences of sparse measures in a modified version of the well-known Wasserstein distance, see e.g. [259]. This new measure of convergence is computationally accessible which also allows to verify the obtained theoretical results in practice. While the Wasserstein distance is a common tool in the theory of optimal transport, it was, to the best our knowledge, not yet considered in sparse optimal control problems.

The last big topic that is covered in this work is the design and analysis of efficient solution algorithms for optimization problems with measure-valued variables. As already pointed out at an earlier point, this is a challenging problem for several reasons. First, the objective functional contains the typically nonsmooth term $G(\|u\|_{\mathcal{M}})$. Second, the space of Borel measures lacks reflexivity, smoothness and strict convexity. Most well-known methods do not yield a direct extension to this setting. While this difficulty can be overcome by simply discretizing the problem, such reasoning harbors the danger of yielding *mesh-dependent* optimization algorithms i.e. their convergence behavior critically depends on the discretization parameters. For this reason, our interest lies in the formulation and analysis of iterative solution algorithms in the function space setting. Concrete practical realizations of such methods can be expected to show a *mesh-independent* convergence behavior i.e. the number of necessary steps to reach some convergence criterion will be essentially independent of the number of degrees of freedom in the discretization. One possibility to tackle this problem, is to consider another, closely related, approach promoting sparse solutions given by

$$\min_{u\in L^2(\Omega_o),u\geq 0}\left[\Psi(\mathcal{I}(u))+G(\|u\|_{L^1(\Omega_o)})+\frac{\varepsilon}{2}\|u\|^2_{L^2(\Omega_o)}\right]$$

where $L^1(\Omega_o)$ and $L^2(\Omega_o)$ denote the spaces of integrable and square integrable functions on Ω_o. By $\varepsilon>0$ we denote an additional regularization parameter. Note that for $u\in L^1(\Omega_o)$ there holds $\|u\|_{L^1(\Omega_o)}=\|u\|_{\mathcal{M}}$. Obviously, this formulation no longer allows for optimal solutions supported on finitely many points. However, since the total variation of u is still present in the problem, the additional regularization still enhances sparsity to some extend i.e. optimal solutions will be zero outside of a small subset of Ω_o. The uniform convexity of the squared $L^2(\Omega_o)$ norm facilitates the application of efficient function space based solution algorithms such as semi-smooth Newton methods, [248, 257]. A solution to the original problem can then be obtained by applying a continuation strategy for driving the regularization parameter ε to 0 as outlined in [208].

In this thesis we follow a different route and consider a solution algorithm for the original problem based on the sequential addition of new Dirac delta functions. This is motivated by the method presented in [50]. We identify the algorithm as a generalization of the well-known conditional gradient method due to Frank and Wolfe, [112]. While its implementation is simple, the characteristic slow convergence behavior of first-order optimization methods diminishes its practical utility. We present an accelerated version of this algorithm alternating between inserting new Dirac delta functions and optimizing the associated coefficients. This new *Primal-Dual-Active-Point* algorithm is reminiscent of the methods presented in [44, 50]. However, we are the first to provide improved convergence rates if additional structural requirements on the problem are met. The derived results are not restricted to optimal sensor placement but also hold for far more general problems involving vector-valued measures as optimization variable. We point out to Section 6.3. Due to the tight connection between measure-valued optimization problems and state constraints, these new results also shed some new light on classical algorithms such as the exchange method in semi-infinite optimization, see [97, 278] and Example 6.9. We also compare the new method to the aforementioned continuation strategy and conditional gradient methods without acceleration in order to highlight its practical efficiency.

This thesis is structured as follows. In the first chapter, Chapter 2, we provide a more profound and mathematical introduction to inverse problems for parameter identification and the difficulties caused by measurement errors. We sketch approaches to quantify the uncertainty induced by random perturbations of the measurement data on approximate solutions of the inverse problem. Finally, this reasoning leads to the formulation of sensor placement problems based on the Fisher information operator in order to mitigate their influence and to obtain reliable estimates for the unknown parameter.

In Chapter 3, we formulate the task of optimally planning measurement experiments as minimization problem in the space of Borel measures over the set Ω_o. Well-posedness of sparse sensor placement problems as well as necessary and sufficient first-order optimality conditions are in the focus of Section 3.2.3. Based upon these results, we derive structural properties of optimal design measures in Section 3.2.4. Sufficient conditions for the existence of sparse solutions consisting of finitely many Dirac delta functions are discussed. In particular, if the unknown parameter is finite dimensional, optimal design measures with support size bounded in dependence of the dimension of the parameter space exist.

Optimal sensor placement for the inverse problem of identifying a finite dimensional parameter entering into a PDE is in the focus of Chapter 4. Inference on its true value is possible based on finitely many pointwise measurements of the associated state variable. In Section 4.2 we discuss this setting in the context of the proposed sparse minimization framework with $G(\|u\|_{\mathcal{M}}) = \beta\|u\|_{\mathcal{M}}$. Section 4.3 draws a parallel between sparse sensor placement for finite dimensional parameters and semi-infinite optimization problems. Section 4.4 is devoted to the formulation and analysis of a numerical solution algorithm based on the sequential insertion of a new Dirac delta function into the iterated design measure. Worst-case convergence results are concluded from the discussions in Chapter 6. By augmenting the method with an additional post-processing step, convergence of the iterated design measures towards a sparse optimal one can be shown. Moreover, we provide an accelerated version of the algorithm for which improved convergence results can be proven under suitable conditions. Stability and sensitivity analysis for sparse sensor placement problems is presented in Section 4.5. For the practical computation of optimal design measures, the underlying PDE is discretized by linear finite elements. We derive estimates for the discretization error in the cost functional as well as the Fisher information matrix associated to optimal designs. Furthermore, a priori error estimates for the optimal positions of measurement sensors and their optimal diligence factors are derived under additional structural assumptions on the problem. Numerical examples confirm their optimality. To the best of our knowledge, we are not aware of any comparable results. We point out that most of the results in Sections 4.1 and 4.2, Section 4.4, with exception of Sections 4.4.3 and 4.4.5, as well as Section 4.6.1 and the numerical examples of Sections 4.7.1 and 5.4.2 are contained in similar form in the scientific paper [200] which is loosely based on the author's master thesis, [262].

Chapter 5 deals with sparse optimal sensor placement in the context of infinite-dimensional Bayesian inversion with partial differential equations. Again, it is assumed that finitely many pointwise measurements of the state variable are available in order to infer on an unknown distributed function entering into a PDE. The prior uncertainty on a suitable value for this function is taken into account by modeling the unknown parameter as a Gaussian random field. Since Bayesian inversion and optimal design of experiments in this context are a (relatively) new and currently very active area of research, the first part of Section 5.1 provides a concise introduction to these topics. In Section 5.1.4 we embed Bayesian optimal sensor placement into the sparse minimization framework of Chapter 3. As for finite dimensional parameter spaces, we proceed with the presentation of a discretization framework based on a finite element surrogate for the PDE. Furthermore, the infinite-dimensional parameter space is replaced by a finite-dimensional subspace spanned by several eigenfunctions of the prior covariance operator. A priori estimates for the error in the cost functionals as well as the optimal Fisher information operators due to the finite element discretization as well as the approximation of the parameter space are presented. An extension of the sequential point insertion algorithm from the previous chapter to the present setting as well as its efficient numerical realization for a particular choice of the design criterion is in the focus of Section 5.3. Numerical experiments in Section 5.4 highlight its practical efficiency.

In the last part of this thesis, Chapter 6, we take a closer look at efficient solution algorithms for sparse minimization problems in the function space setting. For this purpose, we essentially proceed in two steps. We recall that the space of Borel measure on Ω_o can be identified as the topological dual of the separable Banach space of continuous functions on this set. In the first part of the chapter, we embed sparse minimization into a more general setting and consider the more abstract task of minimizing the sum of a smooth but not necessarily convex function and a convex regularizer over the topological dual space of a separable Banach space. Besides sparse minimization, this composite minimization framework also encompasses e.g. bang-bang and minimum-effort control problems as well as optimization problems in the space of functions with bounded total variation. In Section 6.2 we propose an iterative solution algorithm for this type of problem based on a generalization of the well-known conditional gradient method. This procedure generates a new iterate by taking a convex combination between the current iterate and a solution to a certain (partially) linearized problem. We discuss (subsequential) convergence of the generated iterates towards stationary points of the problem and derive worst-case convergence rates if the smooth part is convex and fulfills additional regularity assumptions. More in detail, we obtain sublinear convergence of the objective function values of the generated iterates towards the global minimum of the problem which is characteristic for first-order optimization methods. This result is sharp. Examples point out to possible applications for the method. The second part of the chapter, Section 6.3, is devoted to the adaption of the generalized conditional gradient method to certain composite minimization problems in spaces of vector-valued measures. It turns out that the linearized subproblems admit solutions supported on a single point. In particular, the method may be realized such that all iterates are comprised of finitely many Dirac delta functions. We discuss augmentations of the methods which e.g. guarantee sparsity of the iterates as well as of the approximated solutions in certain cases. Most important, we propose an accelerated version of the method, the *Primal-Dual-Active-Point* algorithm, which alternates between adding a new Dirac-Delta function in each iteration and optimizing the coefficients of all Diracs in the iterate. For this specific version of the algorithm, we are able to prove a linear rate of convergence for the objective function values as well as linear convergence for the sequence of iterates in certain dual norms if additional structural assumptions on the problem hold. This last chapter is based on the paper [211] which will be soon submitted to a scientific journal. The results of Section 6.3.3 are taken from [209].

2 From inverse problems to optimal sensor placement

This chapter of the thesis constitutes a brief introduction to the mathematical concept of inverse problems. Moreover, it serves as a motivation for and a bridge to the sensor placement problems considered in the remainder of this thesis. At first sight, the reasoning behind inverse problems is simple: Assume that we have given a family of partial differential equations which depend on an unknown parameter. To each equation, we can compute a solution which resembles physical quantities, such as fluxes, concentrations or pressure. In order to select the most suitable mathematical model for the simulation of such phenomena, we take a finite number of measurements on these quantities in an experiment. In the inverse problem, we now provide a mathematical model for the measurement process. Its solution is then given by one particular partial differential equation, or more precisely the associated parameter, whose solution minimizes the misfit between the obtained measurements from the experiment and those predicted by the measurement model. In practice, the solution of inverse problems is aggravated by several factors. For example, the amount of provided measurements may not suffice to uniquely identify the unknown parameter. Moreover, in most practical situations, measurements are subject to perturbations stemming back to the imperfectness of the applied sensors. These defects add an additional bias to the problem which has to be properly addressed. Clearly, these shortcomings are intimately related to the setup of the experiment in which the measurements are collected. This observation suggests that we can e.g. mitigate the influence of measurement errors on the solution of the inverse problem by a sophisticated design of the experiment. In particular, this is possible by optimizing the arrangement of measurement sensors and the overall number of performed measurements. In the following chapter, we outline this reasoning mathematically and consider the task of optimal sensor placement as a minimization problem based on the so-called Fisher information operator of a suitable linearized parameter estimator. Similar formulations will then form the basis for the abstract sensor placement framework presented in the remainder of this thesis. For a profound introduction to the vast topic of inverse problems we refer to [16, 21, 101, 155]. Optimal sensor placement based on the Fisher information of the estimates dates back to the works of Smith, [245] and Kiefer, [165] for linear regression. Extensions of these methods to nonlinear models are based on linearization cf. [217]. An overview on comparable approaches for inverse problems with partial differential equations is given in the monograph [256].

2.1 Notation and function spaces

We briefly introduce some of the notation used throughout this thesis. By $\mathbb{R}$, $\mathbb{R}_+$ we refer to the real and nonnegative real numbers, respectively. The letters $\mathbb{N}$, $\mathbb{Z}$ denote the natural and whole numbers. The euclidean inner product on $\mathbb{R}^n$, $n \in \mathbb{N}$, is denoted by $(\cdot, \cdot)_{\mathbb{R}^n}$. The euclidean norm is given by $|\cdot|_{\mathbb{R}^n}$. Given two Banach spaces X and Y with Banach space norms $\|\cdot\|_X$ and $\|\cdot\|_Y$ as well as a linear mapping $B\colon X \to Y$, we define the operator norm of B as

$$\|B\|_{\mathcal{L}(X,Y)} = \sup_{\|\varphi\|_X=1} \|B\varphi\|_Y.$$

The vector space

$$\mathcal{L}(X,Y) := \left\{\, B \mid B\colon X \to Y \text{ linear},\ \|B\|_{\mathcal{L}(X,Y)} < \infty \,\right\}$$

forms a Banach space together with the operator norm $\|\cdot\|_{\mathcal{L}(X,Y)}$. Furthermore, the topological dual space $\mathcal{L}(X,\mathbb{R})$ of the Banach space X is denoted by X^*. The associated duality pairing is given by $\langle\cdot,\cdot\rangle_{X,X^*}$. The adjoint operator to $B \in \mathcal{L}(X,Y)$ is denoted by $B^* \in \mathcal{L}(Y^*,X^*)$. By $\operatorname{Ker} B$ and $\operatorname{Im} B$ we further refer to the kernel and range of B, respectively. Given a Banach space X, an extended real valued functional $\phi\colon X \to \mathbb{R}\cup\{\,+\infty\,\}$ and a convex subset $M \subset X$ we define the domain of ϕ in M as

$$\operatorname{dom}_M \phi = \{\, u \in M \mid \phi(u) < \infty \,\}\,.$$

The convex indicator function of M is denoted by I_M.

Let Ω be a nonempty set, $\mathcal{F}$ a σ-algebra on Ω and μ a nonnegative measure on the measurable space $(\Omega,\mathcal{F})$. The triple $(\Omega,\mathcal{F},\mu)$ is called a measure space. If $\mu(\Omega)=1$ we speak of a probability space. By $\mathcal{B}(\Omega)$ we denote the Borel σ-algebra on Ω. The Lebesgue measure on $\mathbb{R}^d$, $d\in\mathbb{N}$, is denoted by μ_L and the Lebesgue σ-algebra is $\mathcal{L}(\mathbb{R}^d)$. Let $\Omega \subset \mathbb{R}^d$. The classical Lebesgue spaces $L^p(\Omega)$, $p\in[1,\infty]$, are defined as the space of Lebesgue measurable functions (interpreted in the almost everywhere sense) with finite norm

$$\|\varphi\|_{L^p(\Omega)} = \begin{cases} \left(\int_\Omega |\varphi|^p \,\mathrm{d}\mu_L\right)^{1/p} & p\in[1,\infty) \\ \operatorname{ess\,sup}_{x\in\Omega} |\varphi(x)| & p=\infty. \end{cases}$$

Occasionally, we write

$$\int_O \varphi \,\mathrm{d}\mu_L = \int_O \varphi(x)\,\mathrm{d}\mu_L(x) \quad O\in\mathcal{L}(\Omega),$$

to stress the argument of the integrand. We proceed in the same way for integration of suitable functions with respect to general measures. If it is clear from the context that integration is understood with respect to the Lebesgue measure, we further write $\int_O f\,\mathrm{d}\mu_L = \int_O f\mathrm{d}x$. By $H^1(\Omega)$, we refer to the usual Sobolev space of functions $\varphi\in L^2(\Omega)$ admitting square integrable weak partial derivatives. If we incorporate additional zero boundary conditions (in the trace sense) into the space, we write $H^1_0(\Omega)$.

A measure μ_E on $\mathbb{R}^n$, $n\in\mathbb{N}$, is called a Gaussian measure if there exist $x_0\in\mathbb{R}^N$ and a positive definite matrix $\Sigma\in\mathbb{R}^{n\times n}$ with

$$\mu_E(O) = \frac{1}{\mathcal{Z}}\int_O \exp\left(-\frac{1}{2}(x-x_0,\Sigma^{-1}(x-x_0))_{\mathbb{R}^n}\right)\,\mathrm{d}x \quad \forall O\in\mathcal{B}(\mathbb{R}^n).$$

The scalar $\mathcal{Z}>0$ is a normalization constant ensuring $\mu_E(\mathbb{R}^n)=1$. The vector x_0 is called the mean and Σ is the covariance matrix of μ_E. We adopt the usual convention and write $\mu_E = \mathcal{N}(x_0,\Sigma)$.

2.2 The inverse problem

In order to formulate the inverse problem, we first have to elaborate on the underlying mathematical model and the measurement procedure. For this purpose, let Q denote a separable Hilbert space of parameters with inner product $(\cdot,\cdot)_Q$ and induced norm $\|\cdot\|_Q$. Moreover, by Y and W we refer to the so-called state and test space, respectively. Both are assumed to be reflexive Banach spaces. The duality pairing between Y and its dual space Y^* is denoted by $\langle\cdot,\cdot\rangle_{Y,Y^*}$. Analogously we proceed with W. Last, we consider a mapping $A\colon Q\times Y\to W^*$ describing a parametrized family of differential operators. For a given parameter $q\in Q$ an element $y\in Y$ is called an associated state if

$$A(q,y)=0 \quad \text{in } W^*. \tag{2.1}$$

As commonly done we introduce a semi-linear form as

$$a\colon Q\times Y\times W\to\mathbb{R},\quad (q,y,\varphi)\mapsto\langle\varphi,A(q,y)\rangle_{W,W^*}$$

In the following, we write $a(\cdot,\cdot)(\cdot)$ to stress the, in general, nonlinear dependence of this weak form on its first two arguments while it depends linearly on the element in the second bracket. Now, we reformulate the partial differential equation in (2.1) as a variational problem: Given a parameter $q\in Q$ we search for an element $y\in Y$ fulfilling

$$a(q,y)(\varphi)=0 \quad \forall\varphi\in W. \tag{2.2}$$

Next, we give a mathematical description of the measurement process. To this end consider a compact set $\Omega_o\subset\mathbb{R}^d$, $d\in\mathbb{N}$. We will refer to Ω_o as the candidate set of possible sensor locations. On this subset, we assume the existence of a strongly continuous mapping

$$O\colon\Omega_o\to Y^*$$

where $O(x)\in Y^*$ models the action of a measurement sensor located at a spatial point $x\in\Omega_o$ on the state variable y. We give some examples to clarify this abstract definition.

Example 2.1. *Let Ω be a convex and bounded domain in $\mathbb{R}^d$, $d\le 3$, and $\Omega_o\subset\Omega$ a compact subset. First we discuss pointwise measurements of a state variable y in the Sobolev space $H^2(\Omega)$ of functions $y\in L^2(\Omega)$ admitting square integrable weak derivatives up to order two. For a given $x\in\Omega_o$ the associated point evaluation of the state $y\in H^2(\Omega)$ is realized by the duality pairing with the associated Dirac delta function δ_x. Clearly, δ_x defines a linear and continuous functional on $H^2(\Omega)$ since*

$$\langle y,\delta_x\rangle_{H^2(\Omega),H^2(\Omega)^*}=y(x)\le\|y\|_{\mathcal{C}}\le\|y\|_{H^2(\Omega)}$$

due to the continuous embedding of $H^2(\Omega)$ into the space of continuous functions $\mathcal{C}(\Omega_o)$. Accordingly, define the measurement mapping

$$O_1\colon\Omega_o\to H^2(\Omega)^*,\quad\langle y,O_1(x)\rangle_{H^2(\Omega),H^2(\Omega)^*}=y(x)$$

We check that O is indeed a strongly continuous function. To this end, let a sequence $\{x_k\}_{k\in\mathbb{N}}\subset\Omega_o$ with $x_k\to x$ be given. Denoting by $\mathcal{M}(\Omega)\simeq\mathcal{C}(\Omega_o)^*$ the space of Borel measures on Ω_o, we readily verify that the sequence $\{O_1(x_k)\}_{k\in\mathbb{N}}$ converges with respect to the weak topology on $\mathcal{M}(\Omega_o)$ i.e.*

$$\langle\varphi,O_1(x_k)\rangle_{\mathcal{C}(\Omega_o),\mathcal{M}(\Omega_o)}=\varphi(x_k)\Rightarrow\varphi(x)=\langle\varphi,O_1(x)\rangle_{\mathcal{C}(\Omega_o),\mathcal{M}(\Omega_o)}\quad\forall\varphi\in\mathcal{C}(\Omega_o).$$

Since the space $H^2(\Omega)$ embeds compactly into $\mathcal{C}(\Omega_o)$ we also have $\mathcal{M}(\Omega_o) \overset{c}{\hookrightarrow} H^2(\Omega)^*$. Thus we conclude $\delta_{x_k} \to \delta_x$ strongly in $H^2(\Omega)^*$.

As a second example consider averaged measurements of the state variable over balls with fixed radius around a spatial point $x \in \Omega_o$. The state space is given by the $Y = L^\infty(\Omega)$. The measurement mapping is defined as

$$O_2 \colon \Omega_o \to L^\infty(\Omega), \quad \langle y, O_2(x)\rangle_{L^\infty(\Omega),L^\infty(\Omega)^*} = \frac{1}{\mu_L(B_R(x))} \int_{B_R(x)\cap\Omega} y \,\mathrm{d}\mu_L,$$

for some $R > 0$. Again, it is easy to see that $O_2(x)$ defines a linear and continuous functional on $L^\infty(\Omega)$ for all $x \in \Omega_o$. Moreover observe that for any convergent sequence $\{x_k\}_{k\in\mathbb{N}}$, $x_k \to x$, there holds

$$\begin{aligned} \|O_2(x_k) - O_2(x)\|_{L^\infty(\Omega)^*} &= \sup_{\|\varphi\|_{L^\infty(\Omega)}\leq 1} |\langle \varphi, O_2(x_k) - O_2(x)\rangle_{L^\infty(\Omega),L^\infty(\Omega)^*}| \\ &\leq |\mu_L(B_R(x) \setminus B_R(x_k)) + \mu_L(B_R(x_k) \setminus B_R(x))| \end{aligned}$$

where the right hand side tends to zero as $k \to \infty$. Hence, O_2 is a strongly continuous function.

A priori knowledge on the structure of the unknown parameter is incorporated by restricting the parameter space to an admissible set $Q_{ad} \subset Q$. The inverse problem is now formulated as follows: Given a vector of measurements $\mathbf{y}_d = (\mathbf{y}_d^1, \dots, \mathbf{y}_d^N)^\top \in \mathbb{R}^N$, $N \in \mathbb{N}$, collected at a finite number of distinct sensor locations $\{x_i\}_{i=1}^N$, find a pair $(q, y) \in Q_{ad} \times Y$ fulfilling the system of equations defined as

$$\langle y, O(x_i)\rangle_{Y,Y^*} = \mathbf{y}_d^i, \; i = 1, \dots, N, \quad a(q, y)(\varphi) = 0 \quad \forall \varphi \in W. \tag{2.3}$$

Now, we lay the focus on the solution of the inverse problem and the accompanying difficulties. For this purpose, the following assumptions on the solvability of the underlying partial differential equation are made.

Assumption 2.1. For every $q \in Q_{ad}$ there exists a unique element $y \in Y$ fulfilling (2.2). Furthermore, the parameter-to-state mapping $S \colon Q_{ad} \to Y$ given by

$$S \colon Q_{ad} \to Y, \quad q \mapsto y = S[q]$$

is at least continuously Fréchet differentiable with respect to the norm on Q in a neighborhood of the admissible set Q_{ad}. The Fréchet derivative of S is denoted by $\partial S \colon Q \to \mathcal{L}(Q, Y)$.

If the constituting operator $A \colon Q \times Y \to W^*$ is Fréchet differentiable, so is the induced form and there holds

$$a_y'(y, q)(\delta y, \varphi) = \langle \varphi, A_y'(q, y)\delta y\rangle_{W,W^*}, \quad a_q'(y, q)(\delta q, \varphi) = \langle \varphi, A_q'(q, y)\delta q\rangle_{W,W^*}$$

for $q, \delta q \in Q_{ad}$, $y, \delta y \in Y$ and $\varphi \in W$, respectively. By definition of S we further observe that

$$a(q, S[q])(\varphi) = 0 \quad \forall q \in Q_{ad}, \; \varphi \in W.$$

Taking the total derivative with respect to q in the above equation, we conclude that $\partial S[\hat{q}]\delta q \in Y$ for $\hat{q} \in Q_{ad},\ \delta q \in Q$, fulfills the linearized state equation

$$a'_y(\hat{q}, S[\hat{q}])(\partial S[\hat{q}]\delta q, \varphi) = -a'_q(\hat{q}, S[\hat{q}])(\delta q, \varphi) \quad \forall \varphi \in W.$$

This relation between the parameter and the state variable allows to eliminate the partial differential differential equation as an explicit constraint. We arrive at the reduced formulation

$$\text{find } q \in Q_{ad}\colon \quad \langle S[q], O(x_i)\rangle_{Y,Y^*} = \mathbf{y}_d^i,\ i = 1, \dots, N.$$

Moreover, in the following we assume the availability of a sophisticated a priori guess $\hat{q} \in Q_{ad}$ on the parameter value describing the modeled process most faithfully and the parameter-to-state operator S is well-approximated by a first order approximation around it i.e.

$$S[q] \approx S[\hat{q}] + \partial S[\hat{q}](q - \hat{q}) \quad \forall q \in Q_{ad}.$$

Now, we may also drop the constraints on the admissible set of parameters and consider the linearized inverse problem given by

$$\text{find } q \in Q\colon \quad \langle S[\hat{q}], O(x_i)\rangle_{Y,Y^*} + \langle \partial S[\hat{q}](q - \hat{q}), O(x_i)\rangle_{Y,Y^*} = \mathbf{y}_d^i,\ i = 1, \dots, N. \tag{2.4}$$

Let us introduce some additional notation to rewrite this problem in a more compact way. In order to do so, observe that

$$\langle \partial S[\hat{q}]q, O(x_i)\rangle_{Y,Y^*} = (\partial S[\hat{q}]^* O(x_i), q)_Q \quad \forall q \in Q,\ i = 1, \dots, N,$$

where $\partial S[\hat{q}]^*\colon Y^* \to Q$ denotes the Banach space adjoint of $\partial S[\hat{q}]$. Accordingly, we now introduce the reduced measurement mapping

$$\mathcal{O}\colon \Omega_o \to Q, \quad x \mapsto \partial S[\hat{q}]^* O(x),$$

and the (linearized) parameter-to-observations map $X \in \mathcal{L}(Q, \mathbb{R}^N)$ as

$$X\colon Q \to \mathbb{R}^N, \quad (Xq)_i = (\mathcal{O}(x_i), q)_Q, \quad i = 1\dots, N.$$

Last, define the vector $S[\hat{q}](x) \in \mathbb{R}^N$ with $S[\hat{q}](x)_i = \langle S[\hat{q}], O(x_i)\rangle_{Y,Y^*}$ for $i = 1, \dots, N$. We assemble all N equations from (2.4) in one system to equivalently reformulate the linearized inverse problem as

$$\text{find } q \in Q\colon \quad Xq = X\hat{q} + \mathbf{y}_d - S[\hat{q}](x). \tag{2.5}$$

While this is a linear equation for the unknown parameter q its solution is in no way straightforward and has to be handled with care. To highlight this fact, recall the notion of well-posedness due to Hadamard: A mathematical problem is well-posed in the sense of [129] if it admits a unique solution for all admissible input data. Moreover, the obtained solution has to depend continuously on the data. Translating this definition to the present case, a well-posed linearized inverse problem of the form (2.5) admits a unique solution for every vector of measurements $\mathbf{y}_d \in \mathbb{R}^N$. Clearly, these conditions are violated in most cases. On the one hand, we are often interested in identifying high or even infinite dimensional parameters but only a small number of measurements is available. In these underdetermined cases, i.e. if $\dim Q > N$, the inverse problem in (2.4) may provide an infinite number of solutions since the kernel of X is nonempty. On the other hand, the problem admits no solution if

$$X\hat{q} + \mathbf{y}_d - S[\hat{q}](x) \notin \operatorname{Im} X.$$

Note that the non-existence of a continuous inverse to X causes severe problems in practical applications. There the measurement vector $\mathbf{y}_d$ corresponds to data obtained by performing measurements in a real experiment. In this case, it is reasonable to assume that the true data is perturbed by measurement noise stemming back to the imperfectness of the utilized sensor. However, due to the ill-posed nature of the equation in (2.5), small changes in the input data can cause the non-existence of a solution to the perturbed problem or provide solutions that are far away from the unperturbed one.

This observation suggests that the computation of a solution to the inverse problem without accounting for the aforementioned difficulties leads to a severe misinterpretation of the obtained estimates. Ultimately, this results in wrong conclusions on the most suitable choice for the unknown parameter. In order to allow for a stable solution of the problem, we resort to so-called regularization techniques. One particularly famous method for this task is given by (weighted) Tikhonov regularization, [252]. In this approach, an approximate solution to the inverse problem is obtained by solving the regularized Least-Squares problem

$$\min_{q\in Q} J(q,\mathbf{y}_d) := \left[\frac{1}{2}|\Sigma^{-1/2}(S[\hat{q}](x) + X(q-\hat{q}) - \mathbf{y}_d)|^2_{\mathbb{R}^N} + \frac{1}{2}\|\mathcal{I}_0^{1/2}(q-\hat{q})\|^2_Q\right].$$

Here we minimize the trade-off between the misfit of the measurement data and a regularization term that quantifies the distance of the parameter to the linearization point. Note that both terms incorporate weighted Hilbert space norms induced by a matrix $\Sigma^{-1/2} \in \mathbb{R}^{N\times N}$ and an operator $\mathcal{I}_0^{1/2} \colon Q \to Q$, respectively. For example, these allow to put special emphasis on the measurement obtained by a specific sensor or to enhance expected structural features in the approximate solution to the inverse problem. In particular, the operator $\mathcal{I}_0^{1/2}$ can be unbounded on Q. Its Q-domain $\mathcal{Q} \subset Q$ given by

$$\mathcal{Q} = \left\{ q \in Q \mid \|\mathcal{I}_0^{1/2} q\|_Q < \infty \right\}$$

is a, possibly proper, subspace of Q. The following assumptions are made.

Assumption 2.2. The matrix $\Sigma^{-1/2} \in \mathbb{R}^{N\times N}$ is positive definite and $\mathcal{I}_0^{1/2}$ is a closed linear operator. Moreover there holds $\hat{q} \in \mathcal{Q}$.

Note that $J(q,\mathbf{y}_d) = +\infty$ for all $q \in Q \setminus \mathcal{Q}$. Thus, the search for a minimizer of $J(\cdot,\mathbf{y}_d)$ can be a priori restricted to the domain of $\mathcal{I}_0^{1/2}$. We can define a Hilbert space structure on $\mathcal{Q}$ with respect to the graph norm $\|\cdot\|_{\mathcal{Q}}$ induced by the inner product

$$(\cdot,\cdot)_{\mathcal{Q}} = (\cdot,\cdot)_Q + (\mathcal{I}_0^{1/2}\cdot, \mathcal{I}_0^{1/2}\cdot)_Q.$$

This is a consequence of the closedness assumption on $\mathcal{I}_0^{1/2}$. In general, $\mathcal{Q}$ will not be identified with its topological dual space denoted by $\mathcal{Q}^*$. By definition of the graph norm there holds

$$\mathcal{Q} \hookrightarrow Q \simeq Q^* \hookrightarrow \mathcal{Q}^*.$$

Thus, instead of finding a solution to an ill-posed system of equations, we now compute a minimizer to an optimization problem. Imposing additional assumptions, its unique minimizer can be given in closed form.

Proposition 2.1. *Define* $\Sigma^{-1} = \Sigma^{-1/2}\Sigma^{-1/2}$ *and* $\mathcal{I}_0 = (\mathcal{I}_0^{1/2})^*(\mathcal{I}_0^{1/2})$. *Moreover, assume that*

$$\operatorname{Ker}\mathcal{I}_0^{1/2} \cap \operatorname{Ker} X = \{0\} \quad \textit{and} \quad \operatorname{Im}(X^*\Sigma^{-1}X + \mathcal{I}_0) = \mathcal{Q}^*. \tag{2.6}$$

Then the linear and continuous operator

$$X^*\Sigma^{-1}X + \mathcal{I}_0 \colon \mathcal{Q} \to \mathcal{Q}^* \tag{2.7}$$

admits a linear and continuous inverse

$$(X^*\Sigma^{-1}X + \mathcal{I}_0)^{-1} \colon \mathcal{Q}^* \to \mathcal{Q}.$$

Last, denote by $\mathbf{y}_d \in \mathbb{R}^N$ *an arbitrary vector of measurements. Then the unique minimizer of* $J(\cdot, \mathbf{y}_d)$ *over* $\mathcal{Q}$ *is given by*

$$q^{\mathbf{y}_d} = \hat{q} + (X^*\Sigma^{-1}X + \mathcal{I}_0)^{-1}(X^*\Sigma^{-1}(\boldsymbol{y}_d - S[\hat{q}](x))). \tag{2.8}$$

Proof. Let us first check that the operator from (2.7) is indeed linear and continuous with respect to the correct norms. For this purpose, consider an arbitrary element $q \in \mathcal{Q}$. By definition of the dual norm we obtain

$$\begin{aligned}\|(X^*\Sigma^{-1}X + \mathcal{I}_0)q\|_{\mathcal{Q}^*} &= \sup_{\|\tilde{q}\|_{\mathcal{Q}} \le 1} \langle \tilde{q}, (X^*\Sigma^{-1}X + \mathcal{I}_0)q\rangle_{\mathcal{Q},\mathcal{Q}^*} \\ &= (\Sigma^{-1/2}X\tilde{q}, \Sigma^{-1/2}Xq)_{\mathbb{R}^N} + (\mathcal{I}_0^{1/2}\tilde{q}, \mathcal{I}_0^{1/2}q)_Q \\ &\le (\|\Sigma^{-1/2}X\|^2_{\mathcal{L}(\mathcal{Q},\mathbb{R}^N)} + 1)\|q\|_{\mathcal{Q}}.\end{aligned}$$

Its linearity is obvious. Furthermore, this operator is injective since

$$\langle q, (X^*\Sigma^{-1}X + \mathcal{I}_0)q\rangle_{\mathcal{Q},\mathcal{Q}^*} = |\Sigma^{-1/2}Xq|^2_{\mathbb{R}^N} + \|\mathcal{I}_0^{1/2}q\|^2_Q > 0 \quad \forall q \in \mathcal{Q} \setminus \{0\}.$$

This holds true due to positive definiteness of the matrix $\Sigma^{1/2}$ and $\operatorname{Ker}\mathcal{I}_0^{1/2} \cap \operatorname{Ker} X = \{0\}$. Together with the surjectivity assumption from (2.6) we conclude the existence of its continuous inverse operator from the bounded inverse theorem.

We now calculate the Fréchet derivative of $J(\cdot, \mathbf{y}_d)$ with respect to the parameter at a given element $q \in \mathcal{Q}$. Applying the chain rule, we readily obtain

$$J'(q, \mathbf{y}_d) = X^*\Sigma^{-1}(S[\hat{q}](x) + X(q - \hat{q}) - \mathbf{y}_d) + \mathcal{I}_0(q - \hat{q}) \in \mathcal{Q}^*.$$

Since $J(\cdot, \mathbf{y}_d)$ is a convex functional an element $q^{\mathbf{y}_d}$ is a global minimizer of $J(\cdot, \mathbf{y}_d)$ on $\mathcal{Q}$ if and only if the Fréchet derivative $J'(q^{\mathbf{y}_d}, \mathbf{y}_d)$ vanishes. We make the ansatz

$$X^*\Sigma^{-1}(S[\hat{q}](x) + X(q - \hat{q}) - \mathbf{y}_d) + \mathcal{I}_0(q - \hat{q}) = 0 \in \mathcal{Q}^*$$

and solve this equation for q. Rearranging we get

$$(X^*\Sigma^{-1}X + \mathcal{I}_0)q = (X^*\Sigma^{-1}X + \mathcal{I}_0)\hat{q} + X^*\Sigma^{-1}(\mathbf{y}_d - S[\hat{q}](x)).$$

Inverting the operator on the left we finally conclude

$$q = \hat{q} + (X^*\Sigma^{-1}X + \mathcal{I}_0)^{-1}(X^*\Sigma^{-1}(\mathbf{y}_d - S[\hat{q}](x))).$$

Thus, by construction, the element $q^{\mathbf{y}_d}$ from (2.8) is the unique global minimizer of $J(\cdot, \mathbf{y}_d)$ over $\mathcal{Q}$. □

Remark 2.1. Let us briefly point out that the first condition in (2.6) implies the finite dimensionality of $\operatorname{Ker}\mathcal{I}_0^{1/2}$ and $\dim\operatorname{Ker}\mathcal{I}_0^{1/2} \le \dim(\operatorname{Im} X) \le N$. Moreover, if $\dim Q < \infty$, the second condition in (2.6) is redundant since the injectivity of the operator implies its surjectivity if Q is finite dimensional.

2.3 Uncertainty quantification & optimal design

As already remarked at an earlier point of this chapter, it is, especially from a practical point of view, necessary to discuss and quantify the influence of measurement errors on the obtained approximate solutions to the inverse problem. This topic is in the focus of the following considerations. We make the following assumption on the vector of measurements $\mathbf{y}_d \in \mathbb{R}^N$.

Assumption 2.3. There holds $\mathbf{y}_d = \mathbf{y}_d^\dagger + \varepsilon$, where $\mathbf{y}_d^\dagger \in \mathbb{R}^N$ denotes the unperturbed measurements and $\epsilon \in \mathbb{R}^N$ is a vector of measurement errors.

We emphasize that we can only observe the sum $\mathbf{y}_d$ of both terms i.e. the vectors $\mathbf{y}_d^\dagger$ and ε, respectively, are unknown to us. Classical approaches to the treatment of measurement noise in inverse problems are usually based on an a priori upper bound on the error i.e. $|\mathbf{y}_d - \mathbf{y}_d^\dagger|_{\mathbb{R}^N} \leq \delta$ for some $\delta > 0$. In practical applications, measured data is not exactly reproducible i.e. performing the exact same measurement twice usually leads to slightly different outcomes. These deviations stem back to the inability of the experimenter or the used sensor to take the measurement in the exact same way. Moreover, at rare occasions, measurement devices produce outliers i.e. the difference $|\mathbf{y}_d - \mathbf{y}_d^\dagger|_{\mathbb{R}^N}$ is considerably large. This observation suggests the use of a less restrictive, stochastic model for the measurement errors. For this purpose, we consider a probability space $(D, \mathcal{F}, \mathbb{P})$ and interpret the noise vector $\epsilon \in \mathbb{R}^N$ as realization of a random variable $\epsilon \colon D \to \mathbb{R}^N$. The following assumptions on its distribution are made.

Assumption 2.4. Let $\varepsilon \colon D \to \mathbb{R}^N$ be a N-dimensional Gaussian random variable distributed according to $\mu_E = \mathcal{N}(0, \varSigma)$. The components of ε are mutually independent i.e. the positive definite covariance matrix $\varSigma \in \mathbb{R}^{N \times N}$ is diagonal with $\varSigma_{ij} = \delta_{ij}/\mathbf{u}_i$, where δ_{ij} denotes the Kronecker delta and $\mathbf{u}_i > 0$, $i, j = 1, \ldots, N$.

In the following arguments, the weighting matrix for the measurement misfit term in the Least-Squares estimator is always chosen as the unique positive definite square root of the inverse to the noise covariance $\varSigma$. That is $\varSigma_{ij}^{-1/2} = \delta_{ij}\sqrt{\mathbf{u}_i}$, $i, j = 1, \ldots, N$. Given a vector $\mathbf{y}_d = \mathbf{y}_d^\dagger + \epsilon \in \mathbb{R}^N$ for a particular realization ϵ of the measurement noise, we may now rewrite the solution to the regularized Least-Squares problem as

$$\begin{aligned} q^{\mathbf{y}_d} &= \hat{q} + (X^* \varSigma^{-1} X + \mathcal{I}_0)^{-1} (X^* \varSigma^{-1} (\mathbf{y}_d - S[\hat{q}](x))) \\ &= \hat{q} + (X^* \varSigma^{-1} X + \mathcal{I}_0)^{-1} (X^* \varSigma^{-1} (\mathbf{y}_d^\dagger + \epsilon - S[\hat{q}](x))). \end{aligned}$$

By solving the regularized Least-Squares problem the uncertainty in the measurements is also propagated into the obtained approximate solutions. Thus, we should also adapt a probabilistic interpretation of the Least-squares solution and view $q^{\mathbf{y}_d}$ as a realization of the estimator

$$\bar{q} \colon D \to Q, \quad \omega \mapsto \hat{q} + (X^* \varSigma^{-1} X + \mathcal{I}_0)^{-1} (X^* \varSigma^{-1} (\mathbf{y}_d^\dagger + \varepsilon(\omega) - S[\hat{q}](x)))$$

which is a random variable taking values in the parameter space. An element $\mathbb{E}[\bar{q}] \in Q$ is called the mean of the estimator $\bar{q}$ if

$$\int_D (\delta q, \bar{q}(\omega))_Q \, \mathrm{d}\mathbb{P}(\omega) = (\delta q, \mathbb{E}[\bar{q}])_Q \quad \forall \delta q \in Q.$$

Accordingly, $\mathcal{C} \in \mathcal{L}(Q,Q)$ is called the covariance operator of $\bar{q}$ if

$$\int_D (\delta q_1, \bar{q}(\omega) - \mathbb{E}[\bar{q}])_Q (\delta q_2, \bar{q}(\omega) - \mathbb{E}[\bar{q}])_Q \, \mathrm{d}\mathbb{P}(\omega) = (\delta q_1, \mathcal{C}\delta q_2)_Q \quad \forall \delta q_1, \delta q_2 \in Q.$$

Obviously, these expressions are only meaningful if $\bar{q}\colon D \to Q$ satisfies appropriate integrability conditions. However, due to the affine linearity of $\bar{q}$ with respect to the measurement vector and our assumptions on the random variable ε it is readily verified that

$$\mathbb{E}[\bar{q}] = \hat{q} + (X^* \Sigma^{-1} X + \mathcal{I}_0)^{-1} (X^* \Sigma^{-1} (\mathbf{y}_d^\dagger - S[\hat{q}](x)))$$

and that the covariance operator of $\bar{q}$ is given by

$$\mathcal{C} = (X^* \Sigma^{-1} X + \mathcal{I}_0)^{-1} X^* \Sigma^{-1} X (X^* \Sigma^{-1} X + \mathcal{I}_0)^{-1}.$$

Moreover the expected deviation of $\bar{q}$ from its mean is represented by the trace of the covariance operator:

$$\mathbb{E}[\|\bar{q} - \mathbb{E}[\bar{q}]\|_Q^2] := \int_D \|\bar{q}(\omega) - \mathbb{E}[\bar{q}]\|_Q^2 \, \mathrm{d}\mathbb{P}(\omega) = \mathrm{Tr}_Q(\mathcal{C}) < \infty \quad \text{where} \quad \mathrm{Tr}_Q(\mathcal{C}) = \sum_{i \in \mathbf{I}} (\phi_i, \mathcal{C}\phi_i)_Q \tag{2.9}$$

for an arbitrary orthonormal basis $\{\phi_i\}_{i \in \mathbf{I}}$, $\mathbf{I} \subset \mathbb{N}$, of the parameter space Q.

In the following sections we are interested in quantifying the capability of the random variable $\bar{q}$ for estimating a particular parameter $q^* \in Q$. Second, our interest also lies in the convergence of the estimator in the vanishing noise limit i.e. if the measurement errors tend to zero in some suitable sense. Therefore we first point out that we cannot make useful probabilistic statements about the closedness of a particular realization $q^{\mathbf{y}_d}$ to q^* since such events occur with zero probability. However, we can quantify the expected deviation of $\bar{q}$ from q^* in the squared norm on Q. This corresponds to the so-called mean squared error between $\bar{q}$ and q^*.

Definition 2.1. Let a parameter $q^* \in Q$ be given. The mean squared error between $\bar{q}$ and q^* is defined as

$$\mathrm{MSE}(\bar{q}, q^*) = \mathbb{E}[\|\bar{q} - q^*\|_Q^2] := \int_D \|\bar{q}(\omega) - q^*\|_Q^2 \mathrm{d}\, \mathbb{P}(\omega).$$

This term admits the following alternative representation.

Proposition 2.2. *There holds*

$$\mathrm{MSE}(\bar{q}, q^*) = \|\mathbb{E}[\bar{q}] - q^*\|_Q^2 + \mathrm{Tr}_Q(\mathcal{C}).$$

Proof. By definition we have

$$\begin{aligned} \mathrm{MSE}(\bar{q}, q^*) &= \int_D \|\bar{q}(\omega) - q^*\|_Q^2 \, \mathrm{d}\mathbb{P}(\omega) \\ &= \int_D \left[\|\mathbb{E}[\bar{q}] - q^*\|_Q^2 + 2(\bar{q}(\omega) - \mathbb{E}[\bar{q}], \mathbb{E}[\bar{q}] - q^*)_Q + \|\bar{q}(\omega) - \mathbb{E}[\bar{q}]\|_Q^2 \right] \, \mathrm{d}\mathbb{P}(\omega) \end{aligned}$$

Since the first term no longer depends on ω there holds

$$\int_D \|\mathbb{E}[\bar{q}] - q^*\|_Q^2 \,\mathrm{d}\mathbb{P}(\omega) = \|\mathbb{E}[\bar{q}] - q^*\|_Q^2.$$

Furthermore, using the definition of the mean, the second term vanishes since

$$\int_D (\bar{q}(\omega) - \mathbb{E}[\bar{q}], \mathbb{E}[\bar{q}] - q^*)_Q \,\mathrm{d}\mathbb{P}(\omega) = (\mathbb{E}[\bar{q}] - \mathbb{E}[\bar{q}], \mathbb{E}[\bar{q}] - q^*)_Q = 0.$$

The statement now follows by combining these observations with (2.9). □

Let us give some interpretation to this result. We conclude that the mean squared error captures both, the difference between the expected value of the given estimator and the parameter q^* as well as the variability of the parameter estimator around its mean. As a consequence, a small mean squared error implies on the one hand that the expected value of $\bar{q}$ is close to q^*. On the other hand, we also deduce that realizations of $\bar{q}$ do not scatter significantly and are close to the mean $\mathbb{E}[\bar{q}]$ (and thus also q^*) with a high probability. To sum up these arguments, the mean squared error provides a suitable tool to assess the statistical quality of the parameter estimator $\bar{q}$ for the task of estimating q^*. In particular, if q^* corresponds to the exact value of the unknown parameter in the partial differential equations, i.e. the parameter corresponding to the most suitable mathematical model, the mean squared error provides a measure for the influence of the measurement errors on the obtained parameter estimates.

2.3.1 Overdetermined problems

Quantifying uncertainty

We first consider the identification of a finite dimensional parameter in $Q \simeq \mathbb{R}^n$ from overdetermined observations. That is $N \geq n$ and $X \in \mathbb{R}^{N\times n}$ fulfills $\dim(\operatorname{Im} X) = n$. Consequently, the matrix $X^*\Sigma^{-1}X$ is invertible due to the positive definiteness of Σ^{-1}. For simplification let us assume that no model error is present i.e. there holds

$$\min_{q\in\mathbb{R}^n} |S[\hat{q}](x) + X(q^* - \hat{q}) - \mathbf{y}_d^\dagger|_{\mathbb{R}^N} = 0.$$

Since the kernel of X is trivial, the solution q^* to this minimization problem is unique. In this case, we consider the maximum likelihood estimator, [110], which returns the most plausible parameter value given the measurement vector $\mathbf{y}_d$. It is recovered in the presented Tikhonov regularized setting through choosing $\mathcal{I}_0^{1/2} = 0$. By invertibility of $X^*\Sigma^{-1}X$ we obtain

$$\bar{q}\colon D \to \mathbb{R}^N, \quad \omega \mapsto \hat{q} + (X^*\Sigma^{-1}X)^{-1}(X^*\Sigma^{-1}(\mathbf{y}_d^\dagger + \varepsilon(\omega) - S[\hat{q}](x))).$$

Due to the assumption on the existence of q^* this estimator can be equivalently rewritten as

$$\bar{q}\colon D \to \mathbb{R}^N, \quad \omega \mapsto q^* + (X^*\Sigma^{-1}X)^{-1}X^*\Sigma^{-1}\varepsilon(\omega).$$

Its mean is given by $\mathbb{E}[\bar{q}] = q^*$ i.e. this estimator is unbiased. The associated covariance operator $\mathcal{C}$ is obtained as

$$\mathcal{C} = (X^*\Sigma^{-1}X + \mathcal{I}_0)^{-1}X^*\Sigma^{-1}X(X^*\Sigma^{-1}X + \mathcal{I}_0)^{-1} = (X^*\Sigma^{-1}X)^{-1}.$$

We recall that our primary motivation to consider weighted Least-Squares problems was given by the ill-posedness of (2.5) and the appearance of measurement errors. These defects prevented a stable solution of the problem. Thus, we have to address if and in which sense stability of the maximum likelihood estimator can be expected. For this purpose, we consider the vanishing noise case i.e. the variance $1/\mathbf{u}_i$ of each measurement tends to zero. This condition implies that $\Sigma \to 0$ and that the measurement noise $\varepsilon \sim (0, \Sigma)$ converges to 0 in probability.

Calculating the mean squared error between $\bar{q}$ and q^* reveals

$$\operatorname{MSE}(\bar{q}, q^*) = |\mathbb{E}[\bar{q}] - q^*|^2_{\mathbb{R}^n} + \operatorname{Tr}_{\mathbb{R}^n}((X^*\Sigma^{-1}X)^{-1}) = \operatorname{Tr}_{\mathbb{R}^n}((X^*\Sigma^{-1}X)^{-1}). \tag{2.10}$$

Now we estimate

$$\operatorname{Tr}_{\mathbb{R}^n}((X^*\Sigma^{-1}X)^{-1}) \leq n\|(X^*\Sigma^{-1}X)^{-1}\|_{\mathbb{R}^{n\times n}} \leq \frac{n\|\Sigma\|_{\mathbb{R}^{N\times N}}}{|Xq|^2_{\mathbb{R}^N}}$$

for some $q \in \mathbb{R}^N$ with $Xq \neq 0$, $|q|_{\mathbb{R}^n} = 1$, independent of Σ. As a consequence, we conclude that the maximum likelihood estimator is stable in the mean square sense i.e.

$$\max_{i=1,\ldots,N} 1/\mathbf{u}_i \to 0 \Rightarrow \operatorname{MSE}(\bar{q}, q^*) \to 0.$$

Besides the stability of $\bar{q}$, these arguments also highlight that the mean squared error provides a suitable stochastic tool to quantify the uncertainty on the true parameter q^* caused by the measurement errors in the estimation process. Moreover, its computation can be done without knowledge of q^*.

Remark 2.2. It is worthwhile to note that similar stability results also hold for the case of additional modelling errors i.e. there holds

$$X\hat{q} + \mathbf{y}_d^\dagger - S[\hat{q}](x) \notin \operatorname{Im} X.$$

We briefly outline these ideas. For this purpose, consider a parametrized family of measurement noises $\varepsilon_\sigma \sim \mathcal{N}(0, \Sigma_\sigma)$ where $\Sigma_\sigma = \sigma\widehat{\Sigma}$ for $\sigma > 0$ and some positive definite diagonal matrix $\widehat{\Sigma} \in \mathbb{R}^{N\times N}$. The associated parametrized maximum likelihood estimator is given by

$$\bar{q}_\sigma \colon D \to \mathbb{R}^N, \quad \omega \to \hat{q} + (X^*\widehat{\Sigma}^{-1}X)^{-1}(X^*\widehat{\Sigma}^{-1}(\mathbf{y}_d^\dagger + \varepsilon_\sigma(\omega) - S[\hat{q}](x)))$$

Note that the mean

$$q^\dagger := \mathbb{E}[\bar{q}_\sigma] = \hat{q} + (X^*\widehat{\Sigma}^{-1}X)^{-1}(X^*\widehat{\Sigma}^{-1}(\mathbf{y}_d^\dagger - S[\hat{q}](x)))$$

is independent of $\sigma > 0$ and corresponds to the unique solution of the Least-Squares problem

$$\min_{q\in Q} |\widehat{\Sigma}^{-1/2}(S[\hat{q}](x) + X(q-\hat{q}) - \mathbf{y}_d^\dagger)|^2_{\mathbb{R}^N}.$$

The mean squared error between $\bar{q}_\sigma$ and $q^\dagger$ fulfills

$$\operatorname{MSE}(\bar{q}_\sigma, q^\dagger) = |\mathbb{E}[\bar{q}_\sigma] - q^\dagger|^2_{\mathbb{R}^n} + \sigma\operatorname{Tr}_{\mathbb{R}^n}((X^*\widehat{\Sigma}^{-1}X)^{-1}) = \sigma\operatorname{Tr}_{\mathbb{R}^n}((X^*\widehat{\Sigma}^{-1}X)^{-1}) \to 0$$

as $\sigma \to 0$. Thus, the parametrized random variables $\{\bar{q}_\sigma\}_{\sigma>0}$ converge in the mean square sense towards the solution of a deterministic weighted Least-Squares problem for the unperturbed measurement vector $\mathbf{y}_d^\dagger$.

Optimal sensor placement

Summing up our previous discussions, the mean squared error of the maximum likelihood estimator provides a tool to quantify the deviation of the random variable $\bar{q}$ from the true parameter q^*. Moreover, it tends to zero if the variances of the measurement errors, $\varSigma_{ii} = 1/\mathbf{u}_i > 0$, are small. In this case, realizations of $\bar{q}$ will be close to q^* with high probability. In practical applications, a reduction of the measurement variances can be achieved by performing the same measurement several times and average the different outcomes. Alternatively, a single measurement can be performed with a better sensor. From a practical point of view, both of these possibilities are only viable to some extent since taking repeated measurements and constructing or buying better sensors is always associated with certain costs. Clearly, it is reasonable to assume that the overall monetary budget of an experiment to obtain the measurements is limited. As a consequence, an experimenter is usually interested in providing parameter estimates with small as possible mean squared error while simultaneously keeping the cost of the measurement process low. Following (2.10), the mean squared error between the estimator $\bar{q}$ and the true parameter q^* can be given in closed form as the trace of its covariance operator

$$\operatorname{MSE}(\bar{q}, q^*) = \operatorname{Tr}_{\mathbb{R}^n}((X^* \varSigma^{-1} X)^{-1}).$$

Note that this representation is independent of q^* and solely depends on the so-called Fisher information matrix $X^* \varSigma^{-1} X$ with matrices $X \in \mathbb{R}^{N \times n}$ and $\varSigma^{-1} \in \mathbb{R}^{N \times N}$ given by

$$(Xq)_i = (\mathcal{O}(x_i), q)_Q, \ \varSigma^{-1}_{ij} = \delta_{ij} \mathbf{u}_i \quad \forall q \in \mathbb{R}^n, \ i, j = 1, \dots, N.$$

This crucial observation implies that the mean squared error of the estimator cannot only be influenced by decreasing the variances of the measurements, i.e. by increasing $\mathbf{u}_i > 0$, $i = 1, \dots, N$, but also by a sophisticated choice of the sensor locations $\{x_i\}_{i=1}^N$ and the overall number of measurements N. In particular, we can a priori, i.e. before any measurements are carried out, improve the estimator by an optimal choice of the measurement setup. Mathematically, the task of optimal sensor placement can now be formulated as an optimization problem

$$\min_{x_i \in \varOmega_o, \mathbf{u}_i \in \mathbb{R}_+, N \in \mathbb{N}} [\operatorname{Tr}_{\mathbb{R}^n}((X^* \varSigma^{-1} X)^{-1}) + \mathcal{R}(\mathbf{u})] \quad s.t. \quad (Xq)_i = (\mathcal{O}(x_i), q)_Q, \ \varSigma^{-1}_{ij} = \delta_{ij} \mathbf{u}_i,$$

for all $q \in \mathbb{R}^n$, $i, j = 1, \dots, N$, where we minimize the mean squared error by parametrizing the Fisher information as a function of the number of measurements, their positions and the reciprocal of their variances. The regularization term $\mathcal{R}(\mathbf{u})$ captures the overall cost of the experiment based on the vector of measurement weights $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_N)^\top$. For general admissible sets $\varOmega_o$, we emphasize that this minimization problem poses a serious challenge due to the unknown optimal number of measurements and the possible severe nonlinearity or nonsmoothness of the observation mapping $\mathcal{O}\colon \varOmega_o \to \mathbb{R}^n$.

A geometric interpretation of uncertainty

Another, more geometric way, of describing the influence of measurement errors on the maximum likelihood estimator is based on the computation of its confidence region. For a confidence level $\alpha \in (0, 1)$ and a realization $\epsilon \in \mathbb{R}^N$ of the measurement noise, we set

$$D(\bar{q}, \alpha)(\epsilon) = \left\{ q \in \mathbb{R}^n \mid J(q, \mathbf{y}_d) - \min_{q \in \mathbb{R}^n} J(q, \mathbf{y}_d) \leq \gamma_n^2(\alpha)/2 \right\}$$

where $\mathbf{y}_d = \mathbf{y}_d^\dagger + \epsilon$ and $\gamma_n^2(\alpha)$ denotes the $(1-\alpha)$-quantile of the χ^2-distribution with n degrees of freedom. Note that this set can be rewritten in several equivalent ways:

$$\begin{aligned} D(\bar{q},\alpha)(\epsilon) &= \left\{ q \in \mathbb{R}^n \mid J(q,\mathbf{y}_d) - \min_{q\in\mathbb{R}^n} J(q,\mathbf{y}_d) \le \gamma_n^2(\alpha)/2 \right\} \\ &= \left\{ q \in \mathbb{R}^n \mid (q - q^{\mathbf{y}_d})^\top X^* \varSigma^{-1} X (q - q^{\mathbf{y}_d}) \le \gamma_n^2(\alpha) \right\} \\ &= \left\{ q \in \mathbb{R}^n \mid q = q^{\mathbf{y}_d} + (X^*\varSigma^{-1}X)^{-1} X^* \varSigma^{-1/2} \delta\epsilon,\ |\delta\epsilon|_{\mathbb{R}^n} \le \gamma_n(\alpha) \right\}. \end{aligned} \tag{2.11}$$

The mapping

$$D(\bar{q},\alpha)\colon D \to \mathcal{P}(\mathbb{R}^n), \quad \omega \mapsto D(\bar{q},\alpha)(\varepsilon(\omega))$$

is called the confidence region of $\bar{q}$ to the confidence level $\alpha \in (0,1)$. It is a random variable taking values in $\mathcal{P}(\mathbb{R}^n)$, the power sets of the parameter space. Loosely speaking, confidence regions should be interpreted as follows: If we compute several realizations $\bar{q}(\omega)$ of the maximum likelihood estimator, then the true parameter q^* is contained in $\alpha \cdot 100\%$ of the associated realizations $D(\bar{q},\alpha)(\varepsilon(\omega))$.

Consequently, the size of these sets also provides a measure on the statistic quality of the estimator. If the realizations of the confidence regions are small, we may conclude that realizations of $\bar{q}$ are close to q^* with high probability. At this point, we stress that each realization $D(\bar{q},\alpha)(\varepsilon(\omega))$ is an ellipsoid in the parameter space centered at $\bar{q}(\omega)$, see (2.11). For every fixed $\alpha \in (0,1)$, its shape and size are described by the Fisher information matrix $X^*\varSigma^{-1}X$ which is independent of $\omega \in D$. As for the mean squared error, this observation suggests that the confidence domains of the estimator can be minimized a priori by a sophisticated choice of the sensor positions and the variances of the measurements.

Again, this task is formulated as an optimization problem based on a parametrization of the Fisher information by the measurement setup. For example, we may minimize the sum over the eigenvalues of $(X^*\varSigma^{-1}X)^{-1}$ corresponding to the combined length of the ellipsoid's half-axes. The associated sensor placement problem is

$$\min_{x_i\in\varOmega_o,\mathbf{u}_i\in\mathbb{R}_+,N\in\mathbb{N}} [\mathrm{Tr}_{\mathbb{R}^n}((X^*\varSigma^{-1}X)^{-1}) + \mathcal{R}(\mathbf{u})] \quad s.t. \quad (Xq)_i = (\mathcal{O}(x_i),q)_Q,\ \varSigma_{ij}^{-1} = \delta_{ij}\mathbf{u}_i,$$

Thus, minimizing the half-axes of the confidence ellipsoids corresponds to reducing the mean squared error of the maximum likelihood estimator. Another possible criterion to assess the quality of the obtained estimates can be based on the volume of $D(\bar{q},\alpha)(\varepsilon(\omega))$ which is, up to a measurement independent constant, given by the determinant of the covariance matrix. Formulating a sensor placemennt problem for minimizing this criterion leads to

$$\min_{x_i\in\varOmega_o,\mathbf{u}_i\in\mathbb{R}_+,N\in\mathbb{N}} [\mathrm{Det}((X^*\varSigma^{-1}X)^{-1}) + \mathcal{R}(\mathbf{u})] \quad s.t. \quad (Xq)_i = (\mathcal{O}(x_i),q)_Q,\ \varSigma_{ij}^{-1} = \delta_{ij}\mathbf{u}_i,$$

Minimizing the determinant of the covariance matrix is called the D-optimal design problem, while optimizing its trace is usually referred to as A-optimality. For a reference we point out to [222, Chapter 6]. We stress that both of these problems fit into a more abstract framework of sensor placement problems described by

$$\min_{x_i\in\varOmega_o,\mathbf{u}_i\in\mathbb{R}_+,N\in\mathbb{N}} [\varPsi(X^*\varSigma^{-1}X) + \mathcal{R}(\mathbf{u})] \quad s.t. \quad (Xq)_i = (\mathcal{O}(x_i),q)_Q,\ \varSigma_{ij}^{-1} = \delta_{ij}.\mathbf{u}_i.$$

Here, the optimal design criterion $\varPsi$ is a convex and differentiable function on the cone of positive definite matrices. We adopt such an abstract formulation for the case of general parameter spaces in Chapter 3. To finish these discussions, we point out to the consideration of nondifferentiable optimal design criteria. One particular prominent example is constituted by the largest eigenvalue of the covariance matrix, the E-optimality criterion, resembling the length of the longest half-axis of the confidence ellipsoids, [84]. Such optimal design criteria are beyond the scope of this thesis but represent an interesting topic for future research.

2.3.2 Underdetermined problems

It remains to comment on the situation of underdetermined measurements. That is the number of observations N is strictly smaller than the dimension of the parameter space Q. To make the following discussions more transparent, we restrict them to finite dimensional parameter spaces $Q \simeq \mathbb{R}^n$ with $n > N$. The case of infinite dimensional parameter spaces is briefly addressed at the end of this section. We point out that $X \in \mathbb{R}^{N\times n}$ is not injective. Thus, the matrix $X^*\varSigma^{-1}X$ is not invertible and we have to choose a suitable nonzero regualizer $\mathcal{I}_0^{1/2}$ to ensure the existence of $(X^*\varSigma^{-1}X + \mathcal{I}_0)^{-1}$. The regularized Least-Squares estimator is then defined as

$$\bar{q}\colon D \to Q, \quad \omega \mapsto \hat{q} + (X^*\varSigma^{-1}X + \mathcal{I}_0)^{-1}(X^*\varSigma^{-1}(\mathbf{y}_d^\dagger + \varepsilon(\omega) - S[\hat{q}](x))).$$

Assume that there is no modeling error present i.e.

$$\min_{q\in\mathbb{R}^n} |\partial S[\hat{q}](x) + X(q-\hat{q}) - \mathbf{y}_d^\dagger|_{\mathbb{R}^N} = 0 \tag{2.12}$$

Since X is not injective, we emphasize that the solution to this minimization problem is not unique. In the following, we simply select one particular minimizer q^* and evaluate the capability of $\bar{q}$ to estimate q^*. Due to the lack of sufficient information from the provided measurements and the appearance of the regularization term in the problem, the regularized estimator $\bar{q}$ is in general biased i.e. $\mathbb{E}[\bar{q}] \neq q^*$. More in detail, there holds

$$\begin{aligned} |\mathbb{E}[\bar{q}] - q^*|_{\mathbb{R}^n} &= |((X^*\varSigma^{-1}X + \mathcal{I}_0)^{-1}X^*\varSigma^{-1}X - \mathrm{Id})(q^* - \hat{q})|_{\mathbb{R}^n} \\ &= |(X^*\varSigma^{-1}X + \mathcal{I}_0)^{-1}\mathcal{I}_0(q^* - \hat{q})|_{\mathbb{R}^n}. \end{aligned}$$

Consequently, the mean squared error between $\bar{q}$ and q^* is calculated as

$$\mathrm{MSE}(\bar{q}, q^*) = |(X^*\varSigma^{-1}X + \mathcal{I}_0)^{-1}\mathcal{I}_0(q^* - \hat{q})|^2_{\mathbb{R}^n} + \mathrm{Tr}_{\mathbb{R}^n}(\mathcal{C}),$$

where the covariance matrix $\mathcal{C}$ is given by

$$\mathcal{C} = (X^*\varSigma^{-1}X + \mathcal{I}_0)^{-1}X^*\varSigma^{-1}X(X^*\varSigma^{-1}X + \mathcal{I}_0)^{-1}.$$

In contrast to the overdetermined case, we observe that the mean squared error depends on the unknown parameter q^*. This prevents its numerical evaluation and a sophisticated choice of the measurement setup based on minimizing the mean squared error of the estimator. Moreover, we point out that q^* was more or less chosen arbitrary from the solution set to (2.12). These observations suggest that the mean squared error is only of limited practical utility with regard to optimal sensor placement in the present case.

Nevertheless, it would be desirable to formulate meaningful optimal design criteria to allow for a rigorous and systematic choice of the measurement setup before the actual experiment is carried out. For example, we may base the choice of an optimal measurement procedure on averaging the mean squared error over possible values of q^*, see e.g. [127]. To this end, we consider a probability measure μ_0 on the parameter space with finite second moments. That is $\int_{\mathbb{R}^n} q^2 \, \mathrm{d}\mu_0(q) < \infty$. The μ_0-averaged mean squared error is defined as

$$\int_{\mathbb{R}^n} |(X^* \Sigma^{-1} X + \mathcal{I}_0)^{-1} \mathcal{I}_0 (q - \hat{q})|^2_{\mathbb{R}^n} \, \mathrm{d}\mu_0(q) + \mathrm{Tr}_{\mathbb{R}^n}(\mathcal{C}) < \infty.$$

Clearly, this averaged mean squared error no longer depends on the particular choice of q^* and can be minimized with respect to the measurement setup. However, this comes at the cost of evaluating an integral over the parameter space. By assumption, the linearization point $\hat{q}$ can be interpreted as sophisticated a priori guess for the true unknown parameter. Thus it is reasonable to consider probability measures whose mass is localized around $\hat{q}$. A particularly interesting observation can be made in the Gaussian case.

Proposition 2.3. *Let $\mathcal{I}_0^{1/2} \in \mathbb{R}^{n \times n}$ be invertible and set $\mu_0 = \mathcal{N}(\hat{q}, \mathcal{I}_0^{-1})$. Then there holds*

$$\int_{\mathbb{R}^d} |(X^* \Sigma^{-1} X + \mathcal{I}_0)^{-1} \mathcal{I}_0 (q - \hat{q})|^2_{\mathbb{R}^n} \, \mathrm{d}\mu_0(q) + \mathrm{Tr}_{\mathbb{R}^n}(\mathcal{C}) = \mathrm{Tr}_{\mathbb{R}^n}((X^* \Sigma^{-1} X + \mathcal{I}_0)^{-1}).$$

Proof. See [3, Theorem 2]. □

For general parameter spaces Q, a straightforward adaption of this result remains valid. In the general case, the mean squared error between $\bar{q}$ and q^* is given by

$$\|((X^* \Sigma^{-1} X + \mathcal{I}_0)^{-1} X^* \Sigma^{-1} X - \mathrm{Id})(q^* - \hat{q})\|^2_Q + \mathrm{Tr}_Q(\mathcal{C}),$$

whenever the operator $X^* \Sigma^{-1} X + \mathcal{I}_0$ admits a continuous inverse. As in the finite dimensional case we assume that $\mathcal{I}_0$ admits a continuous inverse and average the mean squared error with respect to a Gaussian probability measure $\mu_0 = \mathcal{N}(\hat{q}, \mathcal{I}_0^{-1})$ centered at the linearization point. The covariance operator is again related to the regularization term in the estimator. In particular, see Section 5.1.1, the Gaussian assumption on μ_0 requires the operator $\mathcal{I}_0^{-1}$ to be of trace class i.e. $\mathrm{Tr}_Q(\mathcal{I}_0^{-1}) < \infty$. For infinite dimensional Q, this imposes a restriction on the decay rate of its eigenvalues and implies that $\mathcal{I}_0^{-1}$ is smoothing. Now, again following [3], we obtain

$$\int_Q \|((X^* \Sigma^{-1} X + \mathcal{I}_0)^{-1} X^* \Sigma^{-1} X - \mathrm{Id})(q - \hat{q})\|^2_Q \, \mathrm{d}\mu_0(q) + \mathrm{Tr}_Q(\mathcal{C}) = \mathrm{Tr}_Q((X^* \Sigma^{-1} X + \mathcal{I}_0^{-1})).$$

Consequently, independent of the parameter dimension, a sophisticated choice of the measurement setup can be based on minimizing the trace of the operator $(X^* \Sigma^{-1} X + \mathcal{I}_0)^{-1} \in \mathcal{L}(Q, Q)$. High-dimensional parameter spaces Q further aggravate the numerical treatment of the associated sensor placement problems. For example, evaluating the design criterion already requires the trace of an inverse of a usually large and dense matrix whose computation is a formidable problem in itself. These type of sensor placement problems and their efficient solution are in the focus of Chapter 5. We point out that the averaged optimal sensor placement formulations also admit an interpretation as a certain bilevel optimization problem. Here, we aim to improve the measurement setup such that the resulting estimator provides, on average, good reconstruction results on a set of training parameters, described by the probability measure μ_0. Since we give a profound discussion of this topic in Chapter 5 we do not go into greater detail at this point.

Remark 2.3. For completeness we again pose the question whether the regularized Least-Squares estimator is stable in the vanishing noise case. In the previous section, see Remark 2.2, we observed that the maximum likelihood estimator converges in the mean square sense towards the unique solution of a deterministic Least-Squares problem if the noise tends to zero. Intuitively, we also expect a similar behavior in the regularized case.
Let us outline these ideas for the case of regularizing with the euclidean norm. That is we set $\mathcal{I}_0^{1/2} = \mathrm{Id}$ where Id denotes the identity matrix. Again, consider a parametrized family of measurement noises $\varepsilon_\sigma \sim \mathcal{N}(0, \Sigma_\sigma)$ where $\Sigma_\sigma = \sigma\widehat{\Sigma}$ for $\sigma > 0$ and some positive definite diagonal matrix $\widehat{\Sigma} \in \mathbb{R}^{N\times N}$. Note that the deterministic Least-Squares problem for the unperturbed measurement vector $\mathbf{y}_d^\dagger$

$$\min_{q\in\mathbb{R}^n} |\widehat{\Sigma}^{-1/2}(S[\hat{q}](x) + X(q-\hat{q}) - \mathbf{y}_d^\dagger)|^2_{\mathbb{R}^N}, \tag{2.13}$$

admits infinitely many solutions since the kernel of X is non-trivial. The associated estimators are given by

$$\bar{q}_\sigma \colon D \to \mathbb{R}^N, \quad \omega \to \hat{q} + (X^*\widehat{\Sigma}^{-1}X + \sigma\,\mathrm{Id})^{-1}(X^*\widehat{\Sigma}^{-1}(\mathbf{y}_d^\dagger + \varepsilon_\sigma(\omega) - S[\hat{q}](x)))$$

Its σ-dependent mean is

$$\mathbb{E}[\bar{q}_\sigma] = \hat{q} + (X^*\widehat{\Sigma}^{-1}X + \sigma\,\mathrm{Id})^{-1}(X^*\widehat{\Sigma}^{-1}(\mathbf{y}_d^\dagger - S[\hat{q}](x))).$$

In the same way, its covariance matrix is determined as

$$\mathcal{C}_\sigma = \sigma(X^*\widehat{\Sigma}^{-1}X + \sigma\,\mathrm{Id})^{-1}X^*\widehat{\Sigma}^{-1}X(X^*\widehat{\Sigma}^{-1}X + \sigma\,\mathrm{Id})^{-1}.$$

For $\sigma \to 0$ we conclude

$$(X^*\widehat{\Sigma}^{-1}X + \sigma\,\mathrm{Id})^{-1}X^*\widehat{\Sigma}^{-1/2} \to (\Sigma^{-1/2}X)^\dagger,$$

where $(\widehat{\Sigma}^{-1/2}X)^\dagger$ denotes the Moore-Penrose inverse of $\widehat{\Sigma}^{-1/2}X$. Accordingly, there holds

$$\mathcal{C}_\sigma \to 0, \quad \mathbb{E}[\bar{q}_\sigma] \to q^\dagger := \hat{q} + (\widehat{\Sigma}^{-1/2}X)^\dagger\widehat{\Sigma}^{-1/2}(\mathbf{y}_d^\dagger - S[\hat{q}](x)),$$

as $\sigma \to 0$. The limiting parameter $q^\dagger$ is the unique minimum norm solution to (2.13) with respect to the euclidean norm shifted by $\hat{q}$, see e.g. [184, Theorem 20.9]. That is

$$q^\dagger = \operatorname*{arg\,min}_{q\in\mathbb{R}^n} \left\{ |q-\hat{q}|_{\mathbb{R}^n} \mid q \in \operatorname*{arg\,min}_{\tilde{q}\in\mathbb{R}^n} |\widehat{\Sigma}^{-1/2}(S[\hat{q}](x) + X(\tilde{q}-\hat{q}) - \mathbf{y}_d^\dagger)|^2_{\mathbb{R}^N} \right\}.$$

Combining all previous observations, we obtain

$$\mathrm{MSE}(\bar{q}_\sigma, q^\dagger) = |\mathbb{E}[\bar{q}_\sigma] - q^\dagger|^2_{\mathbb{R}^n} + \mathrm{Tr}_{\mathbb{R}^n}(\mathcal{C}_\sigma) \to 0$$

as σ tends to zero.

2.3.3 Interlude: The Bayesian approach

Before proceeding to the main part of the thesis, we briefly outline a different regularization strategy for the stable solution of the ill-posed inverse problem (2.5). Here, instead of solving an optimization problem, we encode our prior uncertainty on a suitable choice for the unknown parameter into a probability measure. Thus, besides our already probabilistic description of the measurement error, we now also adopt a stochastic model for the parameter. Consequently, the regularized solution to the inverse problem is not given by a single element $q_{\text{post}}^{\mathbf{y}_d} \in Q$ but a probability distribution $\mu_{\text{post}}^{\mathbf{y}_d}$ on the parameter space. We refer to this method as the Bayesian approach, see e.g. [153, 160]. For convergence results in the vanishing noise limit we point out to [149, 150] and Section 2.3 of [250]. As we will see, the Bayesian approach allows to asses the statistical quality of the obtained regularized solutions based on well-known properties of probability measures. Similar to the case of Tikhonov regularization, this leads to the consideration of scalar-valued optimal design criteria acting on the Fisher information operator $X^* \varSigma^{-1} X$.

The following arguments are restricted to the case of $Q = \mathbb{R}^n$, $n \in \mathbb{N}$. A profound description of the Bayesian approach for inverse problems with infinite dimensional parameter spaces and optimal sensor placement in this context is given in Chapter 5. We briefly recall that the measurement noise is modeled by a random variable ε distributed according to a Gaussian probability measure $\mu_E = \mathcal{N}(0, \varSigma)$. Thus, its density function with respect to the Lebesgue measure on $\mathbb{R}^N$ is, up to a normalization constant, given by

$$\pi_{\text{noise}}(\epsilon) \propto \exp\left(-\frac{1}{2}|\epsilon|^2_{\varSigma^{-1}}\right) \quad \forall \epsilon \in \mathbb{R}^N$$

where the weighted euclidean norm is defined as $|\epsilon|^2_{\varSigma^{-1}} = (\epsilon, \varSigma^{-1}\epsilon)_{\mathbb{R}^N}$. In the Bayesian approach we proceed similarly for the unknown parameter and describe our prior uncertainties by a Gaussian distribution centered at the linearization point $\hat{q}$. In more detail, we assume $q \sim \mu_0 = \mathcal{N}(\hat{q}, \mathcal{I}_0^{-1})$ where $\mathcal{I}_0$ is a positive definite matrix. We refer to μ_0 as the prior distribution of the parameter. The associated density function with respect to the Lebesgue measure on $\mathbb{R}^n$ is

$$\pi_{\text{prior}}(q) \propto \exp\left(-\frac{1}{2}|\mathcal{I}_0^{1/2}(q-\hat{q})|^2_{\mathbb{R}^n}\right) \quad \forall q \in \mathbb{R}^n.$$

The random variables q and ϵ are assumed to be independent. The regularized solution to the inverse problem in (2.5) for a measurement vector $\mathbf{y}_d \in \mathbb{R}^N$ is now given by the posterior distribution $\mu_{\text{post}}^{\mathbf{y}_d}$ which is a probability measure on the parameter space with density function

$$\pi_{\text{post}}(q) \propto \pi_{\text{noise}}(S[\hat{q}](x) + X(q-\hat{q}) - \mathbf{y}_d)\, \pi_{\text{prior}}(q) \quad \forall q \in \mathbb{R}^n. \tag{2.14}$$

Loosely speaking, the posterior distribution combines our prior beliefs on the unknown parameter and the information provided by the measurement data. This intuition is backed up its probability density function which is large at parameters $q \in Q$ that are close to the linearization point $\hat{q}$, with respect to the euclidean norm weighted by $\mathcal{I}_0$, and at which the response of the mathematical model approximately matches the measurement vector. A rigorous justification of this definition can be based on Bayes' Theorem and the notion of conditional density functions. We do not go into greater detail at this point. Note that the statement in (2.14) does neither require a Gaussian distribution for the measurement noise or a Gaussian prior distribution for the parameter. In the present case however, it is readily verified that the posterior distribution $\mu_{\text{post}}^{\mathbf{y}_d}$ is a Gaussian probability measure characterized by

$$\mu_{\text{post}}^{\mathbf{y}_d} = \mathcal{N}(q_{\text{post}}^{\mathbf{y}_d}, (X^* \Sigma^{-1} X + \mathcal{I}_0)^{-1}) \quad \text{with} \quad q_{\text{post}}^{\mathbf{y}_d} = \hat{q} + (X^* \Sigma^{-1} X + \mathcal{I}_0)^{-1} X^* \Sigma^{-1} (\mathbf{y}_d - S[\hat{q}](x)).$$

Observe that its mean is given by the unique global minimizer of

$$\min_{q \in \mathbb{R}^n} \left[\frac{1}{2} |X(q - \hat{q}) + S[\hat{q}] - \mathbf{y}_d|^2_{\Sigma^{-1}} + \frac{1}{2} |\mathcal{I}_0^{1/2} (q - \hat{q})|^2_{\mathbb{R}^n} \right],$$

which is also referred to as the maximum a posteriori probability estimate. Clearly, this is closely related to a Tikhonov regularized solution of (2.5) for the particular case of choosing the weighting matrix in the regularization term as the square root of the inverse covariance operator.

In order to assess the statistical quality of the obtained solution we may now, e.g., quantify the variability of the posterior distribution $\mu_{\text{post}}^{\mathbf{y}_d}$ by computing the expected deviation of the associated random variable from its mean $q_{\text{post}}^{\mathbf{y}_d}$. By definition of the covariance operator this corresponds to

$$\int_{\mathbb{R}^N} |q - q_{\text{post}}^{\mathbf{y}_d}|^2_{\mathbb{R}^n} \, \mathrm{d}\mu_{\text{post}}^{\mathbf{y}_d}(q) = \mathrm{Tr}_{\mathbb{R}^n}((X^* \Sigma^{-1} X + \mathcal{I}_0)^{-1}).$$

Another frequently considered criterion is the negative of the expected information gain between prior and posterior distribution, [249], which is, in the present case, given by

$$\log\left(\mathrm{Det}((X^* \Sigma^{-1} X + \mathcal{I}_0)^{-1})\right).$$

We take a closer look on the derivation of this term in Chapter 5. Please note the similarity of these two criteria to those introduced for Tikhonov regularization in the previous sections.

As for the Tikhonov regularized problems, we stress that both of these exemplary design criteria are independent of the particular measurement vector $\mathbf{y}_d \in \mathbb{R}^N$ but depend on the position of the sensors and the variances of the measurements through the Fisher information operator. Thus, we may again, in a statistical sense, optimize the estimation process before performing any measurements in practice by solving a minimization problem for the optimal measurement setup. For an overview on Bayesian experimental design we point out to [68].

To close these discussions we briefly summarize the most important observations of this chapter. First, the inverse problem of identifying an unknown parameter from finite-dimensional data is in general ill-posed. Thus, it calls for sophisticated regularization strategies. Second, it is reasonable to assume that the provided measurements are subject to random perturbations. Through the estimation process, the uncertainty in the measurement data is also propagated into the parameter space. This has to be properly addressed by e.g. modeling the unknown parameter itself as a random quantity or by viewing the Tikhonov regularized solution as a particular realization of a suitable random estimator. Finally, we have observed that the statistical quality of the obtained regularized solution to the inverse problem can be quantified independent of the measurement vector $\mathbf{y}_d$ based on properties of the Fisher information $X^* \Sigma^{-1} X$. Since this operator depends on the measurement setup we can a priori, i.e. before any measurements are performed in practice, improve the estimation process by solving a minimization problem

$$\min_{x_i \in \Omega_o, \mathbf{u}_i \in \mathbb{R}_+, N \in \mathbb{N}} [\Psi(X^* \Sigma^{-1} X) + \mathcal{R}(\mathbf{u})] \quad s.t. \quad (Xq)_i = (\mathcal{O}(x_i), q)_Q, \; \Sigma^{-1}_{ij} = \delta_{ij} \mathbf{u}_i$$

for all $q \in Q$, $i, j = 1, \ldots, N$. Here we minimize with respect to the optimal number N of performed measurements, the positions of the sensors $\{x_i\}_{i=1}^N$ in the candidate set Ω_o as well as the nonnegative measurement weights $\{\mathbf{u}_i\}_{i=1}^N$ describing how careful each measurement should be taken. The functional Ψ is a usually convex and differentiable function referred to as optimal design criterion and $\mathcal{R}(\mathbf{u})$ is a suitable regularization term representing the cost of the experiment based on the measurement weight vector $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_N)^\top$. This key observation builds a bridge between inverse problems and the sensor placement formulations discussed in the remainder of the present thesis.

3 A sparse control approach to optimal sensor placement

Throughout the course of this chapter we consider a general linear inverse problem given by

$$\text{find } q \in Q\colon \quad (\mathcal{O}(x_i), q)_Q = \mathbf{y}_d^i = (\mathcal{O}(x_i), q^*)_Q + \epsilon_i, \quad i = 1, \dots, N,$$

Here we aim to recover an unknown true parameter q^* in a Hilbert space Q from a finite number of observations $\mathbf{y}_d^i \in \mathbb{R}$. Each of these $N \in \mathbb{N}$ measurements is obtained by taking the inner product on Q between the parameter and an element $\mathcal{O}(x_i) \in Q$, $i = 1, \dots, N$. By $\mathcal{O}\colon \Omega_o \to Q$ we denote a continuous function on a compact set $\Omega_o \subset \mathbb{R}^d$, $d \in \mathbb{N}$. It maps a spatial point $x \in \Omega_o$ to $\mathcal{O}(x)$ in the parameter space which models the action of a measurement device or a sensor located at this point.

As an illustrative example the reader may always think of situations in which inference on the true value of the parameter is only indirectly possible through pointwise measurements of a continuous function $y = Su \in \mathcal{C}(\Omega_o)$. If $S\colon Q \to \mathcal{C}(\Omega_o)$ is a compact linear and continuous operator we define $\mathcal{O}(x) = S^* \delta_x$ where δ_x denotes the Dirac delta function supported on $x \in \Omega$. The resulting function $\mathcal{O}\colon \Omega_o \to Q$ is continuous and fulfills

$$(\mathcal{O}(x), q)_Q = (S^* \delta x, q)_Q = \langle Su, \delta_x \rangle = y(x),$$

for all $x \in \Omega_o$ and $q \in Q$. However we also stress that the following considerations are not limited to this case.

The observation of the sensor at x_i is subject to perturbation by additive noise ϵ_i drawn from a random variable $\varepsilon_i \sim \mathcal{N}(0, 1/\mathbf{u}_i)$. The strictly positive scalar $\mathbf{u}_i \in \mathbb{R}_+ \setminus \{0\}$ may be interpreted as a diligence factor quantifying how carefully the observation at x_i is taken. For example $\mathbf{u}_i$ might be related to the variance of the used sensor or gives the total number of measurements taken at the same location. The measurement errors at two distinct locations are assumed to be independently distributed. Assembling the N equations in one system we arrive at a linear operator equation

$$\text{find } q \in Q\colon \quad Xq = Xq^* + \epsilon = \mathbf{y}_d, \quad i = 1, \dots, N,$$

where ϵ is a realization of the normally distributed random variable $\varepsilon \sim \mathcal{N}(0, \Sigma)$ and the observations are collected in the vector $\mathbf{y}_d \in \mathbb{R}^N$. The parameter-to-observation operator $X \in \mathcal{L}(Q, \mathbb{R}^N)$ and the matrix Σ are given as

$$(Xq)_i = (\mathcal{O}(x_i), q)_Q \quad \forall q \in Q, \quad \Sigma_{ij} = \delta_{ij}/\mathbf{u}_i, \quad i, j = 1, \dots, N.$$

Following the discussion in Section 2.3, the statistic quality of approximate solutions to this inverse problem obtained by e.g. Tikhonov regularization methods, can be measured by scalar-valued criteria Ψ acting on the Fisher information operator

$$X^* \Sigma^{-1} X \in \mathcal{L}(Q, Q).$$

Again, we point out to the crucial observation that this operator is independent of the measurements $\mathbf{y}_d \in \mathbb{R}^N$ but depends on the number and positions of the measurement sensors as well as the statistical quality of the measurements

$$\mathbf{x} = (x_1, \dots, x_N)^\top \in \Omega_o^N, \quad \mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_N)^\top \in \mathbb{R}_+^N.$$

In this chapter we aim to mitigate the influence of the stochastic perturbation in the data on the estimates of the parameter. For this purpose, we optimize the data acquisition process. More in detail we will improve the measurement setup by an optimal choice of the number $N \in \mathbb{N}$ of measurements, their positions x_i in Ω_o as well as the diligence factors $\mathbf{u}_i \in \mathbb{R}_+$ a priori, i.e. before any measurements are performed in practice. We base our discussions on an optimal control formulation of the problem given by

$$\min_{\mathbf{x} \in \Omega_o^N,\, \mathbf{u} \in \mathbb{R}_+^N\ N \in \mathbb{N}} \quad \Psi(X^* \Sigma^{-1} X) + G(\|\mathbf{u}\|_1) \quad s.t. \quad (Xq)_i = (\mathcal{O}(x_i), q)_Q, \quad \Sigma_{ij}^{-1} = \delta_{ij} \mathbf{u}_i, \tag{3.1}$$

for all $q \in Q$ and $i, j = 1, \dots, N$. Here we minimize a given convex optimal design criterion Ψ acting on the Fisher information which is parametrized as a function of $\mathbf{x} \in \Omega_o^N$, $\mathbf{u} \in \mathbb{R}_+^N$ and $N \in \mathbb{N}$. To account for the cost of the experiment in the sensor placement formulation we add a second term to the problem involving the $_1$ norm of the measurement weight vector. This creates a trade-off between the statistic optimality of the measurement setup and its cost. For the specific assumptions on G and Ψ we refer to the next section. If the maximum number N of sensors was fixed and Ω_o consists of finitely many candidate locations such a regularization is known to induce sparsity on the coefficient vector, i.e. an optimal weight vector $\bar{\mathbf{u}}$ will only admit few non-zero entries. For other optimal design approaches involving sparsity promoting regularizations we refer to [4, 71, 127]. In contrast to these prior approaches the set of candidate locations for the sensors does not need to be finite in the context of this thesis. Quite the contrary, our special interest lies in admissible sets Ω_o containing a possibly uncountable number of points. Moreover, we also do not prescribe an a priori upper bound on the possible number of sensors to be used in the measurement process.

At first glance, in spite of the convexity of Ψ, problem (3.1) is non-convex due to the parameterization in terms of the points x_i, and has a combinatorial aspect due to the unknown number of measurements N. The main feature of the approach considered in this thesis is to bypass these difficulties by embedding the problem into a more general abstract formulation. Introducing the set of positive Borel measures $\mathcal{M}^+(\Omega_o)$ on Ω_o, see Section 3.1.2, we determine an optimal design measure from

$$\min_{u \in \mathcal{M}^+(\Omega_o)} \Psi(\mathcal{I}(u)) + G(\|u\|_{\mathcal{M}}), \quad s.t. \quad \mathcal{I}(u) = \int_{\Omega_o} [\mathcal{O}(x) \otimes \mathcal{O}(x)] \, \mathrm{d}u(x), \tag{3.2}$$

where $\|u\|_{\mathcal{M}}$ is the canonical total variation norm. The operator $\mathcal{I}(u)$ is given as the Bochner integral of the pointwise Fisher information

$$I \colon \Omega_o \to \mathrm{SHS}(Q, Q), \quad x \mapsto \mathcal{O}(x) \otimes \mathcal{O}(x),$$

which assumes values in the space of self-adjoint Hilbert-Schmidt operators on Q, see Section 3.1.1. For fixed $x \in \Omega_o$, the operator $\mathcal{O}(x) \otimes \mathcal{O}(x)$ acts on Q via

$$(\delta q_1 [\mathcal{O}(x) \otimes \mathcal{O}(x)] \delta q_2)_Q = (\mathcal{O}(x), \delta q_1)_Q (\mathcal{O}(x), \delta q_2)_Q \quad \forall \delta q_1, \delta q_2 \in Q.$$

Since the operator $\mathcal{I}(u)$ depends linearly on the Borel measure, the new problem in (3.2) is convex. We give a detailed description of the derivation of (3.2) and its connection to (3.1) in Section 3.2. Loosely speaking, instead of minimizing for the positions and the quality of individual sensors, we now optimize the distribution of the measurements over the candidate set Ω_o.

Let us put this work into perspective. By choosing G as the convex indicator function of the interval $[0, K]$, we arrive at

$$\min_{u \in \mathcal{M}^+(\Omega_o)} \Psi(\mathcal{I}(u)) \quad \text{subject to} \quad \|u\|_{\mathcal{M}} \leq K, \tag{3.3}$$

where $K > 0$ denotes the overall maximal cost of the measurements. Under certain conditions on Ψ it can be shown that the inequality constraint in (3.3) is attained for every optimal design. This relates (3.3) closely to the concept of approximate designs introduced by Kiefer and Wolfowitz in [165] for general linear-regression. This approach models possible distributions of measurement sensors by probability measures on Ω_o. We refer also to [9, 105, 107, 198, 205, 222] 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 [17, 256]. Indeed, some key results derived in this context, can also be concluded from our general considerations. Most importantly, we derive several equivalent first order optimality conditions for (3.2), which reduce to the well-known equivalence theorem due to Kiefer and Wolfowitz, see [165], in this special situation. Moreover, we stress that the references above only consider the case of $Q = \mathbb{R}^n$, $n \in \mathbb{N}$. From this point of view our sensor placement formulation can be viewed as a natural generalization of this problem. We further comment on the similarities of our approach to this classical one in the subsequent chapters.

Furthermore, choosing $G(\|u\|_{\mathcal{M}}) = \beta\|u\|_{\mathcal{M}}$ for $\beta > 0$, we end up with a norm-regularized problem

$$\min_{u \in \mathcal{M}^+(\Omega_o)} \Psi(\mathcal{I}(u)) + \beta\|u\|_{\mathcal{M}}. \tag{3.4}$$

Optimization problems with total variation regularization recently received increased attention. We refer e.g. to [50, 74, 95, 210]. In the context of optimal sensor placement a special instance of problem (3.4) was considered in [200] for the task of optimizing the measurement setup in a finite-dimensional, PDE-constrained, inverse problem. For a detailed discussion of sparse sensor placement in this context we also refer to Chapter 4 of the present work.

The aim of this chapter is to provide a rigorous and unified framework to prove well-posedness of (3.2) as well as to analyze the structure of design measures which are obtained from solving it. While it is clear that (3.2) is a more general formulation than (3.1), it can be shown that it admits solutions of the form $u = \sum_{i=1}^{N} \mathbf{u}_i \delta_{x_i}$ under certain conditions, making both approaches essentially equivalent. Applications of this general framework to inverse problems with PDE constraints involving an unknown finite dimensional parameter and to infinite-dimensional Bayesian inverse problems with PDEs can be found in the subsequent chapters.

3.1 Notation

In this section we briefly introduce the additional notation which is needed throughout this chapter. Most important, we summarize the necessary theoretical background on Borel measures and Hilbert-Schmidt operators.

3.1.1 Hilbert-Schmidt operators

Throughout this chapter we consider a real separable Hilbert space Q equipped with the scalar product $(\cdot,\cdot)_Q$. The induced norm is denoted by $\|\cdot\|_Q$. In general Q will not be identified with its topological dual space Q^*. The corresponding duality pairing will be denoted by $\langle\cdot,\cdot\rangle_{Q,Q^*}$. By $T_Q\colon Q \to Q^*$ we denote the Riesz isomorphism

$$\langle \delta q_1, T_Q \delta q_2\rangle_{Q,Q^*} = (\delta q_1, \delta q_2)_Q \quad \forall \delta q_1, \delta q_2 \in Q.$$

The space Q^* is a Hilbert space with respect to the canonical scalar product

$$(\delta q_1^*, \delta q_2^*)_{Q^*} = \langle T_Q^{-1} \delta q_1^*, \delta q_2^*\rangle_{Q,Q^*}, \quad \|\delta q_1^*\|_{Q^*} = \sqrt{(\delta q_1^*, \delta q_1^*)} \quad \forall \delta q_1^*, \delta q_2^* \in Q^*.$$

Given a linear continuous operator B between Q and Q^* we fix the following terminology.

Definition 3.1. Let $B \in \mathcal{L}(Q,Q^*)$ be given. We define:

- B is called non-negative iff

$$\langle \delta q_1, B\delta q_1\rangle_{Q,Q^*} \geq 0 \quad \forall \delta q_1 \in Q.$$

- B is called self-adjoint iff

$$\langle \delta q_1, B\delta q_2\rangle_{Q,Q^*} = \langle \delta q_2, B\delta q_1\rangle_{Q,Q^*} \quad \forall \delta q_1, \delta q_2 \in Q.$$

- B is called positive iff B is self-adjoint and non-negative.

In the course of the following sections we will deal with several subsets in the space of bounded linear operators between Q and Q^*. We first fix the notion of trace class operators from Q into itself, c.f. [243].

Definition 3.2. Let an orthonormal basis $\{\phi_i\}_{i\in\mathbf{I}}$, $\mathbf{I} \subset \mathbb{N}$, of Q and $B \in \mathcal{L}(Q,Q)$ be given. We formally define the trace of B as

$$\operatorname{Tr}_Q(B) = \sum_{i\in\mathbf{I}} (\phi_i, B\phi_i)_Q. \tag{3.5}$$

An operator $B \in \mathcal{L}(Q,Q)$ is called a *trace-class operator* on Q iff

$$\operatorname{Tr}_Q(|B|) = \sum_{i\in\mathbf{I}} (\phi_i, (B^*B)^{\frac{1}{2}}\phi_i)_Q < \infty.$$

Here $|B| = (B^*B)^{\frac{1}{2}} \in \mathcal{L}(Q,Q)$ denotes the uniquely determined positive square root of the positive operator B^*B, [33].

If Q is infinite dimensional the trace of $B \in \mathcal{L}(Q,Q)$ is not finite in general. However if its trace is finite the value is independent of the chosen basis. Following these preparatory steps we introduce the set of Hilbert-Schmidt operators on Q.

Definition 3.3. Let $B \in \mathcal{L}(Q, Q^*)$ be given. We call B *Hilbert-Schmidt* iff

$$\mathrm{Tr}_Q(B^*B) = \sum_{i\in\mathbf{I}} (\phi_i, B^*B\phi_i)_Q = \sum_{i\in\mathbf{I}} \|B\phi_i\|_{Q^*}^2 < \infty.$$

The real vector space of Hilbert-Schmidt operators from Q into Q^* is denoted by

$$\mathrm{HS(Q, Q^*)} := \{\, B \in \mathcal{L}(Q, Q^*) \mid \mathrm{Tr}_Q(B^*B) < \infty \,\}.$$

Analogously we define the vector space of self-adjoint Hilbert-Schmidt operators as

$$\mathrm{SHS}(Q, Q^*) := \{\, B \in \mathrm{HS}(Q, Q^*) \mid B \text{ self-adjoint} \,\}.$$

On $\mathrm{HS}(Q, Q^*)$ we consider the Hilbert-Schmidt scalar product

$$\langle\langle B_1, B_2\rangle\rangle_{\mathrm{HS}(Q,Q^*)} = \mathrm{Tr}_Q(B_1^*B_2) = \sum_{i\in\mathbf{I}} (B_1\phi_i, B_2\phi_i)_{Q^*}, \quad B_1, B_2 \in \mathrm{HS}(Q, Q^*). \tag{3.6}$$

Again, its value is independent on the choice of the orthonormal basis $\{\phi_i\}_{i\in\mathbf{I}}$, see Lemma [12, Lemma 12.1.1.].

Proposition 3.1. *The vector spaces* $\mathrm{HS}(Q, Q^*)$ *and* $\mathrm{SHS}(Q, Q^*)$*, respectively, form separable Hilbert spaces with respect to the norm*

$$\|\cdot\|_{\mathrm{HS}(Q,Q)} = \sqrt{\langle\langle\cdot,\cdot\rangle\rangle_{\mathrm{HS}(Q,Q*)}} = \sqrt{\sum_{i\in\boldsymbol{I}} \|\cdot\phi_i\|_{Q^*}^2},$$

induced by the Hilbert-Schmidt scalar product (3.6).

Proof. For $\mathrm{HS}(Q, Q^*)$ this is stated in, e.g., [12, Theorem 12.1.1]. Since $\mathrm{SHS}(Q, Q^*)$ is a closed subspace of $\mathrm{HS}(Q, Q^*)$ the statement follows. □

Note that Hilbert-Schmidt operators are compact, [12, Proposition 12.1.3.]. The set of positive Hilbert-Schmidt operators

$$\mathrm{Pos}(Q, Q^*) := \{\, B \in \mathrm{SHS}(Q, Q^*) \mid B \text{ is positive} \,\},$$

is a closed subset of $\mathrm{SHS}(Q, Q^*)$. Given two elements $q_1^*,\ q_2^* \in Q^*$ we define the linear continuous operator

$$q_1^* \otimes q_2^* \in \mathcal{L}(Q, Q^*), \quad [q_1^* \otimes q_2^*]q_1 = q_1^*\langle q_1, q_2^*\rangle_{Q,Q^*}, \quad q_1 \in Q. \tag{3.7}$$

The following corollary summarizes some properties of these *rank 1 operators.*

Corollary 3.2. *There holds*

$$q_1^* \otimes q_2^* \in \mathrm{SHS}(Q, Q^*), \quad q_1^* \otimes q_1^* \in \mathrm{Pos}(Q, Q^*), \quad q_1^*,\ q_2^* \in Q^*,$$

with

$$\|q_1^* \otimes q_2^*\|_{\mathrm{HS}(Q,Q^*)} = \|q_1^*\|_{Q^*}\|q_2^*\|_{Q^*}.$$

Furthermore if we identify Q *with its dual space the rank 1 operator* $q_1 \otimes q_2$ *is of trace class on* Q *with*

$$\mathrm{Tr}_Q([q_1 \otimes q_2]) = (q_1, q_2)_Q, \quad q_1,\ q_2 \in Q.$$

Proof. We give a short proof of these facts. Given $q_1^*,\ q_2^* \in Q^*$ the operator induced by (3.7) is obviously self-adjoint and additionally non-negative if $q_1^* = q_2^*$. We calculate

$$\begin{aligned}\operatorname{Tr}_Q([q_1^* \otimes q_2^*][q_1^* \otimes q_2^*]) &= \sum_{i\in\mathbf{I}} \|[q_1^* \otimes q_2^*]\phi_i\|_{Q^*}^2 = \|q_1^*\|_{Q^*}^2 \sum_{i\in\mathbf{I}} (T_Q^{-1} q_2^*, \phi_i)_Q^2 \\ &= \|q_1^*\|_{Q^*}^2 \|T_Q^{-1} q_2^*\|_Q^2 = \|q_1^*\|_{Q^*}^2 \|q_2^*\|_{Q*}^2,\end{aligned}$$

where we used that the Riesz isomorphism is an isometry. Taking the square root yields the result. If $Q \simeq Q^*$ we obtain

$$\operatorname{Tr}_Q([q_1 \otimes q_2]) = \sum_{i\in\mathbf{I}} (q_1, \phi_i)_Q (q_2, \phi_i)_Q = (q_1, q_2)_Q,$$

from Parseval's identity. □

To close this section we consider two special instances of the presented abstract setting.

Example 3.1 (Hilbert-Schmidt on $\mathbb{R}^n$). *Let us first consider the case of $Q \simeq Q^* = \mathbb{R}^n$ equipped with the euclidean scalar product*

$$(q_1, q_2)_Q = (q_1, q_2)_{\mathbb{R}^d} = q_1^\top q_2, \quad q_1,\ q_2 \in \mathbb{R}^n.$$

In this case we readily identify $\mathcal{L}(Q, Q)$ with the space of $n \times n$ matrices $\mathbb{R}^{n\times n}$. Since the parameter space Q is finite dimensional every matrix $B \in \mathbb{R}^{n\times n}$ is Hilbert-Schmidt and of trace-class on $\mathbb{R}^n$. The Hilbert-Schmidt norm corresponds to the Frobenius norm

$$\|B\|_{\mathrm{HS}(\mathbb{R}^n,\mathbb{R}^n)} = \|B\|_{\mathrm{Sym}} = \sqrt{\operatorname{Tr}_{\mathbb{R}^n}(B^\top B)} = \sqrt{\sum_{i,j=1}^n B_{ij}^2}, \quad B \in \mathrm{Sym}(n).$$

The space $\mathrm{SHS}(\mathbb{R}^n, \mathbb{R}^n)$ is given by the symmetric matrices

$$\mathrm{Sym}(n) = \left\{ B \in \mathbb{R}^{n\times n} \mid B^\top = B \right\},$$

and the positive Hilbert-Schmidt operators are identified with the non-negative definite matrices

$$\mathrm{NND}(n) = \{ B \in \mathrm{Sym}(n) \mid (\delta q, B\delta q)_{\mathbb{R}^n} \geq 0 \quad \forall \delta q \in \mathbb{R}^n \}.$$

Last we obtain

$$q_1 \otimes q_2 = q_1 q_2^\top \quad \textit{for } q_1, q_2 \in \mathbb{R}^n.$$

Example 3.2 (Hilbert-Schmidt on $L^2(\Omega)$). *As a second example we consider $Q = L^2(\Omega)$ as the space of square integrable function with respect to the Lebesgue measure on $\Omega \subset \mathbb{R}^d$ open and bounded. We identify $L^2(\Omega)$ with its dual space and consider the canonical scalar product*

$$(q_1, q_2)_Q = (q_1, q_2)_{L^2(\Omega)} = \int_\Omega q_1 q_2 \,\mathrm{d}x \quad \textit{for } q_1,\ q_2 \in L^2(\Omega).$$

Let $B \in \mathrm{HS}(L^2(\Omega), L^2(\Omega))$ be given. From the kernel theorem, [12, Theorem 12.6.1], B is Hilbert-Schmidt on $L^2(\Omega)$ if and only if there exists $k_B \in L^2(\Omega \times \Omega)$ with

$$[Bq](x) = \int_\Omega k_B(x, y) q(y) \,\mathrm{d}y, \quad \|B\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))} = \|k_B\|_{L^2(\Omega\times\Omega)}, \quad q \in L^2(\Omega),$$

and almost all $x \in \Omega_o$. Furthermore B is self-adjoint if and only if $k_B(x,y) = k_B(y,x)$ for almost all $x, y \in \Omega_o$. Given $q_1,\ q_2 \in L^2(\Omega)$ the associated rank 1 operator is identified with $k \in L^2(\Omega \times \Omega)$ where $k(x,y) = q_1(x)q_2(y)$ for almost all $x, y \in \Omega$ and

$$([q_1 \otimes q_2]q)(x) = \int_\Omega q_1(x)q_2(y)q(y)\ \mathrm{d}y = q_1(x)(q_2, q)_Q, \quad \forall q \in Q.$$

3.1.2 Borel measures

In the following we consider an observation set Ω_o in which we allow the collection of measurements. It is assumed to be a compact subset of $\mathbb{R}^d$, $d \in \mathbb{N}$. On Ω_o we define the space of regular Borel measures $\mathcal{M}(\Omega_o)$ as the topological dual of $\mathcal{C}(\Omega_o)$, the space of continuous and bounded functions (see, e.g., [100]), with associated duality pairing $\langle \cdot, \cdot \rangle$ given by

$$\langle \varphi, u \rangle = \int_{\Omega_o} \varphi(x)\ \mathrm{d}u(x) \quad \forall \varphi \in \mathcal{C}(\Omega_o), u \in \mathcal{M}(\Omega_o).$$

Let us recall some properties of this space. Given $u \in \mathcal{M}(\Omega_o)$ we can interpret it as a countably additive function $u\colon \mathcal{B}(\Omega_o) \to \mathbb{R}$, where $\mathcal{B}(\Omega_o)$ denotes the Borel sets on Ω_o. Its associated total variation measure $|u| \in \mathcal{M}^+(\Omega_o)$ is defined as

$$|u|(O) = \sup \left\{ \sum_{i=1}^{\infty} |u|(O_i) \mid O_i \in \mathcal{B}(\Omega_o),\ \text{disjoint partition of } O \right\},$$

for all $O \in \mathcal{B}(\Omega_o)$. The space of Borel measures $\mathcal{M}(\Omega_o)$ forms a Banach space with the norm given by

$$\|u\|_\mathcal{M} = |u|(\Omega_o) = \langle 1, |u| \rangle = \sup_{\varphi \in \mathcal{C}(\Omega_o),\ \|\varphi\|_\mathcal{C} \le 1} \langle y, u \rangle = \sup_{\varphi \in \mathcal{C}(\Omega_o),\ \|\varphi\|_\mathcal{C} \le 1} \int_{\Omega_o} \varphi(x)\ \mathrm{d}u(x),$$

where $\|\cdot\|_\mathcal{C}$ denotes the supremum norm on $\mathcal{C}(\Omega_o)$. Given $K > 0$ the indicator function of the (scaled) unit ball with radius K in $\mathcal{M}(\Omega_o)$ is denoted by $I_{\|u\|_\mathcal{M} \le K}(\cdot)$. By $\mathcal{M}^+(\Omega_o)$ we refer to the set of positive Borel measures on Ω_o (see, e.g., [230, Def. 1.18]),

$$\mathcal{M}^+(\Omega_o) = \{\, u \in \mathcal{M}(\Omega_o) \mid \langle \varphi, u \rangle \ge 0,\ \forall \varphi \in \mathcal{C}(\Omega_o),\ \varphi \ge 0 \,\},$$

with convex indicator function $I_{u \ge 0}(\cdot)$. Given $u \in \mathcal{M}(\Omega_o)$ there exist unique positive measures $u^+, u^- \in \mathcal{M}^+(\Omega_o)$ such that

$$u = u^+ - u^-, \quad \|u\|_\mathcal{M} = \|u^+\|_\mathcal{M} + \|u^-\|_\mathcal{M},$$

c.f. [109]. Furthermore its support is defined as

$$\operatorname{supp} u = \Omega_o \backslash \left(\bigcup \{\, O \in \mathcal{B}(\Omega_o) \mid O \text{ open}, |u|(O) = 0 \,\} \right).$$

Since every $u \in \mathcal{M}(\Omega_o)$ is finite, i.e. $u(\Omega_o) < \infty$, it is a Radon measure and thus its support is a closed set. A sequence $\{u_k\}_{k \in \mathbb{N}} \subset \mathcal{M}(\Omega_o)$ is called convergent with respect to the weak*-topology with limit $u \in \mathcal{M}(\Omega_o)$ if $\langle \varphi, u_k \rangle \to \langle \varphi, u \rangle$ for $k \to \infty$ and for all $\varphi \in \mathcal{C}(\Omega_o)$. This is indicated by $u_k \rightharpoonup^* u$. Throughout the following chapters, we frequently wish to quantify the rate of convergence of a given weak* convergent sequence $\{u_k\}_{k\in\mathbb{N}} \subset \mathcal{M}^+(\Omega_o)$. In general, weak* convergence does not imply norm convergence

$$u_k \rightharpoonup^* u \not\Rightarrow \|u_k - u\|_{\mathcal{M}} \to 0.$$

As an easy example consider a sequence $\{x_k\}_{k\in\subset\mathbb{N}} \subset \Omega_o$ with $x_k \to x$ and $x_k \neq x$ for all $k \in \mathbb{N}$. Then it is readily verified that the corresponding Dirac delta functions fulfill $\delta_{x_k} \rightharpoonup^* \delta_x$ but $\|\delta_{x_k} - \delta_x\|_{\mathcal{M}} = 2$ for all $k \in \mathbb{N}$. As a consequence, the canonical norm is not suitable to quantify weak* convergence. In the following consider a sequence $\{u_k\}_{k\in\mathbb{N}} \subset \mathcal{M}^+(\Omega_o)$ with limit $u \neq 0$. Note that $\|u_k\|_{\mathcal{M}} \to \|u\|_{\mathcal{M}}$ i.e. w.l.o.g we may assume $u_k \neq 0$ for all $k \in \mathbb{N}$. In order to metrize the weak* convergence of such a sequence $\{u_k\}_{k\in\mathbb{N}} \subset \mathcal{M}^+(\Omega_o)$ we observe that

$$u_k \rightharpoonup^* u \Leftrightarrow u_k/\|u_k\|_{\mathcal{M}} \rightharpoonup^* u/\|u\|_{\mathcal{M}},\ \|u_k\|_{\mathcal{M}} \to \|u\|_{\mathcal{M}}.$$

Hence, to quantify the weak* convergence we should account for the convergence of the norms and the weak* convergence of the normalized measures. There are several possibilities to metrize the weak* convergence of a sequence of normalized measures, c.f. the overview in [117]. As an example, given two probability measures μ_1, μ_2, we consider their Wasserstein-1 distance, [259, Definition 6.1.], which is given (in its dual form) by

$$W_1(\mu_1, \mu_2) = \sup\left\{ \langle \varphi, \mu_1 - \mu_2 \rangle \mid \varphi \in \mathcal{C}^{0,1}(\Omega_o),\ \|\varphi\|_{\mathrm{Lip}} \leq 1 \right\},$$

using the Kantorovich-Rubinstein theorem, see [161]. Here, $\mathcal{C}^{0,1}(\Omega_o)$ is the space of Lipschitz continuous functions on Ω_o with $\|\varphi\|_{\mathrm{Lip}}$ denoting the Lipschitz constant, see also Section 4.4.3. We propose to quantify the convergence of a weak* convergent sequence $\{u_k\}_{k\in\mathbb{N}} \subset \mathcal{M}^+(\Omega_o)$ with nonzero limit through the modified Wasserstein distance

$$\bar{W}_1(u_k, u) = W_1(u_k/\|u_k\|_{\mathcal{M}}, u/\|u\|_{\mathcal{M}}) + |\|u_k\|_{\mathcal{M}} - \|u\|_{\mathcal{M}}|. \tag{3.8}$$

We stress that the particular choice of the Wasserstein distance for the metrization of the weak* convergence seems quite arbitrary at first. In the subsequent parts of this thesis our special interest lies in sequences consisting of sparse measures, i.e. measures given as a finite comic combinations of Dirac delta functions. In this situation we largely benefit from the representation of the Wasserstein distance as supremum over Lipschitz continuous functions. This allows to discuss convergence rates for the Wasserstein distance of such sequences based on convergence results for their support points and the associated coefficients. We further establish some kind of equivalence between the modified Wasserstein distance $\bar{W}_1$ and the norm on the dual space of $\mathcal{C}^{0,1}(\Omega_o)$. Additionally, for two probability measures consisting of finitely many Dirac delta functions, the computation of $\bar{W}_1$ can be realized by solving a linear program, see e.g. [206, Section 2.7.], which is feasible if the number of support points is reasonably small. A closer inspection on the choice of the metric and its impact on the convergence results derived in this thesis should be a part of future work.

3.2 Sparse optimal sensor placement

This section is devoted to the derivation of the sparse sensor placement problem defined in (3.2) and to clarify its connection to the formulation given in (3.1). Furthermore we state sufficient and reasonable assumptions on the optimal design criterion Ψ as well as the regularization term to allow for a rigorous analysis of the optimal sensor placement problem.

3.2.1 The Fisher operator

To start, we assume that the measurement $\mathcal{O}(x)$ at a given spatial point depends continuously on the position. Furthermore the parameter space Q is identified with its dual space.

Assumption 3.1. Let $Q \simeq Q^*$ be a real, separable Hilbert space. The observation operator

$$\mathcal{O} \colon \Omega_o \to Q, \quad x \mapsto \mathcal{O}(x),$$

is continuous.

Given the total number of measurements N, a vector of sensor positions $\mathbf{x} = (x_1, \dots, x_N)^\top \subset \Omega_o^N$ and measurement weights $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_N) \in \mathbb{R}^N_+$, we will call the triple $(\mathbf{x}, \mathbf{u}, N)$ a measurement setup in the following. Moreover, we recall the definitions of the associated parameter-to-observation $X \in \mathcal{L}(Q, \mathbb{R}^N)$ and the inverse of the noise covariance matrix $\Sigma^{-1} \in \mathbb{R}^{N \times N}$ as

$$(Xq)_i = (\mathcal{O}(x_i), q)_Q, \quad \Sigma^{-1} = \delta_{ij}\mathbf{u}_i \quad \forall q \in Q,\ i,j = 1, \dots, N.$$

The resulting Fisher information operator $X^*\Sigma^{-1}X$ fulfills

$$(\delta q_1, X^*\Sigma^{-1}X\delta q_2)_Q = (X\delta q_1, \Sigma^{-1}X\delta q_2)_{\mathbb{R}^N} = \sum_{i=1}^{N} \mathbf{u}_i (\mathcal{O}(x_i), \delta q_1)_Q (\mathcal{O}(x_i), \delta q_2)_Q \quad \forall \delta q_1,\ \delta q_2 \in Q.$$

Using the rank 1 operator definition from Section 3.1.1, we now note that the Fisher information operator can be equivalently rewritten as

$$X^*\Sigma^{-1}X = \sum_{i=1}^{N} \mathbf{u}_i [\mathcal{O}(x_i) \otimes \mathcal{O}(x_i)] \in \mathcal{L}(Q, Q),$$

In the following proposition we collect some properties of the pointwise Fisher information mapping

$$I \colon \Omega_o \to \mathcal{L}(Q, Q), \quad x \mapsto \mathcal{O}(x) \otimes \mathcal{O}(x). \tag{3.9}$$

Proposition 3.3. *For every $x \in \Omega_o$ the operator $I(x)$ as defined in* (3.9) *satisfies:*

1. *Given $\delta q_1, \delta q_2 \in Q$ there holds*

$$(\delta q_1, I(x)\delta q_2)_Q = (\mathcal{O}(x), \delta q_1)_Q \, (\mathcal{O}(x), \delta q_2)_Q \, .$$

2. *The operator $I(x)$ is positive, i.e. we have*

$$(\delta q_2, I(x)\delta q_1)_Q = (I(x)\delta q_2, \delta q_1)_Q \, , \quad (\delta q_1, I(x)\delta q_1)_Q \geq 0 \quad \forall \delta q_1, \delta q_2 \in Q.$$

3. *$I(x)$ is Hilbert-Schmidt on Q and of trace class: Given an index set $\boldsymbol{I} \subset \mathbb{N}$ and an orthonormal basis $\{\phi_i\}_{i \in \boldsymbol{I}}$ of Q we have*

$$\operatorname{Tr}_Q(I(x)) = \|I(x)\|_{\mathrm{HS(Q,Q)}} = \sum_{i \in \boldsymbol{I}} (\phi_i, I(x)\phi_i)_Q = \|\mathcal{O}(x)\|_Q^2.$$

Consequently, there holds $I(x) \in \operatorname{Pos}(Q, Q)$.

The mapping $I\colon \Omega_o \mapsto \mathrm{SHS}(Q,Q)$ is uniformly continuous.

Proof. Let $\delta q_1, \delta q_2 \in Q$ and $x \in \Omega_o$ be arbitrary but fixed. By definition of the rank 1 operator we have

$$\begin{aligned}(\delta q_1, I(x)\delta q_2)_Q = (\delta q_2, [\mathcal{O}(x) \otimes \mathcal{O}(x)]\delta q_1)_Q &= \Big(\delta q_1, \mathcal{O}(x)\,(\mathcal{O}(x), \delta q_2)_Q\Big)_Q \\ &= (\mathcal{O}(x), \delta q_1)_Q\,(\mathcal{O}(x), \delta q_2)_Q\,.\end{aligned}$$

Using this characterization we directly conclude

$$(\delta q_1, I(x)\delta q_2)_Q = (\mathcal{O}(x), \delta q_1)_Q\,(\mathcal{O}(x), \delta q_2)_Q = (I(x)\delta q_1, \delta q_2)_Q\,,$$

as well as

$$(\delta q_1, I(x)\delta q_1)_Q = (\mathcal{O}(x), \delta q_1)_Q^2 \geq 0.$$

Hence $I(x)$ is self-adjoint and non-negative.
Since $I(x)$ is a rank 1 operator, it is of trace class in Q with

$$\mathrm{Tr}_Q(I(x)) = \sum_{i\in I} (\phi_i, I(x)\phi_i)_Q = \sum_{i\in I} (\mathcal{O}(x), \phi_i)_Q^2 = (\mathcal{O}(x), \mathcal{O}(x))_Q = \|\mathcal{O}(x)\|_Q^2,$$

where we used Parseval's identity in the penultimate equality. Consequently it is also Hilbert-Schmidt, $I(x) \in \mathrm{HS}(Q,Q)$, with

$$\|I(x)\|^2_{\mathrm{HS(Q,Q)}} = \mathrm{Tr}_Q(I(x)^* I(x)) = \|\mathcal{O}(x)\|_Q^4.$$

Taking the square root yields the desired result. It remains to prove the uniform continuity of I. to this end let $x \in \Omega_o$ and $x_j \subset \Omega_o$ with $\lim_{j\to\infty} x_j = x$ be given. We compute

$$\|I(x) - I(x_j)\|^2_{\mathrm{HS(Q,Q)}} = \|\mathcal{O}(x)\|_Q^4 - 2\,\mathrm{Tr}_Q(I(x)^* I(x_j)) + \|\mathcal{O}(x_j)\|_Q^4.$$

Again, using Parseval's identity we have

$$\begin{aligned}\mathrm{Tr}_Q(I(x)^* I(x_j)) &= \sum_{i\in\mathbf{I}} (I(x)\phi_i, I(x_j)\phi_i)_Q \\ &= \sum_{i\in\mathbf{I}} [(\mathcal{O}(x), \mathcal{O}(x_j))_Q\,(\mathcal{O}(x), \phi_i)_Q\,(\mathcal{O}(x_j), \phi_i)_Q] \\ &= (\mathcal{O}(x), \mathcal{O}(x_j))_Q^2\,.\end{aligned}$$

Due to the continuity of the observation operator $\mathcal{O}$ we conclude

$$\lim_{j\to\infty} \mathrm{Tr}_Q(I(x)^* I(x_j)) = \lim_{j\to\infty} [(\mathcal{O}(x), \mathcal{O}(x_j))_Q^2] = \|\mathcal{O}(x)\|_Q^4,$$

and thus $\lim_{j\to\infty} \|I(x) - I(x_j)\|^2_{\mathrm{HS(Q,Q)}} = 0$. Together with the compactness of Ω_o this implies uniform continuity of I. □

Let an arbitrary measurement setup $(\mathbf{x}, \mathbf{u}, N)$ be given. Associated to this triple we define the sparse design measure

$$u = \sum_{i=1}^{N} \mathbf{u}_i \delta_{x_i} \in \mathcal{M}^+(\Omega).$$

The mapping $I \colon \Omega_o \to \mathrm{SHS}(Q,Q)$ is uniformly continuous and thus strongly measurable with respect to u. Furthermore we have

$$\int_{\Omega} \|I(x)\|_{\mathrm{HS(Q,Q)}} \; \mathrm{d}u(x) \leq \max_{x \in \Omega_o} \|I(x)\|_{\mathrm{HS(Q,Q)}} \, \|u\|_{\mathcal{M}} < \infty.$$

Thus the Bochner integral of I with respect to the design measure is well defined due to the separability of $\mathrm{SHS}(Q,Q)$, [267, Theorem 24.8]. Calculating the integral reveals

$$X^* \Sigma^{-1} X = \sum_{i=1}^{N} \mathbf{u}_i [\mathcal{O}(x_i) \otimes \mathcal{O}(x_i)] = \int_{\Omega_o} [\mathcal{O}(x) \otimes \mathcal{O}(x)] \; \mathrm{d}u(x) = \int_{\Omega_o} I(x) \; \mathrm{d}u(x).$$

Consequently, we make the crucial observation that the Fisher information $X^* \Sigma^{-1} X$ can be represented as the Bochner integral of I with respect to the sparse design measure $u = \sum_{i=1}^{N} \mathbf{u}_i \delta_{x_i}$. Naturally we can extend this representation to every Radon measure $u \in \mathcal{M}(\Omega)$.

Proposition 3.4. *Let $u \in \mathcal{M}(\Omega_o)$ and its Jordan decomposition*

$$u = u^+ - u^-, \quad \|u\|_{\mathcal{M}} = \|u^+\|_{\mathcal{M}} + \|u^-\|_{\mathcal{M}}, \quad u^+,\; u^- \in \mathcal{M}^+(\Omega_o),$$

be given. Then the Bochner integrals of I with respect to u^+ and u^-, respectively, are well defined. Set

$$\mathcal{I}(u) = \mathcal{I}(u^+) - \mathcal{I}(u^-) = \int_{\Omega_o} [\mathcal{O}(x) \otimes \mathcal{O}(x)] \; \mathrm{d}u^+(x) - \int_{\Omega_o} [\mathcal{O}(x) \otimes \mathcal{O}(x)] \; \mathrm{d}u^-(x).$$

Then $\mathcal{I}(u) \in \mathrm{SHS}(Q,Q)$ and the mapping

$$\mathcal{I} \colon \mathcal{M}(\Omega_o) \to \mathrm{SHS}(Q,Q), \quad u \mapsto \mathcal{I}(u), \tag{3.10}$$

is linear and continuous. There holds

$$\|\mathcal{I}\|_{\mathcal{L}(\mathcal{M}(\Omega_o), \mathrm{SHS}(Q,Q))} \leq \max_{x \in \Omega_o} \|\mathcal{O}(x)\|^2$$

To prove these results, we recall a basic property of the Bochner integral.

Lemma 3.5. *Let H be an arbitrary Hilbert space and let $T \in \mathcal{L}(\mathrm{SHS}(Q,Q), H)$ be given. Then the function $TI \colon \Omega_o \to H$ is Bochner integrable with respect to $u \in \mathcal{M}^+(\Omega_o)$ and*

$$T\mathcal{I}(u) = \int_{\Omega_o} TI(x) \; \mathrm{d}u(x), \tag{3.11}$$

i.e. applying T commutes with the integral.

Proof. See [8, Theorem 2.1]. □

Let us now prove Proposition 3.4.

Proof of Proposition 3.4. Let $u \in \mathcal{M}(\Omega_o)$ and its Jordan decomposition $u = u^+ - u^-$ be given. Due to the uniform continuity of I its Bochner integrals with respect to u^+ and u^- respectively are well-defined and thus $\mathcal{I}(u) \in \mathrm{SHS}(Q,Q)$. We proceed to prove the linearity of $\mathcal{I}$. Let two measures $u_1, u_2 \in \mathcal{M}(\Omega_o)$, $\lambda \in \mathbb{R}$ be given. Note that in general $(\lambda u^1 + u^2)^+ \neq \lambda u_1^+ + u_2^+$. To circumvent this problem let $B \in \mathrm{SHS}(Q,Q)$ be arbitrary but fixed. Using Lemma 3.5 we obtain

$$\begin{aligned}\mathrm{Tr}_Q(B^*\mathcal{I}(\lambda u_1 + u_2)) &= \mathrm{Tr}_Q(B^*\mathcal{I}((\lambda u_1 + u_2)^+)) - \mathrm{Tr}_Q(B^*\mathcal{I}((\lambda u_1 + u_2)^-)) \\ &= \int_{\Omega_o} \mathrm{Tr}_Q(B^* I(x)) \,\mathrm{d}(\lambda u_1 + u_2)^+(x) - \int_{\Omega_o} \mathrm{Tr}_Q(B^* I(x)) \,\mathrm{d}(\lambda u_1 + u_2)^-(x) \\ &= \lambda \int_{\Omega_o} \mathrm{Tr}_Q(B^* I(x)) \,\mathrm{d}u_1(x) + \int_{\Omega_o} \mathrm{Tr}_Q(B^* I(x)) \,\mathrm{d}u_2(x) \\ &= \lambda \,\mathrm{Tr}_Q(B^*\mathcal{I}(u_1)) + \mathrm{Tr}_Q(B^*\mathcal{I}(u_2)),\end{aligned}$$

where we used the linearity of the Bochner integral in the second inequality, the linearity of duality pairing between $\mathcal{C}(\Omega_o)$ and $\mathcal{M}(\Omega_o)$ in the third one as well as the continuity of the trace. Since B was chosen arbitrary we conclude $\mathcal{I}(\lambda u_1 + u_2) = \lambda\mathcal{I}(u_1) + \mathcal{I}(u_2)$. This yields the linearity of $\mathcal{I}$. Finally, given $u \in \mathcal{M}(\Omega_o)$ we obtain

$$\begin{aligned}\|\mathcal{I}(u)\|_{\mathrm{HS(Q,Q)}} &\leq \left\|\mathcal{I}(u^+)\right\|_{\mathrm{HS(Q,Q)}} + \left\|\mathcal{I}(u^-)\right\|_{\mathrm{HS(Q,Q)}} \\ &\leq \int_{\Omega_o} \|I(x)\|_{\mathrm{HS(Q,Q)}} \,\mathrm{d}u^+(x) + \int_{\Omega_o} \|I(x)\|_{\mathrm{HS(Q,Q)}} \,\mathrm{d}u^-(x) \\ &\leq \max_{x\in\Omega_o} \|I(x)\|_{\mathrm{HS(Q,Q)}} \|u\|_{\mathcal{M}},\end{aligned}$$

where we used $|u| = u^+ + u^-$ and

$$\|\mathcal{I}(\tilde{u})\|_{\mathrm{HS(Q,Q)}} \leq \int_{\Omega_o} \|I(x)\|_{\mathrm{HS(Q,Q)}} \,\mathrm{d}\tilde{u}(x),$$

for all $\tilde{u} \in \mathcal{M}^+(\Omega_o)$, c.f [8, Theorem 2.1]. Noting that

$$\max_{x\in\Omega_o} \|I(x)\|_{\mathrm{HS(Q,Q)}} = \max_{x\in\Omega_o} \|\mathcal{O}(x)\|_Q^2,$$

see Proposition 3.3, we conclude

$$\|\mathcal{I}\|_{\mathcal{L}(\mathcal{M}(\Omega_o),\mathrm{SHS}(Q,Q))} \leq \max_{x\in\Omega_o} \|\mathcal{O}(x)\|^2.$$

□

Due to the linearity of the Bochner integral, some properties of $I(x)$, $x \in \Omega_o$, are carried over to $\mathcal{I}(u)$ if $u \in \mathcal{M}^+(\Omega_o)$.

Corollary 3.6. *Let $u \in \mathcal{M}^+(\Omega_o)$ be given. Then the Fisher information $\mathcal{I}(u)$ satisfies:*

1. *The operator $\mathcal{I}(u)$ is positive:*

$$(\delta q_2, \mathcal{I}(u)\delta q_1)_Q = (\mathcal{I}(u)\delta q_2, \delta q_1)_Q\,, \quad (\delta q_1, \mathcal{I}(u)\delta q_1)_Q \geq 0 \quad \forall \delta q_1, \delta q_2 \in Q.$$

2. The operator $\mathcal{I}(u)$ is of trace class with

$$\operatorname{Tr}_Q(\mathcal{I}(u)) = \int_{\Omega_o} \|\mathcal{O}(x)\|_Q^2 \; \mathrm{d}u(x).$$

Proof. Since $\mathcal{I}(u) \in \operatorname{SHS}(Q,Q)$ it is self-adjoint. Given $\delta q_1, \delta q_2 \in Q$ we observe

$$(\delta q_1, \mathcal{I}(u)\delta q_1)_Q = \int_{\Omega_o} (\delta q_1, I(x)\delta q_1)_Q \; \mathrm{d}u(x) \geq 0,$$

since $I(x)$ is non-negative for every $x \in \Omega_o$ and $u \in \mathcal{M}^+(\Omega_o)$. Let an index set $\mathbf{I} \subset \mathbb{N}$ and an orthonormal basis $\{\phi_i\}_{i\in\mathbf{I}}$ be given. If $\mathbf{I}$ is finite, i.e. Q is finite dimensional, then we readily obtain

$$\operatorname{Tr}_Q(\mathcal{I}(u)) = \sum_{i\in\mathbf{I}} (\phi_i, \mathcal{I}(u)\phi_i)_Q = \int_{\Omega_o} \operatorname{Tr}_Q(I(x)) \; \mathrm{d}u(x) = \int_{\Omega_o} \|\mathcal{O}(x)\|_Q^2 \; \mathrm{d}u(x).$$

Assume that $\mathbf{I} = \mathbb{N}$. For $n \in \mathbb{N}$ we define the continuous function

$$f_n \colon \Omega_o \to \mathbb{R}, \quad f_n(x) = \sum_{i=1}^{n} (\phi_i, I(x)\phi_i)_Q \, .$$

There holds

$$\int_{\Omega_o} f_n(x) \; \mathrm{d}u(x) = \sum_{i=1}^{n} \int_{\Omega_o} (\phi_i, I(x)\phi_i)_Q \; \mathrm{d}u(x) = \sum_{i=1}^{n} (\phi_i, \mathcal{I}(u)\phi_i)_Q \, .$$

Let us observe that for every $n \in \mathbb{N}$ and $x \in \Omega_o$ we have $f_{n+1}(x) \geq f_n(x) \geq 0$ as well as $\lim_{n\to\infty} f_n(x) = \|\mathcal{O}(x)\|_Q^2$. Consequently, applying the monotone convergence theorem, see [100, Theorem 2.7], we can apply the limit on both sides to obtain

$$\operatorname{Tr}_Q(\mathcal{I}(u)) = \lim_{n\to\infty} \int_{\Omega_o} f_n(x) \; \mathrm{d}u(x) = \int_{\Omega_o} \|\mathcal{O}(x)\|_Q^2 \; \mathrm{d}u(x).$$

This concludes the proof. □

3.2.2 Sparse optimal design

Let us now return to the modeling of the sensor placement problem. While the primary goal of an optimal measurement setup is to minimize the uncertainty in the estimation of the parameter it should also account for the costs of the experiment and either aim to minimize them simultaneously or ensure that overall budget constraints are respected. Assuming that the cost of a single measurement is independent on the position of the measurement sensor and scales linearly with the measurement weight, the cost of the experiment can be modeled by the $_1$ norm of the measurement weight vector $\mathbf{u} \in \mathbb{R}^N_+$. To incorporate these costs in the optimal design problem we will add a general, convex regularization term $G(\|\mathbf{u}\|_1)$ to the design criterion Ψ. For example we may consider

$$G_1(\|\mathbf{u}\|_1) = \beta\|\mathbf{u}\|_1, \; \beta > 0, \quad \text{or} \quad G_2(\|\mathbf{u}\|_1) = I_{[0,K]}(\|\mathbf{u}\|_1), \; K > 0.$$

In this fashion, the experimenter may on the one hand create a trade-off between minimizing the optimal design criterion $\varPsi$ and the cost of the experiment or, on the other hand, the total budget for the experiment can be fixed a priori.

For the optimal inference of the unknown parameter we now propose to choose $(\mathbf{x}, \mathbf{u}, N)$ by minimizing the sum of a convex optimal design criterion $\varPsi$ acting on the parametrized Fisher information operator $X^* \varSigma^{-1} X$ and the cost term:

$$\min_{\mathbf{x}\in\varOmega_o^N,\ \mathbf{u}\in\mathbb{R}_+^N,\ N\in\mathbb{N}} [\varPsi(X^*\varSigma^{-1}X) + G(\|\mathbf{u}\|_1)] \quad s.t. \quad (Xq)_i = (\mathcal{O}(x_i), q)_Q, \quad \varSigma_{ij}^{-1} = \delta_{ij}\mathbf{u}_i, \tag{3.12}$$

for all $q \in Q$ and $i, j = 1, \dots, N$. For the concrete assumptions on $\varPsi$ and G, we refer to the following section. Let N be fixed for the moment. Note that, despite of the convexity of $\varPsi$, the dependence of the Fisher information operator on the pair $(\mathbf{x}, \mathbf{u})$ is in general non-convex. Thus (3.12) may admit a large number of local extrema which are not necessarily minima. Additionally the unknown optimal number of sensors as well as the, possibly complicated, geometry of $\varOmega_o$ may aggravate its algorithmic treatment. As a consequence, even if we knew that this sensor placement problem admits a global minimizer its direct computation is in most cases infeasible.

As a remedy we consider the sparse sensor placement problem

$$\min_{u\in\mathcal{M}^+(\varOmega_o)} [\varPsi(\mathcal{I}(u)) + G(\|u\|_{\mathcal{M}})], \tag{$\mathcal{P}$}$$

where we minimize for the measure $u \in \mathcal{M}^+(\varOmega_o)$ instead of the measurement setup $(x, \mathbf{u}, N)$. In contrast, due to the linearity of the Fisher information operator $\mathcal{I}$, the mapping

$$\varPsi \circ \mathcal{I} \colon \mathcal{M}^+(\varOmega_o) \to \mathbb{R}, \quad u \mapsto \varPsi(\mathcal{I}(u)),$$

is convex. Thus, $(\mathcal{P})$ is a convex optimization problem on the space of Borel measures $\mathcal{M}(\varOmega_o)$ and each of its extrema is a global minimum.

Let us clarify the connection between these two, seemingly different, approaches. Given a measurement setup $(\mathbf{x}, \mathbf{u}, N)$ we obtain that the corresponding sparse design measure $u = \sum_{i=1}^N \mathbf{u}_i \delta_{x_i}$ fulfills

$$u \in \operatorname{cone}\{\, \delta_x \mid x \in \varOmega_o \,\} = \left\{ u = \sum_{i=1}^N \mathbf{u}_i \delta_{x_i} \mid N \in \mathbb{N},\ \mathbf{u} \in \mathbb{R}_+^N,\ x \in \varOmega_o^N \right\}.$$

Furthermore we observe

$$\|u\|_{\mathcal{M}} = \|\mathbf{u}\|_1 = \sum_{i=1}^N \mathbf{u}_i.$$

Consequently, instead of minimizing with respect to the number and positions of the sensors as well as the measurement weights, we can directly minimize for the design measure:

$$\min_{u\in\mathcal{M}(\varOmega_o)} [\varPsi(\mathcal{I}(u)) + G(\|u\|_{\mathcal{M}})] \quad s.t. \quad u \in \operatorname{cone}\{\, \delta_x \mid x \in \varOmega_o \,\}. \tag{3.13}$$

Up to now we have not discussed whether (3.12) or, equivalently, (3.13) admit optimal solutions. As a matter of fact, this is not clear a priori, since the set of admissible design measures is not sequentially compact with respect to a suitable topology on $\mathcal{M}^+(\varOmega_o)$. To obtain an, a priori, well-posed problem we therefore replace the cone of all Dirac delta functions by its closure with respect to the weak* topology obtaining

$$\overline{\operatorname{cone}\{\,\delta_x \mid x \in \Omega_o\,\}}^* = \mathcal{M}^+(\Omega_o).$$

Hence, we arrive at $(\mathcal{P})$. Under reasonable assumptions on Ψ and G, existence of an optimal design measure $\bar{u} \in \mathcal{M}^+(\Omega_o)$ can be proven in this framework, see Section 3.2.3. In this light, the sparse sensor placement reformulation follows naturally from (3.12) by embedding it into a rigorous analytic framework.

To close this section, we briefly comment on some features of the sparse sensor placement approach that should be kept in mind throughout the following chapters. First, we stress that the existence of a sparse optimal solution $\bar{u} \in \operatorname{cone}\{\,\delta_x \mid x \in \Omega_o\,\}$, cannot be ensured in general. However it is straightforward to see that (3.12) admits an optimal solution $(\bar{\mathbf{x}}, \bar{\mathbf{u}}, N)$ if and only if the corresponding design measure $\bar{u} = \sum_{i=1}^N \bar{\mathbf{u}}_i \delta_{\bar{x}_i}$ minimizes in $(\mathcal{P})$. From this perspective both formulations can be seen as equal with the crucial difference that $(\mathcal{P})$ is convex. In Section 3.2.4 we review conditions that guarantee the existence of an optimal design measure consisting of finitely many Dirac delta functions, making both approaches essentially equivalent in these cases. In particular, this is the case if Q is finite dimensional.

Furthermore, recall that all statistical arguments were made under the assumption that the number of measurements is finite and the measurement errors are independently distributed. For sparse senor placement problems we can construct simple examples, see Example 4.2, admitting optimal measurement designs which are distributed functions. From a statistical viewpoint it is up to now unclear how to interpret non-sparse optimal designs. However we stress that any such design can be approximated, in the weak* sense, by a finite combination of Dirac deltas up to arbitrary accuracy.

Last, in many works on optimal sensor placement with finite candidate set Ω_o, additional 0-1 constraints on the measurement weights are imposed. These might result from a binary interpretation of the weight where 1 corresponds to taking a measurement at the sensor location and 0 means neglecting it. Usually, these conditions are relaxed, yielding box constraints on the vector of measurement weights. That is, we require $0 \leq \mathbf{u}_i \leq 1$, $i = 1, \dots, N$. For general sets Ω_o we now outline that such additional constraints on the magnitude of the measurement weights are not meaningful. This stems back to the fact that weight-constrained sensors tend to cluster. To highlight this fact mathematically let us consider the observational domain Ω_o as the closure of a bounded domain in $\mathbb{R}^d$. On Ω_o we consider conic combinations of pairwise different Dirac delta functions with bounded coefficients

$$\mathcal{M}^+_{\text{const}}(\Omega_o) = \left\{ u \in \mathcal{M}^+(\Omega_o) \mid u = \sum_{i=1}^N \mathbf{u}_i \delta_{x_i}, \quad x_i \neq x_j,\ 0 \leq \mathbf{u}_i \leq 1,\ i = 1, \dots, N,\ N \in \mathbb{N} \right\}.$$

Since this set is not sequentially weak* compact the sensor placement problem

$$\min_{u \in \mathcal{M}^+(\Omega_o)} [\Psi(\mathcal{I}(u)) + G(\|u\|_{\mathcal{M}})] \quad s.t. \quad u \in \mathcal{M}^+_{\text{const}}(\Omega_o), \tag{3.14}$$

is not well-posed in general. To identify the weak* closure of $\mathcal{M}^+_{\text{const}}(\Omega_o)$ let

$$u = \sum_{i=1}^N \mathbf{u}_i \delta_{x_i} \in \operatorname{cone}\{\,\delta_x \mid x \in \Omega_o\,\}, \quad x_i \neq x_j,\ i, j \in \{1, \dots, N\},$$

be given. Fix an index $i = 1, \dots, N$. For $R > 0$ the intersection $\Omega_o \cap B_R(x_i)$ is nonempty and has nonzero Lebesgue measure. Thus there exists a measure $u_i^R \in \mathcal{M}^+_{\text{const}}(\Omega_o)$ with

$$\max_{x \in \operatorname{supp} u_i^R} |x - x_i| \leq R, \quad \|u_i^R\|_{\mathcal{M}} = \mathbf{u}_i.$$

Letting R tend to zero we conclude $u_i^R \rightharpoonup^* \mathbf{u}_i \delta_{x_i}$. Repeating this argument for each support point and choosing R small enough we get $\operatorname{supp} u_i^R \cap \operatorname{supp} u_j^R = \emptyset$ for all $i = 1, \dots, N$ as well as

$$u^R = \sum_{i=1}^N u_i^R \in \mathcal{M}^+_{\text{const}}(\Omega_o), \quad u^R \rightharpoonup^* u.$$

We deduce

$$\mathcal{M}^+(\Omega_o) = \overline{\operatorname{cone}\{\, \delta_x \mid x \in \Omega_o \,\}}^* \subset \overline{\mathcal{M}^+_{\text{const}}(\Omega_o)}^* \subset \mathcal{M}^+(\Omega_o).$$

Consequently, replacing $\mathcal{M}^+_{\text{const}}(\Omega_o)$ by its weak* closure in (3.14) we again arrive at $(\mathcal{P})$.

The preceding discussion specifically implies that sensor placement problems with $0-1$ constraints on the measurement weights and no further restrictions on the number and positions of sensors are not well-posed in general. One particular reason for this shortcoming is the assumption on the independence of the measurement errors: Since multiple measurements at the same point do not correlate taking several measurements at a single point is favorable and thus constrained sensors are put arbitrarily close together.

3.2.3 Existence of optimal designs and optimality conditions

In this section we state assumptions on the optimal design criterion Ψ and the regularization term which ensure the well-posedness of the sparse sensor placement problem. Subsequently, the existence of solutions as well as first order necessary and sufficient optimality conditions for the sparse optimal design problem $(\mathcal{P})$ are provided.

Let us first elaborate further on the Fisher information operator $\mathcal{I}$. For a rigorous analysis of the sparse sensor placement problem we will require that $\mathcal{I}$ maps weak* convergent sequences in $\mathcal{M}^+(\Omega_o)$ to norm convergent sequences in its image space. While this trivially holds if Q is finite dimensional this needs additional attention in the general case. First we therefore characterize the Banach space adjoint of the Fisher operator $\mathcal{I}$ defined in (3.4).

Proposition 3.7. *The Fisher operator $\mathcal{I}$ is the Banach space adjoint of the operator*

$$\mathcal{I}^* \colon \operatorname{SHS}(Q,Q) \to \mathcal{C}(\Omega_o), \quad B \mapsto \varphi_B, \tag{3.15}$$

where the continuous function φ_B is given by $\varphi_B(x) = (\mathcal{O}(x), B\mathcal{O}(x))_Q$ for every $x \in \Omega_o$.

Proof. Let $B \in \operatorname{SHS}(Q,Q)$ and $u \in \mathcal{M}^+(\Omega_o)$ be given. Due to the linearity of the trace operator and B we get

$$\langle\langle \mathcal{I}(u), B \rangle\rangle_{\operatorname{HS}(Q,Q)} = \operatorname{Tr}_Q(B\mathcal{I}(u)) = \int_{\Omega_o} \operatorname{Tr}_Q(BI(x))\, \mathrm{d}u(x) = \langle \mathcal{I}^* B, u \rangle,$$

using the properties of the Bochner integral. Denote by $\{\phi_i\}_{i\in\mathbf{I}}$, $\mathbf{I} \subset \mathbb{N}$, an orthonormal basis of Q. We have $B\mathcal{O}(x) \in Q$ and consequently

$$B\mathcal{O}(x) = \sum_{i\in\mathbf{I}} (B\mathcal{O}(x), \phi_i)_Q \phi_i(x) \quad \forall x \in \Omega_o.$$

Further calculations show that

$$\begin{aligned} \operatorname{Tr}_Q(BI(x)) &= \sum_{i\in\mathbf{I}} (\phi_i, BI(x)\phi_i)_Q = \sum_{i\in\mathbf{I}} (B\mathcal{O}(x), \phi_i)_Q (\mathcal{O}(x), \phi_i)_Q \\ &= (B\mathcal{O}(x), \mathcal{O}(x))_Q \end{aligned} \tag{3.16}$$

where we used $I(x) = \mathcal{O}(x) \otimes \mathcal{O}(x)$ for all $x \in \Omega_o$. Hence, we identify $\mathcal{I}^*B$ with the continuous function φ_B, where $\varphi_B(x) = \operatorname{Tr}_Q(BI(x)) = (B\mathcal{O}(x), \mathcal{O}(x))_Q$. This gives the statement. □

As an immediate consequence we obtain the weak*-to-strong continuity of the Fisher information operator $\mathcal{I}$.

Theorem 3.8. *The Fisher-information mapping $\mathcal{I}\colon \mathcal{M}(\Omega_o) \to \mathrm{SHS}(Q,Q)$ is weak*-to-strong sequentially continuous, i.e. given $\{u_k\}_{k\in\mathbb{N}} \subset \mathcal{M}(\Omega_o)$ there holds*

$$u_k \rightharpoonup^* u \Rightarrow \mathcal{I}(u_k) \to \mathcal{I}(u),$$

in $\mathrm{SHS}(Q,Q)$.

Proof. Let any weak* convergent sequence $\{u_k\} \subset \mathcal{M}(\Omega_o)$ with $u_k \rightharpoonup^* u$, $u \in \mathcal{M}(\Omega_o)$ be given. We obtain

$$\langle\langle \mathcal{I}(u_k), B\rangle\rangle_{\mathrm{HS}(Q,Q)} = \operatorname{Tr}_Q(\mathcal{I}(u_k)B) = \langle \mathcal{I}^*B, u_k\rangle,$$

for all $B \in \mathrm{SHS}(Q,Q)$. Since $\mathcal{I}^*B \in \mathcal{C}(\Omega_o)$ we conclude

$$\lim_{k\to\infty} \langle\langle \mathcal{I}(u_k), B\rangle\rangle_{\mathrm{HS}(Q,Q)} = \langle \mathcal{I}^*B, u\rangle = \langle\langle \mathcal{I}(u), B\rangle\rangle_{\mathrm{HS}(Q,Q)}.$$

Thus $\mathcal{I}$ is weak*-to-weak continuous. Due to the linearity of the Bochner integral we further calculate

$$\begin{aligned} \|\mathcal{I}(u_k)\|^2_{\mathrm{HS}(Q,Q)} &= \operatorname{Tr}_Q(\mathcal{I}(u_k)\mathcal{I}(u_k)) \\ &= \int_{\Omega_o} (\mathcal{O}(x), \mathcal{I}(u_k)\mathcal{O}(x))_Q \,\mathrm{d}u_k(x) \\ &= \int_{\Omega_o}\int_{\Omega_o} (\mathcal{O}(x), \mathcal{O}(y))^2_Q \,\mathrm{d}u_k(y)\,\mathrm{d}u_k(x). \end{aligned}$$

Define $j \in \mathcal{C}(\Omega_o \times \Omega_o)$ by

$$z\colon \Omega_o \times \Omega_o \to \mathbb{R}, \quad (x,y) \mapsto (\mathcal{O}(x), \mathcal{O}(y))^2_Q.$$

By $\mathcal{B}(\Omega_o) \otimes \mathcal{B}(\Omega_o)$ we denote the tensor-product σ-algebra on the cartesian product $\Omega_o \times \Omega_o$ and $u_k \times u_k$ is given as the unique product measure of u_k with itself on $(\Omega_o \times \Omega_o, \mathcal{B}(\Omega_o) \otimes \mathcal{B}(\Omega_o))$.

We show that $u_k \times u_k \rightharpoonup^* u \times u$ in $\mathcal{M}(\Omega_o \times \Omega_o)$. Therefore note that the span of all functions $f \in \mathcal{C}(\Omega_o \times \Omega_o)$ given by

$$f(x,y) = g(x)h(y), \quad g,h \in \mathcal{C}(\Omega_o),\ x,y \in \Omega_o,$$

is dense in $\mathcal{C}(\Omega_o \times \Omega_o)$, see [226]. Given a finite linear combination of such functions

$$f_n(x,y) = \sum_{i=1}^{n} g_i(x)h_i(y), \quad n \in \mathbb{N},\ g_i,h_i \in \mathcal{C}(\Omega_o),\ x,y \in \Omega_o,\ i = 1,\dots,n,$$

we obtain

$$\langle f_n, u_k \times u_k \rangle_{\mathcal{C}(\Omega_o \times \Omega_o),\mathcal{M}(\Omega_o \times \Omega_o)} = \sum_{i=1}^{n} [\langle g_i, u_k \rangle \langle h_i, u_k \rangle].$$

Passing to the limit for $k \to \infty$ on both sides yields

$$\lim_{k\to\infty} \langle f_n, u_k \times u_k \rangle_{\mathcal{C}(\Omega_o \times \Omega_o),\mathcal{M}(\Omega_o \times \Omega_o)} = \sum_{i=1}^{n} [[\langle g_i, u \rangle \langle h_i, u \rangle] = \langle f_n, u \times u \rangle_{\mathcal{C}(\Omega_o \times \Omega_o),\mathcal{M}(\Omega_o \times \Omega_o)}$$

This gives the desired statement since weak* convergence was tested against a dense subset. We proceed by calculating the limit

$$\begin{aligned}\lim_{k\to\infty} \langle z, u_k \times u_k \rangle_{\mathcal{C}(\Omega_o \times \Omega_o),\mathcal{M}(\Omega_o \times \Omega_o)} &= \int_{\Omega_o}\int_{\Omega_o} (\mathcal{O}(x), \mathcal{O}(y))_Q^2 \,\mathrm{d}u(y)\mathrm{d}u(x) \\ &= \operatorname{Tr}_Q(\mathcal{I}(u)\mathcal{I}(u)) = \|\mathcal{I}(u)\|^2_{\mathrm{HS}(Q,Q)}.\end{aligned}$$

By expanding we derive

$$\lim_{k\to\infty} \|\mathcal{I}(u_k) - \mathcal{I}(u)\|^2_{\mathrm{HS}(Q,Q)} = \lim_{k\to\infty} [\|\mathcal{I}(u_k)\|^2_{\mathrm{HS}(Q,Q)} - \langle\langle \mathcal{I}(u_k), \mathcal{I}(u) \rangle\rangle_{\mathrm{HS}(Q,Q)} + \|\mathcal{I}(u)\|^2_{\mathrm{HS}(Q,Q)}] = 0,$$

where we used the weak convergence of $\mathcal{I}(u_k)$ and the strong convergence of $\|\mathcal{I}(u_k)\|^2_{\mathrm{HS}(Q,Q)}$. This finishes the proof. □

Concerning the design criterion and the regularization term in the optimal design problem, the following assumptions are made.

Assumption 3.2. The functional $\Psi \colon \mathrm{SHS}(Q,Q) \to \mathbb{R} \cup \{+\infty\}$ satisfies:

A3.1 Ψ is convex and lower semi-continuous on $\mathrm{Pos}(Q,Q)$.

A3.2 The domain of Ψ in $\mathrm{Pos}(Q,Q)$ is nonempty and open in $\mathrm{Pos}(Q,Q)$ with respect to the topology induced by the Hilbert-Schmidt norm i.e. given a sequence $\{B_k\}_{k\in\mathbb{N}} \subset \mathrm{Pos}(Q,Q)$ there holds

$$B_k \to B \in \operatorname{dom}_{\mathrm{Pos}(Q,Q)} \Psi \Rightarrow B_k \in \operatorname{dom}_{\mathrm{Pos}(Q,Q)} \Psi$$

for all $k \in \mathbb{N}$ large enough. Furthermore Ψ is continuously differentiable on its domain. The gradient of Ψ at B is denoted by $\nabla\Psi(B) \in \mathrm{SHS}(Q,Q)$.

A3.3 $\varPsi$ is monotonous in the following sense:

$$B_2 - B_1 \in \operatorname{Pos}(Q,Q) \Rightarrow \varPsi(B_2) \leq \varPsi(B_1),$$

for all $B_1,\ B_2 \in \operatorname{Pos}(Q,Q)$.

Assumption 3.3. The function $G\colon \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is proper, convex and lower semi-continuous. Furthermore it is monotonically increasing on $\mathbb{R}_+$ with $\lim_{t\to\infty} G(t) = +\infty$. There holds $\operatorname{dom} G \subset \mathbb{R}_+$.

While (**A3.1**), (**A3.2**) and Assumption 3.3 will ensure the well-posedness of $(\mathcal{P})$, the third assumption, (**A3.3**), can be practically motivated. For example, given design measures $u_1, u_2 \in \mathcal{M}^+(\Omega_o)$ and $\lambda \geq 1$, we conclude

$$\varPsi(\mathcal{I}(u_1 + u_2)) \leq \varPsi(\mathcal{I}(u_1)), \quad \varPsi(\mathcal{I}(\lambda u_1)) \leq \varPsi(\mathcal{I}(u_1)).$$

Hence, adding new measurements or increasing the measurement weights decreases the value of the design criterion. Thus the monotonicity assumption is reasonable since acquiring more or better data should improve the estimator. A more geometric interpretation of (**A3.3**) is given in Chapter 4.

Remark 3.1. While the assumption on the openness of the domain of $\varPsi$ might seem unusual at first sight the following example demonstrates its necessity. In the finite dimensional case, $Q = \mathbb{R}^n$, we consider the A-optimal design criterion

$$\varPsi_A(B) = \begin{cases} \operatorname{Tr}_{\mathbb{R}^n}(B^{-1}) & B \in \operatorname{PD}(n) \\ +\infty & \text{else} \end{cases},$$

for $B \in \operatorname{Sym}(n)$. Here $\operatorname{PD}(n)$ denotes the set of positive definite matrices. It is readily verified that its domain is given by $\operatorname{dom} \varPsi_A = \operatorname{PD}(n)$ which is open in the set of non-negative definite matrices.

Remark 3.2. Please note that we require the functional G to be equal to $+\infty$ on $(-\infty, 0)$. Clearly, this poses no restriction since for a proper, convex and lower semi-continuous function $\widehat{G}\colon \mathbb{R} \to \mathbb{R}$, monotonically increasing on $\mathbb{R}_+$ with $\lim_{t\to\infty} G(t) = +\infty$, we may define $G = \widehat{G} + I_{[0,\infty)}$ and note that

$$\psi(u) + \widehat{G}(\|u\|_{\mathcal{M}}) = \psi(u) + G(\|u\|_{\mathcal{M}}) \quad \forall u \in \mathcal{M}^+(\Omega_o).$$

The restriction of the domain of G is used to obtain compact necessary first-order necessary and sufficient optimality conditions without distinguishing between the cases $\|\bar{u}_\beta\|_{\mathcal{M}} = 0$ and $\|\bar{u}_\beta\|_{\mathcal{M}} > 0$. In particular, we stress that Assumption 3.3 allows to consider norm regularization

$$G_1(\|u\|_{\mathcal{M}}) = \beta\|u\|_{\mathcal{M}} = \beta\|u\|_{\mathcal{M}} + I_{[0,\infty)}(\|u\|_{\mathcal{M}})$$

for some $\beta > 0$ as well as norm constraints

$$G_2(\|u\|_{\mathcal{M}}) = I_{[0,K]}(\|u\|_{\mathcal{M}}), \quad K > 0,$$

in a unified framework.

Now, we formulate the reduced design problem $(\mathcal{P})$ as

$$\min_{u \in \mathcal{M}^+(\Omega_o)} F(u) = [\psi(u) + G(\|u\|_{\mathcal{M}})],$$

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

Proposition 3.9. *Let Ψ be given and let Assumptions (**A3.1**)–(**A3.3**) be fulfilled. The functional ψ satisfies:*

1. *For every $u \in \mathcal{M}^+(\Omega_o)$ there holds $\mathcal{I}(u) \in \operatorname{Pos}(Q, Q)$.*
2. *There holds*

$$\operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi = \left\{ u \in \mathcal{M}^+(\Omega_o) \mid \mathcal{I}(u) \in \operatorname{dom}_{\operatorname{Pos}(Q,Q)} \Psi \right\}.$$

Furthermore the domain is weak sequentially open: Given $\{u_k\}_{k\in\mathbb{N}} \subset \mathcal{M}^+(\Omega_o)$ we have*

$$u^k \rightharpoonup^* \bar{u} \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi \Rightarrow \exists K \in \mathbb{N}\colon u^k \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi, \quad k \geq K.$$

3. *The functional ψ is continuously Fréchet differentiable on $\operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi$. For a design measure $u \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi$ and $\delta u \in \mathcal{M}(\Omega_o)$ the directional derivative $\psi'(u)(\delta u)$ is given by*

$$\psi'(u)(\delta u) = \langle\langle \mathcal{I}(\delta u), \nabla\Psi(\mathcal{I}(\delta u))\rangle\rangle_{\operatorname{HS}(Q,Q)} = \operatorname{Tr}_Q(\mathcal{I}(u)\nabla\Psi(\mathcal{I}(u))).$$

The derivative $\psi'(u) \in \mathcal{M}(\Omega_o)^*$ can be identified with the continuous function

$$\nabla\psi(u)(x) = \mathcal{I}^*\nabla\Psi(\mathcal{I}(u))(x) = (\mathcal{O}(x), \nabla\Psi(\mathcal{I}(u))\mathcal{O}(x))_Q \quad \forall x \in \Omega_o. \tag{3.17}$$

Moreover the gradient $\nabla\psi\colon \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi \to C(\Omega_o)$ is weak-to-strong continuous.*

4. *ψ is weak* lower semi-continuous and convex on $\mathcal{M}^+(\Omega_o)$.*
5. *ψ is monotone in the sense that*

$$\mathcal{I}(u_2 - u_1) \in \operatorname{Pos}(Q, Q) \Rightarrow \psi(u_1) \geq \psi(u_2) \quad \forall u_1,\ u_2 \in \mathcal{M}^+(\Omega_o).$$

Proof. The first claim can be found in Corollary 3.6. The sequential openness of $\operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi$ follows from the openness of the domain of Ψ in $\operatorname{Pos}(Q,Q)$ and the weak*-to-strong continuity of $\mathcal{I}$. For a given measure $u \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi$ the differentiability of ψ follows from assumption (**A3.2**) by applying the chain rule. Using (3.15) we obtain

$$\psi'(u)(\delta u) = \langle\langle \mathcal{I}(\delta u), \nabla\Psi(\mathcal{I}(u))\rangle\rangle_{\operatorname{HS}(Q,Q)} = \operatorname{Tr}_Q(\nabla\Psi(\mathcal{I}(u))\mathcal{I}(\delta u)) = \langle \mathcal{I}^*\nabla\Psi(\mathcal{I}(u)), \delta u\rangle,$$

for every $\delta u \in \mathcal{M}(\Omega_o)$. Hence we identify $\psi'(u) \in \mathcal{M}(\Omega_o)^*$ with the continuous function

$$\nabla\psi(u) = \mathcal{I}^*\nabla\Psi(\mathcal{I}(u)) \in \mathcal{C}(\Omega_o).$$

Additionally, we directly see that the mapping

$$\nabla\psi\colon \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi \to \mathcal{C}(\Omega_o), \quad u \mapsto \nabla\psi(u),$$

is weak*-to-strong continuous, using the continuity of $\nabla\Psi$ and $\mathcal{I}^*$ as well as the weak*-to-strong continuity of $\mathcal{I}$. Statements 4. and 5. can be derived directly from Assumptions (**A3.1**) and (**A3.3**) using $\mathcal{I}(u_k) \in \operatorname{Pos}(Q,Q)$ and $\mathcal{I}(u_k) \to \mathcal{I}(u)$ in $\operatorname{SHS}(Q,Q)$ for every sequence $\{u_k\}_{k\in\mathbb{N}} \subset \mathcal{M}^+(\Omega_o)$ with weak* limit u. □

We derive the following result on the gradient $\nabla\psi$ by imposing further regularity assumptions on the design criterion.

Lemma 3.10. *Assume that Ψ is two times continuously Fréchet differentiable on its domain in $\operatorname{Pos}(Q,Q)$. For every $u \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi$ there holds $\nabla\psi(u)(x) \leq 0$ for all $x \in \Omega_o$. We further have*

$$-\nabla\psi(u)(x) = -(\mathcal{O}(x), \nabla\Psi(\mathcal{I}(u))\mathcal{O}(x))_Q = \|(-\nabla\Psi(\mathcal{I}(u)))^{1/2}\mathcal{O}(x)\|_Q^2, \tag{3.18}$$

for all $x \in \Omega_o$. Here, $(-\nabla\Psi(\mathcal{I}(u)))^{1/2} \in \mathcal{L}(Q,Q)$ denotes the uniquely determined positive square root of $-\nabla\Psi(\mathcal{I}(u))$.

Proof. Recall that $\nabla\psi(u)(x) = (\mathcal{O}(x), \nabla\Psi(\mathcal{I}(u))\mathcal{O}(x))_Q$ for all $x \in \Omega_o$. Let an arbitrary but fixed $B \in \operatorname{dom}_{\operatorname{Pos}(Q,Q)} \Psi$ and $z \in Q$ be given. For all $\varepsilon > 0$ small enough we have

$$B + \varepsilon[z \otimes z] \in \operatorname{dom}_{\operatorname{Pos}(Q,Q)} \Psi,$$

due to the openness assumption on the domain of Ψ. Using Taylor approximation, we find

$$\Psi(B + \varepsilon[z \otimes z]) = \Psi(B) + \varepsilon \operatorname{Tr}_Q(\nabla\Psi(B)[z \otimes z]) + r(\varepsilon),$$

where the remainder term fulfills $\lim_{\varepsilon\to0}[r(\varepsilon)/\varepsilon] = 0$. As in (3.16) we derive

$$\operatorname{Tr}_Q(\nabla\Psi(B)[z \otimes z]) = (\nabla\Psi(B)z, z)_Q.$$

Since Ψ is monotone in the sense of (**A**3.3) and $z \otimes z \in \operatorname{Pos}(Q,Q)$, we obtain

$$0 \geq \Psi(B + \varepsilon[z \otimes z]) - \Psi(B) = \varepsilon \operatorname{Tr}_Q(\nabla\Psi(B)[z \otimes z]) + r(\varepsilon).$$

Dividing both sides by $\varepsilon > 0$ and passing to the limit for $\varepsilon \to 0$ we conclude

$$(\nabla\Psi(B)z, z)_Q \leq 0 \quad \forall z \in Q.$$

The first statement follows by setting $B = \mathcal{I}(u)$ and $z = \mathcal{O}(x)$ for every $x \in \Omega_o$.
Furthermore this implies $-\nabla\Psi(\mathcal{I}(u)) \in \operatorname{Pos}(Q,Q)$. Consequently there exists a unique positive operator $(-\Psi(\mathcal{I}(u)))^{1/2}$ with $-\Psi(\mathcal{I}(u)) = \left((-\Psi(\mathcal{I}(u)))^{1/2}\right)^2$, see [33]. Given an arbitrary $x \in \Omega_o$ we obtain

$$\begin{aligned}
-\nabla\psi(u)(x) &= -(\mathcal{O}(x), \nabla\Psi(\mathcal{I}(u))\mathcal{O}(x))_Q \\
&= \Big((-\nabla\Psi(\mathcal{I}(u)))^{1/2}\mathcal{O}(x), (-\nabla\Psi(\mathcal{I}(u)))^{1/2}\mathcal{O}(x)\Big)_Q \\
&= \|(-\nabla\Psi(\mathcal{I}(u)))^{1/2}\mathcal{O}(x)\|_Q^2,
\end{aligned}$$

which finishes the proof. □

In the following we also assume that the sum of the reduced design criterion and the regularization term is radially unbounded. Clearly, this assumption is fulfilled, e.g., if ψ is bounded from below on $\mathcal{M}^+(\Omega_o)$. We will comment on this additional assumption for the most popular choices of Ψ in the subsequent chapters.

Assumption 3.4. The functional F is radially unbounded on $\mathcal{M}^+(\Omega_o)$: Given a sequence $\{u_k\}_{k\in\mathbb{N}} \subset \mathcal{M}^+(\Omega_o)$ we have

$$\|u_k\|_{\mathcal{M}} \to \infty \Rightarrow F(u_k) \to \infty.$$

The existence of at least one global minimizer $\bar{u} \in M^+(\Omega)$ to the reduced formulation is now obtained by standard arguments. We give a proof for the sake of completeness.

Proposition 3.11. *Assume that* $\operatorname{dom}_{\mathcal{M}^+(\Omega_o)} F$ *is not empty, i.e.*

$$\operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi \cap \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} G(\|\cdot\|_{\mathcal{M}}) \neq \emptyset.$$

There exists at least one optimal solution $\bar{u}$ to $(\mathcal{P})$ and the set of minimizers to $(\mathcal{P})$ is uniformly bounded. If Ψ is strictly convex on $\operatorname{Pos}(Q,Q)$ *then the optimal Fisher information is unique.*

Proof. Since $\operatorname{dom}_{\mathcal{M}^+(\Omega_o)} j$ is not empty, there exists $u \in \mathcal{M}^+(\Omega_o)$ and an infimizing sequence of design measures $\{u_k\}_{k\in\mathbb{N}}$ with

$$F(u) < \infty, \quad F(u_k) \to \inf_{u \in M^+(\Omega_o)} F(u) < \infty.$$

W.l.o.g we assume that $u_k \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} j$ for all $k \in \mathbb{N}$. Since j is radially unbounded the sequence $\{u_k\}_{k\in\mathbb{N}}$ is bounded. Applying the sequential version of Banach-Alaoglu theorem, it admits a subsequence denoted by the same symbol with $u^k \rightharpoonup^* \bar{u} \in \mathcal{M}^+(\Omega_o)$. Since Ψ is lower semi-continuous on $\operatorname{Pos}(Q,Q)$ we conclude

$$\Psi(\mathcal{I}(\bar{u})) \leq \liminf_{k\to\infty} \Psi(\mathcal{I}(u_k)), \quad \|u_k\|_{\mathcal{M}} = \langle 1, u_k \rangle \to \langle 1, \bar{u} \rangle = \|\bar{u}\|_{\mathcal{M}},$$

from the weak* convergence of $\{u_k\}_{k\in\mathbb{N}}$ and the weak*-to-strong continuity of $\mathcal{I}$. Combining these results yields

$$F(\bar{u}) \leq \liminf_{k\to\infty} F(u_k) = \inf_{u\in\mathcal{M}^+(\Omega_o)} F(u),$$

and thus the optimality of $\bar{u}$. The uniform bound on the norm of the minimizers follows from the radial unboundedness of F. In the case of strictly convex Ψ uniqueness of the Fisher information follows by a standard argument. □

This proposition does not give any statement on the structure of the optimal design measure $\bar{u}$ as well as its sparsity pattern. Indeed, from the previous discussions, it is not even clear whether there exists an admissible, sparse, design measure. This is however addressed in the following corollary.

Corollary 3.12. *Assume that* $\operatorname{dom}_{\mathcal{M}^+(\Omega_o)} F$ *is not empty. Then there exists*

$$\tilde{u} \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} F \cap \operatorname{cone}\{\, \delta_x \mid x \in \Omega_o \,\}.$$

Proof. Let $u \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi$ be given. Following the arguments in [50, Appendix A], there exists a sequence of positive measures $\{u_k\}_{k\in\mathbb{N}}$ with

$$u_k \in \operatorname{cone}\{\, \delta_x \mid x \in \Omega_o \,\}, \ u_k \rightharpoonup^* u, \quad \|u_k\|_{\mathcal{M}} \leq \|u\|_{\mathcal{M}},$$

for all $k \in \mathbb{N}$. Since the Fisher operator $\mathcal{I}$ is weak*-to-strong continuous we additionally get $\mathcal{I}(u_k) \to \mathcal{I}(u)$ in $\operatorname{SHS}(Q,Q)$ as $k \to \infty$. Thus we have $\mathcal{I}(u_k) \in \operatorname{dom}_{\operatorname{Pos}(Q,Q)} \Psi$ for all k large enough due to the openness assumption on the domain of Ψ in $\operatorname{Pos}(Q,Q)$. Since G is monotonically increasing we conclude $u_k \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} F$ for all $k \in \mathbb{N}$ large enough. □

Using the differentiability and convexity assumptions on Ψ we proceed to derive necessary and sufficient optimality conditions.

Proposition 3.13. *Let $\bar{u} \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} F$ be given. Then $\bar{u}$ is an optimal solution to $(\mathcal{P})$ if and only if*

$$\langle -\nabla\psi(\bar{u}), u-\bar{u}\rangle + G(\|\bar{u}\|_{\mathcal{M}}) \leq G(\|u\|_{\mathcal{M}}) \quad \forall u \in \mathcal{M}^+(\Omega_o). \tag{3.19}$$

Proof. Recall that the Fréchet derivative of ψ at $\bar{u}$ can be identified with the continuous function $\nabla\psi(\bar{u})$. Hence, following Proposition 6.3, a measure $\bar{u} \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} F$ is optimal if and only if

$$-\nabla\psi(\bar{u}) \in \partial(G(\|\cdot\|_{\mathcal{M}}) + I_{u\geq 0}(\cdot))(\bar{u}),$$

where the set on the right hand side denotes the convex subdifferential of the function

$$G(\|\cdot\|_{\mathcal{M}}) + I_{u\geq 0}(\cdot),$$

at $\bar{u}$. By definition this is equivalent to (3.19). □

Equivalently minimizers of $(\mathcal{P})$ are given by the roots of the non-negative primal-dual gap functional $\Psi\colon \mathcal{M}(\Omega) \to \mathbb{R}_+ \cup \{+\infty\}$ which is given by

$$\Phi(u) = \begin{cases} \max_{v\in\mathcal{M}^+(\Omega_o)}[\langle\nabla\psi(u), u-v\rangle + G(\|u\|_{\mathcal{M}}) - G(\|v\|_{\mathcal{M}})] & u \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} F \\ +\infty & \text{else.} \end{cases}.$$

Proposition 3.14. *Let $\bar{u} \in \mathcal{M}^+(\Omega_o)$ be given. Then $\bar{u}$ is an optimal solution of $(\mathcal{P})$ iff*

$$\bar{u} \in \operatorname*{arg\,min}_{u\in\mathcal{M}(\Omega_o)} \Phi(u), \quad \Phi(\bar{u}) = 0.$$

Proof. By construction we have $\Phi(u) \geq 0$ for all $u \in \mathcal{M}(\Omega_o)$ and $\Phi(u) = +\infty$ for $u \notin \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} j$. Rearranging (3.19) yields the optimality of $\bar{u}$ if and only if

$$\langle\nabla\psi(\bar{u}), \bar{u}-u\rangle + G(\|\bar{u}\|_{\mathcal{M}}) - G(\|u\|_{\mathcal{M}}) \leq 0 \quad \forall u \in \mathcal{M}^+(\Omega_o).$$

Maximizing with respect to $u \in \mathcal{M}^+(\Omega_o)$ on both sides, we conclude that this is equivalent to $\Phi(\bar{u}) = 0$ and thus also $\Phi(\bar{u}) \leq \Phi(u)$ for all $u \in \mathcal{M}(\Omega_o)$. □

3.2.4 Structure of optimal measurement designs

In this section we will provide results on the structure of optimal measurement designs. In particular we provide a generalized version of the famous equivalence theorem due to Kiefer and Wolfowitz, see [164, 165], in Theorem 3.17 and prove the existence of optimal designs comprising finitely many points under certain conditions.

Due to the positive homogeneity of the norm, structural properties of $\bar{u}$ can be derived from the variational inequality (3.19). For this purpose, given a function $\varphi \in \mathcal{C}(\Omega_o)$ we recall the definition of its negative part as $[\varphi]^-(x) = -\min\{\varphi(x), 0\}$ for all $x \in \Omega_o$.

Proposition 3.15. *Let an optimal design $\bar{u} \in \mathcal{M}^+(\Omega_o)$ be given. The variational inequality* (3.19) *is equivalent to*

$$\|[\nabla\psi(\bar{u})]^-\|_{\mathcal{C}} \in \partial G(\|\bar{u}\|_{\mathcal{M}}), \quad \langle -\nabla\psi(\bar{u}), \bar{u}\rangle = \|[\nabla\psi(\bar{u})]^-\|_{\mathcal{C}}\|\bar{u}\|_{\mathcal{M}}, \tag{3.20}$$

where $\partial G(\|\bar{u}\|_{\mathcal{M}})$ denotes the subdifferential of the convex functional G at $\|\bar{u}\|_{\mathcal{M}}$ i.e.

$$\partial G(\|\bar{u}\|_{\mathcal{M}}) = \{\, \bar{m} \in \mathbb{R} \mid \bar{m}(m - \|\bar{u}\|_{\mathcal{M}}) + G(\|\bar{u}\|_{\mathcal{M}}) \leq G(m) \quad \forall m \in \mathbb{R} \,\}.$$

Proof. Assume that $\bar{u}$ satisfies (3.20). Then for an arbitrary measure $u \in \mathcal{M}^+(\Omega_o)$ there holds

$$\begin{aligned}\langle -\nabla\psi(\bar{u}), u - \bar{u}\rangle + G(\|\bar{u}\|_{\mathcal{M}}) &= -\|[\nabla\psi(\bar{u})]^-\|_{\mathcal{C}}\|\bar{u}\|_{\mathcal{M}} - \langle \nabla\psi(\bar{u}), u\rangle + G(\|\bar{u}\|_{\mathcal{M}}) \\ &\leq \|[\nabla\psi(\bar{u})]^-\|_{\mathcal{C}}(\|u\|_{\mathcal{M}} - \|\bar{u}\|_{\mathcal{M}}) + G(\|\bar{u}\|_{\mathcal{M}}) \\ &\leq G(\|u\|_{\mathcal{M}}),\end{aligned}$$

where we used $\langle -\nabla\psi(\bar{u}), \bar{u}\rangle = \|[\nabla\psi(\bar{u})]^-\|_{\mathcal{C}}\|\bar{u}\|_{\mathcal{M}}$ in the first equality and

$$\|[\nabla\psi(\bar{u})]^-\|_{\mathcal{C}} \in \partial G(\|\bar{u}\|_{\mathcal{M}}),$$

in the last inequality. This implies (3.19).
Conversely, assume that $\bar{u}$ fulfills (3.19). Due to the monotonicity of G there holds

$$\langle -\nabla\psi(\bar{u}), u - \bar{u}\rangle \leq 0 \quad \forall u \in \mathcal{M}^+(\Omega_o),\ \|u\|_{\mathcal{M}} \leq \|\bar{u}\|_{\mathcal{M}}.$$

Hence we conclude

$$-\nabla\psi(\bar{u}) \in \partial\left(I_{\|u\|_{\mathcal{M}} \leq \|\bar{u}\|_{\mathcal{M}}}(\cdot) + I_{u \geq 0}(\cdot)\right)(\bar{u}).$$

Using Proposition 6.4, this implies

$$\bar{u} \in \partial\left(I_{\|u\|_{\mathcal{M}} \leq \|\bar{u}\|_{\mathcal{M}}}(\cdot) + I_{u \geq 0}(\cdot)\right)^*(-\nabla\psi(\bar{u})),$$

as well as

$$\left(I_{\|u\|_{\mathcal{M}} \leq \|\bar{u}\|_{\mathcal{M}}}(\cdot) + I_{u \geq 0}(\cdot)\right)^*(-\nabla\psi(\bar{u}))) = \langle -\nabla\psi(\bar{u}), \bar{u}\rangle.$$

Let us calculate the convex conjugate

$$\begin{aligned}\left(I_{\|u\|_{\mathcal{M}} \leq \|\bar{u}\|_{\mathcal{M}}}(\cdot) + I_{u \geq 0}(\cdot)\right)^*(-\nabla\psi(\bar{u})) &= \sup_{\substack{u \in \mathcal{M}^+(\Omega_o) \\ \|u\|_{\mathcal{M}} \leq \|\bar{u}\|_{\mathcal{M}}}} \langle -\nabla\psi(\bar{u}), u\rangle = \sup_{\substack{u \in \mathcal{M}^+(\Omega_o) \\ \|u\|_{\mathcal{M}} \leq \|\bar{u}\|_{\mathcal{M}}}} \langle [\psi(\bar{u})]^-, u\rangle \\ &= \|[\nabla\psi(\bar{u})]^-\|_{\mathcal{C}}\|\bar{u}\|_{\mathcal{M}}.\end{aligned}$$

This gives the second part of (3.20). Consequently there holds

$$\langle -\nabla\psi(\bar{u}), u\rangle - \|[\nabla\psi(\bar{u})]^-\|_{\mathcal{C}}\|\bar{u}\|_{\mathcal{M}} + G(\|\bar{u}\|_{\mathcal{M}}) \leq G(\|u\|_{\mathcal{M}}) \quad \forall u \in \mathcal{M}^+(\Omega_o). \tag{3.21}$$

We distinguish the following cases. First assume that $\bar{u} \neq 0$. By testing (3.21) with the measure $\bar{u}_m = m/\|\bar{u}\|_{\mathcal{M}}\bar{u}$ for every $m \in \mathbb{R}_+$ we arrive at

$$\|[\nabla\psi(\bar{u})]^-\|_{\mathcal{C}}(m - \|\bar{u}\|_{\mathcal{M}}) + G(\|\bar{u}\|_{\mathcal{M}}) \leq G(m) \quad \forall m \in \mathbb{R}_+. \tag{3.22}$$

Since $\operatorname{dom} G \subset \mathbb{R}_+$ this yields $\|[\nabla\psi(\bar{u})]^-\|_{\mathcal{C}} \in \partial G(\|\bar{u}_\beta\|_{\mathcal{M}})$. If we have

$$\|\bar{u}\|_{\mathcal{M}} = \|[\nabla\psi(\bar{u})]^-\|_{\mathcal{C}} = 0,$$

then there holds $0 \in \partial G(0)$ due to the monotonicity of G on $\mathbb{R}_+$. Last we assume that $\|\bar{u}\|_{\mathcal{M}} = 0$ and $\|[\nabla\psi(\bar{u})]^-\|_{\mathcal{C}} \neq 0$. Then there holds

$$\|[\nabla\psi(\bar{u})]^-\|_{\mathcal{C}} = \max_{x\in\Omega_o} -\nabla\psi(\bar{u})(x).$$

Choose $\hat{x} \in \Omega_o$ with $-\nabla\psi(\bar{u})(\hat{x}) = \|[\nabla\psi(\bar{u})]^-\|_{\mathcal{C}}$. Testing (3.21) with $\bar{u}_m = m\delta_{\hat{x}}$ for $m \in \mathbb{R}_+$ we again arrive at (3.22). In all cases we thus conclude

$$\|[\nabla\psi(\bar{u})]^-\|_{\mathcal{C}} \in \partial G(\|\bar{u}\|_{\mathcal{M}}),$$

finishing the proof. □

Going one step further, the second condition in (3.20) can be equivalently reformulated as a condition on the support of the design measure.

Lemma 3.16. *Let $\varphi \in \mathcal{C}(\Omega_o)$ and $u \in \mathcal{M}^+(\Omega_o)$ be given. Then there holds*

$$\langle -\varphi, u\rangle = \|[\varphi]^-\|_{\mathcal{C}}\|u\|_{\mathcal{M}} \Leftrightarrow \operatorname{supp} u \subset \left\{\, x \in \Omega_o \mid -\varphi(x) = \|[\varphi]^-\|_{\mathcal{C}} \,\right\}. \tag{3.23}$$

Proof. Assume that the right side of the equivalence holds. Then we have

$$\langle -\varphi, u\rangle = \int_{\Omega_o} -\varphi \,\mathrm{d}u(x) = \int_{\Omega_o} \|[\varphi]^-\|_{\mathcal{C}} \,\mathrm{d}u(x) = \|[\varphi]^-\|_{\mathcal{C}}\|u\|_{\mathcal{M}}.$$

This proves the first direction. Conversely assume that $\langle -\varphi, u\rangle = \|[\varphi]^-\|_{\mathcal{C}}\|u\|_{\mathcal{M}}$ holds. Assume that $[\varphi]^- \neq 0$. In this case we obtain

$$\|[\varphi]^-\|_{\mathcal{C}} = -\min_{x\in\Omega_o} \varphi = \max_{x\in\Omega_o} -\varphi.$$

Let an arbitrary $x \in \Omega_o$ with $-\varphi < -\min_{x\in\Omega_o}\varphi$ be given. Due to the continuity of φ and a compactness argument there exists $\delta > 0$ with $-\varphi < -\min_{x\in\Omega_o}\varphi$ on $B_\delta(x) \subset \Omega_o$. For an arbitrary nonnegative $y \in \mathcal{C}_0(B_\delta(x))$ there exists $t > 0$ such that $\varphi - ty - \min_{x\in\Omega_o}\varphi \geq 0$. From this we conclude

$$0 \leq \langle \varphi - ty - \min_{x\in\Omega_o}\varphi, u\rangle = -\langle ty, u\rangle \leq 0,$$

due to the positivity of y and u. Therefore $u|_{B_\delta(x)} = 0$ and $B_\delta(x) \subset \Omega_o \backslash \operatorname{supp} u$. If $[\varphi]^- = 0$ we have $\varphi \geq 0$. Argumenting similar as before we conclude

$$-\varphi(x) = 0 \quad u - a.e.\ x \in \Omega_o.$$

By distinguishing between the two cases $u = 0$ and $u \neq 0$ we again arrive at the right hand side of (3.23). □

Collecting all the previous results the optimality of a design measure can be characterized through the following series of equivalences.

Theorem 3.17. *Assume that Ψ is two times Fréchet differentiable on its domain in* $\operatorname{Pos}(Q,Q)$*. Then the following statements are equivalent:*

- *The measure $\bar{u} \in \mathcal{M}^+(\Omega_o)$ is an optimal solution to $(\mathcal{P})$.*
- *There holds*

$$\langle -\nabla\psi(\bar{u}), u - \bar{u}\rangle + G(\|\bar{u}\|_{\mathcal{M}}) \leq G(\|u\|_{\mathcal{M}}) \quad \forall u \in \mathcal{M}^+(\Omega_o).$$

- *There holds*

$$-\min_{x\in\Omega_o} \nabla\psi(\bar{u})(x) \in \partial G(\|\bar{u}\|_{\mathcal{M}}), \quad \langle -\nabla\psi(\bar{u}), \bar{u}\rangle + \min_{x\in\Omega_o} \nabla\psi(\bar{u})(x)\|\bar{u}\|_{\mathcal{M}} = 0.$$

- *There holds*

$$-\min_{x\in\Omega_o} \nabla\psi(\bar{u})(x) \in \partial G(\|\bar{u}\|_{\mathcal{M}}), \quad \operatorname{supp}\bar{u} \subset \left\{ \hat{x} \in \Omega_o \mid \nabla\psi(\bar{u})(\hat{x}) = \min_{x\in\Omega_o} \nabla\psi(\bar{u})(x) \right\}.$$

- *There holds $\Phi(\bar{u}) \leq \Phi(u)$ for all $u \in \mathcal{M}(\Omega_o)$ and*

$$\Phi(\bar{u}) = \max_{u\in\mathcal{M}^+(\Omega_o)} [\langle \nabla\psi(\bar{u}), \bar{u} - u\rangle + G(\|\bar{u}\|_{\mathcal{M}}) - G(\|u\|_{\mathcal{M}})] = 0.$$

Proof. Due to the regularity assumption on Ψ there holds $\nabla\psi(\bar{u})(x) \leq 0$ for all $x \in \Omega_o$, see Lemma 3.10. Thus we have $[\nabla\psi(\bar{u})]^- = -\nabla\psi(\bar{u})$. The equivalence now follows from Propsition 3.13, Proposition 3.14, Proposition 3.15, and Lemma 3.16. □

We illustrate the abstract results of Theorem 3.17 for two important choices of G.

Example 3.3 (Optimality/cost trade-off)**.** *Consider $G(\|u\|_{\mathcal{M}}) = \beta\|u\|_{\mathcal{M}}$ where*

$$G\colon \mathbb{R} \to \mathbb{R}, \quad m \mapsto \beta m + I_{[0,\infty)}(m)$$

for some positive cost parameter $\beta > 0$. Clearly, this functional fulfills Assumption 3.3. We first calculate the set $\partial G(\|\bar{u}\|_{\mathcal{M}})$. If $\|\bar{u}\|_{\mathcal{M}} > 0$ we readily obtain

$$\partial G(\|\bar{u}\|_{\mathcal{M}}) = \{\beta\}.$$

In the second case, for $\|\bar{u}\|_{\mathcal{M}} = 0$, we get

$$m \in \partial G(0) \Leftrightarrow mc \leq \beta c, \quad \forall c \in \mathbb{R}_+ \Leftrightarrow m \in (-\infty, \beta].$$

We conclude $\partial G(0) = [0, \beta]$. Applying Theorem 3.17 to both cases yields the optimality of $\bar{u} \in \mathcal{M}^+(\Omega_o)$ if and only if

- *there holds*

$$-\min_{x\in\Omega_o} \nabla\psi(\bar{u})(x) \begin{cases} = \beta & \|\bar{u}\|_{\mathcal{M}} > 0 \\ \in [0,\beta] & \|\bar{u}\|_{\mathcal{M}} = 0 \end{cases}, \quad \langle -\nabla\psi(\bar{u}), \bar{u}\rangle - \beta\|\bar{u}\|_{\mathcal{M}} = 0.$$

- *there holds*

$$-\min_{x\in\Omega_o} \nabla\psi(\bar{u})(x) \begin{cases} = \beta & \|\bar{u}\|_{\mathcal{M}} > 0 \\ \in [0,\beta] & \|\bar{u}\|_{\mathcal{M}} = 0 \end{cases}, \quad \operatorname{supp}\bar{u} \subset \{\, x \in \Omega_o \mid \nabla\psi(\bar{u})(x) = -\beta \,\}.$$

- *there holds*

$$\bar{u} \in \operatorname*{arg\,min}_{u\in\mathcal{M}^+(\Omega_o)} \Phi(u), \qquad \Phi(\bar{u}) = \max_{u\in\mathcal{M}^+(\Omega_o)} [\langle \nabla\psi(\bar{u}), \bar{u}-u\rangle + \beta\|\bar{u}\|_{\mathcal{M}} - \beta\|u\|_{\mathcal{M}}] = 0,$$

noting that for $\bar{u} = 0$ the conditions

$$\langle -\nabla\psi(\bar{u}), \bar{u}\rangle + \beta\|\bar{u}\|_{\mathcal{M}} = 0 + 0 = 0, \quad \emptyset = \operatorname{supp}\bar{u} \subset \{\, x\in\Omega_o \mid \nabla\psi(\bar{u})(x) = -\beta \,\},$$

are trivially fulfilled.

Example 3.4 (Fixed budget)**.** *In this example we fix the overall cost for the experiment. We choose $G(\|u\|_{\mathcal{M}}) = I_{[0,K]}(\|u\|_{\mathcal{M}})$. The parameter $K \in \mathbb{R}_+\backslash 0$ denotes the budget for the experiment. Straightforward computations yield*

$$\partial G(\|\bar{u}\|_{\mathcal{M}}) = \begin{cases} \mathbb{R}_- & \|\bar{u}\|_{\mathcal{M}} = 0 \\ \{0\} & 0 < \|\bar{u}\|_{\mathcal{M}} < K \\ \mathbb{R}_+ & \|\bar{u}\|_{\mathcal{M}} = K \end{cases},$$

where $\mathbb{R}_-$ denotes the non-positive part of the real axis. We calculate the primal-dual gap for a design measure $u \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} F$ in this case to obtain

$$\Phi(u) = \langle \nabla\psi(u), u\rangle + \max_{v\in\mathcal{M}^+(\Omega_o), \|v\|_{\mathcal{M}}\leq K} \langle -\nabla\psi(u), v\rangle = \langle \nabla\psi(u), u\rangle - K \min_{x\in\Omega_o} \nabla\psi(u)(x),$$

where we used $\nabla\psi(\bar{u}) \leq 0$. Furthermore if $(\mathcal{P})$ admits optimal solutions there exists at least one with $\|\bar{u}\|_{\mathcal{M}} = K$ due to the monotonicity of Ψ. By application of Theorem 3.17 optimality of such design measures is characterized through the following equivalent statements.

- *There holds*

$$\langle -\nabla\psi(\bar{u}), \bar{u}\rangle + K \min_{x\in\Omega_o} \nabla\psi(\bar{u})(x) = 0.$$

- *There holds*

$$\operatorname{supp}\bar{u} \subset \left\{\, \hat{x} \in \Omega_o \mid \nabla\psi(\bar{u})(\hat{x}) = \min_{x\in\Omega_o} \psi(\bar{u})(x) \right\}$$

- *There holds*

$$0 = \langle \nabla\psi(\bar{u}), \bar{u}\rangle - K \min_{x\in\Omega_o} \nabla\psi(\bar{u})(x) \leq \langle \nabla\psi(u), u\rangle - K \min_{x\in\Omega_o} \nabla\psi(u)(x),$$

  for all $u \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} F$.

Example 3.5 (Kiefer-Wolfowitz Theorem)**.** *In this last example we illustrate the results of Theorem 3.17 in the case of $Q = \mathbb{R}^n$, $n \in \mathbb{N}$, $G(\|u\|_{\mathcal{M}}) = I_{[0,1]}(\|u\|_{\mathcal{M}})$ and the logarithmic D-optimal design criterion*

$$\Psi_D(B) = \begin{cases} \log(\det(B^{-1})) & B \in \mathrm{PD}(n) \\ +\infty & \mathit{else} \end{cases}.$$

The corresponding optimal design problem is given by

$$\min_{u \in \mathcal{M}^+(\Omega_o)} \Psi_D(\mathcal{I}(u)) \quad s.t. \quad \|u\|_{\mathcal{M}} \leq 1. \tag{3.24}$$

In what follows we assume that an optimal design $\bar{u}$ exists. Observe that

$$\Psi_D(\mathcal{I}(r\bar{u})) = \Psi_D(\mathcal{I}(\bar{u})) - n\log(r) \quad \forall r \in \mathbb{R}_+ \setminus \{0\}.$$

Thus we conclude $\|\bar{u}\|_{\mathcal{M}} = 1$. The gradient of the reduced functional $\psi_D(u) = \Psi_D(\mathcal{I}(u))$ at $u \in \mathrm{dom}_{\mathcal{M}^+(\Omega_o)} \psi_D$ is given by

$$\nabla\psi_D(u)(x) = -\mathcal{O}(x)^\top \mathcal{I}(u)^{-1}\mathcal{O}(x) \quad \forall x \in \Omega_o.$$

Calculating the primal-dual gap in this case gives

$$\Phi(u) = \langle \nabla\psi_D(u), u\rangle + \max_{x\in\Omega_o} -\nabla\psi_D(u)(x) = -n + \max_{x\in\Omega_o} -\nabla\psi_D(u)(x).$$

This leads to the following characterization:

- *The measure $\bar{u} \in \mathcal{M}^+(\Omega_o)$ is a D-optimal design.*
- *There holds*
$$\max_{x\in\Omega_o} -\nabla\psi_D(\bar{u})(x) = n.$$
- *There holds*
$$\operatorname{supp}\bar{u} \subset \{\, x \in \Omega_o \mid -\nabla\psi_D(\bar{u})(x) = n \,\}.$$
- *There holds*
$$n = \max_{x\in\Omega_o} -\nabla\psi_D(\bar{u})(x) \leq \max_{x\in\Omega_o} -\nabla\psi_D(u)(x), \quad u \in \mathrm{dom}_{\mathcal{M}^+(\Omega_o)} \psi_D,\ \|u\|_{\mathcal{M}} \leq 1.$$

This is exactly the statement of the well-known Kiefer-Wolfowitz theorem, [166] and [256, Theorem 3.2]. From this point of view our results can be interpreted as a natural extension of this classical result to a more general setting and the case of infinite dimensional Q.

We recall that the sparse sensor placement problem ($\mathcal{P}$) was introduced to avoid the non-convexity and combinatorial nature of (3.2). Therefore it remains to comment on conditions that ensure the existence of optimal measurement designs given as conic combination of finitely many Dirac delta functions. If such an optimal design exists the measurement setup is described by the number of support points, their positions and the associated coefficients. To this end we will mainly rely on the characterization of the support of an optimal design measure from Theorem 3.17 as well as the compactness properties of the Fisher information operator $\mathcal{I}$. We start by concluding the sparsity of an optimal design $\bar{u}$ if the set of global minimizers to $\nabla\psi(\bar{u})$ is finite.

Corollary 3.18. *Let an optimal design $\bar{u} \in \mathcal{M}^+(\Omega_o)$ be given. Assume that*

$$\operatorname{Ext}(\bar{u}) = \left\{ \hat{x} \in \Omega_o \mid \nabla\psi(\bar{u})(\hat{x}) = \min_{x \in \Omega_o} \nabla\psi(\bar{u})(x) \right\} = \{\bar{x}_i\}_{i=1}^N.$$

Then we have:

- *The optimal design measure $\bar{u} \in \mathcal{M}^+(\Omega_o)$ is sparse*

$$\bar{u} = \sum_{i=1}^N \bar{\mathbf{u}}_i \delta_{\bar{x}_i}, \quad \bar{\mathbf{u}}_i \in \mathbb{R}_+,\ i = 1, \ldots N$$

- *Additionally assume that the optimal gradient is unique, i.e. $\nabla\psi(\bar{u}_1) = \nabla\psi(\bar{u}_2)$ for arbitrary optimal designs $\bar{u}_1 \neq \bar{u}_2$. Then every minimizer $\bar{u}$ of $(\mathcal{P})$ is sparse and there holds* $\operatorname{supp} \bar{u} \subset \{\bar{x}_i\}_{i=1}^N$.

Proof. From the support condition (3.23) we conclude $\operatorname{supp} \bar{u} \subset \{\bar{x}_i\}_{i=1}^N$. Hence there exists $\bar{\mathbf{u}}_i \geq 0$, $i = 1, \ldots, N$, with $\bar{u} = \sum_{i=1}^N \bar{\mathbf{u}}_i \delta_{\bar{x}_i}$. This gives the first statement. For the second claim, we observe that the uniqueness of the optimal gradient implies

$$\operatorname{Ext}(\bar{u}_1) = \operatorname{Ext}(\bar{u}_2) = \{\bar{x}_i\}_{i=1}^N,$$

for arbitrary optimal solutions $\bar{u}_1$, $\bar{u}_2 \in \mathcal{M}^+(\Omega_o)$ to $(\mathcal{P})$. The second statement now readily follows from the first. □

The uniqueness of the optimal gradient holds for example if Ψ is strictly convex on its domain. If the set of its global minimizers consists of finitely many points, the optimal design is unique under an additional linear assumption condition.

Corollary 3.19. *Assume that $(\mathcal{P})$ admits at least one optimal solution $\bar{u}$ and that Ψ is strictly convex on its domain in* $\operatorname{Pos}(Q, Q)$. *Then we have:*

- *The optimal gradient $\nabla\psi(\bar{u})$ is unique.*
- *If* $\operatorname{Ext}(\bar{u}) = \{\bar{x}_i\}_{i=1}^N$ *and the set $\{\mathcal{I}(\delta_{\bar{x}_i})\}_{i=1}^N$ is linearly independent then the optimal measurement design is unique.*

Proof. Since Ψ is strictly convex on its domain in $\operatorname{Pos}(Q, Q)$, the optimal Fisher-information $\mathcal{I}(\bar{u})$ and thus also $\nabla\psi(\bar{u}) = \mathcal{I}^*\nabla\Psi(\mathcal{I}(\bar{u}))$ are unique. If, in addition, the second condition holds, every optimal design $\bar{u}$ is given by $\bar{u} = \sum_{i=1}^N \bar{\mathbf{u}}_i \delta_{\bar{x}_i}$ for some $\bar{\mathbf{u}}_i \in \mathbb{R}_+$, $i = 1, \ldots, N$ following Corollary 3.18. Obviously the corresponding weight vector $\bar{\mathbf{u}} = (\bar{\mathbf{u}}_1, \ldots, \bar{\mathbf{u}}_N) \in \mathbb{R}_+^N$ is a solution to

$$\min_{\mathbf{u} \in \mathbb{R}_+^N} \mathbf{F}(\mathbf{u}) := \left[\Psi \left(\sum_{i=1}^N \mathbf{u}_i \mathcal{I}(\delta_{\bar{x}_i}) \right) + G\left(\|\mathbf{u}\|_1 \right) \right], \tag{3.25}$$

where we fix the number and the position of the sensors and minimize only with respect to the measurement weights. Due to the strict convexity of Ψ and the linear independence assumption, the functional F is strictly convex. Thus it admits a unique global minimizer. This gives the statement. □

Let us now discuss situations in which the existence of sparse minimizers can be ensured. First, we consider cases in which the image of $\mathcal{I}$ is finite dimensional. Loosely speaking, in this situation, the information obtained through an arbitrary design measure can be obtained by a sparse one with bounded support size at a lower cost. Since a more general statement is provided in Chapter 6, see Theorem 6.32, we omit the proof at this point.

Theorem 3.20. *Assume that* $\dim \operatorname{Im} \mathcal{I} = n \in \mathbb{N}$. *Let* $u \in \mathcal{M}^+(\Omega_o)$ *be given. Then there exists* $\tilde{u} \in \mathcal{M}^+(\Omega_o)$ *with*

$$\mathcal{I}(u) = \mathcal{I}(\tilde{u}), \quad \|\tilde{u}\|_{\mathcal{M}} \le \|u\|_{\mathcal{M}}, \quad \# \operatorname{supp} \tilde{u} \le n.$$

Additionally, if there exists an optimal solution to $(\mathcal{P})$*, then there exists an optimal solution* $\bar{u}$ *with* $\# \operatorname{supp} \bar{u} \le n$.

As a special instance of the previous theorem, we conclude the existence of sparse minimizers if $Q = \mathbb{R}^n$ for some $n \in \mathbb{N}$.

Corollary 3.21. *Assume that* $Q = \mathbb{R}^n$ *for some* $n \in \mathbb{N}$ *and* $(\mathcal{P})$ *admits an optimal solution. Then there exists an optimal solution* $\bar{u}$ *to* $(\mathcal{P})$ *with* $\# \operatorname{supp} \bar{u} \le n(n+1)/2$.

Proof. The statement readily follows from the previous theorem by noting that $\operatorname{Im} \mathcal{I} \subset \operatorname{Sym}(n)$ and $\dim \operatorname{Sym}(n) = n(n+1)/2$. □

In contrast, the situation is certainly more involved if the image of $\mathcal{I}$ is not finite dimensional. However, in certain situations the smoothness of $\nabla\psi(\bar{u})$ implies that the set of its global minimizers is a Lebesgue zero set. A similar argument has been used in e.g. [67]. As a consequence, for one dimensional observation domains, all optimal measurement designs are sparse in this case.

Proposition 3.22. *Let* Ω_o *be the closure of a nonempty open and bounded domain in* $\mathbb{R}^d$. *Assume that* $(\mathcal{P})$ *admits at least one minimizer* $\bar{u}$. *Furthermore assume that*

- *the optimal gradient* $\nabla\psi(\bar{u})$ *is unique.*
- *the optimal gradient is non-constant on* Ω_o *and analytic in* $\operatorname{int} \Omega_o$ *with*

$$\operatorname*{arg\,min}_{x \in \Omega_o} \nabla\psi(\bar{u})(x) < \operatorname*{arg\,min}_{x \in \partial\Omega_o} \nabla\psi(\bar{u})(x), \; i = 1, \dots N.$$

Denote by μ_L *the Lebesgue measure on* Ω_o. *Then there holds:*

- *For every* $\bar{u}$ *we have* $\mu_L(\operatorname{supp} \bar{u}) = 0$.
- *If* $\Omega_o = [a, b]$ *for some* $a < b$ *there exists a set* $\{x_i\}_{i=1}^N \subset \operatorname{int} \Omega_o$ *such that every optimal design is given by*

$$\bar{u} = \sum_{i=1}^N \bar{\mathbf{u}}_i \delta_{x_i}, \quad \bar{\mathbf{u}}_i \in \mathbb{R}_+, \; i = 1, \dots, N$$

Proof. Define $\hat{p} = \nabla\psi(\bar{u}) - \arg\min_{x\in\Omega_o} \nabla\psi(\bar{u})$. Then $\hat{p}$ is non-constant and analytic on $\operatorname{int}\Omega_o$. Let an arbitrary optimal design $\bar{u}$ be given. We define $Z(\hat{p}) = \{\, x \in \Omega_o \mid \hat{p} = 0 \,\}$. Obviously we have $Z(\hat{p}) = \operatorname{Ext}(\bar{u}) \subset \operatorname{int}\Omega_o$. Since $\hat{p}$ is analytic we have

$$\mu_L(Z(\hat{p})) = \mu_L(\operatorname{Ext}(\bar{u})) = 0,$$

see e.g. [196]. Due to the support condition (3.23), we conclude $\mu(\operatorname{supp}\bar{u})_L = 0$, which gives the first part of the proof. Secondly assume that $\Omega_o = [a, b]$ for some $a < b$. Then it is well-known that the zeros of $\hat{p}$, and thus the global minimizers of $\nabla\psi(\bar{u})$, in $\operatorname{int}\Omega_o$ are isolated. Assume now that $\operatorname{Ext}(\bar{u})$ consist of at least countably many elements. Then there exists a sequence $\{x_i\}_{i\in\mathbb{N}}$ with $\hat{p}(x_i) = 0$. Due to Bolzano-Weierstrass there exists $\bar{x} \in \Omega_o$ with $x_i \to \bar{x}$ and, by continuity, $\hat{p}(\bar{x}) = 0$. Therefore $\bar{x} \in \operatorname{int}\Omega_o$ and $\bar{x}$ is an accumulation point of $Z(\hat{p})$. This gives a contradiction, i.e. $\operatorname{Ext}(\bar{u})$ contains only finitely many points. The statement now follows from Corollary 3.18. □

4 Sparse sensor placement for PDE-constrained inverse problems

This chapter is devoted to the inverse problem of identifying a finite dimensional parameter $q \in Q_{ad} \subset \mathbb{R}^n$ entering the weak form of a partial differential equation

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

We refer to the next section for the precise assumptions on the underlying model. As an example, we might consider the combustion process from [28] which is modeled as

$$a(q, y)(\varphi) = (\nabla y, \nabla \varphi)_{L^2} + (\alpha \nabla y, \varphi)_{L^2} + (D \exp\{-E/(d-y)\}\, y(c-y), \varphi)_{L^2}\,, \tag{4.1}$$

where the unknown parameter $q = (D, E)$ is given in terms of the activation energy E and the pre-exponential factor D. On the one hand, the general setting covers problems in which the unknown parameter q represents scalar unknown physical quantities such as material parameters or artificial constants that arise in the modelling process. On the other hand, the parameter of interest may also be a distributed function which is parametrized through finitely many degrees of freedom. In both cases, to obtain an appropriate mathematical surrogate for the simulation of the underlying physical process these parameters have to be well calibrated.

In what follows we assume that it is not possible to measure q directly and inference on its true value can only be made through measurements of the corresponding state $y = S[q]$. More concretely, the measured data $\mathbf{y}_d \in \mathbb{R}^N$ will be obtained through finitely many pointwise measurements of y at a set of points $\{\, x_i \,\}_i^N \subset \Omega_o$, where $\Omega_o \subset \bar{\Omega} \subset \mathbb{R}^d$, $d \in \mathbb{N}$, is a closed subset of the spatial domain covering the possible observation locations. The data $\mathbf{y}_d$ is assumed to be additively perturbed by normally i.i.d distributed noise ε, $\varepsilon_i \sim \mathcal{N}(0, 1/\mathbf{u}_i)$ stemming from the sensors. Estimates for the unknown parameter are obtained through realizations of a suitable Least-Squares estimator, see (4.6).

To mitigate the influence of the measurement errors on the estimator we will formulate and analyze an optimal sensor placement problem based on a linearization of the underlying PDE-model around a sophisticated a priori guess $\hat{q} \in \mathbb{R}^n$. 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, $k = 1, \ldots, n$ at an initial guess $\hat{q} \in Q_{ad}$, stemming either from prior knowledge or obtained from previous experiments. No restrictions on the maximum number of measurements nor their positions are made. Consequently, the optimal number of sensors N, their positions and the measurement weights $\mathbf{u}_i$ will be obtained through solving

$$\min_{x_i \in \Omega_o,\, \mathbf{u}_i \geq 0,\, i=1,\ldots,N, N \in \mathbb{N}} \quad \Psi(X^\top \Sigma^{-1} X + \mathcal{I}_0) + \beta \sum_{i=1}^{N} \mathbf{u}_i, \tag{4.2}$$

where the design dependent matrix $X^\top \Sigma^{-1} X \in \operatorname{Sym}(n)$ is given by

$$X^\top \Sigma^{-1} X = \sum_{i=1}^{N} \mathbf{u}_i \partial S[\hat{q}](x_i) \partial S[\hat{q}](x_i)^\top, \quad \partial S[\hat{q}](x) = (\partial_1 S[\hat{q}](x), \dots, \partial_n S[\hat{q}](x))^\top, \quad x \in \Omega_o.$$

We incorporated the design independent matrix $\mathcal{I}_0 \in \operatorname{NND}(n)$ in the formulation. This can be interpreted as a priori knowledge on the covariance matrix of the estimator stemming from previously collected data. Alternatively we may also adopt a Bayesian viewpoint and take $\mathcal{I}_0$ as the inverse of the covariance operator corresponding to a Gaussian prior. Optimal design approaches based on first-order approximations have been studied for and successfully applied to ordinary differential equations [13], differential-algebraic equations [20], and also partial differential equations [137]. Additionally, they also arise in sequential design approaches, see e.g. [170, 172]. Here, the experimenter alternates between estimating the unknown parameter and optimizing the experiment based on a linearization of the underlying system around the current estimate. The data that has been acquired in the previous experiments can thereby be included in a straightforward fashion by choosing $\mathcal{I}_0 = \mathcal{I}(u_{\text{old}})$. The design measure u_{old} is chosen to represent the previous experiments.

The first aim of this chapter is to demonstrate how we can fit this optimization problem into the general framework presented in Chapter 3 to get rid of the combinatorial aspect as well-as potential non-convexity arising in (4.2). This leads to a convex sensor placement problem

$$\min_{u \in \mathcal{M}^+(\Omega_o)} \Psi(\mathcal{I}(u) + \mathcal{I}_0) + \beta \|u\|_{\mathcal{M}}, \tag{P_β}$$

where we optimize for a design measure u in the space of Borel measures rather than the individual sensors. Here the matrix $\mathcal{I}(u) \in \mathbb{R}^{n \times n}$ is given by

$$\mathcal{I}(u) = \int_{\Omega_o} \partial S[\hat{q}](x) \partial S[\hat{q}](x)^\top \mathrm{d}u(x), \quad \mathcal{I}(u)_{ij} = \langle \partial_i S[\hat{q}] \partial_j S[\hat{q}], u \rangle,$$

where the integration has to be understood in the sense of Bochner. While (P_β) appears to be more general as the original problem we will show that it admits solutions given by a linear combination of Dirac delta functions. Their support points together with the associated coefficients of the corresponding Dirac delta then constitute an optimal solution to the original problem (4.2), making both approaches essentially equivalent.

As an alternative to the regularization term we may instead put constraints on the total cost of the measurement process leading to

$$\min_{u \in \mathcal{M}^+(\Omega_o)} \Psi(\mathcal{I}(u) + \mathcal{I}_0) \quad s.t. \quad \|u\|_{\mathcal{M}} \leq K, \tag{P^K}$$

for some budget $K > 0$. Both formulations, (P_β) and (P^K), are closely linked (see Section 4.2): 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 β. On the other hand, incorporating a priori knowledge, both problem formulations parameterize the same solution manifold. The parameters β 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. For practically relevant design criteria Ψ the inequality constraint will be active at every optimal solution. This links (P^K) closely to the concept of approximate design introduced by Kiefer and

Wolfowitz in the context of linear regression, see [165] and the discussion in the previous chapters. For an extension to sensor placement problems based on first order approximations of nonlinear models see [107, 217]. In the context of partial differential equations such an approach to optimal sensor placement was pursued in [17, 256].

The results in this chapter delimit themselves from these previous approaches in several ways by taking recent advances in the theory of measure-valued optimization problems into account. Besides proving well-posed our main focus lies on three different aspects of problem (P_β). First we provide a suitable solution algorithm for (P_β) based on alternating between adding single sensors to the design and optimizing their measurement weights. Here we essentially generalize the algorithm presented in [44, 50]. By a careful convergence analysis, see Chapter 6, we improve upon the convergence results in these references and further derive convergence rates for the optimal design measure with respect to a suitably chosen and computable metric. Since all considerations are taken at the function space level we observe stability of this convergence behaviour with respect to discretization of the observational domain and the underlying PDE model. In this context we also discuss solution algorithms and improved convergence rates for (P^K), such as the well-known Fedorov-Wynn algorithm, see [105, 272]. Second we consider perturbations of the optimal design criterion or the underlying PDE and study stability and sensitivity of the optimal design measure. Note that these questions are of practical importance since the sensor placement problem itself is based on a first-order approximation of the PDE model. However we are not aware of any results in this direction. Finally, to solve (P_β) or (P^K) one has to compute the state $y = S[q]$ as well as the sensitivities $\{\partial_k S[q]\}_{k=1}^n$ of the state with respect to the parameters. In general, the state and sensitivity PDEs cannot be solved analytically, but only numerically. We therefore analyze a discretization scheme for (P_β) based on a finite element discretization of the underlying PDEs and a variational discretization approach for the design measure, see [59, 148]. Sharp a priori error estimates with respect to the discretization parameter for the optimal design functionals as well as the optimal design measure are provided.

The outline of this chapter is as follows. In Section 4.2 we focus on the existence and the structure of optimal design measurements obtained through (P_β). In Section 4.3 we shed light on the connection between (P_β) and a class of semi-infinite optimization problems, giving a geometric interpretation of the optimal sensor placement problem. Section 4.4 is devoted to the numerical treatment of the sparse sensor placement problem by accelerated conditional gradient methods, see also Chapter 6. Stability and sensitivity analysis of the optimal design measure is in the focus of Section 4.5. In Section 4.6 discretization of (P_β) and a priori error estimation are considered. To underline our results we present some numerical evidence in Section 4.7. We note that parts of this chapter have been submitted for publication, see [200].

4.1 Parameter estimation and optimal design

4.1.1 Parameter estimation

Within the scope of this chapter we consider the identification of a parameter q entering a weak form $a(\cdot,\cdot)(\cdot)\colon\ Q_{ad} \times \hat{Y} \times Y \to \mathbb{R}$, which can be non-linear in its first two arguments but is linear in the last one. Here, we denote by $Q_{ad} \subset \mathbb{R}^n$, $n \in \mathbb{N}$, a set of admissible parameters, Y denotes a suitable Hilbert space of functions on a spatial domain $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, which is assumed to be open and bounded. We consider the state space $\hat{Y} = \hat{y} + Y$, where the function

$\hat{y}$ models (potentially) non-homogeneous (Dirichlet-type) boundary conditions in the model. For every $q \in Q_{ad}$ we introduce the state $y = S[q] \in \hat{Y}$ as a solution to

$$y \in \hat{Y}\colon \quad a(q,y)(\varphi) = 0 \quad \forall \varphi \in Y. \tag{4.3}$$

The operator $S\colon Q_{ad} \to \hat{Y}$ mapping a parameter q to the associated state is called the parameter-to-state operator. We make the following general regularity assumption.

Assumption 4.1. For every $q \in Q_{ad}$ there exists a unique solution $y \in \hat{Y} \cap \mathcal{C}(\Omega_o)$ to (4.3). The parameter-to-state mapping S with

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

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 $\partial_k S[q] \in \mathcal{C}(\Omega_o)$ and by $\partial S[q] \in \mathcal{C}(\Omega_o, \mathbb{R}^n)$ the vector of partial derivatives.

We emphasize that under suitable differentiability assumptions on the form $a(\cdot,\cdot)$ and Assumption 4.1 the k-th partial derivative $\delta y_k = \partial_k S[q] \in Y \cap \mathcal{C}(\Omega_o)$, $k = 1, \ldots, n$, is the unique solution of the sensitivity equation

$$a'_y(q,y)(\delta y_k, \varphi) = -a'_{q_k}(q,y)(\varphi), \quad \forall \varphi \in Y, \tag{4.4}$$

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., [255, 258].

To estimate the unknown parameter we consider measurement data $\mathbf{y}_d$ collected at a set of N distinct sensor locations $\{x_j\}_{j=1}^N \subset \Omega_o$, where $\Omega_o \subset \bar{\Omega}$ is a closed set. In order to take measurement errors into account we assume that the data $\mathbf{y}_d^j \approx S[q^*](x_j)$ is additively perturbed by independently unit normally distributed noise; see, e.g., [19]. Here $S[q^*](x_j)$ denotes the response of the model to an unknown parameter q^*. Taking into account that multiple measurements can be performed at the same location, we obtain that

$$\mathbf{y}_d^j = S[q^*](x_j) + \epsilon_j, \;\; \epsilon_j \sim \mathcal{N}(0, 1/\mathbf{u}_j), \;\; \operatorname{Cov}(\epsilon_j, \epsilon_i) = 0,$$

for all $i, j = 1, \ldots, N$, and $j \neq i$, where $\mathbf{u}_j \in \mathbb{N} \setminus \{0\}$ denotes the number of measurements taken at the j-th location. More generally, we assume that $\mathbf{u}_j$ can be chosen arbitrarily in $\mathbb{R}_+ \setminus \{0\}$ in the following. In this case the measurement weights $\mathbf{u}_j > 0$ should be interpreted as diligence factors giving information on how carefully the data should be collected at the corresponding measurement point.

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

$$J(q,\varepsilon) = \frac{1}{2} \sum_{j=1}^{N} \mathbf{u}_j (S[q](x_j) - \mathbf{y}_d^j(\varepsilon))^2 \tag{4.5}$$

as well as the possibly multi-valued least squares estimator

$$\tilde{q}\colon \mathbb{R}^N \to \mathcal{P}(\mathbb{R}^n), \;\; \tilde{q}(\varepsilon) = \operatorname*{arg\,min}_{q \in Q_{ad}} J(q,\varepsilon), \tag{4.6}$$

where $\mathcal{P}(\mathbb{R}^n)$ denotes the power set of $\mathbb{R}^n$. Note that this estimator is the usual Maximum-Likelihood estimator using the assumption on the distribution measurement errors $\varepsilon_j \sim \mathcal{N}(0, 1/\mathbf{u}_j)$.

4.1.2 Optimal design

Since the measurement errors are modelled as random variables, the uncertainty in the data is also propagated to the estimator. Consequently we interpret $\tilde{q}$ 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

$$D(\tilde{q},\alpha)(\epsilon) = \left\{ p \in Q_{ad} \mid J(p,\epsilon) - \min_{q \in Q_{ad}} J(q,\epsilon) \le \gamma_n^2(\alpha)/2 \right\}, \tag{4.7}$$

where $\gamma_n^2(\alpha)$ denotes the $(1-\alpha)$-quantile of the χ^2-distribution with n degrees of freedom; see, e.g., [23,36]. 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 performance indicator for 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 a high probability. Given a realization $D(\bar{q},\alpha)(\bar{\epsilon})$ 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, e.g. the total number of measurements carried out, their positions x_j, and the measurement weights $\mathbf{u}_j$ 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 $D(\tilde{q},\alpha)(\varepsilon)$.

To circumvent this problem we follow the approach proposed in, e.g., [107, 216] 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, $\epsilon \in \mathbb{R}^N$ denotes an arbitrary vector of measurement errors, and $\mathbf{x} \in \Omega_o^N$, $\mathbf{x} = (x_1,\dots,x_N)$, with $x_j \in \Omega_o$, $j = 1,\dots,N$, stands for the measurement locations. For abbreviation we write $S[\hat{q}](x) \in \mathbb{R}^N$ for the vector of observations with $S[\hat{q}](x)_j = S[\hat{q}](x_j)$, $j = 1,\dots,N$. Moreover the matrices $X \in \mathbb{R}^{N\times n}$ and $\Sigma^{-1} \in \mathbb{R}^{N\times N}$ are defined as

$$X_{jk} = \partial_k S[\hat{q}](x_j), \quad \Sigma^{-1}_{ij} = \delta_{ij}\mathbf{u}_i, \quad i,j = 1,\dots,N,\ k = 1,\dots,n,$$

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

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

which can be equivalently written as

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

where $\|v\|_{\Sigma^{-1}} = v^\top \Sigma^{-1} v$ for $v \in \mathbb{R}^n$. In contrast to the estimator $\tilde{q}$ from (4.6), the associated linearised estimator

$$\tilde{q}_{\text{lin}} \colon \mathbb{R}^N \to \mathbb{R}^n, \quad \tilde{q}_{\text{lin}}(\epsilon) = \arg\min_{q\in\mathbb{R}^n} J_{\text{lin}}(q,\epsilon), \tag{4.8}$$

is single-valued and its realizations can be calculated explicitly (see, e.g., [251]), as

$$\tilde{q}_{\text{lin}}(\epsilon) = \hat{q} + (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} \left(\mathbf{y}_d(\epsilon) - S[\hat{q}](x)\right). \tag{4.9}$$

Due to the assumptions on the noise ϵ 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., [36]) are thus given by

$$D(\tilde{q}_{\text{lin}}, \alpha)(\epsilon) = \left\{ q \in \mathbb{R}^n \mid q = \tilde{q}_{\text{lin}} + (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1/2} \delta\epsilon,\ |\delta\epsilon|_{\mathbb{R}^N} \leq \gamma_n(\alpha) \right\}, \tag{4.10}$$

where $|\cdot|_{\mathbb{R}^N}$ denotes the Euclidean norm on $\mathbb{R}^N$. We point out that the linearised 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 linearised confidence domains as a function of the measurement setup, which leads to (4.2).

4.2 Theoretical results

Motivated through the considerations in the previous section we propose to improve the estimator by minimizing a design criterion acting on the matrix $X^\top \Sigma^{-1} X$ as a function of the experimental setup

$$\min_{x_i \in \Omega_o, \mathbf{u}_i \in \mathbb{R}_+, i=1,\ldots,N, N \in \mathbb{N}} [\Psi(X^\top \Sigma^{-1} X + \mathcal{I}_0) + \beta \|\mathbf{u}\|_1], \tag{4.11}$$

where the matrix $\mathcal{I}_0 \in \text{NND}(n)$ (e.g. $\mathcal{I}_0 = 0$) incorporates prior knowledge on the parameter, as described in the introduction of this chapter.

Let us put this problem into the perspective of Chapter 3. We choose the parameter space as $Q = \mathbb{R}^n$. From the discussion in Section 3.1.1 we recall that $\text{SHS}(\mathbb{R}^n, \mathbb{R}^n)$ can be identified with the symmetric matrices $\text{Sym}(n)$ together with the Frobenius scalar-product. Since there won't be any ambiguities in this chapter we drop the indices and write

$$\|A\|_{\text{Sym}} = \sqrt{(A, A)_{\text{Sym}}} = \sqrt{\text{Tr}(A^* A)}, \quad A \in \text{Sym}(n).$$

Given vectors $v, z \in \mathbb{R}^n$, the tensor $v \otimes z \in \mathbb{R}^{n \times n}$ is simply given as the rank 1 matrix $v \otimes z = vz^\top$. In the same way we identify $\text{Pos}(\mathbb{R}^n, \mathbb{R}^n)$ with the set of non-negative definite matrices $\text{NND}(n)$. On $\text{NND}(n)$ we consider the Löwner ordering given by

$$B_1 \leq_L B_2 \Leftrightarrow B_2 - B_1 \in \text{NND}(n).$$

Furthermore, due to $\partial S[\hat{q}] \in \mathcal{C}(\Omega_o, \mathbb{R}^n)$, the pointwise Fisher-information

$$I \colon \Omega_o \to \text{Sym}(n), \quad x \mapsto \partial S[\hat{q}](x) \partial S[\hat{q}](x)^\top,$$

is continuous.

In this light, given $\mathbf{x} = (x_1, \dots, x_N) \in \Omega_o^N$ and $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_N) \in \mathbb{R}_+^N$ we rewrite the Fisher-information matrix as

$$X^\top \Sigma^{-1} X = \sum_{i=1}^{N} \mathbf{u}_i \partial S[\hat{q}](x_i) \partial S[\hat{q}](x_i)^\top = \int_{\Omega_o} \partial S[\hat{q}](x) \otimes \partial S[\hat{q}](x) \, \mathrm{d}u(x),$$

where the measure $u \in \mathcal{M}^+(\Omega_o)$ is given by $u = \sum_{i=1}^N \mathbf{u}_i \delta_{x_i}$. Thus (4.2) can be viewed as a special instance of the general problem (3.12) by choosing the observation operator $\mathcal{O} \colon \Omega_o \to \mathbb{R}^n$ as $\mathcal{O}(x) = \partial S[\hat{q}](x) \in \mathbb{R}^n$ for all $x \in \Omega_o$. We introduce the linear and continuous Fisher-operator $\mathcal{I}$, see (3.10), by

$$\mathcal{I} \colon \mathcal{M}(\Omega_o) \to \mathrm{Sym}(n), \quad \mathcal{I}(u) = \int_{\Omega_o} \partial S[\hat{q}](x) \partial S[\hat{q}](x)^\top \mathrm{d}u(x),$$

where for $u \in \mathcal{M}(\Omega_o)$ the entries of the matrix $\mathcal{I}(u) \in \mathrm{Sym}(n)$ are given as

$$\mathcal{I}(u)_{ij} = \langle \partial_i S[\hat{q}] \partial_j S[\hat{q}], u \rangle \quad \forall i, j \in \{ 1, \dots, n \}.$$

In the following, we consider the sparse sensor placement problem, c.f. also $(\mathcal{P})$,

$$\min_{u \in \mathcal{M}^+(\Omega_o)} [\Psi(\mathcal{I}(u) + \mathcal{I}_0) + \beta \|u\|_{\mathcal{M}}].$$

Concerning the function Ψ the following assumptions are made.

Assumption 4.2. The function $\Psi \colon \mathrm{Sym}(n) \to \mathbb{R} \cup \{+\infty\}$ satisfies:

A4.1 There holds $\mathrm{dom}\, \Psi = \mathrm{PD}(n)$.

A4.2 Ψ is two times continuously differentiable at every $B \in \mathrm{PD}(n)$.

A4.3 Ψ is lower semi-continuous and convex on $\mathrm{NND}(n)$.

A4.4 Ψ is monotone with respect to the Löwner ordering on $\mathrm{NND}(n)$, i.e. there holds

$$B_1 \leq_L B_2 \Rightarrow \Psi(B_1) \geq \Psi(B_2) \quad \forall B_1, \ B_2 \in \mathrm{NND}(n).$$

Accordingly, the functional $\hat{\Psi}(B) = \Psi(B + \mathcal{I}_0)$ fulfills Assumption 3.2 with

$$\mathrm{dom}_{\mathrm{NND}(n)} \hat{\Psi} = \{ B \in \mathrm{NND}(n) \mid B + \mathcal{I}_0 \in \mathrm{Pos}(n) \}.$$

While Assumptions (**A4.1**) to (**A4.3**) are important for the existence of optimal designs and the derivation of first order optimality conditions, Assumption (**A4.4**) admits a geometric interpretation. Given two design measures $u_1, u_2 \in \mathcal{M}^+(\Omega_o)$ with $\mathcal{I}(u_1), \mathcal{I}(u_1) \in \mathrm{PD}(n)$ and $\mathcal{I}(u_1) \leq_L \mathcal{I}(u_2)$ the corresponding ellipsoids fulfill

$$\mathcal{E}_2 = \{ \delta q \in \mathbb{R}^n \mid \delta q^\top \mathcal{I}(u_2) \delta q \leq r \} \subset \mathcal{E}_1 = \{ \delta q \in \mathbb{R}^n \mid \delta q^\top \mathcal{I}(u_1) \delta q \leq r \}$$

for any $r > 0$. This ensures that Ψ is indeed a suitable criterion for the size of the linearised confidence ellipsoids (4.10). For a similar set of conditions; see [256, p. 41]. The given assumptions can be verified for a large class of classical optimality criteria, among them the A and D criterion

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

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 $\Psi_A^w(B) = \mathrm{Tr}(WB^{-1}W)$ allows to put special emphasis on particular parameters by virtue of the weight matrix $W \in \mathrm{NND}(n)$. However, we emphasize that the results presented in this chapter cannot be applied to other non-differentiable popular criteria such as the E criterion defined by

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

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.

4.2.1 Existence of optimal solutions 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_β). Additionally, results on the sparsity pattern of optimal designs are derived. Let us first take a closer look on the Fisher operator $\mathcal{I}$. It is readily verified that it is the Banach space adjoint of the operator

$$\mathcal{I}^*\colon \mathrm{Sym}(n) \to \mathcal{C}(\Omega_o), \quad \text{with} \quad \mathcal{I}^*(B) = \varphi_B,$$

where $\varphi_B \in \mathcal{C}(\Omega_o)$, given $B \in \mathrm{Sym}(n)$, is the continuous function defined by

$$\varphi_B(x) = \mathrm{Tr}\left(\partial S[\hat{q}](x)\partial S[\hat{q}](x)^\top B\right) = \partial S[\hat{q}](x)^\top B\, \partial S[\hat{q}](x) \quad \forall x \in \Omega_o, \tag{4.12}$$

see Proposition 3.7. Now, we formulate the reduced design problem (P_β) as

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

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

Proposition 4.1. *Let Assumptions (**A4.1**)–(**A4.4**) be fulfilled and let $\mathcal{I}_0 \in \mathrm{NND}(n)$ be given. The operator $\mathcal{I}$ and the functional ψ satisfy:*

1. *For every $u \in \mathcal{M}^+(\Omega_o)$ there holds $\mathcal{I}(u) \in \mathrm{NND}(n)$.*
2. *There holds $\mathrm{dom}_{\mathcal{M}^+(\Omega_o)}\, \psi = \{\, u \in \mathcal{M}^+(\Omega_o) \mid \mathcal{I}(u) + \mathcal{I}_0 \in \mathrm{PD}(n)\,\}$. The domain $\mathrm{dom}_{\mathcal{M}^+(\Omega_o)}\, \psi$ is weak* sequentially open in $M^+(\Omega_o)$.*
3. *ψ is two times continuously differentiable on its domain with derivative*

$$\nabla\psi(u) = \mathcal{I}^*\left(\nabla\Psi(\mathcal{I}(u) + \mathcal{I}_0)\right) \in \mathcal{C}(\Omega_o)$$

for every $u \in \mathrm{dom}_{\mathcal{M}^+(\Omega_o)}\, \psi$. The derivative can be identified with the non-positive continuous function

$$\left[\nabla\psi(u)\right](x) = \partial S[q](x)^\top \nabla\Psi(\mathcal{I}(u) + \mathcal{I}_0)\, \partial S[q](x) \quad \forall x \in \Omega_o. \tag{4.13}$$

Moreover the gradient $\nabla\psi\colon \mathrm{dom}_{\mathcal{M}^+(\Omega_o)}\, \psi \to \mathcal{C}(\Omega_o)$ is weak-to-strong continuous. Given $u \in \mathrm{dom}_{\mathcal{M}^+(\Omega)}\, \psi$, the second derivative $\nabla^2\psi(u) \in \mathcal{L}(\mathcal{M}(\Omega_o), \mathcal{M}(\Omega_o)^*)$ is characterized as*

$$\langle \delta u_1, \nabla^2\psi(u)\delta u_2\rangle_{\mathcal{M},\mathcal{M}^*} = \mathrm{Tr}(\mathcal{I}(\delta u_1)\nabla^2\Psi(\mathcal{I}(u))\mathcal{I}(\delta u_2)), \quad \forall \delta u_1, \delta u_2 \in \mathcal{M}(\Omega_o).$$

4. *ψ is weak* lower semi-continuous and convex on $\mathcal{M}^+(\Omega_o)$.*

5. *ψ is monotone in the sense that*

$$\mathcal{I}(u_1) \leq_L \mathcal{I}(u_2) \Rightarrow \psi(u_1) \geq \psi(u_2) \quad \forall u_1,\ u_2 \in \mathcal{M}^+(\Omega_o).$$

Proof. Second order Fréchet differentiability differentiability follows from the differentiability assumptions on Ψ by applying the chain rule. The rest of the claimed statements can be inferred from Proposition 3.9. □

To ensure existence of optimal designs we make the following assumption on the objective functional.

Assumption 4.3. The functional $F(u) = \psi(u) + \beta\|u\|_{\mathcal{M}}$ is radially unbounded.

Remark 4.1. This additional assumption is fulfilled for the A and D-optimal design criterion considered before, since

$$\beta\|u\|_{\mathcal{M}} \leq \operatorname{Tr}((\mathcal{I}(u) + \mathcal{I}_0)^{-1}) + \beta\|u\|_{\mathcal{M}},$$

as well as

$$\beta\|u\|_{\mathcal{M}} - c_1 \log(c_2\|u\|_{\mathcal{M}} + \|\mathcal{I}_0\|_{\mathrm{Sym}}) \leq -\log(\det(\mathcal{I}(u) + \mathcal{I}_0)) + \beta\|u\|_{\mathcal{M}},$$

for some positive constant $c_1, c_2 > 0$.

Since the regularization term is given by $G_\beta(\|u\|_{\mathcal{M}}) = \beta\|u\|_{\mathcal{M}}$ and the function

$$G_\beta \colon \mathbb{R} \to \mathbb{R} \cup +\infty, \quad m \mapsto \beta m + I_{[0,\infty)}(m),$$

fulfills Assumption 3.3 the following existence result is due to Proposition 3.11.

Proposition 4.2. *Assume that $\operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi \neq \emptyset$ and $\beta > 0$. Then there exists at least one optimal solution $\bar{u}_\beta$ to (P_β). Moreover the set of optimal solutions is bounded. If Ψ is strictly convex on $\mathrm{PD}(n)$ then the optimal Fisher-information matrix $\mathcal{I}(\bar{u}_\beta)$ is unique.*

Next we give conditions for the domain of ψ to be non-empty.

Proposition 4.3. *Assume that $\beta > 0$ and*

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

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

Proof. According to Proposition 4.2 we have to show that there exists an admissible design measure. By assumption we can choose a set of $n - \operatorname{rank}\mathcal{I}_0$ distinct points $x_j \in \Omega_o$ such that

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

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

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

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

By applying standard results from convex analysis we derive necessary and sufficient first order optimality conditions. Most important, we link the support points of an optimal design $\bar{u}_\beta$ to the maximizers of $-\nabla\psi(\bar{u}_\beta) \geq 0$.

Lemma 4.4. *Let $\beta > 0$ be given. A measure $\bar{u}_\beta \in \mathcal{M}^+(\Omega_o)$ is a minimizer of (P_β) if and only if one of the following (equivalent) conditions holds*

- *There holds*

$$-\nabla\psi(\bar{u}_\beta) \in \partial(\beta \| \cdot \|_{\mathcal{M}} + I_{u\geq 0}(\cdot))(\bar{u}_\beta).$$

- *There holds*

$$\sup_{v \in \mathcal{M}^+(\Omega_o)} [\langle \nabla\psi(\bar{u}_\beta), \bar{u}_\beta - v \rangle + \beta \|\bar{u}_\beta\|_{\mathcal{M}} - \beta \|v\|_{\mathcal{M}}] = 0.$$

- *We have*

$$-\min_{x\in\Omega_o} \nabla\psi(\bar{u}_\beta)(x) \begin{cases} = \beta & \|\bar{u}_\beta\|_{\mathcal{M}} > 0 \\ \leq \beta & \|\bar{u}_\beta\|_{\mathcal{M}} = 0 \end{cases}, \quad -\langle \nabla\psi(\bar{u}_\beta), \bar{u}_\beta \rangle = \beta \|\bar{u}_\beta\|_{\mathcal{M}}.$$

- *For all $x \in \Omega_o$ we have*

$$-\min_{x\in\Omega_o} \nabla\psi(\bar{u}_\beta)(x) \begin{cases} = \beta & \|\bar{u}_\beta\|_{\mathcal{M}} > 0 \\ \leq \beta & \|\bar{u}_\beta\|_{\mathcal{M}} = 0 \end{cases}, \quad \operatorname{supp} \bar{u}_\beta \subset \{\, x \in \Omega_o \mid -\nabla\psi(\bar{u}_\beta)(x) = \beta \,\}. \tag{4.14}$$

Proof. Since ψ is two times differentiable and monotone we have $-\nabla\psi(u)(x) \geq 0$ and thus also

$$-\min_{x\in\Omega_o} \nabla\psi(u)(x) \geq 0$$

for all measures $u \in \mathcal{M}^+(\Omega_o)$ and $x \in \Omega_o$. Calculating the subdifferential of G_β at $\|\bar{u}_\beta\|_{\mathcal{M}}$ gives

$$\partial G_\beta(\|\bar{u}_\beta\|_{\mathcal{M}}) = \{\beta\} + \partial I_{[0,\infty)}(\|\bar{u}_\beta\|_{\mathcal{M}}) = \begin{cases} (-\infty, \beta] & \|\bar{u}_\beta\|_{\mathcal{M}} = 0 \\ \{\beta\} & \|\bar{u}_\beta\|_{\mathcal{M}} > 0 \end{cases}.$$

Furthermore we note that $\bar{u}_\beta$ is optimal if and only if

$$-\nabla\psi(\bar{u}_\beta) \in \partial(\beta \| \cdot \|_{\mathcal{M}} + I_{u\geq 0}(\cdot))(\bar{u}_\beta) = \beta\partial \|\bar{u}_\beta\|_{\mathcal{M}} + \partial I_{u\geq 0}(\bar{u}_\beta)$$

where the last equality holds due to the continuity of the norm. Thus we obtain the result by applying Theorem 3.17 as in Example 3.3. □

Remark 4.2. For (P^K) a similar optimality condition can be derived by the same techniques. A measure $\bar{u}^K \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi$ is an optimal solution of (P^K) if and only if

$$\operatorname{supp} \bar{u}^K \subset \left\{ x \in \Omega_o \;\middle|\; \nabla\psi(\bar{u}^K)(x) = \operatorname*{arg\,min}_{x \in \Omega_o} \nabla\psi(\bar{u}^K)(x) \right\},$$

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

$$-\langle \nabla\psi(\bar{u}^K), \bar{u}^K \rangle + \operatorname*{arg\,min}_{x \in \Omega_o} \nabla\psi(\bar{u}^K)(x) \|\bar{u}^K\|_{\mathcal{M}} = 0,$$

yielding again the well-known Kiefer-Wolfowitz equivalence theorem; see [164,165] and [256, 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 follows due to the finite dimensionality of the parameter space. This is addressed in the following theorem. Moreover if the optimal Fisher-information matrix $\mathcal{I}(\bar{u}_\beta) \in \operatorname{Sym}(n)$ is unique and

$$\{ x \in \Omega_o \mid -\nabla\psi(\bar{u}_\beta) = \beta \} = \{x_i\}_{i=1}^N, \tag{4.15}$$

for some $x_i \in \Omega_o$, $i = 1, \ldots, N$, then every optimal design is sparse and uniqueness of the design holds under an additional linear independence assumption.

Theorem 4.5. *Let $u \in \mathcal{M}^+(\Omega_o)$ be given. Then there exists $\tilde{u} \in M^+(\Omega_o)$ with*

$$\mathcal{I}(u) = \mathcal{I}(\tilde{u}), \quad \|\tilde{u}\|_{\mathcal{M}} \le \|u\|_{\mathcal{M}}, \quad \#\operatorname{supp} \tilde{u} \le n(n+1)/2.$$

Additionally, if there exists an optimal solution to (P_β), then there exists an optimal solution $\bar{u}_\beta$ with $\#\operatorname{supp} \bar{u}_\beta \le n(n+1)/2$.

Proof. Since $\dim \operatorname{Sym}(n) = n(n+1)/2$ this result is due to Theorem 3.20 and Corollary 3.21. □

Corollary 4.6. *Let Ψ be strictly convex on its domain and assume that (4.15) holds. Then every optimal design $\bar{u}_\beta$ is of the form $\bar{u}_\beta = \sum_{i=1}^N \mathbf{u}\delta_{x_i}$, $\mathbf{u}_i \in \mathbb{R}_+$. If $\{\mathcal{I}(\delta_{x_i})\}_{i=1}^N$ is linear independent then the optimal design is unique.*

Proof. For a proof see Corollary 3.18 and Corollary 3.19. □

The proof of Theorem 4.5 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. The procedure is summarized in Algorithm 1.

Proposition 4.7. *Let $u = \sum_{i=1}^m \mathbf{u}_i\delta_{x_i}$ be given and assume that $\{\mathcal{I}(\delta_{x_i})\}_{i=1}^m$ is linearly dependent. Denote by $u_{\text{new}} = \sum_{\{ i \mid \mathbf{u}_{\text{new},i} > 0 \}} \mathbf{u}_{\text{new},i}\delta_{x_i}$ the measure that is obtained after one execution of the loop in Algorithm 1. Then there holds*

$$F(u_{\text{new}}) \le F(u), \quad \#\operatorname{supp} u_{\text{new}} \le \#\operatorname{supp} u - 1, \quad \operatorname{supp} u_{\text{new}} \subset \operatorname{supp} u.$$

Algorithm 1 Support-point removal

```
  1. Let $u = \sum_{i=1}^m \mathbf{u}_i \delta_{x_i}$ be given.
  while $\{\mathcal{I}(\delta_{x_i})\}_{i=1}^m$ linearly dependent do
    2. Find $0 \neq \bar{\mathbf{u}}$ with $0 = \sum_{i=1}^m \bar{\mathbf{u}}_i \mathcal{I}(\delta_{x_i})$.
    3. Set $\mu = \max_i \{\, \bar{\mathbf{u}}_i / \mathbf{u}_i \,\}$, $\mathbf{u}_{\text{new},i} = \mathbf{u}_i - \bar{\mathbf{u}}_i / \mu$.
    4. Update $u_{\text{new}} = \sum_{\{\, i \mid \mathbf{u}_{\text{new},i} > 0 \,\}} \mathbf{u}_{\text{new},i} \delta_{x_i}$.
  end while
```

Proof. This is a special case of Proposition 6.33. □

In the last part of this section we will further discuss structural properties of solutions to (P_β), mainly focusing on their connection to (P^K) and their behaviour for $\beta \to \infty$.

Proposition 4.8. *The problems (P^K) and (P_β) are equivalent in the following sense: Given, for fixed $K > 0$, a solution $\bar{u}^K$ to (P^K), there exists a $\beta \geq 0$, such that $\bar{u}^K$ is an optimal solution to (P_β) and vice versa.*

Furthermore, assuming that Ψ is strictly monotone with respect to the Löwner ordering in the sense that

$$B_2 - B_1 \in \mathrm{PD}(n) \Rightarrow \Psi(B_1) > \Psi(B_2), \quad B_1, B_2 \in \mathrm{PD}(n),$$

we additionally obtain the following:

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

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

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

Proof. Fix an arbitrary $K > 0$. By well established results from convex analysis (see, e.g., [43, Proposition 2.153]) the norm-constrained problem (P^K) is calm. Define the Lagrangian L as

$$L \colon \mathcal{M}^+(\Omega_o) \times \mathbb{R}_+ \to \mathbb{R}_+ \quad L(u, \beta) = \psi(u) + \beta\left(\|\omega\|_{\mathcal{M}} - K\right).$$

A given measure $\bar{u}^K \in \mathcal{M}^+(\Omega_o)$ is optimal for (P^K) if and only if there exists a Lagrange multiplier $\beta \geq 0$ with

$$\bar{u}^K \in \operatorname*{arg\,min}_{u \in \mathcal{M}^+(\Omega_o)} L(u, \beta), \quad \beta(\|\bar{u}^K\|_{\mathcal{M}} - K) = 0. \tag{4.16}$$

The set of Lagrange multipliers is independent of the choice of the optimizer $\bar{u}^K$, i.e. given two arbitrary optimal solutions $\bar{u}_1^K, \bar{u}_2^K \in \mathcal{M}^+(\Omega_o)$ to (P^K) and $\beta \geq 0$ such that the pair $(\bar{u}_1^K, \beta)$ fulfills (4.16), then so does $(\bar{u}_2^K, \beta)$. For a proof we refer to, e.g., [43, Theorem 3.4]. This proves the first statement.

Assume that Ψ is strictly monotone. Let $\bar{u}^K$ be an arbitrary optimal solution to (P^K) with $\|\bar{u}^K\|_{\mathcal{M}} < K$. Using the strict monotonicity of Ψ we deduce that $\bar{u}^K \neq 0$. Defining $\tilde{u} =$

$(K/\|\bar{u}^K\|_{\mathcal{M}})\bar{u}^K$ there holds $\psi(\tilde{u}) < \psi(\bar{u}^K)$ since $(K/\|\bar{u}^K\|_{\mathcal{M}}) > 1$. This gives a contradiction and $\|\bar{u}^K\|_{\mathcal{M}} = K$.

It remains to show that for a given K the associated Lagrange multiplier denoted by $\beta(K)$ is positive and unique. To prove the positivity, assume that $\beta(K) = 0$. Then we obtain

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

Given $u \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi$, we have $\psi(2u) < \psi(u)$ 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 $\bar{u}^K$ of (P^K) is also a minimizer of $L(\cdot, \beta_1(K))$ and $L(\cdot, \beta_2(K))$ over $\mathcal{M}^+(\Omega_o)$. First we note again that $0 \in \mathcal{M}^+(\Omega_o)$ is not an optimal solution to (P^K) due to the strict monotonicity of Ψ. Additionally it holds $\|\bar{u}^K\|_{\mathcal{M}} = K$. 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 (4.14), we then obtain

$$-\nabla\psi(\bar{u}^K) \leq \beta_1(K) < \beta_2(K), \quad \operatorname{supp} \bar{u}^K \subset \left\{ x \in \Omega_o \mid -\nabla\psi(\bar{u}^K)(x) = \beta_2(K) \right\},$$

implying $\bar{u}^K = 0$ which gives a contradiction. □

Many commonly used optimality criteria Ψ are positively homogeneous in the sense that there exists a convex, strictly decreasing, and positive function γ fulfilling

$$\Psi(rB) = \gamma(r)\, \Psi(B) \quad \forall r > 0,\ B \in \mathrm{PD}(n); \tag{4.17}$$

cf. also [106, p. 26]. For example, both the A-optimal design criterion $\Psi_A(B) = \operatorname{Tr}(B^{-1})$ and the (non-logarithmic) D-criterion $\Psi_D(B) = \det(B^{-1})$ fulfill this homogeneity with γ_A and γ_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_β).

Proposition 4.9. *Assume that $\mathcal{I}_0 = 0$ and Ψ is positive homogeneous in the sense of* (4.17). *Let $\bar{u}_\beta$ be a solution to (P_β) for some fixed $\beta > 0$. Then*

$$K\, \bar{u}_\beta / \|\bar{u}_\beta\|_{\mathcal{M}} \quad \textit{solves} \quad (P^K). \tag{4.18}$$

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

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

by using the positive homogeneity of Ψ. Thus, the solutions of (P^K) are given by Ku^1, where u^1 are solutions of (P^1). Now, using the fact that

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

the solutions $\bar{u}_\beta$ of (P_β) can be computed as $\bar{u}_\beta = Ku^1$, where K minimizes the above expression and $u^1 \in \arg\min(P^1)$. Together, this directly implies (4.18). □

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 β (resp, K), which only affects the scaling of the coefficients $\mathbf{u}$. 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. $\mathcal{I}_0 \in \mathrm{PD}(n)$) 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 4.10. *Let $\mathcal{I}_0 \in \mathrm{PD}(n)$. Then the zero function $\bar{u} = 0$ is an optimal solution to (P_β) if and only if $\beta > \beta_0 = -\min_{x\in\Omega_o} \nabla\psi(0)(x)$.*

Proof. We first note that $0 \in \operatorname{dom}\psi$ and $\beta_0 = -\min_{x\in\Omega_o} \nabla\psi(0)(x) < \infty$. Clearly, for $\beta \geq \beta_0$, the zero function fulfills the optimality conditions from Lemma 4.4. Thus, it is a solution to (P_β). Conversely, for $\beta < \beta_0$, the optimality conditions are violated. □

4.3 An approach by convex duality

To conclude the discussion on the structure of optimal design measurements we mention a different approach for the functional analytic treatment of (P_β) by convex duality. For a dual viewpoint on sparse optimal control problems we refer to [73, 74]. In the context of optimal design problems similar arguments have been used in, e.g., [1, 102, 221]. For simplicity set $\mathcal{I}_0 = 0$. We rewrite the sparse sensor placement problem (P_β) as an unconstrained convex minimization problem

$$\min_{u\in\mathcal{M}(\Omega_o)} \Psi(\mathcal{I}(u)) + \beta\|u\|_{\mathcal{M}} + I_{u\geq 0}(u). \tag{4.19}$$

By applying the Fenchel-Rockefellar duality theorem, see e.g. [229, Section 31], we can identify its dual problem as

$$\min_{B\in\mathrm{Sym}(n)} \Psi^*(-B) \quad s.t. \quad \partial S[q](x)^\top B \partial S[q](x) \leq \beta \quad \forall x \in \Omega_o, \tag{4.20}$$

where $\Psi^*\colon \mathrm{Sym}(n) \to \mathbb{R} \cup \{+\infty\}$ denotes the convex conjugate of Ψ, see (6.5). Any optimal solution $\bar{u}_\beta$ to (4.19) corresponds to a Lagrange multiplier for the pointwise constraint in (4.20). These results are formalized in the following proposition.

Proposition 4.11. *The following statements are equivalent:*

- *The measure $\bar{u} \in \mathcal{M}(\Omega_o)$ is optimal for (4.19) and $\bar{B} \in \mathrm{Sym}(n)$ is optimal for (4.20).*
- *There holds $\bar{B} = -\nabla\Psi(\mathcal{I}(\bar{u}))$, $\partial S[q](x)^\top \bar{B} \partial S[q](x) \leq \beta$ for all $x \in \Omega_o$ and*

$$\langle \mathcal{I}^* \bar{B}, \bar{u}\rangle = \int_{\Omega_o} \partial S[q](x)^\top \bar{B} \partial S[q](x) \mathrm{d}\bar{u}(x) = \beta\|\bar{u}\|_{\mathcal{M}}.$$

Proof. The statement readily follows from applying [98, Proposition 4.1]. □

Hence, the results of Lemma 4.4 can be interpreted as a complementarity condition for the pointwise constraint on $\partial S[q]^\top \bar{B} \partial S[q]$ and its associated multiplier $\bar{u}_\beta$. To illustrate this result we consider a concrete example.

Example 4.1. *We consider the D-optimal design criterion with no a priori knowledge*

$$\Psi_D(B) = \begin{cases} -\log(\det(B)) & B \in \mathrm{PD}(n), \\ +\infty & \textit{else} \end{cases},$$

and the associated sensor placement problem

$$\min_{u \in \mathcal{M}^+(\Omega_o)} [-\log(\det(\mathcal{I}(u))) + \beta \|u\|_{\mathcal{M}}]. \tag{4.21}$$

Let us calculate the convex conjugate of the log-determinant

$$\Psi_D^*\colon \mathrm{Sym}(n) \to \mathbb{R} \cup \{+\infty\}, \quad B \mapsto \sup_{B_1 \in \mathrm{PD}(n)} [\mathrm{Tr}(B^\top B_1) + \log(\det(B_1))].$$

Let $B \in \mathrm{Sym}(n)$ be given. First assume that there exists $\tilde{B} \in \mathrm{PD}(n)$ with $\mathrm{Tr}(B^\top \tilde{B}) \geq 0$. For $t \in \mathbb{R}_+$ define $\tilde{B}_t = t\tilde{B}$. Then we obtain

$$\begin{aligned} \Psi_D^*(B) \geq \sup_{t \in \mathbb{R}_+} [\mathrm{Tr}(B^\top \tilde{B}_t) + \log(\det(\tilde{B}_t))] &= \sup_{t \in \mathbb{R}_+} [t\,\mathrm{Tr}(B^\top \tilde{B}) + \log(\det(\tilde{B}_t))] \\ &= \sup_{t \in \mathbb{R}_+} [t\,\mathrm{Tr}(B^\top \tilde{B}) + n\log(t) + \log(\det(\tilde{B}))] = +\infty. \end{aligned}$$

Thus we conclude that a necessary condition for $B \in \mathrm{dom}\,\Psi_D^*$ is given by $\mathrm{Tr}(B^\top B_1) < 0$ for all $B_1 \in \mathrm{PD}(n)$ or, equivalently, $-B \in \mathrm{PD}(n)$. Recall that for $B_1 \in \mathrm{PD}(n)$ the gradient of the log-determinant criterion is given by $\nabla \Psi_D(B_1) = -B_1^{-1}$. Since the set of positive definite matrices is open in $\mathrm{Sym}(n)$ we conclude

$$\bar{B} \in \arg\max_{B_1 \in \mathrm{PD}} [\mathrm{Tr}(B^\top B_1) - \Psi_D(B_1)] \Rightarrow \nabla \Psi(\bar{B}) = -\bar{B}^{-1} = B^\top.$$

Inserting this into the definition of the convex conjugate we get

$$\Psi_D^*(B) = -n - \Psi_D(-B^{-1}) = -n + \Psi_D(-B) = -n - \log(\det(-B)).$$

. *The dual problem* (4.20) *is now readily given as*

$$\min_{B \in \mathrm{PD}(n)} -\log(\det(B)) \quad \textit{s.t.} \quad \partial S[\hat{q}](x)^\top B \partial S[\hat{q}](x) \leq \beta \quad \forall x \in \Omega_o. \tag{4.22}$$

We give some geometrical interpretation to this problem. Note that if $B \in \mathrm{PD}(n)$ is admissible for (4.20) *there holds*

$$\{\, \partial S[\hat{q}](x) \mid x \in \Omega_o \,\} \subset \mathcal{E}(B) = \left\{ q \in \mathbb{R}^n \mid q^\top B q \leq \beta \right\},$$

where the set on the right hand side is an ellipsoid, centred at the origin. Its shape is described by B. Furthermore we have $\mathrm{vol}(\mathcal{E}(B)) = c(n)\det(B)^{-1/2}$. Therefore, for the log-determinant criterion, the dual to the sensor placement problem is given by finding the ellipsoid of minimal volume which covers all possible observation vectors $\partial S[\hat{q}](x) \in \mathbb{R}^n$, $x \in \Omega_o$, see also [253]. Minimum volume enclosing ellipsoid problems have been discussed in e.g. [175, 219]. Due to the support condition on the optimal measurement design we further conclude that the observation vector $\partial S[\hat{q}](\bar{x})$ corresponding to an optimal measurement at $\bar{x}$ lies at the boundary of the associated ellipsoid.

These duality results establish an important connection between sparse optimal sensor placement and so called semi-infinite programming, see e.g. [140, 187]. Here, the space of optimization variables is finite dimensional, but an infinite number of constraints is imposed. Optimality conditions and existence of a sparse Lagrange multiplier for these kind of problems has been discussed in e.g. [43]. In this chapter we have chosen a primal approach to discuss the sparse optimal sensor problem to demonstrate the applicability of the general framework presented in the previous chapter. Moreover, we are confident that many of the ideas presented in the following sections can be extended to more general and non-convex sparse optimization problems in a straightforward fashion. However, this work largely benefits from the advanced level of research on semi-infinite problems. We mention for example the a priori error estimates for the design measure in Section 4.6 which partly rely on techniques developed for a semi-infinite problem, c.f. [190,191]. On the other hand, semi-infinite programming might also benefit from results obtained from the study of sparse optimization problems. For instance, recently, c.f. [97], the equivalence of an accelerated conditional gradient method for sparse optimization problems, see Section 4.4, and an exchange method for semi-infinite problems, see e.g. [278], has been shown. This allowed to derive worst-case convergence rates for the latter one. In this light, numerical methods for semi-infinite optimization might eventually also profit from the improved convergence results for accelerated conditional gradient methods derived in the following section.

4.4 Optimization aspects

In this section we will elaborate on the algorithmic solution of (P_β). 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 [50] and [49]. By a closer inspection, the resulting algorithm guarantees convergence of the generated sequence of measures towards a minimizer of (P_β) together with a sub-linear convergence rate of the objective function values. Additionally we propose to alternate between point insertion and point deletion steps to enhance the sparsity of the iterates and to speed up the convergence of the algorithm. 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. If the finite-dimensional sub-problems are solved up to optimality in every iteration, we are further able to show improved convergence rates for the objective functional as well as rates for the iterates in a suitable metric. Moreover the resulting algorithms can be combined with Algorithm 1 in a straightforward manner, guaranteeing a sparse structure of the computed optimal design. Finally, the algorithm is compared to variants of the Fedorov-Wynn algorithm for the algorithmic solution of (P^K).

Secondly, we adapt an approach based on a Hilbert space regularization of the original sparse optimization problem. Here, the optimal design problem (P_β) is replaced by a sequence of regularized optimization problems, which are amenable to proximal point or semismooth Newton methods (which converge locally superlinearly) in function space. Algorithmic approaches for the solution of non-smooth optimization problems based on Hilbert space regularizations have recently increased in interest in the context of PDE-constrained optimization; see, e.g., [73,248]. Since such an approach seems to be new in the context of sensor placement problems, we briefly describe it for the sake of comparison at the end of this section.

4.4.1 A generalized conditional gradient method

For the direct solution of (P_β) on the admissible set $\mathcal{M}^+(\Omega_o)$ we adapt the numerical procedure presented in [50], which relies on finitely supported iterates. A general description of the method is given in Algorithm 2. For convenience of the reader we give a detailed description of the individual steps and their derivation below. Basically, the algorithm can be split into two parts. The first part (steps 2.–4. in Algorithm 2) consists of adding a new sensor to the current measurement design. In the second part (step 5.), 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. 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 and, as we will see, also in theory.

Algorithm 2 Sequential point insertion algorithm

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

This section is structured as follows: First, we focus only on the point insertion step and prove its connection to a generalized conditional method as described in Chapter 6. Thus, by a suitable choice of the stepsize s^k in each step of the procedure we are able to prove a sub-linear convergence rate for the objective functional value. In the second part, we consider two concrete examples for the point removal step 5. and discuss the applicability of Algorithm 1 in the context of the successive point insertion algorithm. Since most of the statements in this section are obtained through applying the general theory in Chapter 6 we omit the majority of proofs in the following.

Let us first recall that the set of optimal solutions to (P_β) is bounded by a constant $M_0 > 0$. For example, if ψ is nonnegative on its domain, we can choose an arbitrary but fixed $u \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi$ to obtain

$$\beta\|\bar{u}_\beta\|_{\mathcal{M}} \leq F(\bar{u}_\beta) \leq F(u),$$

for every optimal design $\bar{u}_\beta \in \mathcal{M}^+(\Omega_o)$. Hence we can set $M_0 = F(u)/\beta$. We now consider the slightly modified problem

$$\min_{u\in\mathcal{M}^+(\Omega_o), \|u\|_{\mathcal{M}}\leq M_0} [\psi(u) + \beta\|u\|_{\mathcal{M}}]. \tag{$P_\beta^{M_0}$}$$

Connected to this auxiliary problem we further define the primal-dual gap

$$\Phi\colon \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi \to [0,\infty),$$

which is given by

$$\Phi(u) = \sup_{v\in\mathcal{M}^+(\Omega_o), \|v\|_{\mathcal{M}}\leq M_0} \left[\langle\nabla\psi(u), u - v\rangle + \beta\|u\|_{\mathcal{M}} - \beta\|v\|_{\mathcal{M}})\right].$$

In the next proposition we collect several results to establish the connection between the optimal design problems (P_β) and ($P_\beta^{M_0}$). Furthermore, the primal-dual gap gives an upper bound for the error in the objective function value at the k-th iterate.

Proposition 4.12. *Given $\bar{u}_\beta \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi$ the following three statements are equivalent:*

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

Furthermore there holds

$$\Phi(u) \geq F(u) - F(\bar{u}_\beta) =: r_F(u), \tag{4.23}$$

for all $u \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi$ with $\|u\|_{\mathcal{M}} \leq M_0$ and all minimizers $\bar{u}_\beta$ of ($P_\beta^{M_0}$).

Following relation (4.23), the primal-dual gap Φ is suitable to monitor the convergence of Algorithm 2. Its numerical computation is discussed in an instance. We now connect the definition of the new sensor v^k (see step 2.–3.) to the minimization of a partial linearization of ($P_\beta^{M_0}$).

Lemma 4.13. *Let $u^k \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi$ 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 2 is a minimizer of*

$$\min_{v \in \mathcal{M}^+(\Omega_o), \|v\|_{\mathcal{M}} \leq M_0} [\langle \nabla\psi(u^k), v\rangle + \beta \|v\|_{\mathcal{M}}]. \tag{P_β^{lin}}$$

Moreover, v^k realizes the supremum in the definition of the primal-dual gap:

$$\Phi(u^k) = \langle \nabla\psi(u^k), u^k - v^k\rangle + \beta\|u^k\|_{\mathcal{M}} - \beta\|v^k\|_{\mathcal{M}}.$$

Proof. We note that (P_β^{lin}) can be equivalently expressed as

$$\min_{r \in [0,M_0]} \min_{\substack{\tilde{v} \in \mathcal{M}^+(\Omega_o),\\ \|\tilde{v}\|_{\mathcal{M}} = 1}} [r\langle \nabla\psi(u^k), \tilde{v}\rangle + \beta r] = \min_{r \in [0,M_0]} [r \min_{x \in \Omega_o} \nabla\psi(u^k)(x) + \beta r].$$

The concrete expression of v^k follows now by a straightforward computation. Clearly, $\Phi(u^k)$ agrees to $-\min$ (P_β^{lin}) up to a constant value. □

We make the following two observations: First, we can interpret Algorithm 2 as a generalized conditional gradient method as described in Chapter 6. Second, as a by-product of the last result, the convergence criterion $\Phi(u^k)$ can be evaluated cheaply once the current gradient $\nabla\psi(u^k)$ and its minimum point $\hat{x}^k$ are calculated.

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

$$\min_{v \in \mathcal{M}^+(\Omega_o)} [\langle \nabla\psi(u), v\rangle + \beta\|v\|_{\mathcal{M}}],$$

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

In the following, we will determine the stepsize s^k according to a Quasi-Armijo-Goldstein condition. We set $s^k = \gamma^{n_k}$, where $\gamma \in (0,1), \alpha \in (0,1/2)$, $u^k_{s^k} = u^k + s^k(v^k - u^k)$ and n_k is the smallest integer fulfilling

$$\alpha s^k \Phi(u^k) \leq F(u^k) - F(u^k_{s^k}). \tag{4.24}$$

For more information on the feasibility for this choice of the stepsize see the discussions around Lemma 6.9.

We now turn our attention to the convergence of Algorithm 2. Therefore we note that $\nabla\psi$ is Lipschitz continuous on the sublevel sets of F.

Proposition 4.14. *Let $u^1 \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi$ be given. Define the associated sub-level set $E_F(u^1)$ as*

$$E_F(u^1) = \left\{ u \in \mathcal{M}^+(\Omega_o) \mid F(u) \leq F(u^1) \right\}.$$

Then there exists L_{u^1} such that

$$\sup_{\substack{u_1,u_2 \in E_F(u^1)\\ u_1 \neq u_2}} \frac{\|\nabla\psi(u_1) - \nabla\psi(u_2)\|_{\mathcal{C}}}{\|u_1 - u_2\|_{\mathcal{M}}} \leq L_{u^1}. \tag{4.25}$$

Proof. Since Ψ is two times continuously differentiable on its domain, its gradient $\nabla\Psi$ is Lipschitz continuous on compact subsets. We observe that $E_F(u^1)$ is convex, bounded, and weak* closed. Consequently, the set of associated information matrices

$$\mathcal{I}(E_F(u^1)) = \left\{ \mathcal{I}(u) + \mathcal{I}_0 \mid u \in E_F(u^1) \right\} \subset \operatorname{dom}\Psi,$$

is compact. For $u_1, u_2 \in E_F(u^1)$ we obtain

$$\begin{aligned}\|\nabla\psi(u_1) - \nabla\psi(u_2)\|_{\mathcal{C}} &= \|\mathcal{I}^*\nabla\Psi(\mathcal{I}(u_1)+\mathcal{I}_0) - \mathcal{I}^*\nabla\Psi(\mathcal{I}(u_2)+\mathcal{I}_0)\|_{\mathcal{C}}\\ &\leq \|\mathcal{I}^*\|_{\mathcal{L}(\mathrm{Sym}(n),\mathcal{C}(\Omega_o))}\|\nabla\Psi(\mathcal{I}(u_1)+\mathcal{I}_0) - \nabla\Psi(\mathcal{I}(u_2)+\mathcal{I}_0)\|_{\mathrm{Sym}}\\ &\leq L_{\mathcal{I}(E_F(u^1))}\|\mathcal{I}^*\|_{\mathcal{L}(\mathrm{Sym}(n),\mathcal{C}(\Omega_o))}\|\mathcal{I}(u_1) - \mathcal{I}(u_2)\|\\ &\leq L_{\mathcal{I}(E_F(u^1))}\|\mathcal{I}^*\|_{\mathcal{L}(\mathrm{Sym}(n),\mathcal{C}(\Omega_o))}\|\mathcal{I}\|_{\mathcal{L}(\mathcal{M}(\Omega_o),\mathrm{Sym}(n))}\|u_1 - u_2\|_{\mathcal{M}},\end{aligned}$$

where $L_{\mathcal{I}(E_F(u^1))}$ denotes the Lipschitz constant of $\nabla\Psi$ on $\mathcal{I}(E_F(u^1))$. This completes the proof. □

Combining all the previous results we conclude the following worst-case convergence result.

Theorem 4.15. *Let the sequence $\{u^k\}_{k\in\mathbb{N}}$ be generated by Algorithm 2 with s^k chosen according to the Quasi-Armijo-Goldstein condition. Then $\{u^k\}_{k\in\mathbb{N}}$ is a minimizing sequence of F and admits a weak* accumulation point $\bar{u}_\beta$. Every such point is an optimal solution to (P_β). Additionally there holds*

$$r_F(u^k) \leq \frac{r_F(u^1)}{1 + q(k-1)}, \quad q = \alpha \min\left\{ \frac{c_1}{L_{u^1}(2M_0)^2}, 1 \right\}. \tag{4.26}$$

Here, L_{u^1} is the Lipschitz-constant of $\nabla\psi$ on $E_F(u^1)$ and $c_1 = 2\gamma(1-\alpha)r_F(u_1)$.

Proof. It is readily verified that the problem fulfills the prerequisites of Theorem 6.29. Thus, the statement follows. □

4.4.2 Acceleration and sparsification strategies

As we have seen in the previous section, an iterative application of steps 2.–4. in Algorithm 2 is sufficient to obtain weak* convergence of the iterates u^k, as well as a sublinear convergence rate for the objective function. However, it is obvious that the support size of the iterates u^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_β). In the following we mitigate those effects by augmenting the point insertion steps by point removal steps, where we incorporate ideas from [44, 50]. Without loss of generality we can assume that $M_0 > 0$ is chosen large enough such that

$$F(u^{k+1}) \leq F(u^{k+1/2}) \leq F(u^k) \Rightarrow \max\{\|u^k\|_{\mathcal{M}}, \|u^{k+1/2}\|_{\mathcal{M}}\} \leq M_0,$$

due to the radial unboundedness of F. Given an ordered set of pairwise distinct points $\mathcal{A} = \{x_1, \dots, x_N\}$, we define the parameterization:

$$\boldsymbol{u}_{\mathcal{A}}(\mathbf{u}) := \sum_{x_i \in \mathcal{A}} \mathbf{u}_i \delta_{x_i} \quad \forall \mathbf{u} \in \mathbb{R}^N. \tag{4.27}$$

Now, we set $\mathcal{A} = \mathcal{A}_k = \operatorname{supp} u^{k+1/2}$, $m_k = \#\mathcal{A}$ and $u^{k+1} = \boldsymbol{u}_{\mathcal{A}}(\mathbf{u}^{k+1})$, where the improved vector $\mathbf{u}^{k+1} \in \mathbb{R}^{m_k}$ is chosen as an approximate solution to the (finite dimensional) coefficient optimization problem

$$\min_{\mathbf{u} \in \mathbb{R}^{m_k}_+} F(\boldsymbol{u}_{\mathcal{A}}(\mathbf{u})) = [\psi(\boldsymbol{u}_{\mathcal{A}}(\mathbf{u})) + \beta \|\mathbf{u}\|_{l_1}], \tag{4.28}$$

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

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

$$\mathbf{u}^{k+1}(\sigma_k)_i = \max\left\{ \mathbf{u}_i^{k+1/2} - \sigma_k \left[\nabla\psi(u^{k+1/2})(x_i) + \beta \right],\ 0 \right\} \quad \forall i \in \{1, \dots, m_k\}, \tag{4.29}$$

where $\sigma_k > 0$ is a suitably chosen step size that avoids ascend in the objective function value. This corresponds to performing one step of a projected gradient method on (4.28) using the previous coefficient vector $\mathbf{u}^{k+1/2}$ as a starting point. Thus, step 5. in Algorithm 2 subtracts or adds mass at support point x_i for $-\nabla\psi(u^{k+1/2})(x_i) < \beta$ or $-\nabla\psi(u^{k+1/2})(x_i) > \beta$, respectively. Furthermore, the new coefficient $\mathbf{u}_i^{k+1}$ of the Dirac delta function δ_{x_i} is set to zero if

$$\mathbf{u}_i^{k+1/2} - \sigma_k(\nabla\psi(u^{k+1/2})(x_i) + \beta) \leq 0,$$

removing the point measure from the iterate.

Secondly, we suppose that the finite-dimensional sub-problems (4.28) can be solved exactly and choose

$$\mathbf{u}^{k+1} \in \operatorname*{arg\,min}_{\mathbf{u} \in \mathbb{R}^{m_k}_+} F(\boldsymbol{u}_{\mathcal{A}}(\mathbf{u})). \tag{4.30}$$

In this case, the conditions

$$\operatorname{supp} u^{k+1} \subset \operatorname{supp} u^{k+1/2}, \quad F(u^{k+1}) \leq F(u^{k+1/2}), \quad \|u^{k+1}\|_{\mathcal{M}} \leq M_0,$$

Algorithm 3 Primal-Dual-Active-Point strategy for (P_β)

```
while $\Phi(u^k) \geq \mathrm{TOL}$ do
    1. Calculate $\nabla\psi_k = \nabla\psi(u^k)$. Determine $\hat{x}^k \in \arg\min_{x\in\Omega_o} \nabla\psi_k(x)$.
    2. Set $\mathcal{A}_k = \operatorname{supp} u^k \cup \{\, \hat{x}^k \,\}$, compute a solution to $\mathbf{u}^{k+1}$ of (4.28) for $\mathcal{A} = \mathcal{A}_k$, and set
    $u^{k+1} = \boldsymbol{u}_{\mathcal{A}}(\mathbf{u}^{k+1})$.
end while
```

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: In each iteration, the minimizer $\hat{x}^k$ of the current gradient $\nabla\psi_k$ is added to the support set to obtain $\mathcal{A}_k = \operatorname{supp} u^k \cup \{\, \hat{x}^k \,\}$. Then, the problem (4.30) is solved on the new support set to obtain the next iterate u^{k+1}. Note that the next active set is given by $\mathcal{A}_{k+1} = \operatorname{supp} u^{k+1} \cup \{\hat{x}^{k+1}\}$, which automatically removes support points corresponding to zero coefficients in each iteration. Furthermore the method terminates after finitely many steps if $\mathcal{A}_k = \mathcal{A}_{k+1}$ for some $k \in \mathbb{N}$. Since the subproblems are solved up to optimality we conclude

$$\nabla\psi_{k+1}(x) \geq -\beta, \quad \forall x \in \operatorname{supp} u^{k+1/2}, \quad -\langle \nabla\psi_{k+1}, u^{k+1}\rangle = \beta\|u^{k+1}\|_{\mathcal{M}}.$$

Especially, the primal-dual gap at a non-optimal u^k coincides, up to a constant, with the constraint violation of the associated gradient, $\Phi(u^k) = -M_0(\min_{x\in\Omega_o} \nabla\psi_k(x) + \beta)$.

Finally, to be able to guarantee the a priori bound $\#\operatorname{supp} u^k \leq n(n+1)/2$ for the algorithmic solutions, we can apply Algorithm 1 to the intermediate iterate $u^{k+1/2}$ in step 5. of Algorithm 2. This ensures the convergence of the presented procedure towards a sparse minimizer of (P_β).

Proposition 4.16. *Assume that $\#\operatorname{supp} u^1 \leq n(n+1)/2$ and let u^{k+1} be obtained by applying Algorithm 1 to $u^{k+1/2}$ in each iteration of Algorithm 2. Then the results of Theorem 4.15 hold. Furthermore we obtain $\#\operatorname{supp} u^k \leq n(n+1)/2$ for all $k \in \mathbb{N}$ and consequently $\#\operatorname{supp} \bar{u}_\beta \leq n(n+1)/2$ for every weak* accumulation point $\bar{u}_\beta$ of u^k.*

Proof. See Theorem 6.36. □

We emphasize that the sparsifying procedure from Algorithm 1 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 (4.29) or (4.30) obtaining an intermediate iterate $u^{k+3/4}$. Subsequently we apply Algorithm 1. Since in both cases, the number of support points cannot increase, the statements of the last proposition remain true.

Remark 4.4. Note that Algorithm 2 can be easily generalized to allow for the insertion of more than one point in every iteration, which yields an additional practical speed up of the method. In detail, the results of Theorem 4.15 and Proposition 4.16 hold true if the search direction $v^k \in \mathcal{M}^+(\Omega_o)$ from Lemma 4.13 is more generally chosen as

$$v^k = \sum_{i=1}^m \mathbf{u}_i \delta_{x_i}, \quad \{x_i\}_{i=1}^m \subset \operatorname*{arg\,min}_{x\in\Omega_o} \nabla\psi_k(x), \quad \|v^k\|_{\mathcal{M}} = M_0$$

if $\min_{x\in\Omega_o} \nabla\psi_k \leq -\beta$. Moreover all results remain valid for Algorithm 3 if we compute $\mathbf{u}^{k+1}$ as the solution of the coefficient minimization problem (4.30) with with a general active set $\mathcal{A}_k$ which contains $\operatorname{supp} u^k \cup \{\, \hat{x}_k \,\}$.

4.4.3 Improved convergence results

In this section we will improve on the convergence result of Theorem 4.15 in the special case of Algorithm 3. Therefore we briefly recall that given a minimizer $\bar{u}_\beta$ of (P_β) fulfills

$$-\nabla\psi(\bar{u}_\beta)(x) \leq \beta, \quad \forall x \in \Omega_o, \quad \operatorname{supp} \bar{u}_\beta \subset \{\, x \in \Omega_o \mid -\nabla\psi(\bar{u}_\beta)(x) = \beta \} \,.$$

We make the following assumptions on the optimal design criterion Ψ, the maximizers of $-\nabla\psi(\bar{u}_\beta)$ and the local regularity of $\partial S[\hat{q}]$.

Assumption 4.4. The design criterion Ψ is strictly convex on its domain and the interior of Ω_o is not empty. Define the unique optimal gradient $\bar{p} = -\nabla\psi(\bar{u}_\beta)$ and assume that there exists $N \in \mathbb{N}$ and $\bar{x}_i \in \operatorname{int} \Omega_o$, $i = 1, \ldots, N$, with

$$\operatorname{supp} \bar{u}_\beta \subset \{\, x \in \Omega_o \mid \bar{p}(x) = \beta\} = \{\bar{x}_i\}_{i=1}^N.$$

Furthermore, the set $\{\mathcal{I}(\delta_{\bar{x}_i})\}_{i=1}^N$ is linear independent and there exists a constant $R > 0$ with

$$\Omega_R := \bigcup_{i=1}^N B_R(\bar{x}_i) \subset \operatorname{int} \Omega_o, \quad \bar{B}_R(\bar{x}_i) \cap \bar{B}_R(\bar{x}_j) = \emptyset, \quad \partial S[\hat{q}] \in \mathcal{C}^2(\bar{\Omega}_R, \mathbb{R}^n) \cap \mathcal{C}(\Omega_o, \mathbb{R}^n).$$

for all $i, j \in \{1, \ldots, N\}$.

The assumption on the global maximizers of $\bar{p}$ together with the linear independence of the associated Fisher information matrices guarantee the sparsity as well as the uniqueness of the optimal design, see Corollaries 3.21 and 3.19. Moreover, we conclude that $\mathcal{I}^*$ maps continuously to $\mathcal{C}^2(\bar{\Omega}_R) \cap \mathcal{C}(\Omega_o)$ since

$$\mathcal{I}^* B = \partial S[\hat{q}](x)^\top B \partial S[\hat{q}](x) \quad \forall B \in \operatorname{Sym}(n).$$

In particular, since $\bar{p}(x) = -\mathcal{I}^* \nabla\Psi(\mathcal{I}(\bar{u}_\beta) + \mathcal{I}_0)$, we immediately get $\bar{p} \in \mathcal{C}^2(\bar{\Omega}_R) \cap \mathcal{C}(\Omega_o)$. To derive improved convergence rates we demand that the following second order conditions for the optimal Fisher information matrix $\mathcal{I}(\bar{u}_\beta)$ and the optimal sensor positions $\{\bar{x}_i\}_{i=1}^N$ hold.

Assumption 4.5. There holds $\bar{u}_\beta = \sum_{i=1}^N \bar{\mathbf{u}}_i \delta_{\bar{x}_i}$ for some $\bar{\mathbf{u}}_i > 0$ and

$$\operatorname{Tr}(B \nabla^2 \Psi(\mathcal{I}(\bar{u}_\beta) + \mathcal{I}_0) B) \geq \gamma_0 \|B\|_{\text{Sym}}^2, \quad \forall B \in \operatorname{Sym}(n).$$

Moreover the Hessian of $\bar{p} \in \mathcal{C}(\Omega_o) \cap \mathcal{C}^2(\bar{\Omega}_R)$ is negative-definite at each $\bar{x}_i$: there exists $\theta > 0$ with

$$-(\zeta, \nabla^2 \bar{p}(\bar{x}_i) \zeta)_{\mathbb{R}^d} \geq \theta |\zeta|_{\mathbb{R}^d}^2, \quad \forall \zeta \in \mathbb{R}^d$$

for all $i = 1, \ldots, N$.

For the rest of this section let Assumptions 4.4 and 4.5 hold. Obviously, the above assumptions guarantee uniform convexity of Ψ around $\mathcal{I}(\bar{u}_\beta) + \mathcal{I}_0$.

Corollary 4.17. *There exists a neighbourhood $N(\mathcal{I}(\bar{u}_\beta))$ of $\mathcal{I}(\bar{u}_\beta)$ in* Sym(n) *with*

$$(\nabla\Psi(B_1 + \mathcal{I}_0) - \nabla\Psi(B_2 + \mathcal{I}_0), B_1 - B_2)_{\text{Sym}} \geq \frac{\gamma_0}{2} \|B_1 - B_2\|_{\text{Sym}}^2 \quad \forall B_1, B_2 \in N(\mathcal{I}(\bar{u}_\beta)).$$

Proof. Let $B_1, B_2 \in \mathrm{NND}(n)$ with $B_i + \mathcal{I}_0 \in \operatorname{dom} \Psi$, $i = 1, 2$, be given. Using Taylor's expansion we get

$$(\nabla\Psi(B_1 + \mathcal{I}_0) - \nabla\Psi(B_2 + \mathcal{I}_0), B_1 - B_2)_{\mathrm{Sym}} = (B_1 - B_2, \nabla^2\Psi(B_\zeta + \mathcal{I}_0)(B_1 - B_2))_{\mathrm{Sym}}$$

for some $B_\zeta = B_1 + \zeta(B_2 - B_1)$, $\zeta \in (0,1)$. Since Ψ is two times continuously Fréchet differentiable its Hessian is uniformly continuous on compact sets. Hence by choosing $N(\mathcal{I}(\bar{u}_\beta))$ small enough there holds $\bar{N}(\mathcal{I}(\bar{u}_\beta)) \subset \operatorname{dom} \Psi$ due to the openess of the domain and

$$\|\nabla^2\Psi(B_\zeta + \mathcal{I}_0) - \nabla^2\Psi(\mathcal{I}(\bar{u}_\beta) + \mathcal{I}_0)\|_{\mathcal{L}(\mathrm{Sym}(n),\mathrm{Sym}(n))} \leq \frac{\gamma_0}{2} \quad \forall B_1, B_2 \in N(\mathcal{I}(\bar{u}_\beta)), \zeta \in (0,1).$$

Combining both results and using the definiteness of $\nabla^2\Psi$ at $\mathcal{I}(\bar{u}_\beta) + \mathcal{I}_0$ we conclude the statement. □

We arrive at the following improved convergence rate for the objective function values.

Theorem 4.18. *Denote by $\{u^k\}_{k\in\mathbb{N}}$ the sequence obtained through Algorithm 3. Let the assumptions of Theorem 4.15, Assumption 4.4 as well as Assumption 4.5 hold. Assume that the algorithm does not terminate after finitely many steps. Then there exists $0 < \zeta < 1$ and a constant $c > 0$ independent of k with*

$$r_F(u^k) = F(u^k) - F(\bar{u}_\beta) \leq c\zeta^k,$$

for all $k \in \mathbb{N}$ large enough.

Proof. Due to the stated assumptions, the functionals $\Psi(\cdot + \mathcal{I}_0)$, G, the operator $\mathcal{I}$, the unique minimizer $\bar{u}_\beta$ and $\bar{p}$ fulfill Assumptions 6.3, 6.4, 6.5 and 6.6 noting that $\bar{p} = |\nabla\psi(\bar{u}_\beta)|$. Thus, we can apply Theorem 6.70 taking Remark 6.14 into account. This yields the desired result. □

The rest of this section is devoted to the establishment of convergence rates for the iterates u^k produced by Algorithm 3. Therefore let us briefly gather the facts so far. Due to the uniqueness of the optimal design we have $u^k \rightharpoonup^* \bar{u}_\beta$ for the whole sequence. Furthermore we have $-\nabla\psi_k \to \bar{p}$ in $\mathcal{C}(\Omega_o)$ and

$$-\nabla\psi_k(x) = \beta, \quad \forall x \in \operatorname{supp} u^k,$$

due to the optimality of $\mathbf{u}^k$ for (4.28) and $k > 1$. Combining this with the fact that by assumption $\bar{p}(x) < \beta$ for all $x \notin \operatorname{supp} \bar{u}_\beta$ we conclude $\operatorname{supp} u^k \subset \Omega_R$ for all k large enough. Hence, by denoting with $u_i^k \in \mathcal{M}^+(\Omega_o)$ the restriction of u^k to $B_R(\bar{x}_i)$, we also obtain $u_i^k \rightharpoonup^* \bar{\mathbf{u}}_i\delta_{\bar{x}_i}$, $i = 1, \ldots N$, by testing u^k with suitable continuous functions.

In general, norm convergence of $\{u^k\}_{k\in\mathbb{N}}$ on $\mathcal{M}(\Omega_o)$ cannot be expected. We recall that the iterates as well as the optimal design $\bar{u}_\beta$ are given as conic combinations of finitely many Dirac delta functions which each correspond to the setup of a measurement experiment. From a practical point of view, the most important question concerns the convergence of the sensor positions and the associated measurement weights. In view of the aforementioned clustering effects, a sensor located at an optimal position $\bar{x}_i$ is in most cases approximated by several sensors in the iterated design u^k. Consequently, the convergence of the sensors in the restricted design u_i^k towards the sensor represented by $\bar{\mathbf{u}}_i\delta_{\bar{x}_i}$ has to be addressed. Moreover, on a more abstract level, quantitative convergence results for the sequence $\{u^k\}_{k\in\mathbb{N}}$ can be obtained when resorting to weaker metrics.

These topics are covered by the following discussion For this purpose, let us first define the set of Lipschitz continuous functions $\mathcal{C}^{0,1}(\Omega_o)$ on Ω_o by

$$\mathcal{C}^{0,1}(\Omega_o) = \{\, \varphi \in \mathcal{C}(\Omega_o) \mid \exists L > 0\colon |\varphi(x_1) - \varphi(x_2)| \leq L|x_1 - x_2|_{\mathbb{R}^d} \quad \forall x_1, x_2 \in \Omega_o \,\}\,.$$

For $\varphi \in \mathcal{C}^{0,1}(\Omega_o)$, the quotient

$$\|\varphi\|_{\mathrm{Lip}} = \sup_{\substack{x_1,x_2 \in \Omega_o,\\ x_1 \neq x_2}} \frac{|\varphi(x_1) - \varphi(x_2)|}{|x_1 - x_2|_{\mathbb{R}^d}},$$

is finite and will be called its Lipschitz constant. The set $\mathcal{C}^{0,1}(\Omega_o)$ together with the norm $\|\varphi\|_{\mathcal{C}^{0,1}} = \|\varphi\|_{\mathcal{C}} + \|\varphi\|_{\mathrm{Lip}}$ forms a Banach space. Since Lipschitz continuous functions are in particular continuous, we have $\mathcal{C}^{0,1}(\Omega_o) \hookrightarrow \mathcal{C}(\Omega_o)$ and thus $\mathcal{M}(\Omega_o) \hookrightarrow \mathcal{C}^{0,1}(\Omega_o)^*$, where the duality pairing is realized as

$$\langle \varphi, u \rangle_{\mathcal{C}^{0,1},\mathcal{C}^{0,1*}} = \langle \varphi, u \rangle = \int_{\Omega_o} \varphi(x) \mathrm{d}u(x),$$

for all $u \in \mathcal{M}(\Omega_o)$ and $\varphi \in \mathcal{C}^{0,1}(\Omega_o)$. Moreover, if e.g. Ω_o is quasi-convex, the space of Lipschitz continuous functions on Ω_o can be identified with $W^{1,\infty}(\Omega_o)$, the space of essentially bounded functions with essentially bounded weak derivative, see [134, Theorem 4.1]. We now define the modified Wasserstein distance of two measures as follows.

Definition 4.1. Given two probability measures μ_1 and μ_2 we define their Wasserstein-1 Distance as

$$W_1(\mu_1, \mu_2) = \sup \left\{\, \langle \varphi, \mu_1 - \mu_2 \rangle \mid \varphi \in \mathcal{C}^{0,1}(\Omega_o),\ \|\varphi\|_{\mathrm{Lip}} \leq 1 \,\right\}.$$

Let now $u_1, u_2 \in \mathcal{M}^+(\Omega_o)$, $u_1, u_2 \neq 0$ be given. We define the modified Wasserstein distance between u_1 and u_2 by

$$\bar{W}_1(u_1, u_2) = W_1(u_1/\|u_1\|_{\mathcal{M}}, u_2/\|u_2\|_{\mathcal{M}}) + |\|u_1\|_{\mathcal{M}} - \|u_2\|_{\mathcal{M}}|.$$

Since the Wasserstein-1 distance metrizes weak* convergence, see [117], we have

$$u_k \rightharpoonup^* u \Rightarrow W_1(u_k/\|u_k\|_{\mathcal{M}}, u/\|u\|_{\mathcal{M}}) \to 0, \quad \|u_k\|_{\mathcal{M}} \to \|u\|_{\mathcal{M}}$$

for every sequence $\{u_k\}_{k\in\mathbb{N}} \subset \mathcal{M}^+(\Omega_o)$ with $u \neq 0$, k large enough. The following result relates the modified Wasserstein distance to the norm on the dual space of $\mathcal{C}^{0,1}(\Omega_o)$

$$\|\varphi\|_{\mathcal{C}^{0,1*}} = \sup \left\{ \langle \varphi, u \rangle_{\mathcal{C}^{0,1},\mathcal{C}^{0,1*}} \mid \varphi \in \mathcal{C}^{0,1}(\Omega_o),\ \|\varphi\|_{\mathcal{C}^{0,1}} \leq 1 \right\}.$$

Proposition 4.19. *Let $u_1, u_2 \in \mathcal{M}^+(\Omega_o)$ with $u_1, u_2 \neq 0$ be given. Then there holds*

$$\|u_1 - u_2\|_{\mathcal{C}^{0,1*}} \leq c_{\|u_2\|_{\mathcal{M}}} \bar{W}_1(u_1, u_2),$$

for some constant $c_{\|u_2\|_{\mathcal{M}}} > 0$ depending on the norm of u_2.

Proof. Let $u_1,\ u_2 \in \mathcal{M}^+(\Omega_o)$ be given. First we note that $\|u_i\|_{\mathcal{C}^{0,1*}} \leq \|u_i\|_{\mathcal{M}},\ i = 1, 2$. We estimate

$$\|u_1 - u_2\|_{\mathcal{C}^{0,1*}} \leq |\|u_1\|_{\mathcal{M}} - \|u_2\|_{\mathcal{M}}| + \|u_2\|_{\mathcal{M}}\|u_1/\|u_1\|_{\mathcal{M}} - u_2/\|u_2\|_{\mathcal{M}}\|_{\mathcal{C}^{0,1*}}.$$

Define $\tilde{u}_1 = u_1/\|u_1\|_{\mathcal{M}}$ and $\tilde{u}_2 = u_2/\|u_2\|_{\mathcal{M}}$. We obtain

$$\|\tilde{u}_1 - \tilde{u}_2\|_{\mathcal{C}^{0,1*}} = \sup\{\, \langle \varphi, \tilde{u}_1 - \tilde{u}_2\rangle \mid \|\varphi\|_{\mathcal{C}^{0,1}} \leq 1 \,\} \leq \sup\{\, \langle \varphi, \tilde{u}_1 - \tilde{u}_2\rangle \mid \|\varphi\|_{\text{Lip}} \leq 1 \,\} = W_1(\tilde{u}_1, \tilde{u}_2).$$

The statement readily follows. □

We now turn our attention to the modified Wasserstein distance between a sequence of sparse measures and its limit. Consider sequences $\{u_i^k\}_{k\in\mathbb{N}} \subset \mathcal{M}^+(\Omega_o), i = 1, \ldots, N$, fulfilling $\#\operatorname{supp} u_i^k < \infty$ for all $k \in \mathbb{N},\ i = 1, \ldots, N$ as well as $u_i^k \rightharpoonup^* \mathbf{u}_i\delta_{x_i}$ for some $x_i \in \Omega_o$ and $\mathbf{u}_i > 0$. Define $u^k = \sum_{i=1}^N u_i^k$ and the limit measure $u = \sum_{i=1}^N \mathbf{u}_i\delta_{x_i}$. The following theorem gives an upper bound on the modified Wasserstein distance between u^k and u in terms of their support points and coeffcients, respectively.

Theorem 4.20. *Without loss of generality assume that $u_i^k \neq 0$ for all $k \in \mathbb{N}$ and $i = 1, \ldots, N$. Then there exists a constant $c_{\|u\|_{\mathcal{M}},N} > 0$ depending on the number of Diracs in the limit u and its norm such that the estimate*

$$\bar{W}(u^k, u) \leq c_{\|u\|_{\mathcal{M}},N}(\max_{i=1,\ldots,N}\ \max_{x\in\operatorname{supp} u_i^k} |x - x_i|_{\mathbb{R}^d} + \max_{i=1,\ldots,N} |\|u_i^k\|_{\mathcal{M}} - \mathbf{u}_i| + |\|u^k\|_{\mathcal{M}} - \|u\|_{\mathcal{M}}|), \tag{4.31}$$

is valid.

Proof. We establish an upper bound for $W_1(u^k/\|u^k\|_{\mathcal{M}}, u/\|u\|_{\mathcal{M}})$. Therefore note that given an arbitrary but fixed $x_0 \in \Omega_o$ there holds

$$\langle \varphi, u^k/\|u^k\|_{\mathcal{M}} - u/\|u\|_{\mathcal{M}}\rangle = \langle \varphi - \varphi(x_0), u^k/\|u^k\|_{\mathcal{M}} - u/\|u\|_{\mathcal{M}}\rangle.$$

Hence the Wasserstein distance can be restricted to

$$W_1(\mu_1, \mu_2) = \sup\left\{\, \langle \varphi, \mu_1 - \mu_2\rangle \mid \varphi \in \mathcal{C}^{0,1}(\Omega_o),\ \|\varphi\|_{\text{Lip}} \leq 1,\ \varphi(x_0) = 0 \,\right\}.$$

Each such $\varphi \in \mathcal{C}^{0,1}(\Omega_o)$ is uniformly bounded due to

$$\|\varphi\|_{\mathcal{C}} = \max_{x\in\Omega_o} |\varphi(x) - \varphi(x_0)| \leq \|\varphi\|_{\text{Lip}}|x - x_0|_{\mathbb{R}^d} \leq 2M,$$

where $M > 0$ is a constant bounding the elements of Ω_o. We estimate

$$|\langle \varphi, u^k/\|u^k\|_{\mathcal{M}} - u/\|u\|_{\mathcal{M}}\rangle| \leq c_{\|u\|_{\mathcal{M}}}(|\|u^k\|_{\mathcal{M}} - \|u\|_{\mathcal{M}}|\|\varphi\|_{\mathcal{C}} + |\langle \varphi, u^k - u\rangle|),$$

for some constant $c_{\|u\|_{\mathcal{M}}} > 0$ only depending on the norm of u if k is large enough. We partition the second term to obtain

$$|\langle \varphi, u^k - u\rangle| \leq \sum_{i=1}^N |\mathbf{u}_i\varphi(x_i) - \langle \varphi, u_k^i\rangle| \leq \sum_{i=1}^N [|\mathbf{u}_i - \|u_i^k\|_{\mathcal{M}}|\|\varphi\|_{\mathcal{C}} + |\|u_i^k\|_{\mathcal{M}}\varphi(x_i) - \langle \varphi, u_i^k\rangle|].$$

We proceed to

$$\sum_{i=1}^{N} [|\mathbf{u}_i - \|u_k^i\|_{\mathcal{M}}| \|\varphi\|_{\mathcal{C}} + |\|u_i^k\|_{\mathcal{M}} \varphi(x_i) - \langle \varphi, u_i^k \rangle|].$$

$$\leq N(\max_{i=1,\dots,N} |\mathbf{u}_i - \|u_i^k\|_{\mathcal{M}}| \|\varphi\|_{\mathcal{C}} + \|u^k\|_{\mathcal{M}} \max_{i=1,\dots,N} \max_{x \in \operatorname{supp} u_i^k} |x - x_i|_{\mathbb{R}^d}).$$

Since $\|u^k\|_{\mathcal{M}}$ is uniformly bounded, taking the supremum with respect to φ and combining the previous estimates yields the result. □

Remark 4.5. Combining the previous results we specifically conclude

$$\|u^k - u\|_{\mathcal{C}^{0,1*}} \leq c_{\|u\|_{\mathcal{M}},N} \max_{i=1,\dots,N} \left(\max_{x \in \operatorname{supp} u_i^k} |x - x_i|_{\mathbb{R}^d} + \max_{i=1,\dots,N} |\mathbf{u}_i - \|u_i^k\|_{\mathcal{M}}| \right).$$

In fact, a similar estimate from below is also valid. Since the support points $\{x_i\}_{i=1}^N$ of u are distinct there exists $R > 0$ such that $\bar{B}_R(x_i) \cap \bar{B}_R(x_j) = \emptyset$, $i \neq j$. Assume that $\operatorname{supp} u_i^k \subset \bar{B}_R(x_i)$ for all $k \in \mathbb{N}$ and $i = 1, \dots, N$. Let $k \in \mathbb{N}$ and $i \in \{1, \dots, N\}$ be arbitrary but fixed. By ξ_i we denote a smooth function with $\xi_i(x) = 1$ for all $x \in \bar{B}_R(x_i)$ and $\xi_i(x) = 0$ for all $x \in \bigcup_{j=1, j\neq i}^N \bar{B}_R(x_i)$ and assume that $u_i^k \neq \mathbf{u}_i \delta_{x_i}$. Then the function

$$\varphi_i^k(x) = \operatorname{sgn}(\mathbf{u}_i - \|u_i^k\|_{\mathcal{M}}) \xi_i(x) + |x - x_i| \xi_i(x),$$

is not equal to zero and Lipschitz-continuous with Lipschitz norm $\|\varphi_i^k\|_{\mathcal{C}^{0,1}}$ bounded independently of $k \in \mathbb{N}$ and i from above and below. Testing with $u - u_i^k$ we obtain

$$\begin{aligned} &\langle \varphi_i^k, u - u^k \rangle = \langle \varphi_i^k, \mathbf{u}_i \delta_{x_i} - u_i^k \rangle \\ &\quad = \sum_{x \in \operatorname{supp} u_i^k} [u_i^k(\{x\}) |x - x_i|_{\mathbb{R}^d}] + |\mathbf{u}_i - \|u_i^k\|_{\mathcal{M}}| \geq \|u_i^k\|_{\mathcal{M}} \min_{x \in \operatorname{supp} u_i^k} |x - x_i|_{\mathbb{R}^d} + |\mathbf{u}_i - \|u_i^k\|_{\mathcal{M}}|. \end{aligned}$$

Since i was chosen arbitrary we conclude

$$\|u^k - u\|_{\mathcal{C}^{0,1*}} \geq \max_{i=1,\dots,N} \langle \varphi_i^k / \|\varphi_i^k\|_{\mathcal{C}^{0,1}}, u^k - u \rangle \geq c \max_{i=1,\dots,N} [\min_{x \in \operatorname{supp} u_i^k} |x - x_i|_{\mathbb{R}^d} + |\mathbf{u}_i - \|u_i^k\|_{\mathcal{M}}|],$$

for some constant $c \leq 1/\|\varphi_i^k\|_{\mathcal{C}^{0,1}}$ for all $k \in \mathbb{N}$ and $i = 1, \dots, N$.

We apply these results to the sequence $\{u^k\}_{k \in \mathbb{N}}$ generated by Algorithm 3.

Theorem 4.21. *Denote by $\{u^k\}_{k \in \mathbb{N}}$ the sequence generated by Algorithm 3 and let the assumptions of Theorem 4.18 hold. Then there holds $u^k = \sum_{i=1}^N u_i^k$ with $\operatorname{supp} u_i^k \subset \bar{B}_R(\bar{x}_i)$, $u_i^k \neq 0$, for all $k \in \mathbb{N}$ large enough and all $i = 1, \dots, N$. Furthermore there exists $1 > \zeta > 0$ and a constant $c_{\|\bar{u}_\beta\|_{\mathcal{M}},N}$ depending on the norm of $\bar{u}_\beta$ and its support size N with*

$$\bar{W}_1(u^k, \bar{u}_\beta) + \max_{i=1,\dots,N} \max_{x \in \operatorname{supp} u_i^k} |x - \bar{x}_i|_{\mathbb{R}^d} + \max_{i=1,\dots,N} |\mathbf{u}_i - \|u_i^k\|_{\mathcal{M}}| \leq c_{\|\bar{u}_\beta\|_{\mathcal{M}},N} \zeta^k$$

for all $k \in \mathbb{N}$ large enough.

Proof. The statement on the linear convergence of the support points and the coefficients again follows by applying Theorem 6.70 with noting that $u^k(\bar{B}_R(\bar{x}_i)) = \|u_i^k\|_{\mathcal{M}}$ due to $u^k \in \mathcal{M}^+(\Omega_o)$. The convergence result for the modified Wasserstein distance $\bar{W}_1$ then follows from the estimate in Theorem 4.20. □

Remark 4.6. Similar to the previous section, all convergence results remain valid if the intermediate active set $\mathcal{A}_k$ is more generally chosen such that

$$\operatorname{supp} u^k \cup \{\hat{x}^k\} \subset \mathcal{A}_k, \quad \#\mathcal{A}_k < \infty,$$

i.e. more than one new sensor can be added in each iteration to further improve the convergence behavior.

4.4.4 Computational cost of the sparsification steps

It remains to comment on the computational cost associated with the various point removal steps presented in this section. First, we address the costs for the point removal steps based on the approximate solution of the finite dimensional subproblems. Computing the new coefficient vector $\mathbf{u}^k$ from (4.29) requires the computation of the pointwise evaluation of $\nabla\psi(u^{k+1/2})$ at the current support points once. In our numerical experiments a suitable step size σ_k is found by a simple backtracking line search to avoid ascend. Consequently, for each trial step size, the max-operator in (4.29) as well as the objective function is evaluated once. This can be done efficiently with cost scaling linearly with the current support size m_k.

Secondly, if $\mathbf{u}^k$ is determined from (4.30), we have to solve a finite-dimensional convex optimization problem in every iteration. Since the most common choices for the optimal design criterion Ψ are twice continuously differentiable, we choose to implement a semi-smooth Newton method. To benefit from the fast local convergence behavior for this class of methods we warm-start the algorithm using the coefficient vector $\mathbf{u}^{k+1/2}$ of the intermediate iterate $u^{k+1/2}$. This choice of the starting point often gives a good initial guess for $\mathbf{u}^{k+1}$. However, we emphasize that essentially any algorithm for smooth convex problems with positivity constraints on the optimization variables can be employed instead. In particular, interior point methods provide complexity bounds for the solution up to machine precision in terms of the support size m_k; see, e.g., [46, Section 11.5]. In light of this fact, the computational cost for the point removal steps can be regarded as a constant, assuming that m_k is uniformly bounded through the iterations, e.g., by employing Algorithm 1. However, interior point methods cannot be warm-started in general, which is why we prefer semi-smooth Newton methods in practice.

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

$$[\boldsymbol{I}(\omega)]_{j,i} = \boldsymbol{I}(\delta_{x_i})_j, \quad i = 1, \dots, N,\ j = 1, \dots, n(n+1)/2.$$

This can be done efficiently employing either a SVD-decomposition or a rank-revealing QR-decomposition. Furthermore, assuming that Algorithm 1 is applied to $u^{k+1/2}$ for every k, this loop will run at most once in each iteration. This can be seen in the following way: Let the

k-th iterate u^k in Algorithm 2 be given such that $\operatorname{rank} \boldsymbol{I}(u^k) = \# \operatorname{supp} u^k$. Note that this implies $\# \operatorname{supp} u^k \leq n(n+1)/2$. Consequently we have either $\operatorname{rank} \boldsymbol{I}(u^{k+1/2}) = \# \operatorname{supp} u^{k+1/2}$ or $\operatorname{rank} \boldsymbol{I}(u^{k+1/2}) = \# \operatorname{supp} u^{k+1/2} - 1$. In the first case no sparsification by the post-processing can be achieved. In the second case $u^{k+1/2} = \sum_{j=1}^m \mathbf{u}_j \delta_{x_j}$ is at least sparsified once. After the first execution of the sparsification loop, we obtain the measure $u_{\text{new}} = \sum_{\{\, i \mid \mathbf{u}_{\text{new},i} > 0 \,\}} \mathbf{u}_{\text{new},i} \delta_{x_i}$ with $\operatorname{rank} \boldsymbol{I}(u_{\text{new}}) = \# \operatorname{supp} u_{\text{new}}$, i.e. Algorithm 1 terminates.

4.4.5 A comparison to the Fedorov-Wynn Algorithm

To close the discussion on sequential point insertion algorithms we briefly describe an analogous approach for the solution of design problems with equality constraints

$$\min_{u \in \mathcal{M}^+(\Omega_o)} \Psi(\mathcal{I}(u) + \mathcal{I}_0) \quad s.t. \quad \|u\|_{\mathcal{M}} = K. \tag{4.32}$$

as they appear in the theory of approximate designs due to Kiefer and Wolfowitz. Due to the monotonicity of Ψ every optimal design obtained through (4.32) is also a minimizer of the inequality constrained problem

$$\min_{u \in \mathcal{M}^+(\Omega_o)} \Psi(\mathcal{I}(u) + \mathcal{I}_0) \quad s.t. \quad \|u\|_{\mathcal{M}} \leq K. \tag{4.33}$$

In the case of strict monotonicity the solution sets of both problems coincide, see Proposition 4.8. Note that (4.33) can be equivalently rewritten as a composite minimization problem

$$\min_{u \in \mathcal{M}(\Omega_o)} [\psi(u) + I_{u \geq 0}(u) + I_{\|u\|_{\mathcal{M}} \leq K}(u)].$$

Hence, to find a minimizer of we apply a (generalized) conditional gradient method. The procedure is described in Algorithm 4. Again, the new sensor v^k is found as a solution of the linearized problem

$$\min_{v \in \mathcal{M}^+(\Omega_o)} \langle \nabla \psi_k, v \rangle \quad s.t. \quad \|v\|_{\mathcal{M}} \leq K,$$

and convergence is monitored by evaluation of the primal-dual gap

$$\Phi(u^k) = \min_{v \in \mathcal{M}^+(\Omega_o)} \langle \nabla \psi_k, u^k - v \rangle = \langle \nabla \psi_k, u^k - v^k \rangle = \langle \nabla \psi_k, u^k \rangle - K \min_{x \in \Omega_o} \nabla \psi_k.$$

By a closer inspection, in the case $K = 1$, the resulting algorithm resembles the, at least among statisticians, well-known Fedorov-Wynn algorithm, see [76, 105, 272], which is one of the fundamental pillars of approximate design theory. Its properties are a well-studied subject, albeit in

Algorithm 4 Fedorov-Wynn algorithm for (4.33)

```
1. Choose $u^1 \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi$, $\# \operatorname{supp} u^1 \leq n(n+1)/2$, $\|u^1\|_{\mathcal{M}} = K$.
while $\Phi(u^k) \geq \mathrm{TOL}$ do
    2. Compute $\nabla \psi_k = \nabla \psi(u^k)$. Determine $\hat{x}^k \in \arg\min_{x \in \Omega_o} \nabla \psi_k(x)$.
    3. Set $v^k = K \delta_{\hat{x}^k}$.
    4. Select a step size $s^k \in (0, 1]$ and set $u^{k+1/2} = (1 - s^k) u^k + s^k v^k$.
    5. Find $u^{k+1}$ with $\operatorname{supp} u^{k+1} \subseteq \operatorname{supp} u^{k+1/2}$ and $F(u^{k+1}) \leq F(u^{k+1/2})$, $\|u^{k+1}\|_{\mathcal{M}} \leq M_0$.
end while
```

most cases only the D-optimal design criterion is considered. In the same fashion as for the norm-penalized problem (P_β) however, the general theory presented in Chapter 6 implies the following worst-case convergence rates for Algorithm 4 and a general optimal design criterion Ψ. Again we consider stepsizes $s^k = \gamma^{n_k}$, where $n_k \in \mathbb{N}$ is the smallest integer fulfilling

$$\alpha s^k \Phi(u^k) \leq \psi(u^k) - \psi(u^k_{s^k}), \tag{4.34}$$

for some fixed $\alpha \in (0, 1/2)$, $u^k_{s^k} = u^k - s^k(v^k - u^k)$.

Proposition 4.22. *Assume that the sequence $\{u^k\}_{k\in\mathbb{N}}$ is generated using Algorithm 4 with s^k chosen according to* (4.34)*. Then there exists at least one weak* accumulation point $\bar{u}$ of $\{u^k\}_{k\in\mathbb{N}}$ and every such point is a minimizer of* (4.33)*. Furthermore, defining $r(u) = \psi(u) - \psi(\bar{u})$ there holds*

$$r(u^k) \leq \frac{r(u^1)}{1 + q(k-1)}, \quad q = \alpha \min\left\{ \frac{c_1}{L_{u^1}(2K)^2},\ 1 \right\}.$$

Here, L_{u^1} is the Lipschitz-constant of $\nabla\psi$ on the sublevel set

$$E_{\psi,K}(u^1) = \{\, u \in \mathcal{M}^+(\Omega_o) \mid \psi(u) \leq \psi(u^1),\ \|u\|_{\mathcal{M}} \leq K \,\},$$

and $c_1 = 2\gamma(1-\alpha)r(u^1)$. Moreover, the algorithm can be implemented such that $\#\operatorname{supp} u^k \leq n(n+1)/2$ holds for all $k \in \mathbb{N}$ and all accumulation points $\bar{u}$.

Throughout the years, numerous modifications of Fedorov's and Wynn's original algorithm were made to enhance the sparsity of the iterates and to improve its convergence behavior. We only name a few here. For example, in [132, 216] the authors provide inequalities which have to be fulfilled by the gradient of the optimal design criterion for an arbitrary probability measure evaluated at optimal support points. Thus, non suitable candidate locations can be identified in each iteration and left out of the problem. Heuristically, sensors at old support points could be moved to a newly added one if they are sufficiently close, see [106, 256]. More recently, Yu, see [276], proposed to couple point insertion steps with moving mass between adjacent sensors according to a nearest neighborhood exchange method, see [41]. However, to our best knowledge, we are not aware of any modifications guaranteeing a uniform bound on the number of support points as done by the method proposed in Algorithm 1. Shortly after the initial papers, Atwood, c.f. [11], proposed to augment Fedorov's algorithm by Wolfe's away steps, [268]. Instead of adding a new sensor, an away step removes measurement weights from non-optimal support points and distributes it among more promising ones, [247], similarly to the projected gradient update described in (4.29).

Given an ordered set of distinct points $\mathcal{A}$ we recall the definition of the parametrization $\boldsymbol{u}_\mathcal{A}$ from (4.27). Early on, Wu, [269, 270], followed by several authors, [44, 275, 276], proposed to alternate between adding a new sensor at $\hat{x}^k$ to the iterated design u^k in Fedorov's algorithm and updating the measurement weights by (approximately) solving

$$\min_{\mathbf{u}\in\mathbb{R}^{m_k}_+, \|\mathbf{u}\|_{l_1}\leq K} \psi(\boldsymbol{u}_\mathcal{A}(\mathbf{u})), \quad \mathcal{A}_k = \operatorname{supp} u^k \cup \{\hat{x}^k\},\ m_k = \#\mathcal{A}_k, \tag{4.35}$$

by e.g. a Newton-like algorithm. One realization of the proposed procedure is summarized in Algorithm 5. Again, since the subproblems are solved up to optimality, no convex combination

Algorithm 5 Accelerated Fedorov-Wynn for (4.33)

```
while $\Phi(u^k) \geq \mathrm{TOL}$ do
    1. Calculate $\nabla\psi_k = \nabla\psi(u^k)$. Determine $\hat{x}^k \in \arg\min_{x\in\Omega_o} \nabla\psi_k(x)$.
    2. Set $\mathcal{A}_k = \operatorname{supp} u^k \cup \{\, \hat{x}^k \,\}$, compute a solution $\mathbf{u}^{k+1}$ of (4.35) with $\|\mathbf{u}^{k+1}\|_{l_1} = K$, and set
    $u^{k+1} = \boldsymbol{u}_{\mathcal{A}}(\mathbf{u}^{k+1})$.
end while
```

has to be formed to ensure admissible iterates and convergence of the procedure. Due to the monotonicity of Ψ we can furthermore guarantee $\|u^k\|_{\mathcal{M}} = K$ for all $k \in \mathbb{N}$. Thus we obtain

$$-\langle \nabla\psi_k, u^k\rangle = -K \min_{x\in\operatorname{supp} u^k} \nabla\psi_k(x), \quad \Phi(u^k) = K\big(\min_{x\in\operatorname{supp} u^k} \nabla\psi_k(x) - \min_{x\in\Omega_o} \nabla\psi_k(x)\big).$$

Since $\|u^k\|_{\mathcal{M}} = K$ for all $k \in \mathbb{N}$, this method is reminiscent of a simplicial decomposition method, see [204,260] and [138], and of the accelerated method proposed in Algorithm 3. Indeed, imposing similar second order conditions, a linear rate of convergence for the objective function value can be shown in the same way as for the norm regularized problem. To our best knowledge, no comparable results for the Fedorov-Wynn algorithm have been achieved so far in this direction.

Theorem 4.23. *Let Ψ be strictly convex on its domain and uniformly convex around $\mathcal{I}(\bar{u}) + \mathcal{I}_0$. Define $\bar{p} = -\partial S[\hat{q}]^\top \nabla\Psi(\mathcal{I}(\bar{u}) + \mathcal{I}_0)\partial S[\hat{q}]$ and assume that $\max_{x\in\Omega_o}\bar{p}(x) > 0$ with*

$$\left\{ x \in \Omega_o \mid \bar{p}(x) = \max_{x\in\Omega_o} \bar{p}(x) \right\} = \{\bar{x}_i\}_{i=1}^N \subset \operatorname{int}\Omega_o.$$

Assume that the set $\{\mathcal{I}(\delta_{\bar{x}_i})\}_{i=1}^N$ is linearly independent and let the unique optimal design be denoted by $\bar{u} = \sum_{i=1}^N \bar{\mathbf{u}}_i \delta_{\bar{x}_i}$ for some $\bar{\mathbf{u}}_i > 0$, $i = 1, \dots, N$. Let there be $R > 0$ with

$$\Omega_R = \bigcup_{i=1}^N B_R(\bar{x}_i), \quad \bar{B}_R(\bar{x}_i) \cap \bar{B}_R(\bar{x}_j) = \emptyset, \quad \partial S[\hat{q}] \in \mathcal{C}^2(\Omega_R, \mathbb{R}^n),$$

for all $i, j \in \{1, \dots, N\}$, $i \neq j$. Furthermore, the Hessian of $\bar{p}$ at the support points is supposed to satisfy

$$-(\zeta, \nabla^2\bar{p}(\bar{x}_i)\zeta)_{\mathbb{R}^d} \geq \theta|\zeta|^2_{\mathbb{R}^d}, \quad \forall \zeta \in \mathbb{R}^d,$$

for all $i = 1, \dots, N$ and some $\theta > 0$.

Finally, denote by $\{u^k\}_{k\in\mathbb{N}} \subset \mathcal{M}^+(\Omega_o)$ the sequence generated by Algorithm 5. Then we have $u^k \rightharpoonup^* \bar{u}$ and there exists $0 < \zeta < 1$ and a constant $c > 0$ with

$$\psi(u^k) - \psi(\bar{u}) \leq c\zeta^k,$$

for all $k \in \mathbb{N}$ large enough.

Proof. This result follows again from Theorem 6.70. □

Analogously, convergence rates for the optimal design measure as in Theorem 4.21 can be deduced. For brevity, we resign from stating them here again.

4.4.6 Algorithmic solution by path-following

As an alternative to Algorithm 2, we briefly describe a path-following approach. For simplicity assume that $\varPsi$ is nonnegative on its domain. To compute a minimizer of (P_β) we solve a sequence of regularized problems given by:

$$\min_{u \in L^2(\Omega_o), u \geq 0} F_\varepsilon(u) := [\psi(u) + \beta \|u\|_{L^1(\Omega_o)} + \frac{\varepsilon}{2} \|u\|^2_{L^2(\Omega_o)}]. \tag{P^ε_β}$$

In the limiting case for $\varepsilon \to 0$, the regularized optimal solutions approximate solutions of (P_β). We state first order optimality conditions for solutions of the regularized problem and investigate the case $\varepsilon \to 0$. For the sake of brevity, we omit most proofs.

Proposition 4.24. *Let the assumptions of Proposition 4.2 be fulfilled. Then the following statements hold:*

1. *For every $\varepsilon > 0$ there exists a unique solution $\bar{u}^\varepsilon_\beta \in L^2(\Omega_o)$ to (P^ε_β).*
2. *A non-negative function $\bar{u}^\varepsilon_\beta \in L^2(\Omega_o)$ is optimal if and only if*

$$\bar{u}^\varepsilon_\beta = \max\left\{ -\frac{1}{\varepsilon}(\nabla\psi(\bar{u}^\varepsilon_\beta) + \beta), 0 \right\}. \tag{4.36}$$

   Consequently there holds $\bar{u}^\varepsilon_\beta \in \mathcal{C}(\Omega_o)$ and

$$\bar{u}^\varepsilon_\beta(x) > 0 \text{ if and only if } -\nabla\psi(\bar{u}^\varepsilon_\beta)(x) > \beta.$$

3. *Given any sequence $\{\varepsilon_k\}_{k\in\mathbb{N}}$ with $\varepsilon_k > 0$, $\varepsilon_k \to 0$, the associated sequence $\bar{u}^{\varepsilon_k}_\beta$ admits at least one weak* accumulation point and every such point is an optimal solution to (P_β).*

Proof. By assumption there exists $u \in \mathcal{M}^+(\Omega_o)$ with $\psi(u) < \infty$, i.e. $\mathcal{I}(u) + \mathcal{I}_0 \in \mathrm{PD}(n)$. Following [208, Appendix A.1], there exists a sequence $\{u_k\}_{k\in\mathbb{N}} \subset L^2(\Omega_o)$ with $u_k \geq 0$ and $u_k \rightharpoonup^* u$. Consequently there also holds

$$\mathcal{I}(u_k) + \mathcal{I}_0 \to \mathcal{I}(u) + \mathcal{I}_0 \in \mathrm{PD}(n),$$

due to the weak*-to-strong continuity of $\mathcal{I}$. Thus we observe $\mathcal{I}(u) + \mathcal{I}_0 \in \mathrm{PD}(n)$ and $\psi(u_k) < \infty$ for all k large enough. The existence of at least one optimal solution $\bar{u}^\varepsilon_\beta$ now follows by similar arguments as in Proposition 4.2. Its uniqueness follows due to the strict convexity of F_ε. The necessary and sufficient optimality condition can be derived as in [248] and [61]. For the last result we observe that given an arbitrary positive null sequence $\{\varepsilon_k\}_{k\in\mathbb{N}}$ there holds

$$\beta\|\bar{u}^\varepsilon_\beta\|_{L^1(\Omega_o)} \leq F_{\varepsilon_k}(\bar{u}^{\varepsilon_k}_\beta) \leq F(u) + \frac{1}{2}\|u\|^2_{L^2(\Omega_o)},$$

for an arbitrary but fixed $u \in \mathrm{dom}_{\mathcal{M}^+(\Omega_o)} \psi \cap L^2(\Omega_o)$ and all k large enough. Following the lines of the proof in [208, Section 2.5] existence of at least one weak* accumulation point of $\bar{u}^{\varepsilon_k}_\beta$ as well as its optimality for (P_β) can now easily be deduced. □

Note that for fixed $\varepsilon > 0$ the unique minimizer $\bar{u}^{\varepsilon}_{\beta}$ is a solution of

$$u - \max\left\{-\frac{1}{\varepsilon}(\nabla\psi(u)(x) + \beta), 0\right\} = 0. \tag{4.37}$$

Under additional regularity assumptions on the optimal design criterion Ψ, the solution of this non-smooth operator equation can be computed by a semi-smooth Newton method in function space; see, e.g., [257]. To compute a solution for the original problem (P_β) we employ a continuation strategy for the regularization parameter ε. For an initial small value ε we compute the unique minimizer $\bar{u}^{\varepsilon}_{\beta}$ to $(P^{\varepsilon}_{\beta})$ by solving (4.37). Then, in an outer loop, we decrease ε, and use the previous optimal solution as an initial guess for the next iteration. The procedure is summarized in Algorithm 6. For further references on path-following we refer to, e.g., [144–146]. We briefly address that Algorithm 6, in contrast to the post-processed version of Algorithm 2, might fail to approximate any sparse minimizer of (P_β), see also [22, Theorem 26.20].

Algorithm 6 Path-following

```
1. Choose $\varepsilon_1 > 0$ and initial guess $u^1_\varepsilon \in \mathrm{dom}_{\mathcal{M}^+(\Omega_o)}\,\psi \cap L^2(\Omega_o)$.
while residual (4.37) large do
    2. Compute $\bar{u}^{\varepsilon_l}_{\beta}$ from (4.37) using $\bar{u}^{\varepsilon_{l-1}}_{\beta}$ as initial guess.
    3. Get $\varepsilon_{l+1} < \varepsilon_l$, $l = l + 1$.
end while
```

Proposition 4.25. *Assume that (P_β) admits an optimal solution $\bar{u}_\beta \in L^2(\Omega_o)$. Then there holds $\bar{u}^{\varepsilon}_{\beta} \to \bar{u}$, where $\bar{u} \in L^2(\Omega_o)$ is the unique optimal solution to*

$$\min_{u \in L^2(\Omega_o)} \|u\|^2_{L^2(\Omega_o)} \quad s.t. \quad u \in L^2(\Omega_o) \cap \operatorname*{arg\,min}_{v \in \mathcal{M}^+(\Omega_o)} F(v). \tag{4.38}$$

Proof. By assumption, the admissible set in (4.38) is not empty, convex and weak* compact. Since the norm $\|\cdot\|_{L^2(\Omega_o)}$ is strictly convex, (4.38) admits a unique solution $\bar{u}$. Let $\hat{u} \in L^2(\Omega_o) \cap \arg\min_{v \in \mathcal{M}^+(\Omega_o)} F(v)$ be arbitrary. From the optimality of $\bar{u}^{\varepsilon}_{\beta}$ we conclude

$$F_\varepsilon(\bar{u}^{\varepsilon}_{\beta}) \le F(\hat{u}) + \frac{\varepsilon}{2}\|\hat{u}\|^2_{L^2(\Omega_o)} \le F(\bar{u}^{\varepsilon}_{\beta}) + \frac{\varepsilon}{2}\|\hat{u}\|^2_{L^2(\Omega_o)}.$$

We conclude $\|\bar{u}^{\varepsilon}_{\beta}\|_{L^2(\Omega_o)} \le \|\hat{u}\|_{L^2(\Omega_o)}$ for every $\varepsilon > 0$ small enough and all $\hat{u} \in L^2(\Omega_o) \cap \arg\min_{u \in \mathcal{M}^+(\Omega_o)} F(u)$. Since $\bar{u}^{\varepsilon}_{\beta}$ is bounded in $L^2(\Omega_o)$ we extract a subsequence denoted by the same symbol which converges weakly to $\tilde{u} \in L^2(\Omega_o)$. This implies $\bar{u}^{\varepsilon}_{\beta} \rightharpoonup^* \tilde{u}$, i.e. $\tilde{u}$ is an optimal solution to (P_β), as well as $\|\tilde{u}\|_{L^2(\Omega_o)} \le \|\hat{u}\|_{L^2(\Omega_o)}$. Consequently $\tilde{u} = \bar{u}$ for every accumulation $\tilde{u}$ of $\bar{u}^{\varepsilon}_{\beta}$. The strong convergence follows from the weak convergence and $\|\bar{u}^{\varepsilon}_{\beta}\|_{L^2(\Omega_o)} \to \|\bar{u}\|_{L^2(\Omega_o)}$. □

To end this section, we provide a simple example to illustrate the findings of the previous proposition.

Example 4.2. *We consider $n = 1$, $\mathcal{I}_0 = 0$ and $\partial S[\hat{q}]\delta q = \delta q g$ for some $g \in \mathcal{C}(\Omega_o)$, $g \neq 0$ and all $\delta q \in \mathbb{R}$. In this case, the A-optimal design problem is given by*

$$\min_{u \in \mathcal{M}^+(\Omega_o)} \frac{1}{\langle g^2, u\rangle} + \beta\|u\|_{\mathcal{M}}. \tag{4.39}$$

From the necessary and sufficient first order optimality conditions it is easily deduced that a given $\bar{u}_\beta \in \mathcal{M}^+(\Omega_o)$ *is an optimal design if and only if*

$$\|\bar{u}_\beta\|_{\mathcal{M}} = \sqrt{\frac{1}{\|g\|_{\mathcal{C}}^2 \beta}}, \quad \operatorname{supp} \bar{u}_\beta \subset \{\, x \in \Omega_o \mid |g(x)| = \|g\|_{\mathcal{C}} \,\}.$$

As a consequence, if $\{\, x \in \Omega_o \mid |g(x)| = \|g\|_{\mathcal{C}} \,\}$ *has non-zero Lebesgue measure, Problem* (4.39) *admits both, solutions consisting of exactly one Dirac delta function as well as solutions in* $L^2(\Omega_o)$*. Following the previous Proposition, applying the path-following approach in this situation leads to an optimal design* $\bar{u}_\beta \in L^2(\Omega_o)$*, while Algorithm 3 gives a minimizer consisting of a single Dirac delta function in one iteration.*

4.5 Stability and sensitivity analysis

In this section we will elaborate on the stability and sensitivity of optimal measurement design under perturbations of the sensor placement problem. To set the stage, we consider

$$\min_{u \in \mathcal{M}^+(\Omega_o)} F_\Delta(u) = \Psi(\mathcal{I}[\Delta_1](u) + \mathcal{I}_0[\Delta_2]) + (\beta_0 + \Delta_3)\|u\|_{\mathcal{M}}, \tag{P_Δ}$$

where the cost parameter, the a priori knowledge as well as the Fisher information operator are subject to a triple of perturbations $\Delta = (\Delta_1, \Delta_2, \Delta_3)$. In the following we will explore and quantify how this bias in the data of the sensor placement problem influences the positions and measurement weights of optimally placed sensors.

These questions naturally arise in the context of optimal sensor placement. Recall for example that all previous considerations are based on a first order approximation of the parameter-to-state mapping S. Ideally the linearization point should be chosen as the true parameter q^*. This is clearly impossible in practice since these quantities are unknown. For this reason one has to resort to well-educated a priori guesses $\hat{q}$ stemming e.g. from previous experiments. Since the Fisher-operator $\mathcal{I}$ depends on $\hat{q}$ we can interpret the difference between the true parameter q^* and $\hat{q}$ as a perturbation of the problem. Other perturbations of the Fisher information operator may be induced by a low-rank approximation of the, possibly high-dimensional, parameter itself or the parameter-to-state mapping S, see e.g. [6, 30, 137]. Moreover, in cases in which the number of measurements should be kept small, sensitivities of the optimal design and the optimal design criterion with respect to perturbations in the cost parameter might help to identify less important sensors. Finally, if several perturbed sensor placement problems have to be solved sequentially, sensitivity results on the optimal design may be used to obtain a good initial iterate for the algorithmic procedure presented in Section 4.7.1.

Our main contributions are the following: Under mild assumptions we show that the support of perturbed optimal designs is localized in the vicinity of the unperturbed optimal design points. For vanishing perturbations, convergence results are presented. If, in addition, the unique optimal design measure consists of finitely many Dirac delta functions and the curvature of the optimal gradient does not degenerate in the vicinity of their positions we can prove additional results. In this case the positions as well as the measurement weights of the perturbed optimal sensors depend continuously differentiable on Δ, at least in the asymptotic regime. This allows for a Taylor expansion of these quantities as well as the objective function with respect to the perturbation. In particular, this implies Lipschitz-stability of the optimal design measure in the modified Wasserstein distance introduced in Definition 4.1.

Let us discuss related work in this direction. To take perturbations of the model into account, robust optimal design approaches based on min-max or average formulations were developed, see [37, 218] and [217, Chapter 8]. We stress that in general no stability of the optimal design measure with respect to the norm on $\mathcal{M}(\Omega_o)$ can be expected, c.f. also [43, Remark 5.120]. Hence, the abstract sensitivity results presented in [70] cannot be applied directly in this situation. Again, we recall the dual relation between the sparse sensor placement problem (P_β) and the semi-infinite problem (4.20). Stability analysis of semi-infinite optimization problems is a well studied subject, see e.g. [167,238,240]. The techniques presented in this section are closest related to the so called local reduction ansatz, see [141, 142]. By imposing a suitable second order condition, we will show that the perturbed optimal design problem (P_Δ) can be reduced to a finite-dimensional optimization problem for small perturbations Δ. Sensitivity results for the optimal positions as well as the measurement weights are then obtained by applying perturbation theory for finite dimensional nonlinear optimization problems. A similar route was taken in [95] in the context of sparse deconvolution problems. While stability with respect to the canonical norm on $\mathcal{M}(\Omega_o)$ cannot be expected, the sensitivity results obtained through the reduction ansatz imply stability with respect to the modified Wasserstein distance, see Definition 4.1. In [239] a similar idea was pursued, deriving stability results for the minimizer in a semi-infinite program by embedding the space of Borel measures $\mathcal{M}^+(\Omega_o)$ into the dual space of the Lipschitz continuous functions on Ω_o. For completion, we also mention the works of [63,64] which discuss stability in the context of sparse control of non-linear partial differential equations by embedding $\mathcal{M}(\Omega_o)$ into Sobolev spaces of negative order. However, no stability results for the optimal control beyond weak* convergence are given. Finally, to the best of our knowledge, we are not aware of any results concerning stability and sensitivity analysis in the context of approximate designs.

4.5.1 The perturbed optimal design problem

We start by collecting the general assumptions on the family of perturbed optimal sensor placement problems. We assume that the interior int Ω_o of the observational domain $\Omega_o \subset \bar{\Omega}$ is non-empty. Furthermore let V_1, V_2 denote separable Banach spaces and set $V = V_1 \times V_2 \times \mathbb{R}$. The norm on V is denoted by $\|\cdot\|_V = \|\cdot\|_{V_1} + \|\cdot\|_{V_2} + |\cdot|$ and the perturbations are assembled in a vector $\Delta = (\Delta_1, \Delta_2, \Delta_3)$. To take perturbations of the Fisher information into account we consider a parametrization of the $i-$th sensitivity $\partial_i S[\hat{q}]$ given by

$$\partial_i S[\hat{q}] \colon V_1 \times \Omega_o \to \mathbb{R} \quad (\Delta_1, x) \mapsto \partial_i S[\Delta_1][\hat{q}](x). \tag{4.40}$$

To improve readability we will drop the dependence of the corresponding sensitivity vector on the linearization point $\hat{q}$ in the following and denote $\partial_i S[\Delta_1](x) := \partial_i S[\Delta_1][\hat{q}](x)$ for all $i = 1, \dots, n$ and $x \in \Omega_o$. The associated perturbation mappings for the sensitivity vector and the pointwise Fisher information are denoted by

$$\partial S \colon V_1 \times \Omega_o \to \mathbb{R}^n, \quad (\Delta_1, x) \mapsto \partial S[\Delta_1](x) := (\partial_1 S[\Delta_1](x), \dots, \partial_n S[\Delta_1](x))^\top,$$

as well as

$$I \colon V_1 \times \Omega_o \to \operatorname{Sym}(n), \quad (\Delta_1, x) \mapsto \partial S[\Delta_1](x) \partial S[\Delta_1](x)^\top,$$

respectively. The corresponding perturbed Fisher information operator is given by

$$\mathcal{I}[\varDelta_1](u) = \int_{\varOmega_o} \partial S[\varDelta_1](x) \partial S[\varDelta_1](x)^\top \mathrm{d}u(x) \quad \text{for } \varDelta_1 \in V_1,\ u \in \mathcal{M}(\varOmega_o). \tag{4.41}$$

As in Proposition 4.1 we readily verify that $\mathcal{I}[\varDelta_1](u) \in \mathrm{NND}(n)$ for every positive measure $u \in \mathcal{M}^+(\varOmega_o)$. Similarly, we define the perturbations of the a priori information matrix and the cost parameter as

$$\mathcal{I}_0\colon V_2 \to \mathrm{Sym}(n), \quad \varDelta_2 \mapsto \mathcal{I}_0[\varDelta_2], \quad \beta\colon \mathbb{R} \to \mathrm{Sym}(n), \quad \varDelta_3 \mapsto \beta[\varDelta_3] := \beta_0 + \varDelta_3,$$

where $\beta_0 > 0$ is a given reference cost parameter. Without loss of generality, the unperturbed sensor placement problem is recovered for $\varDelta = (0,0,0) \in V$. Furthermore, $c > 0$ will denote a generic constant which is independent of the perturbation $\varDelta$. We make the following regularity assumptions on the perturbations and the optimal design criterion $\varPsi$.

Assumption 4.6. Assume that there exists $u \in \mathcal{M}^+(\varOmega_o)$ with $\mathcal{I}[0](u) + \mathcal{I}_0[0] \in \mathrm{PD}(n)$ as well as a neighbourhood $N_V = N_{V_1} \times N_{V_2} \times N_{V_3}$ of $0 \in V$ such that:

A4.5 There holds $\partial S \in \mathcal{C}(N_{V_1} \times \varOmega_o, \mathbb{R}^n)$ and

$$\|\partial_i S[\varDelta_1] - \partial_i S[0]\|_{\mathcal{C}} \leq c \|\varDelta_1\|_{V_1} \quad \forall \varDelta_1 \in N_{V_1},$$

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

A4.6 There holds $\mathcal{I}_0 \in \mathcal{C}(N_{V_2}, \mathrm{Sym}(n))$ with $\mathcal{I}_0(\varDelta_2) \in \mathrm{NND}(n)$ for all $\varDelta_2 \in N_{V_2}$ as well as

$$\|\mathcal{I}_0[\varDelta_2] - \mathcal{I}_0[0]\|_{\mathrm{Sym}} \leq c \|\varDelta_2\|_{V_2}.$$

A4.7 For $\varDelta_3 \in N_{V_3}$ there holds $\beta[\varDelta_3] > c_0 > 0$ for a positive constant c_0.

Remark 4.7. We are especially interested in perturbations of the sensitivity vector caused by disturbances in the expansion point $\hat{q}$ of the first order approximation in the underlying model. Hence we consider $V_1 = \mathbb{R}^n$ and

$$\partial S\colon V_1 \to \mathcal{C}(\varOmega, \mathbb{R}^n), \quad \varDelta_1 \mapsto \partial S[\hat{q} + \varDelta_1].$$

In this case, the regularity assumptions on the mapping ∂S, see (**A4.5**), can be directly inferred from the continuous Fréchet differentiability of the parameter-to-state mapping $S\colon Q_{ad} \to \mathcal{C}(\varOmega_o)$. Higher order regularity of ∂S can be concluded similarly, imposing additional regularity assumptions on the parameter-to-state operator S.

For $\varDelta \in N_V$ we consider the reduced form of the (perturbed) optimal sensor placement problem

$$\min_{u \in \mathcal{M}^+(\varOmega_o)} F_\varDelta(u) = [\psi(u, \varDelta) + \|u\|_{\mathcal{M}}],$$

where the reduced design criterion $\psi(u, \varDelta)$ is given by

$$\psi(u, \varDelta) = \frac{1}{\beta_0 + \varDelta_3} \varPsi(\mathcal{I}[\varDelta_1](u) + \mathcal{I}_0[\varDelta_2]).$$

Note that we have incorporated the cost parameter in the smooth part of the objective function for now. We first provide a stability result for the Fisher information.

Lemma 4.26. *For all $\varDelta_1 \in N_{V_1}$ we have*

$$\max_{x\in\varOmega_o} \|I[\varDelta_1](x) - I[0](x)\|_{\mathrm{Sym}} + \|\mathcal{I}[\varDelta_1] - \mathcal{I}[0]\|_{\mathcal{L}(\mathcal{M}(\varOmega_o),\mathrm{Sym}(n))} \le c\|\varDelta_1\|_V.$$

Proof. For $x \in \varOmega_o$ we readily obtain

$$\begin{aligned}\|I[\varDelta_1](x) - I[0](x)\|^2_{\mathrm{Sym}} &= \mathrm{Tr}((I[\varDelta_1](x) - I[0](x))^\top (I[\varDelta_1](x) - I[0](x))) \\ &= \sum_{i=1}^n \sum_{j=1}^n (\partial_i S[\varDelta_1](x)\partial_j S[\varDelta_1](x) - \partial_i S[0](x)\partial_j S[0](x))^2.\end{aligned}$$

Now we estimate

$$\begin{aligned}&|\partial_i S[\varDelta_1](x)\partial_j S[\varDelta_1](x) - \partial_i S[0](x)\partial_j S[0](x)| \\ &\qquad \le \|\partial_i S[\varDelta_1]\|_{\mathcal{C}}\|\partial_j S[\varDelta_1] - \partial_j S[0]\|_{\mathcal{C}} + \|\partial_j S[0]\|_{\mathcal{C}}\|\partial_i S[\varDelta_1] - \partial_i S[0]\|_{\mathcal{C}},\end{aligned}$$

for all $i, j = 1, \dots n$. From Assumption (**A4.5**) we now conclude

$$|\partial_i S[\varDelta_1](x)\partial_j S[\varDelta_1](x) - \partial_i S[0](x)\partial_j S[0](x)| \le c(\|\partial_i S[\varDelta_1]\|_{\mathcal{C}} + \|\partial_j S[0]\|_{\mathcal{C}})\|\varDelta_1\|_{V_1},$$

as well as

$$\|\partial_i S[\varDelta_1]\|_{\mathcal{C}} \le \|\partial_i S[0]\|_{\mathcal{C}} + \|\partial_i S[\varDelta_1] - \partial_i S[0]\|_{\mathcal{C}} \le \|\partial_i S[0]\|_{\mathcal{C}} + c\|\varDelta_1\|_{V_1},$$

for all $i = 1, \dots, n$. Combining the previous results gives

$$\|I[\varDelta_1](x) - I[0](x)\|^2_{\mathrm{Sym}} \le c\|\varDelta_1\|^2_{V_1},$$

which yields the first statement after taking the square root on both sides and maximizing for x. Concerning the estimate for the operator norm of the Fisher information operator we observe that

$$\|\mathcal{I}[\varDelta_1](u) - \mathcal{I}[0](u)\|_{\mathrm{Sym}} \le c \max_{x\in\varOmega_o} \|I[\varDelta_1](x) - I[0](x)\|_{\mathrm{Sym}}\|u\|_{\mathcal{M}} \le c\|\varDelta_1\|_{V_1}\|u\|_{\mathcal{M}},$$

for all $u \in \mathcal{M}(\varOmega_o)$. Hence the second statement readily follows. □

By use of the triangle inequality we immediately derive the following perturbation result.

Corollary 4.27. *For all $\varDelta_1 \in N_{V_1}$ and all $u_1, u_2 \in \mathcal{M}^+(\varOmega_o)$ there holds*

$$\|\mathcal{I}[\varDelta_1](u_1) - \mathcal{I}[0](u_2)\|_{\mathrm{Sym}} \le c\|\varDelta_1\|_{V_1}\|u_1\|_{\mathcal{M}} + \|\mathcal{I}[0](u_1) - \mathcal{I}[0](u_2)\|_{\mathrm{Sym}},$$

where the constant $c > 0$ is independent of $u_1, u_2 \in \mathcal{M}^+(\varOmega_o)$.

Proof. Let $\varDelta_1 \in N_{V_1}$ and $u_1, u_2 \in \mathcal{M}^+(\varOmega_o)$ be given. We split the difference up as

$$\|\mathcal{I}[\varDelta_1](u_1) - \mathcal{I}[0](u_2)\|_{\mathrm{Sym}} \le \|\mathcal{I}[\varDelta_1] - \mathcal{I}[0]\|_{\mathcal{L}(\mathcal{M}(\varOmega_o),\mathrm{Sym}(n))}\|u_1\|_{\mathcal{M}} + \|\mathcal{I}[0](u_1) - \mathcal{I}[0](u_2)\|_{\mathrm{Sym}}.$$

Applying the estimate of the previous lemma yields the statement. □

Due to the continuity of the perturbation mappings the domain of $\psi(\cdot,\Delta)$ will be non-empty for small perturbations. In order to prove the existence of a perturbed optimal design we recall that there exists $M_0 > 0$ such that $\|\bar{u}_\beta\|_{\mathcal{M}} \leq M_0$. We consider the auxiliary problem

$$\min_{u \in \mathcal{M}^+(\Omega_o)} F_\Delta(u) \quad s.t. \quad \|u\|_{\mathcal{M}} \leq 2M_0. \tag{$P_\Delta^{M_0}$}$$

Proposition 4.28. *For all $\Delta \in N_V$ small enough there exists at least one optimal solution $\bar{u}_\Delta$ to $(P_\Delta^{M_0})$ with $\#\operatorname{supp}\bar{u}_\Delta \leq n(n+1)/2$. If Ψ is strictly convex then the optimal Fisher-information matrix $\mathcal{I}[\Delta_1](\bar{u}_\Delta)$ is unique.*

Proof. By assumption, there exists $u \in \mathcal{M}^+(\Omega_o)$, $\|u\|_{\mathcal{M}} \leq M_0$, with $\mathcal{I}[0] + \mathcal{I}_0[0] \in \mathrm{PD}(n)$. Hence $u \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi(\cdot,0)$. We estimate

$$\|\mathcal{I}[0](u) + \mathcal{I}_0[0] - \mathcal{I}[\Delta_1](u) - \mathcal{I}_0[\Delta_2]\|_{\mathrm{Sym}} \leq c(\|\Delta_1\|_{V_1}\|u\|_{\mathcal{M}} + \|\Delta_2\|_{V_2}),$$

for $\Delta_1 \in N_{V_1}$ and $\Delta_2 \in N_{V_2}$. Since $\mathrm{PD}(n)$ is open we conclude $\mathcal{I}[\Delta_1](u) + \mathcal{I}_0[\Delta_2] \in \mathrm{PD}(n)$ for all $\Delta \in N_V$ with $\|\Delta\|_V$ small enough and thus $u \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi(\cdot,\Delta)$. Thus the admissible set in $(P_\Delta^{M_0})$ is not empty. The existence of a minimizer to $(P_\Delta^{M_0})$ can now be concluded from the weak* compactness of the unit ball in $\mathcal{M}(\Omega_o)$ and the weak* lower semi-continuity of F_Δ on $\mathcal{M}^+(\Omega_o)$. The claim on the sparsity of the minimizer and the uniqueness of the Fisher information matrix follow as in Proposition 4.2 and Theorem 4.5. □

We proceed to prove the convergence of the design measures $\{\bar{u}_\Delta\}_{\Delta \in N_V}$ towards minimizers of (P_β) as well as the stability of the optimal objective function value in (P_Δ). Since the norm constraint in $(P_\Delta^{M_0})$ is inactive at unperturbed optimal designs we conclude the existence of minimizers to (P_Δ).

Proposition 4.29. *Consider a null sequence $\{\Delta_k\}_{k\in\mathbb{N}} \subset N_V$ with $\Delta_k = (\Delta_1^k, \Delta_2^k, \Delta_3^k)$. For each $k \in \mathbb{N}$ let $\bar{u}_{\Delta_k} \in \mathcal{M}^+(\Omega_o)$ denote a minimizer to $(P_{\Delta_k}^{M_0})$. Then there exists at least one weak* convergent subsequence of $\{\bar{u}_{\Delta_k}\}_{k\in\mathbb{N}}$ denoted by the same symbol with weak* limit $\bar{u}_0 \in \mathcal{M}^+(\Omega_o)$. There holds*

$$\psi(\bar{u}_{\Delta_k}, \Delta_k) \to \psi(\bar{u}_0, 0), \quad \|\bar{u}_{\Delta_k}\|_{\mathcal{M}} \to \|\bar{u}_0\|_{\mathcal{M}}, \quad \mathcal{I}[\Delta_2^k](\bar{u}_{\Delta_k}) \to \mathcal{I}[0](\bar{u}_0)$$

for $k \to \infty$. Consequently, $\bar{u}_{\Delta_k}$ is a minimizer of (P_{Δ_k}) for k large enough. Furthermore, every accumulation point of $\{\bar{u}_{\Delta_k}\}_{k\in\mathbb{N}}$ is a minimizer of $(\mathcal{P}_0)$. If there holds

$$\#\operatorname{supp}\bar{u}_{\Delta_k} \leq n(n+1)/2, \quad k \in \mathbb{N},$$

then the same holds for every accumulation point $\bar{u}_0$.

Proof. The sequence $\bar{u}_{\Delta_k}$ is uniformly bounded. Hence, by the Banach-Alaoglu theorem we can extract a subsequence denoted by the same symbol with $\bar{u}_{\Delta_k} \rightharpoonup^* \bar{u}_0$ for some $\bar{u}_0 \in \mathcal{M}^+(\Omega_o)$. Since the norm of $\bar{u}_{\Delta_k}$ is uniformly bounded we conclude

$$\|\mathcal{I}[\Delta_1^k](\bar{u}_{\Delta_k}) - \mathcal{I}[0](\bar{u}_0)\|_{\mathrm{Sym}} \leq c(\|\Delta_1\|_{V_1} + \|\mathcal{I}[0](\bar{u}_{\Delta_k} - \bar{u}_0)\|_{\mathrm{Sym}}),$$

see Lemma 4.26, and $\mathcal{I}[\Delta_1^k](\bar{u}_{\Delta_k}) \to \mathcal{I}[0](\bar{u}_0)$ in $\mathrm{Sym}(n)$. Moreover, due to Assumption 4.6 we obtain

$$\mathcal{I}[\Delta_k^1](\bar{u}_{\Delta_k}) + \mathcal{I}_0[\Delta_2^k] \to \mathcal{I}[0](\bar{u}_0) + \mathcal{I}_0[0] \in \mathrm{PD}(n),$$

and thus $\bar{u}_{\Delta_k} \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi(\cdot, 0)$ for k large enough. We note that

$$\|\bar{u}_{\Delta_k}\|_{\mathcal{M}} = \langle 1, \bar{u}_{\Delta_k} \rangle \to \|\bar{u}_0\|_{\mathcal{M}} = \langle 1, \bar{u}_0 \rangle = \|\bar{u}_0\|_{\mathcal{M}} \leq M_0.$$

As a consequence we have $\|\bar{u}_{\Delta_k}\|_{\mathcal{M}} < 2M_0$ for all k large enough. Since the norm constraint in $(P^{M_0}_{\Delta_k})$ is inactive at $\bar{u}_{\Delta_k}$ it is also a minimizer of the unconstrained problem (P_{Δ_k}). Let $\bar{u} \in \mathcal{M}^+(\Omega_o)$ denote a solution to $(\mathcal{P}_0)$. From the lower semi-continuity of F on $\operatorname{NND}(n)$ we deduce

$$\psi(\bar{u}_0, 0) + \|\bar{u}_0\|_{\mathcal{M}} \leq \liminf_{k \to \infty} [\psi(\bar{u}_{\Delta_k}, \Delta_k) + \|\bar{u}_{\Delta_k}\|_{\mathcal{M}}] \leq \psi(\bar{u}, 0) + \|\bar{u}\|_{\mathcal{M}}.$$

Therefore $\bar{u}_0$ is a minimizer of the unperturbed problem. It follows that

$$\psi(\bar{u}_{\Delta_k}, \Delta_k) = \frac{1}{\beta_0 + \Delta_3^k} \Psi(\mathcal{I}[\Delta_1^k](\bar{u}_{\Delta_k}) + \mathcal{I}_0[\Delta_2^k]) \to \frac{1}{\beta_0} \Psi(\mathcal{I}[0](\bar{u}_0) + \mathcal{I}_0[0]) = \psi(\bar{u}_0, 0),$$

Since $\bar{u}_0$ was arbitrary, the same can be shown for every accumulation point. The result on the number of support points follows due to Proposition 6.34. □

These results especially yield the existence of at least one sparse minimizer of (P_Δ) for $\Delta \in N_V$ if the set of perturbations is chosen small enough. Analogously to the previous section, we characterize a perturbed optimal design measure $\bar{u}_\Delta$ by a condition on the gradient of the optimal design criterion. To avoid distraction we will denote the gradient of $\psi(u, \Delta)$ with respect to u by $\nabla_u \psi(u, \Delta) \in \mathcal{C}(\Omega_o)$ in the following.

Corollary 4.30. *Let $\Delta \in N_V$ be given and let $\bar{u}_\Delta \in \mathcal{M}^+(\Omega_o)$ be an optimal solution to (P_Δ). Then there holds*

$$-\nabla_u \psi(\bar{u}_\Delta, \Delta) \leq 1, \quad \operatorname{supp} \bar{u}_\Delta \subset \{x \in \Omega_o \mid -\nabla_u \psi(\bar{u}_\Delta, \Delta)(x) = 1\}, \tag{4.42}$$

where $-\nabla_u \psi(\bar{u}_\Delta, \Delta) \in \mathcal{C}(\Omega_o)$ is given by

$$\begin{aligned} -\nabla_u \psi(\bar{u}_\Delta, \Delta)(x) &= -\frac{1}{\beta_0 + \Delta_3} \mathcal{I}[\Delta_1]^* \nabla \Psi(\mathcal{I}[\Delta_1](\bar{u}_\Delta) + \mathcal{I}_0[\Delta_2]) \\ &= -\frac{1}{\beta_0 + \Delta_3} \operatorname{Tr}(I[\Delta_1](x) \nabla \Psi(\mathcal{I}[\Delta_1](\bar{u}_\Delta) + \mathcal{I}_0[\Delta_2])) \\ &= -\frac{1}{\beta_0 + \Delta_3} \partial S[\Delta_1](x)^\top \nabla \Psi(\mathcal{I}[\Delta_1](\bar{u}_\Delta) + \mathcal{I}_0[\Delta_2]) \partial S[\Delta_1](x), \end{aligned}$$

for all $x \in \Omega_o$.

Due to the lipschitzian dependence on the perturbations we derive the following stability result for the design criterion.

Lemma 4.31. *Let a null sequence $\{\Delta_k\}_{k \in \mathbb{N}} \subset N_V$ and an associated sequence $u_{\Delta_k} \rightharpoonup^* u$ for some $u \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi(\cdot, \Delta_k) \cap \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi(\cdot, 0)$ and $u_{\Delta_k} \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi(\cdot, \Delta_k)$ for all $k \in \mathbb{N}$ be given. Then there holds*

$$|\psi(u_{\Delta_k}, \Delta_k) - \psi(u_{\Delta_k}, 0)| \leq c_u (|\psi(u_{\Delta_k}, 0)| |\Delta_3^k| + \|\Delta_1^k\|_{V_1} + \|\Delta_2^k\|_{V_2}), \tag{4.43}$$

where the constant $c_u > 0$ may depend on the weak limit $u \in \mathcal{M}^+(\Omega_o)$.*

Proof. Let such a sequence be given. We expand

$$\psi(u_{\Delta_k},\Delta_k)-\psi(u_{\Delta_k},0)=\frac{1}{\beta_0+\Delta_3^k}\Psi(\mathcal{I}[\Delta_1^k](u_{\Delta_k})+\mathcal{I}_0[\Delta_2^k])-\frac{1}{\beta_0}\Psi(\mathcal{I}[0](u_{\Delta_k})+\mathcal{I}_0[0]).$$

We start by estimating

$$\begin{aligned}&\left|\frac{1}{\beta_0+\Delta_3^k}\right||\Psi(\mathcal{I}[\Delta_1^k](u_{\Delta_k})+\mathcal{I}_0[\Delta_2^k])-\frac{1}{\beta_0}\Psi(\mathcal{I}[0](u_{\Delta_k})+\mathcal{I}_0[0])|\\ \leq&\left|\frac{\Delta_3^k}{\beta_0(\beta_0+\Delta_3^k)}\right||\Psi(\mathcal{I}[0](u_{\Delta_k})+\mathcal{I}_0[0])|+\frac{1}{\beta_0+\Delta_3^k}|\Psi(\mathcal{I}[\Delta_1](u_{\Delta_k})+\mathcal{I}_0[\Delta_2^k])-\Psi(\mathcal{I}[0](u_{\Delta_k})+\mathcal{I}_0[0])|.\end{aligned}$$

By Taylor's expansion we obtain

$$\begin{aligned}&\Psi(\mathcal{I}[\Delta_1^k](u_{\Delta_k})+\mathcal{I}_0[\Delta_2^k])-\Psi(\mathcal{I}[0](u_{\Delta_k})+\mathcal{I}_0[0])=\\ &\qquad\operatorname{Tr}(\nabla\Psi(\mathcal{I}_{\zeta_k}(u_{\Delta_k})+\mathcal{I}_{0,\zeta_k})^\top(\mathcal{I}[\Delta_1^k](u_{\Delta_k})+\mathcal{I}_0[\Delta_2^k]-\mathcal{I}[0](u_{\Delta_k})-\mathcal{I}_0[0])),\end{aligned}$$

where $\mathcal{I}_{\zeta_k}(u_{\Delta_k})+\mathcal{I}_{0,\zeta_k}=\mathcal{I}[0](u_{\Delta_k})+\mathcal{I}_0[0]+\zeta_k(\mathcal{I}[\Delta_1^k](u_{\Delta_k})+\mathcal{I}_0[\Delta_2^k]-\mathcal{I}[0](u_{\Delta_k})-\mathcal{I}_0[0])$ for some $\zeta_k\in(0,1)$. Since the Fisher information and $\mathcal{I}_0$ are stable with respect to the perturbation we conclude

$$\begin{aligned}&\|\mathcal{I}[0](u)+\mathcal{I}_0[0]-\mathcal{I}_{\zeta_k}(u_{\Delta_k})-\mathcal{I}_{0,\zeta_k}\|_{\mathrm{Sym}}\\ &\quad\leq\|\mathcal{I}[0](u)-\mathcal{I}[0](u_{\Delta_k})\|_{\mathrm{Sym}}+\|\mathcal{I}[\Delta_1^k](u_{\Delta_k})+\mathcal{I}_0[\Delta_2^k]-\mathcal{I}[0](u_{\Delta_k})-\mathcal{I}_0[0]\|_{\mathrm{Sym}}\to 0,\end{aligned}\tag{4.44}$$

due to $u_{\Delta_k}\rightharpoonup^* u$ and the stability of the mapping $\mathcal{I}[\cdot]$. Finally, we estimate

$$\begin{aligned}&|\Psi(\mathcal{I}[\Delta_1^k](u_{\Delta_k})+\mathcal{I}_0[\Delta_2^k])-\Psi(\mathcal{I}[0](u_{\Delta_k})+\mathcal{I}_0[0])|\\ &\quad\leq\|\nabla\Psi(\mathcal{I}_{\zeta_k}(u_{\Delta_k})+\mathcal{I}_{0,\zeta_k})\|_{\mathrm{Sym}}(\|\mathcal{I}[\Delta_1^k](u_{\Delta_k})-\mathcal{I}[0](u_{\Delta_k})\|_{\mathrm{Sym}}+\|\mathcal{I}_0[\Delta_2^k]-\mathcal{I}_0[0]\|_{\mathrm{Sym}})\\ &\quad\leq\|\nabla\Psi(\mathcal{I}_{\zeta_k}(u_{\Delta_k})+\mathcal{I}_{0,\zeta_k})\|_{\mathrm{Sym}}(\|\Delta_1^k\|_{V_1}\|u_{\Delta_k}\|_{\mathcal{M}}+\|\Delta_2^k\|_{V_2}).\end{aligned}$$

Since $\nabla\Psi\colon\operatorname{dom}\Psi\to\mathrm{Sym}(n)$ is continuous, we have

$$\|\nabla\Psi(\mathcal{I}_{\zeta_k}(u_{\Delta_k})+\mathcal{I}_{0,\zeta_k})\|_{\mathrm{Sym}}\leq c_u,$$

for all $k\in\mathbb{N}$ and some constant $c_u>0$. Combining all the previous results and noting that $\beta_0+\Delta_3^k>c_0$ yields

$$|\psi(u_{\Delta_k},\Delta_k)-\psi(u_{\Delta_k},0)|\leq\frac{1}{\beta_0c_0}|\Delta_3^k||\psi(u_{\Delta_k},0)|+c_u(\|\Delta_1^k\|_{V_1}\|u_{\Delta_k}\|_{\mathcal{M}}+\|\Delta_2^k\|_{V_2}),$$

for some constant c_u only depending on u. This finishes the proof since $\{\|u_{\Delta_k}\|_{\mathcal{M}}\}_{k\in\mathbb{N}}$ and $\psi(u_{\Delta_k},0)$ are bounded. □

To close this section, we provide a Lipschitz stability result for the objective function value.

Theorem 4.32. *Let a sequence of perturbations* $\{\Delta_k\}_{k\in\mathbb{N}}\subset N_V$ *with* $\lim_{k\to\infty}\Delta_k=0$ *be given. For* $k\in\mathbb{N}$ *let* $\bar{u}_{\Delta_k}$ *denote an optimal solution to* (P_{Δ_k})*. Assume that* $\bar{u}_{\Delta_k}\rightharpoonup^*\bar{u}_0\in\mathcal{M}^+(\Omega_o)$*. Then there holds*

$$|F_{\Delta_k}(\bar{u}_{\Delta_k})-F_0(\bar{u}_0)|\leq c\|\Delta_k\|_V,$$

for all k *large enough.*

Proof. First, we mention that $\bar{u}_0$ is an optimal solution to $(\mathcal{P}_0)$, see Proposition 4.29. Moreover, due to the convergence of the Fisher information matrices, there holds

$$\bar{u}_{\Delta_k}, \bar{u}_0 \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi(\cdot, \Delta_k) \cap \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi(\cdot, 0),$$

for all k large enough. By optimality of $\bar{u}_{\Delta_k}$ and $\bar{u}_0$ respectively we obtain

$$F_{\Delta_k}(\bar{u}_{\Delta_k}) - F_0(\bar{u}_{\Delta_k}) \leq F_{\Delta_k}(\bar{u}_{\Delta_k}) - F_0(\bar{u}_0) \leq F_{\Delta_k}(\bar{u}_0) - F_0(\bar{u}_0).$$

Taking the absolute value we thus conclude

$$|F_{\Delta_k}(\bar{u}_{\Delta_k}) - F_0(\bar{u}_0)| \leq \max\{|F_{\Delta_k}(\bar{u}_{\Delta_k}) - F_0(\bar{u}_{\Delta_k})|, |F_{\Delta_k}(\bar{u}_0) - F_0(\bar{u}_0)|\}.$$

Note that $F_{\Delta_k}(u) - F_0(u) = \psi(u, \Delta_k) - \psi(u, 0)$ for all $u \in \mathcal{M}^+(\Omega_o)$ and $k \in \mathbb{N}$. By Lemma 4.31 we obtain

$$\begin{aligned} &\max\{|F_{\Delta_k}(\bar{u}_{\Delta_k}) - F_0(\bar{u}_{\Delta_k})|, |F_{\Delta_k}(\bar{u}_0) - F_0(\bar{u}_0)|\} \\ &\qquad\qquad \leq c_{\bar{u}_0}(\max\{|\psi(\bar{u}_0, 0)|, |\psi(\bar{u}_{\Delta_k}, 0)|\}|\Delta_3^k| + \|\Delta_1^k\|_{V_1} + \|\Delta_2^k\|_{V_2}). \end{aligned}$$

Due to the weak* convergence of $\bar{u}_{\Delta_k}$ and the continuity of $\psi(\cdot, 0)$ the sequence $\psi(\bar{u}_{\Delta_k}, 0)$ is bounded. Thus, the statement follows. □

4.5.2 Stability and sensitivity of the design measure

In the previous section we have proven Lipschitz stability of the optimal function values with respect to the perturbation. Concerning the optimal design measure however, only (subsequential) weak* convergence has been shown. The aim of this section is to close this gap by providing qualitative and quantitative statements on the location of the support points and the convergence of the measurement weights.

Let us fix some additional notation and collect some general observations. To focus on the ideas behind the proofs in this section we will assume that

A4.8 The functional Ψ is strictly convex on its domain.

Again, this is for example the case for the A as well as the D optimal design criterion. This implies the uniqueness of the optimal Fisher information matrix $\mathcal{I}[\Delta_1](\bar{u}_\Delta)$ and the optimal gradient $\nabla_u\psi(\bar{u}_\Delta, \Delta)$ for all $\Delta \in N_V$. In the following, $\{\Delta_k\}_{k\in\mathbb{N}} \subset N_V$ will always denote a sequence of perturbations with $\Delta_k \to 0$, while $\{\bar{u}_{\Delta_k}\}_{k\in\mathbb{N}} \subset \mathcal{M}^+(\Omega_o)$ is a sequence of associated optimal design measures obtained from $(\mathcal{P}_{\Delta_k})$. W.l.o.g. we assume that $\bar{u}_{\Delta_k} \rightharpoonup^* \bar{u}_0$ for some optimal solution $\bar{u}_0$ of $(\mathcal{P}_0)$.

Given a perturbation $\Delta \in N_V$ and a solution $\bar{u}_\Delta$ to (P_Δ), we recall that

$$\nabla_u\psi(\bar{u}_\Delta, \Delta)(x) \leq 1, \quad \operatorname{supp} \bar{u}_\Delta \subset \{\, x \in \Omega_o \mid -\nabla_u\psi(\bar{u}_\Delta, \Delta)(x) = 1 \,\},$$

see Corollary 4.30. Due to the strict convexity of Ψ, the set of global maximizers to $-\nabla_u\psi(\bar{u}_\Delta, \Delta)$ only depends on the perturbation Δ and not on the particular optimal design measure $\bar{u}_\Delta$. Hence we will denote it by $\operatorname{Ext}(\Delta)$ in the following. Furthermore, again for abbreviation we define the mapping

$$\bar{p}_\cdot \colon N_V \to \mathcal{C}(\Omega_o), \quad \Delta \mapsto \bar{p}_\Delta = -\nabla_u\psi(\bar{u}_\Delta, \Delta)$$

Due to the assumptions on the perturbations and the weak*-to-strong continuity of the Fisher information operator, $\bar{p}_\Delta$ depends continuously on Δ. For $R > 0$, we define the R-extensions of the support of an optimal design $\bar{u}_\Delta \in \mathcal{M}^+(\Omega_o)$ by

$$\operatorname{supp}_R \bar{u}_\Delta = \bigcup_{x \in \operatorname{supp} \bar{u}_\Delta} \{\, \tilde{x} \in \Omega_o \mid |x - \tilde{x}| < R \,\}$$

and of the corresponding set of global maximizers $\operatorname{Ext}(\Delta)$ as

$$\operatorname{Ext}_R(\Delta) := \bigcup_{x \in \operatorname{Ext}(\Delta)} \{\, \tilde{x} \in \Omega_o \mid |x - \tilde{x}| < R \,\},$$

respectively. First, we address stability results for the positions of optimally placed sensors.

Lemma 4.33. *Let $R > 0$ be given. For all $\Delta \in N_V$ with $\|\Delta\|_V$ small enough, every optimal solution $\bar{u}_\Delta$ to (P_Δ) fulfills*

$$\operatorname{supp} \bar{u}_\Delta \subset \operatorname{Ext}(\Delta) \subset \operatorname{Ext}_R(0).$$

Proof. Given $R > 0$ the extended set $\operatorname{Ext}_R(\bar{u}_0)$ is open in Ω_o. Consequently, its complement in Ω_o, $K_R := \Omega_o \backslash \operatorname{Ext}_R(\bar{u}_0)$, is compact. By construction we have $\bar{p}_0 \in \mathcal{C}(K_R)$ and $\max_{x \in K_{R_1}} \bar{p}_0 < 1$. Define $r = 1 - \max_{x \in K_R} \bar{p}_0$. Given an arbitrary $x \in \operatorname{Ext}(\Delta)$ we have

$$\bar{p}_0(x) = 1 + \bar{p}_0(x) - \bar{p}_\Delta(x) \geq 1 - \|\bar{p}_\Delta - \bar{p}_0\|_{\mathcal{C}(\Omega_o)}. \tag{4.45}$$

Since the mapping

$$\bar{p} \colon N_V \to \mathcal{C}(\Omega_o) \quad \delta \mapsto \bar{p}_\Delta,$$

is continuous at zero, there exists $c_1 > 0$ with

$$\|\bar{p}_\Delta - \bar{p}_0\|_{\mathcal{C}(\Omega_o)} < \frac{r}{2} \quad \forall \delta \in N_V,\ \|\Delta\|_V \leq c_1.$$

We conclude

$$\bar{p}_0(x) > 1 - \frac{r}{2} > 1 - r = \max_{x \in K_R} \bar{p}_0,$$

and thus $x \in \operatorname{Ext}_R(0)$. Since $x \in \operatorname{Ext}(\Delta)$ was chosen arbitrarily and $\operatorname{supp} \bar{u}_\Delta \subset \operatorname{Ext}(\Delta)$ this concludes the proof. □

Remember that our special interest lies in optimal design measures consisting of finitely many Dirac delta functions corresponding to point measurements at their respective positions. From now on, let us assume that $\operatorname{Ext}(0)$ consists of finitely many distinct points:

A4.9 $\operatorname{Ext}(0) = \{\, x \in \Omega_o \mid \bar{p}_0(x) = 1 \,\} = \{\bar{x}_{i,0}\}_{i=1}^N \subset \Omega_o,$

for some $N \in \mathbb{N}$. In virtue of Corollary 4.30, every optimal solution $\bar{u}_0 \in \mathcal{M}^+(\Omega_o)$ to (P_0) is sparse, i.e.

$$\bar{u}_0 = \sum_{i=1}^N \bar{u}_{i,0} \delta_{\bar{x}_{i,0}} \quad \bar{u}_{i,0} \in \mathbb{R}_+,\ i \in \{1, \ldots, N\}.$$

Let us discuss the results of Lemma 4.33 in this case. For this purpose, choose $R_1 > 0$ with

$$\bar{B}_{R_1}(\bar{x}_{i,0}) \cap \bar{B}_{R_1}(\bar{x}_{j,0}) = \emptyset,\ i \neq j, i, j \in \{1, \dots, N\} \quad \text{where } B_{R_1}(\bar{x}_{i,0}) := \{\, x \in \Omega_o \mid |x - \bar{x}_{i,0}| < R_1 \,\}. \tag{4.46}$$

The statement of the previous lemma readily translates to this situation as

$$\operatorname{supp} \bar{u}_\Delta \subset \bigcup_{i=1}^{N} B_{R_1}(\bar{x}_{i,0}) \subset \bigcup_{i=1}^{N} \bar{B}_{R_1}(\bar{x}_{i,0}),$$

if the perturbation Δ is small enough. From this we conclude that the support of an arbitrary optimal solution $\bar{u}_\Delta$ to the perturbed problem (P_Δ) is localized in small balls around the possible support points of $\bar{u}_0$.
To prove a stability result for the optimal measurement weights we combine the localization results on the support of the optimal measurement design and the weak* convergence result for vanishing perturbations. To this end, let us denote the restriction of $\bar{u}_{\Delta_k}$ onto $\bar{B}_{R_1}(\bar{x}_{i,0})$ by $\bar{u}^i_{\Delta_k}$, $i = 1, \dots, N$. Loosely speaking, if $x_i \in \operatorname{supp} \bar{u}_0$, we will prove that $\|\bar{u}^i_{\Delta_k}\|_\mathcal{M}$ approximates $\bar{\mathbf{u}}_i$ in the limit while $\bar{u}_{\Delta_k}$ converges strongly to 0 on the complement of the extended support

$$\operatorname{supp}_{R_1} \bar{u}_0 = \bigcup_{\bar{x}_i \in \operatorname{supp} \bar{u}_0} B_{R_1}(\bar{x}_i).$$

Our findings are summarized in the following proposition.

Proposition 4.34. *Assume that $\bar{u}_{\Delta_k} \rightharpoonup^* \bar{u}_0$ with $\bar{u}_0 = \sum_{i=1}^{N} \bar{\mathbf{u}}_{i,0}\delta_{\bar{x}_{i,0}}$. Then there holds*

$$\operatorname{supp} \bar{u}_{\Delta_k} \subset \bigcup_{i=1}^{N} \bar{B}_{R_1}(\bar{x}_{i,0}), \quad \bar{B}_{R_1}(\bar{x}_{i,0}) \cap \bar{B}_{R_1}(\bar{x}_{j,0}) = \emptyset,\ i \neq j, \quad i, j \in \{1, \dots, N_d\}, \tag{4.47}$$

for all k large enough. Furthermore we have

$$\|\bar{u}^i_{\Delta_k}\|_\mathcal{M} \to \bar{\mathbf{u}}_i \quad \forall i \in \{1, \dots, N_d\}. \tag{4.48}$$

In particular, if $\bar{\mathbf{u}}_{i,0} > 0$ and k is large enough, then

$$\operatorname{supp} \bar{u}_{\Delta_k} \cap \bar{B}_{R_1}(\bar{x}_i) \neq \emptyset.$$

Conversely we have

$$\|\bar{u}_{\Delta_k}\|_{\mathcal{M}(\Omega_o \setminus \overline{\operatorname{supp}_{R_1} \bar{u}_0})} = \int_{\Omega_o \setminus \overline{\operatorname{supp}_{R_1} \bar{u}_0}} \mathrm{d}\bar{u}_{\Delta_k}(x) \to 0,$$

as $k \to \infty$. Here $\overline{\operatorname{supp}_{R_1} \bar{u}_0}$ denotes the closure of $\operatorname{supp}_{R_1} \bar{u}_0$ in Ω_o.

Proof. Let $j \in \{1, \dots, N\}$ be arbitrary but fixed. Since the sets $\bar{B}_{R_1}(\bar{x}_i)$, $i = 1, \dots, N$, are closed and pairwise disjoint, Urysohn's Lemma, see [266, 15.6], yields the existence of a function $\varphi_j \in \mathcal{C}(\Omega_o)$ with

$$\varphi_j(x) = 1, \quad \forall x \in \bar{B}_{R_1}(\bar{x}_{j,0}) \quad \varphi_j(x) = 0, \quad \forall x \in \bar{B}_{R_1}(\bar{x}_{i,0}),\ i \in \{1, \dots, N\},\ i \neq j.$$

By testing φ_j with $\bar{u}_{\Delta_k}$ we obtain

$$\langle \varphi_j, \bar{u}_{\Delta_k} \rangle = \int_{\Omega_o} \varphi_j \mathrm{d}\bar{u}_{\Delta_k}(x) = \int_{\bar{B}_{R_1}(\bar{x}_{j,0})} \mathrm{d}\bar{u}_{\Delta_k}(x).$$

Due to the weak* convergence of $\bar{u}_{\Delta_k}$ we thus conclude

$$\int_{\bar{B}_{R_1}(\bar{x}_{j,0})} \mathrm{d}\bar{u}_{\Delta_k}(x) = \langle \varphi_j, \bar{u}_{\delta_k} \rangle \to \langle \varphi_j, \bar{u}_0 \rangle = \bar{\mathbf{u}}_{j,0},$$

as $k \to \infty$, which gives (4.47). To prove the last statement, (4.48), we note that

$$\int_{\Omega_o \setminus \overline{\mathrm{supp}_{R_1} \bar{u}_0}} \mathrm{d}\bar{u}_{\Delta_k}(x) = \sum_{\substack{i \in \{1,\ldots,N\} \\ \bar{u}_{i,0}=0}} \int_{\bar{B}_{R_1}(\bar{x}_{i,0})} \mathrm{d}\bar{u}_{\Delta_k}(x) \to \sum_{\substack{i \in \{1,\ldots,N_d\} \\ \bar{u}_{i,0}=0}} \bar{\mathbf{u}}_{i,0} = 0,$$

where we used (4.47). □

Up to now we have not discussed the stability of the optimal number of sensors. In the light of Proposition 4.28 we can readily assume that $\# \operatorname{supp} \bar{u}_{\Delta_k} \leq n(n+1)/2$. By the results of Proposition 4.34, at least one perturbed optimal sensor is placed in the vicinity of each unperturbed sensor. However it is possible that a single, optimally placed, Dirac delta function in the unperturbed problem $(\mathcal{P}_0)$ is approximated by a larger number of perturbed ones since the design measures only converge in the weak* sense. Hence, in most situations, we expect $\# \operatorname{supp} \bar{u}_\Delta > \# \operatorname{supp} \bar{u}_0$. In the following we will focus on the question whether we can expect stability for the number of optimally placed sensors under small perturbations in a more restrictive setting. For this purpose, we again impose additional assumptions on $\operatorname{Ext}(0)$ and on the (local) regularity of $\partial S[\Delta]$ for a given $\Delta \in N_V$.

Assumption 4.7. Assume that $\operatorname{Ext}(0) = \{\bar{x}_{i,0}\}_{i=1}^N \subset \operatorname{int} \Omega_o$ and there exists $R > 0$ with

$$\Omega_R := \bigcup_{i=1}^N B_R(\bar{x}_i) \subset \operatorname{int} \Omega_o, \quad \bar{B}_R(\bar{x}_i) \cap \bar{B}_R(\bar{x}_j) = \emptyset, \quad \partial S \in \mathcal{C}^2(N_V \times \bar{\Omega}_R, \mathbb{R}^n) \cap \mathcal{C}(N_V \times \Omega_o, \mathbb{R}^n),$$

for all $i, j \in \{1, \ldots, N\}$, $i \neq j$

Throughout the following considerations we tacitly assume that $\|\Delta\|_V$ is chosen small enough such that $\operatorname{supp} \bar{u}_\Delta \subset \Omega_R$ for every optimal design measure $\bar{u}_\Delta$ obtained from (P_Δ). We immediately derive the following regularity result for $\bar{p}_\Delta$.

Corollary 4.35. *Let $\Delta \in N_V$ be given. Then there holds $\bar{p}_\Delta \in \mathcal{C}^2(\bar{\Omega}_R)$. Furthermore the mapping $\bar{p}.\colon N_V \to \mathcal{C}^2(\bar{\Omega}_R)$, $\Delta \mapsto \bar{p}_\Delta$ is continuous at* 0.

Proof. Let $\Delta \in N_V$ be given. By assumption, the mapping

$$\partial S\colon N_{V_1} \to \mathcal{C}^2(\bar{\Omega}_R, \mathbb{R}^n), \quad \Delta_1 \mapsto \partial S[\Delta_1],$$

is continuous. Hence the same holds for the perturbation mapping of the pointwise Fisher information $I \in \mathcal{C}(N_{V_1} \times \Omega_o, \operatorname{Sym}(n))$. The claimed statements follows from the definition of $\bar{p}_\Delta$ as

$$\bar{p}_\Delta = -\frac{1}{\beta_0 + \Delta_3} \operatorname{Tr}(I[\Delta_1](x) \nabla \Psi(\mathcal{I}[\Delta_1](\bar{u}_\Delta) + \mathcal{I}_0[\Delta_2])),$$

and $\mathcal{I}_0[\Delta_2] \to \mathcal{I}_0[0]$, $\beta_0 + \Delta_3 \to \beta_0$ as well as $\mathcal{I}[\Delta_1](\bar{u}_\Delta) \to \mathcal{I}[0](\bar{u}_0)$ for $\Delta \to 0$. □

In the following, we need a perturbation result for the Hessian of $\bar{p}_0$.

Lemma 4.36. *Let $\hat{x} \in \operatorname{Ext}(0)$ be given and assume that $\nabla^2 \bar{p}_0(\hat{x})$ is negative definite, i.e. there exists $\theta > 0$ with*

$$-(\zeta, \nabla^2 \bar{p}_0(\hat{x})\zeta)_{\mathbb{R}^d} \geq \theta |\zeta|^2_{\mathbb{R}^d}, \quad \forall \zeta \in \mathbb{R}^d. \tag{4.49}$$

Then there exist $R_2 > 0$ such that for all $x \in B_{R_2}(\hat{x}) \subset \Omega_R$ and all $\Delta \in N_V$ with $\|\Delta\|_V$ small enough, we have

$$-(\zeta, \nabla^2 \bar{p}_\Delta(x)\zeta)_{\mathbb{R}^d} \geq \frac{\theta}{2} |\zeta|^2_{\mathbb{R}^d}, \quad \forall \zeta \in \mathbb{R}^d. \tag{4.50}$$

Proof. Let $x \in \Omega_R$ and $\zeta \in \mathbb{R}^d$ be given. We expand

$$(\zeta, \nabla^2 \bar{p}_\Delta(x)\zeta)_{\mathbb{R}^d} = (\zeta, \nabla^2 \bar{p}_0(\hat{x})\zeta)_{\mathbb{R}^d} + \left(\zeta, (\nabla^2 \bar{p}_0(x) - \nabla^2 \bar{p}_0(\hat{x}) + \nabla^2 \bar{p}_\Delta(x) - \nabla^2 \bar{p}_0(x))\zeta\right)_{\mathbb{R}^d}.$$

Using the negative definiteness of the Hessian of $\bar{p}_0$, we estimate

$$(\zeta, \nabla^2 \bar{p}_\Delta(x)\zeta)_{\mathbb{R}^d} \leq \left(-\theta + \|\nabla^2 \bar{p}_0(x) - \nabla^2 \bar{p}_0(\hat{x})\|_{\mathbb{R}^{d\times d}} + \|\bar{p}_0 - \bar{p}_\Delta\|_{\mathcal{C}^2(\Omega_R)}\right) |\zeta|^2_{\mathbb{R}^d}.$$

Since $\bar{p}_0 \in \mathcal{C}^2(\Omega_R)$ and $\bar{p}_\cdot \in \mathcal{C}(N_V, \mathcal{C}^2(\Omega_R))$ there exist constants $R_2 > 0$ and $\varepsilon > 0$ with

$$\|\nabla^2 \bar{p}_0(x) - \nabla^2 \bar{p}_0(\hat{x})\|_{\mathbb{R}^{d\times d}} \leq \frac{\theta}{4}, \quad \|\bar{p}_0 - \bar{p}_\Delta\|_{\mathcal{C}^2(\Omega_R)} \leq \frac{\theta}{4},$$

for all $x \in \Omega_R$, $|x - \hat{x}|_{\mathbb{R}^d} < R_2$ and $\Delta \in N_V$, $\|\Delta\|_V < \varepsilon$. Combining these results with the previous estimate we conclude (4.50), finishing the proof. □

We are now able to strengthen the stability result on the global maximizers of $\bar{p}_0$ if their curvature is not degenerate.

Proposition 4.37. *Let $\hat{x} \in \operatorname{Ext}(0)$ be given such that (4.49) holds for some $\theta > 0$. Then there exists $R_2 > 0$, with $B_{R_2}(\hat{x}) \subset \Omega_R$, $B_{R_2}(\hat{x}) \cap \operatorname{Ext}(0) = \{\hat{x}\}$, such that for all $\Delta \in N_V$, $\|\Delta\|_V$ small enough, we either have*

$$\operatorname{Ext}(\Delta) \cap B_{R_2}(\hat{x}) = \emptyset \quad \text{or} \quad \operatorname{Ext}(\Delta) \cap B_{R_2}(\hat{x}) = \{\hat{x}_\Delta\}, \tag{4.51}$$

for some $\hat{x}_\Delta \in B_{R_2}(\hat{x})$. Furthermore, if $\bar{u}_{\Delta_k} \rightharpoonup^* \bar{u}_0$ with $\hat{x} \in \operatorname{supp} \bar{u}_0$ then only the second case is possible and there exists a sequence $\{\hat{x}_{\Delta_k}\}_{k\in\mathbb{N}} \subset B_{R_2}(\bar{x}_i)$ with

$$\operatorname{Ext}(\Delta_k) \cap B_{R_2}(\hat{x}) = \{\hat{x}_{\Delta_k}\},$$

for all k large enough.

Proof. Let such a $\hat{x} \in \operatorname{Ext}(0)$ be given. We start by proving the existence of $R_2 > 0$ such that (4.51) holds for small perturbations. Since $\operatorname{Ext}(\bar{u})$ consists of finitely many points and Ω_R is open in Ω_o we can choose $R_2 > 0$ with

$$B_{R_2}(\hat{x}) \subset \Omega_R, \quad B_{R_2}(\hat{x}) \cap \operatorname{Ext}(\bar{u}_0) = \{\hat{x}\}.$$

By choosing R_2 small enough, see Lemma 4.36, we conclude from (4.50) that $\bar{p}_\Delta$ is strictly concave and thus admits at most one of its global maximizers in $B_{R_2}(\hat{x})$ for all $\Delta \in N_V$ small enough. Due to $\bar{p}_\Delta(x) \leq 1$, $x \in \Omega_o$, and $\|\bar{p}_\Delta\|_{\mathcal{C}(\Omega_o)} = 1$, the statement in (4.51) readily follows. □

Roughly speaking, the previous result states that an unperturbed optimal sensor at a position $\hat{x}$ is approximated by exactly one sensor in the perturbed problem if the curvature of $\bar{p}_0$ at $\hat{x}$ does not degenerate. Hence, non-degeneracy of the curvature of $\bar{p}_0$ at every optimal sensor positions guarantees one-to-one approximation of the sensors in the unperturbed problem by the perturbed optimal ones. We adopt this condition as a standing assumption in the following. Additionally, we impose regularity assumptions on the perturbations.

Assumption 4.8. Assume that there exists $\theta > 0$ with

$$-(\zeta, \nabla^2 \bar{p}_0(\bar{x}_{i,0})\zeta)_{\mathbb{R}^d} \geq \theta |\zeta|^2_{\mathbb{R}^d} \quad \forall \zeta \in \mathbb{R}^d,$$

for all $i = 1, \dots, N$. Furthermore, the mapping $\mathcal{I}_0 \colon N_{V_2} \to \operatorname{Sym}(n)$ is two times continuously Fréchet differentiable, the set $\{\mathcal{I}[0](\delta_{\bar{x}_{i,0}})\}_{i=1}^N$ is linearly independent and there exists $\gamma_0 > 0$ with

$$\operatorname{Tr}(\delta B \nabla^2 \Psi(\mathcal{I}[0](\bar{u}_0) + \mathcal{I}_0[0])\delta B) \geq \gamma_0 \|\delta B\|_{\mathrm{Sym}}, \quad \forall \delta B \in \operatorname{Sym}(n).$$

Note that as a consequence of these assumptions, the unperturbed optimal design measurement $\bar{u}_0$ is unique. In the following two technical lemmas we will prove that, as a consequence, the optimal solution $\bar{u}_\Delta$ to (P_Δ) is unique for small perturbations. In the light of Proposition 4.37 we thus conclude

$$\#\operatorname{supp} \bar{u}_0 \leq \#\operatorname{supp} \bar{u}_\Delta \leq \#\operatorname{Ext}(\Delta) \leq \#\operatorname{Ext}(0).$$

Moreover, the positions of the optimal sensors and the measurement weights depend continuously on the perturbation. Consequently, if $\operatorname{supp} \bar{u}_0 = \operatorname{Ext}(0)$, all the inequalities above become equalities yielding

$$\#\operatorname{supp} \bar{u}_0 = \#\operatorname{supp} \bar{u}_\Delta = \#\operatorname{Ext}(0).$$

Hence, the number of optimal sensors is stable for small perturbations in this situation.

Lemma 4.38. *For all $\Delta \in N_V$, $\|\Delta\|_V$ small enough there exist $\bar{x}_{i,\Delta} \in B_{R_2}(\bar{x}_{i,0})$, $i = 1, \dots N$, with* $\operatorname{Ext}(\Delta) \subset \{\bar{x}_{i,\Delta}\}_{i=1}^N$. *Moreover there exists a small neighborhood $\hat{N}_V$ of 0 such that the mapping*

$$\bar{\boldsymbol{x}}_\cdot \colon \hat{N}_V \to \Omega_R^N, \quad \Delta \mapsto \bar{\boldsymbol{x}}_\Delta = (\bar{x}_{1,\Delta}, \cdots, \bar{x}_{N,\Delta})^\top,$$

is well-defined and continuous at zero.

Proof. Due to its continuity, $\bar{p}_\Delta$ admits at least one global maximum on $\bar{B}_{R_2}(\bar{x}_{i,0})$ for all $\Delta \in N_V$ and $i = 1, \dots, N$. From the uniform convergence of $\bar{p}_\Delta$ towards $\bar{p}_0$ and $\operatorname{Ext}(0) = \{\bar{x}_{i,0}\}_{i=1}^N$ we conclude the existence of $\sigma > 0$ such that

$$\bar{p}_\Delta(x) \leq 1 - \sigma, \quad \forall x \in \Omega_o \setminus \bigcup_{i=1}^N B_{R_2}(\bar{x}_{i,0}),$$

for all $\Delta \in N_V$ small enough. Let $i \in \{1, \dots, N\}$ be arbitrary but fixed. Again using the continuity of $\bar{p}_\cdot$ we conclude $\max_{x \in \bar{B}_{R_2}} \bar{p}_\Delta(x) > \bar{p}_\Delta(\bar{x}_{i,0}) > 1 - \sigma$. As a consequence, each global maximum of $\bar{p}_\Delta$ lies in the interior of the ball. Consequently it is unique since $\bar{p}_\Delta$ is strictly concave on $B_{R_2}(\bar{x}_{i,0})$. We denote it by $\bar{x}_{i,\Delta}$. It remains to prove the continuity of $\bar{\mathbf{x}}_\cdot$. Assume that $\bar{\mathbf{x}}_\cdot$ is not continuous at 0. Then there exists $\varepsilon > 0$ such that for all $\sigma_k = 1/k$ there is a perturbation

Δ_k and $|\mathbf{x}_{\Delta_k} - \mathbf{x}_{i,0}| > \varepsilon$. Let $i \in \{1, \dots, N\}$ be given. By boundedness of Ω_R we can extract a subsequence of $\bar{x}_{i,\Delta_k}$ denoted by the same symbol with $\bar{x}_{i,\Delta_k} \to \bar{x} \in \bar{B}_{R_2}(\bar{x}_{i,0})$. We conclude

$$1 - \bar{p}_0(\bar{x}_{i,\Delta_k}) \leq \bar{p}_0(\bar{x}_{i,0}) - \bar{p}_\Delta(\bar{x}_{i,0}) + \bar{p}_{\Delta_k}(\bar{x}_{i,\Delta_k}) - \bar{p}_0(\bar{x}_{i,\Delta_k}) \leq c\|\bar{p}_{\Delta_k} - \bar{p}_0\|_C,$$

and consequently $\bar{p}_0(\bar{x}) = 1$. By the uniqueness of the global maximum of $\bar{p}_0$ on $\bar{B}_{R_2}(\bar{x}_{i,0})$, we have $\bar{x} = \bar{x}_{i,0}$. This gives a contradiction since $\bar{x}$ was an arbitrary accumulation point and i was chosen arbitrarily. □

In the following we consider w.l.o.g perturbations in the smaller set $\hat{N}_V \subset N_V$.

Lemma 4.39. *For all $\Delta \in \hat{N}_V$, the optimal solution $\bar{u}_\Delta = \sum_{i=1}^N \bar{\mathbf{u}}_{i,\Delta}\delta_{\bar{x}_{i,\Delta}}$ to (P_Δ) is unique. Furthermore, the mapping*

$$\bar{\mathbf{u}}_\cdot \colon \hat{N}_V \to \mathbb{R}^N_+, \quad \Delta \mapsto \bar{\mathbf{u}}_\Delta = (\bar{\mathbf{u}}_{1,\Delta}, \dots, \bar{\mathbf{u}}_{N,\Delta})^\top,$$

is continuous at 0.

Proof. Let us first proof the uniqueness of the perturbed optimal design $\bar{u}_\Delta$. From the previous lemma we recall that $\operatorname{Ext}(\Delta) \subset \{\bar{x}_{i,\Delta}\}_{i=1}^N$. Thus, every perturbed optimal design is of the form $\bar{u}_\Delta = \sum_{i=1}^N \bar{\mathbf{u}}_i\delta_{\bar{x}_{i,\Delta}}$ for some $\mathbf{u}_i \in \mathbb{R}_+$. For $i = 1, \dots, N$, we interpret $I[0](\bar{x}_{i,0}) \in \operatorname{Sym}(n)$ as a vector in $\mathbb{R}^{n(n+1)/2\times N}$. We assemble these vectors in a matrix

$$V_0 = (I[0](\bar{x}_{1,0})| \cdots |I[0](\bar{x}_{N,0})) \in \mathbb{R}^{n(n+1)/2\times N}.$$

Note that $\operatorname{rank} V_0 = N$ due to the linear independence assumption. Similarly we proceed for the perturbed problem, defining the matrix

$$V_\Delta = (I[\Delta_1](\bar{x}_{1,\Delta})| \cdots |I[\Delta_1](\bar{x}_{N,\Delta})) \in \mathbb{R}^{n(n+1)/2\times N}, \quad \forall \Delta \in \hat{N}_V.$$

From the continuity of the pointwise Fisher information and of $\bar{\mathbf{x}}_\cdot$ we obtain $\lim_{\Delta\to 0} \|V_\Delta - V_0\|_{\mathbb{R}^{d\times d}} = 0$. Since the rank of a matrix is a lower semi-continuous function we conclude

$$N = \operatorname{rank} V_0 \leq \liminf_{\Delta\to 0} \operatorname{rank} V_\Delta \leq N.$$

W.l.o.g we can thus assume that $\hat{N}_V$ is chosen small enough such that $\operatorname{rank} V_\Delta = N$ for all $\Delta \in \hat{N}_V$. Consequently, the set $\{I[\Delta_1](\bar{x}_{i,\Delta})\}_{i=1}^N$ is linearly independent and the optimal design $\bar{u}_\Delta$ is unique, see Corollary 3.19. It remains to discuss the continuity of the mapping $\bar{\mathbf{u}}_\cdot$. We prove it by contradiction. Assume that there exists $\varepsilon > 0$ such that for all $\sigma > 0$ there is an element Δ with $|\bar{\mathbf{u}}_\Delta - \bar{\mathbf{u}}_0|_{\mathbb{R}^N} > \varepsilon$ and $\|\Delta\|_V < \sigma$. Now choose a sequence of such perturbations $\Delta_k \in \hat{N}_V$ with $\|\Delta_k\|_V = 1/k$ and the associated coefficient vectors $\bar{\mathbf{u}}_{\Delta_k}$. By assumption there exists $\varepsilon > 0$ with $|\bar{\mathbf{u}}_{\Delta_k} - \bar{\mathbf{u}}_0|_{\mathbb{R}^N} > \varepsilon$ for all $k \in \mathbb{N}$. This however contradicts $\bar{u}_{\Delta_k} \rightharpoonup^* \bar{u}_0$ and (4.48). Hence the mapping is continuous. □

The rest of this section focusses on properties of the mappings $\bar{\mathbf{x}}_\cdot$ and $\bar{\mathbf{u}}_\cdot$ beyond continuity. We make the following additional assumption.

Assumption 4.9. There holds $\operatorname{supp} \bar{u}_0 = \operatorname{Ext}(0)$, i.e. there exists $\bar{\mathbf{u}}_{i,0} > 0$, $i = 1, \dots N$ with

$$\bar{u}_0 = \sum_{i=1}^N \bar{\mathbf{u}}_{i,0}\delta_{\bar{x}_{i,0}}.$$

The aim of the following technical discussion is to establish Fréchet differentiability of the mappings $\bar{\mathbf{x}}.$ and, $\bar{\mathbf{u}}.$, respectively, in a neighborhood of 0. Choosing the neighborhood $\hat{N}_V$ small enough we start by concluding the following corollary from the continuity of $\bar{\mathbf{u}}.$.

Corollary 4.40. *Let Assumption 4.9 hold. For all $\Delta \in \hat{N}_V$ there holds $\bar{\mathbf{u}}_{i,\Delta} > 0$, $i = 1, \ldots, N$.*

Proof. The claim follows immediately from the continuity of $\bar{\mathbf{u}}.$ and $\bar{\mathbf{u}}_{i,0} > 0$ for all $i = 1, \ldots, N$. □

We make an important observation. Let us define the admissible set X_{ad} as the cartesian product $X_{ad} = \bar{B}_{R_2}(\bar{x}_{1,0}) \times \cdots \bar{B}_{R_2}(\bar{x}_{N,0})$. Due to the previous results, obtaining a solution $\bar{u}_\Delta$ to (P_Δ) is reduced to solving the finite dimensional, non-linear, optimization problem

$$\min_{\substack{\mathbf{x}=(x_1,\ldots,x_N)^\top \in X_{ad},\\ \mathbf{u}\in\mathbb{R}^N_+}} \mathbf{F}(\mathbf{x},\mathbf{u},\Delta) = \left[\Psi\left(\sum_{i=1}^N \mathbf{u}_i I[\Delta_1](x_i) + \mathcal{I}_0[\Delta_2]\right) + (\beta_0 + \Delta_3)\|\mathbf{u}\|_1\right]. \qquad (\mathcal{P}^N_\Delta)$$

While we will never solve the, potentially non-convex, problem $(\mathcal{P}^N_\Delta)$ in practice, this relation allows to break down the measure-valued problem (P_Δ) to a finite dimensional optimization problem. We will use this fact in the following to infer stability properties of $\bar{u}_\Delta$ by applying ideas for parametric optimization problems in finite dimensions to $(\mathcal{P}^N_\Delta)$, see e.g. [43,88,108,168]. From the regularity assumptions on the perturbations, $\mathcal{I}[0](\bar{u}_0)+\mathcal{I}_0[0] \in \mathrm{PD}(n)$ and $\bar{\mathbf{u}}_i > 0$, $i = 1, \ldots, N$, we conclude that the functional $\mathbf{F}$ is at least two-times continuously differentiable in a neighborhood of $(\bar{\mathbf{x}}_0, \bar{\mathbf{u}}_0, 0)$. We denote the partial derivatives of $\mathbf{F}$ with respect to $\mathbf{x}$, $\mathbf{u}$ and Δ by $\partial_\mathbf{x}\mathbf{F}$, $\partial_\mathbf{u}\mathbf{F}$ and $\partial_\Delta\mathbf{F}$, respectively. Second order derivatives are denoted by $\partial.\partial.\mathbf{F}$. Hence, $(\mathcal{P}^N_\Delta)$ is a nonlinear but smooth optimization problem with additional constraints on the optimization variables. However, since $\bar{\mathbf{x}}_0 \in \operatorname{int} X_{ad}$ and $\bar{\mathbf{u}}_{i,0} > 0$, $i = 1, \ldots, N$, these constraints are also inactive for the perturbed solutions $(\bar{\mathbf{x}}_\Delta, \bar{\mathbf{u}}_\Delta)$ due to their continuity with respect to Δ. As a consequence the tupel $(\bar{\mathbf{x}}_\Delta, \bar{\mathbf{u}}_\Delta)$ fulfils the first order necessary optimality conditions for $\mathbf{F}(\cdot,\cdot,\Delta)$ given by

$$\partial_\mathbf{x}\mathbf{F}(\bar{\mathbf{x}}_\Delta, \bar{\mathbf{u}}_\Delta, \Delta) = 0, \quad \partial_\mathbf{u}\mathbf{F}(\bar{\mathbf{x}}_\Delta, \bar{\mathbf{u}}_\Delta, \Delta) = 0. \qquad (4.52)$$

In order to keep the notation more compact we recall the parameterized design measure and the gradient mapping

$$\boldsymbol{u}\colon X_{ad} \times \mathbb{R}^N_+ \to \mathcal{M}^+(\Omega_o), \quad (\mathbf{x},\mathbf{u}) \mapsto \boldsymbol{u}(\mathbf{x},\mathbf{u}) = \sum_{i=1}^N \mathbf{u}_i \delta_{x_i},$$
$$p\colon \mathcal{M}^+(\Omega_o) \times \hat{N}_V \to \mathcal{C}(\Omega_o) \cap \mathcal{C}^2(\Omega_R), \quad (u,\Delta) \mapsto p(u,\Delta) = -\nabla_u \psi(u,\Delta).$$

Whenever $p(u,\Delta)$ is well-defined, we denote its gradient and its Hessian with respect to the spatial variable x by $\nabla p(u,\Delta)$ and $\nabla^2 p(u,\Delta)$. Furthermore, for $i = 1, \ldots, N$, we (formally) define the mappings

$$G^i_1\colon X_{ad} \times \mathbb{R}^N_+ \times \hat{N}_V \to \mathbb{R}, \quad (\mathbf{x},\mathbf{u},\Delta) \mapsto p(\boldsymbol{u}(\mathbf{x},\mathbf{u}),\Delta)(x_i) - 1,$$
$$G^i_2\colon X_{ad} \times \mathbb{R}^N_+ \times \hat{N}_V \to \mathbb{R}^d, \quad (\mathbf{x},\mathbf{u},\Delta) \mapsto \nabla p(\boldsymbol{u}(\mathbf{x},\mathbf{u}),\Delta)(x_i).$$

As with the objective functional $\mathbf{F}$ we conclude that G^i_1 and G^i_2, respectively, are well-defined and of class $\mathcal{C}^1$ in a neighbourhood of $(\bar{\mathbf{x}}_0, \bar{\mathbf{u}}_0, 0)$, $i = 1, \ldots, N$. Differentiating $\mathbf{F}$ at $(\bar{\mathbf{x}}_\Delta, \bar{\mathbf{u}}_\Delta, \Delta)$ it is straightforward to verify that (4.52) is equivalent to

$$G^i_1(\mathbf{x}_\Delta, \mathbf{u}_\Delta, \Delta) = \bar{p}_\Delta(\bar{x}_{i,\Delta}) + 1 = 0, \quad G^i_2(\mathbf{x}_\Delta, \mathbf{u}_\Delta, \Delta) = \nabla \bar{p}_\Delta(\bar{x}_{i,\Delta}) = 0,$$

for all $i = 1, \dots, N$. We consider the following system of non-linear equations

$$G\colon X_{ad} \times \mathbb{R}^N_+ \times \hat{N}_V \to \mathbb{R}^N \times (\mathbb{R}^d)^N, \quad (\mathbf{x}, \mathbf{u}, \Delta) \mapsto \begin{pmatrix} G^1_1(\mathbf{x}_\Delta, \mathbf{u}_\Delta, \Delta) \\ \cdot \\ G^N_1(\mathbf{x}_\Delta, \mathbf{u}_\Delta, \Delta) \\ G^1_2(\mathbf{x}_\Delta, \mathbf{u}_\Delta, \Delta) \\ \cdot \\ G^N_2(\mathbf{x}_\Delta, \mathbf{u}_\Delta, \Delta) \end{pmatrix}.$$

Due to its optimality for $(\mathcal{P}^N_\Delta)$, $(\bar{\mathbf{x}}_\Delta, \bar{\mathbf{u}}_\Delta)$ is the unique root of $G(\cdot, \cdot, \Delta)$ in $X_{ad} \times \mathbb{R}^N_+$. We are now ready to state the main theorem of this section. Exploiting the equivalence between (P_Δ) and $(\mathcal{P}^N_\Delta)$, the optimal positions of the measurement sensors as well as the measurement weights depend at least (locally) continuously differentiable on the perturbation of the problem.

Theorem 4.41. *Let Assumption 4.6, (**A**4.8), (**A**4.9), Assumptions 4.7,4.8 as well as 4.9 hold. The mappings $\bar{\boldsymbol{x}}_\cdot$ and $\bar{\mathbf{u}}_\cdot$ from Lemma 4.38 and Lemma 4.39, respectively, are at least continuously Fréchet differentiable with*

$$\begin{pmatrix} \nabla_\Delta \bar{\boldsymbol{x}}_{\hat{\Delta}} \\ \nabla_\Delta \bar{\mathbf{u}}_{\hat{\Delta}} \end{pmatrix} = -\left(\partial_{(\boldsymbol{x},\mathbf{u})} G(\bar{\boldsymbol{x}}_{\hat{\Delta}}, \bar{\mathbf{u}}_{\hat{\Delta}}, \hat{\Delta})\right)^{-1} \partial_\Delta G(\bar{\boldsymbol{x}}_{\hat{\Delta}}, \bar{\mathbf{u}}_{\hat{\Delta}}, \hat{\Delta}), \tag{4.53}$$

for $\hat{\Delta} \in \hat{N}_V$. Here $\partial_{(\boldsymbol{x},\mathbf{u})} G$ and $\partial_\Delta G$ denote the Jacobian of G with respect to $(\boldsymbol{x}, \mathbf{u})$ and Δ. Moreover, there holds

$$\sum_{i=1}^N [|\bar{\boldsymbol{x}}_{i,\Delta} - \bar{\boldsymbol{x}}_{i,0}|_{\mathbb{R}^d} + |\bar{\mathbf{u}}_{i,\Delta} - \bar{\mathbf{u}}_{i,0}|] \le c\|\Delta\|_V,$$

for some $c > 0$.

Proof. If G is Fréchet differentiable at a given $(\mathbf{x}, \mathbf{u}, \Delta) \in X_{ad} \times \mathbb{R}^N_+$ its partial derivative are given in terms of $\partial_{(\mathbf{x},\mathbf{u})} G(\mathbf{x}, \mathbf{u}, \Delta) = H_1(\mathbf{x}, \mathbf{u}, \Delta) + H_2(\mathbf{x}, \mathbf{u}, \Delta)$. Let us characterize the matrices $H_1(\mathbf{x}, \mathbf{u}, \Delta)$, $H_2(\mathbf{x}, \mathbf{u}, \Delta) \in \mathrm{Sym}(dN + N)$. Given $\delta\mathbf{x} = (\delta x_1, \dots, \delta x_N) \in (\mathbb{R}^d)^N$ and $\delta\mathbf{u} = (\delta\mathbf{u}_1, \dots, \delta\mathbf{u}_N) \in \mathbb{R}^N$ there holds

$$\begin{aligned} &\begin{pmatrix} \delta\mathbf{x} & \delta\mathbf{u} \end{pmatrix} H_1(\mathbf{x}, \mathbf{u}, \Delta) \begin{pmatrix} \delta\mathbf{x} \\ \delta\mathbf{u} \end{pmatrix} \\ &\qquad = \sum_{i=1}^N [2\delta\mathbf{u}_i (\nabla p(\boldsymbol{u}(\mathbf{x}, \mathbf{u}), \Delta)(x_i), \delta x_i)_{\mathbb{R}^d} + \mathbf{u}_i (\delta x_i, \nabla^2 p(\boldsymbol{u}(\mathbf{x}, \mathbf{u}), \Delta)(x_i) \delta x_i)_{\mathbb{R}^d}], \end{aligned}$$

as well as

$$\begin{aligned} &\begin{pmatrix} \delta\mathbf{x} & \delta\mathbf{u} \end{pmatrix} H_2(\mathbf{x}, \mathbf{u}, \Delta) \begin{pmatrix} \delta\mathbf{x} \\ \delta\mathbf{u} \end{pmatrix} \\ &\qquad = -\frac{1}{\beta_0 + \Delta_3} \mathrm{Tr}(A(\mathbf{x}, \mathbf{u}, \delta\mathbf{x}, \delta\mathbf{u}, \Delta) \nabla^2 \Psi(\mathcal{I}[\Delta_1](\boldsymbol{u}(\mathbf{x}, \mathbf{u})) + \mathcal{I}_0([\Delta_2])) A(\mathbf{x}, \mathbf{u}, \delta\mathbf{x}, \delta\mathbf{u}, \Delta)) \end{aligned}$$

with the matrix $A(\mathbf{x}, \mathbf{u}, \delta\mathbf{x}, \delta\mathbf{u}, \Delta) \in \mathrm{Sym}(n)$ given by

$$A(\mathbf{x}, \mathbf{u}, \delta\mathbf{x}, \delta\mathbf{u}, \Delta) = \sum_{i=1}^N [\delta\mathbf{u}_i I[\Delta_1](x_i) + \mathbf{u}_i I'[\Delta_1](x_i) \delta x_i],$$

where $I'[\varDelta](x_i)$ denotes the first Fréchet derivative of $I \in \mathcal{C}^2(N_{V_1} \times \varOmega_R, \mathrm{Sym}(n)$ with respect to the spatial variable. We show that $\partial_{(\mathbf{x},\mathbf{u})}G(\bar{\mathbf{x}}_0, \bar{\mathbf{u}}_0, 0)$ is invertible. First, note that due to the optimality of $(\bar{\mathbf{x}}_0, \bar{\mathbf{u}}_0)$ for $(\mathcal{P}_0^N)$, the matrix $H_1(\bar{\mathbf{x}}, \bar{\mathbf{u}}_0, 0)$ simplifies to

$$\begin{pmatrix} \delta\mathbf{x} \; \delta\mathbf{u} \end{pmatrix} H_1(\bar{\mathbf{x}}_0, \bar{\mathbf{u}}_0, 0) \begin{pmatrix} \delta\mathbf{x} \\ \delta\mathbf{u} \end{pmatrix} = \sum_{i=1}^{N} [\mathbf{u}_{i,0}(\delta x_i, \nabla^2 \bar{p}_0(\bar{x}_{i,0})\delta x_i)_{\mathbb{R}^d}].$$

From Assumption 4.8 we conclude

$$\begin{pmatrix} \delta\mathbf{x} \; \delta\mathbf{u} \end{pmatrix} H_1(\bar{\mathbf{x}}_0, \bar{\mathbf{u}}_0, 0) \begin{pmatrix} \delta\mathbf{x} \\ \delta\mathbf{u} \end{pmatrix} \leq -\theta \sum_{i=1}^{N} \bar{\mathbf{u}}_{i,0} |\delta x_i|^2_{\mathbb{R}^d}.$$

Similarly we obtain

$$\begin{pmatrix} \delta\mathbf{x} \; \delta\mathbf{u} \end{pmatrix} H_2(\mathbf{x}, \mathbf{u}, \varDelta) \begin{pmatrix} \delta\mathbf{x} \\ \delta\mathbf{u} \end{pmatrix} \leq -\gamma_0 \|A(\bar{\mathbf{x}}_0, \bar{\mathbf{u}}_0, \delta\mathbf{x}, \delta\mathbf{u}, 0)\|^2_{\mathrm{Sym}}.$$

We distinguish two cases in the following.

Case 1: Assume that $\delta\mathbf{x} \neq 0$ and $\delta\mathbf{u}$ is arbitrary. From the previous discussion we readily deduce

$$\begin{pmatrix} \delta\mathbf{x} \; \delta\mathbf{u} \end{pmatrix} \partial_{(\mathbf{x},\mathbf{u})}G(\bar{\mathbf{x}}_0, \bar{\mathbf{u}}_0, 0) \begin{pmatrix} \delta\mathbf{x} \\ \delta\mathbf{u} \end{pmatrix} \leq -\theta \sum_{i=1}^{N} \bar{\mathbf{u}}_{i,0} |\delta x_i|^2_{\mathbb{R}^d} < 0.$$

Case 2: Assume that $\delta\mathbf{x} = 0$ and $\delta\mathbf{u} \neq 0$. In this situation, we have

$$A(\bar{\mathbf{x}}_0, \bar{\mathbf{u}}_0, \delta\mathbf{x}, \delta\mathbf{u}, 0) = \sum_{i=1}^{N} [\delta\mathbf{u}_i I[0](\bar{x}_{i,0}) + \bar{\mathbf{u}}_{i,0} I'[0](\bar{x}_{i,0})\delta x_i] = \sum_{i=1}^{N} \delta\mathbf{u}_i I[0](\bar{x}_{i,0}).$$

Since the set $\{I[0](\bar{x}_{i,0})\}_{i=1}^N$ is linear independent, $A(\bar{\mathbf{x}}_0, \bar{\mathbf{u}}_0, \delta\mathbf{x}, \delta\mathbf{u}, 0) = 0$ if and only if $\delta\mathbf{u} = 0$. Hence we have

$$\begin{pmatrix} \delta\mathbf{x} \; \delta\mathbf{u} \end{pmatrix} \partial_{(\mathbf{x},\mathbf{u})}G(\bar{\mathbf{x}}_0, \bar{\mathbf{u}}_0, 0) \begin{pmatrix} \delta\mathbf{x} \\ \delta\mathbf{u} \end{pmatrix} = \begin{pmatrix} \delta\mathbf{x} \; \delta\mathbf{u} \end{pmatrix} H_2(\bar{\mathbf{x}}_0, \bar{\mathbf{u}}_0, 0) \begin{pmatrix} \delta\mathbf{x} \\ \delta\mathbf{u} \end{pmatrix} < 0.$$

Combining both statements, we conclude that $\partial_{(\mathbf{x},\mathbf{u})}G(\bar{\mathbf{x}}_0, \bar{\mathbf{u}}_0, 0)$ is invertible, since its kernel is trivial. In virtue of the implicit function theorem, there exist neighborhoods $\bar{N}_V$, $N(\bar{\mathbf{x}}_0)$ and $N(\bar{\mathbf{u}}_0)$ of $0 \in \hat{N}_V$, $\bar{\mathbf{x}}_0 \in X_{ad}$ and $\bar{\mathbf{u}} \in \mathbb{R}^N_+$ respectively as well as $\mathcal{C}^1$ mapping

$$\begin{aligned} \hat{\mathbf{x}}_\cdot &: \bar{N}_V \to N_{X_{ad}}, \quad \mapsto \hat{\mathbf{x}}_\varDelta = (\hat{x}_{1,\varDelta}, \ldots, \hat{x}_{N,\varDelta})^\top, \\ \hat{\mathbf{x}}_\cdot &: \bar{N}_V \to N_{X_{ad}}, \quad \mapsto \hat{\mathbf{u}}_\varDelta = (\hat{\mathbf{u}}_{1,\varDelta}, \ldots, \hat{\mathbf{u}}_{N,\varDelta})^\top, \end{aligned}$$

such that $\hat{\mathbf{u}}_{i,\varDelta} > 0$ for all $i = 1, \ldots, N$ and $\hat{\mathbf{x}}_0 = \bar{\mathbf{x}}_0$, $\hat{\mathbf{u}}_0 = \bar{\mathbf{u}}_0$. Furthermore $(\hat{\mathbf{x}}_\varDelta, \hat{\mathbf{u}}_\varDelta)$ is the unique element in $N(\bar{\mathbf{x}}_0) \times N(\bar{\mathbf{u}}_0)$ fulfilling

$$G(\hat{\mathbf{x}}_\varDelta, \hat{\mathbf{u}}_\varDelta, \varDelta) = 0, \quad \forall \varDelta \in \bar{N}_V. \tag{4.54}$$

Recall that the mappings $\bar{\mathbf{x}}.$ and $\bar{\mathbf{u}}.$ from Lemma 4.38 and Lemma 4.39 are continuous and

$$G(\bar{\mathbf{x}}_\varDelta, \bar{\mathbf{u}}_\varDelta, \varDelta) = 0 \quad \forall \varDelta \in \hat{N}_V.$$

As a consequence, for small perturbations $\varDelta$ we obtain $\bar{\mathbf{x}}_\varDelta \in N(\bar{\mathbf{x}}_0)$ and $\bar{\mathbf{u}}_\varDelta \in N(\bar{\mathbf{u}}_0)$. Thus $\bar{\mathbf{x}}_\varDelta = \hat{\mathbf{x}}_\varDelta$ and $\bar{\mathbf{u}}_\varDelta = \hat{\mathbf{u}}_\varDelta$ for all $\varDelta \in \bar{N}_V$. Thus, the restricitions of $\bar{\mathbf{x}}.$ and $\bar{\mathbf{u}}.$ to $\bar{N}_V$ are at least of class $\mathcal{C}^1$. Without loss of generality we can assume that $\hat{N}_V$ was chosen small enough such that $\hat{N}_V \subset \bar{N}_V$ yielding the statement. The formula for the derivative readily follows by taking the total derivative in (4.54) and applying the chain rule. □

To close this section, we discuss several results which are implications of the previous theorem. Assume that the prerequisites of Theorem 4.41 hold. From the Lipschitz continuity of the optimal sensor positions and the measurement weights we first infer a stability result for the optimal measurement designs in the modified Wasserstein distance.

Proposition 4.42. *There exist constants $c^1_{\|\bar{u}_0\|_\mathcal{M},N}, c^2_{\|\bar{u}_0\|_\mathcal{M},N}$ depending on the norm of $\bar{u}_0$ and N such that*

$$\|\bar{u}_\varDelta - \bar{u}_0\|_{\mathcal{C}^{0,1*}} \leq c^1_{\|\bar{u}_0\|_\mathcal{M},N} \bar{W}_1(\bar{u}_\varDelta, \bar{u}_0) \leq c^2_{\|\bar{u}_0\|_\mathcal{M},N} \|\varDelta\|_V,$$

for all $\varDelta$ small enough.

Proof. The result directly follows from the Lipschitz stability of $\bar{\mathbf{x}}.$ and $\bar{\mathbf{u}}.$, respectively, and applying Proposition 4.19 and Theorem 4.20. □

Second, we provide a Lipschitz stability result for the optimal Fisher information matrix.

Corollary 4.43. *Let $\bar{u}_0$ and $\bar{u}_\varDelta$ be the unique solutions to (P_0) and $(P_\varDelta)$ for $\varDelta \in N_V$, $\varDelta$ small enough. Then there holds*

$$\|\mathcal{I}[0](\bar{u}_0) - \mathcal{I}[\varDelta_1](\bar{u}_\varDelta)\|_{\mathrm{Sym}} \leq c\|\varDelta\|_V,$$

for some $c > 0$ independent of $\varDelta$.

Proof. Let such a $\varDelta$ be given. We start by splitting up the difference as

$$\|\mathcal{I}[0](\bar{u}_0) - \mathcal{I}[\varDelta_1](\bar{u}_\varDelta)\|_{\mathrm{Sym}} \leq \|\mathcal{I}[0](\bar{u}_\varDelta) - \mathcal{I}[\varDelta_1](\bar{u}_\varDelta)\|_{\mathrm{Sym}} + \|\mathcal{I}[0](\bar{u}_0) - \mathcal{I}[0](\bar{u}_\varDelta)\|_{\mathrm{Sym}}.$$

The first term is estimated by

$$\|\mathcal{I}[0](\bar{u}_\varDelta) - \mathcal{I}[\varDelta_1](\bar{u}_\varDelta)\|_{\mathrm{Sym}} \leq \|\varDelta_1\|_{V_1}\|\bar{u}_\varDelta\|_\mathcal{M}.$$

For the second term we expand

$$\|\mathcal{I}[0](\bar{u}_0) - \mathcal{I}[0](\bar{u}_\varDelta)\|_{\mathrm{Sym}} \leq \sum_{i=1}^N \|\bar{\mathbf{u}}_{i,0} I[0](\bar{x}_{i,0}) - \bar{\mathbf{u}}_{i,\varDelta} I[0](\bar{x}_{i,\varDelta})\|_{\mathrm{Sym}}.$$

Since $I[0]$ is at least two times continuously differentiable around $\bar{x}_{i,0}$, $i = 1, \ldots, N$, it is also locally Lipschitz continuous. Hence, for each $i = 1, \ldots N$, we estimate

$$\begin{aligned}&\|\bar{\mathbf{u}}_{i,0} I[0](\bar{x}_{i,0}) - \bar{\mathbf{u}}_{i,\varDelta} I[0](\bar{x}_{i,\varDelta})\|_{\mathrm{Sym}} \\ &\qquad \leq |\bar{\mathbf{u}}_{i,0} - \bar{\mathbf{u}}_{i,\varDelta}|\|I[0](\bar{x}_{i,0})\|_{\mathrm{Sym}} + c|\bar{\mathbf{u}}_{i,\varDelta}||\bar{x}_{i,0} - \bar{x}_{i,\varDelta}|_{\mathbb{R}^d} \leq c(1 + |\bar{\mathbf{u}}_{i,\varDelta}|)\|\varDelta\|_V.\end{aligned}$$

Summing up yields

$$\|\mathcal{I}[0](\bar{u}_0) - \mathcal{I}[0](\bar{u}_\varDelta)\|_{\mathrm{Sym}} \leq c(N + \|\bar{u}_\varDelta\|_\mathcal{M})\|\varDelta\|_V$$

Since $\bar{u}_\varDelta$ is uniformly bounded, combining both estimates yields the statement. □

Finally we define the optimal value function v associated to (P_Δ) as

$$v\colon N_V \to \mathbb{R}, \quad \Delta \mapsto \min_{u\in\mathcal{M}^+(\Omega_o)} F_\Delta(u).$$

Following Theorem 4.32 we conclude that v is Lipschitz stable at 0. Moreover, using the differentiability of $\bar{\mathbf{x}}_\cdot$ and $\bar{\mathbf{u}}_\cdot$, the optimal value function admits a second order Taylor approximation, c.f. also [122, 188].

Proposition 4.44. *For $\hat{\Delta} \in V$ and τ small enough there holds*

$$v(\tau\hat{\Delta}) = v(0) + \tau Dv(0)(\hat{\Delta}) + \frac{\tau^2}{2} D^2 v(0)(\hat{\Delta}, \hat{\Delta}) + o(\tau^2),$$

where

$$Dv(0)(\hat{\Delta}) = \partial_\Delta \boldsymbol{F}(\bar{\boldsymbol{x}}_0, \bar{\mathbf{u}}_0, 0)\hat{\Delta},$$

and

$$\begin{aligned} D^2 v(0)(\hat{\Delta}, \hat{\Delta}) = \partial_\Delta \partial_\Delta \boldsymbol{F}(\bar{\boldsymbol{x}}_0, \bar{\mathbf{u}}_0, 0)(\hat{\Delta}, \hat{\Delta}) \\ + \partial_{\boldsymbol{x}} \partial_\Delta \boldsymbol{F}(\bar{\boldsymbol{x}}_0, \bar{\mathbf{u}}_0, 0)(\hat{\Delta}, \nabla_\Delta \bar{\boldsymbol{x}}_0 \hat{\Delta}) + \partial_{\mathbf{u}} \partial_\Delta \boldsymbol{F}(\bar{\boldsymbol{x}}_0, \bar{\mathbf{u}}_0, 0)(\hat{\Delta}, \nabla_\Delta \bar{\mathbf{u}}_0 \hat{\Delta}). \end{aligned}$$

Proof. Following the previous arguments we obtain

$$v(\tau\hat{\Delta}) = \min_{u\in\mathcal{M}^+(\Omega_o)} F_{\tau\hat{\Delta}}(u) = \min_{\substack{\mathbf{x}=(x_1,\dots,x_N)^\top \in X_{ad}, \\ \mathbf{u}\in\mathbb{R}^N_+}} \mathbf{F}(\mathbf{x}, \mathbf{u}, \tau\hat{\Delta}) = \mathbf{F}(\bar{\mathbf{x}}_{\tau\hat{\Delta}}, \bar{\mathbf{u}}_{\tau\hat{\Delta}}, \tau\hat{\Delta})$$

Differentiating $v(\tau\hat{\Delta})$ with respect to τ and setting $\tau = 0$, we obtain

$$Dv(0)(\hat{\Delta}) = \partial_{\mathbf{x}} \mathbf{F}(\bar{\mathbf{x}}_0, \bar{\mathbf{u}}_0, 0)\nabla_\Delta \bar{\mathbf{x}}_0 \hat{\Delta} + \partial_{\mathbf{u}} \mathbf{F}(\bar{\mathbf{x}}_0, \bar{\mathbf{u}}_0, 0)\nabla_\Delta \bar{\mathbf{u}}_0 \hat{\Delta} + \partial_\Delta \mathbf{F}(\bar{\mathbf{x}}_0, \bar{\mathbf{u}}_0, 0)\hat{\Delta}.$$

Due to the optimality of $(\bar{\mathbf{x}}_0, \bar{\mathbf{u}}_0)$ for $(\mathcal{P}_0^N)$, we have

$$\partial_{\mathbf{x}} \mathbf{F}(\bar{\mathbf{x}}_0, \bar{\mathbf{u}}_0, 0) = 0, \quad \partial_{\mathbf{u}} \mathbf{F}(\bar{\mathbf{x}}_0, \bar{\mathbf{u}}_0, 0) = 0,$$

and consequently

$$Dv(0)(\hat{\Delta}) = \partial_\Delta \mathbf{F}(\bar{\mathbf{x}}_0, \bar{\mathbf{u}}_0, 0)\hat{\Delta}. \tag{4.55}$$

Analogously, the formula for the second derivative $D^2 v(0)(\hat{\Delta}, \hat{\Delta})$ can be established by taking the total derivative in (4.55). The statement now follows directly from Taylor's formula. □

Most remarkably, a first order Taylor approximation of v can be obtained without an evaluation of the sensitivities $\nabla_\Delta \bar{\mathbf{x}}_0$ and $\nabla_\Delta \bar{\mathbf{u}}_0$. Similarly, first order Taylor approximations of the the sensor positions and measurement weights are given through

$$\bar{\mathbf{x}}_{\tau\hat{\Delta}} = \bar{\mathbf{x}}_0 + \tau\nabla_\Delta \bar{\mathbf{x}}_0 \hat{\Delta} + o(\tau), \quad \bar{\mathbf{u}}_{\tau\hat{\Delta}} = \bar{\mathbf{u}}_0 + \tau\nabla_\Delta \bar{\mathbf{u}}_0 \hat{\Delta} + o(\tau).$$

Let us briefly summarize the findings of this section. A first attempt on establishing sensitivity results for sensor placement problems with measure-valued designs was taken. Under mild assumptions, convergence of the perturbed measurement designs and Lipschitz stability of the optimal value function were established. By additional regularity assumptions on the global maximizers of the unperturbed optimal gradient $-\nabla\psi(\bar{u}_0, 0)$, stability of the optimal number of measurements can be proven. Hence the sparse optimization problem $(P_\varDelta)$ can be broken down to a finite-dimensional dimensional problem $(\mathcal{P}^N_\varDelta)$. Exploiting the equivalence between those two problem formulations revealed that the optimal sensor positions as well as the measurement weights depend differentiable on the perturbation. Additionally, this led to Lipschitz stability result for the optimal design measure in a modified Wasserstein distance.

As mentioned in the beginning, literature on sensitivity analysis in the context of optimal sensor placement problems in particular, and for sparse optimization problems in general, seems to be scarce. Hence, the results in this section should be seen as a first step in this direction, leaving room for further investigations. A first natural question is to ask whether the regularity assumptions on the perturbations and on the spatial regularity of the sensitivity vector $\partial S[\varDelta_1]$ can be lowered while maintaining the Lipschitz stability results for the optimal measurement design. Furthermore it would be worth to investigate whether the differentiability of the positions and measurement weights implies some kind of differentiable dependence of the optimal design on the perturbation. Finally, from a practical point of view, the efficient numerical evaluation of the sensitivities $\partial_\varDelta \bar{\mathbf{x}}_0$ and $\partial_\varDelta \bar{\mathbf{u}}_0$ should be the topic of further research. Following the formula in Theorem 4.41, one needs at least matrix-vector products between a given $\delta\mathbf{x}$ and the Hessian of the continuosuly differentiable function $-\nabla_u\psi(\bar{u}_0, 0)$. Thereby we note that the sensitivities $\partial S[\varDelta_1]$ admit no closed form in general, but are replaced by a discrete approximation over a grid on $\varOmega$, see also the following section. In the context of this chapter, a discrete surrogate of the k-th sensitivity is obtained from a piecewiese linear finite element ansatz for the corresponding sensitivity equation. Therefore the optimal discrete gradient $-\nabla_u\psi_h(\bar{u}_{h,0}, 0)$ is given as a sum over products between piecewise linear functions and thus it is especially not of class $\mathcal{C}^2$. Two possible strategies to circumvent these difficulties could consist in either choosing higher order finite elements for the discretization of the sensitivity equations or applying a gradient recovery type algorithm, see e.g. [279], for the derivatives of $-\nabla\psi_h(\bar{u}_{h,0}, 0)$.

4.6 Discretization and error estimates

The aim of this section is twofold: First, we provide an approximation framework for the sparse sensor placement problem based on a finite element discretization of the state as well as the sensitivity equations. In contrast, the space of design measures $\mathcal{M}^+(\varOmega_o)$ is not discretized. This corresponds to the variational discretization approach in optimal control, see e.g [148]. As for the continuous sensor placement problem we prove the existence of discrete optimal designs and provide a discrete version of the first order optimality condition. Finally we prove the convergence of the discrete optimal designs for a vanishing meshsize. For an application of variational discretization in the context of optimal control problems with measure-valued controls we refer to [59, 136, 176].

Second, assuming a suitable second order condition, we derive a priori error estimates for the discretization error between a continuous sparse optimal measurement design and its discrete counterpart. To be more concrete, given an optimal design $\bar{u}_\beta = \sum_{i=1}^N \bar{\mathbf{u}}_i \delta_{\bar{x}_i}$ we will provide convergence rates for the positions $\bar{x}_i$ of the optimal sensors as well as the diligence factors $\bar{\mathbf{u}}_i$. This implies a priori error estimates for the modified Wasserstein distance introduced in Definition 4.1. A priori error estimates for optimal control problems with measure-valued controls and elliptic PDE constraints are for example considered in [176] and [210]. In the latter one the authors prove convergence rates for the control in the norm on the dual space of $H^2(\Omega)$. From Section 4.3 we also recall that an optimal measurement design can be interpreted as Lagrangian multiplier associated to a pointwise imposed constraint in the dual problem, see (4.20). A priori error analysis for the Lagrangian multiplier in this context has been, e.g., conducted in [186] and [190,191]. In the latter ones, convergence rates for the support points of the optimal Lagrange multiplier are provided. Last we recall that optimal control problems involving both, $\|\cdot\|_{L^1(\Omega_o)}$ and $\|\cdot\|_{L^2(\Omega_o)}$, as regularization, admit optimal solutions supported on small sets, see also the discussion in Section 4.4.6. A priori error estimates for these kind of problems are provided in [61,62,208,261]. To our best knowledge this work constitutes the first rigorous approach to the discretization and error analysis of sparse sensor placement problems in the context of optimal design of experiments.

4.6.1 Finite element discretization

In the following, the sets Ω as well as Ω_o are assumed to be polytopal (i.e. polygonal in two dimensions and polyhedral in three dimensions). We discuss the approximation of (P_β) by linear finite elements. For this purpose we consider a family of triangulations $\{\mathcal{T}_h\}_{h>0}$ of Ω with

$$\Omega = \bigcup_{T \in \mathcal{T}_h} \bar{T}, \quad \Omega_o = \bigcup_{T \in \mathcal{T}_h^o} \bar{T}, \tag{4.56}$$

where $\mathcal{T}_h^o \subset \mathcal{T}_h$ denotes the union of all cells making up the observational domain. To each $T \in \mathcal{T}_h$ we assign two numbers $\rho(T)$ and $\sigma(T)$ denoting the diameter of T and the diameter of the largest ball inside of T, respectively. The size of the mesh is defined by $h = \max_{T \in \mathcal{T}_h} \rho(T)$. We assume that the triangulation fulfills the usual regularity conditions (cf., e.g., [62]) , i.e. there exist constants ρ, $\sigma > 0$ such that

$$\frac{\rho(T)}{\sigma(T)} \leq \sigma, \ \frac{h}{\rho(T)} \leq \rho.$$

By $\mathcal{N}_h$ we denote the set of nodes of the triangulation. For each $h > 0$ we now define the space of continuous piecewise linear finite elements V_h on $\mathcal{T}_h$ and its dual space $V_h^* \simeq \mathcal{M}_h$ as

$$V_h = \{\, y_h \in C(\bar{\Omega}) \mid y_{h_{|T}} \in P_1 \ \forall T \in \mathcal{T}_h \,\}, \quad \mathcal{M}_h = \{\, u_h \in \mathcal{M}(\bar{\Omega}) \mid \operatorname{supp} u_h \subset \mathcal{N}_h \,\}.$$

In the following assume that $Y_h = V_h \cap Y$ is not empty. For each $x_i \in \mathcal{N}_h$ we denote by $e_i^h \in V_h$ the associated nodal basis function. Finally, we introduce the nodal interpolation operators $i_h \colon \mathcal{C}(\bar{\Omega}) \to V_h$ and $\Lambda_h \colon \mathcal{M}(\bar{\Omega}) \to \mathcal{M}_h$ as

$$i_h(y) = \sum_{x_i \in \mathcal{N}_h} y(x_i) e_i^h, \quad \Lambda_h(u) = \sum_{x_i \in \mathcal{N}_h} \langle e_i^h, u \rangle \delta_{x_i}$$

see, e.g., [59]. Note that $\Lambda_h u \in \mathcal{M}^+(\Omega_o) \cap \mathcal{M}_h$ for all $u \in \mathcal{M}^+(\Omega_o)$ due to (4.56). We define the discrete state space $\hat{Y}_h = \hat{y}_h + Y_h$ where $\hat{y}_h$ denotes an approximation of the Dirichlet boundary data $\hat{y}$. For a given $q \in Q_{ad}$ the discrete state equation $y^h = S^h[q]$ is defined as

$$y^h \in \hat{Y}_h \ \text{ such that } \ a(q, y^h)(\varphi_h) = 0 \quad \forall \varphi_h \in Y_h. \tag{4.57}$$

Analogously, for all $k \in \{1, \dots, n\}$, the discrete sensitivity $\delta y^h = \partial_k S^h[\hat{q}] \in Y_h \cap C(\Omega_o)$ at the given a priori guess $\hat{q}$ is given as the solution to

$$a'_y(\hat{q}, y^h)(\delta y^h, \varphi_h) = -a'_{q_k}(\hat{q}, \hat{y}^h)(\varphi_h) \quad \forall \varphi_h \in Y_h, \tag{4.58}$$

where $\hat{y}^h = S^h[\hat{q}]$. For the remainder of this section we make the following assumption.

Assumption 4.10. There exists $h_0 > 0$ such that for all $h \le h_0$ and $\hat{q} \in Q_{ad}$ the discrete state and sensitivity equations, (4.57) and (4.58), admit unique solutions. Moreover the discrete sensitivities fulfill

$$\lim_{h \to 0} \max_k \|\partial_k S[\hat{q}] - \partial_k S^h[\hat{q}]\|_{\mathcal{C}} = 0.$$

Note that these assumptions can be verified for a variety of settings, in particular the ones considered in Section 4.7. In the following, $c > 0$ denotes a generic constant which is independent of the meshsize h.

Discretization of (P_β)

We define the discrete approximation to (P_β) by

$$\min_{u \in \mathcal{M}^+(\Omega_o)} F_h(u) = [\psi_h(u) + \beta \|u\|_{\mathcal{M}}], \tag{$P_{\beta,h}$}$$

where $\psi_h(u) = \Psi(\mathcal{I}_h(u) + \mathcal{I}_0)$ and the operator $\mathcal{I}_h$ results from the discretization of the Fisher operator $\mathcal{I}$ as

$$\mathcal{I}_h \colon \mathcal{M}(\Omega_o) \to \mathrm{Sym}(n), \quad \mathcal{I}_h(u_h)_{i,j} = \langle \partial_i S^h[\hat{q}] \partial_j S^h[\hat{q}], u_h \rangle. \tag{4.59}$$

As in the continuous case, given $u \in \mathcal{M}(\Omega_o)$, the Fisher-Information $\mathcal{I}_h(u)$ admits an interpretation as a Bochner integral

$$\mathcal{I}_h(u) = \int_{\Omega_o} \partial S^h[\hat{q}](x) \partial S^h[\hat{q}](x)^\top \mathrm{d}x.$$

Define the discrete pointwise Fisher-information as

$$I_h \colon \Omega_o \to \mathrm{Sym}(n), \quad x \mapsto \partial S^h[\hat{q}](x) \partial S^h[\hat{q}](x)^\top.$$

Initially, we do not discretize the optimal design space $\mathcal{M}^+(\Omega_o)$, which corresponds to a variational discretization approach; cf. [61,148]. However, we will show below that this is essentially equivalent to an additional discretization of the measure space by $\mathcal{M}_h$.

Turning to the study of $(P_{\beta,h})$, we observe that the discrete problem admits admissible points provided that the discrete sensitivities fulfill

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

Due to Assumption 4.10, this property of the discrete problem follows from the analogous property of the continuous problem for h small enough. In the next theorem we prove existence of a discrete optimal design. In addition we show that there exists at least one discrete optimal solution located in the nodes of the triangulation.

Theorem 4.45. *Assume that $\partial S[\hat q]$ fulfills the assumptions from Proposition 4.3 and let Assumption 4.10 hold. Then there exists $h_0 > 0$ such that for every $h \le h_0$ the problem $(P_{\beta,h})$ admits at least one optimal solution $\bar u_{\beta,h} \in \mathcal{M}^+(\Omega_o)$ fulfilling*

$$\nabla\psi_h(\bar u_{\beta,h}) \ge -\beta, \quad \operatorname{supp} \bar u_{\beta,h} \subset \{x \in \Omega_o | -\nabla\psi_h(\bar u_{\beta,h})(x) = \beta\},$$

and $\#\operatorname{supp}\bar u_{\beta,h} \le n(n+1)/2$. Moreover, for every optimal solution $\bar u_{\beta,h}$ of $(P_{\beta,h})$ the interpolated measure $\Lambda_h(\bar u_{\beta,h}) \in \mathcal{M}_h$ is also optimal.

Proof. Let a constant $M_0 > 0$ bounding the norm of continuous optimal designs be given. To show the existence of at least one discrete optimal design we proceed as in the previous section by considering the auxiliary problem

$$\min_{u\in\mathcal{M}^+(\Omega_o)} F_h(u) \quad s.t. \quad \|u\|_{\mathcal{M}} \le 2M_0. \tag{4.60}$$

We have to show that the domain of F_h on $\mathcal{M}^+(\Omega_o)$ is not empty for all h small enough. Existence of at least one minimizer $\bar u_{\beta,h}$ to (4.60) for h small enough then follows immediately. Moreover the sequence $\{\bar u_{\beta,h}\}_{h>0}$ is uniformly bounded and $\|\bar u_{\beta,h}\|_{\mathcal{M}} < 2M_0$ for all h small enough. Consequently $\bar u_{\beta,h}$ is also a minimizer of the unconstrained problem $(P_{\beta,h})$.

By assumption there exists $u \in \mathcal{M}^+(\Omega_o)$, $\|u\|_{\mathcal{M}} \le M_0$ with $\mathcal{I}(u) + \mathcal{I}_0 \in \mathrm{PD}(n)$. Due to the uniform convergence of the sensitivities $\partial S_h[q]$, we have $\mathcal{I}_h(u) \to \mathcal{I}(u)$ for $h \to 0$. Therefore, for h small enough there holds $\mathcal{I}_h(u) + \mathcal{I}_0 \in \mathrm{PD}(n)$, since the set of positive definite matrices is open. Thus $u \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} F_h$ for all h small enough. The necessary and sufficient condition on the gradient as well as the upper bound on the number of support points can be derived as in the continuous case.

It remains to prove the existence of a solution supported in $\mathcal{N}_h$. Given an arbitrary but fixed $u \in \mathcal{M}^+(\Omega_o)$ we have

$$\mathcal{I}_h(\Lambda_h u)_{ik} = \left\langle \partial_i S^h[\hat q]\partial_k S^h[\hat q], \Lambda_h u\right\rangle = \left\langle i_h\left(\partial_i S^h[\hat q]\partial_k S^h[\hat q]\right), u\right\rangle$$

for all $i, k \in \{1, \dots, n\}$, by using properties of Λ_h; see [59, Theorem 3.5]. Let $z \in \mathbb{R}^n$ be arbitrary. Then there holds

$$\begin{aligned} z^T\mathcal{I}_h(u)z &= \left\langle z^\top \partial S^h[\hat q]\partial S^h[\hat q]^\top z, u\right\rangle = \left\langle \left(\partial S^h[\hat q]^\top z\right)^2, u\right\rangle \\ &= \left\langle \left(\sum_{x_j\in\mathcal{N}_h} e_j^h\, \partial S^h[\hat q](x_j)^\top z\right)^2, u\right\rangle. \end{aligned}$$

Now, we estimate

$$\left\langle \left(\sum_{x_j\in\mathcal{N}_h} e_j^h z^\top \partial S^h[\hat q](x_j)\right)^2, u\right\rangle \le \left\langle \sum_{x_j\in\mathcal{N}_h} e_j^h \left(z^\top \partial S^h[\hat q](x_j)\right)^2, u\right\rangle,$$

with Jensen's inequality, using the convexity of the square function and $\sum_{x_i \in \mathcal{N}_h} e_i^h(x) = 1$ for all $x \in \Omega_o$. Expanding and rearranging yields

$$\left\langle \sum_{x_j \in \mathcal{N}_h} e_j^h \left(z^\top \partial S^h[\hat{q}](x_j) \right)^2, u \right\rangle = \left\langle \sum_{x_j \in \mathcal{N}_h} e_j^h \, z^\top \partial S^h[\hat{q}](x_j) \partial S^h[\hat{q}](x_j)^\top z, u \right\rangle$$
$$= \left\langle i_h \left(z^\top \partial S^h[\hat{q}] \partial S^h[\hat{q}]^\top z \right), u \right\rangle = \left\langle z^\top \partial S^h[\hat{q}] \partial S^h[\hat{q}]^\top z, \Lambda_h u \right\rangle = z^T \mathcal{I}_h(\Lambda_h u) z.$$

Since $z \in \mathbb{R}^n$ was arbitrary, this implies $\mathcal{I}_h(u) \leq_L \mathcal{I}_h(\Lambda_h u)$ and therefore also

$$\Psi(\mathcal{I}_h(u) + \mathcal{I}_0) \geq \Psi(\mathcal{I}_h(\Lambda_h u) + \mathcal{I}_0),$$

due to the monotonicity of Ψ with respect to the Löwner ordering. Let $\bar{u}_{\beta,h}$ be an optimal solution of $(P_{\beta,h})$. From this and $\|\Lambda_h \bar{u}_{\beta,h}\|_{\mathcal{M}} \leq \|\bar{u}_{\beta,h}\|_{\mathcal{M}}$ we deduce that $\Lambda_h \bar{u}_{\beta,h}$ is an optimal solution to $(P_{\beta,h})$. □

Note that this result, together with a straightforward adaption of Theorem 4.5 and Proposition 4.7, implies in particular that there exists an optimal solution to $(P_{\beta,h})$ in $\mathcal{M}_h \cap \mathcal{M}(\Omega_o)$ which is comprised of at most $n(n+1)/2$ distinct support points. Finally, we prove subsequential convergence of discrete optimal solutions for $h \to 0$.

Proposition 4.46. *For $h \leq h_0$ denote by $\bar{u}_{\beta,h}$ an arbitrary optimal solution to $(P_{\beta,h})$. There exists at least one subsequence of $\{ \bar{u}_{\beta,h} \}_{h>0}$ (denoted in the same way), converging in the weak* topology for $h \to 0$. Every accumulation point $\bar{u}_\beta$ of $\{ \bar{u}_{\beta,h} \}_{h>0}$ is a minimizer of (P_β) and*

$$\|\bar{u}_{\beta,h}\|_{\mathcal{M}} \to \|\bar{u}_\beta\|_{\mathcal{M}}, \quad \psi_h(\bar{u}_{\beta,h}) \to \psi(\bar{u}_\beta).$$

Furthermore, if there holds

$$\# \operatorname{supp} \bar{u}_{\beta,h} \leq n(n+1)/2, \quad \forall h > 0$$

then the same holds for every accumulation point.

Proof. The sequence $\{ \bar{u}_{\beta,h} \}_{h>0}$ is uniformly bounded by M_0 in h. Thus, there exists a subsequence denoted in the same way and a measure $\bar{u}_\beta \in \mathcal{M}^+(\Omega_o)$ with $\bar{u}_{\beta,h} \rightharpoonup^* \bar{u}_\beta$ for $h \to 0$. Due to the weak* lower semi-continuity of the norm and the uniform convergence of the sensitivities there holds

$$\psi(\bar{u}_\beta) + \beta\|\bar{u}_\beta\|_{\mathcal{M}} \leq \liminf_{h \to 0} [\psi_h(\bar{u}_{\beta,h}) + \beta\|\bar{u}_{\beta,h}\|_{\mathcal{M}}] \leq \psi(\bar{u}) + \beta\|\bar{u}\|_{\mathcal{M}}.$$

Therefore $\bar{u}_\beta$ is also an optimal solution of (P_β) and

$$\psi_h(\bar{u}_{\beta,h}) + \beta\|\bar{u}_{\beta,h}\|_{\mathcal{M}} \to \psi(\bar{u}_\beta) + \beta\|\bar{u}_\beta\|_{\mathcal{M}}.$$

Furthermore, due to the weak* convergence of $\bar{u}_{\beta,h}$, we obtain

$$\|\bar{u}_{\beta,h}\|_{\mathcal{M}} = \langle 1, \bar{u}_{\beta,h} \rangle \to \|\bar{u}_\beta\|_{\mathcal{M}} = \langle 1, \bar{u}_\beta \rangle = \|\bar{u}_{\beta,h}\|_{\mathcal{M}}.$$

The convergence of $\psi_h(\bar{u}_{\beta,h})$ is a direct consequence of the convergence of the objective function values as well as the the convergence of the norms. The result on the number of support points follows from Proposition 6.32, again using that $\dim \operatorname{Sym}(n) = n(n+1)/2$. □

Observe that the different implementations of Algorithms 2 which are presented in Section 4.4.1 can be directly applied to $(P_{\beta,h})$. Following Theorem 4.45 the position $\hat{x}^k$ of the new Dirac delta function can be chosen from $\mathcal{N}_h$. Therefore step 2. in Algorithm 2 amounts to the computation of the discrete gradient $\nabla\psi_h(u^k)$ and the determination of its maximum in $\mathcal{N}_h$. The latter one can be done efficiently by $O(\#\mathcal{N}_h)$ operations.

Post-processing of the discrete design measure

By Theorem 4.5 the support of an optimal design $\bar{u}_\beta$ can be limited to $n(n+1)/2$ points. In practice, this upper bound is often rather pessimistic. However, due to discretization error, the support of a discrete solution $\bar{u}_{\beta,h} \in \mathcal{M}_h$ of $(P_{\beta,h})$ can be bigger than that of the continuous counterpart $\bar{u}_\beta$, while still respecting the upper bound $n(n+1)/2$. Usually, a sensor at a specific location in the continuous solution appears spread out over several adjacent grid points in the numerical solution. A similar effect has been observed and theoretically investigated in the context of sparse deconvolution in the presence of noise; cf. [95]. As a remedy, we employ the following heuristic post-processing of the discrete solution: First, we cluster the support of $\bar{u}_{\beta,h}$ into $N_c \leq \#\operatorname{supp}\bar{u}_{\beta,h}$ sets $S_i \subset \Omega_o$, with $\operatorname{diam}(S_i) \leq Ch$. Then, we construct a new design $\bar{u}^S = \sum_{i=1,\dots,N_c} \mathbf{u}_i^S \delta_{x_i^S}$ with $\mathbf{u}_i^S = \int_{S_i} \mathrm{d}\bar{u}_{\beta,h}$ summing up the coefficients of each cluster, and $x_i^S = \int_{S_i} x\mathrm{d}\bar{u}_{\beta,h}/\mathbf{u}_i^S$ the locations by the center of mass. Note that this introduces an additional error in the location of the support points of order h, which is not worse than what we can expect from $\bar{u}_h$. Additionally, the weak*-convergence result for $h \to 0$ from Proposition 4.46 is not affected by this post-processing.

Discretization of (P_β^ε)

We briefly comment on the discretization of the regularized sub-problems (P_β^ε). We adapt the approach from [61, 208] and discretize the design by piece-wise linear finite elements on the observation set, denoted by U_h. We endow this space with the lumped inner product defined for any $\varphi, \psi \in U_h \subset \mathcal{C}(\Omega_o)$ in the usual way as

$$(\varphi, \psi)_{\Omega_o,h} = \int_{\Omega_o} i_h(\varphi\psi)(x)\mathrm{d}x.$$

The approximation of (P_β^ε) is then defined as

$$\min_{u_h \in U_h, u_h \geq 0} \left[\psi_h(\Lambda_h u_h) + \beta\|u_h\|_{L^1(\Omega_o)} + \frac{\varepsilon}{2}\|u_h\|^2_{L^2(\Omega_o),h}\right], \qquad (P_{h,\beta}^\varepsilon)$$

where $\|u_h\|^2_{L^2(\Omega_o),h} = (u_h, u_h)_{\Omega_o,h}$ is the lumped regularization term. Here, the appearance of $\Lambda_h\omega_h$ turns integrals involving the finite element function u_h into appropriate lumped integrals, i.e., we obtain

$$\mathcal{I}_h(\Lambda_h u_h)_{ij} = (\partial_i S^h[\hat{q}]\partial_j S^h[\hat{q}], u_h)_h.$$

Note also that $\|u_h\|_{L^1(\Omega_o)} = \|u_h\|_{\mathcal{M}} = \|\Lambda_h u_h\|_{\mathcal{M}}$. The existence of an optimal solution to $(P_{h,\beta}^\varepsilon)$, for h small enough, can be shown by similar arguments as for the unregularized discrete problem. Additionally uniqueness of the solution follows using the strict convexity of the regularization term.

The necessary and sufficient optimality conditions can be derived in a straightforward manner and are equivalent to the point-wise projection formula

$$\bar{u}^{\varepsilon}_{\beta,h}(x_i) = \max\left\{ -\frac{1}{\varepsilon}(\nabla\psi_{l,h}(\bar{u}^{\varepsilon}_{\beta,h})(x_i) + \beta), 0 \right\} \quad \forall x_i \in \mathcal{N}_h \cap \Omega_o, \tag{4.61}$$

where $\psi_{l,h}(\bar{u}^{\varepsilon}_{\beta,h}) = \psi_h(\Lambda_h \bar{u}^{\varepsilon}_{\beta,h})$. For a discussion and comparison of different discretization schemes of the regularized problem we refer to [208, Section 4.5.3].

4.6.2 A priori error estimates

This section is devoted to the derivation of a priori error estimates for the sparse sensor placement problem. Therefore we strengthen our assumptions on the design criterion Ψ and the convergence of the discrete sensitivities.

Assumption 4.11. There exists a positive, strict monotonically increasing and continuous function $\gamma\colon \mathbb{R}_+ \to \mathbb{R}_+$ with $\lim_{h\to+0}\gamma(h) = 0$ and

$$\max_k \|\partial_k S[\hat{q}] - \partial_k S^h[\hat{q}]\|_{\mathcal{C}} \leq \gamma(h),$$

for all $h \leq h_0$. Furthermore, Ψ is strictly convex on its domain and there exists $\gamma_0 > 0$ with

$$\operatorname{Tr}(B\nabla^2\Psi(\mathcal{I}(\bar{u}_\beta) + \mathcal{I}_0)B) \geq \gamma_0 \|B\|^2_{\mathrm{Sym}}, \quad \forall B \in \mathrm{Sym}(n). \tag{4.62}$$

Here $\mathcal{I}(\bar{u}_\beta)$ denotes the unique Fisher-information matrix.

Note that the continuous optimal Fisher information $\mathcal{I}(\bar{u}_\beta)$ and $\nabla\psi(\bar{u}_\beta)$ as well as their discrete counterparts $\mathcal{I}_h(\bar{u}_{\beta,h})$ and $\nabla\psi_h(\bar{u}_{\beta,h})$ are unique due to the strict convexity of Ψ. We briefly recall that (4.62) implies uniform convexity of Ψ in a neighbourhood $N(\mathcal{I}(\bar{u}_\beta))$ of the optimal Fisher-information $\mathcal{I}(\bar{u}_\beta)$, i.e

$$(\nabla\Psi(B_1 + \mathcal{I}_0) - \nabla\Psi(B_2 + \mathcal{I}_0), B_1 - B_2)_{\mathrm{Sym}} \geq \frac{\gamma_0}{2}\|B_1 - B_2\|^2_{\mathrm{Sym}} \quad \forall B_1, B_2 \in N(\mathcal{I}(\bar{u}_\beta)),$$

see Corollary 4.17. Additionally, since Ψ is two-times continuously Fréchet differentiable on its domain, the gradient $\nabla\Psi\colon \operatorname{dom}\Psi \to \mathrm{Sym}(n)$ is Lipschitz continuous on compact sets, i.e. given a compact set $M \subset \operatorname{dom}\Psi$ there exists a constant $L_M > 0$ with

$$\|\nabla\Psi(B_1) - \nabla\Psi(B_2)\|_{\mathrm{Sym}} \leq L_M \|B_1 - B_2\|_{\mathrm{Sym}} \quad \forall B_1, B_2 \in M.$$

Error estimates for the objective function

Let us first collect some perturbation results for the Fisher information $\mathcal{I}$ and the optimal design criterion ψ.

Lemma 4.47. *There exists a constant $c > 0$ such that for all h small enough we have:*

- $\max_{x\in\Omega_o} \|I(x) - I_h(x)\|_{\mathrm{Sym}} + \|\mathcal{I} - \mathcal{I}_h\|_{\mathcal{L}(\mathcal{M}(\Omega_o),\mathrm{Sym}(n))} \leq c\gamma(h)$.
- *For all $B_1, B_2 \in \mathrm{Sym}(n)$ there holds*

$$\|\mathcal{I}^* B_1 - \mathcal{I}^*_h B_2\|_{\mathcal{C}} \leq c(\|B_1\|_{\mathrm{Sym}}\gamma(h) + \|B_1 - B_2\|_{\mathrm{Sym}}).$$

- *For all $u_1, u_2 \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi \cap \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi_h$ we have*

$$\begin{aligned}&\|\nabla\psi(u_1) - \nabla\psi_h(u_2)\|_{\mathcal{C}}\\ &\qquad\le c(\|\nabla\Psi(\mathcal{I}(u_1)+\mathcal{I}_0)\|_{\mathrm{Sym}}\gamma(h) + \|\nabla\Psi(\mathcal{I}(u_1)+\mathcal{I}_0) - \nabla\Psi(\mathcal{I}_h(u_2)+\mathcal{I}_0)\|_{\mathrm{Sym}}).\end{aligned}$$

Proof. Let $x \in \Omega_o$ be given. We calculate

$$\begin{aligned}\|I(x) - I_h(x)\|_{\mathrm{Sym}}^2 &= \operatorname{Tr}((I(x)-I_h(x))^\top (I(x)-I_h(x)))\\ &= \sum_{i=1}^n \sum_{j=1}^n (\partial_i S[\hat{q}](x)\partial_j S[\hat{q}](x) - \partial_i S^h[\hat{q}](x)\partial_j S^h[\hat{q}](x))^2.\end{aligned}$$

For $i, j = 1, \ldots, n$ we estimate

$$\begin{aligned}&|\partial_i S[\hat{q}](x)\partial_j S[\hat{q}](x) - \partial_i S^h[\hat{q}](x)\partial_j S^h[\hat{q}](x)|\\ &\qquad\le \|\partial_i S[\hat{q}]\|_{\mathcal{C}}\|\partial_j S[\hat{q}] - \partial_j S^h[\hat{q}]\|_{\mathcal{C}} + \|\partial_j S^h[\hat{q}]\|_{\mathcal{C}}\|\partial_i S[\hat{q}] - \partial_i S^h[\hat{q}]\|_{\mathcal{C}}.\end{aligned}$$

Due to the uniform convergence of the sensitivities we conclude

$$\max_{x\in\Omega_o} \|I(x) - I_h(x)\|_{\mathrm{Sym}}^2 \le c\gamma(h)^2.$$

Furthermore, let an arbitrary $u \in \mathcal{M}(\Omega_o)$ be given. Using the properties of the Bochner integral there holds

$$\|\mathcal{I}(u) - \mathcal{I}_h(u)\|_{\mathrm{Sym}} \le \max_{x\in\Omega_o} \|I(x) - I_h(x)\|_{\mathrm{Sym}}\|u\|_{\mathcal{M}} \le \gamma(h)\|u\|_{\mathcal{M}}.$$

The first statement now follows by taking the supremum over all $u \in \mathcal{M}(\Omega_o)$. Next, let $B_1,\ B_2 \in \mathrm{Sym}(n)$ be given. We obtain

$$\|\mathcal{I}^* B_1 - \mathcal{I}_h^* B_2\|_{\mathcal{C}} \le \|\mathcal{I}^* B_1 - \mathcal{I}_h^* B_1\|_{\mathcal{C}} + \|\mathcal{I}_h^* B_1 - \mathcal{I}_h^* B_2\|_{\mathcal{C}}.$$

Using $\|\mathcal{I}^* - \mathcal{I}_h^*\|_{\mathcal{L}(\mathrm{Sym}(n),\mathcal{C}(\Omega_o))} = \|\mathcal{I} - \mathcal{I}_h\|_{\mathcal{L}(\mathcal{M}(\Omega_o),\mathrm{Sym}(n))}$ there holds

$$\|\mathcal{I}^* B_1 - \mathcal{I}_h^* B_1\|_{\mathcal{C}} \le \|\mathcal{I} - \mathcal{I}_h\|_{\mathcal{L}(\mathcal{M}(\Omega_o),\mathrm{Sym}(n))}\|B_1\|_{\mathrm{Sym}} \le c\gamma(h)\|B_1\|_{\mathrm{Sym}}.$$

In the same way we conclude

$$\|\mathcal{I}_h^* B_1 - \mathcal{I}_h^* B_2\|_{\mathcal{C}} \le \|\mathcal{I}_h\|_{\mathcal{L}(\mathcal{M}(\Omega_o),\mathrm{Sym}(n))}\|B_1 - B_2\|_{\mathrm{Sym}} \le c\|B_1 - B_2\|_{\mathrm{Sym}},$$

since $\|\mathcal{I}_h\|_{\mathcal{L}(\mathcal{M}(\Omega_o),\mathrm{Sym}(n))}$ is uniformly bounded as $h \to 0$. Combining both estimates yields the second statement. The final statement follows directly, noting that

$$\|\nabla\psi(u_1) - \nabla\psi_h(u_2)\|_{\mathcal{C}} = \|\mathcal{I}^*\nabla\Psi(\mathcal{I}(u_1)+\mathcal{I}_0) - \mathcal{I}_h^*\nabla\Psi(\mathcal{I}_h(u_2)+\mathcal{I}_0)\|_{\mathcal{C}}.$$

□

Lemma 4.48. *Let a sequence $\{u_h\}_{h>0} \subset \mathcal{M}^+(\Omega_o)$ with $u_h \rightharpoonup^* u \in \mathcal{M}^+(\Omega_o)$ as $h \to 0$. Assume that $u, u_h \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi \cap \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi_h$ for h small enough. For all $h \le h_0$ small enough there holds*

$$|\psi_h(u_h) - \psi(u_h)| \le c_u\gamma(h)\|u_h\|_{\mathcal{M}},$$

with some constant $c_u > 0$ depending on $u \in \mathcal{M}^+(\Omega_o)$.

Proof. Let such a sequence be given. By Taylor's expansion we have

$$\psi_h(u_h) - \psi(u_h) = \operatorname{Tr}(\nabla\Psi(\mathcal{I}_{\zeta_h}(u_h) + \mathcal{I}_0)^\top(\mathcal{I}(u_h) - \mathcal{I}_h(u_)))$$

where $\mathcal{I}_{\zeta_h}(u_h) = \mathcal{I}(u_h) + \zeta(\mathcal{I}_h(u_h) - \mathcal{I}(u_h))$ for some $\zeta_h \in (0,1)$. Obviously there holds

$$\|\mathcal{I}(u) - \mathcal{I}_{\zeta_h}(u_h)\|_{\mathrm{Sym}} \leq \|\mathcal{I}(u) - \mathcal{I}(u_h)\|_{\mathrm{Sym}} + c\gamma(h)\|u_h\|_{\mathcal{M}} \to 0,$$

for some constant $c > 0$ independent of h and ζ as $h \to 0$. By using Lemma 4.47 we obtain

$$\begin{aligned}\operatorname{Tr}(\nabla\Psi(\mathcal{I}_{\zeta_h}(u_h) + \mathcal{I}_0)^\top(\mathcal{I}(u_h) - \mathcal{I}_h(u_h))) &\leq \|\nabla\Psi(\mathcal{I}_{\zeta_h}(u_h) + \mathcal{I}_0)\|_{\mathrm{Sym}}\|\mathcal{I}(u_h) - \mathcal{I}_h(u_h)\|_{\mathrm{Sym}} \\ &\leq \|\nabla\Psi(\mathcal{I}_{\zeta_h}(u_h) + \mathcal{I}_0)\|_{\mathrm{Sym}}\gamma(h)\|u_h\|_{\mathcal{M}}.\end{aligned}$$

Since $\nabla\Psi\colon \operatorname{dom}\Psi \to \mathrm{Sym}(n)$ is continuous the norm $\|\nabla\Psi(\mathcal{I}_{\zeta_h}(u_h) + \mathcal{I}_0)\|_{\mathrm{Sym}}$ stays bounded. The statement now readily follows. □

Since we do not discretize the set of admissible designs we conclude the following convergence result for the optimal objective function values.

Theorem 4.49. *Let arbitrary solutions $\bar{u}_\beta$ to (P_β) and $\bar{u}_{\beta,h}$ to $(P_{\beta,h})$,respectively, be given. Then there exists a constant $c > 0$ with*

$$|F_h(\bar{u}_{\beta,h}) - F(\bar{u}_\beta)| \leq c\gamma(h). \tag{4.63}$$

Proof. Let a continuous optimal design $\bar{u}_\beta$ as well as a discrete one $\bar{u}_{\beta,h}$ be given. Then there holds $\bar{u}_{\beta,h} \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)}\psi$ and $\bar{u}_\beta \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)}\psi_h$ due to the convergence of the Fisher information matrices. Exploiting optimality we obtain

$$F_h(\bar{u}_{\beta,h}) - F(\bar{u}_{\beta,h}) \leq F_h(\bar{u}_{\beta,h}) - F(\bar{u}_\beta) \leq F_h(\bar{u}_\beta) - F(\bar{u}_\beta).$$

Consequently we conclude

$$|F_h(\bar{u}_{\beta,h}) - F(\bar{u})| \leq \max\{|F_h(\bar{u}_{\beta,h}) - F(\bar{u}_{\beta,h})|, |F_h(\bar{u}_\beta) - F(\bar{u}_\beta)|\}.$$

Note that $F_h(u) - F(u) = \psi_h(u) - \psi(u)$ for all $u \in \mathcal{M}^+(\Omega_o)$. Using Lemma 4.48 we arrive at

$$\max\{|F_h(\bar{u}_{\beta,h}) - F(\bar{u}_{\beta,h})|, |F_h(\bar{u}_\beta) - F(\bar{u}_\beta)|\} \leq c\gamma(h)\max\{\|\bar{u}_\beta\|_{\mathcal{M}}, \|\bar{u}_{\beta,h}\|_{\mathcal{M}}\}.$$

Due to the weak* convergence of the optimal designs, $\|\bar{u}_{\beta,h}\|_{\mathcal{M}}$ is uniformly bounded. The statement now readily follows. □

Additionally, the strict convexity of Ψ implies the following quadratic growth behavior.

Proposition 4.50. *Let an optimal design $\bar{u}_\beta$ be given. For every $u \in \mathcal{M}^+(\Omega_o)$ with $u \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)}\psi$ we have*

$$\frac{\gamma_0}{2}\|\mathcal{I}(\bar{u}_\beta) - \mathcal{I}(u)\|^2_{\mathrm{Sym}} \leq F(u) - F(\bar{u}_\beta).$$

Proof. Let $u \in \mathcal{M}^+(\Omega_o)$ with $u \in \operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi$ be given. Due to the the coercivity assumptions on the Hessian we have

$$F(u) - F(\bar{u}_\beta) \geq (\nabla\Psi(\mathcal{I}(\bar{u}_\beta) + \mathcal{I}_0), \mathcal{I}(u) - \mathcal{I}(\bar{u}_\beta))_{\mathrm{Sym}} + \beta\|u\|_{\mathcal{M}} - \beta\|\bar{u}_\beta\|_{\mathcal{M}} + \frac{\gamma_0}{2}\|\mathcal{I}(u) - \mathcal{I}(\bar{u}_\beta)\|^2_{\mathrm{Sym}}.$$

Due to the optimality of $\bar{u}_\beta$ we further obtain

$$\begin{aligned}\langle \nabla\psi(\bar{u}_\beta), u - \bar{u}_\beta\rangle &+ \beta\|u\|_{\mathcal{M}} - \beta\|\bar{u}_\beta\|_{\mathcal{M}} \\ &= (\nabla\Psi(\mathcal{I}(\bar{u}_\beta) + \mathcal{I}_0), \mathcal{I}(u) - \mathcal{I}(\bar{u}_\beta))_{\mathrm{Sym}} + \beta\|u\|_{\mathcal{M}} - \beta\|\bar{u}_\beta\|_{\mathcal{M}} \geq 0.\end{aligned}$$

This proves the claimed result. □

Combining the previous statements we arrive at the following convergence result for the Fisher information matrix $\mathcal{I}(\bar{u}_\beta)$ and the optimal gradient $\nabla\psi(\bar{u}_\beta)$.

Corollary 4.51. *For all h small enough there holds*

$$\|\mathcal{I}(\bar{u}_\beta) - \mathcal{I}_h(\bar{u}_{\beta,h})\|_{\mathrm{Sym}} + \|\nabla\Psi(\mathcal{I}(\bar{u}_\beta) + \mathcal{I}_0) - \nabla\Psi(\mathcal{I}_h(\bar{u}_{\beta,h}) + \mathcal{I}_0)\|_{\mathrm{Sym}} \leq c\sqrt{\gamma(h)},$$

as well as

$$\|\nabla\psi(\bar{u}_\beta) - \nabla\psi_h(\bar{u}_{\beta,h})\|_{\mathcal{C}} \leq c\sqrt{\gamma(h)},$$

for some constant $c > 0$.

Proof. Let a continuous optimal design $\bar{u}_\beta$ as well as a discrete one $\bar{u}_{\beta,h}$ be given. We split up the error as

$$\|\mathcal{I}(\bar{u}_\beta) - \mathcal{I}_h(\bar{u}_{\beta,h})\|_{\mathrm{Sym}} \leq \|\mathcal{I}(\bar{u}_{\beta,h}) - \mathcal{I}_h(\bar{u}_{\beta,h})\|_{\mathrm{Sym}} + \|\mathcal{I}(\bar{u}_\beta) - \mathcal{I}(\bar{u}_{\beta,h})\|_{\mathrm{Sym}}.$$

The first term can be estimated by

$$\|\mathcal{I}(\bar{u}_{\beta,h}) - \mathcal{I}_h(\bar{u}_{\beta,h})\|_{\mathrm{Sym}} \leq \|\mathcal{I} - \mathcal{I}_h\|_{\mathcal{L}(\mathcal{M}(\Omega_o),\mathrm{Sym}(n))}\|\bar{u}_{\beta,h}\|_{\mathcal{M}} \leq c\gamma(h)\|\bar{u}_{\beta,h}\|_{\mathcal{M}}.$$

Furthermore, for an arbitrary discrete optimal design $\bar{u}_{\beta,h}$ we have

$$\|\mathcal{I}(\bar{u}_\beta) - \mathcal{I}(\bar{u}_{\beta,h})\|_{\mathrm{Sym}} \leq \|\mathcal{I}(\bar{u}_\beta) - \mathcal{I}_h(\bar{u}_{\beta,h})\|_{\mathrm{Sym}} + c\gamma(h)\|\bar{u}_{\beta,h}\|_{\mathcal{M}}.$$

Therefore, we conclude $\mathcal{I}(\bar{u}_{\beta,h}) \in N(\mathcal{I}(\bar{u}_\beta))$ for all h small enough and all discrete optimal designs $\bar{u}_{\beta,h}$. Thus there holds,

$$\frac{\gamma_0}{2}\|\mathcal{I}(\bar{u}_\beta) - \mathcal{I}(\bar{u}_{\beta,h})\|^2_{\mathrm{Sym}} \leq F(\bar{u}_{\beta,h}) - F(\bar{u}_\beta) = F(\bar{u}_{\beta,h}) - F_h(\bar{u}_{\beta,h}) + F_h(\bar{u}_{\beta,h}) - F(\bar{u}_\beta).$$

Consequently

$$\|\mathcal{I}(\bar{u}_\beta) - \mathcal{I}(\bar{u}_{\beta,h})\|^2_{\mathrm{Sym}} \leq c(\gamma(h) + \gamma(h)\|\bar{u}_{\beta,h}\|_{\mathcal{M}}), \tag{4.64}$$

using Theorem 4.49 and Lemma 4.47. Combining both estimates and taking the square root yields the first statement due to the uniform boundedness of $\|\bar{u}_{\beta,h}\|_{\mathcal{M}}$. The remaining results are now obtained from the Lischitz continuity of $\nabla\Psi$ on compact sets and Lemma 4.47. □

Error estimates for the optimal design measure

In the following we will derive a priori error estimates for the optimal sensors. We impose the following assumption on the set of global maximizers to $-\nabla\psi(\bar{u})$ and on the smoothness of $\partial S[\hat{q}]$ in its vicinity.

Assumption 4.12. Assume that the interior of Ω_o is non-empty and there exist $\bar{x}_i \in \operatorname{int}\Omega_o$, $i = 1, \dots, N$, such that the set $\{\mathcal{I}(\delta_{\bar{x}_i})\}_{i=1}^N$ is linear independent and

$$\operatorname{supp}\bar{u}_\beta \subset \{\, x \in \Omega_o \mid -\nabla\psi(\bar{u}_\beta)(x) = \beta \,\} = \{\bar{x}_i\}_{i=1}^N.$$

Furthermore there exists $R > 0$ with

$$\Omega_R := \bigcup_{i=1}^N B_R(\bar{x}_i) \subset \operatorname{int}\Omega_o, \quad \bar{B}_R(\bar{x}_i) \cap \bar{B}_R(\bar{x}_j) = \emptyset, \quad \partial_i S[\hat{q}] \in \mathcal{C}^2(\bar{\Omega}_R),$$

for all $i, j = 1, \dots, N$, $i \neq j$.

From this additional assumption we immediately derive that the adjoint of the Fisher information maps continuously to (locally) smooth functions, i.e.

$$\mathcal{I}^*\colon \operatorname{Sym}(n) \to \mathcal{C}(\Omega_o) \cap \mathcal{C}^2(\bar{\Omega}_R) \quad A \mapsto \partial S[\hat{q}]^\top A \partial S[\hat{q}],$$

is linear and continuous. Secondly, due to the linear independence assumption, the optimal design $\bar{u}_\beta$ is unique. For abbreviation we define the continuous functions $\bar{p} \in \mathcal{C}(\Omega_o) \cap \mathcal{C}^2(\bar{\Omega}_R)$ and, for every $h \leq h_0$, $\bar{p}_h \in \mathcal{C}(\Omega_o)$ as

$$\bar{p}\colon \Omega_o \to \mathbb{R} \quad x \mapsto -\nabla\psi(\bar{u}_\beta)(x), \quad \bar{p}_h\colon \Omega_o \to \mathbb{R}, \quad x \mapsto -\nabla\psi_h(\bar{u}_{\beta,h})(x),$$

respectively. For the rest of this section we will denote the gradient and the Hessian of $\bar{p}$ by $\nabla\bar{p}$ and $\nabla^2\bar{p}$ respectively. Note that due to the to the optimality of $\bar{u}_\beta \neq 0$ we have

$$\bar{p}(x) \leq \beta \quad \forall x \in \Omega_o, \quad \bar{p}(\bar{x}_i) = \beta, \quad \nabla\bar{p}(\bar{x}_i) = 0,$$

since $\bar{x}_i \in \operatorname{int}\Omega_o$, $i = 1, \dots, N$. To derive error estimates for the position of the sensors we further impose assumptions on the curvature of $\bar{p}$ in the support points, see also Section 4.4 and Section 4.5. For convenience of the reader we restate them.

Assumption 4.13. Let $\bar{u}_\beta$ be the unique optimal solution to (P_β). Assume that $\bar{u}_\beta = \sum_{i=1}^N \bar{u}_i \delta_{\bar{x}_i}$ for some $\bar{u}_i > 0$, $i = 1, \dots, N$ and there exists $\theta > 0$ with

$$-(\zeta, \nabla^2\bar{p}(\bar{x}_i)\zeta)_{\mathbb{R}^d} \geq \theta|\zeta|_{\mathbb{R}^d} \quad \forall\zeta \in \mathbb{R}^d,$$

for all $i = 1, \dots, N$.

Based on this assumption, we conclude the following quadratic grow condition for the optimal gradient $\bar{p}$.

Proposition 4.52. *There exists* $0 < R_1 \leq R$ *with*

$$\bar{p}(x) \leq \beta - \frac{\theta}{4}|x - \bar{x}_i|^2_{\mathbb{R}^d} \quad \forall x \in \bigcup_{i=1}^{N} \bar{B}_{R_1}(\bar{x}_i), \tag{4.65}$$

for all $i = 1, \ldots, N$. *Moreover there exists* $\sigma > 0$ *with*

$$\bar{p}(x) \leq \beta - \sigma \quad \forall x \in \Omega_o \setminus \bigcup_{i=1}^{N} B_{R_1}(\bar{x}_i). \tag{4.66}$$

Proof. Let us fix $i \in \{1, \ldots, N\}$. For $x \in B_R(\bar{x}_i)$ we apply Taylor's expansion to obtain

$$\bar{p}(x) = \bar{p}(\bar{x}_i) + (\nabla \bar{p}(\bar{x}_i), x - \bar{x}_i)_{\mathbb{R}^d} + \frac{1}{2}(x - \bar{x}_i, \nabla^2 \bar{p}(x_\zeta)(x - \bar{x}_i))_{\mathbb{R}^d}$$

where $x_\zeta = \bar{x}_i - \zeta(x - \bar{x}_i)$ for some $0 < \zeta < 1$. Since $\bar{x}_i$ is an optimal sensor position we have $\bar{p}(\bar{x}_i) = \beta$ and $\nabla \bar{p}(\bar{x}_i) = 0$ respectively. We proceed by estimating the second order term as

$$\begin{aligned}
&\big(x - \bar{x}_i, \nabla^2 \bar{p}(x_\zeta)(x - \bar{x}_i)\big)_{\mathbb{R}^d} \\
&\qquad = \big(x - \bar{x}_i, \nabla^2 \bar{p}(\bar{x}_i)(x - \bar{x}_i)\big)_{\mathbb{R}^d} + \big(x - \bar{x}_i, \nabla^2 \bar{p}(x_\zeta) - \nabla^2 \bar{p}(\bar{x}_i)(x - \bar{x}_i)\big)_{\mathbb{R}^d} \\
&\qquad\qquad \leq \big(\|\nabla^2 \bar{p}(x_\zeta) - \nabla^2 \bar{p}(\bar{x}_i)\|_{\mathbb{R}^{d\times d}} - \theta\big)\, |x - \bar{x}_i|^2_{\mathbb{R}^d}.
\end{aligned}$$

Due to the continuity of $\nabla^2 \bar{p}$ on $B_R(\bar{x}_i)$ there exists $R > R_i > 0$ with

$$|x - \bar{x}_i|_{\mathbb{R}^d} \leq R_i \Rightarrow \|\nabla^2 \bar{p}(x) - \nabla^2 \bar{p}(\bar{x}_i)\|_{\mathbb{R}^{d\times d}} \leq \frac{\theta}{2}.$$

Noting that $|x_\zeta - \bar{x}_i|_{\mathbb{R}^d} \leq |x - \bar{x}_i|_{\mathbb{R}^d}$ we conclude

$$\bar{p}(x) \leq \beta - \frac{\theta}{4}|x - \bar{x}_i|^2_{\mathbb{R}^d} \quad \forall x \in \bar{B}_{R_i}(\bar{x}_i).$$

Since $\bar{p}$ admits its global maximum only in finitely many points $\bar{x}_i$, $i = 1, \ldots, N$ we choose R_1 as the maximum over the R_i, $i = 1, \ldots N$. This gives (4.65). The existence of $\sigma > 0$ such that (4.66) holds follows due to the continuity of $\bar{p}$ and $\bar{p}(\bar{x}_i) = \beta$, $i = 1, \ldots, N$ as well as $\bar{p}(x) < \beta$ for $x \in \Omega_o \setminus \bigcup_{i=1}^{N} \{\bar{x}_i\}$. □

In the next corollary we localize the support of a discrete optimal design $\bar{u}_{\beta,h}$ in the vicinity of the continuous optimal sensor positions $\bar{x}_i$, $i = 1, \ldots, N$.

Corollary 4.53. *For all* h *small enough there holds*

$$\bar{p}_h(x) \leq \beta - \frac{\sigma}{2} \quad \forall x \in \Omega_o \setminus \bigcup_{i=1}^{N} B_{R_1}(\bar{x}_i). \tag{4.67}$$

Furthermore, given an arbitrary discrete optimal design $\bar{u}_{\beta,h}$ *we have* $\operatorname{supp} \bar{u}_{\beta,h} \subset \bigcup_{i=1}^{N} \bar{B}_{R_1}(\bar{x}_i)$.

Proof. Let $x \in \Omega_o \setminus \bigcup_{i=1}^{N} B_{R_1}(\bar{x}_i)$ be given. We estimate

$$\bar{p}_h(x) = \bar{p}(x) + \bar{p}_h(x) - \bar{p}(x) \leq \beta - \sigma + \|\bar{p}_h(x) - \bar{p}\|_{\mathcal{C}}.$$

For all h small enough we have $\|\bar{p}_h(x) - \bar{p}\|_{\mathcal{C}} \leq \sigma/2$, see Corollary 4.51. This gives (4.67). Let now an arbitrary optimal solution $\bar{u}_{\beta,h}$ to $(P_{\beta,h})$ be given. If $\bar{x}_h \in \operatorname{supp} \bar{u}_{\beta,h}$, then there holds $\bar{p}_h = \beta$, see Theorem 4.45. Following (4.67), this is only possible if $\bar{x}_h \in \bar{B}_{R_1}(\bar{x}_i)$ for some $i \in \{1, \ldots, N\}$. □

In the following we will also make use of the auxiliary function $\tilde{p} \in \mathcal{C}(\Omega_o) \cap \mathcal{C}^2(\bar{\Omega}_R)$ defined as

$$\tilde{p} \colon \Omega \to \mathbb{R} \quad x \mapsto -\mathcal{I}^* \nabla \Psi(\mathcal{I}_h(\bar{u}_{\beta,h}) + \mathcal{I}_0)(x)$$

Corollary 4.54. *For all h small enough there holds*

$$\|\bar{p}_h - \tilde{p}\|_{\mathcal{C}} + \|\tilde{p} - \bar{p}\|^2_{\mathcal{C}^2(\Omega_R)} \leq c\gamma(h),$$

for some $c > 0$.

Proof. We directly obtain

$$\|\bar{p}_h - \tilde{p}\|_{\mathcal{C}} \leq c\|\mathcal{I}_h - \mathcal{I}\|_{\mathcal{L}(\mathcal{M}(\Omega_o),\mathrm{Sym}(n))} \|\nabla\Psi(\mathcal{I}_h(\bar{u}_{\beta,h}) + \mathcal{I}_0)\|_{\mathrm{Sym}}$$

as well as

$$\|\tilde{p} - \bar{p}\|_{\mathcal{C}^2(\Omega_R)} \leq c\|\mathcal{I}^*\|_{\mathcal{L}(\mathrm{Sym}(n),\mathcal{C}^2(\Omega_R))} \|\nabla\Psi(\mathcal{I}(\bar{u}_\beta) + \mathcal{I}_0) - \nabla\Psi(\mathcal{I}_h(\bar{u}_{\beta,h}) + \mathcal{I}_0)\|_{\mathrm{Sym}}.$$

Following Lemma 4.47 and Corollary 4.51 we have

$$\|\nabla\Psi(\mathcal{I}(\bar{u}_\beta) + \mathcal{I}_0) - \nabla\Psi(\mathcal{I}_h(\bar{u}_{\beta,h}) + \mathcal{I}_0)\|^2_{\mathrm{Sym}} + \|\mathcal{I}_h - \mathcal{I}\|_{\mathcal{L}(\mathcal{M}(\Omega_o),\mathrm{Sym}(n))} \leq c\gamma(h).$$

Hence the desired estimates directly follow since $\|\nabla\Psi(\mathcal{I}_h(\bar{u}_{\beta,h}) + \mathcal{I}_0)\|_{\mathrm{Sym}}$ is uniformly bounded. □

As for $\bar{p}$ the gradient and the Hessian of $\tilde{p}$ with respect to the spatial variable will be denoted by $\nabla\tilde{p}$ and $\nabla^2\tilde{p}$ respectively. After these preliminary preparations we are now able to derive a first intermediate estimate for the distances between continuous and discrete optimal sensor positions respectively.

Lemma 4.55. *Let h be small enough and let $\bar{x}_h$ with $\bar{p}_h(\bar{x}_h) = \beta$ be given. Then there exists an index $i \in \{1, \ldots, N\}$ with*

$$|\bar{x}_h - \bar{x}_i|_{\mathbb{R}^d} \leq c\sqrt[4]{\gamma(h)},$$

for some $c > 0$. Given a discrete optimal design $\bar{u}_{\beta,h}$ we get

$$\max_{i=1,\ldots N} \max_{x \in \operatorname{supp} \bar{u}_{\beta,h} \cap B_{R_1}(\bar{x}_i)} |x - \bar{x}_i|_{\mathbb{R}^d} \leq c\sqrt[4]{\gamma(h)}, \tag{4.68}$$

for some constant $c > 0$ independent of $\bar{u}_{\beta,h}$.

Proof. Let such a $\bar{x}_h$ be given. Then we have $\bar{x}_h \in B_{R_1}(\bar{x}_i)$ for some index $i \in \{1, \ldots, N\}$ due to Corollary 4.53. We estimate

$$\beta = \bar{p}_h(\bar{x}_h) = \bar{p}(\bar{x}_h) + \bar{p}_h(\bar{x}_h) - \bar{p}(\bar{x}_h) \leq \bar{p}(\bar{x}_h) + c\sqrt{\gamma(h)} \leq \beta - \frac{\theta}{4}|\bar{x}_h - \bar{x}_i|^2_{\mathbb{R}^d} + \sqrt{\gamma(h)}.$$

Rearranging and taking the square root yields

$$|\bar{x}_h - \bar{x}_i| \leq c\sqrt[4]{\gamma(h)}.$$

This implies the first assertion. The statement in (4.68) readily follows from $\bar{p}_h(x) = \beta$ for all $x \in \operatorname{supp} \bar{u}_{\beta,h}$. □

In the following we improve the error estimate for the optimal sensor positions. To do so, we proceed similarly to [191] and derive auxiliary results for the growth behaviour of the function $\tilde{p}$ in a neighbourhood of $\bar{x}_i$, $i = 1, \dots, N$. These are summarized in the following two lemmas.

Lemma 4.56. *For each $i = 1, \dots, N$ the function $\tilde{p}$ admits a unique local maximum $\tilde{x}_i^h$ on $B_{R_1}(\bar{x}_i)$. Moreover there holds*

$$|\bar{x}_i - \tilde{x}_i^h|_{\mathbb{R}^d} \leq c\sqrt{\gamma(h)},$$

for some constant $c > 0$.

Proof. By Assumption 4.11 the pointwise Fisher information

$$I \colon \Omega_o \to \operatorname{Sym}(n) \quad x \mapsto \partial S[\hat{q}](x) \partial S[\hat{q}](x)^\top,$$

is two times continuously differentiable on Ω_R with Fréchet derivatives $I' \in \mathcal{C}^1(\Omega_R, \mathcal{L}(\mathbb{R}^d, \operatorname{Sym}(n)))$ and $I'' \in \mathcal{C}(\Omega_R, \mathcal{L}(\mathbb{R}^d, \mathcal{L}(\mathbb{R}^d, \operatorname{Sym}(n)))$. Given $A \in \operatorname{Sym}(n)$ we consider the continuous function

$$\nabla p[A] \colon \Omega_R \to \mathbb{R}^d \quad \nabla p[A](x)_i = -\operatorname{Tr}(A^\top (I'(x)e_i)),$$

for $x \in \Omega_o$. Here $e_i \in \mathbb{R}^d$ denotes the i-th canonical basis vector of $\mathbb{R}^d$, $i = 1, \dots, d$. Note that $p[A]$ is continuously differentiable on Ω_R for every $A \in \operatorname{Sym}(n)$ and

$$\nabla p\left[\nabla \Psi(\mathcal{I}(\bar{u}_\beta) + \mathcal{I}_0)\right] = \nabla \bar{p}, \quad \nabla p\left[\nabla \Psi(\mathcal{I}_h(\bar{u}_{\beta,h}) + \mathcal{I}_0)\right] = \nabla \tilde{p},$$

for all $h > 0$ small enough. Define now the function $P \in \mathcal{C}^1(\Omega_R \times \operatorname{Sym}(n), \mathbb{R}^d)$ as

$$P \colon \Omega_R \times \operatorname{Sym}(n) \to \mathbb{R}^d, \quad (x, A) \mapsto \nabla p[A](x).$$

Fix an arbitrary index $i \in \{1, \dots, N\}$ and denote by $\nabla_x P(x, A) \in \operatorname{Sym}(n)$ the partial derivative of P at (x, A) with respect to x. Then there holds

$$P\left(\bar{x}_i, \nabla \Psi(\mathcal{I}(\bar{u}_\beta) + \mathcal{I}_0)\right) = \nabla \bar{p}(\bar{x}_i) = 0, \quad \nabla_x P\left(\bar{x}_i, \nabla \Psi(\mathcal{I}(\bar{u}_\beta) + \mathcal{I}_0)\right) = \nabla^2 \bar{p}(\bar{x}_i).$$

By assumption, the Hessian of $\bar{p}$ at $\bar{x}_i$ is positive definite and thus $\nabla_x P\left(\bar{x}_i, \nabla \Psi(\mathcal{I}(\bar{u}_\beta) + \mathcal{I}_0\right)$ is invertible. From the implicit function theorem we get $R_i, c_i, \rho > 0$ such that for every $A \in \operatorname{Sym}(n)$ with $\|\nabla \Psi(\mathcal{I}(\bar{u}_\beta) - A\|_{\operatorname{Sym}} \leq \rho$ there exists a unique $\tilde{x}_i(A) \in B_{R_i}(\bar{x}_i)$ with $P(\tilde{x}_i, A) = 0$ and

$$|\tilde{x}_i(A) - \bar{x}_i|_{\mathbb{R}^d} \leq c_i \|\nabla \Psi(\mathcal{I}(\bar{u}_\beta) + \mathcal{I}_0) - A\|_{\operatorname{Sym}}.$$

We apply this result to $A = \nabla \Psi(\mathcal{I}_h(\bar{u}_{\beta,h}) + \mathcal{I}_0)$ to obtain the existence of $\tilde{x}_i^h := \tilde{x}_i(A)$ with $P(\tilde{x}_i^h, A) = \nabla \tilde{p}(\tilde{x}_i^h) = 0$ and

$$|\tilde{x}_i^h - \bar{x}_i|_{\mathbb{R}^d} \leq c\|\nabla \Psi(\mathcal{I}(\bar{u}_\beta) + \mathcal{I}_0) - \nabla \Psi(\mathcal{I}_h(\bar{u}_{\beta,h}) + \mathcal{I}_0)\|_{\operatorname{Sym}} \leq c\sqrt{\gamma(h)}.$$

Hence we have $\tilde{x}_i^h \in B_{R_1}(\bar{x}_i)$ for all h small enough. It remains to show that $\tilde{x}_i^h$ is a local maximum of $\tilde{p}$. For $x \in \Omega_R$ we estimate

$$\begin{aligned}
-\left(\zeta, \nabla^2 \tilde{p}(x)\zeta\right)_{\mathbb{R}^d} &= -\left(\zeta, \nabla^2 \bar{p}(x)\zeta\right)_{\mathbb{R}^d} - \left(\zeta, (\nabla^2 \tilde{p}(x) + \nabla^2 \bar{p}(x))\zeta\right)_{\mathbb{R}^d} \\
&\geq -\left(\zeta, \nabla^2 \bar{p}(\bar{x}_i)\zeta\right)_{\mathbb{R}^d} - (\|\nabla^2 \bar{p}(x) - \nabla^2 \bar{p}(\bar{x}_i)\|_{\operatorname{Sym}} + c\sqrt{\gamma(h)})|\zeta|^2_{\mathbb{R}^d} \\
&\geq (\theta - \|\nabla^2 \bar{p}(x) - \nabla^2 \bar{p}(\bar{x}_i)\|_{\operatorname{Sym}} - c\sqrt{\gamma(h)})|\zeta|^2_{\mathbb{R}^d}.
\end{aligned}$$

for all $\zeta \in \mathbb{R}^d$. W.l.o.g the radius R_1 can be chosen small enough such that for all $x \in \Omega_R$ we have

$$|x - \bar{x}_i|_{\mathbb{R}^d} \leq R_1 \Rightarrow \|\nabla^2 \bar{p}(x) - \nabla^2 \bar{p}(\bar{x}_i)\|_{\mathrm{Sym}} \leq \frac{\theta}{2},$$

due to the continuity of $\nabla^2 \bar{p}$. Thus we conclude

$$-\left(\zeta, \nabla^2 \tilde{p}(x)\zeta\right)_{\mathbb{R}^d} \geq \left(\frac{\theta}{2} - c\sqrt{\gamma(h)}\right) |\zeta|^2_{\mathbb{R}^d} \geq \frac{\theta}{4}|\zeta|^2_{\mathbb{R}^d}, \tag{4.69}$$

for all $x \in B_{R_1}(\bar{x}_i)$, all $\zeta \in \mathbb{R}^d$ and all h small enough. Therefore $\tilde{p}$ is strictly concave on $B_{R_1}(\bar{x}_i)$ and the Hessian $\nabla^2 \tilde{p}(\tilde{x}^h_i)$ is negative definite. Consequently, $\tilde{p}$ admits a strict local maximum at $\tilde{x}^h_i$ which is unique on $B_{R_1}(\bar{x}_i)$. Since i was chosen arbitrary and there are only finitely many $\bar{x}_i$ all constants can be considered as uniformly bounded in i. This gives the statement. □

Lemma 4.57. *There exist* $0 < R_2 \leq R_1$ *and a constant* $c > 0$ *such that*

- $\tilde{p}(x) \leq \beta + c\gamma(h) \quad \forall x \in \Omega_o.$
- $\tilde{p}(x) \leq \beta - \frac{\sigma}{2} \quad \forall x \in \Omega_o \backslash \bigcup_{i=1}^N B_{R_1}(\bar{x}_i).$
- $\tilde{p}(x) \leq \tilde{p}(\tilde{x}^h_i) - \frac{\theta}{8}|x - \tilde{x}^h_i|^2_{\mathbb{R}^d} \quad \forall x \in B_{R_2}(\tilde{x}^h_i).$

for all h *small enough and* $i = 1, \dots, N$.

Proof. Let $x \in \Omega_o$ be given. We immediately obtain

$$\tilde{p}(x) = \bar{p}_h(x) + \tilde{p}(x) - \bar{p}_h(x) \leq \beta + \|\tilde{p} - \bar{p}_h\|_{\mathcal{C}} \leq \beta + c\gamma(h),$$

see Corollary 4.54. This proves the first result. The second statement readily follows due to the uniform convergence of $\tilde{p}$ towards $\bar{p}$. Concerning the third claim we observe that $\tilde{x}^h_i \in \operatorname{int} \Omega_o$ for all h small enough and thus $\nabla \tilde{p}(\tilde{x}^h_i) = 0$ for all $i = 1, \dots, N$. Fix an arbitrary index $i \in \{1, \dots, N\}$. Again by applying Taylor's expansion we deduce

$$\tilde{p}(x) = \tilde{p}(\tilde{x}^h_i) + \frac{1}{2}\left(x - \tilde{x}^h_i, \nabla^2 \tilde{p}(x_\zeta)(x - \tilde{x}^h_i)\right)_{\mathbb{R}^d},$$

where $x_\zeta = \tilde{x}^h_i + \zeta(x - \tilde{x}^h_i)$ for some $0 < \zeta < 1$. Observe that

$$|x_\zeta - \bar{x}_i|_{\mathbb{R}^d} \leq |x - \tilde{x}^h_i|_{\mathbb{R}^d} + |\bar{x}_i - \tilde{x}^h_i|_{\mathbb{R}^d} \leq c\sqrt{\gamma(h)} + |x - \tilde{x}^h_i|_{\mathbb{R}^d}$$

Hence by choosing $R_2 = R_1/2$ we have $x_\zeta \in B_{R_1}(\bar{x}_i)$ for all $x \in B_{R_2}(\tilde{x}^h_i)$ and all h small enough. From the concavity of $\tilde{p}$ on $B_{R_1}(\bar{x}_i)$, see (4.69), we conclude

$$\tilde{p}(x) = \tilde{p}(\tilde{x}^h_i) + \frac{1}{2}\left(x - \tilde{x}^h_i, \nabla^2 \tilde{p}(x_\zeta)(x - \tilde{x}^h_i)\right)_{\mathbb{R}^d} \leq \tilde{p}(\tilde{x}^h_i) - \frac{\theta}{8}|x - \tilde{x}^h_i|^2_{\mathbb{R}^d} \quad \forall x \in B_{R_2}(\tilde{x}^h_i).$$

As before, since the index i was chosen arbitrary, all estimates can be derived with constants uniformly bounded in i. □

Using these additional results on $\tilde{p}$ we can now improve on the a priori estimate for the support points $\bar{x}_i$ derived in Lemma 4.55.

Theorem 4.58. *Let Assumptions 4.11, 4.12 and 4.13 hold. For all h small enough and every discrete solution $\bar{u}_{\beta,h}$ to $(P_{\beta,h})$ we have*

$$\max_{i=1,\dots,N} \max_{x\in \operatorname{supp} \bar{u}_{\beta,h}\cap B_{R_1}(\bar{x}_i)} |x-\bar{x}_i|_{\mathbb{R}^d} \leq c\sqrt{\gamma(h)} \tag{4.70}$$

for some $c>0$ independent of $\bar{u}_{\beta,h}$ and h.

Proof. Let an arbitrary discrete optimal design $\bar{u}_{\beta,h}$ be given and fix an index $i\in\{1,\dots,N\}$. W.l.o.g assume that $\operatorname{supp}\bar{u}_{\beta,h}\cap B_{R_1}(\bar{x}_i)\neq\emptyset$. For $x\in\operatorname{supp}\bar{u}_{\beta,h}\cap B_{R_1}(\bar{x}_i)$ we readily obtain

$$|x-\tilde{x}_i^h|_{\mathbb{R}^d} \leq |\tilde{x}_i^h-\bar{x}_i|_{\mathbb{R}^d} + |x-\bar{x}_i|_{\mathbb{R}^d} \leq c(\sqrt{\gamma(h)}+\sqrt[4]{\gamma(h)}),$$

from Lemma 4.56 and Lemma 4.55. Consequently, if h is chosen small enough, there holds $x\in B_{R_2}(\tilde{x}_i^h)$. Using Corollary 4.54 we furthermore observe that there holds

$$\beta=\bar{p}_h(x)\leq\tilde{p}(x)+c\gamma(h).$$

In virtue of Lemma 4.57 we conclude

$$\beta-c\gamma(h)\leq\tilde{p}(x)\leq\tilde{p}(\tilde{x}_i^h)-\frac{\theta}{8}|x-\tilde{x}_i^h|^2_{\mathbb{R}^d},$$

and thus, by rearranging

$$|x-\tilde{x}_i^h|^2_{\mathbb{R}^d}\leq\tilde{p}(\tilde{x}_i^h)-\beta+c\gamma(h)\leq\|\tilde{p}-\bar{p}_h\|_{\mathcal{C}}+c\gamma(h)+\bar{p}_h(\tilde{x}_i^h)-\beta\leq c\gamma(h).$$

Here, the last inequality is obtained by applying Corollary 4.54 and $\bar{p}^h(x)-\beta\leq 0$ for all $x\in\Omega_o$ and all h small enough. Since all derived estimates do not depend on the chosen point x, we can maximize both sides of the inequality with respect to $x\in\operatorname{supp}\bar{u}_{\beta,h}\cap B_{R_1}(\bar{x}_i)$ and $i\in\{1,\dots,N\}$. Taking the square root concludes the proof. □

Based on the improved convergence rate for the optimal positions of the measurement sensors, we proceed to prove an a priori error estimate for the diligence factors $\bar{\mathbf{u}}_i$. In the following, given $\bar{u}_{\beta,h}$, its restriction to $\bar{B}_{R_1}(\bar{x}_i)$ will be denoted by $\bar{u}^i_{\beta,h}$, $i=1,\dots,N$. First, note that, up to now, we have not discussed whether there is a discrete optimal sensor in a neighborhood of a continuous one. Mathematically this reduces to the question whether $\operatorname{supp}\bar{u}_{\beta,h}\cap B_{R_1}(\bar{x}_i)\neq\emptyset$ for a given discrete optimal design $\bar{u}_{\beta,h}$ and $i\in\{1,\dots,N\}$. This issue is discussed in the following lemma.

Lemma 4.59. *Consider a sequence $\{\bar{u}_{\beta,h}\}_{h>0}$ of optimal solutions to $(P_{\beta,h})$. For all h small enough we have*

$$\operatorname{supp}\bar{u}_{\beta,h}\subset\bigcup_{i=1}^N\bar{B}_{R_1}(\bar{x}_i),\quad \operatorname{supp}\bar{u}_{\beta,h}\cap\bar{B}_{R_1}(\bar{x}_i)\neq\emptyset,$$

as well as $\|\bar{u}^i_{\beta,h_k}\|_{\mathcal{M}}\to\bar{\mathbf{u}}_i$, for all $i=1,\dots N$.

Proof. The localization result on the support of $\bar{u}_{\beta,h}$ readily follows from Corollary 4.53. Fix an arbitrary index $i\in\{1,\dots,N\}$. Using Urysohn's Lemma, there exists $\varphi_i\in\mathcal{C}(\Omega_o)$ with

$$\varphi_i(x)=1\quad\forall x\in\bar{B}_{R_1}(\bar{x}_i),\quad\varphi_i(x)=0\quad\forall x\in\bigcup_{j=1,j\neq i}^N\bar{B}_{R_1}(\bar{x}_j).$$

Due to the weak* convergence of $\bar{u}_{\beta,h}$ towards $\bar{u}_\beta$ we have

$$\|\bar{u}^i_{\beta,h}\|_{\mathcal{M}} = \langle \varphi_i, \bar{u}_{\beta,h}\rangle \to \langle \varphi_i, \bar{u}_\beta\rangle = \bar{\mathbf{u}}_i,$$

as $h \to 0$. Since $\bar{\mathbf{u}}_i > 0$ we thus obtain that $\operatorname{supp} \bar{u}^i_{\beta,h} \neq \emptyset$ for all h small enough. This concludes the proof. □

While the last lemma ensures the existence of a discrete optimal sensor in the neighborhood of a continuous one it does not make a statement on the number of approximating points. In fact, an optimal sensor at $\bar{x}_i$ in the continuous measurement design might be approximated by a larger number of discrete ones clustering in $\bar{B}_{R_1}(\bar{x}_i)$. Hence, the error between $\bar{\mathbf{u}}_i$ and the norm of $\bar{u}^i_{\beta,h}$ should be quantified. Furthermore, recall that for a given discrete design $\bar{u}_{\beta,h}$ the interpolated design measure $\Lambda_h \bar{u}_{\beta,h} \in \mathcal{M}_h$ is also optimal, c.f Theorem 4.45. From now on, we assume that $\bar{u}_{\beta,h} \in \mathcal{M}_h$ for all $h \leq h_0$.

Let us introduce the operator $\hat{\mathcal{I}}$ and the vector of measurement weights $\bar{\mathbf{u}} \in \mathbb{R}^N_+ \setminus \{0\}$ as

$$\hat{\mathcal{I}} \colon \mathbb{R}^N \to \operatorname{Sym}(n), \quad \mathbf{u} \mapsto \sum_{i=1}^N \mathbf{u}_i I(\bar{x}_i), \quad \bar{\mathbf{u}} = (\bar{\mathbf{u}}_1, \dots, \bar{\mathbf{u}}_N)^\top,$$

respectively. From the improved a priori error estimate for the optimal sensor positions we conclude the following perturbation result.

Lemma 4.60. *Let a sequence $\{\bar{u}_{\beta,h}\}_{h>0} \subset \mathcal{M}_h$ of discrete optimal designs be given. For $h > 0$ define the weight vector $\bar{\mathbf{u}}^h = (\|\bar{u}^1_{\beta,h}\|_{\mathcal{M}}, \dots, \|\bar{u}^N_{\beta,h}\|_{\mathcal{M}})^\top$. For all h small enough there holds*

$$\|\hat{\mathcal{I}}(\bar{\mathbf{u}}^h) - \mathcal{I}(\bar{u}_{\beta,h})\|_{\mathrm{Sym}} \leq c_N \|\bar{u}_{\beta,h}\|_{\mathcal{M}} \sqrt{\gamma(h)},$$

for some $c_N > 0$ depending on the support size of the continuous optimal design.

Proof. We decompose $\bar{u}_{\beta,h}$ to obtain

$$\|\hat{\mathcal{I}}(\bar{\mathbf{u}}^h) - \mathcal{I}(\bar{u}_{\beta,h})\|_{\mathrm{Sym}} \leq \sum_{i=1}^N \|\bar{\mathbf{u}}^h_i I(\bar{x}_i) - \mathcal{I}(\bar{u}^i_{\beta,h})\|_{\mathrm{Sym}}.$$

Fix an arbitrary index $i \in \{1, \dots, N\}$. By assumption, there exists $N^{h,i} \in \mathbb{N} \setminus \{0\}$ with

$$\bar{u}^i_{\beta,h} = \sum_{j=1}^{N^{h,i}} \bar{\mathbf{u}}^h_{j,i} \delta_{\bar{x}^h_{i,j}}, \quad \bar{x}^h_{i,j} \in \mathcal{N}_h, \quad \bar{\mathbf{u}}^h_{j,i} \in \mathbb{R}^+ \setminus \{0\},$$

for $j = 1, \dots, N^{h,i}$. Due to optimality and Theorem 4.58 we have

$$\bar{p}(\bar{x}^h_{i,j}) = \beta, \quad |\bar{x}^h_{i,j} - \bar{x}_i|^2_{\mathbb{R}^d} \leq c\sqrt{\gamma(h)}.$$

From the regularity assumptions on $\partial S[\hat{q}]$ we conclude that the mapping

$$I \colon \Omega_o \to \operatorname{Sym}(n), \quad x \mapsto \partial S[\hat{q}](x) \partial S[\hat{q}](x)^\top,$$

is Lipschitz continuous around $\bar{x}_i$. Using $\|\bar{u}^i_{\beta,h}\|_{\mathcal{M}} = \sum_{j=1}^{N^{h,i}} \bar{\mathbf{u}}^h_{j,i}$ we estimate

$$\|\bar{\mathbf{u}}^h_i I(\bar{x}_i) - \mathcal{I}(\bar{u}^i_{\beta,h})\|_{\mathrm{Sym}} \leq \sum_{j=1}^{N^{h,i}} \bar{\mathbf{u}}^h_{j,i} \|I(\bar{x}_i) - I(\bar{x}^h_{i,j})\|_{\mathrm{Sym}} \leq c\|\bar{u}^i_{\beta,h}\|_{\mathcal{M}} |\bar{x}_i - \bar{x}^h_{i,j}|_{\mathbb{R}^d}$$

Since i was arbitrary and $\|\bar{u}_{\beta,h}\|_{\mathcal{M}} = \sum_{i=1}^N \|\bar{u}^i_{\beta,h}\|_{\mathcal{M}}$ we obtain

$$\|\hat{\mathcal{I}}(\bar{\mathbf{u}}^h) - \mathcal{I}(\bar{u}_{\beta,h})\|_{\mathrm{Sym}} \leq cN\|\bar{u}_{\beta,h}\|_{\mathcal{M}} \max_{i=1,\ldots,N} \max_{x \in \operatorname{supp} \bar{u}_{\beta,h} \cap B_{R_1}(\bar{x}_i)} |x - \bar{x}_i| \leq c\|\bar{u}_{\beta,h}\|_{\mathcal{M}} \sqrt{\gamma(h)}.$$

This finishes the proof. □

We are now ready to prove an a priori estimate for the discretization error between the measurement weight $\bar{\mathbf{u}}_i$ and the sum of the discrete optimal measurement weights corresponding to sensors in $\bar{B}_{R_1}(\bar{x}_i)$.

Theorem 4.61. *Let Assumptions 4.11, 4.12 and 4.13 hold. Let a sequence $\{\bar{u}_{\beta,h}\}_{h>0} \subset \mathcal{M}_h$ of discrete optimal designs be given. If h is small enough we have*

$$\sum_{i=1}^N |\bar{\mathbf{u}}_i - \|\bar{u}^i_{\beta,h}\|_{\mathcal{M}}| \leq c\sqrt{\gamma(h)}, \tag{4.71}$$

for some $c > 0$.

Proof. First note that since $\{\mathcal{I}(\delta_{\bar{x}_i})\}_{i=1}^N$ is linear independent, the operator $\hat{\mathcal{I}}$ has full rank and $\hat{\mathcal{I}}^*\hat{\mathcal{I}} \in \mathrm{Sym}(N)$ is invertible. We estimate

$$\begin{aligned}\sum_{i=1}^N |\bar{\mathbf{u}}_i - \|\bar{u}^i_{\beta,h}\|_{\mathcal{M}}| &\leq c|\bar{\mathbf{u}} - \bar{\mathbf{u}}^h|_{\mathbb{R}^N} \leq c\|(\hat{\mathcal{I}}^*\hat{\mathcal{I}})^{-1}\|_{\mathbb{R}^{d\times d}} |\hat{\mathcal{I}}^*\hat{\mathcal{I}}(\bar{\mathbf{u}} - \bar{\mathbf{u}}^h)|_{\mathbb{R}^N} \\ &\leq c\|(\hat{\mathcal{I}}^*\hat{\mathcal{I}})^{-1}\|_{\mathbb{R}^{d\times d}} \|\hat{\mathcal{I}}^*\|_{\mathcal{L}(\mathrm{Sym}(n),\mathbb{R}^N)} \|\hat{\mathcal{I}}(\bar{\mathbf{u}} - \bar{\mathbf{u}}^h)\|_{\mathrm{Sym}}.\end{aligned}$$

By construction we have $\hat{\mathcal{I}}(\bar{\mathbf{u}}) = \mathcal{I}(\bar{u}_\beta)$ and thus

$$\|\hat{\mathcal{I}}(\bar{\mathbf{u}} - \bar{\mathbf{u}}^h)\|_{\mathrm{Sym}} = \|\mathcal{I}(\bar{u}_\beta) - \hat{\mathcal{I}}(\bar{\mathbf{u}}^h)\|_{\mathrm{Sym}} \leq \|\mathcal{I}(\bar{u}_\beta) - \mathcal{I}(\bar{u}_{\beta,h})\|_{\mathrm{Sym}} + \|\hat{\mathcal{I}}(\bar{\mathbf{u}}^h) - \mathcal{I}(\bar{u}_{\beta,h})\|_{\mathrm{Sym}}.$$

From (4.64), Lemma 4.60 and the boundedness of $\|\bar{u}_{\beta,h}\|_{\mathcal{M}}$ we obtain

$$\|\mathcal{I}(\bar{u}_\beta) - \mathcal{I}(\bar{u}_{\beta,h})\|_{\mathrm{Sym}} + \|\hat{\mathcal{I}}(\bar{\mathbf{u}}^h) - \mathcal{I}(\bar{u}_{\beta,h})\|_{\mathrm{Sym}} \leq c\sqrt{\gamma(h)}.$$

□

Due to the sparsity of $\bar{u}_\beta$ and $\bar{u}_{\beta,h}$ for all $h \leq h_0$ we also derive an error estimate for the optimal measurement design in the modified Wasserstein distance as well as for the norm in $\mathcal{C}^{0,1}(\Omega_o)^*$, see Section 4.4.3.

Theorem 4.62. *Let Assumptions 4.11, 4.12 and 4.13 hold. Let a sequence $\{\bar{u}_{\beta,h}\}_{h>0} \subset \mathcal{M}_h$ of discrete optimal designs be given. For all h small enough we have*

$$\|\bar{u}_{\beta,h} - \bar{u}_\beta\|_{\mathcal{C}^{0,1*}} \leq c^1_{\|\bar{u}_\beta\|_{\mathcal{M}},N} \bar{W}_1(\bar{u}_{\beta,h}, \bar{u}) \leq c^2_{\|\bar{u}_\beta\|_{\mathcal{M}},N} \sqrt{\gamma(h)},$$

where the constants $c^1_{\|\bar{u}_\beta\|_{\mathcal{M}},N}$, $c^2_{\|\bar{u}_\beta\|_{\mathcal{M}},N}$ depend on the norm of $\bar{u}_\beta$ as well as its support size N.

Proof. The statement readily follows from applying Proposition 4.19, Theorem 4.20 and the a priori error estimates in Theorem 4.58 and Theorem 4.61. □

4.7 Numerical examples

We end this chapter with a numerical study of two test examples. In the following we consider the unit square $\bar{\Omega} = [0,1]^2$ and a sequence $\mathcal{T}_{h_k}$, $k \in \{1,2,\dots,9\}$, of uniform triangulations of $\bar{\Omega}$ with $h_k = \sqrt{2}/2^k$. 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 different 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. Since the small number of Dirac delta functions in these examples aid the practical performance of Algorithm 3 we postpone a comparison between sequential point insertion algorithms and the path-following strategy to the following chapter. In all examples we consider the A-optimal design problem, i.e. $\Psi = \operatorname{Tr}(\cdot + \mathcal{I}_0)^{-1}$ and the discrete state and the associated sensitivities $\partial S[\hat{q}]$ are computed for a fixed $\hat{q}$ once at the beginning. During the execution of the different variants of Algorithms 2 and 6 no additional PDEs need to be solved. Moreover, the gradient of the reduced cost functional is given by

$$[\nabla\psi(u)](x) = -\operatorname{Tr}((\mathcal{I}(u)+\mathcal{I}_0)^{-1}\mathcal{I}(\delta_x)(\mathcal{I}(u)+\mathcal{I}_0)^{-1}) = -\|(\mathcal{I}(u)+\mathcal{I}_0)^{-1}\partial S[\hat{q}](x)\|^2_{\mathbb{R}^n} \quad \forall x \in \Omega_o$$

which relates the pointwise value of the gradient directly to the corresponding sensitivity vector $\partial S[\hat{q}](x) \in \mathbb{R}^n$. A corresponding computation on the discrete level allows for an efficient implementation based on a single Cholesky-decomposition of $\mathcal{I}(u) + \mathcal{I}_0$ in each iteration. Moreover, a corresponding expression for the Hessian-vector-product $\left[\nabla^2\psi(u)(\delta u)\right](x)$ for $\delta u \in \mathcal{M}(\Omega_o)$ can be derived by differentiating the above expression. In both examples, the assumptions on the continuous and discrete state equation, see Assumption 4.1 and Assumption 4.10, respectively, can be easily verified.

4.7.1 Estimation of diffusion and convection coefficients

As a first example for the state equation (4.3), we take a convection-diffusion process where for a given $q \in Q_{ad} = \{\, q \in \mathbb{R}^3 \mid 5 \geq q_1 \geq 0.25 \,\}$ the associated state $y = S[q] \in H^1_0(\Omega) \cap \mathcal{C}(\Omega_o)$ is the unique solution to

$$a(q,y)(\varphi) = \int_\Omega \left[q_1 \nabla y \cdot \nabla\varphi + q_2\varphi\frac{\partial y}{\partial x_1} + q_3\varphi\frac{\partial y}{\partial x_2} \right] \mathrm{d}x = \int_\Omega f\varphi \, \mathrm{d}x, \tag{4.72}$$

for all $\varphi \in H^1_0(\Omega)$. The forcing term f is chosen as $\exp(3(x_1^2 + x_2^3))$. This corresponds to the linear elliptic equation

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

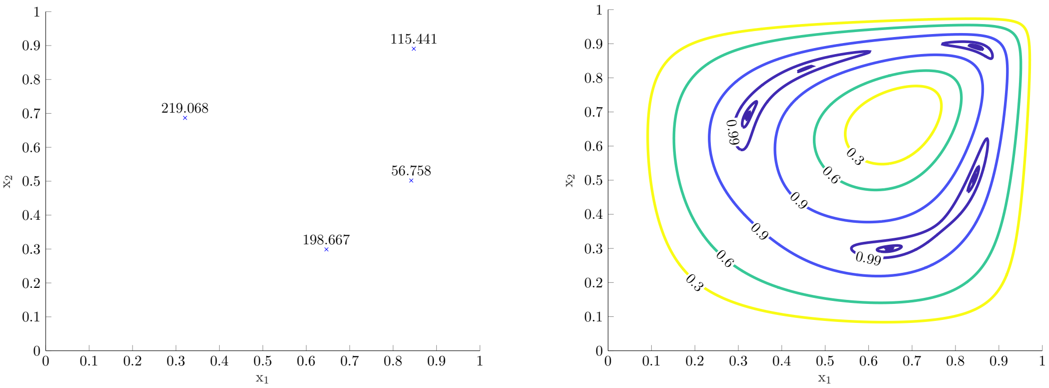
together with homogeneous Dirichlet boundary conditions on $\partial\Omega$. Here, the parameter q contains the scalar diffusion and convection coefficients of the elliptic operator. The observation domain is chosen as $\Omega_o = \bar{\Omega} = [0,1]^2$. As a priori guess for the parameter we choose $\hat{q} = (3, 0.5, 0.25)^\top$. Note that while (4.72) is a linear equation, the state $y \in H^1_0(\Omega) \cap \mathcal{C}(\Omega_o)$ depends non-linearly but differentiably on q. For each $k \in \{1,2,3\}$ the sensitivity $\delta y_k = \partial_k S[\hat{q}] \in H^1_0(\Omega) \cap \mathcal{C}(\Omega_o)$ can be computed from (4.4). Due to the tri-linearity of the form $a(\cdot,\cdot)(\cdot)$ it fulfills

$$a(\hat{q}, \delta y_k)(\varphi) = a(\mathbf{e}_k, \hat{y})(\varphi) \quad \forall \varphi \in H^1_0(\Omega),$$

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

First order optimality condition

In this section we numerically verify the discrete first-order necessary and sufficient optimality conditions from Theorem 4.45. 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 2 (together with Algorithm 1 and a full resolution of the arising finite-dimensional subproblems), until the residual is below machine precision. We obtain a discrete optimal design $\bar{u}_h$ in $\mathcal{M}^+(\Omega_o) \cap \mathcal{M}_h$ with five support points. By closer inspection we observe that two of the computed support points are located in adjacent nodes of the triangulation. Applying the post-processing from Section 4.6.1, we obtain the design given in Figure 4.1. Alongside we plot the isolines of $-\nabla\psi_h(\bar{u}_{\beta,h})$. As predicted by Theorem 4.45, $-\nabla\psi_h(\bar{u}_{\beta,h})$ is bounded from above by the cost parameter $\beta = 1$ and the support points of $\bar{u}_{\beta,h}$ align themselves with those points in which this upper bound is achieved.



(a) Optimal design $\bar{u}_{\beta,h}$.

(b) Isolines of $-\nabla\psi_h(\bar{u}_{\beta,h})$.

Figure 4.1: Optimal design and isolines of the gradient.

Confidence domains of the optimal estimator

Given the optimal design $\bar{\omega}_h$ from Figure 4.1a, and $K > 0$ we note that the measure $\bar{u}_h^K = (K/\|\bar{u}_{\beta,h}\|_{\mathcal{M}})\bar{u}_{\beta,h}$ is an optimal solution to

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

since the A-optimal design criterion is positive homogeneous; see Proposition 4.9. In this section we compute the linearised confidence domains (4.10) of the least-squares estimator $\tilde{q}$ from (4.6) corresponding to $\bar{u}_h^K$ for $K = 3 \cdot 10^4$.

Note that, given a sparse design measure u, and the associated linearised estimator $\tilde{q}_{\text{lin}} = (\tilde{q}_{\text{lin}}^1, \tilde{q}_{\text{lin}}^2, \tilde{q}_{\text{lin}}^3)^T$, see (4.9), there holds $\operatorname{Cov}[\tilde{q}_{\text{lin}}, \tilde{q}_{\text{lin}}] = \mathcal{I}_h(u)^{-1}$; see the discussion in Section 4.1. Consequently we have

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

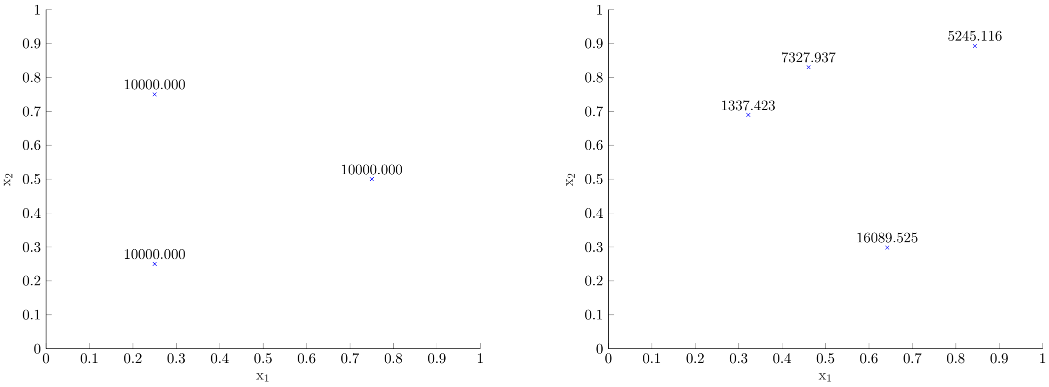


Figure 4.2: Reference measures u_1 (left) and $\bar{u}_h^{K,W}$ (right).

As a comparison, we also consider the estimators corresponding to two reference designs of the same norm. The first measure u_1 is chosen as a linear combination of three Dirac delta functions with equal coefficients while the second measure $\bar{u}_h^{K,W}$ is a solution to

$$\min_{u \in M^+(\Omega_o)} \operatorname{Tr}(W \mathcal{I}_h(u_h)^{-1} W) \quad \text{subject to } \|u_h\|_{\mathcal{M}} \leq K, \tag{4.73}$$

where $W = \operatorname{diag}(1, 1, 4)$, i.e. we place more weight on the variance for the estimation of q_3. The designs u_1 and $\bar{u}_h^{K,W}$ are depicted in Figure 4.2.

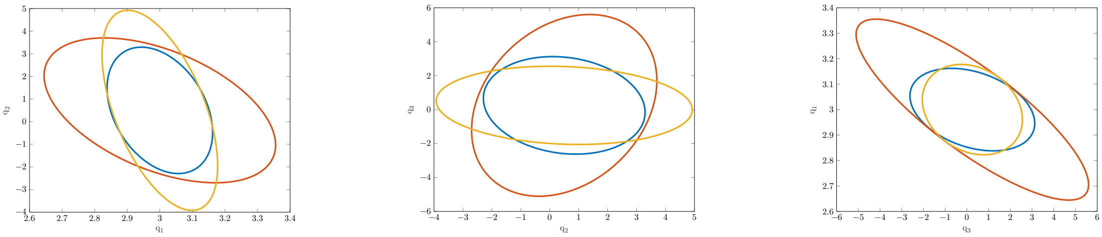
For a better visualization we plot the 50%-linearised confidence domains of the obtained estimators for the two dimensional parameter vectors $(q_1, q_2)^T$, $(q_2, q_3)^T$, and $(q_3, q_1)^T$ in Figure 4.3. Additionally, for each design we report $\operatorname{Tr}(\mathcal{I}_h(u)^{-1})$ as well as the diagonal entries of $\mathcal{I}_h(u)^{-1}$ in Table 4.1. As expected, since $\bar{u}_{\beta,h}$ is chosen by the A-optimal design criterion, we observe that

Table 4.1: Trace and diagonal entries of $\mathcal{I}_h(u)^{-1}$

u	$\mathcal{I}_h(u)^{-1}_{11}$	$\mathcal{I}_h(u)^{-1}_{22}$	$\mathcal{I}_h(u)^{-1}_{33}$	$\operatorname{Tr}(\mathcal{I}_h(u)^{-1})$
$\bar{u}_h^K$	0.019	5.627	5.955	11.601
u_1	0.091	7.388	20.678	28.157
$\bar{u}_h^{K,W}$	0.023	14.12	3.831	17.974

$$\operatorname{Tr}(\mathcal{I}_h(\bar{u}_h^K)^{-1}) \leq \operatorname{Tr}(\mathcal{I}_h(\bar{u}_h^{K,W})^{-1}) \leq \operatorname{Tr}(\mathcal{I}_h(u_1)^{-1}). \tag{4.74}$$

Moreover we note that $\mathcal{I}_h(\bar{u}_h^K)^{-1}_{kk} < \mathcal{I}_h(u_1)^{-1}_{kk}$ for all k, i.e. the optimal estimator estimates all unknown parameters with a smaller variance than the estimator associated to the reference design u_1. As a consequence, the linearised confidence domains of the optimal estimator are contained in those of the one corresponding to u_1; see Figure 4.3. In contrast, considering $\bar{u}_h^{K,W}$, we have $\mathcal{I}_h(\bar{u}_h^{K,W})^{-1}_{33} < \mathcal{I}_h(\bar{u}_h^K)^{-1}_{33}$ and $\mathcal{I}_h(\bar{u}_h^K)^{-1}_{kk} < \mathcal{I}_h(\bar{u}_h^{K,W})^{-1}_{kk}$ for $k = 1, 2$, i.e. the third parameter is estimated more accurately by choosing the measurement locations and weights according to $\bar{u}_h^{K,W}$ 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 (4.74). On the one hand, the obtained results show the efficiency of an optimally chosen measurement design at least for the linearised 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 Ψ.



(a) Ellipsoids for $(q_1, q_2)^T$.

(b) Ellipsoids for $(q_2, q_3)^T$.

(c) Ellipsoids for $(q_3, q_1)^T$.

Figure 4.3: Confidence ellipsoids for the estimators associated to $\bar{u}_h^K$ (blue), u_1 (red) and $\bar{u}_h^{K,W}$ (yellow).

Comparison of point insertion algorithms

In this section we investigate the performance of the successive point insertion algorithm presented in Section 4.4.1. We consider the same setup as in Section 4.7.1, 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 α and γ in (6.24) are both chosen as $1/2$ throughout the experiments and the iteration is terminated if either $\varPhi(u^k) \leq 10^{-9}$ or if the iteration number k exceeds $2 \cdot 10^4$. The aim of this section is to confirm the theoretical convergence results for Algorithms 2 and 3 as well as 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 $\mathbf{u}^{k+1}$ 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 2:

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

SPINAT Here, we employ the procedure suggested in [50], termed “Sequential Point Insertion and Thresholding”. In step 5., $\mathbf{u}^{k+1}$ is determined from a proximal gradient iteration (4.29). 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 $F(u(\mathbf{u}^{k+1}(\sigma_k))) \leq F(u(\mathbf{u}^{k+1/2}))$. In particular, given $u^{k+1/2} = \sum_i \mathbf{u}_i^{k+1/2} \delta_{x_i}$, we choose $\sigma_{0,k}$ as

$$\sigma_{0,k} = \max \left\{ 100, -2 \min_i \left\{ \frac{\mathbf{u}_i}{-\nabla\psi(u^{k+1/2})(x_i) - \beta} \right\} \right\}.$$

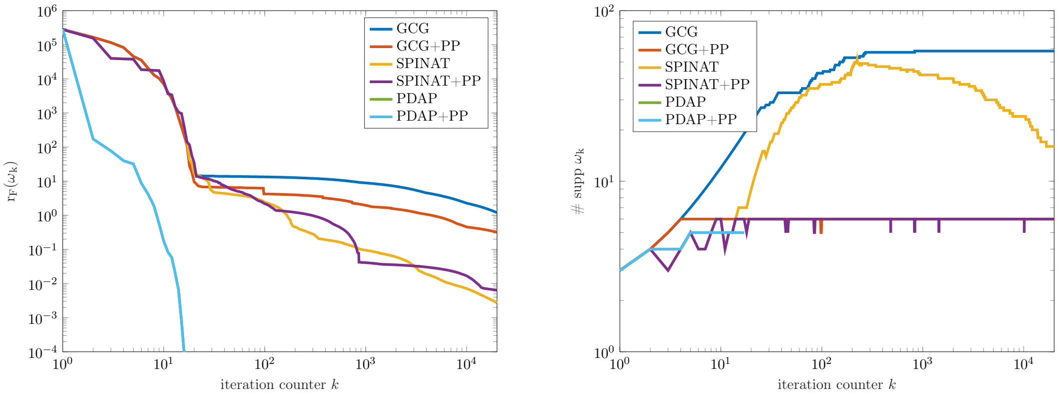
Note that by this choice of $\sigma_{0,k}$, the coefficients of all points $x \in \operatorname{supp} u^{k+1/2}$ with

$$-\nabla\psi(u^{k+1/2})(x) < \beta,$$

are set to zero in the first trial step (i.e. for $n = 0$).

PDAP Here, we consider Algorithm 3, i.e. the coefficient vector $\mathbf{u}^{k+1}$ is chosen as in (4.30) by solving the finite dimensional sub-problem (4.28) 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. As already discussed, this method can be interpreted as a method operating on a set of active points $\mathcal{A}_k = \operatorname{supp} u^k$ (see section 4.4.2), 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 in Algorithm 1 at the end of each iteration. In the following this will be denoted by an additional "+PP".



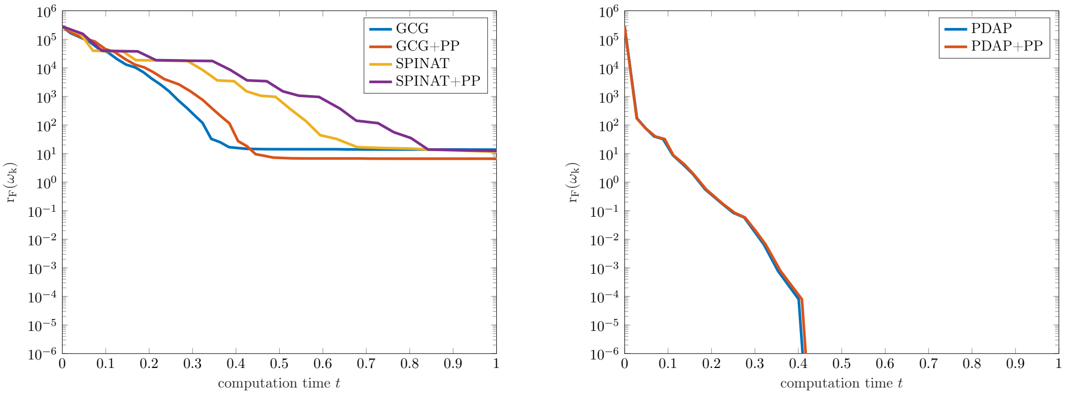
(a) Residual $r_F(u^k)$ over k.

(b) Support size $\#\operatorname{supp} u^k$ over k.

Figure 4.4: Residual and support size plotted over iteration number k.

In Figure 4.4a we plot the residual $r_F(u^k)$ 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 4.15 is sharp in this instance. The additional application of Algorithm 1 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 $r_F(u^k)$ and thus also the primal-dual gap $\Phi(u^k)$ remain above 10^{-3}. In contrast, PDAP terminates after few iterations within the tolerance backing the findings of Theorem 4.18. Note however that this is far from being conclusive since Theorem 4.18 cannot be applied to the discrete problem due to the piecewise linearity of $-\nabla\psi_h(\bar{u}_{\beta,h})$. Additionally, for fixed h, Algorithm 3 always converges in finitely many steps since possible support points are only chosen from $\mathcal{N}_h$ and the subproblems are solved up to optimality. We examine the convergence behaviour on a sequence of meshes in a following section.

Next, we study the influence of the different point removal steps on the sparsity pattern of the obtained iterates in Figure 4.4b. 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 $\bar{u}_{\beta,h}$ which is possibly caused by the small curvature of $-\nabla\psi_h(\bar{u}_{\beta,h})$ (see Figure 4.1b) 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 grows slower due to the additional projected gradient step in every iteration. Additionally, after reaching a threshold at approximately $k = 110$, the support size decreases monotonically in the remaining iterations. Concerning the application of Algorithm 1, we observe that the support remains bounded by $6 = 3(3 + 1)/2$ as predicted by Proposition 4.16. 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 1 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.



(a) Residual for GCG and SPINAT.

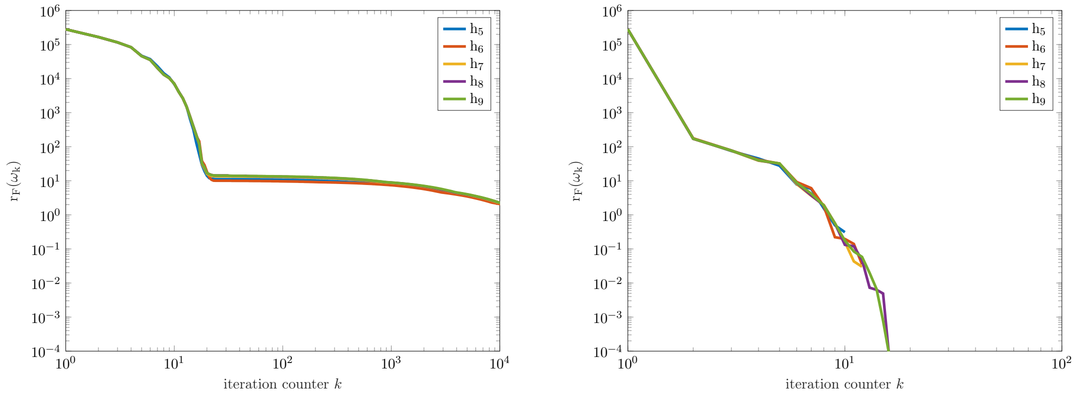
(b) Residual for PDAP.

Figure 4.5: Residual $r_F(u^k)$ plotted over the first second of the running time.

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 Figure 4.5a. We observe that neither the additional projected gradient steps nor the additional application of Algorithm 1 lead to a significant increase of the computational time. For PDAP, the measurement times and residuals for all iterations are shown in Figure 4.5b. 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 (4.30). On the other hand, using the intermediate iterate $u^{k+1/2}$ 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.

Mesh-independence

To finish our numerical studies on Example 1 we examine the influence of the mesh-size h on the performance of Algorithm 2. We again consider the A-optimal design problem for $\beta = 1$ and $\mathcal{I}_0 = 0$ on consecutively refined meshes $\mathcal{T}_{h_l}$, $l = 5, \ldots, 9$. On each refinement level l the optimal design problem is solved using GCG and PDAP, respectively. The computed residuals are shown in Figure 4.6. For both versions we observe that the convergence rate of the objective function value is stable with respect to mesh-refinement. A theoretical investigation of this mesh-independence property should be the subject of future work. Moreover the observed rate seems to be linear which backs up the theoretical results on the improved convergence behaviour of Algorithm 3, see Theorem 4.18. However, since the continuous sensitivities as well as an analytic solution $\bar{u}_\beta$ to the continuous optimal are not available its requirements on the curvature of the optimal gradient can not be checked straightforward. Additionally, in Figure 4.7, we plot the support size over the iteration counter for each refinement level. For GCG we observe a monotonic growth of the

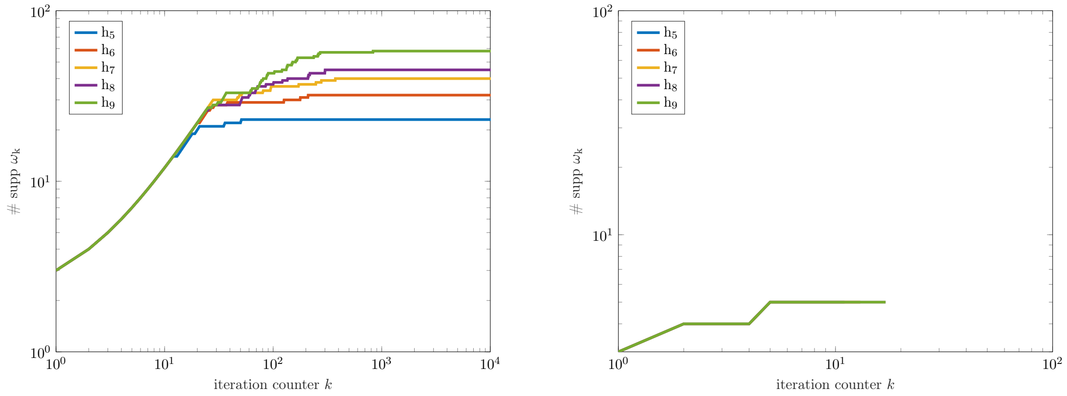


(a) Residuals for GCG.

(b) Residuals for PDAP.

Figure 4.6: Evolution of residuals $r_F(u^k)$ over iterations k on different refinement levels.

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.



(a) $\#\,\mathrm{supp}\,u^k$ over k for GCG.

(b) $\#\,\mathrm{supp}\,u^k$ over k for PDAP.

Figure 4.7: Evolution of the support size on different refinement levels.

4.7.2 Estimation of a parameterized source term

In this section we consider the A-optimal for a linear elliptic PDE with a linear parameter-to-state dependence. More concretely, for a given $q = (q_1, q_2) \in \mathbb{R}^2$ the associated state $y = S[q] \in H_0^1(\Omega) \cap \mathcal{C}(\Omega_o)$ is the unique solution to

$$a(q, y)(\varphi) = \int_{\Omega} \nabla y \cdot \nabla \varphi \; \mathrm{d}x = \int_{\Omega} (q_1 f_1 + q_2 f_2) \varphi \; \mathrm{d}x, \tag{4.75}$$

for all $\varphi \in H_0^1(\Omega)$. Here, the soure term is given as a linear combination between

$$f_1(x_1, x_2) = \sin(x_1) \sin((7/3)x_2), \quad f_2(x_1, x_2) = -\cos(1.777 * x_1) \sin((7/3)x_2).$$

Obviously, this corresponds to a Laplacian equation with homogeneous Dirichlet boundary conditions in which the right hand side is parameterized by q:

$$-\Delta y = q_1 f_1 + q_2 f_2 \quad \text{in } \Omega, \quad y = 0 \quad \text{on } \partial\Omega.$$

Due to the linearity of the parameter-to-state map the $k-th$ sensitivity $\delta y_k = \partial_k S,\ k = 1,2$ fulfills

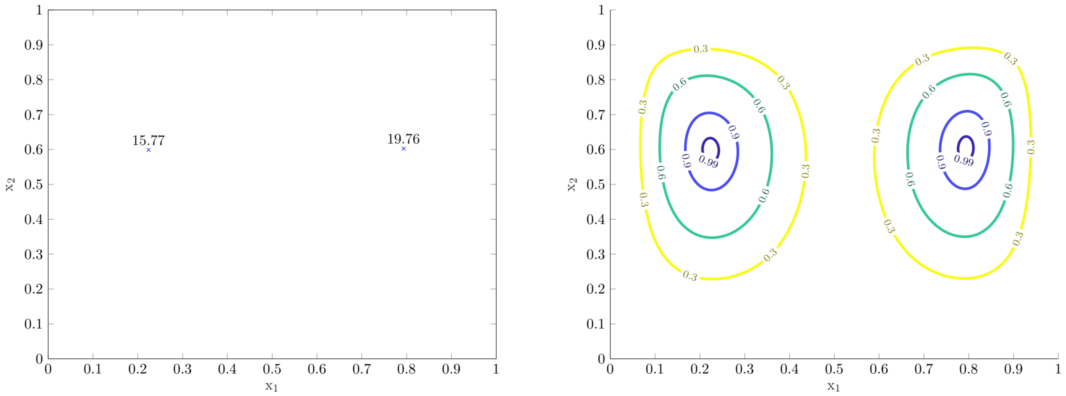
$$\int_\Omega \nabla\delta y_k \cdot \nabla\varphi \mathrm{d}x = \int_\Omega f_k \varphi \mathrm{d}x, \quad k = 1,2.$$

We choose $\Omega = [0,1]^2$, $\Omega_o = [0.1, 0.9]^2$ and $\beta = 1$ as well as $\mathcal{I}_0 = 0$. The computed optimal design $\bar{u}_{\beta,h} = \bar{\mathbf{u}}_1^h \delta_{\bar{x}_1^h} + \bar{\mathbf{u}}_2^h \delta_{\bar{x}_2^h}$ on $\mathcal{T}_{h_{11}}$ comprising two optimal sensors is depicted in Figure 4.8 alongside the isolines of $-\nabla\psi_h(\bar{u}_{\beta,h})$.

This example is geared towards the verification of the a priori error estimates for the objective functional and the optimal design which were presented in Section 4.6.2. Therefore let us briefly discuss the assumptions made for their derivation. First we note that $-\nabla\psi_h(\bar{u}_{\beta,h})$ admits exactly two global maximizers which align themselves with the support points of $\bar{u}_{\beta,h}$. Additionally, we verify that the Fisher information matrices $\{\mathcal{I}_h(\delta_{\bar{x}_i^h})\}_{i=1}^2$ are linearly independent. Hence the discrete optimal design $\bar{u}_{\beta,h}$ is unique. Due to the weak* convergence of the discrete design measures this may also indicate a similar behavior in the continuous problem. From the smoothness of the source terms f_1 and f_2, respectively, we conclude $\partial_k S \in \mathcal{C}^2(K)$, see [118, 124], as well as

$$\|\partial_k S - \partial_k^h S\|_{\mathcal{C}(\bar{K})} \leq c|\ln(h)|h^2, \quad k = 1,2,$$

for every open subset $K \subset\subset \Omega$ and for all $h \leq h_0$ small enough, c.f. [225]. Furthermore the Hessian of the A-optimal design criterion is positively definite at $\mathcal{I}(\bar{u}_\beta) + \mathcal{I}_0$. Consequently, since $\Omega_o \subset\subset \Omega$, Assumption 4.11 is fulfilled with $\gamma(h) = |\ln(h)|h^2$. However since the continuous optimal gradient $-\nabla\psi(\bar{u}_\beta) \in \mathcal{C}^2(\Omega_o)$ is unknown and $-\nabla\psi_h(\bar{u}_{\beta,h}) \notin \mathcal{C}^2(\Omega_o)$ the verification of the assumptions on its curvature is not directly possible and is therefore left for future work.

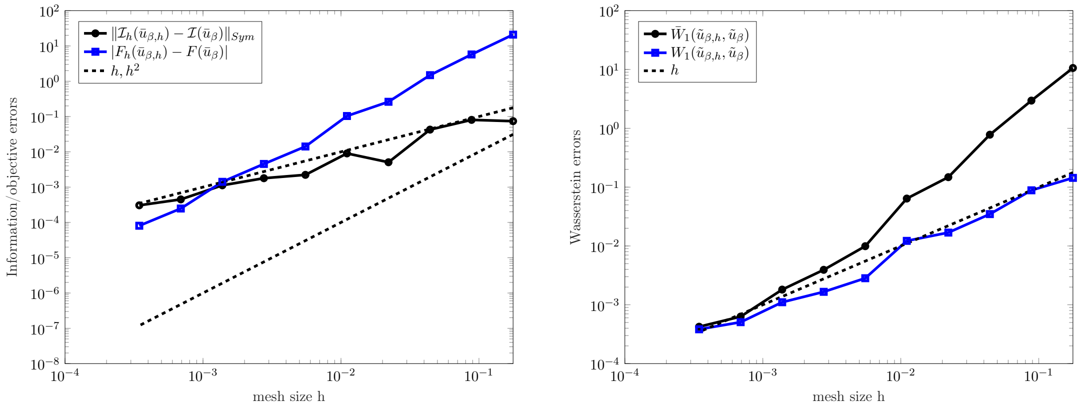


(a) Optimal design $\bar{u}_{\beta,h}$. (b) Isolines of $-\nabla\psi_h(\bar{u}_{\beta,h})$.

Figure 4.8: Optimal design and isolines of the gradient.

To verify the a priori estimates, we compute discrete optimal designs $\bar{u}_{\beta,h_k} \in \mathcal{M}_{h_k} \cap \mathcal{M}^+(\Omega_o)$ on a sequence of triangulations $\mathcal{T}_{h_k}$, $k = 2, \ldots, 11$. No analytic reference solution is available for this example. Therefore we consider a sequence of uniform triangulations $\hat{\mathcal{T}}_{h_k}$, $k = 1, \ldots, 12$, of Ω, where $\hat{\mathcal{T}}_{h_1}$ is obtained from $\mathcal{T}_{h_1}$ by a slight perturbation of the node at $(0.5, 0.5)$. As a reference $\bar{u}_\beta = \bar{\mathbf{u}}_1\delta_{\bar{x}_1} + \bar{\mathbf{u}}_2\delta_{\bar{x}_2}$ we compute an optimal design on the finest grid $\hat{\mathcal{T}}_{h_{12}}$. We emphasize that the support points of the reference measure are not included in the set of nodes $\mathcal{N}_{h_{11}}$ corresponding to the finest grid $\mathcal{T}_{h_{11}}$.

We evaluate the numerical results. Note that we do not expect to see the influence of the logarithmic factor $|\ln(h)|$ in the computations. In Figure 4.9a we display the convergence rates of the optimal objective function values as well as the Fisher information matrices. As predicted by Theorem 4.49 we observe the full order of convergence for the optimal objective function values $F_h(\bar{u}_{\beta,h})$ and a reduced order of $h \approx \gamma(h)$ for the error of the Fisher information matrices $\mathcal{I}_h(\bar{u}_{\beta,h})$, see Corollary 4.51.



(a) Errors for $F_h(\bar{u}_{\beta,h})$ and $\mathcal{I}_h(\bar{u}_{\beta,h})$.

(b) Errors for $\bar{W}_1$(blue) and W_1(black)

Figure 4.9: Convergence rates with respect to h.

The convergence rates for the modified Wasserstein distance $\bar{W}_1(\bar{u}_{\beta,h}, \bar{u}_\beta)$ and the Wasserstein-1 distance $W_1(\bar{u}_{\beta,h}/\|\bar{u}_{\beta,h}\|_{\mathcal{M}}, \bar{u}_\beta/\|\bar{u}_\beta\|_{\mathcal{M}})$ of the normalized optimal designs are displayed in Figure 4.9b. The Wasserstein distance between the sparse measures is computed by solving a linear program following [206, p. 64]. As predicted by the theory, see Theorem 4.62 the quantities $W_1(\bar{u}_{\beta,h}/\|\bar{u}_{\beta,h}\|_{\mathcal{M}}, \bar{u}_\beta/\|\bar{u}_\beta\|_{\mathcal{M}})$ as well as $\bar{W}_1(\bar{u}_{\beta,h}, \bar{u}_\beta)$ admit an asymptotic linear rate of convergence $h \approx \sqrt{\gamma(h)}$. However, for the latter one, the convergence rate on coarser grids appears to be better. To explain this observation we recall that $\bar{W}_1(\bar{u}_{\beta,h}, \bar{u}_\beta) = W_1(\bar{u}_{\beta,h}/\|\bar{u}_{\beta,h}\|_{\mathcal{M}}, \bar{u}_\beta/\|\bar{u}_\beta\|_{\mathcal{M}}) + |\|\bar{u}_{\beta,h}\|_{\mathcal{M}} - \|\bar{u}_\beta\|_{\mathcal{M}}|$. For the special case of the A-optimal design problem with $\mathcal{I}_0 = 0$ we obtain

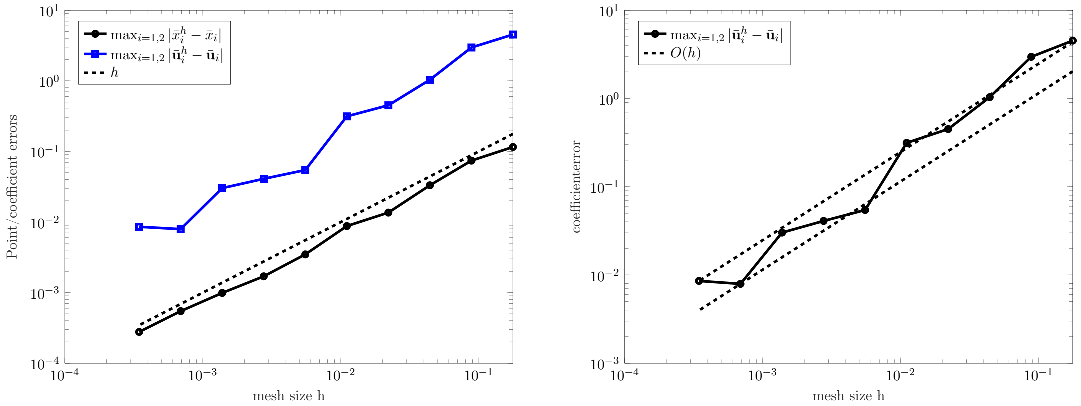
$$\beta\|\bar{u}\|_{\mathcal{M}} = -\langle \nabla\psi(\bar{u}_\beta), \bar{u}_\beta \rangle = \operatorname{Tr}(\mathcal{I}(\bar{u}_\beta)^{-1}) = \psi(\bar{u}_\beta),$$

due to the optimality conditions. Analogously we deduce $\psi_h(\bar{u}_{\beta,h}) = \beta\|\bar{u}_{\beta,h}\|_{\mathcal{M}}$. Hence, in this situation, we obtain

$$2\beta|\|\bar{u}_\beta\|_{\mathcal{M}} - \|\bar{u}_{\beta,h}\|_{\mathcal{M}}| = |F(\bar{u}_\beta) - F_h(\bar{u}_{\beta,h})| \leq |\ln(h)|h^2.$$

This explains the apparently better behavior of $\bar{W}_1(\bar{u}_{\beta,h}, \bar{u}_\beta)$ on coarser grids, while its asymptotic convergence rate is dominated by $W_1(\bar{u}_{\beta,h}/\|\bar{u}_{\beta,h}\|_{\mathcal{M}}, \bar{u}_\beta/\|\bar{u}_\beta\|_{\mathcal{M}})$.

Finally, we consider the convergence of the support points and the measurement weights as discussed in Theorem 4.58 and 4.61, respectively. For every $k = 2, \ldots, 11$, the discrete optimal design



(a) Errors for $\bar{x}_i^h$ and $\bar{\mathbf{u}}_i^h$

(b) Linear convergence on different meshes

Figure 4.10: Convergence rates for support points and measurement weights

consists of two distinct support points $\bar{u}_{\beta,h_k} = \bar{\mathbf{u}}_1^{h_k}\delta_{\bar{x}_1^{h_k}} + \bar{\mathbf{u}}_1^{h_k}\delta_{\bar{x}_2^{h_k}}$ such that

$$\max_{i=1,2} \max_{x \in \operatorname{supp} \bar{u}^i_{\beta,h_k}} |x - \bar{x}_i|_{\mathbb{R}^d} = \max_{i=1,2} |\bar{x}_i^{h_k} - \bar{x}_i|_{\mathbb{R}^d}.$$

Hence we compute the errors $\max_{i=1,2}|\bar{x}_i^h - \bar{x}_i|_{\mathbb{R}^d}$ and $\max_{i=1,2}|\bar{\mathbf{u}}_i^h - \bar{\mathbf{u}}_i|$. The results are shown in Figure 4.10a. For the distance of the support points we obtain, as predicted by Theorem 4.58, the reduced rate of $h \approx \sqrt{\gamma(h)}$. Note that since $\operatorname{supp} \bar{u}_{\beta,h} \subset \mathcal{N}_h$ and h denotes the mesh size, this estimate is in some sense optimal. The same rate can be concluded for the error of the coefficients, see Theorem 4.61, albeit the computed rate seems to be somewhat wiggly. Let us shortly elaborate on this seemingly strange behavior. In Figure 4.10b we plot the convergence rate of the coefficients and two lines indicating convergence of order h with different constants. As one can see, the error alternates between both reference lines. This implies a linear convergence behavior of the error with constants depending on the sequence of grids. For example the constants might depend on the barycentric coordinates of the reference support points within the cells of the triangulation $\mathcal{T}_{h_k}$. A similar behavior, that strengthens this conjecture, has been observed and examined for the convergence of the optimal control in a semi-infinite optimization problem, see [190].

Putting all previous observations into a nutshell, we conclude that the a priori error estimates for the sparse sensor placement problem from Section 4.6.2 are sharp in general and thus optimal. A next natural step is to consider meshes obtained by adaptive refinement based on a posteriori error estimates instead of uniformly refined triangulations. As for a priori error estimation, a starting point for such considerations may be provided by studying known concepts in state-constrained and semi-infinite optimization. For references on adaptivity in this context we point out [29, 126, 192, 208].

5 Sparse sensor placement for infinite-dimensional Bayesian inverse problems with PDEs

While we gave an in-depth discussion of sparse sensor placement for unknown finite dimensional parameter vectors in the preceding chapter, many complex processes rely on mathematical models incorporating unknown distributed functions. For example they may describe boundary and initial conditions of the model or they directly enter the definition of the differential operator describing the system. A study of optimal sensor placement problems in this context is in the main focus of this chapter. We consider real-life processes, e.g. from physics or biology, which are mathematically modeled by the weak form of a partial differential equation

$$a(q, y)(\varphi) = 0 \quad \forall \varphi \in W, \tag{5.1}$$

This equation relates the state variable $y \in Y$ to an unknown parameter function $q \in L^2(\Omega)$. The state space Y and the test space W are suitable sets of functions on a spatial domain $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, which we assume to be open and bounded. For the precise assumptions on the underlying PDE model we refer to Section 5.1.2. In the following, our interest lies in the inverse problem of identifying a distributed function q^* such that the associated partial differential equation and its solutions provide a reasonable mathematical surrogate for the modeled process.

As in the previous chapters a standing assumption of the following discussion is that the parameter cannot be measured directly. Inference on q^* is only possible through a vector $\mathbf{y}_d \in \mathbb{R}^N$ containing pointwise measurements of the quantity represented by the state y. These are taken at a finite set of sensor sites $\{x_i\}_{i=1}^N \subset \Omega_o$ in an experiment. The set $\Omega_o \subset \bar{\Omega}$ denotes a compact subset of possible sensor locations. Furthermore, the data is assumed to be subject to perturbation by random additive noise $\varepsilon \sim \mathcal{N}(0, \Sigma)$ where $\Sigma \in \mathbb{R}^{N \times N}$, $\Sigma_{ij} = \delta_{ij}/\mathbf{u}_i$, $i, j = 1, \ldots, N$. The positive scalar $\mathbf{u}_i > 0$ models the diligence of the measurement taken at the point x_i. An estimate for the unknown parameter is then obtained by matching the expected response of the mathematical model with the obtained measurements

$$\text{find } q \in L^2(\Omega),\ y \in Y: \quad y(x_i) = \mathbf{y}_d^i, \quad a(q, y)(\varphi) = 0 \quad \forall \varphi \in W$$

where $\mathbf{y}_d^i$ denotes the measurement obtained at x_i, $i = 1, \ldots, N$.

We point out that the measurements in this problem are found in a finite dimensional observation space whilst the unknown parameter is given by a distributed function in the infinite dimensional parameter space $L^2(\Omega)$. Without further knowledge on the parameter, e.g. its structure or smoothness, this discrepancy implies that the described inverse problem is inherently ill-posed. Thus it may admit infinitely many solutions or no solution at all. Moreover, the presence of measurement noise may lead to a severe misinterpretation of the obtained results since solutions do not depend continuously on the measurements. In order to circumvent these pathological cases one usually resorts to sophisticated regularization techniques allowing for a stable solution of the problem. Through this estimation process, the uncertainty of the measurements is also propagated into the obtained results. In particular, a full discussion of the inverse problem requires to quantify the influence of the measurement errors on the solution and to critically evaluate its reliability. As already discussed in the previous chapters, the quality of the solution critically depends on the measurement process. A careful choice of the used sensors and their positioning in the spatial domain may on the one hand yield measurements from which we can, in some sense, optimally draw conclusions on the parameter while mitigating the stochastic variability due to the measurement errors. On the other hand, restricting the measurements only to informative locations also keeps the overall cost of the experiment low.

The aim of this chapter is to provide and rigorously analyse an optimization framework which allows for a systematic choice of the measurement setup before any measurements are performed in practice. Obviously such a sensor placement formulation should also take the applied regularization strategy for the solution of the inverse problem into account. For optimal sensor placement based on Tikhonov regularization for high-dimensional parameters and the mean-squared error of the least-squares estimator, we refer to [127,128] and the METER method, [18]. Engineering applications of these approaches include the optimal monitoring of gravity dams, [180], and impedance imaging, [152]. More recently, probabilistic regularization approaches for inverse problems with infinite dimensional parameter spaces have received considerable attention. In this context, the uncertainty on the true value of the parameter is modeled by a probability measure on the parameter space, the so-called prior distribution. Instead of trying to compute a single function satisfying the constraints in the inverse problem we update the prior knowledge on the unknown parameter based on the obtained measurements and our assumptions on the measurement model. This is done by applying Bayes' Theorem. The solution to the inverse problem is then given by a new measure, the posterior distribution, which reflects our remaining degree of uncertainty on the parameter after observing the provided measurements. This framework allows to assess the statistical quality of the obtained solution in a natural way by comparing properties of the prior and the posterior distribution. For example, if its finite, we may compute the posterior variance which quantifies the stochastic variability of the probability measure around its mean. For a deeper discussion on the Bayesian approach to infinite dimensional inverse problems we refer to [80, 111, 250]. Simultaneously to the advances in the theory of Bayesian inverse problems with PDE constraints, the interest in optimal sensor placement for this type of regularization rose. Similar to the finite dimensional situation of the previous chapter these approaches are based on minimizing scalar-valued optimal design criteria acting on the Fisher information operator of the parameter estimates. This includes e.g. infinite dimensional analogues of the A and D optimal design criteria. For references on this highly active line of research we point out to [3–6,10]. In [138] a similar reasoning is applied to optimally place temperature sensors in a thermo-mechanical system.

Throughout the course of this chapter we will adopt this probabilistic view on inverse problems and model the uncertainties on the true value of the parameter as a Gaussian probability measure. Moreover we again resort to a linearization of the model equation around a given a priori guess $\hat{q} \in L^2(\Omega)$ and define the sensitivity operator $\partial S[\hat{q}]\colon L^2(\Omega) \to Y$ describing the influence of changes in the linearization point on the associated state variable. From the Bayesian viewpoint, this linearization leads to a Gaussian approximation to the posterior distribution. In order to improve the estimation results we propose to optimize the measurement process by solving the sensor placement problem

$$\min_{x\in\Omega_o^N,\ \mathbf{u}\in\mathbb{R}^N,\ N\in\mathbb{N}}[\Psi(X^*\Sigma^{-1}X)+\beta\sum_{i=1}^{N}\mathbf{u}_i]. \tag{5.2}$$

Here, Ψ denotes a convex optimal design criterion based on the Fisher information $X^*\Sigma^{-1}X$ which is parametrized by the measurement setup as

$$(\delta q_1, X^*\Sigma^{-1}X\delta q_2)_{L^2(\Omega)}=\sum_{i=1}^{N}\mathbf{u}_i\partial S[\hat{q}]\delta q_1\,(x_i)\partial S[\hat{q}]\delta q_2\,(x_i)\quad\forall\delta q_1,\ \delta q_2\in L^2(\Omega).$$

The additional term involving the total amount of measurements and the cost parameter $\beta>0$ models the total cost of the measurement process. While this formulation is clearly motivated by the problems presented in [3–6, 10] we point out that all of the previously mentioned works discuss the problem of selecting optimal sensor positions in an a priori given finite set of possible candidate locations in Ω_o. This reduces the sensor placement problem to a convex optimization problem for the diligence factor $\mathbf{u}$ associated to the respective sensor. However, this problem is still computationally challenging due to the infinite dimensional parameter space. A major novelty of the present work is that sensors can be placed everywhere in a set of possible sensor locations Ω_o which may contain an infinite number of candidate locations. Furthermore the optimal number of placed measurement sensors is also subject to optimization and is thus not a priori fixed. In particular, the possibly complicated dependence of the Fisher information operator on the sensor positions renders the present problem non-convex. Obviously this fact complicates the algorithmic solution of the sensor placement problem. We refer e.g. to [116] where the authors aim to place a fixed number of sensors with prescribed diligence factors at optimal positions. In order to do so they first discuss the computation of derivatives of the design criterion with respect to the sensor positions. This is a computationally challenging problem in itself but crucial for derivative-based optimization routines. Moreover we point out that the a priori unknown optimal number of measurements additionally introduces a combinatorial aspect to the problem in our case. Clearly, this further aggravates the algorithmic treatment of the problem.

In order to bypass these difficulties we consider the proposed sensor placement problem in the framework presented in Chapter 3. Instead of optimizing for the individual sensors we rewrite the problem and minimize with respect to their distribution. Mathematically these are modeled as positive Borel measures on the set of possible sensor locations. This leads to the sparse sensor placement problem

$$\min_{u\in\mathcal{M}^+(\Omega_o)}[\Psi(\mathcal{I}(u))+\beta\|u\|_{\mathcal{M}}].$$

Here, given $u\in\mathcal{M}^+(\Omega_o)$, the generalized Fisher information operator $\mathcal{I}(u)$ is characterized by

$$(\delta q_1,\mathcal{I}(u)\delta q_2)_{L^2(\Omega)}=\langle\partial S[\hat{q}]\delta q_1\partial S[\hat{q}]\delta q_2,u\rangle\quad\forall\delta q_1,\ \delta q_2\in L^2(\Omega).$$

Note that this measure-based sensor placement problem is convex due to the linear dependence of $\mathcal{I}(u)$ on the optimization variable u. Thus we may study existence of solutions and sufficient optimality conditions by resorting to results from convex analysis. Moreover its efficient numerical solution can be based on a generalization of the Primal-Dual-Active-Point strategy presented in the previous chapter. These methods only require the derivative of the optimal design criterion with respect to the measure u and not with respect to the individual sensors. At last, in order

to numerically compute an optimal design, the PDE constraints as well as the parameter have to be discretized. We base our discretization on a finite element surrogate for the PDE constraints and a variational discretization of the measurement setup. The parameter space $L^2(\Omega)$ is replaced by a finite dimensional subspace spanned by eigenfunctions of the prior covariance operator. As already done for the finite dimensional case in the previous chapter, convergence of these discretization schemes is proven and a priori error estimates between the discrete and continuous optimal solutions are provided. Again, we are not aware of any comparable results in this direction. As in the finite dimensional case we stress the similarity between our approach and the notion of approximate design theory dating back to Kiefer and Wolfowitz, [165]. However, to the best of our knowledge, there are no previous works on the extension of their reasoning to infinite dimensional parameters entering in a partial differential equation. Thus the results of this chapter should be interpreted as a first step towards optimal sensor placement accounting for both, the infinite dimensional nature of the distributed parameter and a possibly infinite number of candidate locations for the sensors.
To illustrate this rather abstract setting and to highlight the practical relevance of the proposed method we give a short example. Given a suitable triple of unknown functions $(q_1, q_2, q_3) \in L^2(\Omega)^3$ we consider the elliptic diffusion equation

$$a(q,y)(\varphi) = (\exp(q_1)\nabla y, \nabla\phi)_{L^2(\Omega)} + (q_2\partial_{x_1}y, \varphi)_{L^2(\Omega)} + (q_3\partial_{x_1}y, \varphi)_{L^2(\Omega)} - (f, \varphi)_{L^2(\Omega)} = 0, \quad (5.3)$$

for all φ in a suitable test space W. Similar models are frequently encountered in different research disciplines. In geophysical sciences, for example, diffusion models such as the Darcy equation, [235], are simple surrogates for the subsurface flow of fluids. The diffusion coefficient models the permeability of the underlying rock which is inferred from measurements of the fluid pressure y. Knowledge of this quantity is critical to make reliable predictions on the diffusion of nuclear waste due to a washout by groundwater or to optimize the recovery of underground oil resources. Diffusion equations incorporating unknown distributed functions are also encountered in oceanographie, [189, 265]. The state variable y models the concentration of a tracer substance diffusing in the ocean. Point measurements of this quantity are then used to infer on the unknown horizontal water velocities $(q_2, q_3) \in L^2(\Omega)^2$ as well as the diffusion coefficient q_1. In both of these examples researchers are faced with the problem of identifying a distributed function based on a limited amount of available data.

This chapter is organized as follows. We do not assume that the reader is familiar with the concept of probability measures on separable Hilbert spaces. In the following section we therefore briefly elaborate on the necessary theoretical background on Gaussian measures on $L^2(\Omega)$ and linear Bayesian inversion. Thereafter we apply the Bayesian methodology to inverse problems involving linearized PDE constraints in Section 5.1.2. Before proceeding to the optimal placement of measurement sensors we first have to define suitable design criteria quantifying the statistical properties of the obtained solution to the inverse problem. Several suitable examples and their mathematical properties are studied in Section 5.1.3. In Section 5.1.4 we finally formulate the optimal sensor placement problem based on the framework presented in Chapter 3. Existence results as well as a structural characterization of optimal measurement designs are provided. Their efficient numerical computation is in the focus of Section 5.3. Here we propose an extension of the Primal-Dual-Active-Point method from Section 4.4.2. We point out that these discussions are all based on the continuous problem which is formulated on the space of Borel measures. In order to compute an optimal measurement setup in practice, the problem has to be discretized. To this end, Section 5.2 puts the focus on suitable discretization strategies and the associated a priori error analysis. The presentation is complemented by numerical experiments which highlight the practical efficiency of the proposed method.

5.1 Sparse Bayesian optimal design

5.1.1 A primer on Gaussian random fields and linear Bayesian inference

In this section we provide the necessary background on Gaussian measures on $L^2(\Omega)$ and Bayesian inference in infinite dimensional Hilbert spaces. Readers familiar with this concept may skip this section and proceed directly to Section 5.1.2. Since the focus of this thesis lies on the analysis and the numerical treatment of the associated sensor placement problems we tend to keep this presentation short and concise, providing additional references where necessary. In the following we consider a probability space $(D, \mathcal{F}, \mathbb{P})$. Here D denotes a set of samples, $\mathcal{F}$ denotes a σ-algebra over D (a set of events) and $\mathbb{P}\colon \mathcal{F} \to [0,1]$ is a probability measure

$$\mathbb{P}(D) = 1, \quad \mathbb{P}(\emptyset) = 0, \quad \mathbb{P}\left(\bigcup_{i\in I} O_i\right) = \sum_{i\in I} \mathbb{P}(O_i), \quad O_i \in \mathcal{F},\ I \subseteq \mathbb{N}.$$

Furthermore we need some tools from measure theory. Let us consider two measurable spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$. For a $(\mathcal{A}, \mathcal{B})$-measurable mapping $f\colon X \to Y$ we recall the definition of its preimage as

$$f^{-1}\colon \mathcal{B} \to \mathcal{A}, \quad f^{-1}(O) = \{\, x \in X \mid f(x) \in O \,\}, \quad \forall O \in \mathcal{B}.$$

Given a measure $\mu\colon \mathcal{A} \to \mathbb{R}_* \cup \{+\infty\}$ its push-forward under f is defined as

$$f^{\#}\mu\colon \mathcal{B} \to \mathbb{R}_+ \cup \{+\infty\}, \quad O \mapsto \mu(f^{-1}(O)).$$

Let us first recall the definition of Gaussian measures on the real line, [77, Section 1.2.].

Definition 5.1. Consider the measurable spaces $(D, \mathcal{F})$ and $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. A probability measure $\mu\colon \mathcal{B}(\mathbb{R}) \to \mathbb{R}_+$ is called Gaussian if there exists $m \in \mathbb{R}$ and $\sigma \geq 0$ such that, if $\sigma > 0$, we have

$$\mu(O) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_O \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) \mathrm{d}x, \quad \forall O \in \mathcal{B}(\mathbb{R}),$$

or, whenever $\sigma = 0$,

$$\mu(O) = \begin{cases} 0, & m \notin O \\ 1, & m \in O \end{cases}, \quad \forall O \in \mathcal{B}(\mathbb{R}).$$

To stress the characterization of μ by $m \in \mathbb{R}$ and $\sigma \geq 0$ we write $\mu = \mathcal{N}(m, \sigma^2)$ A $(\mathcal{F}, \mathcal{B}(\mathbb{R}))$-measurable mapping $\zeta\colon D \to \mathbb{R}$ is called a Gaussian random variable if the probability measure $\mu = \zeta^{\#}\mathbb{P}$ is Gaussian.

In the following definition, we now fix the notion of a Gaussian probability measure on $L^2(\Omega)$. We refer e.g. to [215, Definition 2.1.1.].

Definition 5.2. Consider the measurable spaces $(D, \mathcal{F})$ and $(L^2(\Omega), \mathcal{B}(L^2(\Omega)))$. Furthermore given $v \in L^2(\Omega)$ we denote the associated linear form on $L^2(\Omega)$ by

$$v' \colon L^2(\Omega) \to \mathbb{R}, \quad z \mapsto (v, z)_{L^2(\Omega)}.$$

We call μ a Gaussian measure on $(L^2(\Omega), \mathcal{B}(L^2(\Omega)))$ if for every $v \in L^2(\Omega)$ the probability measure $\mu_v = (v')^{\#}\mu$ is a Gaussian measure on $\mathbb{R}$.

Further let a $(\mathcal{F}, \mathcal{B}(L^2(\Omega)))$-measurable mapping $q \colon D \to L^2(\Omega)$ be given and set $\mu = q^{\#}\mathbb{P}$. The mapping q is called a Gaussian random variable distributed according to μ if μ is a Gaussian measure on $L^2(\Omega)$. In short we denote this by $q \sim \mu$.

In the following we will always think of $q \colon D \to L^2(\Omega)$ as a random field distributed according to a Gaussian measure μ. Let us now consider the measure space $(L^2(\Omega), \mathcal{B}(L^2(\Omega)), \mu)$. Given $f \in L^1(L^2(\Omega), \mathcal{B}(L^2(\Omega)), \mu)$ define

$$\mathbb{E}^\mu[f(q)] = \int_D f(q(\omega)) \, \mathrm{d}\mathbb{P}(\omega) = \int_{L^2(\Omega)} f(z) \, \mathrm{d}\mu(z).$$

Here the change-of-variables formula, [40, A.3.1.], was used in the second equality. The following proposition is due to Theorem 2.1.2. and Proposition 2.1.4 in [215].

Proposition 5.1. *Let $q \colon D \to L^2(\Omega)$ be a Gaussian random variable distributed according to μ. Then there exist a unique function $q_\mu \in L^2(\Omega)$ and a unique positive trace class operator T_μ, i.e.*

$$(\delta q_1, T_\mu \delta q_2)_{L^2(\Omega)} = (T_\mu \delta q_1, \delta q_2)_{L^2(\Omega)}, \quad (\delta q_1, T_\mu \delta q_1)_{L^2(\Omega)} \geq 0, \quad \mathrm{Tr}_{L^2(\Omega)}(T_\mu) < \infty$$

for all δq_1 $\delta q_2 \in L^2(\Omega)$, with the following properties:

- $\mathbb{E}^\mu[(v, q)_{L^2(\Omega)}] = (v, q_\mu)_{L^2(\Omega)}$ *for all $v \in L^2(\Omega)$.*
- $\mathbb{E}^\mu[(v, q - q_\mu)_{L^2(\Omega)}(z, q - q_\mu)_{L^2(\Omega)}] = (v, T_\mu z)_{L^2(\Omega)}$ *for all $v, z \in L^2(\Omega)$.*
- $\mathrm{Var}(q) := \mathbb{E}^\mu[\|q - q_\mu\|^2_{L^2(\Omega)}] = \mathrm{Tr}_{L^2(\Omega)}(T_\mu)$.

Furthermore a Gaussian measure is uniquely defined by these properties.

Proposition 5.2. *Let two Gaussian probability measures μ_1 and μ_2 be given. Denote by q_{μ_i} and T_{μ_i} the associated function and positive trace class operator from Proposition 5.1, respectively, $i = 1, 2$. Then there holds*

$$\mu_1 = \mu_2 \Leftrightarrow T_{\mu_1} = T_{\mu_2}, \quad q_{\mu_1} = q_{\mu_2}.$$

Definition 5.3. We call q_μ the *mean* and T_μ the *covariance operator* of μ.

Remark 5.1. It is worthwhile to note that given a random variable $q \sim \mu$ the associated covariance operator T_μ quantifies the uncertainty of q in some appropriate sense. For simplicity assume that $q_\mu = 0$ and let $\delta q \in L^2(\Omega)$ be given. Following Definition 5.2 the mapping

$$m_{\delta q} \colon D \to \mathbb{R}, \quad \omega \mapsto (q(\omega), \delta q)_{L^2(\Omega)},$$

is a scalar-valued random variable distributed according $\mu_{\delta q} = \mathcal{N}(0, \sigma_{\delta q}^2)$. Its variance $\sigma_{\delta q}^2 \geq 0$ is calculated from the second property in Proposition 5.1 as

$$\sigma_{\delta q}^2 = \mathbb{E}^{\mu}[(q, \delta q)_{L^2(\Omega)}] = (\delta q, T_\mu \delta q)_{L^2(\Omega)}.$$

Thus the weighted scalar-product

$$(\delta q, T_\mu \delta q)_{L^2(\Omega)} = \|T_\mu^{1/2} \delta q\|_{L^2(\Omega)}^2,$$

provides a measure to quantify the uncertainty on q into the direction of $\delta q \in L^2(\Omega)$.

Since T_μ is positive and of trace class on $L^2(\Omega)$ there holds $T_\mu \in \operatorname{Pos}(L^2(\Omega), L^2(\Omega))$. Following the discussion in Section 3.1.1 there exists a function $k_{T_\mu} \in L^2(\Omega \times \Omega)$ with

$$[T_\mu v](x) = \int_\Omega k_{T_\mu}(x,y) v(y) \,\mathrm{d}y, \quad \forall v \in L^2(\Omega), \quad k_{T_\mu}(x,y) = k_{T_\mu}(y,x), \text{ a.e. } x, y \in \Omega.$$

The function $k_{T_\mu} \in L^2(\Omega \times \Omega)$ is called the *covariance function* of μ. Recalling the Hilbert-Schmidt theorem, see [264, Theorem VI.3.2], and Lidskii's theorem, [213], we deduce the following results.

Proposition 5.3. *There exists a sequence of scalars $\{\lambda_i\}_{i \in \mathbb{N}}$, $\lambda_i \geq \lambda_{i+1} \geq 0$, $i \in \mathbb{N}$, and an orthonormal system $\{\phi_i\}_{i \in \mathbb{N}}$ of $L^2(\Omega)$ with*

$$T_\mu \phi_i = \lambda_i \phi_i, \quad v = q_v + \sum_{i=1}^{\infty} (v, \phi_i)_{L^2(\Omega)} \phi_i, \quad \operatorname{Tr}_{L^2(\Omega)}(T_\mu) = \sum_{i=1}^{\infty} \lambda_i < \infty$$

for some $q_v \in \operatorname{Ker} T_\mu$ and all $v \in L^2(\Omega)$.

The following result allows for an integral representation of the trace, [53, Theorem 3.1.].

Proposition 5.4. *There exists a function $\tilde{k}_{T_\mu} \in L^1(\Omega)$ with*

$$\operatorname{Tr}_{L^2(\Omega)}(T_\mu) = \sum_{i=1}^{\infty} \lambda_i = \int_\Omega \tilde{k}_{T_\mu} \,\mathrm{d}x,$$

where $\{\lambda_i\}_{i \in \mathbb{N}}$ denotes the sequence of nonnegative scalars from Proposition 5.3.

Remark 5.2. At first sight it might seem tempting to define the kernel function $\tilde{k}_{T_\mu}$ as

$$\tilde{k}_{T_\mu}(x) = k_{T_\mu}(x,x) \quad \text{for a.e. } x \in \Omega_o$$

Indeed this holds true if $k_{T_\mu} \in \mathcal{C}(\bar{\Omega} \times \bar{\Omega})$. However this definition is in general not meaningful since the set

$$\Omega_d = \{\, (x_1, x_2) \in \Omega \times \Omega \mid x_1 = x_2 \,\},$$

has zero Lebesgue measure. The kernel $\tilde{k}_{T_\mu} \in L^1(\Omega)$ is obtained through a pointwise averaging of k_{T_μ} on the diagonal set Ω_d. For a deeper discussion on this subject we refer to Section 3 of [53].

From the boundedness of $\mathrm{Var}(q)$ in Proposition 5.1 we conclude $q \in L^2(D, \mathcal{F}, \mathbb{P}; L^2(\Omega))$ and thus also

$$q \in L^2(D \times \Omega, \mathcal{F} \otimes \mathcal{L}(\Omega), \mathbb{P} \times \mu_L; \mathbb{R}),$$

see [90, III.11, Theorem 17]. Here $\mathcal{F} \otimes \mathcal{L}(\Omega)$ denotes the tensor σ-algebra on the cartesian product $D \times \Omega$, μ_L denotes the Lebesgue measure and $\mathbb{P} \times \mu_L$ is the uniquely defined product measure. This allows for a pointwise discussion of the random field q in an almost everywhere sense. For a.e $x \in \Omega$ we have

$$q(\cdot, x) \in L^2(D, \mathcal{F}, \mathbb{P}; \mathbb{R}), \quad q(\cdot, x) \sim \mu_x = \mathbb{P} \circ q(\cdot, x)^{-1} = \mathcal{N}(q_\mu(x), \tilde{k}_{T_\mu}(x, x)),$$

i.e. $q(\cdot, x)$ is a scalar valued Gaussian random variable. In the same fashion, looking at it the other way round, there holds $q(\omega, \cdot) \in L^2(\Omega)$ for $\omega \in D$ $\mathbb{P}$-almost surely. We call $q(\omega, \cdot) \in L^2(\Omega)$ a *realization* of q or a *draw* from μ.

The *pointwise variance* Var_q of q is defined by

$$\mathrm{Var}_q \colon \Omega_o \to \mathbb{R}_+, \quad x \mapsto \int_D |q(\omega, x) - q_\mu(x)|^2 \, \mathrm{d}\mathbb{P}(\omega). \tag{5.4}$$

Since $q \in L^2(D, \mathcal{F}, \mathbb{P}; L^2(\Omega))$ there holds $\mathrm{Var}_q \in L^1(\Omega)$ and, due to Fubini-Tonelli, its norm is given as

$$\begin{aligned}\int_\Omega \mathrm{Var}_q(x) \, \mathrm{d}x &= \int_\Omega \int_D |q(\omega, x) - q_\mu(x)|^2 \, \mathrm{d}\mathbb{P}(\omega)\mathrm{d}x = \int_\Omega \tilde{k}_{T_\mu} \, \mathrm{d}x = \int_D \|q(w, \cdot) - q_\mu\|^2_{L^2(\Omega)} \, \mathrm{d}\mathbb{P}(\omega) \\ &= \mathbb{E}^\mu[\|q - q_\mu\|^2_{L^2(\Omega)}] = \mathrm{Tr}_{L^2(\Omega)}(T_\mu).\end{aligned} \tag{5.5}$$

In the remainder of this chapter, covariance operators defined through the inverse of an unbounded operator will play a central role.

Lemma 5.5. *Let $\mathcal{I}_0 \colon \mathrm{dom}_{L^2(\Omega)} \mathcal{I}_0 \to L^2(\Omega)$ be a not necessarily bounded but closed operator with dense domain. Assume that $\mathcal{I}_0$ is self-adjoint and nonnegative on its domain i.e*

$$(q_1, \mathcal{I}_0 q_2)_{L^2(\Omega)} = (\mathcal{I}_0 q_1, q_2)_{L^2(\Omega)}, \quad (q_1, \mathcal{I}_0 q_1)_{L^2(\Omega)} \geq 0 \quad \forall q_1, q_2 \in \mathrm{dom}_{L^2(\Omega)} \mathcal{I}_0,$$

as well as $\mathrm{Im}\, \mathcal{I}_0 = L^2(\Omega)$. Then $\mathcal{I}_0$ is a bijection. Its inverse

$$\mathcal{I}_0^{-1} \colon L^2(\Omega) \to L^2(\Omega),$$

is a bounded, self-adjoint and positive operator on $L^2(\Omega)$. Assume that $\mathcal{I}_0^{-1}$ is compact. Then there exists a sequence of positive scalars $\{\lambda_i\}_{i \in \mathbb{N}}$, $\lambda_i > \lambda_{i+1} > 0$, $i \in \mathbb{N}$, and an orthonormal basis $\{\phi_i\}_{i \in \mathbb{N}}$ of $L^2(\Omega)$ with

$$\mathcal{I}_0^{-1} \phi_i = \lambda_i \phi_i, \quad v = \sum_{i=1}^{\infty} (v, \phi_i)_{L^2(\Omega)} \phi_i, \quad \forall v \in L^2(\Omega).$$

If $\sum_{i=1}^{\infty} \lambda_i < \infty$, then $T_\mu = \mathcal{I}_0^{-1}$ is a covariance operator.

Proof. The existence of $\mathcal{I}_0^{-1}$ and its boundedness follows from [52, Theorem 2.21]. Furthermore we readily obtain

$$(\mathcal{I}_0^{-1}q_1, q_2)_{L^2(\Omega)} = (\mathcal{I}_0^{-1}q_1, \mathcal{I}_0\mathcal{I}_0^{-1}q_2)_{L^2(\Omega)} = (q_1, \mathcal{I}_0^{-1}q_2)_{L^2(\Omega},$$

as well as

$$(\mathcal{I}_0^{-1}q_1, q_1)_{L^2(\Omega)} = (\mathcal{I}_0^{-1}q_1, \mathcal{I}_0\mathcal{I}_0^{-1}q_1)_{L^2(\Omega)} \geq 0,$$

for all $q_1, q_2 \in L^2(\Omega)$. The existence of a sequence of positive eigenvalues $\{\lambda_i\}_{i\in\mathbb{N}}$ and associated eigenfunctions $\{\phi_i\}_{i\in\mathbb{N}}$ forming an orthonormal $L^2(\Omega)$ basis follows from the spectral theorem, see [52, Theorem 6.11], and $\operatorname{Ker}\mathcal{I}_0 = \{0\}$. If the eigenvalues of $\mathcal{I}_0^{-1}$ are summable it is a positive trace class operator on $L^2(\Omega)$. This gives the last statement. □

For the rest of this chapter we make the following standing assumption.

Assumption 5.1. The compact operator $\mathcal{I}_0^{-1}\colon L^2(\Omega) \to L^2(\Omega)$ is given by the inverse of an operator $\mathcal{I}_0$ as defined in Lemma 5.5. There holds $\operatorname{Tr}_{L^2(\Omega)}(\mathcal{I}_0^{-1}) < \infty$.

Note that the operator $\mathcal{I}_0$ and its inverse are completely characterized through their eigenvalues and the associated eigenfunctions since

$$\mathcal{I}_0 q_1 = \sum_{i=1}^{\infty} \lambda_i^{-1}(q_1, \phi_i)_{L^2(\Omega)}\phi_i, \quad \mathcal{I}_0^{-1}q_2 = \sum_{i=1}^{\infty} \lambda_i (q_2, \phi_i)_{L^2(\Omega)}\phi_i \quad \forall q_1 \in \operatorname{dom}_{L^2(\Omega)}\mathcal{I}_0,\ q_2 \in L^2(\Omega).$$

Throughout the rest of this chapter we adapt this spectral representation of such operators. More general, for $s \in [-1, 1]$ we define the s-th fractional powers of $\mathcal{I}_0$ as

$$\mathcal{I}_0^s\colon \operatorname{dom}_{L^2(\Omega)}\mathcal{I}_0^s \to L^2(\Omega) \quad q \mapsto \sum_{i=1}^{\infty} \lambda^{-s}(q, \phi_i)_{L^2(\Omega)}\phi_i.$$

The $L^2(\Omega)$ domain of $\mathcal{I}_0^s$ is given by

$$\operatorname{dom}_{L^2(\Omega)}\mathcal{I}_0^s = \left\{ q \in L^2(\Omega) \mid \|\mathcal{I}_0^s q\|^2_{L^2(\Omega)} = \sum_{i=1}^{\infty} \lambda_i^{-2s}(q, \phi_i)^2 < \infty \right\}.$$

Associated to a Gaussian measure $\mu = \mathcal{N}(q_\mu, \mathcal{I}_0^{-1})$ we define its Cameron-Martin space.

Definition 5.4. Let a covariance operator $\mathcal{I}_0^{-1}$ in the sense of Lemma 5.5 with eigenpairs $(\lambda_i, \phi_i)_{i\in\mathbb{N}}$ be given. Its Cameron-Martin space is defined as

$$\mathcal{H} = \operatorname{dom}_{L^2(\Omega)}\mathcal{I}_0^{1/2} = \left\{ q \in L^2(\Omega) \mid \|\mathcal{I}_0^{1/2} q\|^2_{L^2(\Omega)} = \sum_{i=1}^{\infty} \lambda_i^{-1}(\phi_i, q)^2_{L^2(\Omega)} < \infty \right\}.$$

Proposition 5.6. *The bilinear form $(\cdot,\cdot)_{\mathcal{H}}\colon \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ with*

$$(q_1, q_2)_{\mathcal{H}} = (\mathcal{I}_0^{1/2}q_1, \mathcal{I}_0^{1/2}q_2)_{L^2(\Omega)} = \sum_{i=1}^{\infty} \lambda_i^{-1}(q_1, \phi_i)(q_2, \phi_i) \quad \forall q_1,\ q_2 \in \mathcal{H},$$

defines an inner product on $\mathcal{H}$. The set $\mathcal{H}$ together with $(\cdot,\cdot)_{\mathcal{H}}$ form a Hilbert space with respect to the induced norm

$$\|q\|_{\mathcal{H}} = \|\mathcal{I}_0^{1/2}q\|_{L^2(\Omega)} = \sqrt{(q, q)_{\mathcal{H}}} \quad \forall q \in \mathcal{H}.$$

Proof. Obviously $(\cdot,\cdot)_{\mathcal{H}}$ defines an inner product on $\mathcal{H}$. Since $\mathcal{I}_0$ is closed so is its square root $\mathcal{I}_0^{1/2}$. As a consequence the domain of $\mathcal{I}_0^{1/2}$ is a Hilbert space with respect to the graph norm

$$\|q\|_G = \sqrt{\|q\|^2_{L^2(\Omega)} + \|q\|^2_{\mathcal{H}}} \quad \forall q \in \mathcal{H},$$

which is induced by the inner product

$$(q_1, q_2)_G = (q_1, q_2)_{L^2(\Omega)} + (q_1, q_2)_{\mathcal{H}} \quad \forall q_1, q_2 \in \mathcal{H}.$$

Again following [52, Theorem 2.21] we conclude

$$\|q\|_{\mathcal{H}} \leq \|q\|_G, \quad \|q\|^2_{L^2(\Omega)} \leq c(q, \mathcal{I}_0^{1/2} q)_{L^2(\Omega)} \leq c\|q\|_{\mathcal{H}}\|q\|_{L^2(\Omega)} \quad \forall q \in \mathcal{H},$$

and some constant $c > 0$ independent of q. Thus the graph norm and the $\mathcal{H}$ norm are equivalent on $\mathcal{H}$ finishing the proof. □

Since $\mathcal{I}_0^{-1}$ the associated Cameron-Martin space and $L^2(\Omega)$ form a rigged Hilbert space

$$\mathcal{H} \overset{c}{\hookrightarrow} L^2(\Omega) \simeq L^2(\Omega)^* \hookrightarrow \mathcal{H}^*,$$

where the first embedding (and thus the second) is compact and dense. We give a concrete example to clarify this abstract definition.

Example 5.1. *Let Ω be a bounded convex domain in $\mathbb{R}^d$, $d \leq 3$. Furthermore denote by $\mathcal{A} = -\Delta$ the Dirichlet Laplacian on Ω. It is well known that $\mathcal{A}$ defines an isomorphism between $L^2(\Omega)$ and its $L^2(\Omega)$-domain*

$$\operatorname{dom}_{L^2(\Omega)} \mathcal{A} = H^2(\Omega) \cap H^1_0(\Omega),$$

equipped with the graph norm. Furthermore its inverse $\mathcal{A}^{-1}$ is compact and positive. Applying the spectral theorem yields the existence of an orthonormal basis $\{\phi_i\}_{i\in\mathbb{N}}$ of $L^2(\Omega)$ and a zero sequence $\{\lambda_i\}_{i\in\mathbb{N}}$ of positive scalars with $0 < \lambda_{i+1} \leq \lambda_i$, $i \in \mathbb{N}$, and

$$\mathcal{A}^{-1} q = \sum_{i=1}^{\infty} \lambda_i (q, \phi_i)_{L^2(\Omega)} \phi_i \quad \forall q \in L^2(\Omega)$$

Recently covariance operators constructed from solution operators to fractional elliptic equations have increased in interest. We consider fractional powers of the operator $\mathcal{A}^{-1}$ defined by

$$\mathcal{A}^{-s} q = \sum_{i=1}^{\infty} \lambda_i^s (q, \phi_i)_{L^2(\Omega)} \phi_i \quad \forall q \in L^2(\Omega),$$

for $s \in [1,2]$. If e.g. $\bar{\Omega} = [0,1]^d$ and $s > d/2$ the eigenvalues of $\mathcal{A}^{-s}$ are summable, see e.g. [250, Theorem 2.10]. Thus $\mathcal{A}^{-s}$ yields a covariance operator. Let us characterize the space

$$\mathcal{H}^s = \left\{ q \in L^2(\Omega) \mid \sum_{i=1}^{\infty} \lambda_i^{-s} (q, \phi_i)^2_{L^2(\Omega)} < \infty \right\},$$

for a general bounded and convex domain Ω and $s \in [1,2]$. In the extremal cases $s = 1$ and $s = 2$ we readily obtain

$$\mathcal{H}^1 = \left\{ q \in H^1_0(\Omega) \mid \|\nabla q\|_{L^2(\Omega)} < \infty \right\} = H^1_0(\Omega),$$

as well as

$$\mathcal{H}^2 = \{\, q \in H_0^1(\Omega) \mid \| -\Delta\, q\|_{L^2(\Omega)} < \infty \,\} = H^2(\Omega) \cap H_0^1(\Omega),$$

by partial integration. The remaining cases for $s \in (1,2)$ can be identified from the Hilbert scale defined by $(-\Delta)^{-1}$, [42], through real-valued interpolation

$$\mathcal{H}^s = [H^2(\Omega) \cap H_0^1(\Omega), H_0^1(\Omega)]_{2-s} = H^s(\Omega) \cap H_0^1(\Omega) \quad s \in (1,2).$$

For the last result we refer to [51, Chapter 14].

Remark 5.3. Following these considerations it is clear that $\mathcal{I}_0$ can be extended to an operator from $\mathcal{H}$ to $\mathcal{H}^*$ with

$$\langle q_1, \mathcal{I}_0 q_2\rangle_{\mathcal{H},\mathcal{H}^*} = (q_1, q_2)_{\mathcal{H}} \quad \forall q_1, q_2 \in \mathcal{H}.$$

The existence of its bounded inverse $\mathcal{I}_0^{-1} \colon \mathcal{H}^* \to \mathcal{H}$ is a direct consequence of the Lax-Milgram Lemma. Since there will not be ambiguities in the following we denote the resulting operator in both cases, as operator on $\mathcal{H}$ and on $\operatorname{dom}_{L^2(\Omega)} \mathcal{I}_0$, by the same letter. Moreover set $s = -1/2$. Then the operator $\mathcal{I}_0^{-1/2}$ maps $L^2(\Omega)$ continuously into the Cameron-Martin space $\mathcal{H}$. Thus its adjoint operator $(\mathcal{I}_0^{-1/2})^*$ is linear and continuous between the topological dual space $\mathcal{H}^*$ of $\mathcal{H}$ and $L^2(\Omega)$. However we will also frequently interpret $\mathcal{I}_0^{-1/2}$ as operator from $L^2(\Omega)$ onto itself. In this situation $\mathcal{I}_0^{-1/2}$ is self-adjoint i.e. $(\mathcal{I}_0^{-1/2})^* = \mathcal{I}_0^{-1/2}$. To improve readability we also write $\mathcal{I}_0^{-1/2}$ for the adjoint operator in both cases.

An useful characterization of a Gaussian random field q is given in terms of its *Karhunen-Loève expansion*, see [215].

Theorem 5.7. *Let a covariance operator $\mathcal{I}_0^{-1}$ in the sense of Lemma 5.5 be given and denote by $(\lambda_i, \phi_i)_{i\in\mathbb{N}}$ the associated eigenpairs. Furthermore let $\{\zeta_i\}_{i\in\mathbb{N}}$ denote a family of i.i.d. random variables with $\zeta_1 \colon D \to \mathbb{R}$, $\zeta_1 \sim \mathcal{N}(0,1)$. Define the function*

$$q \colon D \to L^2(\Omega), \quad q(\omega, x) = q_\mu(x) + \sum_{i=1}^{\infty} \sqrt{\lambda_i}\zeta_i(\omega)\phi_i(x), \tag{5.6}$$

for $\mathbb{P}$-a.e. $\omega \in D$, a.e. $x \in \Omega$ and some $q_\mu \in \operatorname{dom}_{L^2(\Omega)} \mathcal{I}_0^{1/2}$. Then q is distributed according to $\mu = \mathcal{N}(q_\mu, \mathcal{I}_0^{-1})$.

This representation allows to compute (approximate) draws from the measure $\mu = \mathcal{N}(q_\mu, \mathcal{I}_0^{-1})$ by simply truncating the orthogonal expansion in (5.6) after a fixed number of terms n. A realization $q^n(\omega,\cdot) \in L^2(\Omega)$ of the truncated field

$$q^n \colon D \to L^2(\Omega), \quad q^n(\omega, x) = q_\mu(x) + \sum_{i=1}^{n} \sqrt{\lambda_i}\zeta_i(\omega)\phi_i(x),$$

can then be obtained by drawing from the finite dimensional distribution $\mathcal{N}(0, \mathrm{Id})$, $\mathrm{Id} \in \mathbb{R}^{n\times n}$, once the eigenvalues $\{\lambda_i\}_{i=1}^n$ of $\mathcal{I}_0^{-1}$ as well as the associated eigenfunctions $\{\phi_i\}_{i=1}^n$ are known.

Let us now discuss the inverse problem of identifying a distributed function from scarce observations. To this end we consider operator equations of the form

$$\text{find } q \in L^2(\Omega)\colon \quad Xq = \mathbf{y}_d. \tag{5.7}$$

Here $X \in \mathcal{L}(L^2(\Omega), \mathbb{R}^N)$ denotes a linear and continuous operator and $\mathbf{y}_d \in \mathbb{R}^N$ is a given finite dimensional vector containing the collected data. Moreover we assume that there is no systematic modeling error in the equation but the measurements are subject to additive perturbation i.e.

$$\mathbf{y}_d = Xq^* + \epsilon.$$

By $\epsilon \in \mathbb{R}^N$ we denote the measurement noise and $q^* \in L^2(\Omega)$ denotes the unknown parameter which we aim to recover. We adopt a probabilistic description of the measurement error and assume that ϵ is given as realization of a Gaussian random variable $\varepsilon\colon D \to \mathbb{R}^N$ distributed according to $\mu_E = \mathcal{N}(0, \Sigma)$ where Σ is a diagonal matrix with $\Sigma_{ii} > 0$, $i = 1, \dots, N$. Our aim is now to identify q^* based on the observation $\mathbf{y}_d$. Obviously this problem is inherently ill-posed due to the discrepancy between the finite dimensionality of the data and the infinite dimensional nature of the parameter. In particular we stress that the kernel of X is non-empty. Thus the equation in (5.7) may admit infinitely many solutions or no solution at all depending on whether $\mathbf{y}_d \in \operatorname{Im} X$ or not.

In order to obtain a, in some sense, well-defined formulation we resort to a different concept of solutions to the inverse problem. To this end we follow a Bayesian approach and describe our prior believes on e.g. the smoothness of q^* through a Gaussian probability measure $\mu_0 = \mathcal{N}(q_0, \mathcal{I}_0^{-1})$. In the following we give a brief and intuitive introduction to this regularization concept for inverse problems. As before we tend to keep this presentation short. For a more detailed discussion we refer to [80, 111, 250]. Denote by $q\colon D \to L^2(\Omega)$ the Gaussian random variable distributed according to μ_0. Now, instead of trying to compute a *point estimator* $\bar{q}^{\mathbf{y}_d} \in L^2(\Omega)$ fulfilling $X\bar{q}^{\mathbf{y}_d} = \mathbf{y}_d$ we construct a probability measure $\mu_{\text{post}}^{\mathbf{y}_d}$ on $L^2(\Omega)$ which takes into account the prior knowledge on the unknown parameter as well as the information provided by the collected data. As a first step we impose additional assumptions on the relation between the prior distribution of the parameter and the distribution of the measurement noise.

Assumption 5.2. The random field $q\colon D \to L^2(\Omega)$ and the measurement errors $\varepsilon\colon D \to \mathbb{R}^N$ are independent i.e. there holds

$$\mathbb{P}\left(q^{-1}(O_1) \cap \varepsilon^{-1}(O_2)\right) = \mu_0(O_1)\mu_E(O_2) \quad \forall O_1 \in \mathcal{B}(L^2(\Omega)),\ O_2 \in \mathcal{B}(\mathbb{R}^N).$$

Let us recall that the prior distribution $\mu_0 = \mathcal{N}(q_0, \mathcal{I}_0^{-1})$ is given by the push-forward of $\mathbb{P}$ under q. Thus we have

$$\mu_0(O) = \mathbb{P}(q^{-1}(O)) = \mathbb{P}\left(\{\, \omega \in D \mid q(\omega) \in O \,\}\right) \quad \forall O \in \mathcal{B}(L^2(\Omega)).$$

Loosely speaking we should interpret $\mu_0(O)$ as the probability that a particular realization of q is contained in a Borel set O. Formally we now define the probabilistic solution to the inverse problem (5.7) as

$$\mu_{\text{post}}^{\mathbf{y}_d}(O) = \int_O \frac{1}{\mathcal{Z}(\mathbf{y}_d)} \exp\left(-\frac{1}{2}|Xq - \mathbf{y}_d|^2_{\Sigma^{-1}}\right) \,\mathrm{d}\mu_0(q) \quad \forall O \in \mathcal{B}(L^2(\Omega)), \tag{5.8}$$

where $|\cdot|^2_{\Sigma^{-1}} = (\cdot, \Sigma^{-1}\cdot)_{\mathbb{R}^N}$ and the normalization constant $\mathcal{Z}(\mathbf{y}_d) > 0$ is given by

$$\mathcal{Z}(\mathbf{y}_d) = \int_{L^2(\Omega)} \exp\left(-\frac{1}{2}|Xq - \mathbf{y}_d|^2_{\Sigma^{-1}}\right) \,\mathrm{d}\mu_0(q).$$

Note that we have

$$0 \leq \int_O \exp\left(-\frac{1}{2}|Xq - \mathbf{y}_d|^2_{\Sigma^{-1}}\right) \,\mathrm{d}\mu_0(q) \leq \mu_0(O) \leq 1 \quad \forall O \in \mathcal{B}(L^2(\Omega)).$$

Thus $\mu^{\mathbf{y}_d}_{\text{post}}$ is a well-defined probability measure on $L^2(\Omega)$ if the normalization constant $\mathcal{Z}(\mathbf{y}_d)$ is bounded away from zero. In this case $\mu^{\mathbf{y}_d}_{\text{post}}$ is called the *posterior measure* or *posterior distribution* given the data $\mathbf{y}_d$. Let us give some interpretation to this definition. To this end we remark that $\mu^{\mathbf{y}_d}_{\text{post}}(O)$ is, up to a constant, given by the weighted integral of the characteristic function associated to O taken with respect to the prior distribution. The data-dependent weighting function incorporates the discrepancy of the predicted response of the model and the observed data

$$\pi(\mathbf{y}_d, \cdot)\colon L^2(\Omega) \to [0,1] \quad \text{where} \quad \pi(\mathbf{y}_d, q) = \begin{cases} 1 & Xq = \mathbf{y}_d \\ \exp\left(-\frac{1}{2}|Xq - \mathbf{y_d}|^2_{\Sigma^{-1}}\right) & \text{else} \end{cases}, \tag{5.9}$$

for all $q \in L^2(\Omega)$. The function $1/\mathcal{Z}\ \pi(\mathbf{y}_d, \cdot)$ is called the *Radon-Nikodým derivative* of $\mu^{\mathbf{y}_d}_{\text{post}}$ with respect to the prior distribution μ_0. Loosely speaking the weighting of the integral and its normalization lead to a probability measure whose mass is concentrated on Borel sets O with $\mu_0(O) > 0$ and on which $\pi(\mathbf{y}_d, \cdot) \approx 1$. Thus the posterior distribution indeed incorporates both the prior knowledge on the unknown parameter as well as the information provided by the data. This allows to make statements on the relative probability of an event in the parameter space provided that the particular data vector $\mathbf{y}_d$ was observed.

A mathematically rigorous justification of the definition in (5.8) can be based on the notion of conditional probability density functions. We sketch these ideas for the sake of completeness. To this end recall the assumption on the additivity of the measurement noise and its independence on the prior distribution. We define the $(\mathcal{F}, \mathcal{B}(\mathbb{R}^N))$-measurable function y_d with

$$y_d\colon D \to \mathbb{R}^N, \quad \omega \mapsto Xq(\omega) + \varepsilon(\omega). \tag{5.10}$$

We interpret y_d as a random variable. Its distribution is given by

$$\mu_{y_d} = \mathbb{P}(y_d^{-1}(\cdot)) = \mathcal{N}(Xq_0, X^*\mathcal{I}_0^{-1}X + \Sigma).$$

Note that its distribution depends on that of the measurement noise as well as the prior distribution of the random field q. This raises the following central question of Bayesian inference: *Given a realization $\mathbf{y}_d$ of the data y_d which conclusions can be drawn on the distribution of the random field q?* The answer to this question is given by the *conditional probability distribution* $\mu_{q|\mathbf{y}_d}$ describing the relative probability of events in the parameter space if we know that y_d attains the value $\mathbf{y}_d$.

The goal of the following considerations is to compute a closed form expression for this distribution. As a first step we therefore compute the probability measure $\mu_{y_d|\mathbf{q}}$ characterizing the distribution of the random variable y_d given an arbitrary but fixed function $\mathbf{q} \in L^2(\Omega)$ for the parameter. To this end we exploit the additivity of the noise and consider the random variable $y_d|\mathbf{q}\colon D \to \mathbb{R}^N$ given by

$$y_d|\mathbf{q}(\omega) = X\mathbf{q} + \varepsilon(\omega)$$

for $\mathbb{P}$-almost surely all $\omega \in D$. The conditional distribution of the data given knowledge on the parameter is now obtained as $\mu_{y_d|\mathbf{q}} = \mathbb{P}(y_d|\mathbf{q}^{-1}(\cdot))$. Since the noise ε is normally distributed according to $\mu_E = \mathcal{N}(0, \Sigma)$ we conclude that $\mu_{y_d|\mathbf{q}}$ is also a Gaussian with $\mu_{y_d|\mathbf{q}} = \mathcal{N}(X\mathbf{q}, \Sigma)$. In particular this implies

$$\mu_{y_d|\mathbf{q}}(O) = \int_O \frac{1}{\mathcal{Z}} \exp\left(-\frac{1}{2}|y - X\mathbf{q}|^2_{\Sigma^{-1}}\right) \, \mathrm{d}y = \int_O \frac{1}{\mathcal{Z}} \pi(y, \mathbf{q}) \, \mathrm{d}y \quad \forall O \in \mathcal{B}(\mathbb{R}^N),$$

where $\mathcal{Z} > 0$ is a normalization constant independent of $\mathbf{q} \in L^2(\Omega)$. The function $1/\mathcal{Z} \; \pi(\cdot, \mathbf{q})$ with

$$\pi(\cdot, \mathbf{q}) \colon \mathbb{R}^N \to [0,1], \quad \pi(y, \mathbf{q}) = \exp\left(-\frac{1}{2}|y - X\mathbf{q}|^2_{\Sigma^{-1}}\right) \quad \forall y \in \mathbb{R}^N$$

is called the *conditional probability density function* of y_d given $\mathbf{q}$.

For a fixed measurement vector $\mathbf{y}_d \in \mathbb{R}^N$ we now define the *likelihood function*

$$\pi(\mathbf{y}_d, \cdot) \colon L^2(\Omega) \to [0,1], \quad \pi(\mathbf{y}_d, q) = \exp\left(-\frac{1}{2}|Xq - \mathbf{y}_d|^2_{\Sigma^{-1}}\right)$$

for all $q \in L^2(\Omega)$. Note that this definition coincides with that of the weighting function in (5.9). The famous *Theorem of Bayes*, see e.g. [80, Theorem 14] , now states that $\mu_{q|\mathbf{y}_d}$ is absolutely continuous with respect to μ_0 and the Radon-Nikodým derivative is given by the scaled likelihood function. More in detail we obtain

$$\mu_{q|\mathbf{y}_d}(O) = \frac{\int_O \pi(\mathbf{y}_d, q) \, \mathrm{d}\mu_0(q)}{\int_{L^2(\Omega)} \pi(\mathbf{y}_d, q) \, \mathrm{d}\mu_0(q)} \quad \forall O \in \mathcal{B}(L^2(\Omega)).$$

Substituting the definition of the likelihood we recover the posterior distribution from (5.8).

We summarize our findings in the following theorem. In particular the previous observations imply that, in the present case, the posterior measure is Gaussian. Thus it is completely characterized through its mean and covariance operator.

Theorem 5.8. *Let $\mathbf{y}_d \in \mathbb{R}^N$ be given and assume that $q_0 \in \mathcal{H}$. Then $\mu^{\mathbf{y}_d}_{post}$ as given by (5.8) is a well-defined Gaussian probability measure on $L^2(\Omega)$ with*

$$\mu^{\mathbf{y}_d}_{post} = \mathcal{N}(q^{\mathbf{y}_d}_{post}, \mathcal{C}_{post}).$$

The posterior mean $q^{\mathbf{y}_d}_{post} \in \mathcal{H}$ and covariance operator $\mathcal{C}_{post} \in \mathcal{L}(\mathcal{H}^*, \mathcal{H})$ are given by

$$q^{\mathbf{y}_d}_{post} = q_0 + \mathcal{C}_{post} X^* \Sigma^{-1}(\mathbf{y}_d - Xq_0) \in \mathcal{H}, \quad \mathcal{C}_{post} = (X^* \Sigma^{-1} X + \mathcal{I}_0)^{-1} \in \mathcal{L}(L^2(\Omega), L^2(\Omega)).$$

Proof. These statements can be concluded directly from Example 6.23 and Theorem 6.31 in [250] noting that $\mu_0(L^2(\Omega)) = 1$. □

To close this short introduction we briefly recap the Bayesian approach to inverse problems and point out to its limitations. Recall that the starting point of our considerations was given by the ill-posed deterministic inverse problem (5.7). In order to obtain a well-posed formulation we resorted to a description of the prior believes on the parameter in terms of a random field. This can be viewed as a probabilistic regularization of the problem in which we describe the uncertainty on the true value of the parameter by a probability measure. Applying Bayesian inference to the problem should then be interpreted as a *learning process* in which we re-evaluate our current knowledge on the distribution of the random field based on the obtained measurements and thus reduce this uncertainty.

First we again stress that the solution to the Bayesian inverse problem is given by a probability measure over the parameter space and not by a single function. This allows for *probabilistic* statements on the unknown parameter rather than *deterministic* ones. For example the measure of a ball in the parameter space with respect to the posterior describes our degree of certainty that the observed data $\mathbf{y}_d$ corresponds to the response of the mathematical model X for some parameter $\mathbf{q}$ inside this ball. However it does not allow to draw conclusions on the plausibility of particular realizations of the random field since

$$\mu_0(\{\mathbf{q}\}) = \mu_{post}^{\mathbf{y}_d}(\{\mathbf{q}\}) = 0 \quad \forall \mathbf{q} \in L^2(\Omega).$$

Nevertheless, from a practical point of view, it would be desirable to define a point estimator reflecting our belief on the most likely value of the parameter given the obtained data. Following [79] one possibility to do so is the consideration of minimizers to the *Onsager-Machlup* functional

$$\min_{q \in \mathcal{H}} \frac{1}{2}|Xq - \mathbf{y}_d|^2_{\Sigma^{-1}} + \frac{1}{2}\|q - q_0\|^2_{\mathcal{H}}. \tag{5.11}$$

Note that this minimization problem resembles a Tikhonov regularization of the inverse problem (5.7) for the particular case of choosing the Cameron-Martin norm as regularization term. Obviously, in the present case, the global minimizer to this problem is unique and coincides with the mean $q_{\text{post}}^{\mathbf{y}_d}$ of the posterior distribution. We call it the *maximum a posteriori probability* estimator or, to shorten, the *MAP*.

These considerations clearly highlight the importance of properly choosing the prior distribution and its tremendous influence on the obtained results. A first restriction on its choice is given by the well-established assumption of a Gaussian prior distribution. This implies that the eigenvalues of $\mathcal{I}_0^{-1}$ are summable. In particular the elements of the associated Cameron-Martin space will exhibit additional smoothness beyond $L^2(\Omega)$ regularity. To illustrate this fact, we pick up on Example 5.1 and $\Omega = (0,1)^d$. The Cameron-Martin space $\mathcal{H}$ associated to $\mathcal{I}_0^{-1} = (-\Delta)^{-s}$, $s > d/2$, is given by the Sobolev space $H^s \cap H_0^1(\Omega)$. Thus, due to the Sobolev embedding theorem, [85, Theorem 8.2], we get Hölder regularity of the mean $q_{\text{post}}^{\mathbf{y}_d}$. Moreover, while random draws from e.g. $\mathcal{N}(0, (-\Delta)^{-s})$ are almost surely not contained in $\mathcal{H}$, [99, Proposition 4.22], the *Kolmogorov continuity theorem* ensures (almost surely) their Hölder regularity in this case. For a reference we point out to Theorem 6.24 and Lemma 6.25 in [250]. This makes an application of the Bayesian approach based on Gaussian priors questionable if we expect the true parameter q^* to be e.g. piecewise constant. Additionally we emphasize that a prior distribution which encompasses all structural features of the unknown parameter would render Bayesian inference obsolete. Loosely speaking, this observation implies that the choice of the prior has to be, at least partly, arbitrary.

The main focus of this chapter lies on the development of a sensor placement framework for Bayesian inverse problems. In particular we assume that a prior distribution $\mu_0 = \mathcal{N}(q_0, \mathcal{I}_0^{-1})$ which is suitable for the problem at hand is already provided. Thus we do not comment further on this topic. However this critical point on the Bayesian approach should always be kept in mind throughout the following considerations. For further discussion on the matter of choosing the prior distribution in a sophisticated way we direct the reader to Chapters 3 and 10 in [228].

5.1.2 Bayesian inference for PDE constrained problems

In this section we apply the presented Bayesian methodology to the inverse problems of identifying a distributed parameter q entering a partial differential equation from observations $\mathbf{y}_d$ of the state. More concretely, we consider the form

$$a(\cdot,\cdot)(\cdot)\colon Q_{ad}\times Y\times W\to\mathbb{R},$$

which depends linearly on the elements of the second bracket but may dependent nonlinearly on those embraced by the first one. The set of admissible parameters Q_{ad} is given as a subset of $L^2(\Omega)$, where $\Omega\subset\mathbb{R}^d$, $d\in\mathbb{N}$, is open and bounded. The state space Y and the test space W are assumed to be reflexive Banach-spaces. Given $q\in Q_{ad}$ an element $y=S[q]\in Y$ is called the state associated to q if there holds

$$a(q,y)(\varphi)=0\quad\forall\varphi\in W. \tag{5.12}$$

We make the following assumptions on its existence and regularity.

Assumption 5.3. The equation (5.12) admits a unique solution $y=S[q]\in Y$ for every $q\in Q_{ad}$. Furthermore the parameter-to-state operator

$$S\colon Q_{ad}\to Y,\quad q\mapsto S[q],$$

is at least continuously Fréchet differentiable in $L^2(\Omega)$ on a neighbourhood of Q_{ad}.

The observations $\mathbf{y}_d$ will be obtained by taking pointwise measurements of the physical quantity represented by the state y. To this end we assume $Y\overset{c}{\hookrightarrow}\mathcal{C}(\Omega_o)$. Here, $\Omega_o\subset\Omega$ is a compact set of possible sensor locations. As in the finite dimensional situation of the previous chapter given $q\in Q_{ad}$, $y=S[q]$ and $\delta q\in L^2(\Omega)$ the associated sensitivity $\delta y=\partial S[q]\delta q\in Y$ is the unique element fulfilling

$$a'_y(q,y)(\delta y,\varphi)=-a'_q(q,y)(\delta q,\varphi)\quad\forall\varphi\in W, \tag{5.13}$$

given sufficient regularity of the weak form $a(\cdot,\cdot)(\cdot)$. Here a'_y and a'_q denote the partial derivatives of the form a with respect to the state and the parameter respectively. The following examples aim to illustrate this abstract setting. In all of them we consider a bounded domain $\Omega\subset\mathbb{R}^d$, $d\leq 3$ which we assume to be convex and thus with Lipschitz boundary. Furthermore the state and test spaces are chosen as

$$Y=H^2(\Omega)\cap H^1_0(\Omega)\overset{c}{\hookrightarrow}\mathcal{C}(\Omega_o),\quad W=L^2(\Omega).$$

Example 5.2. *We consider the identification of an unknown source term $q\in L^2(\Omega)$ entering the right hand side of a Poisson equation together with homogeneous Dirichlet-boundary conditions*

$$-\Delta y=q\quad in\ \Omega,\quad y=0\quad on\ \partial\Omega.$$

The admissible set of parameters is fixed to $Q_{ad}=L^2(\Omega)$ and the form a is chosen such that

$$a(q,y)(\varphi)=\int_\Omega[(-\Delta y-q)\varphi]\mathrm{d}x=0\quad\forall\varphi\in L^2(\Omega). \tag{5.14}$$

Given $q\in L^2(\Omega)$ it is well-known that (5.14) admits a unique solution $y=S[q]\in H^2(\Omega)\cap H^1_0(\Omega)$, see [124, Theorem 3.2.12]. The parameter-to-state operator

$$S\colon L^2(\Omega)\to H^2(\Omega)\cap H^1_0(\Omega),\quad q\mapsto S[q]$$

is linear and continuous. As a consequence we conclude $\partial S[\hat{q}]\delta q=S[\delta q]$ for all $\hat{q},\ \delta q\in L^2(\Omega)$.

Example 5.3. *As a second example we aim to identify an unknown diffusion coefficient. To this end we define the admissible set of parameters*

$$Q_{ad} = \left\{ q \in \mathcal{C}^{0,1}(\bar{\Omega}) \mid q > 0 \right\} \subset L^2(\Omega).$$

The considered PDE is given in its weak form as

$$a(q,y)(\varphi) = \int_\Omega [(-\nabla \cdot (q\nabla y) - f)\,\varphi]\;\mathrm{d}x = 0 \quad \forall \varphi \in L^2(\Omega).$$

The source term $f \in L^2(\Omega)$ is assumed to be known. Following the arguments in [58] there exists a unique solution $y = S[q] \in H^2(\Omega) \cap H_0^1(\Omega)$ given $q \in Q_{ad}$. The parameter-to-state operator S is Fréchet differentiable in a neighborhood of Q_{ad} with respect to the topology on $\mathcal{C}^{0,1}(\bar{\Omega})$. Given a direction $\delta q \in \mathcal{C}^{0,1}(\bar{\Omega})$ and $\hat{q} \in Q_{ad}$ the sensitivity $\delta y = \partial S[\hat{q}]\delta q$ is the unique element in $H^2(\Omega) \cap H_0^1(\Omega)$ fulfilling

$$\int_\Omega -\nabla \cdot (\hat{q}\nabla \delta y)\varphi\;\mathrm{d}x = \int_\Omega \nabla \cdot (\delta q \nabla S[\hat{q}])\varphi\;\mathrm{d}x \quad \forall \varphi \in L^2(\Omega)$$

At this point we note that the operator $\partial S[\hat{q}]$ cannot be extended to a linear continuous operator on $L^2(\Omega)$. Thus this problem cannot be fit directly into the framework considered in this chapter. As a possible workaround we propose to consider a reparametrization of the parameter $q \in \mathcal{C}^{0,1}(\bar{\Omega})$ as

$$q(x) = \exp([Tp](x)), \quad p \in L^2(\Omega),\ \textit{a.e. } x \in \Omega_o, \quad T\colon L^2(\Omega) \to \mathcal{C}^{0,1}(\bar{\Omega}),$$

where the operator T is e.g. a sufficiently smoothing convolution operator. The exponential function is applied to get rid of the positivity constraints in the parameter space. In the same moment we stress that such discrepancies between the topology on the parameter space $L^2(\Omega)$ and the topology needed to ensure differentiability of S are characteristic for e.g. parameter-to-state mappings corresponding to nonlinear PDEs. Thus a rigorous extension of the approach presented in this chapter in order to cover these cases should be in the focus of future research.

Example 5.4. *Last we consider an unknown parameter q in the reaction term of a linear elliptic PDE given by*

$$-\Delta y + qy = f \quad \textit{in } \Omega, \quad y = 0 \quad \textit{on } \partial\Omega.$$

The source term $f \in L^2(\Omega)$ is again assumed to be known and the set of admissible parameters is defined as

$$Q_{ad} = \left\{ q \in L^2(\Omega) \mid \|q^-\|_{L^2(\Omega)} < c_{Q_{ad}} \right\},$$

where $q^-(x) = -\min\{0, q(x)\}$, for a.e. $x \in \Omega$, denotes the negative part of $q \in L^2(\Omega)$. The value of the constant $c_{Q_{ad}} > 0$ will be fixed in an instance. The associated weak formulation of the PDE is now given by

$$a(q,y)(\varphi) = \int_\Omega [(-\Delta y + qy - f)\varphi]\;\mathrm{d}x = 0 \quad \forall \varphi \in L^2(\Omega). \tag{5.15}$$

In the following we will briefly prove that (5.15) *admits a unique solution in $H^2(\Omega) \cap H_0^1(\Omega)$. To this end let us first consider the form*

$$b\colon Q_{ad} \times H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}, \quad b(q)(y,\varphi) = \int_\Omega [(\nabla y \cdot \nabla \varphi + qy\varphi]\;\mathrm{d}x.$$

For $y,\ \varphi \in H_0^1(\Omega)$ and $q \in Q_{ad}$ we immediately infer

$$b(q)(y,y) = \|y\|^2_{H_0^1(\Omega)} + \int_\Omega qy^2 \,\mathrm{d}x \geq \|y\|^2_{H_0^1(\Omega)} - \|q^-\|_{L^2(\Omega)}\|y\|^2_{L^4(\Omega)} \geq (1 - c_\Omega\|q^-\|_{L^2(\Omega)})\|y\|^2_{H_0^1(\Omega)},$$

where the constant $c_\Omega > 0$ depends on the domain. Second we obtain

$$b(q)(y,\varphi) \leq \|y\|_{H_0^1(\Omega)}\|\varphi\|_{H_0^1(\Omega)} + \|q\|_{L^2(\Omega)}\|y\|_{L^4(\Omega)}\|\varphi\|_{L^4(\Omega)} \leq (1 + c_\Omega\|q\|_{L^2(\Omega)})\|y\|_{H_0^1(\Omega)}\|\varphi\|_{H_0^1(\Omega)}.$$

If we choose $c_{Q_{ad}} \leq 1/c_\Omega$ then an application of the Lax-Milgram lemma yields the existence of a unique function $y_f \in H_0^1(\Omega)$ fulfilling

$$b(q)(y_f,\varphi) = \langle \varphi, f\rangle_{H_0^1(\Omega),H^{-1}} \quad \forall \varphi \in H_0^1(\Omega),$$

for every $f \in H^{-1}$. By a bootstrapping argument it is now readily verified that $y_f \in H^2(\Omega) \cap H_0^1(\Omega)$ whenever $f \in L^2(\Omega)$. Furthermore it is readily verified that the operator

$$-\Delta + q\,\mathrm{Id}\colon H^2(\Omega) \cap H_0^1(\Omega) \to L^2(\Omega),$$

is an isomorphism. Applying the implicit function theorem, see e.g. [86], yields the existence of an operator

$$S\colon Q_{ad}\colon H^2(\Omega) \cap H_0^1(\Omega), \quad q \mapsto S[q],$$

where $y = S[q]$ is the unique solution to (5.15). *The mapping S is at least of class $\mathcal{C}^1$ in a neighborhood of Q_{ad}. Given a linearization point $\hat{q} \in Q_{ad}$ and a direction $\delta q \in L^2(\Omega)$ the associated sensitivity $\delta y = \partial S[\hat{q}]\delta q$ is the unique element in $H^2(\Omega) \cap H_0^1(\Omega)$ fulfilling*

$$\int_\Omega [(-\Delta\,\delta y + q\delta y)\varphi] \,\mathrm{d}x = \int_\Omega \delta q S[\hat{q}]\varphi \,\mathrm{d}x \quad \forall \varphi \in L^2(\Omega)$$

Let us now return to the discussion of the general case. The true parameter, i.e. the distributed function describing the model most faithfully, will be denoted by $q^* \in Q_{ad}$. The point measurements of the state y are taken at a finite number of sensors located at $\{x_i\}_{i=1}^N \subset \Omega_o$, $N \in \mathbb{N}$, and the obtained measurements are assembled in a vector $\mathbf{y}_d \in \mathbb{R}^n$. To take measurement errors into account we assume that no systematic model errors are present, i.e. the "true" measurement at a point $x \in \Omega_o$ is given by $S[q^*](x)$, and the measurements are perturbed by additive noise stemming from the sensors. For abbreviation, given $q \in Q_{ad}$, we will write $S[q](x) \in \mathbb{R}^N$ for the vector of observations with $S[q](x)_i = S[q](x_i)$, $i = 1, \ldots, N$, in the following and define

$$y_d\colon Q_{ad} \times \mathbb{R}^N \to \mathbb{R}^N, \quad (q,\epsilon) \mapsto S[q](x) + \epsilon. \tag{5.16}$$

The obtained measurements are given by $\mathbf{y}_d = y_d(q^*,\epsilon)$ for some $\epsilon \in \mathbb{R}^N$. Let $(D, \mathcal{F}, \mathbb{P})$ be a probability space. We adopt a probabilistic description of the measurement error and interpret ϵ as a realization of an N-dimensional Gaussian random variable $\varepsilon\colon D \to \mathbb{R}^N$ with $\varepsilon \sim \mathcal{N}(0,\Sigma)$ where $\Sigma \in \mathrm{Sym}(N)$, $\Sigma_{ij} = \delta_{ij}/\mathbf{u}_i$. The constant $\mathbf{u}_i$ describes how carefully the measurement at x_i should be performed. For example, if $\mathbf{u}_i$ is an integer, it might resemble the total number of measurements at the same position. More general $\mathbf{u}_i$ corresponds to the quality of the used sensor i.e. the reciprocal of its measurement error. The unknown parameter is now determined by matching the collected data with the predicted response of the mathematical model

$$\text{find } q \in Q_{ad}\colon \quad S[q](x) = \mathbf{y}_d \tag{5.17}$$

In the following we simplify the problem by considering a first-order approximation of the underlying model around a sophisticated a priori guess $\hat{q} \in Q_{ad}$

$$S[q] \approx S[\hat{q}] + \partial S[\hat{q}](q - \hat{q}), \quad q \in Q_{ad}.$$

In the same manner we linearize the mapping in (5.16) and assume

$$y_d(q, \epsilon) = S[\hat{q}](x) + X(q - \hat{q}) + \epsilon,$$

where the operator $X \in \mathcal{L}(L^2(\Omega_o), \mathbb{R}^N)$ is defined through

$$X \colon L^2(\Omega) \to \mathbb{R}^N, \quad (Xq)_i = \partial S[\hat{q}]q(x_i).$$

Finally we drop the constraints on the admissible set of parameters and formulate the linearized inverse problem as

$$\text{find } q \in L^2(\Omega)\colon \quad S[\hat{q}](x) + X(q - \hat{q}) = \mathbf{y}_d. \tag{5.18}$$

Despite its linearity the inverse problem in (5.18) is still ill-posed due to the finite dimensionality of the collected data. To obtain a well-defined problem, we adopt the Bayesian viewpoint discussed in the previous section to the problem. The uncertainty on the true value of the parameter is modeled as a Gaussian random field $q\colon D \to L^2(\Omega)$ distributed according to $\mu_0 = \mathcal{N}(\hat{q}, \mathcal{I}_0^{-1})$. Here, $\mathcal{I}_0^{-1}$ is a known covariance operator given by the inverse of an unbounded operator, see Lemma 5.5 and Assumption 5.1. Its eigenvalues and the associated eigenfunctions are denoted by $\{\lambda_i\}_{i\in\mathbb{N}}$ and $\{\phi_i\}_{i\in\mathbb{N}}$, respectively. We silently assume that the linearization point $\hat{q} \in Q_{ad}$ is an element of the corresponding Cameron-Martin space $\mathcal{H}$ given by

$$\mathcal{H} = \left\{ q \in L^2(\Omega) \mid \|\mathcal{I}_0^{1/2} q\|_{L^2(\Omega)} < \infty \right\}.$$

We assume that the random field distributed according to μ_0 and the measurement noise ε are independent. As an illustrative example the reader may recall the situation discussed in Example 5.1 and consider the prior covariance operators defined through the inversion of (fractional) differential operator, e.g. $\mathcal{I}_0^{-1} = (-\Delta)^{-s}$, $s > d/2$, on $\bar{\Omega} = [0,1]^2$. Here $-\Delta$ denotes the Dirichlet Laplacian. The corresponding Cameron-Martin space is given by $H^s(\Omega) \cap H_0^1(\Omega)$. However we also stress that the following analysis is not restricted to this setting.

Subsequently, the knowledge on the parameter is updated based on the collected data $\mathbf{y}_d$. The probabilistic solution of (5.18) is given by the posterior distribution

$$\begin{aligned}
\mu_{\text{post}}^{\mathbf{y}_d}(O) &= \int_O \frac{1}{\mathcal{Z}(\mathbf{y}_d)} \exp\left(-\frac{1}{2} |S[\hat{q}](x) + X(q - \hat{q}) - \mathbf{y}_d|^2_{\Sigma^{-1}} \right) \, \mathrm{d}\mu_0(q) \quad \forall O \in \mathcal{B}(L^2(\Omega)),\\
\mathcal{Z}(\mathbf{y}_d) &= \int_{L^2(\Omega)} \exp\left(-\frac{1}{2} |S[\hat{q}](x) + X(q - \hat{q}) - \mathbf{y}_d|^2_{\Sigma^{-1}} \right) \, \mathrm{d}\mu_0(q).
\end{aligned}$$

Due to the linearity of the model it is again a Gaussian measure, cf. Theorem 5.8, with

$$\mu_{post}^{\mathbf{y}_d} = \mathcal{N}(q_{post}^{\mathbf{y}_d}, \mathcal{C}_{post}) \quad \text{where} \quad \mathcal{C}_{post} = (X^* \Sigma^{-1} X + \mathcal{I}_0)^{-1} \tag{5.19}$$

and the posterior mean $q_{post}^{\mathbf{y}_d} \in \mathcal{H}$ is the unique minimizer to

$$\min_{q\in\mathcal{H}} \frac{1}{2} |S[\hat{q}](x) + X(q - \hat{q}) - \mathbf{y}_d|^2_{\Sigma^{-1}} + \frac{1}{2} \|\mathcal{I}_0^{1/2}(q - \hat{q})\|^2_{L^2(\Omega)}.$$

It admits an explicit representation as

$$q^{\mathbf{y}_d}_{post} = \hat{q} + \mathcal{C}_{post}(X^* \varSigma^{-1}(\mathbf{y}_d - S[\hat{q}](x))). \tag{5.20}$$

We emphasize that for a fixed a priori guess $\hat{q}$ the posterior covariance operator does not rely on the measurement vector $\mathbf{y}_d \in \mathbb{R}^N$. However its depends on the measurement points $\{x_i\}_{i=1}^N$ and $\mathbf{u}_i$, $i = 1, \dots, N$, through the Fisher information operator

$$\mathcal{I}(\boldsymbol{u}(\mathbf{x}, \mathbf{u})) = X^* \varSigma^{-1} X \in \mathcal{L}(L^2(\varOmega), L^2(\varOmega)).$$

To stress this dependence we denote the posterior covariance by $\mathcal{C}_{post}(\mathcal{I}(\boldsymbol{u}(\mathbf{x}, \mathbf{u})))$ in the following. Note that the Fisher information is positive and Hilbert-Schmidt on $L^2(\varOmega)$. The latter property follows due to the finite number of measurements.

Remark 5.4. We point out that it is also possible to consider the nonlinear inverse problem from (5.17) in the Bayesian context. As in the linear case we formally define the posterior measure of q given $\mathbf{y}_d$ by

$$\begin{aligned} \mu^{\mathbf{y}_d}_{\text{post}}(O) &= \int_O \frac{1}{\mathcal{Z}(\mathbf{y}_d)} \exp\left(-\frac{1}{2}|S[q](x) - \mathbf{y}_d|^2_{\varSigma^{-1}}\right) \, \mathrm{d}\mu_0(q) \quad \forall O \in \mathcal{B}(L^2(\varOmega)), \\ \mathcal{Z}(\mathbf{y}_d) &= \int_{L^2(\varOmega)} \exp\left(-\frac{1}{2}|S[q](x) - \mathbf{y}_d|^2_{\varSigma^{-1}}\right) \, \mathrm{d}\mu_0(q). \end{aligned} \tag{5.21}$$

Imposing additional assumptions on S one can show that the relation in (5.21) indeed defines a probability measure, see [250, Chapter 4]. However, since S is non-linear, this posterior measure is in general not Gaussian. While the optimal design criteria presented in the upcoming section still remain meaningful in this situation, see e.g. [5], they usually do not admit a closed form representation adding an additional level of complexity to the problem. Since such formulations are out of the scope of this thesis, we will not comment further on this topic, however we stress their relevance for future research. In this light, the proposed approach based on a linearization of the underlying PDE can be interpreted as a Gaussian approximation to the true posterior measure.

5.1.3 Optimal design criteria for distributed parameters

As already stressed at several points in this chapter the solution to the Bayesian inverse problem is given by the posterior distribution. This probability measure summarizes the current knowledge or, equivalently, the remaining degree of uncertainty on the unknown parameter given the vector of measurements $\mathbf{y}_d$. A complete discussion of the Bayesian inverse problem requires the quantification of both the uncertainty in the estimate and the obtained amount of information. This section aims to illustrate this process of *uncertainty quantification.*

To this end, a first goal of this section is to present several scalar-valued functions which quantify the statistic properties of the posterior distribution. Surprisingly these well-established measures of uncertainty depend exclusively on the prior distribution and the Fisher information operator. In particular, they do not depend on the vector of measurements but are parametrized by the positions $\{x_i\}_{i=1}^N$ of the measurement sensors and the diligence factors $\{\mathbf{u}_i\}_{i=1}^N$. Thus, similar to the finite dimensional situation in the previous chapter, they can serve as design criteria to compare the statistic quality of different sensor configurations *before* any measurements are carried out in practice.

In the following sections we then proceed to the formulation and analysis of sensor placement problems associated to the discussed Bayesian inverse problem. In this context we improve the estimation process a priori, i.e. before any measurements are carried out, by optimizing one of the presented design criteria with respect to the measurement setup. To this end we identify mathematical properties that are common to all of the considered functionals. This will enable us to treat the corresponding sensor placement problems in a rigorous and unified way in the following sections.

The a posteriori covariance operator

As in the finite dimensional setting of Chapter 4 the covariance operator $\mathcal{C}_{\text{post}}(\mathcal{I}(\boldsymbol{u}(\mathbf{x},\mathbf{u})))$ will play a major role in the following discussions. Let us first fix some notation. The topological dual space of $\mathcal{H}$ will be denoted by $\mathcal{H}^*$ in the following. By definition of $\mathcal{H}$ the Riesz-isomorphism $T_{\mathcal{H}}\colon \mathcal{H} \to \mathcal{H}^*$ is readily identified with $\mathcal{I}_0$ and

$$\langle \delta q_1, \delta q_1^* \rangle_{\mathcal{H},\mathcal{H}^*} = (\delta q_1, T_{\mathcal{H}}^{-1} \delta q_1^*)_{\mathcal{H}} = (\delta q_1, \mathcal{I}_0^{-1} \delta q_1^*)_{\mathcal{H}} \quad \forall \delta q_1 \in \mathcal{H},\ \delta q_1^* \in \mathcal{H}^*.$$

On $\mathcal{H}^*$ a Hilbert space structure is induced by the inner product

$$(\delta q_1^*, \delta q_2^*)_{\mathcal{H}^*} = \langle \mathcal{I}_0^{-1} \delta q_1^*, \delta q_2^* \rangle_{\mathcal{H},\mathcal{H}^*} = (\mathcal{I}_0^{-1} \delta q_1^*, \mathcal{I}_0^{-1} \delta q_2^*)_{\mathcal{H}} = (\mathcal{I}_0^{-1/2} \delta q_1^*, \mathcal{I}_0^{-1/2} \delta q_2^*)_{L^2(\Omega)},$$

for all $\delta q_1^*,\ \delta q_2^* \in \mathcal{H}^*$. The space $\mathcal{H}$ together with its topological dual and $L^2(\Omega)$ form a Gelfand-triple

$$\mathcal{H} \overset{c}{\hookrightarrow} L^2(\Omega) \simeq L^2(\Omega)^* \hookrightarrow \mathcal{H}^*,$$

where the first embedding, and thus the second, is compact and dense. As a consequence, given the eigenfunctions $\{\phi_i\}_{i\in\mathbb{N}}$ of $\mathcal{I}_0^{-1}$, the sets

$$\{\mathcal{I}_0^{-1/2} \phi_i\}_{i\in\mathbb{N}} \subset \mathcal{H}, \quad \{\phi_i\}_{i\in\mathbb{N}} \subset L^2(\Omega), \quad \{\mathcal{I}_0^{1/2} \phi_i\}_{i\in\mathbb{N}} \subset \mathcal{H}^*,$$

form orthonormal bases with respect to the inner product on the respective spaces. Thus the Hilbert-Schmidt norm of $B \in \mathcal{L}(\mathcal{H},\mathcal{H}^*)$ is given by

$$\|B\|^2_{\text{HS}(\mathcal{H},\mathcal{H}^*)} = \sum_{i=1}^{\infty} \|B\mathcal{I}_0^{-1/2}\phi_i\|^2_{\mathcal{H}^*} = \|\mathcal{I}_0^{-1/2} B \mathcal{I}_0^{-1/2}\|_{\text{HS}(L^2(\Omega),L^2(\Omega))}.$$

Given a Hilbert-Schmidt operator $B \in \text{HS}(L^2(\Omega), L^2(\Omega))$ we immediately infer

$$\|B\|_{\text{HS}(\mathcal{H},\mathcal{H}^*)} = \|\mathcal{I}_0^{-1/2} B \mathcal{I}_0^{-1/2}\|_{\text{HS}(L^2(\Omega),L^2(\Omega))} \leq \|\mathcal{I}_0^{-1}\|_{\mathcal{L}(L^2(\Omega),L^2(\Omega))} \|B\|_{\text{HS}(L^2(\Omega),L^2(\Omega))},$$

i.e the spaces $\text{HS}(L^2(\Omega), L^2(\Omega))$ and $\text{SHS}(L^2(\Omega), L^2(\Omega))$ continuously embed into $\text{HS}(\mathcal{H},\mathcal{H}^*)$ and $\text{SHS}(\mathcal{H},\mathcal{H}^*)$, respectively.

We start by taking a closer look at properties of the posterior covariance operator and study well-posedness of the mapping

$$\mathcal{C}_{\text{post}}\colon\ \text{SHS}(\mathcal{H},\mathcal{H}^*) \to \mathcal{L}(\mathcal{H}^*,\mathcal{H}), \quad B \mapsto (B+\mathcal{I}_0)^{-1}, \tag{5.22}$$

as well as its differentiability properties. To this end we adopt a variational description of $B + \mathcal{I}_0$ given $B \in \mathrm{SHS}(\mathcal{H}, \mathcal{H}^*)$. This operator induces a symmetric bilinear form

$$a[B] \colon \mathcal{H} \times \mathcal{H} \to \mathbb{R}, \quad a[B](q_1, q_2) = \langle q_1, Bq_2 \rangle_{\mathcal{H},\mathcal{H}^*} + (q_1, q_2)_{\mathcal{H}} \quad \forall q_1,\ q_2 \in \mathcal{H}.$$

Given $f \in \mathcal{H}^*$ we consider the variational problem of finding $q_f \in \mathcal{H}$ with

$$a[B](q_f, q_2) = \langle q_2, f \rangle_{\mathcal{H},\mathcal{H}^*} \quad \forall q_2 \in \mathcal{H}. \tag{5.23}$$

As a first step we establish well-posedness of this covariance equation under mild assumptions.

Proposition 5.9. *Let $B \in \mathrm{SHS}(\mathcal{H}, \mathcal{H}^*)$ be given. Then there exists a constant $c_B \geq 0$ with*

$$\langle q_1, Bq_1 \rangle_{\mathcal{H},\mathcal{H}^*} \geq -c_B \|q_1\|^2_{\mathcal{H}} \quad \forall q_1 \in \mathcal{H}.$$

If $c_B < 1$ then equation (5.23) admits a unique solution for every $f \in \mathcal{H}^*$. The operator

$$\mathcal{C}_{post}(B) \colon \mathcal{H}^* \to \mathcal{H}, \quad f \mapsto q_f,$$

is linear and continuous with $\|\mathcal{C}_{post}(B)\|_{\mathcal{L}(\mathcal{H}^*,\mathcal{H})} \leq 1/(1 - c_B)$. If $B \in \mathrm{Pos}(\mathcal{H}, \mathcal{H}^*)$ we can choose $c_B = 0$ and there holds $\|\mathcal{C}_{post}(B)\|_{\mathcal{L}(\mathcal{H}^*,\mathcal{H})} \leq 1$.

Furthermore, given $B_1, B_2 \in \mathrm{SHS}(\mathcal{H}, \mathcal{H}^*)$ with $c_{B_1}, c_{B_2} < 1$, there holds

$$\|\mathcal{C}_{post}(B_1) - \mathcal{C}_{post}(B_2)\|_{\mathcal{L}(\mathcal{H}^*,\mathcal{H})} \leq \frac{\|B_1 - B_2\|_{\mathcal{L}(\mathcal{H},\mathcal{H}^*)}}{(1 - c_{B_1})(1 - c_{B_2})}.$$

Proof. Let $B \in \mathrm{SHS}(\mathcal{H}, \mathcal{H}^*)$ be given. The claimed existence of $c_B \geq 0$ follows immediately. For $f \in \mathcal{H}^*$ and $q_1, q_2 \in \mathcal{H}$ we have

$$a[B](q_1, q_2) = \langle q_1, Bq_2 \rangle_{\mathcal{H},\mathcal{H}^*} + (q_1, q_2)_{\mathcal{H}} \leq (\|B\|_{\mathcal{L}(\mathcal{H},\mathcal{H}^*)} + 1)\|q_1\|_{\mathcal{H}}\|q_2\|_{\mathcal{H}},$$

as well as

$$a[B](q_1, q_1) = \langle q_1, Bq_1 \rangle_{\mathcal{H},\mathcal{H}^*} + (q_1, q_1)_{\mathcal{H}} \geq (1 - c_B)\|q_1\|^2_{\mathcal{H}},$$

by the assumptions on B. Hence applying Lax-Milgram Lemma, see [52, Corollary 5.8], yields the existence of a unique solution $q_f = \mathcal{C}_{\mathrm{post}}(B)f \in \mathcal{H}$ to equation (5.23) with

$$(1 - c_B)\|q_f\|^2_{\mathcal{H}} \leq a[B](q_f, q_f) = \langle q_f, f \rangle_{\mathcal{H},\mathcal{H}^*} \leq \|q_f\|_{\mathcal{H}}\|f\|_{\mathcal{H}^*}.$$

This implies the desired estimate. If $B \in \mathrm{Pos}(\mathcal{H}, \mathcal{H}^*)$ we can choose $c_B = 0$ yielding the estimate $\|\mathcal{C}_{\mathrm{post}}(B)\|_{\mathcal{L}(\mathcal{H}^*,\mathcal{H})} \leq 1$.

Let $B_1, B_2 \in \mathrm{SHS}(\mathcal{H}, \mathcal{H}^*)$ with $c_{B_1},\ c_{B_2} < 1$ and $f \in \mathcal{H}^*$ be given. Define $q_f^{B_1} = \mathcal{C}_{\mathrm{post}}(B_1)f$, $q_f^{B_2} = \mathcal{C}_{\mathrm{post}}(B_2)f$ and the difference $\delta q_f = q_f^{B_1} - q_f^{B_2}$, respectively. We conclude

$$\begin{aligned}
(1 - c_{B_1})\|\delta q_f\|^2_{\mathcal{H}} &\leq a[B_1](\delta q_f, \delta q_f) = a[B_2](q_f^{B_2}, \delta q_f) - a[B_1](q_f^{B_2}, \delta q_f) \\
&= \langle q_f^{B_2}, (B_2 - B_1)\delta q_f \rangle_{\mathcal{H},\mathcal{H}^*} \\
&\leq \|q_f^{B_2}\|_{\mathcal{H}}\|\delta q_f\|_{\mathcal{H}}\|B_1 - B_2\|_{\mathcal{L}(\mathcal{H},\mathcal{H}^*)} \\
&\leq \frac{\|f\|_{\mathcal{H}^*}\|\delta q_f\|_{\mathcal{H}}}{1 - c_{B_2}}\|B_1 - B_2\|_{\mathcal{L}(\mathcal{H},\mathcal{H}^*)}.
\end{aligned}$$

This proves the Lipschitz-stability of the covariance mapping. □

Note that if $c_B < 1$ the operator $\mathcal{C}_{\text{post}}(B) \in \mathcal{L}(\mathcal{H}^*, \mathcal{H})$ can be decomposed as

$$\mathcal{C}_{\text{post}}(B) = \mathcal{I}_0^{-1/2} \widehat{\mathcal{C}}_{\text{post}}(B) \mathcal{I}_0^{-1/2} \quad \text{where} \quad \widehat{\mathcal{C}}_{\text{post}}(B) = (\mathcal{I}_0^{-1/2} B \mathcal{I}_0^{-1/2} + \text{Id})^{-1} \in \mathcal{L}(L^2(\Omega), L^2(\Omega)).$$

We will refer to $\widehat{\mathcal{C}}_{\text{post}}(B)$ as the *prior preconditioned covariance operator*. Due to the continuous embedding of $\mathcal{H}$ into $L^2(\Omega)$ the covariance operator $\mathcal{C}_{\text{post}}(B)$ can further be considered as an element of $\mathcal{L}(L^2(\Omega), L^2(\Omega))$. The following lemma characterizes its properties.

Lemma 5.10. *Let $B \in \text{SHS}(\mathcal{H}, \mathcal{H}^*)$ be given. If $c_B < 1$ the operator $\mathcal{C}_{post}(B) \in \mathcal{L}(L^2(\Omega), L^2(\Omega))$ is positive and Hilbert-Schmidt on $L^2(\Omega)$. Furthermore it is of trace class in $L^2(\Omega)$ and there holds*

$$B_2 - B_1 \in \text{Pos}(\mathcal{H}, \mathcal{H}^*) \Rightarrow \text{Tr}_{L^2(\Omega)}(\mathcal{C}_{post}(B_2)) \leq \text{Tr}_{L^2(\Omega)}(\mathcal{C}_{post}(B_1)),$$

for all $B_1,\ B_2 \in \text{Pos}(\mathcal{H}, \mathcal{H}^*)$.

Proof. Positivity of $\mathcal{C}_{\text{post}}(B)$ on $L^2(\Omega)$ follows from the symmetry and coercivity of the form $a[B]$. We prove that $\mathcal{C}_{\text{post}}(B)$ is of trace class. The Hilbert-Schmidt property then follows immediately. Denote by $\{\phi_i\}_{i\in\mathbb{N}}$ the orthonormal basis of $L^2(\Omega)$ given by the eigenfunctions of $\mathcal{I}_0^{-1}$. By definition we have $\|\phi_i\|^2_{L^2(\Omega)} = 1$. Fix an arbitrary index $i \in \mathbb{N}$. Calculating the $\mathcal{H}^*$ norm of ϕ_i reveals

$$\|\phi_i\|^2_{\mathcal{H}^*} = \langle \mathcal{I}_0^{-1} \phi_i, \phi_i \rangle_{\mathcal{H},\mathcal{H}^*} = \lambda_i (\phi_i, \phi_i) = \lambda_i.$$

We obtain

$$(\phi_i, \mathcal{C}_{post}(B)\phi_i)_{L^2(\Omega)} = \langle \mathcal{C}_{\text{post}}(B)\phi_i, \phi_i \rangle_{\mathcal{H},\mathcal{H}^*} \leq \|\phi_i\|^2_{\mathcal{H}^*} \|\mathcal{C}_{post}(B)\|_{\mathcal{L}(\mathcal{H}^*,\mathcal{H})} \leq \frac{\lambda_i}{1 - c_B}.$$

Summing over all indices we get

$$\text{Tr}_{L^2(\Omega)}(\mathcal{C}_{post}(B)) = \sum_{i=1}^{\infty} (\phi_i, \mathcal{C}_{post}(B)\phi_i)_{L^2(\Omega)} \leq \frac{1}{1 - c_B} \sum_{i=1}^{\infty} \lambda_i = \frac{\text{Tr}_{L^2(\Omega)}(\mathcal{I}_0^{-1})}{1 - c_B}.$$

Thus $\mathcal{C}_{post}(B) \in \mathcal{L}(L^2(\Omega), L^2(\Omega))$ is of trace class.

It remains to prove the last claim. Let $B_1,\ B_2 \in \text{Pos}(\mathcal{H}, \mathcal{H}^*)$ with $B_2 - B_1 \in \text{Pos}(\mathcal{H}, \mathcal{H}^*)$ be given and fix an arbitrary index $i \in \mathbb{N}$. Recalling the definition of the preconditioned operator $\widehat{\mathcal{C}}_{\text{post}}(B) \in \mathcal{L}(L^2(\Omega), L^2(\Omega))$ we arrive at

$$(\phi_i, \mathcal{C}_{post}(B_2)\phi_i)_{L^2(\Omega)} = \lambda_i (\phi_i, \widehat{\mathcal{C}}_{post}(B_2)\phi_i)_{L^2(\Omega)} = \lambda_i (\phi_i, (\mathcal{I}_0^{-1/2} B_2 \mathcal{I}_0^{-1/2} + \text{Id})^{-1} \phi_i)_{L^2(\Omega)}.$$

Expanding yields

$$\begin{aligned}
&(\phi_i, \widehat{\mathcal{C}}_{post}(B_2)\phi_i)_{L^2(\Omega)} \\
&\quad = (\widehat{\mathcal{C}}_{\text{post}}(B_1)^{1/2}\phi_i, (\widehat{\mathcal{C}}_{\text{post}}(B_1)^{1/2} \mathcal{I}_0^{-1/2} (B_2 - B_1) \mathcal{I}_0^{-1/2} \widehat{\mathcal{C}}_{\text{post}}(B_1)^{1/2} + \text{Id})^{-1} \widehat{\mathcal{C}}_{\text{post}}(B_1)^{1/2}\phi_i)_{L^2(\Omega)} \\
&\quad \leq \|\widehat{\mathcal{C}}_{\text{post}}(B_1)^{1/2}\phi_i\|^2_{L^2(\Omega)} = (\phi_i, \widehat{\mathcal{C}}_{\text{post}}(B_1)\phi_i)_{L^2(\Omega)}.
\end{aligned}$$

Here the second inequality follows since the operator

$$D = \widehat{\mathcal{C}}_{\text{post}}(B_1)^{1/2} \mathcal{I}_0^{-1/2} (B_2 - B_1) \mathcal{I}_0^{-1/2} \widehat{\mathcal{C}}_{\text{post}}(B_1)^{1/2} + \text{Id} \in \mathcal{L}(L^2(\Omega), L^2(\Omega)),$$

is self-adjoint and coercive with constant one due to $(B_2 - B_1) \in \operatorname{Pos}(\mathcal{H}, \mathcal{H}^*)$. Thus its inverse exists and

$$\|D^{-1}\|_{\mathcal{L}(L^2(\Omega),L^2(\Omega))} \leq 1.$$

Since $i \in \mathbb{N}$ was chosen arbitrary we obtain

$$\begin{aligned}\operatorname{Tr}_{L^2(\Omega)}(\mathcal{C}_{post}(B_2)) &= \sum_{i=1}^{\infty} \lambda_i (\phi_i, \widehat{\mathcal{C}}_{post}(B_2)\phi_i)_{L^2(\Omega)} \\ &\leq \sum_{i=1}^{\infty} \lambda_i (\phi_i, \widehat{\mathcal{C}}_{post}(B_1)\phi_i)_{L^2(\Omega)} = \operatorname{Tr}_{L^2(\Omega)}(\mathcal{C}_{post}(B_1)),\end{aligned}$$

by factoring in $\mathcal{I}_0^{-1/2}$. This proves the claimed statement. □

We close this section by establishing differentiability properties of the covariance mapping.

Proposition 5.11. *Let $B \in \operatorname{SHS}(\mathcal{H}, \mathcal{H}^*)$ with $c_B < 1$ be given. Then the mapping*

$$\mathcal{C}_{post}\colon \operatorname{SHS}(\mathcal{H}, \mathcal{H}^*) \to \mathcal{L}(\mathcal{H}^*, \mathcal{H}),$$

is at least two times continuosly Fréchet differentiable at B. Given $\delta B_1, \delta B_2 \in \operatorname{SHS}(\mathcal{H}, \mathcal{H}^*)$ its first and second derivatives are characterized by

$$\begin{aligned}\nabla \mathcal{C}_{post}(B)\delta B_1 &= -\mathcal{C}_{post}(B)\, \delta B_1\, \mathcal{C}_{post}(B) \in \mathcal{L}(\mathcal{H}^*, \mathcal{H}), \\ \nabla^2 \mathcal{C}_{post}(B)(\delta B_1, \delta B_2) &= 2\mathcal{C}_{post}(B)\, \delta B_1\, \mathcal{C}_{post}(B)\, \delta B_2\, \mathcal{C}_{post}(B) \in \mathcal{L}(\mathcal{H}^*, \mathcal{H}).\end{aligned}$$

Proof. We only provide the proof for the first derivative, the formula for the second derivative can be be established analogously. Let B, $\delta B \in \operatorname{SHS}(\mathcal{H}, \mathcal{H}^*)$ with $c_B < 1$ be given. We have

$$\langle q_1, (B + \delta B) q_1 \rangle_{\mathcal{H},\mathcal{H}^*} \geq (-(c_B + \|\delta B\|_{\operatorname{HS}(\mathcal{H},\mathcal{H}^*)}) \|q_1\|_{\mathcal{H}}^2 \quad \forall q_1 \in \mathcal{H}.$$

Thus $\mathcal{C}_{\text{post}}(B + \delta B)$ is well-defined if $(c_B + \|\delta B\|_{\operatorname{HS}(\mathcal{H},\mathcal{H}^*)}) < 1$. Let $f \in \mathcal{H}^*$ be given and set

$$q_f^B = \mathcal{C}_{\text{post}}(B) f, \quad q_f^{B+\delta B} = \mathcal{C}_{\text{post}}(B + \delta B) f, \quad \nabla q_f^B = -\mathcal{C}_{\text{post}}(B)\delta B \mathcal{C}_{\text{post}}(B) f.$$

Note that $\nabla q_f^B \in \mathcal{H}$ is the unique element fulfilling

$$a[B](\nabla q_f^B, q_2) + a'_B[\delta B](q_f^B, q_2) = a[B](\nabla q_f^B, q_2) + \langle q_f^B, \delta B q_2 \rangle_{\mathcal{H},\mathcal{H}^*} = 0 \quad \forall q_2 \in \mathcal{H}.$$

Set $\delta q_f = q_f^{B+\delta B} - q_f^B - \nabla q_f^B$. We estimate

$$\begin{aligned}(1 - c_B)\|\delta q_f\|_{\mathcal{H}}^2 &\leq a[B](\delta q_f, \delta q_f) \\ &= a[B](q_f^{B+\delta B}, \delta q_f) - a[B + \delta B](q_f^{B+\delta B}, \delta q_f) + \langle q_f^B, \delta B \delta q_f \rangle_{\mathcal{H},\mathcal{H}^*} \\ &= \langle q_f^B - q_f^{B+\delta B}, \delta B \delta q_f \rangle_{\mathcal{H},\mathcal{H}^*} \\ &\leq \frac{\|f\|_{\mathcal{H}^*} \|\delta B\|_{\operatorname{HS}(\mathcal{H},\mathcal{H}^*)}^2 \|\delta q_f\|_{\mathcal{H}}}{(1 - c_B)(1 - c_B - \|\delta B\|_{\operatorname{HS}(\mathcal{H},\mathcal{H}^*)})},\end{aligned}$$

where we used the Lipschitz stability of $\mathcal{C}_{post}$ and

$$\|\delta B\|_{\mathcal{L}(\mathcal{H},\mathcal{H}^*)} \leq \|\delta B\|_{\operatorname{HS}(\mathcal{H},\mathcal{H}^*)}.$$

Dividing by $\|\delta B\|_{\operatorname{HS}(\mathcal{H},\mathcal{H}^*)}, \|\delta q_f\|_{\mathcal{H}} \neq 0$ and taking the supremum over $f \in \mathcal{H}^*$ on both sides we deduce the Fréchet differentiability of $\mathcal{C}_{\text{post}}$ by performing the limit $\|\delta B\|_{\operatorname{HS}(\mathcal{H},\mathcal{H}^*)} \to 0$. □

Sparse A and D-optimal design

After these preparatory steps we are ready to formulate suitable optimal design criteria for Bayesian inverse problems. We define the set $\mathbf{B}^+(L^2(\Omega))$ as

$$\left\{ B \in \mathrm{SHS}(L^2(\Omega), L^2(\Omega)) \mid \exists c_B \in [0,1) \colon (\mathcal{I}_0^{-1/2} q_1, B\mathcal{I}_0^{-1/2} q_1)_{L^2(\Omega)} \geq -c_B \|q_1\|^2_{L^2(\Omega)},\ q_1 \in L^2(\Omega) \right\}.$$

Note that this set is open in $\mathrm{SHS}(L^2(\Omega), L^2(\Omega))$ and $\mathrm{Pos}(L^2(\Omega), L^2(\Omega)) \subset \mathbf{B}^+(L^2(\Omega))$. Moreover if we interpret $B \in \mathbf{B}^+(L^2(\Omega))$ as Hilbert-Schmidt operator from $\mathcal{H}$ into $\mathcal{H}^*$ we have

$$\langle q_1, Bq_1 \rangle_{\mathcal{H},\mathcal{H}^*} = (q_1, Bq_1)_{L^2(\Omega)} \geq -c_B \|q_1\|^2_{\mathcal{H}}$$

for all $q_1 \in \mathcal{H}$. Here we used $\mathcal{H} = \mathrm{dom}_{L^2(\Omega)} \mathcal{I}_0^{1/2}$ and the definition of the norm on $\mathcal{H}$. Thus the covariance operator $\mathcal{C}_{post}(B) \in \mathcal{L}(\mathcal{H}^*, \mathcal{H})$ in the sense of Proposition 5.1 is well-defined.

First we discuss an infinite dimensional analogue of the A-optimal design criterion which is given by the trace of the posterior covariance operator

$$\Psi_A \colon \mathrm{SHS}(L^2(\Omega), L^2(\Omega)) \to \mathbb{R} \cup \{+\infty\} \quad B \mapsto \begin{cases} \mathrm{Tr}_{L^2(\Omega)}(\mathcal{C}_{post}(B)) & B \in \mathbf{B}^+(L^2(\Omega)) \\ +\infty & \text{else.} \end{cases} \qquad (5.24)$$

We give some interpretation to this choice of the optimal design criterion. To this end we recall the definition of the posterior measure $\mu^{\mathbf{y}_d}_{\mathrm{post}} = \mathcal{N}(q^{\mathbf{y}_d}_{\mathrm{post}}, \mathcal{C}_{\mathrm{post}}(\mathcal{I}(\boldsymbol{u}(\mathbf{x}, \mathbf{u}))))$ given a vector of measurement data $\mathbf{y}_d \in \mathbb{R}^N$, see (5.19). By $q^{\mathbf{y}_d} \colon D \to L^2(\Omega)$ we denote the random field distributed according to it. A first indicator for the quality of the obtained posterior measure is its variability around the mean. An optimal measurement setup should lead to posterior measures whose draws are close to $q^{\mathbf{y}_d}_{\mathrm{post}}$, at least on average. To make these considerations rigorous we calculate the variance of the posterior distribution as

$$\begin{aligned} \mathrm{Var}(q^{\mathbf{y}_d}) = \mathbb{E}^{\mu^{\mathbf{y}_d}_{\mathrm{post}}}[\|q^{\mathbf{y}_d} - q^{\mathbf{y}_d}_{\mathrm{post}}\|^2_{L^2(\Omega)}] &= \int_{L^2(\Omega)} \|q - q^{\mathbf{y}_d}_{\mathrm{post}}\|^2_{L^2(\Omega)} \,\mathrm{d}\mu^{\mathbf{y}_d}_{\mathrm{post}}(q) = \int_\Omega \mathrm{Var}_{q^{\mathbf{y}_d}} \,\mathrm{d}x \\ &= \mathrm{Tr}_{L^2(\Omega)}(\mathcal{C}_{\mathrm{post}}(\mathcal{I}(\boldsymbol{u}(\mathbf{x}, \mathbf{u})))), \end{aligned}$$

where $\mathrm{Var}_{q^{\mathbf{y}_d}}$ denotes the pointwise variance of $q^{\mathbf{y}_d}$, see (5.4) and (5.5). In particular, this implies that the left hand side of this equation is independent of the data vector $\mathbf{y}_d \in \mathbb{R}^N$ and corresponds to the averaged posterior variance. Furthermore it only depends on the measurement setup through the posterior covariance operator. Hence we can a priori, i.e. before the measurements are carried out, improve it by minimizing the A-optimal design criterion for the Fisher information with respect to the measurement setup.

A second motivation to consider the trace of the posterior covariance operator is given by the *mean squared error* (MSE) of the posterior mean $q^{\mathbf{y}_d}_{\mathrm{post}}$. We recall the assumptions on the data model

$$y_d \colon L^2(\Omega) \times \mathbb{R}^N \to \mathbb{R}^N, \quad (q, \epsilon) \mapsto S[\hat{q}](x) + X(q - \hat{q}) + \epsilon.$$

The obtained measurement vector is given by $\mathbf{y}_d = y_d(q^*, \epsilon)$ where q^* denotes the true value of the unknown parameter and ϵ is drawn from a Gaussian distribution $\mu_E = \mathcal{N}(0, \Sigma)$. Given a function $q \in L^2(\Omega)$ we define the estimator

$$q^{y_d(q,\cdot)}_{\mathrm{post}} \colon \mathbb{R}^N \to L^2(\Omega), \quad \epsilon \mapsto q^{y_d(q,\epsilon)}_{\mathrm{post}}.$$

A properly chosen measurement setup for the estimation of q^* should result in an estimator $q_{\text{post}}^{y_d(q^*,\cdot)}$ whose realizations are close to q^*, e.g., with respect to the norm on $L^2(\Omega)$. Changes in the measurement data due to noise should only lead to small changes in the estimated parameter.

Again we give some mathematical rigor to this intuition. Given $q \in L^2(\Omega)$ and the associated estimator $q_{\text{post}}^{y_d(q,\cdot)}$ we consider its mean squared error

$$\mathrm{MSE}(q_{\text{post}}^{y_d(q,\cdot)}, q) = \mathbb{E}^{\mu_E}[\|q_{\text{post}}^{y_d(q,\cdot)} - q\|^2_{L^2(\Omega)}] = \int_{\mathbb{R}^N} \|q_{\text{post}}^{y_d(q,\varepsilon)} - q\|^2_{L^2(\Omega)} \, \mathrm{d}\mu_E(\varepsilon).$$

Evaluating the integral yields

$$\begin{aligned}\mathrm{MSE}(q_{\text{post}}^{y_d(q,\cdot)}, q) = \mathbb{E}^{\mu_E}[\|q_{\text{post}}^{y_d(q,\cdot)} - q\|^2_{L^2(\Omega)}] &= \int_{\mathbb{R}^N} \|q_{\text{post}}^{y_d(q,\varepsilon)} - q\|^2_{L^2(\Omega)} \, \mathrm{d}\mu_E(\varepsilon) \\ &= \|(\mathcal{C}_{\text{post}}(\mathcal{I}(\boldsymbol{u}(\mathbf{x},\mathbf{u})))\mathcal{I}(\boldsymbol{u}(\mathbf{x},\mathbf{u})) - \mathrm{Id})(q - \hat{q})\|^2_{L^2(\Omega)} \\ &\quad + \mathrm{Tr}_{L^2(\Omega)}(\mathcal{C}_{\text{post}}(\mathcal{I}(\boldsymbol{u}(\mathbf{x},\mathbf{u})))^2\mathcal{I}(\boldsymbol{u}(\mathbf{x},\mathbf{u}))).\end{aligned} \tag{5.25}$$

A derivation of this equality can be found in [3]. The measurement setup should be chosen to minimize the mean squared error for the true parameter q^*. However, from the calculations in (5.25) we infer that $\mathrm{MSE}(q_{\text{post}}^{y_d(q^*,\cdot)}, q^*)$ depends on the unknown parameter itself and therefore cannot be evaluated. As a remedy we demand that an optimal estimator should provide good estimates for draws taken from μ_0 in an average sense. Averaging over the prior distribution gives the *expected mean squared error*

$$\int_{L^2(\Omega)} \int_{\mathbb{R}^N} \|q_{\text{post}}^{y_d(q,\varepsilon)} - q\|^2_{L^2(\Omega)} \, \mathrm{d}\mu_E(\varepsilon)\mathrm{d}\mu_0(q) = \mathrm{Tr}_{L^2(\Omega)}(\mathcal{C}_{\text{post}}(\mathcal{I}(\boldsymbol{u}(\mathbf{x},\mathbf{u})))), \tag{5.26}$$

which again corresponds to the trace of the posterior covariance operator. For a derivation of this last step see again [3].

Let us take a closer look on the left hand side of the last equation. Given a vector of measurements $\mathbf{y}_d = y_d(q,\epsilon)$ for some $q \in L^2(\Omega)$ and $\epsilon \in \mathbb{R}^N$ we recall that the associated MAP estimator $q_{\text{post}}^{\mathbf{y}_d} \in \mathcal{H}$ is found as the unique solution to the linear-quadratic problem

$$\min_{q \in \mathcal{H}} \frac{1}{2}|Xq - \mathbf{y}_d|^2_{\Sigma^{-1}} + \|q\|^2_{\mathcal{H}}$$

This allows for an interpretation of the minimization of the A-optimal design criterion with respect to the measurement setup as a *learning* or *bilevel* problem for an optimal sensor distribution. We briefly shed some light on this connection. Given a fixed measurement setup we may consider the associated estimator given by

$$q_{\text{post}}^{\cdot} \colon \mathbb{R}^N \to \mathcal{H}, \quad q_{\text{post}}^{\mathbf{y}_d} = \arg\min_{q \in \mathcal{H}} \frac{1}{2}|Xq - \mathbf{y}_d|^2_{\Sigma^{-1}} + \|q\|^2_{\mathcal{H}}.$$

Thus for every vector $\mathbf{y}_d$ of measurements we find $q_{\text{post}}^{\mathbf{y}_d}$ by minimizing a linear quadratic functional in the *lower level* problem. In order to assess the quality of this estimator we first generate a test set of parameters (described by the prior distribution of the random field). For each test parameter we then obtain a set of artifical measurement data based on our assumptions on the measurement noise. Subsequently the discrepancy between the expected result and the parameter proposed by the estimator is calculated. A solution to the sensor placement problem is then found in the *upper* *level* problem as one particular sensor configuration whose associated estimator yields, on average, the best reconstruction results.

In the following proposition we elaborate on the mathematical properties of the A-optimal design criterion.

Proposition 5.12. *The mapping* $\Psi_A\colon \operatorname{SHS}(L^2(\Omega),L^2(\Omega)) \to \mathbb{R}\cup\{+\infty\}$ *has the following properties.*

- *On* $\operatorname{Pos}(L^2(\Omega),L^2(\Omega))$, Ψ_A *is non-negative and strictly convex.*
- *On* $\operatorname{Pos}(L^2(\Omega),L^2(\Omega))$, Ψ_A *is at least two times continuously Fréchet differentiable. Given* $\delta B_1,\delta B_2 \in \operatorname{SHS}(L^2(\Omega),L^2(\Omega))$ *the Fréchet derivatives can be identified as*

$$\begin{aligned}\langle\langle \nabla\Psi_A(B),\delta B_1\rangle\rangle_{HS(L^2(\Omega),L^2(\Omega))} &= -\operatorname{Tr}_{L^2(\Omega)}(\mathcal{C}_{post}(B)\delta B_1\mathcal{C}_{post}(B)),\\ \langle\langle \delta B_1,\nabla^2\Psi_A(B)\delta B_2\rangle\rangle_{HS(L^2(\Omega),L^2(\Omega))} &= 2\operatorname{Tr}_{L^2(\Omega)}(\mathcal{C}_{post}(B)\delta B_1\mathcal{C}_{post}(B)\delta B_2\mathcal{C}_{post}(B)).\end{aligned}$$

- *The A-optimal design criterion is monotone. Given* $B_1,B_2\in\operatorname{Pos}(L^2(\Omega),L^2(\Omega))$ *we have*

$$B_2-B_1\in\operatorname{Pos}(L^2(\Omega),L^2(\Omega)) \Rightarrow \Psi_A(B_2)\le\Psi_A(B_1).$$

Proof. Following Lemma 5.10 the A-optimal design criterion is non-negative on $\operatorname{Pos}(L^2(\Omega),L^2(\Omega))\subset \mathbf{B}^+(L^2(\Omega))$. We further note that the $L^2(\Omega)$ trace defines a linear continuous functional on $\mathcal{L}(\mathcal{H}^*,\mathcal{H})$ since

$$\operatorname{Tr}_{L^2(\Omega)}(\tilde{B}) \le \sum_{i=1}^{\infty}\|\phi_i\|^2_{\mathcal{H}^*}\|\tilde{B}\|_{\mathcal{L}(\mathcal{H}^*,\mathcal{H})} = \operatorname{Tr}_{L^2(\Omega)}(\mathcal{I}_0^{-1})\|\tilde{B}\|_{\mathcal{L}(\mathcal{H}^*,\mathcal{H})}\quad \forall\tilde{B}\in\mathcal{L}(\mathcal{H}^*,\mathcal{H}).$$

Continuity and Fréchet-differentiability of Ψ_A at $B\in\operatorname{Pos}(L^2(\Omega),L^2(\Omega))$ now follows from the linearity of the trace and the results of Proposition 5.11. In particular we obtain

$$\begin{aligned}\langle\langle \delta B,\nabla^2\Psi_A(B)\delta B\rangle\rangle_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))} &= 2\operatorname{Tr}_{L^2(\Omega)}(\mathcal{C}_{\mathrm{post}}(B)\delta B\mathcal{C}_{\mathrm{post}}(B)\delta B\mathcal{C}_{\mathrm{post}}(B))\\ &= 2\|\mathcal{C}_{\mathrm{post}}(B)^{1/2}\delta B\mathcal{C}_{\mathrm{post}}(B)\|^2_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))}.\end{aligned}$$

Fix an arbitrary index $i\in\mathbb{N}$. We proceed to estimate

$$\begin{aligned}\|\mathcal{C}_{\mathrm{post}}(B)^{1/2}\delta B\mathcal{C}_{\mathrm{post}}(B)\phi_i\|_{L^2(\Omega)} &= \|\widehat{\mathcal{C}}_{\mathrm{post}}(B)^{1/2}\mathcal{I}_0^{-1/2}\delta B\mathcal{I}_0^{-1/2}\widehat{\mathcal{C}}_{\mathrm{post}}(B)\mathcal{I}_0^{-1/2}\phi_i\|^2_{L^2(\Omega)}\\ &\ge \frac{1}{\|\widehat{\mathcal{C}}_{\mathrm{post}}(B)^{-1}\|_{\mathcal{L}(L^2(\Omega),L^2(\Omega))}}\|\widehat{\mathcal{C}}_{\mathrm{post}}(B)\mathcal{I}_0^{-1}\delta B\mathcal{I}_0^{-1/2}\phi_i\|^2_{L^2(\Omega)}\\ &\ge \frac{1}{\|\widehat{\mathcal{C}}_{\mathrm{post}}(B)^{-1}\|^3_{\mathcal{L}(L^2(\Omega),L^2(\Omega))}}\|\mathcal{I}_0^{-1}\delta B\mathcal{I}_0^{-1/2}\phi_i\|^2_{L^2(\Omega)},\end{aligned}$$

where we used that

$$\|q_1\|_{L^2(\Omega)} = \|\widehat{\mathcal{C}}_{\mathrm{post}}(B)^{-1}\widehat{\mathcal{C}}_{\mathrm{post}}(B)q_1\|_{L^2(\Omega)} \le \|\widehat{\mathcal{C}}_{\mathrm{post}}(B)^{-1}\|_{\mathcal{L}(L^2(\Omega),L^2(\Omega))}\|\widehat{\mathcal{C}}_{\mathrm{post}}(B)q_1\|_{L^2(\Omega)},$$

for all $q_1\in L^2(\Omega)$. The norm in the denominator is further bounded by

$$\|\widehat{\mathcal{C}}_{\mathrm{post}}(B)^{-1}\|_{\mathcal{L}(L^2(\Omega),L^2(\Omega))} = \|\operatorname{Id}+\mathcal{I}_0^{-1/2}B\mathcal{I}_0^{-1/2}\|_{\mathcal{L}(L^2(\Omega),L^2(\Omega))} \le 1+\|B\|_{\mathrm{HS}(\mathcal{H},\mathcal{H}^*)}$$

Finally summing over all indices $i \in \mathbb{N}$ we conclude

$$2\,\mathrm{Tr}_{L^2(\Omega)}(\mathcal{C}_{\mathrm{post}}(B)\delta B\mathcal{C}_{\mathrm{post}}(B)\delta B\mathcal{C}_{\mathrm{post}}(B)) \geq \frac{2}{(1+\|B\|_{\mathrm{HS}(\mathcal{H},\mathcal{H}^*)})^3}\|\mathcal{I}_0^{-1}\delta B\mathcal{I}_0^{-1/2}\|^2_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))}. \tag{5.27}$$

Thus Ψ_A is strictly convex on $\mathrm{Pos}(L^2(\Omega), L^2(\Omega))$. Monotonicity of the A-optimal design criterion on $\mathrm{Pos}(L^2(\Omega), L^2(\Omega))$ follows from Lemma 5.10. □

As a second example we comment on the infinite dimensional D-optimal design criterion. Given a trace class operator $T \in \mathcal{L}(L^2(\Omega), L^2(\Omega))$ we define its Fredholm determinant by

$$\mathrm{Det}(T + \mathrm{Id}) = \prod_{i=1}^{\infty}(1+\mu_i),$$

where $\{\mu_i\}_{i=1}^\infty$ denote the eigenvalues of $|T| = (T^*T)^{1/2}$, see [119, Chapter IV]. For a given Hilbert-Schmidt operator $B \in \mathrm{SHS}(L^2(\Omega), L^2(\Omega))$ the operator $\mathcal{I}_0^{-1/2}B\mathcal{I}_0^{-1/2}$ is of trace class on $L^2(\Omega)$. The D-optimal design criterion is now defined as the negative logarithm of the Fredholm determinant of $\mathcal{I}_0^{-1/2}B\mathcal{I}_0^{-1/2}$. Recalling the definition of the prior-preconditioned covariance operator

$$\widehat{\mathcal{C}}_{\mathrm{post}}(B) = (\mathcal{I}_0^{-1/2}B\mathcal{I}_0^{-1/2} + \mathrm{Id})^{-1} \in \mathcal{L}(L^2(\Omega), L^2(\Omega)),$$

this is, in a more compact way, stated as

$$\Psi_D\colon\ \mathrm{SHS}(L^2(\Omega), L^2(\Omega)) \to \mathbb{R}\cup\{+\infty\} \quad B \mapsto \begin{cases} -\log(\mathrm{Det}(\widehat{\mathcal{C}}_{\mathrm{post}}(B)^{-1})) & B \in \mathbf{B}^+(L^2(\Omega)) \\ +\infty & \text{else.}\end{cases}. \tag{5.28}$$

As for the A-optimal design criterion we clarify the interpretation of this definition. To this end let a vector of measurement data $\mathbf{y}_d \in \mathbb{R}^N$ and the associated posterior measure $\mu_{\mathrm{post}}^{\mathbf{y}_d}$ be given. In Section 3.1.2 we already mentioned that there are several possibilities to compare probability measures, cf. [117]. In the following we quantify the distance between the prior measure μ_0 and the posterior through their *Kullback-Leibler divergence* or *relative entropy* defined as

$$d_I(\mu_{\mathrm{post}}^{\mathbf{y}_d}, \mu_0) = \int_{L^2(\Omega)} \log\left(\frac{d\mu_{\mathrm{post}}^{\mathbf{y}_d}}{d\mu_0}\right)\,\mathrm{d}\mu_{\mathrm{post}}^{\mathbf{y}_d}(q) = \int_{L^2(\Omega)} \log\left(\frac{d\mu_{\mathrm{post}}^{\mathbf{y}_d}}{d\mu_0}\right)\frac{d\mu_{\mathrm{post}}^{\mathbf{y}_d}}{d\mu_0}\,\mathrm{d}\mu_0(q), \tag{5.29}$$

see, e.g, [174]. Here, the Radon-Nikodým derivative of the posterior with respect to the prior is given by

$$\frac{d\mu_{\mathrm{post}}^{\mathbf{y}_d}}{d\mu_0}\colon L^2(\Omega) \to [0,1], \quad q \mapsto \frac{1}{\mathcal{Z}_0(\mathbf{y}_d)}\exp\left(-\frac{1}{2}|S[\hat{q}](x) + X(q-\hat{q}) - \mathbf{y}_d|^2_{\Sigma^{-1}}\right),$$

with the constant $\mathcal{Z}_0(\mathbf{y}_d) > 0$ defined as

$$\mathcal{Z}_0(\mathbf{y}_d) = \int_{L^2(\Omega)} \exp\left(-\frac{1}{2}|S[\hat{q}](x) + X(q-\hat{q}) - \mathbf{y}_d|^2_{\Sigma^{-1}}\right)\,\mathrm{d}\mu_0(q).$$

While the Kullback-Leibler divergence fulfills some intuitive notions of a distance such as non-negativity and

$$\mu_1 = \mu_2 \Leftrightarrow d_I(\mu_1, \mu_2) = 0,$$

it does not define a metric on the probability measures as it lacks symmetry and does not fulfill the triangle inequality.

Intuitively if the measured data $\mathbf{y}_d \in \mathbb{R}^N$ provides a lot of information on the unknown parameter the prior and the posterior measures differ significantly, i.e. their relative entropy $d_I(\mu_{\text{post}}^{\mathbf{y}_d}, \mu_0)$ should be large. A measurement setup could now be chosen in order to maximize this distance. However, as already pointed out for the mean squared error, this quantity depends on the concrete realization of the data vector $\mathbf{y}_d$. In order to obtain a criterion that is computable before the measurements are carried out, we average over the prior distribution of the parameter and the measurement noise. Following [3] we therefore calculate the *expected information gain*

$$\int_{L^2(\varOmega)} \int_{\mathbb{R}^N} \frac{1}{\mathcal{Z}_1(q)} d_I(\mu_{\text{post}}^{\mathbf{y}_d}, \mu_0) \exp\left(-\frac{1}{2}|S[\hat{q}](x) + X(q-\hat{q}) - \mathbf{y}_d|^2_{\varSigma^{-1}}\right) \, \mathrm{d}\mathbf{y}_d \mathrm{d}\mu_0(q),$$

where for fixed $q \in L^2(\varOmega)$ the normalization constant $\mathcal{Z}_1(q) > 0$ is given by

$$\mathcal{Z}_1(q) = \int_{\mathbb{R}^N} \exp\left(-\frac{1}{2}|S[\hat{q}](x) + X(q-\hat{q}) - \mathbf{y}_d|^2_{\varSigma^{-1}}\right) \, \mathrm{d}\mathbf{y}_d.$$

Calculating the integral leads to

$$\begin{aligned}\int_{L^2(\varOmega)} \int_{\mathbb{R}^N} \frac{1}{\mathcal{Z}_1(q)} d_I(\mu_{\text{post}}^{\mathbf{y}_d}, \mu_0) \exp\left(-\frac{1}{2}|S[\hat{q}](x) + X(q-\hat{q}) - \mathbf{y}_d|^2_{\varSigma^{-1}}\right) \, \mathrm{d}\mathbf{y}_d \mathrm{d}\mu_0(q) \\ = \log(\operatorname{Det}(\widehat{\mathcal{C}}_{\text{post}}(\mathcal{I}(\boldsymbol{u}(\mathbf{x}, \mathbf{u})))^{-1})).\end{aligned}$$

Thus we might consider a measurement setup optimal if it maximizes the averaged Kullback-Leibler divergence or equivalently the logarithm of the Fredholm determinant of the Fisher information preconditioned by $\mathcal{I}_0^{-1/2}$. Since through the course of this thesis minimization problems are studied we take its negative to arrive at the D-optimal design criterion. The following proposition summarizes some of its properties.

Proposition 5.13. *The mapping $\varPsi_D \colon \operatorname{SHS}(L^2(\varOmega), L^2(\varOmega)) \to \mathbb{R} \cup \{+\infty\}$ has the following properties.*

- *On $\operatorname{Pos}(L^2(\varOmega), L^2(\varOmega))$, $\varPsi_D$ is strictly convex and weakly lower semicontinuous.*
- *On $\operatorname{Pos}(L^2(\varOmega), L^2(\varOmega))$, $\varPsi_D$ is at least two times continuously Fréchet differentiable. Given $\delta B_1, \delta B_2 \in \operatorname{SHS}(L^2(\varOmega), L^2(\varOmega))$ the derivatives are characterized by*

$$\begin{aligned}\langle\langle \nabla \varPsi_D(B), \delta B_1 \rangle\rangle_{HS(L^2(\varOmega), L^2(\varOmega))} &= -\operatorname{Tr}_{L^2(\varOmega)}(\mathcal{C}_{post}(B)\delta B_1), \\ \langle\langle \delta B_1, \nabla^2 \varPsi_D(B)\delta B_2 \rangle\rangle_{HS(L^2(\varOmega), L^2(\varOmega))} &= \operatorname{Tr}_{L^2(\varOmega)}(\mathcal{C}_{post}(B)\delta B_1 \mathcal{C}_{post}(B)\delta B_2).\end{aligned}$$

- *The D-optimal design criterion is monotone. Given $B_1, B_2 \in \operatorname{Pos}(L^2(\varOmega), L^2(\varOmega))$ we have*

$$B_2 - B_1 \in \operatorname{Pos}(L^2(\varOmega), L^2(\varOmega)) \Rightarrow \varPsi_D(B_2) \leq \varPsi_D(B_1).$$

Proof. Following [242], the logarithm of the Fredholm determinant as a function on trace class operators is Gâteaux differentiable. Following this result, the Gâteaux derivative $\nabla_{\delta B}\Psi_D(B)$ of Ψ_D at $B \in \mathrm{Pos}(L^2(\Omega), L^2(\Omega))$ in the direction of $\delta B \in \mathrm{SHS}(L^2(\Omega), L^2(\Omega))$ is given by

$$\nabla_{\delta B}\Psi_D(B) = \mathrm{Tr}_{L^2(\Omega)}(\widehat{\mathcal{C}}_{\mathrm{post}}(B)\mathcal{I}_0^{-1/2}\delta B\mathcal{I}_0^{-1/2}) = \mathrm{Tr}_{L^2(\Omega)}(\mathcal{C}_{\mathrm{post}}(B)\delta B),$$

applying the chain rule. Here, the second equality follows since the trace allows for cyclic permutations. Hence the Gateaux differential is linear and continuous. Invoking the results of Lemma 5.11 we conclude its continuous dependence on B. Thus Ψ_D is Fréchet differentiable on $\mathrm{Pos}(L^2(\Omega), L^2(\Omega))$. The existence of the second Fréchet derivative now follows immediately from the results for Ψ_A. Given a direction $\delta B \in \mathrm{SHS}(L^2(\Omega), L^2(\Omega))$ we further conclude

$$\mathrm{Tr}_{L^2(\Omega)}(\mathcal{C}_{\mathrm{post}}(B)\delta B\mathcal{C}_{\mathrm{post}}(B)\delta B) \geq \frac{1}{(1+\|B\|_{\mathrm{HS}(\mathcal{H},\mathcal{H}^*)})^2}\|\mathcal{I}_0^{-1/2}\delta B\mathcal{I}_0^{-1/2}\|^2_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))},$$

following the same steps as in the proof of Proposition 5.12. As a consequence, Ψ_D is strictly convex on $\mathrm{Pos}(L^2(\Omega), L^2(\Omega))$.

It remains to prove the monotonicity of Ψ_D. Therefore let $B_1, B_2 \in \mathrm{Pos}(L^2(\Omega), L^2(\Omega))$ with $\delta B = B_2 - B_1 \in \mathrm{Pos}(L^2(\Omega), L^2(\Omega))$ be given. By Taylor expansion we obtain

$$\Psi_D(B_1) - \Psi_D(B_2) = \mathrm{Tr}_{L^2(\Omega)}(\mathcal{C}_{post}(B_\zeta)\delta B) \geq 0,$$

for some $B_\zeta = B_1 + \zeta(B_2 - B_1) \in \mathrm{Pos}(L^2(\Omega), L^2(\Omega))$, $\zeta \in (0,1)$. Here we used

$$\mathrm{Tr}_{L^2(\Omega)}(\mathcal{C}_{post}(B_\zeta)\delta B) = \mathrm{Tr}_{L^2(\Omega)}(\mathcal{C}_{post}(B_\zeta)^{1/2}\delta B\mathcal{C}_{post}(B_\zeta)^{1/2}) \geq 0$$

This finishes the proof. □

Sparse goal oriented design

In many applications, the interest of an experimenter may not lie on the infinite-dimensional parameter q itself but on a finite dimensional quantity of interest ρ depending on the random field. Such goal oriented inverse problems are considered, e.g., in [123, 246]. In this situation, optimal design criteria should reflect this fact and aim for uncertainty reduction in the quantity of interest rather than the distributed parameter itself. We consider a linear dependence $\rho = Mq$ for $M \in \mathcal{L}(L^2(\Omega), \mathbb{R}^m)$, $m \in \mathbb{N}$. Consequently ρ is a normally distributed random variable with

$$\rho \sim \mathcal{N}(\rho_0, \mathcal{I}_{0,M}^{-1}), \quad \rho_0 = Mq_0, \quad \mathcal{I}_{0,M}^{-1} = M^*\mathcal{I}_0^{-1}M.$$

Moreover the posterior measure of ρ given a vector of measured data $\mathbf{y}_d \in \mathbb{R}^N$ is also a Gaussian $\mathcal{N}(\rho_{\mathrm{post}}^{\mathbf{y}_d}, \mathcal{C}_{\mathrm{post}}^M(\mathcal{I}(\boldsymbol{u}(\mathbf{x},\mathbf{u}))))$ where

$$\rho_{\mathrm{post}}^{\mathbf{y}_d} = Mq_{\mathrm{post}}^{\mathbf{y}_d}, \quad \mathcal{C}_{\mathrm{post}}^M(\mathcal{I}(\boldsymbol{u}(\mathbf{x},\mathbf{u}))) = M^*\mathcal{C}_{\mathrm{post}}(\mathcal{I}(\boldsymbol{u}(\mathbf{x},\mathbf{u})))M = M^*(\mathcal{I}(\boldsymbol{u}(\mathbf{x},\mathbf{u})) + \mathcal{I}_0)^{-1}M.$$

As in the previous section, optimal designs may now be determined as to minimize the expected mean squared error or as to maximize the expected information gain for the quantity of interest, see [10,138]. This leads to the formulation of the goal-oriented A and D-optimal design criteria

$$\Psi_A^G\colon \operatorname{SHS}(L^2(\Omega),L^2(\Omega)) \to \mathbb{R}\cup\{+\infty\} \quad B \mapsto \begin{cases} \operatorname{Tr}_{\mathbb{R}^m}(M^*\mathcal{C}_{post}(B)M) & B \in \mathbf{B}^+(L^2(\Omega)) \\ +\infty & \text{else} \end{cases}, \tag{5.30}$$

$$\Psi_D^G\colon \operatorname{SHS}(L^2(\Omega),L^2(\Omega)) \to \mathbb{R}\cup\{+\infty\} \quad B \mapsto \begin{cases} \log(\operatorname{Det}_{\mathbb{R}^m}(M^*\mathcal{C}_{\text{post}}(B)^{-1}M)) & B \in \mathbf{B}^+(L^2(\Omega)) \\ +\infty & \text{else} \end{cases}, \tag{5.31}$$

respectively. Here, to avoid ambiguities, $\operatorname{Tr}_{\mathbb{R}^m}$ and $\operatorname{Det}_{\mathbb{R}^m}$ denote the trace and determinant of a matrix in $\mathbb{R}^{m\times m}$. The following properties of these functionals can be inferred from well-known results for the finite-dimensional trace and determinant as well as the differentiability of $\mathcal{C}_{\text{post}}$. We omit the proof here for brevity.

Proposition 5.14. *The goal oriented A and D-optimal design criteria as defined by* (5.30) *and* (5.31)*, respectively, are convex, weak*-to-strong continuous, and at least two times continuously Fréchet differentiability. Moreover they are monotone in the sense that*

$$B_2 - B_1 \in \operatorname{Pos}(L^2(\Omega),L^2(\Omega)) \Rightarrow \Psi_A^G(B_2) \le \Psi_A^G(B_1), \quad \Psi_D^G(B_2) \le \Psi_D^G(B_1),$$

for all $B_1,\ B_2 \in \operatorname{Pos}(L^2(\Omega),L^2(\Omega))$.

5.1.4 Sparse sensor placement

We are now prepared to formulate optimal sensor placement problems for the Bayesian inverse problem discussed in Section 5.1.2. Motivated by the discussions of the previous section we propose to determine an optimal number of measurements $N\in\mathbb{N}$, their positions $\mathbf{x}\in\Omega_o^N$ and a vector of measurement weights $\mathbf{u}\in\mathbb{R}_+^N$ by solving an optimization problem

$$\min_{\mathbf{x}\in\Omega_o^N,\ \mathbf{u}\in\mathbb{R}_+^N,\ N\in\mathbb{N}} [\Psi(\mathcal{I}(\boldsymbol{u}(\mathbf{x},\mathbf{u}))) + \beta\|\mathbf{u}\|_1], \tag{5.32}$$

based on a parametrization of the Fisher information operator by the measurement setup. As in the previous chapters, the parameter $\beta>0$ models the cost of a single measurement and the optimal design criterion Ψ is a convex, scalar-valued function acting on the Fisher-Information operator $\mathcal{I}(\boldsymbol{u}(\mathbf{x},\mathbf{u})) = X^*\Sigma^{-1}X$. The operator $X\in\mathcal{L}(L^2(\Omega),\mathbb{R}^N)$ and $\Sigma^{-1}\in\mathbb{R}^{N\times N}$ are given in terms of the measurement setup as

$$(Xq)_i = \partial S[\hat{q}]\delta q\,(x_i) \quad \forall \delta q \in L^2(\Omega), \quad \Sigma_{ij}^{-1} = \delta_{ij}\mathbf{u}_i, \quad i,j=1,\ldots,N. \tag{5.33}$$

Again, we will avoid the combinatorial and non-convex aspect of the minimization problem in (5.32) by replacing the admissible set of measurement setups with the set of positive Radon measures $\mathcal{M}^+(\Omega_o)$. To this end we first take a closer look at the Fisher-information operator. Given $x\in\Omega_o$ and $\delta q\in L^2(\Omega)$ there holds

$$\partial S[\hat{q}]\delta q\,(x) = \langle \partial S[\hat{q}]\delta q, \delta_x\rangle = (\delta q, \partial S[\hat{q}]^*\delta_x)_{L^2(\Omega)},$$

where $\partial S[\hat{q}]^*$ denotes the adjoint of the solution operator to the sensitivity equation (5.13). For abbreviation, we set $G^x = \partial S[\hat{q}]^*\delta_x \in L^2(\Omega)$ and refer to it as the *Green's function* of $\partial S[\hat{q}]^*$ at the point $x\in\Omega_o$. In the following lemma its continuity with respect to x is studied.

Lemma 5.15. *The mapping*

$$G^{\cdot}: \Omega_o \to L^2(\Omega), \quad x \mapsto G^x,$$

is uniformly continuous.

Proof. By assumption the operator $\partial S[\hat{q}]\colon L^2(\Omega) \to \mathcal{C}(\Omega_o)$ is linear continuous and compact. Due to Schauder's Theorem the same holds for its adjoint. In particular this implies weak*-to-strong continuity of $\partial S[\hat{q}]^*$. Let $\{x_k\}_{k\in\mathbb{N}} \subset \Omega_o$ with $\lim_{k\to\infty} x_k = x \in \Omega_o$ be given. Then the corresponding Dirac delta functions converge in the weak* sense and thus

$$\lim_{k\to\infty} |x_k - x|_{\mathbb{R}^d} = 0 \Rightarrow \lim_{k\to\infty} \|G^x - G^{x_k}\|_{L^2(\Omega)} = \lim_{k\to\infty} \|\partial S[\hat{q}]^*(\delta_{x_k} - \delta_x)\|_{L^2(\Omega)} = 0.$$

Together with the compactness of Ω_o this completes the proof. □

Remark 5.5. We pause for a moment to take a closer look at the Green's function $G^x \in L^2(\Omega)$ of the adjoint operator $\partial S[\hat{q}]^*$ at $x \in \Omega_o$ and its computation. Assume that the partial derivatives

$$a'_y(\hat{q}, S[\hat{q}])(\cdot,\cdot)\colon Y \times W \to \mathbb{R}, \quad a'_q(\hat{q}, S[\hat{q}])(\cdot,\cdot)\colon L^2(\Omega) \times W \to \mathbb{R}$$

of a at $(\hat{q}, S[\hat{q}])$ give continuous bilinear forms. Furthermore for $f \in W^*$ there exists a unique element $g_f \in Y$ with

$$a'_y(\hat{q}, S[\hat{q}])(g_f, \varphi) = \langle \varphi, f\rangle_{W,W^*} \quad \forall \varphi \in W,$$

and the mapping

$$T\colon W^* \to Y, \quad f \mapsto g_f,$$

is linear and continuous. Since W is reflexive there holds $T^*\colon Y^* \to W$. Furthermore we recall that $Y \overset{c}{\hookrightarrow} \mathcal{C}(\Omega_o)$ and consequently $\mathcal{M}(\Omega_o) \overset{c}{\hookrightarrow} Y^*$. Let $\delta q \in L^2(\Omega)$ be fixed for the moment. Then

$$f_{\delta q}\colon W \to \mathbb{R}, \quad \varphi \mapsto -a'_q(\hat{q}, S[\hat{q}])(\varphi, \delta q),$$

defines a linear and continuous functional on W, i.e. $f_{\delta_q} \in W^*$. We have $\partial S[\hat{q}]\delta q = T f_{\delta q}$. Now, given $x \in \Omega_o$ set $\mathcal{G}^x = T^*\delta_x \in W$. By construction, $\mathcal{G}^x$ fulfills the adjoint equation

$$a'_y(\hat{q}, S[\hat{q}])(\tilde{\varphi}, \mathcal{G}^x) = \langle \tilde{\varphi}, \delta_x\rangle_{Y,Y^*} \quad \forall \tilde{\varphi} \in \operatorname{Im} T. \tag{5.34}$$

Thus we obtain

$$(G^x, \delta q)_{L^2(\Omega)} = \partial S[\hat{q}]\delta q(x) = \langle \mathcal{G}^x, f_{\delta_q}\rangle_{W,W^*} = a'_y(\hat{q}, S[\hat{q}])(\partial S[\hat{q}]\delta q, \mathcal{G}^x) = -a'_q(\hat{q}, S[\hat{q}])(\delta q, \mathcal{G}^x).$$

From the continuity assumptions on the partial derivatives we infer that

$$-a'_q(\hat{q}, S[\hat{q}])(\cdot, \mathcal{G}^x)\colon L^2(\Omega) \to \mathbb{R}, \quad \delta q \mapsto -a'_q(\hat{q}, S[\hat{q}])(\delta q, \mathcal{G}^x),$$

gives a linear continuous functional on $L^2(\Omega)$. Applying the Riesz representation theorem it is identified with G^x. In particular, this implies that the evaluation of the mapping $G^{\cdot}$ at a spatial point $x \in \Omega_o$ requires the computation of the function $\mathcal{G}^x$, the Green's function of the operator T^*, fulfilling the partial differential equation in (5.34).

We proceed with the characterization of the Fisher-Information operator. For $x \in \Omega_o$ define

$$k^x \in L^2(\Omega \times \Omega) \quad \text{with} \quad k^x(y,z) = G^x(y)G^x(z), \quad \text{for } a.e.\ y, z \in \Omega.$$

From Section 3.1.1 we recall that $L^2(\Omega \times \Omega)$ is isometrically isomorphic to the space of Hilbert-Schmidt operators $\mathrm{HS}(L^2(\Omega), L^2(\Omega))$. We recall that this space and the space of self-adjoint Hilbert-Schmidt operators $\mathrm{SHS}(L^2(\Omega), L^2(\Omega))$ are Hilbert spaces with respect to the norm induced by the Hilbert-Schmidt inner product

$$\langle\langle B_1, B_2 \rangle\rangle_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))} = \mathrm{Tr}_{L^2(\Omega)}(B_1^* B_2) = \sum_{i=1}^{\infty} (B_1, \phi_i)_{L^2(\Omega)} (B_2, \phi_i)_{L^2(\Omega)},$$

for all $B_1, B_2 \in \mathrm{HS}(L^2(\Omega), L^2(\Omega))$. If $B \in \mathrm{SHS}(L^2(\Omega), L^2(\Omega))$ there exists a square-integrable function $k_B \in L^2(\Omega \times \Omega)$ with

$$[B\delta q](z) = \int_\Omega k_B(y,z)\delta q(y)\mathrm{d}y, \quad k_B(y,z) = k_B(z,y) \quad a.e.\ y, z \in \Omega.$$

The positive Hilbert-Schmidt operator corresponding to k^x is given by the rank 1 operator

$$I(x) = G^x \otimes G^x \in \mathrm{Pos}(L^2(\Omega), L^2(\Omega)),$$

which acts on $L^2(\Omega)$ via

$$\begin{aligned}(\delta q_1, I(x)\delta q_2)_{L^2(\Omega)} &= (\delta q_1, [G^x \otimes G^x]\delta q_2)_{L^2(\Omega)} = (G^x, \delta q_1)_{L^2(\Omega)}(G^x, \delta q_2)_{L^2(\Omega)} \\ &= \partial S[\hat{q}]\delta q_1(x)\, \partial S[\hat{q}]\delta q_2(x),\end{aligned}$$

for all $\delta q_1,\ \delta q_2 \in L^2(\Omega)$. We make the following observations.

Proposition 5.16. *The function*

$$I \colon \Omega_o \to \mathrm{SHS}(L^2(\Omega), L^2(\Omega)), \quad x \mapsto G^x \otimes G^x,$$

is uniformly continuous and thus Bochner-integrable with respect to $u \in \mathcal{M}(\Omega_o)$. There holds

$$\mathcal{I}(\boldsymbol{u}(\mathbf{x},\mathbf{u})) = X^* \Sigma^{-1} X = \int_{\Omega_o} I(x)\ \mathrm{d}u(x) = \int_{\Omega_o} G^x \otimes G^x\ \mathrm{d}u(x),$$

for every measurement setup

$$\mathbf{x} = (x_1, \dots, x_N)^\top \in \Omega_o^N, \quad \mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_N)^\top, \quad u = \sum_{i=1}^N \mathbf{u}_i \delta_{x_i} \in \mathcal{M}^+(\Omega_o).$$

Proof. The uniform continuity of the mapping $I(x)$ follows from Proposition 3.3. Thus it is Bochner integrable with respect to $u \in \mathcal{M}(\Omega_o)$. Let a vector of measurement positions $\mathbf{x} \in \Omega_o^N$, a vector of measurement weights $\mathbf{u} \in \mathbb{R}_+^N$ and $\delta q_1, \delta q_2 \in L^2(\Omega)$ be given. Define $u = \sum_{i=1}^N \mathbf{u}_i \delta_{x_i}$. By definition of the operator $X \in \mathcal{L}(L^2(\Omega), \mathbb{R}^N)$ and $\Sigma^{-1} \in \mathbb{R}^{N \times N}$, see (5.33), we obtain

$$(\delta q_1, \mathcal{I}(\boldsymbol{u}(\mathbf{x},\mathbf{u}))\delta q_2)_{L^2(\Omega)} = (X\delta q_1, \Sigma^{-1} X \delta q_2)_{\mathbb{R}^N} = \sum_{i=1}^N \mathbf{u}_i \partial S[\hat{q}]\delta q_1(x_i)\, \partial S[\hat{q}]\delta q_2(x_i).$$

In the same manner we compute

$$\begin{aligned}\left(\delta q_1, \left[\int_{\Omega_o} G^x \otimes G^x \,\mathrm{d}u(x)\right] \delta q_2\right)_{L^2(\Omega)} &= \sum_{i=1}^{N} \mathbf{u}_i(\delta q_1, [G^{x_i} \otimes G^{x_i}]\delta q_2)_{L^2(\Omega)} \\ &= \sum_{i=1}^{N} \mathbf{u}_i \partial S[\hat{q}]\delta q_1(x_i)\, \partial S[\hat{q}]\delta q_2(x_i).\end{aligned}$$

Since $\delta q_1,\ \delta q_2 \in L^2(\Omega)$ where chosen arbitrary combining both results yields the statement. □

Thus the sensor placement problem (5.32) fits into the general framework of Chapter 3 by choosing $Q = L^2(\Omega)$ and

$$\mathcal{O}\colon \Omega_o \to L^2(\Omega), \quad x \mapsto G^x.$$

In order to determine an optimal measurement setup we now interpret the distibution of the sensors on the spatial domain as a Radon measure and solve the sparse sensor placement problem

$$\min_{u \in \mathcal{M}^+(\Omega_o)} [\Psi(\mathcal{I}(u)) + \beta \|u\|_{\mathcal{M}}]. \tag{5.35}$$

Here, the Fisher operator $\mathcal{I}$ maps a given $u \in \mathcal{M}(\Omega_o)$ to the associated Bochner integral:

$$\mathcal{I}\colon \mathcal{M}(\Omega_o) \to \mathrm{SHS}(L^2(\Omega), L^2(\Omega)), \quad u \mapsto \int_{\Omega_o} G^x \otimes G^x \mathrm{d}u(x),$$

which fulfills

$$[\mathcal{I}(u)\delta q_1](y) = \int_{\Omega_o} G^x(y)(G^x, \delta q_1)_{L^2(\Omega)}\, \mathrm{d}u(x), \quad (\delta q_1, \mathcal{I}(u)\delta q_2)_{L^2(\Omega)} = \langle \partial S[\hat{q}]\delta q_1\, \partial S[\hat{q}]\delta q_2, u\rangle,$$

for all $\delta q_1, \delta q_2 \in L^2(\Omega)$ and almost every $y \in \Omega$. Its definition is formalized through the following proposition.

Proposition 5.17. *The mapping*

$$\mathcal{I}\colon \mathcal{M}(\Omega_o) \to \mathrm{SHS}(L^2(\Omega), L^2(\Omega)), \quad u \mapsto \int_{\Omega_o} G^x \otimes G^x \mathrm{d}u(x),$$

is linear continuous with

$$\|\mathcal{I}(u)\|_{\mathrm{HS}(L^2(\Omega), L^2(\Omega))} \le \max_{x \in \Omega_o} \|G^x\|^2_{L^2(\Omega)} \|u\|_{\mathcal{M}}.$$

Furthermore it is weak-to-strong continuous.*

Proof. For a proof of this statement see Proposition 3.4 and Theorem 3.8, setting $Q = L^2(\Omega)$ and $\mathcal{O}(x) = G^x$. □

We make the following general assumptions on the optimal design criterion Ψ.

Assumption 5.4. The function $\Psi\colon\ \mathrm{SHS}(L^2(\Omega), L^2(\Omega)) \to \mathbb{R} \cup \{+\infty\}$ satisfies:

A5.1 There holds $\mathrm{Pos}(L^2(\Omega), L^2(\Omega)) \subset \operatorname{dom} \Psi$.

A5.2 Ψ is two times continuously differentiable on $\operatorname{Pos}(L^2(\Omega), L^2(\Omega))$.

A5.3 Ψ is convex on $\operatorname{Pos}(L^2(\Omega), L^2(\Omega))$.

A5.4 Ψ is monotone in the sense that

$$B_2 - B_1 \in \operatorname{Pos}(L^2(\Omega), L^2(\Omega)) \Rightarrow \Psi(B_2) \leq \Psi(B_1) \quad \forall B_1,\ B_2 \in \operatorname{Pos}(L^2(\Omega), L^2(\Omega)).$$

We emphasize that all examples considered in Section 5.1.3 fit into these general assumptions. In the following lemma the adjoint operator of $\mathcal{I}$ is characterized.

Lemma 5.18. *The Banach-space adjoint of the operator* $\mathcal{I}\colon \mathcal{M}(\Omega_o) \to \operatorname{SHS}(L^2(\Omega), L^2(\Omega))$ *as defined in Proposition 5.17 is given by*

$$\mathcal{I}^*\colon\ \operatorname{SHS}(L^2(\Omega), L^2(\Omega)) \to \mathcal{C}(\Omega_o), \quad B \mapsto \varphi_B.$$

Here, the continuous function $\varphi_B \in \mathcal{C}(\Omega_o)$ *is given by* $\varphi_B(x) = (G^x, BG^x)_{L^2(\Omega)}$ *for all* $x \in \Omega_o$.

Proof. The statement is directly obtained by applying Proposition 3.7, setting $Q = L^2(\Omega)$ and $\mathcal{O}(x) = G^x$. □

We introduce the reduced problem formulation

$$\min_{u \in \mathcal{M}^+(\Omega_o)} F(u) = [\psi(u) + \beta \|u\|_{\mathcal{M}}], \tag{$\mathcal{P}_\beta$}$$

where the functional ψ is defined as $\psi = \Psi \circ \mathcal{I}$. The following proposition summarizes some key properties of ψ. These can be inferred from the general theory of Chapter 3, see Proposition 3.9 and Lemma 3.10.

Proposition 5.19. *Let Assumptions (**A5.1**)–(**A5.4**) be fulfilled. The operator* $\mathcal{I}$ *and the functional* ψ *satisfy:*

1. *For every* $u \in \mathcal{M}^+(\Omega_o)$ *there holds* $\mathcal{I}(u) \in \operatorname{Pos}(L^2(\Omega), L^2(\Omega))$.
2. *There holds* $\operatorname{dom}_{\mathcal{M}^+(\Omega_o)} \psi = \mathcal{M}^+(\Omega_o)$.
3. ψ *is at least two times continuously differentiable on* $\mathcal{M}^+(\Omega_o)$*. For* $u \in \mathcal{M}^+(\Omega_o)$ *its first derivative* $\psi'(u) = \mathcal{I}^* \nabla \Psi(\mathcal{I}(u)) \in \mathcal{C}(\Omega_o)$ *can be identified with the continuous function*

$$[\nabla \psi(u)](x) = (G^x, \nabla \Psi(\mathcal{I}(u)) G^x)_{L^2(\Omega)} = -\|(-\nabla \Psi(\mathcal{I}(u)))^{1/2} G^x\|_{L^2(\Omega)} \quad \forall x \in \Omega_o. \tag{5.36}$$

Moreover the gradient $\nabla \psi\colon \mathcal{M}^+(\Omega_o) \to \mathcal{C}(\Omega_o)$ *is weak*-to-strong continuous.*

Given $u \in \mathcal{M}^+(\Omega_o)$*, the second derivative* $\nabla^2 \psi(u) \in \mathcal{L}(\mathcal{M}(\Omega_o), \mathcal{M}(\Omega_o)^*)$ *is characterized as*

$$\langle \delta u_1, \nabla^2 \psi(u) \delta u_2 \rangle_{\mathcal{M}, \mathcal{M}^*} = \operatorname{Tr}_{L^2(\Omega)}(\mathcal{I}(\delta u_1) \nabla^2 \Psi(\mathcal{I}(u)) \mathcal{I}(\delta u_2)), \quad \forall \delta u_1, \delta u_2 \in \mathcal{M}(\Omega_o).$$

5. ψ *is convex on* $\mathcal{M}^+(\Omega_o)$.
6. ψ *is monotone in the sense that*

$$\mathcal{I}(u_2 - u_1) \in \operatorname{Pos}(L^2(\Omega), L^2(\Omega)) \Rightarrow \psi(u_2) \leq \psi(u_1) \quad \forall u_1,\ u_2 \in \mathcal{M}^+(\Omega_o).$$

To ensure the existence of a solution to $(\mathcal{P}_\beta)$ we impose additional assumptions on the objective functional F.

Assumption 5.5. The objective functional F is radially unbounded on $\mathcal{M}^+(\Omega_o)$, i.e. given a sequence $\{u_k\}_{k\in\mathbb{N}} \subset \mathcal{M}^+(\Omega_o)$ there holds

$$\|u_k\|_{\mathcal{M}} \to \infty \Rightarrow F(u_k) \to \infty.$$

Observe that the regularization term in $(\mathcal{P}_\beta)$ can be written as

$$G_\beta(\|u\|_{\mathcal{M}}) = \beta\|u\|_{\mathcal{M}} \quad \text{where} \quad G_\beta \colon \mathbb{R} \to \mathbb{R}, \quad m \mapsto \beta m + I_{[0,\infty)}(m).$$

The functional G_β obviously fulfills Assumption 3.3. Hence the following results on the existence and characterization of optimal measurement designs can be derived from the theory presented in Chapter 3. We therefore omit most of the standard proofs and give references to the general results where necessary.

Theorem 5.20. *Let $\beta > 0$ be given and let Assumption 5.5 hold. Then there exists at least one optimal solution $\bar{u}_\beta \in \mathcal{M}^+(\Omega_o)$ to $(\mathcal{P}_\beta)$ and the set of optimal solutions is bounded. If the design criterion Ψ is strictly convex on* $\operatorname{Pos}(L^2(\Omega), L^2(\Omega))$ *then the optimal Fisher information $\mathcal{I}(\bar{u}_\beta)$ is the same for every optimal solution.*

Proof. Given Assumptions 5.4 and 5.5, this statement follows from Proposition 3.11. □

The following example discusses Assumption 5.5 in the context of A and D-optimality.

Example 5.5. *Obviously Assumption 5.5 is fulfilled if Ψ is nonnegative on* $\operatorname{Pos}(L^2(\Omega), L^2(\Omega))$ *since then*

$$\beta\|u\|_{\mathcal{M}} \leq \Psi(\mathcal{I}(u)) + \beta\|u\|_{\mathcal{M}} = F(u) \quad \forall u \in \mathcal{M}^+(\Omega_o)$$

This is e.g. the case for the A-optimal design criterion

$$\Psi_A(\mathcal{I}(u)) = \operatorname{Tr}_{L^2(\Omega)}(\mathcal{C}_{post}(\mathcal{I}(u))).$$

For the D-optimal design criterion

$$\Psi_D(\mathcal{I}(u)) = -\log(\operatorname{Det}(\mathcal{I}_0^{-1/2}\mathcal{I}(u)\mathcal{I}_0^{-1/2} + \operatorname{Id})),$$

the situation is more involved. To prove the radial unboundedness of $\Psi_D(\mathcal{I}(u)) + \beta\|u\|_{\mathcal{M}}$ we proceed as follows. Given $u \in \mathcal{M}^+(\Omega_o)$ we conclude

$$-\log(\operatorname{Det}(\max_{x\in\Omega_o} \|G^x\|^2_{L^2(\Omega)}\|u\|_{\mathcal{M}}\mathcal{I}_0^{-1} + \operatorname{Id})) \leq -\log(\operatorname{Det}(\mathcal{I}_0^{-1/2}\mathcal{I}(u)\mathcal{I}_0^{-1/2} + \operatorname{Id})).$$

by a first-order Taylor approximation. Define the differentiable function

$$f \colon \mathbb{R}_+ \to \mathbb{R}, \quad m \mapsto -\log(\operatorname{Det}(\max_{x\in\Omega_o} \|G^x\|^2_{L^2(\Omega)} m\mathcal{I}_0^{-1} + \operatorname{Id})).$$

We calculate

$$f'(m) = -\max_{x\in\Omega_o} \|G^x\|^2_{L^2(\Omega)} \operatorname{Tr}_{L^2(\Omega)}((\max_{x\in\Omega_o} \|G^x\|^2_{L^2(\Omega)} m\mathcal{I}_0^{-1} + \operatorname{Id})^{-1}\mathcal{I}_0^{-1}).$$

Denoting by $\{\lambda_i\}_{i\in\mathbb{N}}$ *the eigenvalues of* $\mathcal{I}_0^{-1}$ *ordered by decreasing magnitude we can calculate the trace explicitly to arrive at*

$$f'(m) = -\max_{x\in\Omega_o} \|G^x\|^2_{L^2(\Omega)} \sum_{i=1}^{\infty} \lambda_i (1 + \max_{x\in\Omega_o} \|G^x\|^2_{L^2(\Omega)} m\lambda_i)^{-1}.$$

Let an arbitrary $\varepsilon > 0$ *be given and fix* $M_1 > 0$*. There exists an index* $K \in \mathbb{N}$ *with*

$$\max_{x\in\Omega_o} \|G^x\|^2_{L^2(\Omega)} \sum_{i=K+1}^{\infty} \lambda_i (1 + \max_{x\in\Omega_o} \|G^x\|^2_{L^2(\Omega)} m\lambda_i)^{-1} < \frac{\varepsilon}{2},$$

for all $m \geq M_1$*. Furthermore there exists* $M_2 > 0$ *such that*

$$\max_{x\in\Omega_o} \|G^x\|^2_{L^2(\Omega)} \sum_{i=1}^{K} \lambda_i (1 + \max_{x\in\Omega_o} \|G^x\|^2_{L^2(\Omega)} m\lambda_i)^{-1} < \frac{\varepsilon}{2},$$

for all $m \geq M_2$*. Combining both statements yields*

$$|f'(m)| = \max_{x\in\Omega_o} \|G^x\|^2_{L^2(\Omega)} \sum_{i=1}^{\infty} \lambda_i (1 + \max_{x\in\Omega_o} \|G^x\|^2_{L^2(\Omega)} m\lambda_i)^{-1} < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

for all $m \geq \max\{M_1, M_2\}$*. Since* ε *was chosen arbitrary we conclude* $\lim_{m\to\infty} f'(m) = 0$ *and, applying L'Hôspital's rule,*

$$0 = \lim_{\|u\|_{\mathcal{M}}\to\infty} \frac{f(\|u\|_{\mathcal{M}})}{\|u\|_{\mathcal{M}}} = \lim_{\|u\|_{\mathcal{M}}\to\infty} f'(\|u\|_{\mathcal{M}}) \leq \lim_{\|u\|_{\mathcal{M}}\to\infty} \frac{\Psi_D(\mathcal{I}(u))}{\|u\|_{\mathcal{M}}} \leq 0.$$

Consequently for $\|u\|_{\mathcal{M}}$ *large enough we have*

$$\frac{\beta}{2}\|u\|_{\mathcal{M}} \leq \Psi_D(u) + \beta\|u\|_{\mathcal{M}},$$

and we deduce that $F(u) = \Psi_D(\mathcal{I}(u)) + \beta\|u\|_{\mathcal{M}}$ *is radially unbounded.*

The rest of this section focuses on the structure of optimal design measures and their behaviour for large $\beta > 0$. We start by deriving necessary and sufficient first-order optimality conditions characterizing the support of $\bar{u}_\beta$.

Theorem 5.21. *Let* $\beta > 0$ *be given. A measure* $\bar{u}_\beta \in \mathcal{M}^+(\Omega_o)$ *is a minimizer of* $(\mathcal{P}_\beta)$ *if and only if one of the following (equivalent) conditions holds*

- *There holds*

$$-\nabla\psi(\bar{u}_\beta) \in \beta\partial\|\bar{u}_\beta\|_{\mathcal{M}} + \partial I_{u\geq 0}(\bar{u}_\beta).$$

- *There holds*

$$\sup_{v\in\mathcal{M}^+(\Omega_o)} [\langle\nabla\psi(\bar{u}_\beta), \bar{u}_\beta - v\rangle + \beta\|\bar{u}_\beta\|_{\mathcal{M}} - \beta\|v\|_{\mathcal{M}}] = 0.$$

- *There holds*

$$-\min_{x\in\Omega_o}\nabla\psi(\bar{u}_\beta)(x)\begin{cases}=\beta & \|\bar{u}_\beta\|_{\mathcal{M}}>0\\ \leq\beta & \|\bar{u}_\beta\|_{\mathcal{M}}=0\end{cases},\quad -\langle\nabla\psi(\bar{u}_\beta),\bar{u}_\beta\rangle=\beta\|\bar{u}_\beta\|_{\mathcal{M}}.$$

- *There holds*

$$-\min_{x\in\Omega_o}\nabla\psi(\bar{u}_\beta)(x)\begin{cases}=\beta & \|\bar{u}_\beta\|_{\mathcal{M}}>0\\ \leq\beta & \|\bar{u}_\beta\|_{\mathcal{M}}=0\end{cases},\quad \operatorname{supp}\bar{u}_\beta\subset\{\,x\in\Omega_o\mid -\nabla\psi(\bar{u}_\beta)(x)=\beta\,\}.$$

Proof. Since ψ is two times differentiable and monotone we have $-\nabla\psi(u)(x)\geq 0$ and thus also

$$-\min_{x\in\Omega_o}\nabla\psi(u)(x)\geq 0$$

for all measures $u\in\mathcal{M}^+(\Omega_o)$ and $x\in\Omega_o$. Calculating the subdifferential of G_β at $\|\bar{u}_\beta\|_{\mathcal{M}}$ gives

$$\partial G_\beta(\|\bar{u}_\beta\|_{\mathcal{M}})=\{\beta\}+\partial I_{[0,\infty)}(\|\bar{u}_\beta\|_{\mathcal{M}})=\begin{cases}(-\infty,\beta] & \|\bar{u}_\beta\|_{\mathcal{M}}=0\\ \{\beta\} & \|\bar{u}_\beta\|_{\mathcal{M}}>0\end{cases}.$$

Furthermore we note that

$$\partial(\beta\|\cdot\|_{\mathcal{M}}+I_{u\geq 0}(\cdot))(\tilde{u})=\beta\partial\|\tilde{u}\|_{\mathcal{M}}+\partial I_{u\geq 0}(\tilde{u})$$

for all $\tilde{u}\in\mathcal{M}(\Omega)$ due to the continuity of the norm. Thus we obtain the result by applying Theorem 3.17 as in Example 3.3. □

Remark 5.6. As in the finite dimensional case we stress that similar equivalent optimality conditions can be derived for the norm constrained problem

$$\min_{u\in\mathcal{M}^+(\Omega_o)}\psi(u)\quad s.t.\quad \|u\|_{\mathcal{M}}\leq K,$$

given a maximum cost $K>0$ for the measurements. For the sake of brevity we resign from stating them here and refer to the general case in Theorem 3.17 as well as Example 3.4.

In contrast to the situation discuessed in the previous chapter, existence of sparse minimizers to $(\mathcal{P}_\beta)$ may not be guaranteed since the parameter space is no longer finite dimensional. This issue is addressed in the following corollaries. In general conclusions on the sparsity pattern of minimizers to $(\mathcal{P}_\beta)$ can be based on the support condition stated in Theorem 5.21. Additionally the choice of the cost parameter $\beta>0$ provides some indirect control on the support size of optimal designs and thus the number of measurements. Last we emphasize that all optimal designs are well-approximated by suboptimal sparse design measures up to arbitrary accuracy in a sense made clear below.

Corollary 5.22. *Denote by $\bar{u}_\beta\in\mathcal{M}^+(\Omega_o)$ an optimal design such that*

$$\{\,x\in\Omega_o\mid -\nabla\psi(\bar{u}_\beta)(x)=\beta\,\}=\{\bar{x}_i\}_{i=1}^N$$

Then $\bar{u}_\beta$ is given as a conic linear combination $\bar{u}_\beta=\sum_{i=1}^N\bar{\mathbf{u}}_i\delta_{\bar{x}_i}$ for some $\bar{\mathbf{u}}_i\in\mathbb{R}_+$, $i=1,\dots,N$.

Proof. Since $\bar{u}_\beta$ minimizes in $(\mathcal{P}_\beta)$ we infer

$$\operatorname{supp} \bar{u}_\beta \subset \{\, x \in \Omega_o \mid -\nabla\psi(\bar{u}_\beta)(x) = \beta \,\} = \{\bar{x}_i\}_{i=1}^N,$$

from Theorem 5.21. This finishes the proof. □

Corollary 5.23. *There exists $\beta_0 > 0$ such that for all $\beta \geq \beta_0$ the unique solution to $(\mathcal{P}_\beta)$ is given by the zero measure.*

Proof. The statement can be derived along the lines of proof in Proposition 4.10. □

Corollary 5.24. *Let an arbitrary minimizer $\bar{u}_\beta$ to $(\mathcal{P}_\beta)$ be given. For all $\varepsilon > 0$ there exists $\bar{u}_\varepsilon \in \mathcal{M}^+(\Omega_o)$ with*

$$\bar{u}_\varepsilon \in \operatorname{cone}\{\, \delta_x \mid x \in \Omega_o \,\}, \quad F(\bar{u}_\varepsilon) - F(\bar{u}_\beta) < \varepsilon.$$

Proof. Let an optimial design $\bar{u}_\beta$ to $(\mathcal{P}_\beta)$ and $\varepsilon > 0$ be given. Following [50, Appendix A] there exists a sequence $\{u_k\}_{k\in\mathbb{N}} \subset \mathcal{M}^+(\Omega_o)$ with

$$u_k \in \operatorname{cone}\{\, \delta_x \mid x \in \Omega_o \,\} \quad \forall k \in \mathbb{N}, \quad u_k \rightharpoonup^* \bar{u}_\beta.$$

The claimed statement now follows noting that

$$\|u_k\|_{\mathcal{M}} = \langle 1, u_k\rangle \to \langle 1, \bar{u}_\beta\rangle = \|\bar{u}_\beta\|_{\mathcal{M}}, \quad \psi(u_k) \to \psi(\bar{u}_\beta),$$

due to the weak* convergence of $\{u_k\}_{k\in\mathbb{N}}$, weak*-to-strong continuity of $\mathcal{I}$ and continuity of Ψ on $\operatorname{Pos}(L^2(\Omega), L^2(\Omega))$. □

5.2 Discretization and error estimates

In the following section we present a suitable approximation framework for the Bayesian sensor placement problem $(\mathcal{P}_\beta)$. Therefore we proceed in two steps, first starting with a discretization of the underlying state and sensitivity equations by linear finite elements. In contrast neither the parameter space $L^2(\Omega)$ nor the space of design measures $\mathcal{M}^+(\Omega)$ is discretized. Again, this can be interpreted as a variational discretization approach. We discuss well-posedness of the FE-discretized sensor placement problem and derive first-order optimality conditions. Most important a careful study of the discrete Fisher information operator reveals that the FE-discretized problem is equivalent to an additional discretization of the parameter space and the restriction of possible sensor locations to the grid nodes of the mesh. Finally we prove convergence of the discrete optimal design measures towards optimal solutions of $(\mathcal{P}_\beta)$ and present a priori error estimates.

While a finite element discretization of the sensitivity equation implicitly leads to a finite dimensional optimization problem the discretized parameter space is in general high-dimensional. This makes a direct evaluation of the design criterion and its derivatives computationally infeasible and thus prohibits its numerical solution, see also Section 5.3.2. Furthermore it may depend on the linearization point $\hat{q}$ which may differ significantly from the true value of the parameter. Therefore, in a second step, we consider the approximation of the parameter space $L^2(\Omega)$ through the subspace spanned by the first n eigenvectors of the a priori covariance operator. This approach corresponds to a truncation of the Karhunen-Lòeve expansion corresponding to q after n terms, i.e. we only consider the directions in which the prior distribution admits the largest uncertainty. Subsequently, sensors are placed to optimally infer the coefficients in this basis representation.

Last, to obtain a computationally feasible problem, we combine both discretization concepts and analyse the resulting fully discrete problem. Convergence results for the discrete optimal design measurements as well as a priori error estimates with respect to the spatial mesh-size h and the tail sum of the eigenvalues corresponding to the neglected eigenvectors are derived. The results are illustrated on the A and D-optimal design problem highlighting the practical relevance of the proposed approach.

Efficient computational methods for discrete approximations of sensor placement problems associated to infinite dimensional Bayesian inference are e.g. considered in [4, 6]. In contrast to the present work the authors consider only finitely many candidate locations for the placement of the sensors. The PDE constraints as well as the underlying parameter space are approximated using a finite element ansatz and the coefficients of the unknown parameter in the corresponding basis expansion are treated as random variables. This results in high-dimensional discrete parameter spaces and large covariance matrices. Evaluating e.g. the A-optimal design criterion in this situation requires calculating the trace of the inverse to a large and dense matrix, usually stemming from the discretization of the fractional power of an elliptic differential operator. In order to make the resulting discrete sensor placement problems computationally feasible, different tools from stochastic linear algebra such as randomized trace estimation, [231], are applied and low-rank approximations of the design-dependent posterior covariance operator are considered. In particular the authors exploit the low-rank rank structure of the parameter-to-observable map due to the finite number of sensors. A comprehensive comparison between several existing approaches including their computational costs is provided in [6] together with stability results for the evaluation of the optimal design criterion and its gradient. However, we are not aware of any pre-existing works dealing with the case of vanishing discretization parameters or a priori error estimation.

5.2.1 Finite element discretization

We first discuss a discretization of $(\mathcal{P}_\beta)$ based on a finite element ansatz for the underlying state and sensitivity equations. In the following, the sets Ω and Ω_o are assumed to be polytopal (i.e. polygonal in two dimensions and polyhedral in three dimensions). We consider a family of triangulations $\{\, \mathcal{T}_h \,\}_{h>0}$ of Ω which resolve the spatial domain Ω as well as the observational domain Ω_o

$$\Omega = \bigcup_{T \in \mathcal{T}_h} \bar{T}, \quad \Omega_o = \bigcup_{T \in \mathcal{T}_h^o} \bar{T}. \tag{5.37}$$

Here $\mathcal{T}_h^o \subset \mathcal{T}_h$ denotes the union of all cells making up the observational domain. To each triangulation we assign a positive scalar $h > 0$ denoting the maximal diameter of a cell $K \in \mathcal{T}_h$.

By $\mathcal{N}_h$ we denote the set of nodes of the triangulation. For each $h > 0$ the space of continuous piecewise linear finite elements V_h on $\mathcal{T}_h$ and its dual space $V_h^* \simeq \mathcal{M}_h$ are defined as before by

$$V_h = \{\, y_h \in \mathcal{C}(\bar{\Omega}) \mid y_{h_{|T}} \in P_1 \ \forall T \in \mathcal{T}_h \,\}, \quad \mathcal{M}_h = \{\, u_h \in \mathcal{M}(\bar{\Omega}) \mid \operatorname{supp} u_h \subset \mathcal{N}_h \,\}.$$

By $Y_h \subset V_h$ as well as $W_h \subset V_h$ we denote the discretized state and solution spaces, respectively. We recall the nodal interpolation operators $i_h \colon \mathcal{C}(\bar{\Omega}) \to V_h$ and $\Lambda_h \colon \mathcal{M}(\bar{\Omega}) \to \mathcal{M}_h$ as

$$i_h(y) = \sum_{x_i \in \mathcal{N}_h} y(x_i) e_i^h, \quad \Lambda_h(u) = \sum_{x_i \in \mathcal{N}_h} \langle e_i^h, u \rangle \delta_{x_i}$$

where e_i^h, $i \in \{1, \dots, \#\mathcal{N}_h\}$ denotes the nodal basis function associated to a node $x_i \in \mathcal{N}_h$. The discretized state equation is described by a continuously differentiable form

$$a_h \colon Q_{ad} \times Y_h \times W_h \to \mathbb{R}.$$

For a given $q \in Q_{ad}$ an element $y^h \in Y_h$ is called an associated state if

$$a_h(q, y^h)(\varphi_h) = 0 \quad \forall \varphi_h \in W_h. \tag{5.38}$$

In the following we assume existence and uniqueness of the state $y^h = S^h[\hat{q}]$. Analogously, given $\delta q \in L^2(\Omega)$, the discrete sensitivity $\delta y^h \in Y_h$ at the a priori guess $\hat{q} \in Q_{ad}$ is a solution to

$$a'_{h,y}(\hat{q}, y^h)(\delta y^h, \varphi_h) = -a'_{h,q}(\hat{q}, \hat{y}^h)(\delta q, \varphi_h) \quad \forall \varphi_h \in W_h, \tag{5.39}$$

where $\hat{y}^h = S^h[\hat{q}]$. The forms $a'_{h,y}$, $a'_{h,q}$ denote the partial derivatives of a_h with respect to the state and the parameter. For the remainder of this section we make the following existence and stability assumptions for the considered discretization.

Assumption 5.6. There exists $h_0 > 0$ such that for all $h \leq h_0$, $\hat{q} \in Q_{ad}$ and $\delta q \in L^2(\Omega)$ the discrete state and sensitivity equations, (5.38) and (5.39), admit unique solutions $y^h = S^h[\hat{q}]$ and $\delta y^h = \partial S^h[\hat{q}]\delta q$. Moreover the operator $\partial S^h[\hat{q}] \colon L^2(\Omega) \to \mathcal{C}(\Omega_o)$ is linear and continuous and there exists a positive, strictly monotonically increasing and continuous function $\gamma \colon \mathbb{R}_+ \to \mathbb{R}_+$ with $\lim_{h \to +0} \gamma(h) = 0$ and a constant $c > 0$ independent of h such that

$$\|(\partial S[\hat{q}] - \partial S^h[\hat{q}])\delta q\|_{\mathcal{C}} \leq c\gamma(h)\|\delta q\|_{L^2(\Omega)},$$

holds for every $\delta q \in L^2(\Omega)$.

Discretization of $(\mathcal{P}_\beta)$ and stability estimates

Let $h \leq h_0$ be given. As in the continuous case we observe

$$\partial S^h[\hat{q}]\delta q\,(x) = \langle \partial S^h[\hat{q}]\delta q, \delta_x \rangle = (\partial S^h[\hat{q}]^* \delta_x, \delta q_2)_{L^2(\Omega)} = (G_h^x, \delta q)_{L^2(\Omega)},$$

for all $x \in \Omega_o$, $\delta q \in L^2(\Omega)$ and $G_h^x = \partial S^h[\hat{q}]^* \delta_x$. Due to the compactness of the operator $\partial S^h[\hat{q}] \colon L^2(\Omega) \to \mathcal{C}(\Omega_o)$ the mapping

$$G_h^{\cdot} \colon \Omega_o \to L^2(\Omega), \quad x \mapsto G_h^x,$$

is uniformly continuous. We now define the finite element approximation to $(\mathcal{P}_\beta)$ by

$$\min_{u_h \in \mathcal{M}^+(\Omega_o)} F_h(u_h) = [\psi_h(u_h) + \beta \|u_h\|_{\mathcal{M}}], \tag{$\mathcal{P}_{\beta,h}$}$$

where the reduced functional is given as $\psi_h(u) = \Psi(\mathcal{I}_h(u))$. Here, the discrete Fisher information operator $\mathcal{I}_h$ stems from a straightforward discretization of $\mathcal{I}$ by

$$\mathcal{I}_h \colon \mathcal{M}(\Omega_o) \to \mathrm{SHS}(L^2(\Omega), L^2(\Omega)), \quad u \mapsto \int_{\Omega_o} [G_h^x \otimes G_h^x] \, \mathrm{d}u(x).$$

Accordingly we define the discrete pointwise Fisher information as

$$I_h \colon \Omega_o \to \mathrm{SHS}(L^2(\Omega), L^2(\Omega)), \quad x \mapsto G_h^x \otimes G_h^x.$$

Note that we neither discretize the space of design measures $\mathcal{M}^+(\Omega_o)$ nor the parameter space $L^2(\Omega)$. This corresponds to a variational discretization approach.

Before proving the existence of minimizers to $(\mathcal{P}_{\beta,h})$ we present several stability results for the Green's function G_h^x and the Fisher information operator $\mathcal{I}_h$.

Lemma 5.25. *For all $h \leq h_0$ there holds*

$$\max_{x \in \Omega_o} \|G^x - G_h^x\|_{L^2(\Omega)} \leq c\gamma(h), \tag{5.40}$$

for some constant $c > 0$ independent of h.

Proof. By definition of G^x and G_h^x we obtain

$$\|G^x - G_h^x\|_{L^2(\Omega)} = \sup_{\substack{\delta q \in L^2(\Omega), \\ \|\delta q\|_{L^2(\Omega)} = 1}} (G^x - G_h^x, \delta q)_{L^2(\Omega)} = \langle \partial S[\hat{q}]\delta q - \partial S^h[\hat{q}]\delta q, \delta x \rangle \leq c\gamma(h),$$

for all $x \in \Omega_o$ using the estimate from Assumption 5.6. □

Proposition 5.26. *For all $h \leq h_0$ small enough we have*

$$\max_{x \in \Omega_o} \|I(x) - I_h(x)\|_{\mathrm{HS}(L^2(\Omega), L^2(\Omega))} + \|\mathcal{I} - \mathcal{I}_h\|_{\mathcal{L}(\mathcal{M}(\Omega_o), \mathrm{HS}(L^2(\Omega), L^2(\Omega)))} \leq c\gamma(h),$$

for some constant $c > 0$ independent of h.

Proof. Denote by $\{\phi_i\}_{i \in \mathbb{N}}$ an orthonormal basis of $L^2(\Omega)$ and fix $x \in \Omega_o$. By definition we have

$$\begin{aligned} \|I(x) - I_h(x)\|^2_{\mathrm{HS}(L^2(\Omega), L^2(\Omega))} &= \mathrm{Tr}_{L^2(\Omega)}((I(x) - I_h(x))(I(x) - I_h(x))) \\ &= \sum_{i=1}^{\infty} \|(I(x) - I_h(x))\phi_i\|^2_{L^2(\Omega)}. \end{aligned}$$

Fix an an arbitrary index $i \in \mathbb{N}$. We estimate

$$\begin{aligned} \|(I(x) - I_h(x))\phi_i\|_{L^2(\Omega)} &= \|G^x(G^x, \phi_i)_{L^2(\Omega)} - G^x(G^x, \phi_i)_{L^2(\Omega)}\|_{L^2(\Omega)} \\ &\leq |(G_h^x, \phi_i)_{L^2(\Omega)}| \|G^x - G_h^x\|_{L^2(\Omega)} + \|G^x\|_{L^2(\Omega)} |(G^x - G_h^x, \phi_i)_{L^2(\Omega)}|. \end{aligned}$$

Squaring both sides and applying Young's inequality we conclude

$$\|(I(x) - I_h(x))\phi_i\|^2_{L^2(\Omega)} \leq 2((G_h^x, \phi_i)^2_{L^2(\Omega)} \|G^x - G_h^x\|^2_{L^2(\Omega)} + \|G^x\|^2_{L^2(\Omega)} (G^x - G_h^x, \phi_i)^2_{L^2(\Omega)}).$$

Recall that due to Parseval's identity there holds $\|v\|^2_{L^2(\Omega)} = \sum_{i=1}^{\infty}(v,\phi_i)^2_{L^2(\Omega)}$ for all $v \in L^2(\Omega)$. Summing over all $i \in \mathbb{N}$ in the above inequality we arrive at

$$\|(I(x) - I_h(x))\|^2_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))} \leq 2(\|G^x\|^2_{L^2(\Omega)} + \|G^x_h\|^2_{L^2(\Omega)})\|G^x - G^x_h\|^2_{L^2(\Omega)}.$$

The norms of G^x and G^x_h are uniformly bounded with respect to $x \in \Omega_o$ and h. Applying estimate (5.40) thus yields

$$\max_{x\in\Omega_o} \|I(x) - I_h(x)\|^2_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))} \leq c\gamma(h)^2,$$

for some $c > 0$. The first result is now obtained through taking the square root. The stability estimate for $\mathcal{I}$ is deduced immediately from

$$\|\mathcal{I}(u) - \mathcal{I}_h(u)\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))} \leq \max_{x\in\Omega_o} \|I(x) - I_h(x)\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))}\|u\|_{\mathcal{M}},$$

for all $u \in \mathcal{M}(\Omega_o)$. □

Roughly speaking, the following proposition states that given an arbitrary design measure a better Fisher information can be obtained at a lower cost by only placing sensors in the nodes of the triangulation.

Proposition 5.27. *Let $h > 0$ and $u \in \mathcal{M}^+(\Omega_o)$ be given. Then there holds*

$$\mathcal{I}_h(\Lambda_h u) - \mathcal{I}_h(u) \in \mathrm{Pos}(L^2(\Omega), L^2(\Omega)), \quad \|\Lambda_h u\|_{\mathcal{M}} \leq \|u\|_{\mathcal{M}}. \tag{5.41}$$

Proof. We proceed similarly to the proof of Theorem 4.45. Let an arbitrary but fixed $u \in \mathcal{M}^+(\Omega_o)$ and $z \in L^2(\Omega)$ be given. The second statement, $\|\Lambda_h u\|_{\mathcal{M}} \leq \|u\|_{\mathcal{M}}$, follows from elementary properties of Λ_h, see [59, Theorem 3.5]. Let us proof the first one. Testing with $z \in L^2(\Omega)$ we obtain

$$(z, \mathcal{I}_h(u)z)_{L^2(\Omega)} = \left\langle (G^{\cdot}_h, z)^2_{L^2(\Omega)}, u \right\rangle = \left\langle \left(\partial S^h[\hat{q}]z\right)^2, u \right\rangle = \left\langle \left(\sum_{x_j\in\mathcal{N}_h} e^h_j\, \partial S^h[\hat{q}]z(x_j)\right)^2, u \right\rangle.$$

Now, we estimate

$$\left\langle \left(\sum_{x_j\in\mathcal{N}_h} e^h_j \partial S^h[\hat{q}]z(x_j)\right)^2, u \right\rangle \leq \left\langle \sum_{x_j\in\mathcal{N}_h} e^h_j \left(\partial S^h[\hat{q}]z(x_j)\right)^2, u \right\rangle,$$

with Jensen's inequality, using the convexity of the square function and $\sum_{x_i\in\mathcal{N}_h} e^h_i(x) = 1$ for all $x \in \Omega_o$. From this point on we follow exactly the steps in the proof of Theorem 4.45 obtaining

$$\left\langle \sum_{x_j\in\mathcal{N}_h} e^h_j \left(\partial S^h[\hat{q}]z(x_j)\right)^2, u \right\rangle = \left\langle i_h \left(\partial S^h[\hat{q}]z\right)^2, u \right\rangle = \left\langle \left(\partial S^h[\hat{q}]z\right)^2, \Lambda_h u \right\rangle.$$

Thus we conclude

$$(z, \mathcal{I}_h(u)z)_{L^2(\Omega)} \leq \left\langle \left(\partial S^h[\hat{q}]z\right)^2, \Lambda_h u \right\rangle = \left\langle (G^{\cdot}_h, z)^2_{L^2(\Omega)}, \Lambda_h u \right\rangle = (z, \mathcal{I}_h(\Lambda_h u)z)_{L^2(\Omega)}.$$

Since $z \in L^2(\Omega)$ was arbitrary, this implies $\mathcal{I}_h(\Lambda_h u) - \mathcal{I}_h(u) \in \mathrm{Pos}(L^2(\Omega), L^2(\Omega))$ finishing the proof. □

We are now ready to prove well-posedness of $(\mathcal{P}_{\beta,h})$. In addition, using the results of the previous proposition, there exist optimal measurement designs supported in $\mathcal{N}_h$.

Theorem 5.28. *Let $\beta > 0$ be given. For all $h \leq h_0$ small enough there exists at least one minimizer $\bar{u}_{\beta,h} \in \mathcal{M}^+(\Omega_o)$ to $(\mathcal{P}_{\beta,h})$ fulfilling*

$$-\nabla\psi_h(\bar{u}_{\beta,h}) \leq \beta, \quad \operatorname{supp} \bar{u}_{\beta,h} \subset \{x \in \Omega_o | -\nabla\psi_h(\bar{u}_{\beta,h})(x) = \beta\}.$$

Here the discrete gradient is given by

$$-\nabla\psi_h(\bar{u}_{\beta,h})(x) = -(G_h^x, \nabla\Psi(\mathcal{I}_h(\bar{u}_{\beta,h})G_h^x)_{L^2(\Omega)} = \|(-\nabla\Psi(\mathcal{I}_h(\bar{u}_{\beta,h}))^{1/2}G_h^x\|^2_{L^2(\Omega)},$$

for all $x \in \Omega_o$. Moreover the set of minimizers to $(\mathcal{P}_{\beta,h})$ is bounded uniformly in h. Given a sequence of discrete optimal designs $\{\bar{u}_{\beta,h}\}_{h>0}$ it admits at least one weak accumulation point and every accumulation point $\bar{u}_\beta$ is an optimal solution to $(\mathcal{P}_\beta)$.*

Proof. Let $h \leq h_0$. From Theorem 5.20 we recall that the set of continuous optimal designs is bounded by a constant M_0. Consider the auxiliary problem

$$\min_{u \in \mathcal{M}^+(\Omega_o)} [\Psi(\mathcal{I}_h(u)) + \beta\|\bar{u}_{\beta,h}\|_\mathcal{M}] \quad s.t. \quad \|u\|_\mathcal{M} \leq 2M_0. \tag{5.42}$$

Since F_h is weak*-to-strong continuous on $\mathcal{M}^+(\Omega_o)$ this problem admits at least one minimizer $\bar{u}_{\beta,h}$. We proceed to show that the additional norm constraint is inactive if $h > 0$ is chosen small enough. By construction the sequence $\{\bar{u}_{\beta,h}\}_{h>0}$ is uniformly bounded. Extracting a weak* convergent subsequence $\bar{u}_{\beta,h} \rightharpoonup^* \bar{u} \in \mathcal{M}^+(\Omega_o)$ denoted by the same symbol, we note

$$\mathcal{I}_h(\bar{u}_{\beta,h}) \to \mathcal{I}(\bar{u}), \quad \|\bar{u}_{\beta,h}\|_\mathcal{M} \to \|\bar{u}\|_\mathcal{M}, \quad \psi_h(\bar{u}_{\beta,h}) \to \psi(\bar{u}),$$

as h tends to 0. Let an arbitrary minimizer $\bar{u}_\beta$ of $(\mathcal{P}_\beta)$ be given. Since $\|\bar{u}_\beta\|_\mathcal{M} < 2M_0$ we conclude

$$F(\bar{u}) = \lim_{h\to 0} F_h(\bar{u}_{\beta,h}) \leq \lim_{h\to 0} F_h(\bar{u}_\beta) = F(\bar{u}_\beta).$$

Thus $\bar{u}$ is a minimizer of $(\mathcal{P}_\beta)$. In particular this yields $\|\bar{u}\|_\mathcal{M} < 2M_0$. From the weak* convergence of $\{\bar{u}_{\beta,h}\}_{h>0}$ the same holds for $\bar{u}_{\beta,h}$, choosing h small enough, since

$$\|\bar{u}_{\beta,h}\|_\mathcal{M} = \langle 1, \bar{u}_{\beta,h}\rangle \to \langle 1, \bar{u}\rangle = \|\bar{u}\|_\mathcal{M}.$$

As a consequence, the norm constraint in (5.42) is inactive at $\bar{u}_{\beta,h}$ from which we infer its optimality for $(\mathcal{P}_{\beta,h})$. As the subsequence as well as the accumulation point were chosen arbitrary the statement on the existence of discrete optimal designs, their uniform boundedness and their convergence follows.

The necessary and sufficient condition on the discrete optimal gradient $\nabla\psi_h(\bar{u}_\beta)$ as well as its representation are derived as in the continuous case. □

Proposition 5.29. *For every discrete optimal design $\bar{u}_{\beta,h}$ the measure $\Lambda_h\bar{u}_{\beta,h} \in \mathcal{M}^+(\Omega_o) \cap \mathcal{M}_h$ is also optimal.*

Proof. For an arbitrary discrete optimal design $\bar{u}_{\beta,h}$, $h > 0$ we have

$$\Psi(\mathcal{I}_h(\bar{u}_{\beta,h})) + \beta\|\bar{u}_{\beta,h}\|_\mathcal{M} \geq \Psi(\mathcal{I}_h(\Lambda_h\bar{u}_{\beta,h})) + \beta\|\Lambda\bar{u}_{\beta,h}\|_\mathcal{M},$$

due to the monotonicity of Ψ and Proposition 5.27. The statement follows. □

Remark 5.7. A straightforward combination of the previous results immediately yields

$$F_h(u) \geq F_h(\Lambda_h u) \quad \forall u \in \mathcal{M}^+(\Omega_o). \tag{5.43}$$

At first sight it might seem strange that any optimal design measure is outperformed by a sparse one supported in the grid nodes given that the parameter space $L^2(\Omega)$ is infinite dimensional. Therefore it is worthwhile noting that the parameter space is implicitly approximated due to the finite dimensionality of the discrete state space Y_h. More in detail defining the $L^2(\Omega)$ complement of the kernel as $Q_h = \ker \partial S^h[\hat{q}]^\top$ we can decompose

$$L^2(\Omega) = Q_h \otimes Q_h^\top, \quad \delta q = (\mathrm{Id} - P_{Q_h})\delta q + P_{Q_h}\delta q,$$

for all δq in $L^2(\Omega)$. Here P_{Q_h} denotes the orthogonal $L^2(\Omega)$ projection onto Q_h. Thus we conclude

$$\partial S^h[\hat{q}]\delta q = \partial S^h[\hat{q}]P_{Q_h}\delta q, \quad G_h^x \in Q_h, \quad \mathcal{I}_h(u) \in \mathrm{SHS}(Q_h, Q_h),$$

for all $\delta q \in L^2(\Omega)$, $x \in \Omega_o$ and $u \in \mathcal{M}(\Omega_o)$. Especially this implies $\dim(\mathrm{Im}\,\mathcal{I}_h) \leq \dim(Q_h)$. In this light, Proposition 5.27 and (5.43) can be interpreted as a stronger version of Theorem 3.20.

We illustrate the implicit discretization of the parameter space for several examples. Here we emphasize that some of the discretized PDEs considered in the following correspond to parameter-to-state operators S which are not differentiable in $L^2(\Omega)$ but only with respect to a stronger topology. In particular, the corresponding linearized operator $\partial S[\hat{q}]$ is not continuous on $L^2(\Omega)$. However after discretizing the model the discrete operator $\partial S^h[\hat{q}]$ can be extended to a linear and continuous operator between $L^2(\Omega)$ and $\mathcal{C}(\Omega_o)$. We include these examples for the purpose of highlighting the different outcomes of the implicit discretization on the parameter space and the dependence of Q_h on the underlying PDE. The spatial domain $\Omega \subset \mathbb{R}^d$, $d \leq 3$, is assumed to be a bounded convex domain.

Example 5.6. *Let us first consider the finite element discretization of Example 5.4. Here, given a direction $\delta q \in L^2(\Omega)$ and $\hat{q} \in L^2(\Omega)$ the discrete sensitivity $\delta y^h = \partial S^h[\hat{q}]\delta q = S^h[\delta q] \in V_h \cap H_0^1(\Omega)$ is given as the unique element in $V_h \cap H_0^1(\Omega)$ fulfilling*

$$\int_\Omega \nabla \delta y_h \cdot \nabla \varphi_h \mathrm{d}x = \int_\Omega \delta q \varphi_h \mathrm{d}x, \quad \forall \varphi_h \in V_h \cap H_0^1.$$

We characterize the kernel of $\partial S^h[\hat{q}]$ as

$$\ker \partial S^h[\hat{q}] = \left\{ \delta q \in L^2(\Omega) \mid \int_\Omega \delta q \varphi_h \mathrm{d}x = 0, \quad \forall \varphi_h \in V_h \cap H_0^1(\Omega) \right\} = (V_h \cap H_0^1(\Omega))^\top.$$

Thus we conclude $Q_h = V_h \cap H_0^1(\Omega)$.

Example 5.7. *In the following example, the unknown parameter q enters in the transportation term of an elliptic PDE. Given $\hat{q} \in W^{1,\infty}(\Omega)$, $\partial_{x_1}\hat{q} = 0$ a.e. in Ω, the associated discrete state $y_h = S^h[\hat{q}] \in V_h \cap H_0^1(\Omega)$ is given by the unique solution to*

$$\int_\Omega [\nabla y_h \cdot \nabla \varphi_h + \hat{q}\varphi \partial_{x_1} y_h - f\varphi_h] \,\mathrm{d}x = 0 \quad \forall \varphi_h \in V_h \cap H_(^1\Omega),$$

for some known source term $f \in L^2(\Omega)$. The discrete sensitivity $\delta y_h = \partial S^h[\hat{q}]\delta q \in V_h \cap H_0^1(\Omega)$ in a direction $\delta q \in L^2(\Omega)$ fulfills

$$\int_\Omega [\nabla \delta y_h \cdot \nabla \varphi_h + \hat{q}\varphi_h \partial_{x_1} \delta y_h] \mathrm{d}x = -\int_\Omega \delta q \varphi_h \partial_{x_1} y_h \mathrm{d}x, \quad \forall \varphi_h \in V_h \cap H_0^1(\Omega).$$

We obtain

$$\begin{aligned}\ker \partial S^h[\hat{q}] &= \left\{ \delta q \in L^2(\Omega) \mid \int_\Omega \delta q \partial_{x_1} y_h \varphi_h \mathrm{d}x = 0, \quad \forall \varphi_h \in V_h \cap H_0^1(\Omega) \right\} \\ &= \left\{ q \in V_h^{\not{c}} \mid \exists \varphi_h \in V_h \cap H_0^1(\Omega) \colon q = \partial_{x_1} y_h \varphi_h \right\}^\top \\ &= Q_h^\top,\end{aligned}$$

where the space of piecewise linear and not necessarily continuous function on $\mathcal{T}_h$ is given by

$$V_h^{\not{c}} = \left\{ \varphi_h \in L^\infty(\Omega) \mid \varphi_{h_{|T}} \in P_1 \quad \forall T \in \mathcal{T}_h \right\}.$$

To prove this note that

$$\delta q \in Q_h^\top \Leftrightarrow \int_\Omega \delta q \partial_{x_1} y_h \varphi_h \mathrm{d}x = 0, \quad \forall \varphi_h \in V_h \cap H_0^1(\Omega) \Leftrightarrow \delta q \in \ker \partial S^h[\hat{q}].$$

Example 5.8. *Last we deal with the identification of a distributed diffusion coefficient, see Example 5.4. For $\hat{q} \in \mathcal{C}^{0,1}(\Omega)$, the discrete state $y_h = S^h[\hat{q}]$ fulfills*

$$\int_\Omega [\exp(\hat{q}) \nabla y_h \cdot \nabla \varphi_h - f \varphi_h] \, \mathrm{d}x = 0 \quad \forall \varphi_h \in V_h \cap H_0^1(\Omega).$$

The sensitivity equation for $\delta y_h = \partial S^h[\hat{q}]\delta q$, $\delta q \in L^2(\Omega)$, is derived as

$$\int_\Omega \exp(\hat{q}) \nabla \delta y_h \cdot \nabla \varphi_h \mathrm{d}x = -\int_\Omega \exp(\hat{q}) \delta q \nabla y_h \cdot \nabla \varphi_h \mathrm{d}x, \quad \forall \varphi_h \in V_h \cap H_0^1(\Omega)$$

Before proceeding we note that $\exp(\hat{q}) > 0$ *on Ω and*

$$(\delta q_1, \delta q_2)_{L^2_{\hat{q}}(\Omega)} = (\exp(\hat{q}) \delta q_1, \delta q_2), \quad \forall \delta q_1, \delta q_2 \in L^2(\Omega),$$

induces an inner product on $L^2(\Omega)$. The induced norm is obviously equivalent to the canonical norm on $L^2(\Omega)$. The kernel of the discrete solution operator $\partial S^h[\hat{q}]$ is now given as

$$\begin{aligned}\ker \partial S^h[\hat{q}] &= \left\{ \delta q \in L^2(\Omega) \mid \int_\Omega \exp(\hat{q}) \delta q \nabla y_h \cdot \nabla \varphi_h \mathrm{d}x = 0, \quad \forall \varphi_h \in V_h \cap H_0^1(\Omega) \right\} \\ &= \left\{ q \in V_h^0 \mid \exists \varphi_h \in V_h \cap H_0^1(\Omega) \colon q = \nabla y_h \cdot \nabla \varphi_h \right\}^{\top_{\hat{q}}} \\ &= Q_h^{\top_{\hat{q}}},\end{aligned}$$

where the orthogonal complement is formed with respect to the $L^2_{\hat{q}}(\Omega)$ inner product. Here we define the space of piecewisse constant finite element functions on $\mathcal{T}_h$ as

$$V_h^0 = \left\{ \varphi_h \in L^\infty(\Omega) \mid \varphi_{h_{|T}} \in P_0 \quad T \in \mathcal{T}_h \right\}.$$

This can be proven analogously to the previous example.

In the following theorem error estimates for the objective functional values are provided.

Theorem 5.30. *Let a sequence of discrete optimal designs $\{\bar{u}_{\beta,h}\}_{h>0}$ with $\bar{u}_{\beta,h} \rightharpoonup^* \bar{u}_\beta$ be given. For $h \le h_0$ small enough there holds*

$$|F_h(\bar{u}_{\beta,h}) - F(\bar{u}_\beta)| \le c\gamma(h), \tag{5.44}$$

for some $c > 0$ independent of h.

Proof. By optimality of $\bar{u}_\beta$ and $\bar{u}_{\beta,h}$ we have

$$F_h(\bar{u}_{\beta,h}) - F(\bar{u}_{\beta,h}) \le F_h(\bar{u}_{\beta,h}) - F(\bar{u}_\beta) \le F_h(\bar{u}_\beta) - F(\bar{u}_\beta).$$

Note that $F_h(u) - F(u) = \psi_h(u) - \psi(u)$ for $u \in \mathcal{M}^+(\Omega_o)$. Thus we obtain

$$|F_h(\bar{u}_{\beta,h}) - F(\bar{u}_\beta)| \le \max\{|\psi_h(\bar{u}_{\beta,h}) - \psi(\bar{u}_{\beta,h})|, |\psi_h(\bar{u}_\beta) - \psi(\bar{u}_\beta)|\}. \tag{5.45}$$

We proceed by Taylor expansion to obtain

$$\begin{aligned}
|\psi_h(\bar{u}_\beta) - \psi(\bar{u}_\beta)| &= |\operatorname{Tr}_{L^2(\Omega)}(\nabla\Psi(\mathcal{I}_{\zeta_h^1}(\bar{u}_\beta))(\mathcal{I}_h(\bar{u}_{\beta,h}) - \mathcal{I}(\bar{u}_\beta))| \\
&\le \|\nabla\Psi(\mathcal{I}_{\zeta_h^1}(\bar{u}_\beta))\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))}\|\mathcal{I}_h(\bar{u}_\beta) - \mathcal{I}(\bar{u}_\beta)\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))} \\
&\le \|\nabla\Psi(\mathcal{I}_{\zeta_h^1}(\bar{u}_\beta))\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))}\|\bar{u}_\beta\|_{\mathcal{M}}\gamma(h),
\end{aligned}$$

where $\mathcal{I}_{\zeta_h^1}(\bar{u}_\beta) = \mathcal{I}(\bar{u}_\beta) + \zeta_h^1(\mathcal{I}_h(\bar{u}_\beta) - \mathcal{I}(\bar{u}_\beta))$ for some $\zeta_h^1 \in (0,1)$ depending on $h \le h_0$. Analogously given $\bar{u}_{\beta,h}$ we get

$$\begin{aligned}
|\psi_h(\bar{u}_{\beta,h}) - \psi(\bar{u}_{\beta,h})| &= |\operatorname{Tr}_{L^2(\Omega)}(\nabla\Psi(\mathcal{I}_{\zeta_h^2}(\bar{u}_{\beta,h}))(\mathcal{I}_h(\bar{u}_{\beta,h}) - \mathcal{I}(\bar{u}_{\beta,h}))| \\
&\le \|\nabla\Psi(\mathcal{I}_{\zeta_h^2}(\bar{u}_{\beta,h}))\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))}\|\mathcal{I}_h(\bar{u}_{\beta,h}) - \mathcal{I}(\bar{u}_{\beta,h})\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))} \\
&\le \|\nabla\Psi(\mathcal{I}_{\zeta_h^2}(\bar{u}_{\beta,h}))\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))}\|\bar{u}_{\beta,h}\|_{\mathcal{M}}\gamma(h),
\end{aligned}$$

with $\mathcal{I}_{\zeta_h^2}(\bar{u}_{\beta,h}) = \mathcal{I}(\bar{u}_{\beta,h}) + \zeta_h^2(\mathcal{I}_h(\bar{u}_{\beta,h}) - \mathcal{I}(\bar{u}_{\beta,h}))$ for some $\zeta_h^2 \in (0,1)$ again depending on h. Observe that there holds

$$\|\mathcal{I}(\bar{u}_\beta) - \mathcal{I}_{\zeta_h^1}(\bar{u}_\beta)\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))} \le \|\mathcal{I}_h(\bar{u}_\beta) - \mathcal{I}(\bar{u}_\beta)|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))} \le c\gamma(h)\|\bar{u}_\beta\|_{\mathcal{M}},$$

as well as

$$\begin{aligned}
&\|\mathcal{I}(\bar{u}_\beta) - \mathcal{I}_{\zeta_h^2}(\bar{u}_{\beta,h})\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))} \\
&\quad\le \|\mathcal{I}(\bar{u}_\beta) - \mathcal{I}(\bar{u}_{\beta,h})\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))} + \|\mathcal{I}(\bar{u}_{\beta,h}) - \mathcal{I}_h(\bar{u}_{\beta,h})\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))} \\
&\quad\le \|\mathcal{I}(\bar{u}_\beta) - \mathcal{I}(\bar{u}_{\beta,h})\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))} + c\gamma(h)\|\bar{u}_{\beta,h}\|_{\mathcal{M}}.
\end{aligned}$$

From the strong convergence of $\{\mathcal{I}_h\}_{h>0}$, the weak* convergence of $\{\bar{u}_{\beta,h}\}_{h>0}$ and the uniform boundedness of $\{\|\bar{u}_{\beta,h}\|_{\mathcal{M}}\}_{h>0}$ we conclude

$$\|\nabla\Psi(\mathcal{I}_{\zeta_h^1}(\bar{u}_\beta)) - \nabla\Psi(\mathcal{I}(\bar{u}_\beta))\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))} + \|\nabla\Psi(\mathcal{I}_{\zeta_h^2}(\bar{u}_{\beta,h})) - \nabla\Psi(\mathcal{I}(\bar{u}_\beta))\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))} \to 0,$$

as h tends to 0. Thus we further estimate (5.45) yielding

$$|F_h(\bar{u}_{\beta,h}) - F(\bar{u}_\beta)| \le c\|\nabla\Psi(\mathcal{I}(\bar{u}_\beta))\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))}\gamma(h).$$

for some $c > 0$ independent of $h > 0$. □

A statistical interpretation of variational parameter discretization

Discretizing the underlying equation but not the parameter space can be interpreted as a variational discretization approach to the Bayesian inverse and the sensor placement problem. We close this section by shedding some light on the statistical consequences of the finite element discretization on the inverse problem. To this end let $q_0\colon D \to L^2(\Omega)$ denote the Gaussian random field distributed according to the prior distribution $\mu_0 = \mathcal{N}(\hat{q}, \mathcal{I}_0^{-1})$. Again, we abbreviate $S^h[\hat{q}](x) = (S^h[\hat{q}](x_1), \dots, S^h[\hat{q}](x_N))^\top$. Consider the discrete (linearized) inverse problem

$$\text{find } q \in L^2(\Omega)\colon \quad S^h[\hat{q}](x) + X_h(q - \hat{q}) = \mathbf{y}_d \quad \text{where} \quad (X_h q)_i = \partial S^h[\hat{q}](x_i),$$

for all $i = 1, \dots, N$ and a given vector of measurements $\mathbf{y}_d \in \mathbb{R}^N$. As in the continuous case its solution is given in terms of the posterior distribution $\mu_{\text{post}}^{h,\mathbf{y}_d}$. This probability measure is a Gaussian whose mean and covariance operator are defined by

$$q_{post}^{h,\mathbf{y}_d} = \hat{q} + \mathcal{C}_{post}^h (X_h^* \Sigma^{-1} (\mathbf{y}_d - S^h[\hat{q}](x))), \quad \mathcal{C}_{post}^h = (X_h^* \Sigma^{-1} X_h + \mathcal{I}_0)^{-1}.$$

Recall that the covariance operator of a Gaussian measure $\mu = \mathcal{N}(q_\mu, T_\mu)$ allows to quantify the uncertainty of the associated random field $q^\mu\colon D \to L^2(\Omega)$ along given directions in the parameter space. More precisely, we observed in Proposition 5.1 that there holds

$$\mathbb{E}^\mu[(\delta q, q^\mu - q_\mu)] = \int_{L^2(\Omega)} (\delta q, q - q_\mu)^2_{L^2(\Omega)} \, \mathrm{d}\mu(q) = (\delta q, T_\mu \delta q)_{L^2(\Omega)}$$

for all $\delta q \in L^2(\Omega)$. As a consequence, the differences

$$\int_{L^2(\Omega)} (\delta q, q - q_0)^2 \, \mathrm{d}\mu_0(q) - \int_{L^2(\Omega)} (\delta q, q - q_{post}^{h,\mathbf{y}_d})^2 \, \mathrm{d}\mu_{\text{post}}^{h,\mathbf{y}_d}(q) = (\delta q, (\mathcal{I}_0^{-1} - \mathcal{C}_{\text{post}}^h)\delta q)_{L^2(\Omega)} \geq 0,$$

for all $\delta q \in L^2(\Omega)$, can be interpreted as a measure of directional uncertainty reduction that was achieved through incorporating the knowledge provided by the measurements $\mathbf{y}_d$.

To illustrate this fact let us consider the prior-preconditioned Fisher information operator

$$\mathcal{I}_0^{-1/2} X_h^* \Sigma^{-1} X_h \mathcal{I}_0^{-1/2} \in \mathcal{L}(L^2(\Omega), L^2(\Omega)).$$

This operator represents the information provided by the mathematical model through the measurement setup filtered by the prior. Obviously it is positive, self-adjoint and has at most rank N. If it is not equal to zero, it admits a strictly positive eigenvalue $\varrho > 0$ with associated eigenfunction $\vartheta \in L^2(\Omega)$. Due to the definition of the preconditioned Fisher information operator there holds $\vartheta \in \mathcal{H}$. Furthermore, since $\mathcal{H} = \operatorname{Im} \mathcal{I}_0^{1/2}$, there exists an element $v \in L^2(\Omega)$ with $\mathcal{I}_0^{-1/2} v = \vartheta$. Let us quantify the uncertainty reduction provided by the measurements in this direction. We readily calculate

$$\begin{aligned} (v, \mathcal{C}_{\text{post}} v)_{L^2(\Omega)} &= (v, (X_h^* \Sigma^{-1} X_h + \mathcal{I}_0)^{-1} v)_{L^2(\Omega)} \\ &= (\mathcal{I}_0^{-1/2} v, (\mathcal{I}_0^{-1/2} X_h^* \Sigma^{-1} X_h \mathcal{I}_0^{-1/2} + \mathrm{Id})^{-1} \mathcal{I}_0^{-1/2} v)_{L^2(\Omega)} \\ &= \frac{1}{\varrho + 1} (v, \mathcal{I}_0^{-1} v)_{L^2(\Omega)}. \end{aligned}$$

Thus the amount of uncertainty reduction that occurs in this direction is given by

$$(v, \mathcal{I}_0^{-1} v)_{L^2(\Omega)} - (v, \mathcal{C}_{\text{post}} v)_{L^2(\Omega)} = \frac{\varrho}{1+\varrho} (v, \mathcal{I}_0^{-1} v)_{L^2(\Omega)}.$$

In particular, if ϱ is large, i.e. $\varrho/(\varrho+1) \approx 1$, the measurements $\mathbf{y}_d$ are highly informative in this direction of the parameter space.

In contrast, if $\delta q \in Q_h^\top$ we readily obtain

$$(\delta q, \mathcal{C}^h_{\text{post}} \delta q)_{L^2(\Omega)} = (\delta q, \mathcal{I}_0^{-1} \delta q)_{L^2(\Omega)}$$

and consequently

$$\int_{L^2(\Omega)} (\delta q, q - q_0)^2 \, \mathrm{d}\mu_0(q) - \int_{L^2(\Omega)} (\delta q, q - q_{post}^{h,\mathbf{y}_d})^2 \, \mathrm{d}\mu_{\text{post}}^{h,\mathbf{y}_d}(q) = (\delta q, (\mathcal{I}_0^{-1} - \mathcal{I}_0^{-1}) \delta q)_{L^2(\Omega)} = 0.$$

Thus, in such directions, no uncertainty reduction can be achieved by solving the FE discretized inverse problem. The variability of the posterior distribution on $Q_h^\top$ is completely characterized by the prior.

Let us put this observation into the context of sparse sensor placement problems. While $(\mathcal{P}_{\beta,h})$ is still formulated as a sensor placement for the Gaussian random field q in $L^2(\Omega)$, any measurement design $u \in \mathcal{M}^+(\Omega_o)$ only provides information for the parameter on the finite dimensional space Q_h. For example, the A-optimal design criterion might be rewritten as

$$\operatorname{Tr}_{L^2(\Omega)}((\mathcal{I}_h(u) + \mathcal{I}_0)^{-1}) = \operatorname{Tr}_{Q_h}((\mathcal{I}_h(u) + \mathcal{I}_0)^{-1}) + \operatorname{Tr}_{Q_h^\top}(\mathcal{I}_0^{-1}),$$

where the second term is independent of the design measure u and cannot be reduced through optimizing the measurement setup. In particular the associated sensor placement problem is equivalent to solving

$$\min_{u \in \mathcal{M}^+(\Omega_o)} [\operatorname{Tr}_{Q_h}((\mathcal{I}_h(u) + \mathcal{I}_0)^{-1}) + \beta \|u\|_{\mathcal{M}}]. \tag{5.46}$$

To give a rigorous statistical interpretation of this observation let us consider the Gaussian random variable obtained through projecting the random parts of q onto Q_h

$$P_{Q_h} q \colon D \to L^2(\Omega), \quad \omega \mapsto \hat{q} + P_{Q_h}(q(\omega) - \hat{q}).$$

Clearly, $P_{Q_h} q$ is an affine linear transformation of a Gaussian random field and thus again Gaussian. Its prior distribution is given by $\mu_0^{Q_h} = \mathcal{N}(q_0^{Q_h}, P_{Q_h} \mathcal{I}_0^{-1} P_{Q_h})$ with mean $q_0^{Q_h} = \hat{q}$. Arguing similarly to our discussion on sparse goal-oriented design criteria, we obtain its posterior distribution $\mu_{\text{post}}^{Q_h,\mathbf{y}_d}$ as

$$\mu_{\text{post}}^{Q_h,\mathbf{y}_d} = \mathcal{N}(q_{post}^{Q_h,\mathbf{y}_d}, \mathcal{C}_{\text{post}}^{Q_h}) \quad \text{where} \quad q_{post}^{Q_h,\mathbf{y}_d} = \hat{q} + P_{Q_h}(q_{post}^{h,\mathbf{y}_d} - \hat{q}),\ \mathcal{C}_{\text{post}}^{Q_h} = P_{Q_h} \mathcal{C}^h_{\text{post}} P_{Q_h}.$$

Now, calculating the averaged pointwise posterior variance of the projected random field yields

$$\begin{aligned} \int_{L^2(\Omega)} \|q - q_{post}^{Q_h,\mathbf{y}_d}\|^2_{L^2(\Omega)} \, \mathrm{d}\mu_{\text{post}}^{Q_h,\mathbf{y}_d} &= \operatorname{Tr}_{L^2(\Omega)}(P_{Q_h} \mathcal{C}^h_{\text{post}} P_{Q_h}) \\ &= \operatorname{Tr}_{Q_h}(P_{Q_h} \mathcal{C}^h_{\text{post}} P_{Q_h}) + \operatorname{Tr}_{Q_h^\top}(P_{Q_h} \mathcal{C}^h_{\text{post}} P_{Q_h}) \\ &= \operatorname{Tr}_{Q_h}(\mathcal{C}^h_{\text{post}}). \end{aligned}$$

This admits an intuitive interpretation. Due to the discretization of the PDE model we cannot obtain any additional knowledge on the unknown parameter on the complement of the finite dimensional subspace Q_h through measurements of the (linearized) state variable and the FE discretized model. Thus, intuitively, optimal sensors should be chosen in order to, at least, provide as much certainty on the projection of the parameter onto Q_h. This intuition is captured by the semi-discretized problem $(\mathcal{P}_{\beta,h})$ in a mathematically rigorous way.

5.2.2 Spectral discretization

In general a straightforward algorithmic solution of $(\mathcal{P}_{\beta,h})$ is infeasible since the discrete parameter space Q_h is usually high-dimensional, see also the discussion in Section 5.3.2. To reduce the dimension of the parameter space we propose to replace the space $L^2(\Omega)$ with a subspace spanned by finitely many eigenvectors of $\mathcal{I}_0^{-1}$. We first discuss this approach for the continuous model i.e. with no additional FE discretization.

Denote by $\{\lambda_i\}_{i\in\mathbb{N}}$ the eigenvalues of $\mathcal{I}_0^{-1}$ ordered by decreasing magnitude and by $\{\phi_i\}_{i\in\mathbb{N}}$ the associated eigenfunctions. Given a truncation parameter $n \in \mathbb{N}$ we define the linear subspace $V_n \subset L^2(\Omega)$ as

$$V_n = \operatorname{span}\{\phi_1, \dots, \phi_n\} = \left\{ \sum_{i=1}^{n} v_i \phi_i \mid v_i \in \mathbb{R},\ i = 1, \dots, n \right\}.$$

The orthogonal $L^2(\Omega)$ projection onto V_n will be denoted by

$$P_n \colon L^2(\Omega) \to V_n, \quad v \mapsto \sum_{i=1}^{n} (v, \phi_i)_{L^2(\Omega)} \phi_i.$$

Discretization of $(\mathcal{P}_\beta)$

The spectral discretized sensor placement problem is now defined by

$$\min_{u\in\mathcal{M}^+(\Omega_o)} F^n(u) = [\psi^n(u) + \beta\|u\|_{\mathcal{M}}], \tag{$\mathcal{P}_\beta^n$}$$

where $\psi^n(u) = \Psi(P_n\mathcal{I}(u)P_n)$. We make the following additional regularity assumption on the optimal design criterion Ψ.

Assumption 5.7. For every $n \in \mathbb{N}$ large enough there holds

A.4.5 Given $M_0 > 0$ and $B \in \operatorname{Pos}(L^2(\Omega), L^2(\Omega))$, $\|B\|_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))} \leq M_0$, we have

$$0 \leq \Psi(P_nBP_n) - \Psi(B) \leq c_{M_0} \sum_{i=n+1}^{\infty} \lambda_i, ,$$

for some constant $c_{M_0} > 0$ which may dependent on M_0 but not on B.

We verify this assumption for A and D optimality.

Example 5.9. *For the A-optimal design criterion $\Psi_A(B) = \operatorname{Tr}_{L^2(\Omega)}(\mathcal{C}_{post}(B))$ we have*

$$\Psi_A(P_nBP_n) = \operatorname{Tr}_{L^2(\Omega)}(\mathcal{C}_{post}(P_nBP_n)) = \sum_{i=1}^{n}[(\phi_i, \mathcal{C}_{post}(P_nBP_n)\phi_i)_{L^2(\Omega)}] + \sum_{i=n+1}^{\infty} \lambda_i,$$

for $n \in \mathbb{N}$ and $B \in \operatorname{Pos}(L^2(\Omega), L^2(\Omega))$. Let us verify Assumption 5.7 in this case. Let an operator $B \in \operatorname{Pos}(L^2(\Omega), L^2(\Omega))$ with $\|B\|_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))} \leq M_0$ be given. Applying Taylor's expansion we obtain

$$\begin{aligned}0 \leq \Psi_A(P_nBP_n) - \Psi_A(B) &= \operatorname{Tr}_{L^2(\Omega)}(\mathcal{C}_{post}(B_\zeta)^2(B - P_nBP_n))\\ &= \operatorname{Tr}_{L^2(\Omega)}((\mathcal{C}_{post}(B_\zeta)^2 - P_n\mathcal{C}_{post}(B_\zeta)^2P_n)B),\end{aligned}$$

where $B_\zeta = B + \zeta(P_nBP_n - B) \in \operatorname{Pos}(L^2(\Omega), L^2(\Omega))$ for some $\zeta \in (0,1)$. Further estimates reveal

$$\begin{aligned}&\operatorname{Tr}_{L^2(\Omega)}((\mathcal{C}_{post}(B_\zeta)^2 - P_n\mathcal{C}_{post}(B_\zeta)^2P_n)B)\\ &\qquad\leq \|B\|_{\mathcal{L}(L^2(\Omega),L^2(\Omega))} \operatorname{Tr}_{L^2(\Omega)}(\mathcal{C}_{post}(B_\zeta)^2 - P_n\mathcal{C}_{post}(B_\zeta)^2P_n)\\ &\qquad= \|B\|_{\mathcal{L}(L^2(\Omega),L^2(\Omega))} \sum_{i=n+1}^{\infty} \|\mathcal{C}_{post}(B_\zeta)\phi_i\|^2_{L^2(\Omega)}.\end{aligned}$$

Recalling that $\|\mathcal{C}_{post}(B_\zeta)\|_{\mathcal{L}(\mathcal{H}^*,\mathcal{H})} \leq 1$, see (5.9), $\|\phi_i\|^2_{\mathcal{H}^*} = \lambda_i$, and that $\operatorname{SHS}(L^2(\Omega), L^2(\Omega))$ embeds continuously into $\mathcal{L}(L^2(\Omega), L^2(\Omega))$ we conclude

$$\Psi_A(P_nBP_n) - \Psi_A(B) \leq M_0 \sum_{n+1}^{\infty} \lambda_i.$$

Example 5.10. *Concerning the D-optimal design criterion we have*

$$\Psi_D(P_nBP_n) = -\log(\operatorname{Det}(\mathcal{I}_0^{-1/2}P_nBP_n\mathcal{I}_0^{-1/2} + \operatorname{Id})),$$

for all $B \in \operatorname{Pos}(L^2(\Omega), L^2(\Omega))$ and $n \in \mathbb{N}$.

Let B with $\|B\|_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))} \leq M_0$ be given. Proceeding as for the A-optimal design criterion we obtain

$$0 \leq \Psi_D(P_nBP_n) - \Psi_D(B) = \operatorname{Tr}_{L^2(\Omega)}(\mathcal{C}_{post}(B_\zeta)(B - P_nBP_n)),$$

with $B_\zeta = B + \zeta(P_nBP_n - B) \in \operatorname{Pos}(L^2(\Omega), L^2(\Omega))$. We further estimate

$$\operatorname{Tr}_{L^2(\Omega)}(\mathcal{C}_{post}(B_\zeta)(B - P_nBP_n)) \leq \operatorname{Tr}_{L^2(\Omega)}(\mathcal{C}_{post}(B_\zeta) - P_n\mathcal{C}_{post}(B_\zeta)P_n)\|B\|_{\mathcal{L}(L^2(\Omega),L^2(\Omega))}.$$

After calculating the trace we end up with

$$\begin{aligned}\Psi_D(P_nBP_n) - \Psi_D(B) &\leq \operatorname{Tr}_{L^2(\Omega)}(\mathcal{C}_{post}(B_\zeta) - P_n\mathcal{C}_{post}(B_\zeta)P_n)\|B\|_{\mathcal{L}(L^2(\Omega),L^2(\Omega))}\\ &= \|B\|_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))} \sum_{i=n+1}^{\infty} (\phi_i\mathcal{C}_{post}(B_\zeta)\phi_i)_{L^2(\Omega)}\\ &\leq M_0 \sum_{i=n+1}^{\infty} \lambda_i.\end{aligned}$$

Existence of a minimizer to $(\mathcal{P}^n_\beta)$ can be concluded from the monotonicity of the design criterion. Due to the discretization of the parameter space the number of optimal sensors can additionally be bounded in dependence of the truncation parameter.

Theorem 5.31. *Let $n \in \mathbb{N}$ large enough be given. Then there exists at least one minimizer $\bar{u}^n_\beta \in \mathcal{M}^+(\Omega_o)$ to $(\mathcal{P}^n_\beta)$ with $\#\operatorname{supp} \bar{u}^n_\beta \leq n(n+1)/2$. Furthermore a measure $\bar{u}^n_\beta \in \mathcal{M}^+(\Omega_o)$ is optimal for $(\mathcal{P}^n_\beta)$ if and only if*

$$-\psi^n(\bar{u}^n_\beta) \leq \beta, \quad \operatorname{supp} \bar{u}^n_\beta \subset \left\{ x \in \Omega_o \mid -\psi^n(\bar{u}^n_\beta)(x) = \beta \right\},$$

where the continuous function $-\psi^n(\bar{u}^n_\beta)$ is given by

$$\begin{aligned} -\nabla\psi^n(\bar{u}^n_\beta)(x) &= -(P_n G^x, \nabla\Psi(P_n \mathcal{I}_h(\bar{u}^n_\beta)P_n), P_n G^x)_{L^2(\Omega)} \\ &= \|(-\nabla\Psi(P_n\mathcal{I}(\bar{u}^n_\beta)P_n)^{1/2} P_n G^x\|^2_{L^2(\Omega)}. \end{aligned}$$

Proof. Let $n \in \mathbb{N}$ large enough be given. Since Ψ is monotone in the sense of Assumption 5.4 we conclude

$$F(u) = \Psi(\mathcal{I}(u)) + \beta\|u\|_{\mathcal{M}} \leq \Psi(P_n\mathcal{I}(u)P_n) + \beta\|u\|_{\mathcal{M}} = F^n(u) \quad \forall u \in \mathcal{M}^+(\Omega_o).$$

In particular, this implies radial unboundedness of F^n. Existence of a minimizer and the conditions on the gradient can now be concluded as in the continuous case. The result on the existence of a minimizer $\bar{u}^n_\beta$ with $\#\operatorname{supp} \bar{u}^n_\beta \leq n(n+1)/2$ follows by a straightforward adaption of Theorem 3.20 noting that $\dim(\operatorname{Im} P_n\mathcal{I}P_n) \leq n(n+1)/2$. □

The following proposition addresses convergence of minimizers to the spectral discretized problem $(\mathcal{P}^n_\beta)$ as $n \to \infty$.

Proposition 5.32. *For $n \in \mathbb{N}$ large enough let $\bar{u}^n_\beta \in \mathcal{M}^+(\Omega_o)$ denote a minimizer of $(\mathcal{P}^n_\beta)$. Then the sequence $\{\bar{u}^n_\beta\}_{n\in\mathbb{N}}$ admits at least one weak* accumulation point $\bar{u}_\beta$ as $n \to \infty$ and every such point is a minimizer of $(\mathcal{P}_\beta)$.*

Proof. Let such a sequence be given. Exploiting the monotonicity of Ψ we conclude

$$F(\bar{u}^n_\beta) \leq F^n(\bar{u}^n_\beta) \leq F^n(u) \leq F(u) - c_u \sum_{n+1}^{\infty} \lambda_i,$$

for some arbitrary but fixed $u \in \mathcal{M}^+(\Omega_o)$, a constant $c_u > 0$ only depending on u, and all $n \in \mathbb{N}$ large enough. As a consequence $\{F(\bar{u}^n_\beta))\}_{n\in\mathbb{N}}$ and thus $\{\|\bar{u}^n_\beta\|_{\mathcal{M}}\}_{n\in\mathbb{N}}$ is bounded. We extract a weak* convergent subsequence $\{\bar{u}^n_\beta\}_{n\in\mathbb{N}}$, denoted by the same symbol, with limit $\bar{u}_\beta \in \mathcal{M}^+(\Omega_o)$. Denote by $\bar{u}$ a minimizer of $(\mathcal{P}_\beta)$. Then there holds

$$F(\bar{u}_\beta) = \lim_{n\to\infty} F(\bar{u}^n_\beta) \leq \lim_{n\to\infty} F^n(\bar{u}^n_\beta) \leq \lim_{n\to\infty} F^n(\bar{u}) = F(\bar{u}).$$

This implies optimality of $\bar{u}_\beta$ for $(\mathcal{P}_\beta)$. Since the weak* accumulation point was chosen arbitrary the statement follows. □

As a straightforward consequence of Assumption 5.7 the approximation error in the optimal objective function values is bounded by the tail sum of the eigenvalues corresponding to the neglected orthonormal basis functions.

Theorem 5.33. *For $n \in \mathbb{N}$ large enough denote by $\bar{u}_\beta^n \in \mathcal{M}^+(\Omega_o)$ an arbitrary optimal solution to $(\mathcal{P}_\beta^n)$. Given a sequence $\{\bar{u}_\beta^n\}_{n\in\mathbb{N}}$ with $\bar{u}_\beta^n \rightharpoonup^* \bar{u}_\beta$ as $n \to \infty$ there holds*

$$|F^n(\bar{u}_\beta^n) - F(\bar{u}_\beta)| \le c \sum_{i=n+1}^{\infty} \lambda_i,$$

for all $n \in \mathbb{N}$ large enough and some constant $c > 0$ independent of $n \in \mathbb{N}$.

Proof. Let such a sequence be given. From its weak* convergence we get the existence of a constant $M_0 > 0$ with $\|\mathcal{I}(\bar{u}_\beta^n)\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))} \le M_0$ for all $n \in \mathbb{N}$. By comparing objective function values we conclude

$$|F^n(\bar{u}_\beta^n) - F(\bar{u}_\beta)| \le \max\{\psi^n(\bar{u}_\beta^n) - \psi(\bar{u}_\beta^n), \psi^n(\bar{u}_\beta) - \psi(\bar{u}_\beta)\} \le c_{M_0} \sum_{n+1}^{\infty} \lambda_i,$$

using Assumption 5.7. This yields the statement. □

A statistical interpretation of spectral discretization

As for the finite element discretized problem $(\mathcal{P}_{\beta,h})$ we give a statistical interpretation of the spectral discretization approach. To this end we denote denote by $q\colon D \to L^2(\Omega)$ the Gaussian random field distributed according to the prior distribution $\mu_0 = \mathcal{N}(\hat{q}, \mathcal{I}_0^{-1})$. As a first step we recall the definition of the space V_n as the span of n eigenfunctions $\{\phi_i\}_{i=1}^n$ corresponding to the largest eigenvalues of the prior covariance operator. Moreover, if ϕ_i is an eigenfunction to $\lambda_i > 0$, we calculate

$$\mathbb{E}^{\mu_0}[(\phi_i, q - \hat{q})^2_{L^2(\Omega)}] = (\phi_i, \mathcal{I}_0^{-1}\phi_i)_{L^2(\Omega)} = \lambda_i,$$

Thus the magnitude of the eigenvalue quantifies the amount of prior uncertainty in the direction of the associated eigenfunctions. In order to measure the amount of information that we obtain on the unknown parameter by solving the inverse problem we compute these directional variances also for the posterior distribution.

To this end, denote by $q^{\mathbf{y}_d}\colon D \to L^2(\Omega)$ the Gaussian random field distributed according to the posterior distribution $\mu_{\text{post}}^{\mathbf{y}_d} = \mathcal{N}(q_{\text{post}}^{\mathbf{y}_d}, \mathcal{C}_{\text{post}})$. Furthermore we recall the definition of the posterior covariance operator as $\mathcal{C}_{\text{post}} = (X^*\Sigma^{-1}X + \mathcal{I}_0)^{-1}$. In the following we compute the difference between directional prior and posterior variances

$$\mathbb{E}^{\mu_0}[(\phi, q - \hat{q})^2_{L^2(\Omega)}] - \mathbb{E}^{\mu_{\text{post}}^{\mathbf{y}_d}}[(\phi, q^{\mathbf{y}_d} - q_{\text{post}}^{\mathbf{y}_d})^2_{L^2(\Omega)}] = (\phi, (\mathcal{I}_0^{-1} - \mathcal{C}_{post})\phi)_{L^2(\Omega)} \tag{5.47}$$

for each eigenfunction ϕ of $\mathcal{I}_0^{-1}$ with associated eigenvalue $\lambda > 0$. Note that this difference is always non-negative and $(\phi, \mathcal{I}_0^{-1}\phi)_{L^2(\Omega)} = \lambda$. Thus if this quantity is approximately λ then the posterior uncertainty of the random field in this direction is small. This means that a significant amount of information on the unknown parameter along this direction is obtained by solving the inverse problem. We readily calculate

$$\begin{aligned}(\phi, \mathcal{C}_{\text{post}}\phi)_{L^2(\Omega)} &= \lambda(\phi, (\mathcal{I}_0^{-1/2}X_n^*\Sigma^{-1}X_n\mathcal{I}_0^{-1/2} + \mathrm{Id})^{-1}\phi)_{L^2(\Omega)} \\ &= \lambda - \lambda^2\,|(\Sigma + X\mathcal{I}_0^{-1}X^*)^{-1/2}X\phi|^2_{\mathbb{R}^N}.\end{aligned}$$

Here we used the Sherman-Morrison-Woodbury formula for the inverse and $\mathcal{I}_0^{-1/2}\phi = \lambda^{1/2}\phi$. As a consequence, we obtain

$$\mathbb{E}^{\mu_0}[(\phi, q-\hat{q})^2_{L^2(\Omega)}] - \mathbb{E}^{\mu_{\text{post}}^{\mathbf{y}_d}}[(\phi, q^{\mathbf{y}_d} - q^{\mathbf{y}_d}_{\text{post}})^2_{L^2(\Omega)}] = \lambda^2 |(\Sigma + X\mathcal{I}_0^{-1}X^*)^{-1/2}X\phi|^2_{\mathbb{R}^N}. \tag{5.48}$$

Let us interpret this statement. In order to do so we introduce the random variable y_d modeling our prior believes on the distribution of the measurements as

$$y_d\colon D \to \mathbb{R}^N, \quad \omega \mapsto S[\hat{q}](x) + X(q(\omega)-\hat{q}) + \varepsilon(\omega).$$

Due to the statistical independence of the measurement noise and the prior distribution of the parameter we conclude that y_d is a Gaussian random vector distributed according to

$$\mu_{y_d} = \mathcal{N}(S[\hat{q}](x), X\mathcal{I}_0^{-1}X^* + \Sigma).$$

In particular, there holds

$$\mu_{y_d}(O) = \int_{\mathbb{R}^N} \exp\left(-\frac{1}{2}|(\Sigma + X\mathcal{I}_0^{-1}X^*)^{-1/2}(y - S[\hat{q}](x))|^2_{\mathbb{R}^N}\right)\, \mathrm{d}y \quad \forall O \in \mathcal{B}(\mathbb{R}^N).$$

Given a concrete vector of measurements $\mathbf{y}_d \in \mathbb{R}^N$, the weighted euclidean inner product

$$|(\Sigma + X\mathcal{I}_0^{-1}X^*)^{-1/2}(\mathbf{y}_d - S[\hat{q}](x))|^2_{\mathbb{R}^N} = ((\mathbf{y}_d - S[\hat{q}](x)), (\Sigma + X\mathcal{I}_0^{-1}X^*)^{-1}(\mathbf{y}_d - S[\hat{q}](x)))_{\mathbb{R}^N} \tag{5.49}$$

is often referred to as the *information* on the distribution of the measurements provided by the vector $\mathbf{y}_d$. Intuitively, this terminology can be justified as follows: Suppose that the vector $\mathbf{y}_d \in \mathbb{R}^N$ corresponds to real -life measurements taken in an experiment. Now, based on these observations, we re-evaluate our prior believes on the distribution of the measurements y_d. If the misfit term in (5.49) is small, i.e. $\mathbf{y}_d$ is close to the mean of y_d, the obtained measurements back up our prior believes but no significant new information is obtained. However, if this term is large, the vector $\mathbf{y}_d$ is far away from $S[\hat{q}](x)$. This either implies that the observed measurements are an outlier or that the prior distribution of the measurements is incorrect and needs to be adjusted. In this case the vector $\mathbf{y}_d$ can be seen as highly informative.

Second, we recall the Karhunen-Loève expansion of the Gaussian random field q as

$$q(\omega, x) = \hat{q}(x) + \sum_{i=1}^{\infty} \sqrt{\lambda_i}\zeta_i(\omega)\phi_i(x),$$

for $\mathbb{P}$-a.e. $\omega \in D$ and almost every $x \in \Omega$. Here the coefficient functions are given by a family of i.i.d scalar-valued random variables $\{\zeta_i\}_{i\in\mathbb{N}}$ with $\zeta_1\colon D \to \mathbb{R}$, $\zeta_1 \sim \mathcal{N}(0,1)$.

Now we split up the right hand side of (5.48) as

$$\lambda^2|(\Sigma + X\mathcal{I}_0^{-1}X^*)^{-1/2}X\phi|^2_{\mathbb{R}^N} = \lambda \cdot \lambda|(\Sigma + X\mathcal{I}_0^{-1}X^*)^{-1/2}X\phi|^2_{\mathbb{R}^N}. \tag{5.50}$$

The first factor, $\lambda > 0$, describes our prior knowledge on the random field. Obviously, significant uncertainty reduction can only be achieved if there is substantial prior uncertainty for the random field along the direction of ϕ. Additionally, reducing uncertainty is only possible if the obtained measurements are sensitive with respect to changes in the random field along this direction. This is captured by the second factor in (5.50). To make these arguments rigorous we consider random perturbations of the measurement mean in the direction of ϕ. More precisely, we define

$$\delta y_d \colon D \to \mathbb{R}^N, \quad \delta y(\omega) = S[\hat{q}](x) + XP_\phi(q(\omega) - \hat{q}) = S[\hat{q}](x) + (\phi, q(\omega) - \hat{q})X\phi,$$

where P_ϕ denotes the $L^2(\Omega)$ orthogonal projection on the one-dimensional subspace spanned by ϕ. Now, we compute the information, in the sense of (5.49), for every realization of δy_d and average over its distribution. This yields

$$\begin{aligned}\int_D |(\Sigma + X\mathcal{I}_0^{-1}X^*)^{-1/2}(\delta y_d(\omega) - S[\hat{q}](x))|^2_{\mathbb{R}^N} \, \mathrm{d}\mathbb{P}(\omega) \\ &= \int_{L^2(\Omega)} (\phi, q - \hat{q})^2_{\mathbb{R}^N} |(\Sigma + X\mathcal{I}_0^{-1}X^*)^{-1/2}X\phi|^2_{\mathbb{R}^N} \, \mathrm{d}\mu_0(q) \\ &= \lambda|(\Sigma + X\mathcal{I}_0^{-1}X^*)^{-1/2}X\phi|^2_{\mathbb{R}^N}\end{aligned}$$

using the formula for the expectation of a quadratic form. In particular, if $\phi \in \operatorname{Ker} X$ there holds

$$\lambda|(\Sigma + X\mathcal{I}_0^{-1}X^*)^{-1/2}X\phi|^2_{\mathbb{R}^N} = 0,$$

i.e. no uncertainty reduction can be achieved in this direction by solving the inverse problem.

Similar to the FE discretized case, we now interpret the spectral discretization approach as a variational discretization of the inverse problem. To clarify this connection let us recall that the set of eigenfunctions $\{\phi_i\}_{i\in\mathbb{N}}$ to $\mathcal{I}_0^{-1}$ forms an orthonormal basis of $L^2(\Omega)$. Now, consider the spectral discretized inverse problem

$$\text{find } q \in L^2(\Omega)\colon \quad S[\hat{q}](x) + X_n(q - \hat{q}) = \mathbf{y}_d, \tag{5.51}$$

where the linearized parameter-to-observation operator X is replaced by the reduced model $X_n = XP_n$. This corresponds to a low-rank approximation of the sensitivity operator $\partial S[\hat{q}]$ in the reduced basis $\{\phi_i\}_{i=1}^n$. Again note that, from this perspective, we only discretize the underlying equation but not the parameter space. The solution to this inverse problem in the Bayesian approach is given by the posterior distribution $\mu_{\text{post}}^{n,\mathbf{y}_d}$ which is characterized by its mean and covariance operator

$$q_{post}^{n,\mathbf{y}_d} = \hat{q} + \mathcal{C}_{post}^n(X_n^*\Sigma^{-1}(\mathbf{y}_d - S^h[\hat{q}](x))), \quad \mathcal{C}_{post}^n = (X_n^*\Sigma^{-1}X_n + \mathcal{I}_0)^{-1}.$$

As for the original problem, let us quantify the amount of uncertainty reduction that we achieve through incorporating the information provided by the measurements in this case. To this end denote by $q^{n,\mathbf{y}_d}\colon D \to L^2(\Omega)$ the Gaussian random field distributed according to $\mu^{n,\mathbf{y}_d}$ and let ϕ be an eigenfunction of $\mathcal{I}_0^{-1}$. The corresponding eigenvalue will be denoted by λ. Evaluating the difference between prior and posterior uncertainty now yields

$$\mathbb{E}^{\mu_0}[(\phi, q - \hat{q})^2_{L^2(\Omega)}] - \mathbb{E}^{\mu_{\text{post}}^{n,\mathbf{y}_d}}[(\phi, q^{n,\mathbf{y}_d} - q_{\text{post}}^{n,\mathbf{y}_d})^2_{L^2(\Omega)}] = \lambda^2|(\Sigma + X_n\mathcal{I}_0^{-1}X_n^*)^{-1/2}X_n\phi|^2_{\mathbb{R}^N}.$$

If ϕ corresponds to a neglected eigenvalue, i.e. $\phi \notin V_n$, we readily obtain

$$\mathbb{E}^{\mu_{\text{post}}^{n,\mathbf{y}_d}}[(\phi, q^{n,\mathbf{y}_d} - q_{\text{post}}^{n,\mathbf{y}_d})^2_{L^2(\Omega)}] = (\phi, \mathcal{C}_{\text{post}}^n\phi)_{L^2(\Omega)} = (\phi, \mathcal{I}_0^{-1}\phi)_{L^2(\Omega)} = \lambda.$$

Thus in such directions no uncertainty reduction can be achieved by solving the spectral discretized inverse problem (5.51). However aggregating the directional pointwise variances in these directions we observe that

$$\operatorname{Tr}_{V_n^\top}(\mathcal{C}^n_{\text{post}}) = \operatorname{Tr}_{V_n^\top}(\mathcal{I}_0^{-1}) = \sum_{i=n+1}^{\infty} \lambda_i < \infty, \quad \operatorname{Tr}_{V_n^\top}(\mathcal{C}^n_{\text{post}}) \to 0$$

as $n \to \infty$. From this perspective we may interpret the presented spectral discretization as a restriction of the inverse problem to the subspace V_n spanned by the directions of largest prior uncertainty. On its complement, if $n \in \mathbb{N}$ is large enough, the variability of the random field is already small due to the provided prior knowledge. This reduces the infinite-dimensional inverse problem to a finite dimensional one.

In the remaining directions, uncertainty reduction can still be achieved by solving the inverse problem but only to a smaller extend than in the full model. More precisely, for $\phi \in V_n$ we obtain

$$\begin{aligned}0 \le \mathbb{E}^{\mu^{n,\mathbf{y}_d}_{\text{post}}}[(\phi, q^{n,\mathbf{y}_d} - q^{n,\mathbf{y}_d}_{\text{post}})^2_{L^2(\Omega)}] &- \mathbb{E}^{\mu^{\mathbf{y}_d}_{\text{post}}}[(\phi, q^{\mathbf{y}_d} - q^{n,\mathbf{y}_d}_{\text{post}})^2_{L^2(\Omega)}] \\ &= (\phi, (\mathcal{C}^n_{\text{post}} - \mathcal{C}_{\text{post}})\phi) \le \lambda_i \|\mathcal{C}^n_{\text{post}} - \mathcal{C}_{\text{post}}\|_{\text{HS}(\mathcal{H}^*,\mathcal{H})} \\ &\le \lambda \|P_n X^* \Sigma^{-1} X P_n - X^* \Sigma^{-1} X\|_{\mathcal{L}(\mathcal{H},\mathcal{H}^*)} \\ &\le \lambda \operatorname{Tr}_{L^2(\Omega)}(\mathcal{I}_0^{-1/2}(X^* \Sigma^{-1} X - P_n X^* \Sigma^{-1} X P_n)\mathcal{I}_0^{-1/2}) \\ &\le c\lambda \sum_{i=n+1}^{\infty} \lambda_i |\Sigma^{-1/2} X \phi_i|^2_{\mathbb{R}^N} \le c\lambda \sum_{i=n+1}^{\infty} \lambda_i\end{aligned}$$

for some constant $c > 0$ independent of ϕ. Here we used the Lipschitz continuity of the posterior covariance mapping, see Proposition 5.19, the continuous embedding of $\text{SHS}(\mathcal{H}, \mathcal{H}^*)$ into $\mathcal{L}(\mathcal{H}, \mathcal{H}^*)$ and the definition of the norm on $\text{SHS}(\mathcal{H}, \mathcal{H}^*)$.

In particular these observations suggest that if the sequence of eigenvalues $\{\lambda_i\}_{i\in\mathbb{N}}$ converges fast enough to zero we can restrict the parameter space and thus the inverse problem to a small number of uncertain directions. On this subspace we can achieve uncertainty reduction comparable to that provided by the full model through solving the spectral discretized inverse problem. In contrast, on the complement of this low dimensional space we are already certain about the behavior of the random field due to the small directional prior variances. Thus they can be left out of the problem.

Clearly, these observations also yield implications for the corresponding sensor placement problem $(\mathcal{P}^n_\beta)$. For example, computing the spectral discretized A-optimal design criterion reveals

$$\begin{aligned}\Psi_A(P_n \mathcal{I}(u) P_n) &= \operatorname{Tr}_{V_n}((P_n \mathcal{I}(u) P_n + \mathcal{I}_0)^{-1}) + \operatorname{Tr}_{V_n^\top}((P_n \mathcal{I}(u) P_n + \mathcal{I}_0)^{-1}) \\ &= \operatorname{Tr}_{V_n}((P_n \mathcal{I}(u) P_n + \mathcal{I}_0)^{-1}) + \sum_{i=n+1}^{\infty} \lambda_i.\end{aligned}$$

As a consequence, finding a measurement setup that minimizes the averaged variance of the posterior distribution obtained by the spectral discretized problem boils down to solving

$$\min_{u \in \mathcal{M}^+(\Omega_o)} [\operatorname{Tr}_{V_n}((P_n \mathcal{I}(u) P_n + \mathcal{I}_0)^{-1}) + \beta \|u\|_{\mathcal{M}}]. \tag{5.52}$$

It is straightforward to see that this is equivalent to minimizing the trace of a $n \times n$ matrix:

$$\min_{u \in \mathcal{M}^+(\Omega_o)} [\mathrm{Tr}_{\mathbb{R}^n}((\mathcal{I}^n(u) + \mathcal{I}_0^n)^{-1}) + \beta \|u\|_{\mathcal{M}}], \tag{5.53}$$

Here $\mathcal{I}_0^n \in \mathrm{Sym}(n)$ is a diagonal matrix with $(\mathcal{I}_0^n)_{ii} = 1/\lambda_i$ for $i = 1, \ldots, n$, and the matrix $\mathcal{I}^n(u) \in \mathrm{Sym}(n)$ is given by

$$\mathcal{I}^n(u) = \int_{\Omega_o} \partial S^n[\hat{q}](x) \partial S^n[\hat{q}](x)^\top \mathrm{d}u(x), \quad \partial S^n[\hat{q}](x) = (\partial S[\hat{q}]\phi_1(x), \ldots, \partial S[\hat{q}]\phi_n(x))^\top,$$

for $x \in \Omega_o$ and $u \in \mathcal{M}^+(\Omega_o)$. This problem fits into the general framework presented in the previous chapter. Note that $\partial S[\hat{q}]\phi_i(x) = (G^x, \phi_i)_{L^2(\Omega)}$ for all $i = 1, \ldots, n$. To clarify the connection between these two problems we introduce the mapping

$$\mathbf{P}_n \colon L^2(\Omega) \to \mathbb{R}^n, \quad q \mapsto ((q, \phi_1)_{L^2(\Omega)}, \ldots, (q, \phi_n)_{L^2(\Omega)})^\top.$$

Now we readily obtain

$$\begin{aligned} \mathrm{Tr}_{V_n}((P_n \mathcal{I}(u) P_n + \mathcal{I}_0)^{-1}) &= \sum_{i=1}^n (\mathbf{e}_i, \mathbf{P}_n (P_n \mathcal{I}(u) P_n + \mathcal{I}_0)^{-1} \mathbf{P}_n^* \mathbf{e}_i)_{\mathbb{R}^n} \\ &= \mathrm{Tr}_{\mathbb{R}^n}((\mathbf{P}_n \mathcal{I}(u) \mathbf{P}_n^* + \mathbf{P}_n \mathcal{I}_0 \mathbf{P}_n^*)^{-1}) \\ &= \mathrm{Tr}_{\mathbb{R}^n}((\mathcal{I}^n(u) + \mathcal{I}_0^n)^{-1}). \end{aligned}$$

As for the finite element discretization these two equivalent problems can be interpreted as sensor placement problems to optimally infer on suitable projections of q. Let us first consider the projection of the random parts of q onto the subspace V_n

$$P_{V_n} q \colon D \to L^2(\Omega), \quad \omega \to \hat{q} + P_n(q(\omega) - \hat{q}).$$

Due to the linearity of the projection, $P_{V_n} q$ is a Gaussian random field distributed according to $\mu_0^{V_n}(\hat{q}, P_n \mathcal{I}_0^{-1} P_n)$. Computing its posterior distribution given the measurements and the spectral discretized inverse problem gives

$$\mu_{\mathrm{post}}^{V_n, \mathbf{y}_d} = \mathcal{N}(q_{post}^{V_n, \mathbf{y}_d}, \mathcal{C}_{\mathrm{post}}^{V_n}) \quad \text{where} \quad q_{post}^{V_n, \mathbf{y}_d} = \hat{q} + P_{V_n}(q_{post}^{n, \mathbf{y}_d} - \hat{q}), \; \mathcal{C}_{\mathrm{post}}^{V_n} = P_{V_n} \mathcal{C}_{\mathrm{post}}^n P_{V_n}.$$

Its averaged posterior is readily calculated as

$$\begin{aligned} \int_{L^2(\Omega)} \|q - q_{post}^{V_n, \mathbf{y}_d}\|_{L^2(\Omega)}^2 \, \mathrm{d}\mu_{\mathrm{post}}^{V_n, \mathbf{y}_d} &= \mathrm{Tr}_{L^2(\Omega)}(P_n \mathcal{C}_{\mathrm{post}}^h P_n) \\ &= \mathrm{Tr}_{V_n}(P_n \mathcal{C}_{\mathrm{post}}^n P_n) + \mathrm{Tr}_{V_n^\top}(P_n \mathcal{C}_{\mathrm{post}}^n P_n) \\ &= \mathrm{Tr}_{V_n}(\mathcal{C}_{\mathrm{post}}^n). \end{aligned}$$

Moreover we observe that

$$P_{V_n} q(\omega) = \hat{q} + P_n(q(\omega) - \hat{q}) = \hat{q} + \sum_{i=1}^n \sqrt{\lambda_i} \zeta_i(\omega) \phi_i \quad \text{for} \quad \mathbb{P} - a.e. \; \omega \in D.$$

From this perspective, the minimization problem in (5.52) corresponds to finding a measurement setup in order to optimally infer on the first n terms in the Karhunen-Loève expansion of the

Gaussian random field q by solving($\mathcal{P}_\beta^n$). In the same way, we can motivate the problem in (5.53) by considering the vector-valued random variable assembling the random scalar coefficients of the first n terms appearing in the KL-expansion of q.

To close this section we point out that all of the preceding discussions concerning the spectral discretization inherently rely on an appropriate choice of the prior distribution for the estimation problem at hand. In particular, we base the construction of the subspace V_n on our belief on the directional prior variances of the random field. However the measurements and the mathematical model may provide significant information in the neglected directions. Following (5.50) the expected uncertainty reduction in such directions is dampened by our already strong prior beliefs , i.e. the small variance of the random field, in these directions. Vice versa, the measurements might not be sensitive with respect to an element $\phi \in V_n$ i.e. perturbations of the unknown parameter along such directions only slightly affect the obtained measurements. Consequently no significant uncertainty reduction can be achieved in these directions by solving the inverse problem. This again stresses the dependence of the Bayesian approach in the presented setting on a sophisticated choice of the prior distribution.

5.2.3 Fully-discrete problem

We proceed by combining the two presented discretization approaches to obtain a sensor placement problem which is amenable to the solution by sequential point insertion algorithms. Therefore we replace the parameter space $L^2(\Omega)$ by V_n, $n \in N$, and the continuous solution operator to the sensitivity equation $\partial S[\hat{q}]$ by its discrete counterpart $\partial S^h[\hat{q}]$ for $h \leq h_0$. We end up with the fully discrete problem

$$\min_{u \in \mathcal{M}^+(\Omega)} F_h^n(u) = [\psi_h^n(u) + \beta \|u\|_{\mathcal{M}}], \qquad (\mathcal{P}_{\beta,h}^n)$$

where the reduced functional is given by $\psi_h^n(u) = \Psi(P_n \mathcal{I}_h(u) P_n)$. The following theorem addresses existence of fully discrete optimal designs as well as their convergence behaviour for vanishing mesh-size and $n \to \infty$.

Theorem 5.34. *Let $\beta > 0$, $h \leq h_0$ small enough and $n \in \mathbb{N}$ large enough be given. Then there exists at least one optimal solution $\bar{u}_{\beta,h}^n$ to $(\mathcal{P}_{\beta,h}^n)$. Every optimal design $\bar{u}_{\beta,h}^n$ fulfills*

$$-\nabla \psi_h^n(\bar{u}_{\beta,h}^n) \leq \beta, \quad \operatorname{supp} \bar{u}_{\beta,h}^n \subset \left\{ x \in \Omega_o \middle| -\nabla \psi_h^n(\bar{u}_{\beta,h}^n)(x) = \beta \right\}.$$

Here the discrete gradient is given by

$$\begin{aligned} -\nabla \psi_h^n(\bar{u}_{\beta,h})(x) &= -(P_n G_h^x, \nabla \Psi(P_n \mathcal{I}_h(\bar{u}_{\beta,h}^n) P_n), P_n G_h^x)_{L^2(\Omega)} \\ &= \|(-\nabla \Psi(P_n \mathcal{I}_h(\bar{u}_{\beta,h}) P_n)^{1/2} P_n G_h^x\|_{L^2(\Omega)}^2, \end{aligned}$$

Given a sequence $\{\bar{u}_{\beta,h}^n\}_{h>0,n\in\mathbb{N}}$ of optimal designs there exists at least one subsequence denoted by the same symbol which converges in the weak sense as $h \to 0$ and $n \to \infty$. Every accumulation point $\bar{u}_\beta$ of $\{\bar{u}_{\beta,h}^n\}_{h>0,n\in\mathbb{N}}$ is an optimal solution to $(\mathcal{P}_\beta)$.*

Proof. First let us consider an arbitrary weak* convergent sequence $\{u_{h,n}\}_{h>0,n\in\mathbb{N}} \subset \mathcal{M}^+(\Omega_o)$, $u_{h,n} \rightharpoonup^* \bar{u}$, for some $\bar{u} \in \mathcal{M}^+(\Omega_o)$, as $h \to 0, n \to \infty$. Then there holds

$$\lim_{h \to 0, n \to \infty} [\|\mathcal{I}_h(u_{h,n}) - \mathcal{I}(\bar{u})\|_{HS(L^2(\Omega),L^2(\Omega))} + |\|u_{h,n}\|_{\mathcal{M}} - \|\bar{u}\|_{\mathcal{M}}|] = 0,$$

as well as

$$|\psi_h^n(u_{h,n}) - \psi(\bar{u})| \leq |\Psi(\mathcal{I}_h(u_{h,n})) - \Psi(\mathcal{I}(\bar{u}))| + M \sum_{i=n+1}^{\infty} \lambda_i,$$

for some constant M with $\|u_{h,n}\|_{\mathcal{M}} \leq M$. Hence we conclude $F_h^n(u_{h,n}) \to F(\bar{u})$ for $h \to 0$, $n \to \infty$. We show the existence of at least one discrete optimal design $\bar{u}_{\beta,h}^n \in \mathcal{M}^+(\Omega_o)$. Therefore we proceed along the lines of proof in Theorem 5.28 and consider the auxiliary problem

$$\min_{u \in \mathcal{M}^+(\Omega_o)} F_h^n(u) \quad s.t. \quad \|u\|_{\mathcal{M}} \leq 2M_0, \tag{5.54}$$

for some constant $M_0 > 0$ bounding the set of optimal solutions to $(\mathcal{P}_\beta)$. Given $h > 0$ and $n \in \mathbb{N}$ there exists a minimizer to (5.54) and the sequence $\{\bar{u}_{\beta,h}^n\}_{h>0,n\in\mathbb{N}}$ admits a weak* convergent subsequence with limit point $\bar{u}_\beta$. Denote by $\bar{u} \in \mathcal{M}^+(\Omega_o)$ an arbitrary optimal solution to $(\mathcal{P}_\beta)$. From the previous discussions we conclude

$$F(\bar{u}_\beta) = \lim_{h\to 0, n\to\infty} F_h^n(\bar{u}_{\beta,h}^n) \leq \lim_{h\to 0, n\to\infty} F_h^n(\bar{u}) = F(\bar{u}).$$

Thus $\bar{u}_\beta$ is an optimal solution to $(\mathcal{P}_\beta)$ and $\|\bar{u}_\beta\|_{\mathcal{M}} < M_0$. Due to the weak* convergence we also have $\|\bar{u}_{\beta,h}\|_{\mathcal{M}} < 2M_0$ yielding the optimality of $\bar{u}_{\beta,h}^n$ for $(\mathcal{P}_{\beta,h}^n)$. Since the weak* convergent subsequence as well as $\bar{u}_\beta$ were chosen arbitrary the same holds for every accumulation point.

The necessary (and sufficient) optimality condition on the gradient are derived as in the previous sections. □

Due to the finite element discretization of the sensitivities the existence of an optimal design supported in the grid nodes is concluded. Combining this observation with the finite dimensionality of the parameter space V_n its support size can be further bounded in dependence on n.

Proposition 5.35. *Assume that there exists an optimal solution to $(\mathcal{P}_{\beta,h}^n)$. Then there exists an optimal design $\bar{u}_{\beta,h}^n \in \mathcal{M}^+(\Omega) \cap \mathcal{M}_h$ with $\#\operatorname{supp} \bar{u}_{\beta,h}^n \leq n(n+1)/2$.*

Proof. Let an arbitrary $\delta q \in L^2(\Omega)$ be given and define $\mathcal{I}_h^n(u) = P_n \mathcal{I}_h(u) P_n$ for $u \in \mathcal{M}^+(\Omega_o)$. First we follow Proposition 5.27 to obtain

$$\begin{aligned}(\delta q, \mathcal{I}_h^n(u)\delta q)_{L^2(\Omega)} = (P_n\delta q, \mathcal{I}_h(u)P_n\delta q)_{L^2(\Omega)} &\leq (P_n\delta q, \mathcal{I}_h(\Lambda_h u)P_n\delta q)_{L^2(\Omega)} \\ &= (\delta q, \mathcal{I}_h^n(\Lambda_h u)\delta q)_{L^2(\Omega)}\end{aligned}$$

and $\|\Lambda_h u\|_{\mathcal{M}} \leq \|u\|_{\mathcal{M}}$. Consequently we have $F_h^n(u) \geq F_h^n(u_h)$ for $u_h = \Lambda_h u \in \mathcal{M}^+(\Omega_o) \cap \mathcal{M}_h$. Furthermore, due to the discretization of the parameter space, we readily infer

$$\dim(\operatorname{Im} \mathcal{I}_h^n) \leq n(n+1)/2.$$

By combining the statements of Theorem 3.20 and Proposition 4.7 this yields the existence of a measure $\tilde{u}_h^n \in \mathcal{M}^+(\Omega_o) \cap \mathcal{M}_h$ with

$$\mathcal{I}_h^n(\tilde{u}_h^n) = \mathcal{I}_h^n(u_h), \quad \|\tilde{u}_h^n\|_{\mathcal{M}} \leq \|u_h\|_{\mathcal{M}}, \quad \#\operatorname{supp} \tilde{u}_h^n \leq \frac{n(n+1)}{2}.$$

Since the design measure $u \in \mathcal{M}^+(\Omega_o)$ was chosen arbitrary all considerations especially apply to optimal designs obtained from $(\mathcal{P}_{\beta,h}^n)$. Therefore we conclude the existence of an optimal design $\bar{u}_{\beta,h}^n \in \mathcal{M}^+(\Omega_o) \cap \mathcal{M}_h$ fulfilling $\operatorname{supp} \bar{u}_{\beta,h}^n \leq n(n+1)/2$. □

Error estimates for the objective functional

The rest of this section is devoted to a priori error estimates between the fully discrete problem $(\mathcal{P}^n_{\beta,h})$ and the continuous one. Based on the stability results for $\mathcal{I}_h$ and Assumption 5.7 we conclude the following estimate for the optimal objective function values.

Theorem 5.36. *For $h \leq h_0$ small enough and $n \in \mathbb{N}$ large enough denote by $\bar{u}^n_{\beta,h}$ an optimal solution to $(\mathcal{P}^n_{\beta,h})$. Given a sequence $\{\bar{u}^n_{\beta,h}\}_{h>0,n\in\mathbb{N}}$ with $\bar{u}^n_{\beta,h} \rightharpoonup^* \bar{u}_\beta$ as $h \to 0$, $n \to \infty$ there holds*

$$|F^n_h(\bar{u}^n_{\beta,h}) - F(\bar{u}_\beta)| \leq c\left(\sum_{i=n+1}^{\infty} \lambda_i + \gamma(h)\right), \tag{5.55}$$

for $h > 0$ small enough, $n \in \mathbb{N}$ large enough and some $c > 0$ independent of h and n.

Proof. Let such a sequence be given and denote by $M_0 > 0$ a constant bounding the norm of its elements. Again, comparing objective function values yields

$$|F^n_h(\bar{u}^n_{\beta,h}) - F(\bar{u}_\beta)| \leq \max\{|\psi^n_h(\bar{u}^n_{\beta,h}) - \psi(\bar{u}^n_{\beta,h})|, |\psi^n_h(\bar{u}_\beta) - \psi(\bar{u}_\beta)|\}.$$

Noting that $\{\mathcal{I}_h(\bar{u}^n_{\beta,h})\}_{h>0,n\in\mathbb{N}}$ is uniformly bounded in $\operatorname{SHS}(L^2(\Omega), L^2(\Omega))$ we proceed to

$$\begin{aligned}\max\{|\psi^n_h(\bar{u}^n_{\beta,h}) - \psi(\bar{u}^n_{\beta,h})|,&|\psi^n_h(\bar{u}_\beta) - \psi(\bar{u}_\beta)|\}\\ &\leq c\sum_{i=n+1}^{\infty}\lambda_i + \max\{|\psi_h(\bar{u}^n_{\beta,h}) - \psi(\bar{u}^n_{\beta,h})|, |\psi_h(\bar{u}_\beta) - \psi(\bar{u}_\beta)|\}.\end{aligned}$$

The remaining term on the right-hand side can be estimated along the lines of proof in Theorem 5.30 yielding

$$\max\{|\psi_h(\bar{u}^n_{\beta,h}) - \psi(\bar{u}^n_{\beta,h})|, |\psi_h(\bar{u}_\beta) - \psi(\bar{u}_\beta)|\} \leq c\gamma(h).$$

for some constant $c > 0$ independent of $n \in \mathbb{N}$ and $h > 0$. Combining all previous results yields the statement. □

Error estimates for the Fisher information operator

Finally we provide a priori error estimates for the convergence of the optimal Fisher information in the case of $\Psi = \Psi_A$ and $\Psi = \Psi_D$ respectively. Recalling the results of Proposition 5.12 we derive the following quadratic growth condition.

Lemma 5.37. *Let $\Psi = \Psi_A$. Then the optimal Fisher information $\mathcal{I}(\bar{u}_\beta)$ is the same for every optimal solution $\bar{u}_\beta$ to $(\mathcal{P}_\beta)$. There exist a neighborhood $N(\mathcal{I}(\bar{u}_\beta))$ of $\mathcal{I}(\bar{u}_\beta)$ in $\operatorname{SHS}(L^2(\Omega), L^2(\Omega))$ as well as a constant $\gamma_0 >$ with*

$$\frac{\gamma_0}{4}\|\mathcal{I}_0^{-1}(\mathcal{I}(u) - \mathcal{I}(\bar{u}_\beta))\mathcal{I}_0^{-1/2}\|^2_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))} \leq F(u) - F(\bar{u}_\beta),$$

for all $u \in \mathcal{M}^+(\Omega_o)$ with $\mathcal{I}(u) \in N(\mathcal{I}(\bar{u}_\beta))$.

Proof. Uniqueness of the Fisher information $\mathcal{I}(\bar{u}_\beta)$ follows from the strict convexity of Ψ_A on $\operatorname{Pos}(L^2(\Omega), L^2(\Omega))$. Given a direction $\delta B \in \operatorname{SHS}(L^2(\Omega), L^2(\Omega))$ we calculate

$$\begin{aligned}
&\langle\langle \delta B, \nabla^2 \Psi_A(\mathcal{I}(\bar{u}_\beta))\delta B\rangle\rangle_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))} \\
&\qquad = 2\operatorname{Tr}_{L^2(\Omega)}(\mathcal{C}_{\text{post}}(\mathcal{I}(\bar{u}_\beta))\delta B\mathcal{C}_{\text{post}}(\mathcal{I}(\bar{u}_\beta))\delta B\mathcal{C}_{\text{post}}(\mathcal{I}(\bar{u}_\beta))) \qquad (5.56)\\
&\qquad \geq \gamma_0 \|\mathcal{I}_0^{-1}\delta B \mathcal{I}_0^{-1/2}\|^2_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))},
\end{aligned}$$

for some $\gamma_0 > 0$, see also the proof of Proposition 5.12. We apply Taylor's formula to obtain

$$\begin{aligned}
\Psi(\mathcal{I}(u)) = \Psi(\mathcal{I}(\bar{u}_\beta)) &+ \langle\langle \nabla\Psi(\mathcal{I}(\bar{u}_\beta)), \mathcal{I}(u) - \mathcal{I}(\bar{u}_\beta)\rangle\rangle_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))} \\
&+ \frac{1}{2}\langle\langle \mathcal{I}(u) - \mathcal{I}(\bar{u}_\beta), \nabla^2\Psi(\mathcal{I}(u_\zeta))(\mathcal{I}(u) - \mathcal{I}(\bar{u}_\beta))\rangle\rangle_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))},
\end{aligned}$$

with $u_\zeta = \bar{u}_\beta + \zeta(u - \bar{u}_\beta)$ for some $\zeta \in (0,1)$. Note that we have

$$\|\mathcal{I}(u_\zeta) - \mathcal{I}(\bar{u}_\beta)\|_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))} \leq \|\mathcal{I}(u) - \mathcal{I}(\bar{u}_\beta)\|_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))},$$

for all $u \in \mathcal{M}^+(\Omega)$, $\zeta \in (0,1)$. From the continuity of $\nabla^2\Psi$ at $\mathcal{I}(\bar{u}_\beta)$ as well as

$$\|\mathcal{I}_0^{-1}\delta B\mathcal{I}_0^{-1/2}\|_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))} \leq c\|\delta B\|_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))} \quad \forall \delta B \in \operatorname{SHS}(L^2(\Omega), L^2(\Omega)),$$

we thus conclude the existence of a neighborhood $N(\mathcal{I}(\bar{u}_\beta))$ of $\mathcal{I}(\bar{u}_\beta)$ in $\operatorname{SHS}(L^2(\Omega), L^2(\Omega))$ with

$$\begin{aligned}
&\langle\langle \delta B, \nabla^2\Psi_A(\mathcal{I}(u_\zeta))\delta B\rangle\rangle_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))} \\
&\quad = \langle\langle \delta B, \nabla^2\Psi_A(\mathcal{I}(\bar{u}_\beta))\delta B\rangle\rangle_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))} \\
&\qquad + \langle\langle \delta B, (\nabla^2\Psi_A(\mathcal{I}(u_\zeta)) - \nabla^2\Psi_A(\mathcal{I}(\bar{u}_\beta)))\delta B\rangle\rangle_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))} \\
&\quad \geq \frac{\gamma_0}{2}\|\mathcal{I}_0^{-1}\delta B\mathcal{I}_0^{-1/2}\|^2_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))},
\end{aligned}$$

for all $u_\zeta = \bar{u}_\beta + \zeta(u - \bar{u}_\beta)$, where $\zeta \in [0,1)$ and $u \in \mathcal{M}^+(\Omega_o)$, $\mathcal{I}(u) \in N(\mathcal{I}(\bar{u}_\beta))$. By optimality of $\bar{u}_\beta$ we further have

$$\langle\langle \nabla\Psi(\mathcal{I}(\bar{u}_\beta)), \mathcal{I}(u) - \mathcal{I}(\bar{u}_\beta)\rangle\rangle_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))} + \beta\|u\|_{\mathcal{M}} - \beta\|\bar{u}_\beta\|_{\mathcal{M}} \geq 0,$$

for all $u \in \mathcal{M}^+(\Omega_o)$. Combining the previous statements we arrive at

$$F(u) - F(\bar{u}_\beta) \geq \frac{\gamma_0}{4}\|\mathcal{I}_0^{-1}(\mathcal{I}(u) - \mathcal{I}(\bar{u}_\beta))\mathcal{I}_0^{-1/2}\|^2_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))},$$

for all $u \in \mathcal{M}^+(\Omega_o)$ with $\mathcal{I}(u) \in N(\mathcal{I}(\bar{u}_\beta))$. Thus the statement follows. □

A similar result holds for the D-optimal design criterion.

Lemma 5.38. *Let $\Psi = \Psi_D$ and consider an optimal solution $\bar{u}_\beta$. Then the optimal Fisher information $\mathcal{I}(\bar{u}_\beta)$ is the same for every optimal solution $\bar{u}_\beta$ to $(\mathcal{P}_\beta)$. There exist a neighborhood $N(\mathcal{I}(\bar{u}_\beta))$ of $\mathcal{I}(\bar{u}_\beta)$ in $\operatorname{SHS}(L^2(\Omega), L^2(\Omega))$ as well as a constant $\gamma_0 >$ with*

$$\frac{\gamma_0}{4}\|\mathcal{I}_0^{-1/2}(\mathcal{I}(u) - \mathcal{I}(\bar{u}_\beta))\mathcal{I}_0^{-1/2}\|^2_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))} \leq F(u) - F(\bar{u}_\beta),$$

for all $u \in \mathcal{M}^+(\Omega_o)$ with $\mathcal{I}(u) \in N(\mathcal{I}(\bar{u}_\beta))$.

Proof. The proof follows along the lines of the previous lemma with the sole difference of noting that

$$\begin{aligned}\langle\langle \delta B, \nabla^2 \varPsi_D(\mathcal{I}(\bar{u}_\beta))\delta B\rangle\rangle_{\mathrm{HS}(L^2(\varOmega),L^2(\varOmega))} &= \mathrm{Tr}_{L^2(\varOmega)}(\mathcal{C}_{\mathrm{post}}(\mathcal{I}(\bar{u}_\beta))\delta B\mathcal{C}_{\mathrm{post}}(\mathcal{I}(\bar{u}_\beta))\delta B)\\ &\geq \gamma_0 \|\mathcal{I}_0^{-1/2}\delta B\mathcal{I}_0^{-1/2}\|^2_{\mathrm{HS}(L^2(\varOmega),L^2(\varOmega))},\end{aligned}$$

for some $\gamma_0 > 0$ and all $\delta B \in \mathrm{SHS}(L^2(\varOmega), L^2(\varOmega))$, see also the discussion in the proof of Proposition 5.13. □

The following proposition provides an a priori error estimate for the optimal Fisher information associated to the A-optimal design problem.

Proposition 5.39. *Let $\varPsi = \varPsi_A$ and denote by $\{\bar{u}^n_{\beta,h}\}_{h>0,n\in\mathbb{N}}$ a sequence of optimal solutions to $(\mathcal{P}^n_{\beta,h})$ with $\bar{u}^n_{\beta,h} \rightharpoonup^* \bar{u}_\beta$ as $h \to 0$, $n \to \infty$. Then there holds*

$$\|\mathcal{I}_0^{-1}(P_n\mathcal{I}_h(\bar{u}^n_{\beta,h})P_n - \mathcal{I}(\bar{u}_\beta))\mathcal{I}_0^{-1/2}\|_{\mathrm{HS}(L^2(\varOmega),L^2(\varOmega))} \leq c\sqrt{\sum_{i=n+1}^{\infty}\lambda_i + \gamma(h)},$$

for all $h > 0$ small, $n \in \mathbb{N}$ large enough and some constant $c > 0$ independent of h and n.

Proof. Let such a sequence be given. We first split the error as

$$\begin{aligned}&\|\mathcal{I}_0^{-1}(P_n\mathcal{I}_h(\bar{u}^n_{\beta,h})P_n - \mathcal{I}(\bar{u}_\beta))\mathcal{I}_0^{-1/2}\|_{\mathrm{HS}(L^2(\varOmega),L^2(\varOmega))}\\ &\qquad\leq \|\mathcal{I}_0^{-1}(P_n\mathcal{I}_h(\bar{u}^n_{\beta,h})P_n - \mathcal{I}_h(\bar{u}^n_{\beta,h}))\mathcal{I}_0^{-1/2}\|_{\mathrm{HS}(L^2(\varOmega),L^2(\varOmega))}\\ &\qquad\quad+ \|\mathcal{I}_0^{-1}(\mathcal{I}_h(\bar{u}^n_{\beta,h}) - \mathcal{I}(\bar{u}_\beta))\mathcal{I}_0^{-1/2}\|_{\mathrm{HS}(L^2(\varOmega),L^2(\varOmega))}.\end{aligned}\tag{5.57}$$

The first term on the right hand side of the inequality above is further divided into

$$\begin{aligned}&\|\mathcal{I}_0^{-1}(P_n\mathcal{I}_h(\bar{u}^n_{\beta,h})P_n - \mathcal{I}_h(\bar{u}^n_{\beta,h}))\mathcal{I}_0^{-1/2}\|_{\mathrm{HS}(L^2(\varOmega),L^2(\varOmega))}\\ &\qquad\leq \|\mathcal{I}_0^{-1}(P_n\mathcal{I}_h(\bar{u}^n_{\beta,h})P_n - P_n\mathcal{I}_h(\bar{u}^n_{\beta,h}))\mathcal{I}_0^{-1/2}\|_{\mathrm{HS}(L^2(\varOmega),L^2(\varOmega))}\\ &\qquad\quad+ \|\mathcal{I}_0^{-1}(P_n\mathcal{I}_h(\bar{u}^n_{\beta,h}) - \mathcal{I}_h(\bar{u}^n_{\beta,h}))\mathcal{I}_0^{-1/2}\|_{\mathrm{HS}(L^2(\varOmega),L^2(\varOmega))}\end{aligned}$$

We estimate

$$\begin{aligned}&\|\mathcal{I}_0^{-1}(P_n\mathcal{I}_h(\bar{u}^n_{\beta,h})P_n - P_n\mathcal{I}_h(\bar{u}^n_{\beta,h}))\mathcal{I}_0^{-1/2}\|_{\mathrm{HS}(L^2(\varOmega),L^2(\varOmega))}\\ &\qquad\leq \|\mathcal{I}_0^{-3/2}P_n\mathcal{I}_h(\bar{u}^n_{\beta,h})(P_n - \mathrm{Id})\|_{\mathrm{HS}(L^2(\varOmega),L^2(\varOmega))}\\ &\qquad\leq \|P_n\mathcal{I}_h(\bar{u}^n_{\beta,h})\|_{\mathrm{HS}(L^2(\varOmega),L^2(\varOmega))}\|\mathcal{I}_0^{-3/2}(P_n - \mathrm{Id})\|_{\mathrm{HS}(L^2(\varOmega),L^2(\varOmega))}\end{aligned}$$

as well as

$$\begin{aligned}&\|\mathcal{I}_0^{-1}(P_n\mathcal{I}_h(\bar{u}^n_{\beta,h}) - \mathcal{I}_h(\bar{u}^n_{\beta,h}))\mathcal{I}_0^{-1/2}\|_{\mathrm{HS}(L^2(\varOmega),L^2(\varOmega))}\\ &\qquad\leq \|\mathcal{I}_h(\bar{u}^n_{\beta,h})\|_{\mathrm{HS}(L^2(\varOmega),L^2(\varOmega))}\|\mathcal{I}_0^{-3/2}(P_n - \mathrm{Id})\|_{\mathrm{HS}(L^2(\varOmega),L^2(\varOmega))}.\end{aligned}$$

Moreover, we conclude

$$\|P_n\mathcal{I}_h(\bar{u}^n_{\beta,h})\|_{\mathrm{HS}(L^2(\varOmega),L^2(\varOmega))} \leq \|\mathcal{I}_h(\bar{u}^n_{\beta,h})\|_{\mathrm{HS}(L^2(\varOmega),L^2(\varOmega))} \leq \|\mathcal{I}_h\|_{\mathcal{L}(\mathcal{M}(\varOmega_o),\mathrm{HS}(L^2(\varOmega),L^2(\varOmega)))}\|\bar{u}^n_{\beta,h}\|_{\mathcal{M}}.$$

Thus, the sequence $\{P_n\mathcal{I}_h(\bar{u}^n_{\beta,h})\}_{h>0,n\in\mathbb{N}}$ is uniformly bounded in $\mathrm{HS}(L^2(\Omega),L^2(\Omega))$ due to the strong convergence of $\{\mathcal{I}_h\}_{h>0}$ and the weak* convergence of $\{\bar{u}^n_{\beta,h}\}_{h>0,n\in\mathbb{N}}$. Last, we estimate

$$\|\mathcal{I}_0^{-3/2}(P_n-\mathrm{Id})\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))}=\sqrt{\sum_{i=n+1}^{\infty}\lambda_i^3}\leq\sum_{i=n+1}^{\infty}\lambda_i^{3/2},$$

using Jensen inequality for concave functions. Putting these results together we obtain

$$\|\mathcal{I}_0^{-1}(P_n\mathcal{I}_h(\bar{u}^n_{\beta,h})P_n-P_n\mathcal{I}_h(\bar{u}^n_{\beta,h}))\mathcal{I}_0^{-1/2}\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))}\leq c\sum_{i=n+1}^{\infty}\lambda_i^{3/2}$$

for some constant $c>0$ independent of $h>0$ and $n\in\mathbb{N}$.

We split up the second term on the right hand side in (5.57) as

$$\begin{aligned}&\|\mathcal{I}_0^{-1}(\mathcal{I}_h(\bar{u}^n_{\beta,h})-\mathcal{I}(\bar{u}_\beta))\mathcal{I}_0^{-1/2}\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))} &(5.58)\\&\leq\|\mathcal{I}_0^{-1}(\mathcal{I}_h(\bar{u}^n_{\beta,h})-\mathcal{I}(\bar{u}^n_{\beta,h}))\mathcal{I}_0^{-1/2}\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))}+\|\mathcal{I}_0^{-1}(\mathcal{I}(\bar{u}^n_{\beta,h})-\mathcal{I}(\bar{u}_\beta))\mathcal{I}_0^{-1/2}\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))}\\&\leq c\|(\mathcal{I}_h(\bar{u}^n_{\beta,h})-\mathcal{I}(\bar{u}^n_{\beta,h}))\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))}+\|\mathcal{I}_0^{-1}(\mathcal{I}(\bar{u}^n_{\beta,h})-\mathcal{I}(\bar{u}_\beta))\mathcal{I}_0^{-1/2}\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))}.\end{aligned}$$

Following Proposition 5.26 the first term is estimated by

$$\begin{aligned}\|(\mathcal{I}_h(\bar{u}^n_{\beta,h})-\mathcal{I}(\bar{u}^n_{\beta,h}))\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))}&\leq\|\mathcal{I}_h-\mathcal{I}\|_{\mathcal{L}(\mathcal{M}(\Omega_o),\mathrm{HS}(L^2(\Omega),L^2(\Omega)))}\|\bar{u}_{\beta,h}\|_{\mathcal{M}}\\&\leq c\gamma(h)\|\bar{u}^n_{\beta,h}\|_{\mathcal{M}}.\end{aligned}$$

From the weak* convergence of $\{\bar{u}^n_{\beta,h}\}_{h>0,n\in\mathbb{N}}$ we further deduce $\mathcal{I}(\bar{u}^n_{\beta,h})\to\mathcal{I}(\bar{u}_\beta)$ strongly in $\mathrm{SHS}(L^2(\Omega),L^2(\Omega))$ and thus $\mathcal{I}(\bar{u}^n_{\beta,h})\in N(\mathcal{I}(\bar{u}_\beta))$ for all $h>0$ small and $n\in\mathbb{N}$ large enough. Hence we obtain

$$\|\mathcal{I}_0^{-1}(\mathcal{I}_h(\bar{u}^n_{\beta,h})-\mathcal{I}(\bar{u}_\beta))\mathcal{I}_0^{-1/2}\|^2_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))}\leq F(\bar{u}^n_{\beta,h})-F(\bar{u}_\beta),$$

from Lemma 5.37. We proceed by estimating

$$\begin{aligned}F(\bar{u}^n_{\beta,h})-F(\bar{u}_\beta)&\leq F(\bar{u}^n_{\beta,h})-F^n_h(\bar{u}^n_{\beta,h})+F^n_h(\bar{u}^n_{\beta,h})-F(\bar{u}_\beta)\\&\leq|F(\bar{u}^n_{\beta,h})-F_h(\bar{u}^n_{\beta,h})|+|F^h_n(\bar{u}^n_{\beta,h})-F(\bar{u}_\beta)|.\end{aligned}$$

Here the second inequality follows due to the monotonicity of Ψ. Using the estimates obtained in Theorem 5.36 and its proof we conclude

$$F(\bar{u}^n_{\beta,h})-F(\bar{u}_\beta)\leq c_1\gamma(h)+c_2\sum_{i=n+1}^{\infty}\lambda_i,$$

for some constants $c_1,c_2>0$ independent of $h>0$ and $n\in\mathbb{N}$. Plugging all previous estimates into (5.58) we obtain

$$\|\mathcal{I}_0^{-1}(\mathcal{I}_h(\bar{u}^n_{\beta,h})-\mathcal{I}(\bar{u}_\beta))\mathcal{I}_0^{-1/2}\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))}\leq c\left(\gamma(h)+\sqrt{\gamma(h)+\sum_{i=n+1}^{\infty}\lambda_i}\right),$$

for some constant $c>0$ independent of $h>0$ and $n\in\mathbb{N}$ due to the uniform boundedness of $\{\bar{u}^n_{\beta,h}\}_{h>0,n\in\mathbb{N}}$. Combining all previous results yields the desired statement. □

For the D-optimal design problem a similar statement can be proven by the same arguments albeit with respect to a stronger norm. We omit the proof for the sake of brevity.

Proposition 5.40. *Let $\Psi = \Psi_D$ and denote by $\{\bar{u}^n_{\beta,h}\}_{h>0,n\in\mathbb{N}}$ a sequence of optimal solutions to $(\mathcal{P}^n_{\beta,h})$ with $\bar{u}^n_{\beta,h} \rightharpoonup^* \bar{u}_\beta$ as $h \to 0$, $n \to \infty$. Then there holds*

$$\|\mathcal{I}_0^{-1/2}(P_n\mathcal{I}_h(\bar{u}^n_{\beta,h})P_n - \mathcal{I}(\bar{u}_\beta))\mathcal{I}_0^{-1/2}\|_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))} \leq c\sqrt{\sum_{i=n+1}^{\infty} \lambda_i + \gamma(h)},$$

for all $h > 0$ small, $n \in \mathbb{N}$ large enough and some constant $c > 0$ independent of h and n.

Conceptually the proof of Propositions 5.39 and 5.40 follows the same steps as the corresponding one for the Fisher information matrices derived in Section 4.6.2. However the obtained results highlight a significant difference between sensor placement problems for finite and infinite dimensional parameters. To make this clear observe that we have $\mathcal{I}(\bar{u}_\beta) \in \mathrm{SHS}(L^2(\Omega), L^2(\Omega))$, but, surprisingly, the derived a priori estimates only hold in weighted Hilbert-Schmidt norms involving fractional powers of the compact operator $\mathcal{I}_0^{-1}$. For the D-optimal design criterion, e.g., we obtain convergence rates in the norm on the weaker space $\mathrm{SHS}(\mathcal{H}, \mathcal{H}^*)$ since

$$\|\mathcal{I}_0^{-1/2}\delta B\mathcal{I}_0^{-1/2}\|^2_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))} = \sum_{i=1}^{\infty} \|\delta B\mathcal{I}_0^{-1/2}\phi_i\|^2_{\mathcal{H}^*} = \|\delta B\|^2_{\mathrm{HS}(\mathcal{H},\mathcal{H}^*)}.$$

This stems back to the fact that while Ψ is two times continuously differentiable with respect to to the norm on $\mathrm{SHS}(L^2(\Omega), L^2(\Omega))$, coercivity of its second derivative at $\mathcal{I}(\bar{u}_\beta)$ is only given in a weaker norm. In sensor placement problems for a finite dimensional parameter this phenomenon does not occur since all norms on the symmetric matrices are equivalent. This can be interpreted as an instance of the well-known *two norm discrepancy* which arises frequently in infinite dimensional optimization problems, see e.g. [65]. Moreover we emphasize that the choice of the weaker norm depends on the optimal design criterion.

We close this part of the thesis by elaborating on the limitations of the results derived above. One of the standing assumptions throughout this section is the availability of the eigenvalues and associated eigenfunctions corresponding to the a priori covariance operator. While for some choices of $\mathcal{I}_0^{-1}$ and Ω analytic expressions are available, this is in general not the case. In these situations we have to resort to discrete approximations of its first n eigenpairs.

Obviously, this introduces an additional approximation error to the problem. Additionally, e.g. if the discrete sensitivities and eigenpairs are obtained on the same spatial mesh, this leads to a coupling between the finite element and the spectral discretization error. This is due to the fact that while we might expect optimal convergence rates for the eigenfunctions in $L^2(\Omega)$, the constants in the necessary stability estimates usually depend on the associated, continuous, eigenvalue, see e.g. [39,156]. However, if the discrete eigenpairs are determined on a sufficiently fine grid, different from the one used to approximate the PDE, we might assume that the overall error is dominated by the FE and spectral approximation error.

Finally we have to make a critical remark on the proposed full discretization scheme. To this end, for better illustration, we consider again the A-optimal design problem. In order to arrive at the fully discrete problem $(\mathcal{P}^n_{\beta,h})$ we can proceed along two paths. We may first replace the parameter space $L^2(\Omega)$ by the truncated space V_n. Proceeding in this direction we obtain $(\mathcal{P}^n_\beta)$

which fits into the framework discussed in Chapter 4. Subsequently we discretize the state and sensitivity equations according to Section 4.6. Changing this order we first arrive at the semi-discrete problem $(\mathcal{P}_{\beta,h})$. Following the discussions in Section 5.2.1 this reduces the A-optimal design problem to minimizing the trace of the posterior covariance operator on the implicitly discretized space Q_h. The additional spectral discretization now amounts to replacing the space Q_h by V_n. From this perspective this step can be interpreted as a non-conforming approximation of Q_h since $V_n \not\subset Q_h$ in general. Obviously this introduces an additional error depending on how well elements in Q_h are approximated through the truncated space V_n. This also indicates that it may be appropriate to consider a full discretizations of $(\mathcal{P}_{\beta,h})$ based, e.g. on further approximations of Q_h and $\partial S^h[\hat{q}]$. We leave this for future research.

Nevertheless the proposed discretization scheme seems reasonable in many situations. If the state y depends nonlinearly on the parameter q we may adopt a sequential viewpoint on optimal sensor placement. That is to say we alternate between the estimation of the unknown parameter and the determination of a new measurement setup based on a linearization of the model around the current point estimate. Especially in the first iterations of this process the linearization points may be far from the true value of the parameter and the linearized models are only of limited utility. In this case it seems appropriate to place sensors in order to reduce the uncertainty that stems from the prior believes. From the concluding remarks of Section 5.2.2 we recall that the directions of highest uncertainty with respect to the prior span the space V_n. It is also worthwhile to note that the FE discretized parameter space Q_h may depend on $\hat{q}$. Thus, in general, it needs to be determined or approximated in every iteration of the sequential procedure. In contrast the prior knowledge $\mathcal{I}_0^{-1}$ and thus the vectors spanning up V_n are independent of the linearized model. As a consequence they may be pre-computed once at the beginning and, if possible, stored for further usage. Altogether, we point out that the discussions in this section do not claim any completeness and should merely be seen as a first attempt to a rigorous discretization concept of sparse sensor placement problems respecting both the infinite dimensional nature of the parameter as well as the possible continuity of the observational set Ω_o.

5.3 Optimization aspects

In this section we briefly cover the algorithmic treatment of $(\mathcal{P}_\beta)$ and extend the Primal-Dual-Active-Point method presented in Section 4.4 to optimal sensor placement problems for infinite dimensional parameters. Furthermore we comment on their practical realization for the A-optimal design criterion.

5.3.1 Algorithmic treatment

For an efficient numerical solution of $(\mathcal{P}_\beta)$ we again exploit the, at least expected, sparse structure of optimal measurement designs and consider algorithms based on the sequential placement of single measurement sensors. To this end, given a sparse initial design measure $u^1 \in \mathcal{M}^+(\Omega_o)$, we recall the definition of the associated sublevel set of F as

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

Since F is radially unbounded there exists a constant $M_0 > 0$ bounding the norm of elements in E_{u_1}. For convenience of the reader the Primal-Dual-Active-Point strategy is now again summarized in Algorithm 7. To monitor its convergence we define the primal-dual gap of the k-th iterate u^k as

$$\Phi(u^k) = \sup_{v \in \mathcal{M}^+(\Omega_o), \|v\|_{\mathcal{M}} \leq M_0} [\langle \nabla\psi(u), u - v\rangle + \beta\|u\|_{\mathcal{M}} - \beta\|v\|_{\mathcal{M}})] = M_0(\beta + \min_{x\in\Omega_o} \nabla\psi(u^k)).$$

As in the previous chapter this quantity provides an upper bound on the error in the objective functional

$$\Phi(u^k) \geq F(u^k) - F(\bar{u}_\beta) \geq 0 \quad \forall u^k \in E_{u_1}.$$

Furthermore there holds $\Phi(\bar{u}_\beta) = 0$ if and only if $\bar{u}_\beta \in \mathcal{M}^+(\Omega_o)$ is a minimizer of $(\mathcal{P}_\beta)$. Given an ordered set of finitely many distinct points $\mathcal{A} = \{x_1, \dots, x_N\} \subset \Omega_o$, $N = \#\mathcal{A}$, we define the parametrization by

Algorithm 7 Primal-Dual-Active-Point strategy for $(\mathcal{P}_\beta)$

```
while $\Phi(u^k) \geq \mathrm{TOL}$ do
    1. Calculate $\nabla\psi_k = \nabla\psi(u^k)$. Determine $\hat{x}^k \in \arg\min_{x\in\Omega_o} \nabla\psi_k(x)$.
    2. Set $\mathcal{A}_k = \mathrm{supp}(u^k) \cup \{\, \hat{x}^k \,\}$, compute a solution to $\mathbf{u}^{k+1}$ of (5.59) for $\mathcal{A} = \mathcal{A}_k$, and set $u^{k+1} = \boldsymbol{u}_\mathcal{A}(\mathbf{u}^{k+1})$.
end while
```

$$\boldsymbol{u}_\mathcal{A} \colon \mathbb{R}^{\#\mathcal{A}} \to \mathcal{M}(\Omega), \quad \mathbf{u} \mapsto \sum_{x_i \in \mathcal{A}} \mathbf{u}_i \delta_{x_i}.$$

As in the finite dimensional setting of the previous chapter we compute a global minimizer of $(\mathcal{P}_\beta)$ by alternating between choosing a new sensor location $\hat{x}^k$ fulfilling

$$\hat{x}^k \in \arg\min_{x\in\Omega_o} \nabla\psi(u^k)(x)$$

and solving the coefficient optimization problem

$$\mathbf{u}^{k+1} \in \arg\min_{\mathbf{u}\in\mathbb{R}_+^{\#\mathcal{A}}} F(\boldsymbol{u}_\mathcal{A}(\mathbf{u})), \tag{5.59}$$

for the choice of $\mathcal{A} = \mathrm{supp}\, u^k \cup \{\hat{x}^k\}$. The new iterate u^{k+1} is then obtained as $u^{k+1} = \boldsymbol{u}(\mathbf{u}^{k+1})$. Note that this definition also ensures that the iterates are pruned after each iteration i.e. all Dirac delta functions with zero coefficient are removed from the iterate.

We can interpret Algorithm 7 as a special instance of an accelerated generalized conditional gradient method. Consequently we derive worst case convergence rates for the objective function values of the generated iterates following Theorems 6.29 and 6.37 applied to the special case of positive measures.

Proposition 5.41. *Let $u^1 \in \mathcal{M}^+(\Omega_o)$ be given. Then $\nabla\psi$ is Lipschitz continuous on the associated sublevel set E_{u^1}, i.e. there exists a constant $L_{u^1} > 0$ with*

$$\sup_{\substack{u_1,\, u_2 \in E_{u^1} \\ u_1 \neq u_2}} \frac{\|\nabla\psi(u_1) - \nabla\psi(u_2)\|_{\mathcal{C}}}{\|u_1 - u_2\|_{\mathcal{M}}} \leq L_{u^1}.$$

Proof. The proof follows the one for Proposition 4.14 noting that the set

$$\mathcal{I}(E_{u^1}) = \{\, \mathcal{I}(u) \mid u \in E_{u^1} \,\} \subset \operatorname{Pos}(L^2(\Omega), L^2(\Omega)),$$

is a compact subset of $\operatorname{SHS}(L^2(\Omega), L^2(\Omega))$ due to the weak* sequential compactness of E_{u^1} and the weak*-to-strong continuity of $\mathcal{I}$. □

Theorem 5.42. *Assume that the sequence $\{u^k\}_{k\in\mathbb{N}}$ is generated using Algorithm 7. Then $\{u^k\}_{k\in\mathbb{N}}$ is a minimizing sequence for F. Furthermore it admits a weak* convergent subsequence denoted by the same symbol. Every weak* accumulation point $\bar{u}_\beta$ of $\{u^k\}_{k\in\mathbb{N}}$ is an optimal solution to $(\mathcal{P}_\beta)$. There holds*

$$r_F(u^k) \leq \frac{r_F(u^1)}{1+q(k-1)}, \quad q = \alpha \min\left\{ \frac{c_1}{L_{u^1}(2M_0)^2},\, 1 \right\}. \tag{5.60}$$

Here, L_{u^1} is the Lipschitz-constant of $\nabla\psi$ on E_{u^1} and $c_1 = 2\gamma(1-\alpha)r_F(u^1)$ for some arbitrary but fixed $\gamma \in (0,1)$, $\alpha \in (1/2, 1)$.

Remark 5.8. We emphasize that Algorithm 7 can be readily applied for the solution of $(\mathcal{P}_{\beta,h})$, $(\mathcal{P}^n_\beta)$ and $(\mathcal{P}^n_{\beta,h})$. Since the design criterion Ψ is not discretized in any of these problems the results of Theorem 5.42 remain valid with an appropriate adaption of the appearing constants. Moreover, applied to $(\mathcal{P}^n_\beta)$ or $(\mathcal{P}^n_{\beta,h})$, the method can be modified to ensure $\#\operatorname{supp} u^k \leq n(n+1)/2$ for all $k \in \mathbb{N}$, see Proposition 4.16. For $(\mathcal{P}_{\beta,h})$ and $(\mathcal{P}^n_{\beta,h})$ the search for the minimizer $\hat{x}^k \in \Omega_o$ in step 2. can be restricted to $\mathcal{N}_h \cap \Omega_o$, see Proposition 5.43 .

We stress that, up to now, we have not been able to improve on the sublinear convergence rate for the Primal-Dual-Active-Point method applied to the continuous problem $(\mathcal{P}_\beta)$. One particular reason for this shortcoming lies in the aforementioned two norm discrepancy. In particular, the standard examples of A- and D-optimality already show that we cannot expect quadratic growth conditions of the form

$$\|\mathcal{I}(u) - \mathcal{I}(\bar{u}_\beta)\|^2_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))} \leq F(u) - F(\bar{u}_\beta) \quad \forall u \in \mathcal{M}^+(\Omega_o),\ \mathcal{I}(u) \in N(\mathcal{I}(\bar{u}_\beta)),$$

to hold in a neighborhood $N(\mathcal{I}(\bar{u}_\beta))$ of the optimal Fisher information since the Hessian of Ψ is in general not coercive with respect to the $L^2(\Omega)$ Hilbert-Schmidt norm. However, following Lemma 5.37 and Lemma 5.38, similar results can be obtained by replacing the Hilbert-Schmidt norm with a weaker norm depending on the concrete choice of the optimal design criterion. This situation is not yet covered by the convergence results derived in Chapter 6 but poses an interesting question for future research. In particular, improved convergence results for the continuous problem $(\mathcal{P}_\beta)$ would suggest the uniform boundedness of the constants appearing in the upcoming improved convergence results for the spectral discretized problem with respect to the truncation parameter.

The remainder of this section is concerned with improved convergence results for the adaptation of Algorithm 7 to the semi-discretized problems and the fully discrete one, respectively.

We start by proving the finite termination of Algorithm 7 when applied to the finite element discretized problems $(\mathcal{P}_{\beta,h})$ and $(\mathcal{P}^n_{\beta,h})$. To improve readability we provide the statement for the first one. All arguments carry over to $(\mathcal{P}^n_{\beta,h})$ in a straightforward way. For the rest of this chapter we define $\mathcal{N}^o_h = \mathcal{N}_h \cap \Omega_o$.

In order to apply the Primal-Dual-Active-Point strategy to $(\mathcal{P}_{\beta,h})$ we first have to discuss the computation of the new sensor location $\hat{x}^k \in \Omega_o$ fulfilling

$$\hat{x}^k \in \operatorname*{arg\,min}_{x \in \Omega_o} \nabla\psi_h(u^k). \tag{5.61}$$

At first sight, this is a challenging problem in itself since the discretized gradient is neither an element of V_h nor differentiable and convex. In Proposition 5.29 we have already proven that $(\mathcal{P}_{\beta,h})$ admits at least one optimal solution $\bar{u}_{\beta,h}$ supported in the nodes of the finite element triangulation. In view of this result it is tempting to circumvent the global minimization of $\nabla\psi_h(u^k)$ by restricting the search for the new Dirac delta position to the grid nodes. However, it is unclear whether the resulting sequential point insertion method still corresponds to a Primal-Dual-Active-Point method on the discretized problem. In the following proposition we give a positive answer to this question by proving

$$\operatorname*{arg\,min}_{x \in \mathcal{N}_h^o} \nabla\psi_h(u^k) \subset \operatorname*{arg\,min}_{x \in \Omega_o} \nabla\psi_h(u^k).$$

This implies that (5.61) boils down to sorting the values of the discretized gradient in the grid nodes. As an immediate consequence we further conclude the (grid-dependent) finite termination of the Primal-Dual-Active-Point strategy when applied to the FE discretized sensor placement problems.

Proposition 5.43. *Consider Algorithm 7 applied to $(\mathcal{P}_{\beta,h})$ and assume that $u^1 \in \mathcal{M}^+(\Omega_o) \cap \mathcal{M}_h$. Then the new sensor location $\hat{x}^k \in \Omega_o$ can be chosen from $\mathcal{N}_h^o$ for all $k \in \mathbb{N}$. Denote by $\{u^k\}_{k\in\mathbb{N}}$ the sequence generated by Algorithm 7 with $\hat{x}^k \in \mathcal{N}_h^o$ for all $k \in \mathbb{N}$. Then there exists $k \in \mathbb{N}$ with $u^{k+1} = u^k$, i.e. the algorithm terminates after finitely many steps.*

Proof. Let us first prove that there holds

$$\min_{x \in \mathcal{N}_h^o} \nabla\psi_h(u)(x) = \min_{x \in \Omega_o} \nabla\psi_h(u)(x) \quad \forall u \in \mathcal{M}^+(\Omega_o).$$

Given $u \in \mathcal{M}^+(\Omega_o)$ we observe that $\nabla\psi_h(u) \leq 0$ on Ω_o and

$$\min_{x \in \Omega_o} \nabla\psi_h(u)(x) = \nabla\psi_h(u)(\hat{x}) = -\|(-\nabla\Psi(\mathcal{I}_h(u)))^{1/2} G_h^{\hat{x}}\|^2_{L^2(\Omega)} = -\sum_{i=1}^{\infty} (\partial S[\hat{q}]\delta q_i\,(\hat{x}))^2,$$

for some $\hat{x} \in \Omega_o$ and $\delta q_i = (-\nabla\Psi(\mathcal{I}_h(u)))^{1/2}\phi_i \in L^2(\Omega)$, $i \in \mathbb{N}$. Recalling the properties of the interpolation operator Λ_h and the measure-theoretic form of Jensen's inequality we estimate

$$(\partial S[\hat{q}]\delta q_i\,(\hat{x}))^2 = (\langle \partial S[\hat{q}]\delta q_i, \Lambda_h \delta_{\hat{x}}\rangle)^2 \leq \langle (\partial S[\hat{q}]\delta q_i)^2, \Lambda_h \delta_{\hat{x}}\rangle.$$

for every $i \in \mathbb{N}$ since $\partial S[\hat{q}]\delta q_i \in V_h$. Combining both results yields

$$\min_{x \in \Omega_o} \nabla\psi_h(u)(x) \geq -\sum_{i=1}^{\infty} \langle (\partial S[\hat{q}]\delta q_i)^2, \Lambda_h \delta_{\hat{x}}\rangle = \langle \nabla\psi_h(u), \Lambda_h \delta_{\hat{x}}\rangle.$$

Using $\Lambda_h \delta_{\hat{x}} \in \mathcal{M}^+(\Omega_o) \cap \mathcal{M}_h$ and $\|\Lambda_h \delta_{\hat{x}}\|_{\mathcal{M}} \leq 1$ we finally arrive at

$$\min_{x \in \mathcal{N}_h^o} \nabla\psi_h(u)(x) \geq \min_{x \in \Omega_o} \nabla\psi_h(u)(x) \geq \langle \nabla\psi_h(u), \Lambda_h \delta_{\hat{x}}\rangle \geq \min_{x \in \mathcal{N}_h^o} \nabla\psi_h(u)(x).$$

Setting $u = u^k$ we now conclude that $\hat{x}^k$ can be chosen from $\mathcal{N}_h^o$.

We proceed to prove the finite termination property. By assumption we know $u^1 \in \mathcal{M}^+(\Omega_o) \cap \mathcal{M}_h$. Thus, by induction, we get $u^{k+1} \in \mathcal{M}^+(\Omega_o) \cap \mathcal{M}_h$ from

$$\operatorname{supp} u^{k+1} \subset \operatorname{supp} u^k \cup \{\hat{x}^k\} \subset \mathcal{N}_h^o,$$

for all $k \in \mathbb{N}$. Since the finite dimensional suproblems in Algorithm 7 are solved up to optimality we have $j(u^{k+1}) < j(u^k)$ if u^k is not optimal. Thus we conclude

$$\operatorname{supp} u^{k+1} \in \mathcal{P}(\mathcal{N}_h^o) \setminus \bigcup_{i=1}^{k} \{\operatorname{supp} u^i\}, \quad k \in \mathbb{N},$$

where $\mathcal{P}(\mathcal{N}_h^o)$ denotes the power sets of $\mathcal{N}_h^o$. Consequently Algorithm 7 converges after at most $\#\mathcal{P}(\mathcal{N}_h^o) < \infty$ iterations. This completes the proof. □

The remainder of this section focuses on improved convergence results for the sequence of iterates $\{u^k\}_{k\in\mathbb{N}}$ generated by applying the Primal-Dual-Active-Point method to the spectral discretized problem $(\mathcal{P}_\beta^n)$. To this end we first recall some additional notation. We consider the continuous operators

$$\mathbf{P}_n \colon L^2(\Omega) \to \mathbb{R}^n, \quad q \mapsto ((q, \phi_1)_{L^2(\Omega)}, \dots, (q, \phi_n)_{L^2(\Omega)})^\top.$$

as well as $\mathcal{I}^n \colon \mathcal{M}(\Omega_o) \to \operatorname{Sym}(n)$ given by

$$\mathcal{I}^n(u) = \int_{\Omega_o} \partial S^n[\hat{q}](x) \partial S^n[\hat{q}](x)^\top \mathrm{d}u(x), \quad \partial S^n[\hat{q}](x) = (\partial S[\hat{q}]\phi_1\,(x), \dots, \partial S[\hat{q}]\phi_n\,(x))^\top,$$

for all $u \in \mathcal{M}(\Omega_o)$ and $x \in \Omega_o$. For abbreviation we set $\partial_i S[\hat{q}] = \partial S[\hat{q}]\phi_i \in \mathcal{C}(\Omega_o)$, $i = 1, \dots, n$. Last, we introduce the optimal design criterion

$$\Psi^n \colon \operatorname{Sym}(n) \to \mathbb{R} \cup \{+\infty\}, \quad B \mapsto \Psi(\mathbf{P}_n^* B \mathbf{P}_n)$$

on the space of symmetric matrices. We make the following observations.

Lemma 5.44. *Let $n \in \mathbb{N}$ large enough be given. Then the functional Ψ^n has the following properties:*

- *There holds* $\operatorname{NND}(n) \subset \operatorname{dom} \Psi^n$.
- *The functional Ψ^n is two times continuously differentiable and convex on* $\operatorname{NND}(n)$.
- *The functional Ψ^n is monotone with respect to the Loewner ordering on* $\operatorname{NND}(n)$.
- *There holds*

$$\Psi(P_n \mathcal{I}(u) P_n) = \Psi^n(\mathcal{I}^n(u)) \quad \forall u \in \mathcal{M}(\Omega).$$

Proof. We only prove the last statement, the remaining claims follow from the assumptions on Ψ and the linearity of $\mathbf{P}_n$. Given two arbitrary functions $q_1,\ q_2 \in L^2(\Omega)$ we readily calculate

$$\begin{aligned}(q_1, P_n\mathcal{I}(u)P_nq_2)_{L^2(\Omega)} &= \sum_{i=1}^{n}\sum_{j=1}^{n}(q_1,\phi_i)_{L^2(\Omega)}(q_2,\phi_j)_{L^2(\Omega)}\langle\partial_i S^n[\hat{q}]\partial_j S^n[\hat{q}],u\rangle \\ &= (\mathbf{P}_nq_1, \mathcal{I}^n(u)\mathbf{P}_nq_2)_{\mathbb{R}^d} \\ &= (q_1, \mathbf{P}_n^*\mathcal{I}^n(u)\mathbf{P}_nq_2)_{L^2(\Omega)},\end{aligned}$$

From this observation we conclude

$$P_n\mathcal{I}(u)P_n = \mathbf{P}_n^*\mathcal{I}^n(u)\mathbf{P}_n \quad \forall u \in \mathcal{M}(\Omega_o).$$

Thus the desired statement follows. □

In order to improve on the sublinear convergence of the Primal-Dual-Active-Point method we interpret the spectral discretized optimal design criterion $\psi^n = \Psi(P_n \cdot P_n) \circ \mathcal{I}$ as the composition of Ψ^n and the operator $\mathcal{I}^n$. Thus we rewrite problem $(\mathcal{P}_\beta^n)$ as

$$\min_{u\in\mathcal{M}^+(\Omega_o)} F^n(u) = [\Psi^n(\mathcal{I}^n(u)) + \beta\|u\|_{\mathcal{M}}].$$

Note that F^n is radially unbounded on $\mathcal{M}^+(\Omega_o)$ since $F(u) \leq F^n(u)$ for all $u \in \mathcal{M}^+(\Omega)$ due to the monotonicity of Ψ. Since $\mathcal{I}^n$ maps continuously into the space of symmetric matrices this rewritten problem resembles the sensor placement problems discussed in the previous chapter. As a consequence, improved convergence results can be concluded by arguing similarly to Section 4.4.3. To this end, we first comment on the coercivity of the Hessian $\nabla^2\Psi^n$.

Lemma 5.45. *If Ψ is strictly convex on $\operatorname{Pos}(L^2(\Omega), L^2(\Omega))$ then Ψ^n is strictly convex on $\operatorname{NND}(n)$. In this case, the projected optimal Fisher information $\mathcal{I}^n(\bar{u}_\beta^n)$ and thus $P_n\mathcal{I}(\bar{u}_\beta^n)P_n$ are the same for every optimal solution $\bar{u}_\beta^n$ to $(\mathcal{P}_\beta^n)$. In addition, if there exist a constant $\gamma_0 > 0$ and a norm $\|\cdot\|$ on $\operatorname{SHS}(L^2(\Omega), L^2(\Omega))$ with*

$$\langle\langle\delta B, \nabla^2\Psi(P_n\mathcal{I}(\bar{u}_\beta^n)P_n)\delta B\rangle\rangle_{\operatorname{HS}(L^2(\Omega),L^2(\Omega))} \geq \gamma_0\|\delta B\|^2 \quad \forall\delta B \in \operatorname{SHS}(L^2(\Omega), L^2(\Omega)), \tag{5.62}$$

then there exists a constant $\gamma_0^n > 0$, possibly depending on $n \in \mathbb{N}$, with

$$\operatorname{Tr}_{\mathbb{R}^n}(\delta\hat{B}\nabla^2\Psi^n(\mathcal{I}^n(u))\delta\hat{B}) \geq \gamma_0^n\|\delta\hat{B}\|_{\operatorname{Sym}}^2 \quad \forall\delta\hat{B} \in \operatorname{Sym}(n).$$

Proof. Let $B_1,\ B_2 \in \operatorname{NND}(n)$ with $B_1 \neq B_2$ be given. It its readily verified that there holds $\mathbf{P}_n^*B_i\mathbf{P}_n \in \operatorname{Pos}(L^2(\Omega), L^2(\Omega))$, $i = 1, 2$. For $s \in (0, 1)$ we conclude

$$\begin{aligned}\Psi^n(B_1 + s(B_2 - B_1)) &= \Psi(\mathbf{P}_n^*(B_1 + s(B_2 - B_1))\mathbf{P}_n) \\ &\leq \Psi(\mathbf{P}_n^*B_1\mathbf{P}_n) + s(\Psi(\mathbf{P}_n^*B_2\mathbf{P}_n) - \Psi(\mathbf{P}_n^*B_1\mathbf{P}_n)) \\ &= \Psi^n(B_1) + s(\Psi^n(B_2) - \Psi^n(B_1))\end{aligned}$$

and strict inequality holds if and only if $\mathbf{P}_n^*B_1\mathbf{P}_n \neq \mathbf{P}_n^*B_2\mathbf{P}_n$. Assume that $\mathbf{P}_n^*(B_1 - B_2)\mathbf{P}_n = 0$. Testing with $\phi_i,\ \phi_j,\ i, j = 1, \dots, n$, from left and right, respectively, we then get

$$0 = (\phi_i, \mathbf{P}_n^*(B_1 - B_2)\mathbf{P}_n\phi_j)_{L^2(\Omega)} = (B_1 - B_2)_{ij}.$$

This contradicts $B_1 \neq B_2$. Thus strict inequality holds and Ψ^n is strictly convex on $\mathrm{NND}(n)$. In particular, this implies that the projected Fisher information matrix $\mathcal{I}^n(\bar{u}^n_\beta)$ is the same for every optimal solution $\bar{u}^n_\beta$ to $(\mathcal{P}^n_\beta)$. By definition, the uniqueness of $P_n\mathcal{I}(\bar{u}^n_\beta)P_n$ also follows.

Now assume that (5.62) holds. For $\delta\hat{B} \in \mathrm{Sym}(n)$ we calculate

$$\mathrm{Tr}_{\mathbb{R}^n}(\delta\hat{B}, \nabla^2\Psi^n(\mathcal{I}^n(u))\delta\hat{B}) = \mathrm{Tr}_{L^2(\Omega)}(\nabla^2\Psi(P_n\mathcal{I}(\bar{u}^n_\beta)P_n)\hat{B}) \geq \gamma_0\|\mathbf{P}^*_n\delta\hat{B}\mathbf{P}_n\|^2.$$

Recalling that

$$\mathbf{P}^*_n\delta\hat{B}\mathbf{P}_n = 0 \in \mathrm{SHS}(L^2(\Omega), L^2(\Omega)) \Leftrightarrow \delta\hat{B} = 0 \in \mathrm{Sym}(n).$$

it is straightforward to verify that $\|\mathbf{P}^*_n\cdot\mathbf{P}_n\|$ defines a norm on the symmetric matrices. Since $\mathrm{Sym}(n)$ is finite dimensional this new norm is equivalent to the Frobenius norm i.e. there exists a constant $\theta^n_0 > 0$, possibly depending on $n \in \mathbb{N}$ with

$$\gamma_0\|\mathbf{P}^*_n\delta\hat{B}\mathbf{P}_n\|^2 \geq \gamma^n_0\|\delta B\|^2_{\mathrm{Sym}} \quad \forall\delta\hat{B} \in \mathrm{Sym}(n).$$

This finishes the proof. □

We adopt the assumption on the strict convexity of Ψ and its curvature around the optimal projected Fisher information operator for the rest of this section.

Assumption 5.8. Let Ψ be strictly convex on $\mathrm{Pos}(L^2(\Omega), L^2(\Omega))$. Moreover there exists a constant $\gamma_0 > 0$ with

$$\langle\langle\delta B, \nabla^2\Psi(P_n\mathcal{I}(\bar{u}^n_\beta)P_n)\delta B\rangle\rangle_{\mathrm{HS}(L^2(\Omega),L^2(\Omega))} \geq \gamma_0\|\delta B\|^2 \quad \forall\delta B \in \mathrm{SHS}(L^2(\Omega), L^2(\Omega)),$$

This assumption together with Lemma 5.45 imply the uniform convexity of Ψ^n in a neighborhood $N(\mathcal{I}^n(\bar{u}_\beta)) \subset \mathrm{NND}(n)$ of the unique optimal projected Fisher information matrix $\mathcal{I}^n(\bar{u}^n_\beta)$ depending on the truncation parameter $n \in \mathbb{N}$. There holds

$$(\nabla\Psi^n(B_1) - \nabla\Psi^n(B_2), B_1 - B_2)_{\mathrm{Sym}} \geq \frac{\gamma^n_0}{2}\|B_1 - B_2\|^2_{\mathrm{Sym}} \quad \forall B_1, B_2 \in N(\mathcal{I}^n(\bar{u}^n_\beta))$$

for some $\gamma^n_0 > 0$. We further assume additional regularity of the sensitivities $\partial_i S[\hat{q}]$, $i = 1, \ldots, N$, and define $\bar{p}^n = -(\mathcal{I}^n)^*\nabla\Psi^n(\mathcal{I}^n(\bar{u}^n_\beta))$.

Assumption 5.9. Assume that there holds

$$\{\, x \in \Omega_o \mid \bar{p}^n(x) = \beta \,\} = \{\bar{x}_i\}_{i=1}^N \subset \operatorname{int}\Omega_o.$$

Moreover the set $\{\mathcal{I}^n(\delta_{\bar{x}_i})\}_{i=1}^N$ is linearly independent and there exists $R > 0$ with

$$\Omega_R := \bigcup_{i=1}^N B_R(\bar{x}_i) \subset \operatorname{int}\Omega_o, \quad \bar{B}_R(\bar{x}_i) \cap \bar{B}_R(\bar{x}_j) = \emptyset,\ i \neq j, \quad \partial S^n[\hat{q}] \in \mathcal{C}^2(\bar{\Omega}_R, \mathbb{R}^n) \cap \mathcal{C}(\Omega_o).$$

Along the lines of Corollary 4.6, we conclude that this assumption guarantees the uniqueness and sparsity of the optimal solution $\bar{u}^n_\beta = \sum_{i=1}^N \bar{\mathbf{u}}_i\delta_{\bar{x}_i}$ to $(\mathcal{P}^n_\beta)$. Moreover it is readily verified that the Banach space adjoint of $\mathcal{I}^n$ maps continuously to $\mathcal{C}^2(\bar{\Omega}_R) \cap \mathcal{C}(\Omega_o)$ since

$$[(\mathcal{I}^n)^*B](x) = \partial S^n[\hat{q}](x)^\top B\partial S^n[\hat{q}](x) \quad \forall B \in \mathrm{Sym}(n),\ x \in \Omega_o.$$

In particular there holds $\bar{p}^n \in \mathcal{C}^2(\bar{\Omega}_R) \cap \mathcal{C}(\Omega)$. As a last ingredient we impose additional assumptions on the curvature of $\bar{p}^n$ and the coefficients of $\bar{u}^n_\beta$.

Assumption 5.10. There holds $\operatorname{supp} \bar{u}^n_\beta = \{\bar{x}_i\}_{i=1}^N$, i.e. $\bar{\mathbf{u}}_i > 0$, $i = 1, \dots, N$, and there exists a constant $\theta^n_0 > 0$ with

$$-(\zeta, \nabla^2 \bar{p}^n(\bar{x}_i)\zeta)_{\mathbb{R}^d} \geq \theta^n_0 |\zeta|^2_{\mathbb{R}^d} \quad \forall \zeta \in \mathbb{R}^d$$

and all $i = 1, \dots, N$.

Interpreting the spectral discretized optimal design criterion as $\psi^n = \Psi^n \circ \mathcal{I}^n$ it is now readily verified that $(\mathcal{P}^n_\beta)$ and the associate optimal solution $\bar{u}_\beta$ fulfill the prerequisites of Theorem 6.70. We conclude the following improved convergence result for the residual in the spectral discretized problem.

Theorem 5.46. *Let the sequence $\{u^k\}_{k \in \mathbb{N}} \subset \mathcal{M}^+(\Omega_o)$ be generated by applying Algorithm 7 to $(\mathcal{P}^n_\beta)$ and let Assumptions 5.8, 5.9 and 5.10 hold. Then there exist constants $c_n, R_n > 0$ and $\zeta_n \in (0,1)$ with*

$$\operatorname{supp} u^k \subset \bigcup_{i=1}^N \bar{B}_{R_n}(\bar{x}_i), \quad \operatorname{supp} u^k \cap \bar{B}_{R_n}(\bar{x}_i) \neq \emptyset, \ i = 1, \dots, N,$$

as well as

$$r_F(u^k) + \max_{i=1,\dots,N} \max_{x \in \operatorname{supp} u^k \cap \bar{B}_{R_n}(\bar{x}_i)} |x - \bar{x}_i|_{\mathbb{R}^d} + \max_{i=1,\dots,N} |\bar{\mathbf{u}}_i - \|u^k|_{\bar{B}_{R_n}(\bar{x}_i)}\|_{\mathcal{M}}| \leq c_n \zeta^k_n,$$

for all $k \geq k_n \in \mathbb{N}$. All appearing constants may depend on the truncation parameter $n \in \mathbb{N}$.

We point out that under the same assumptions convergence rates for the modified Wasserstein distances of the iterates can be obtained along the lines of Theorem 4.21.

5.3.2 Implementation details

To close this section we discuss the practical realization of Algorithm 7 applied to the fully discrete problem $(\mathcal{P}^n_{\beta,h})$. In view of the numerical experiments presented in the following section we focus on the A-optimal design criterion

$$\psi_A(u) = \Psi_A(\mathcal{I}(u)) = \operatorname{Tr}_{L^2(\Omega)}(\mathcal{C}_{\text{post}}(\mathcal{I}(u))), \quad \mathcal{C}_{\text{post}}(\mathcal{I}(u)) = (\mathcal{I}(u) + \mathcal{I}_0)^{-1}, \quad u \in \mathcal{M}^+(\Omega_o),$$

and consider its evaluation as well as the efficient computation of its gradient. Given a design measure $u \in \mathcal{M}^+(\Omega_o)$ we recall that the evaluation of the gradient at a given spatial point $x \in \Omega_o$ can be related to the Green's function G^x by

$$\nabla \psi(u)(x) = -(G^x, \mathcal{C}_{\text{post}}(\mathcal{I}(u))^2 G^x)_{L^2(\Omega)} = -\|\mathcal{C}_{\text{post}}(\mathcal{I}(u)) G^x\|^2_{L^2(\Omega)}.$$

Let us now consider the fully discretized problem $(\mathcal{P}^n_{\beta,h})$ for given $n \in \mathbb{N}$ and $h \leq h_0$ small enough. Throughout the following discussions we assume that the first n eigenpairs $\{(\lambda_i, \phi_i)\}_{i=1}^n$, of $\mathcal{I}_0^{-1}$ are either analytically available or good approximations can be obtained numerically by e.g applying several steps of an Arnoldi iteration to a discretization of $\mathcal{I}_0^{-1}$. Furthermore the truncation parameter $n \in \mathbb{N}$ is reasonably small in the sense that the discrete sensitivities $\partial S_n^h[\hat{q}] \in \mathcal{C}(\Omega_o, V_h^n)$ defined by

$$\partial S_n^h[\hat{q}] \colon \Omega_o \to \mathbb{R}^n, \quad x \mapsto (\partial S^h[\hat{q}]\phi_1\,(x), \dots, \partial S^h[\hat{q}]\phi_n\,(x))^\top,$$

can be pre-computed and stored for further use. The $k-$th component of $\partial S_n^h[\hat{q}]$ is denoted by $\partial_k S_n^h[\hat{q}] \in \mathcal{C}(\Omega_o)$, $k = 1, \dots, n$. We emphasize that besides one solve of the discrete state equation (5.38) and n solutions of the sensitivity equation (5.39) to obtain $\partial_k S_n^h[\hat{q}] = \partial S^h[\hat{q}]\phi_k$, $k = 1, \dots n$, no additional PDE solves will be required in the following. This allows for an efficient and fast solution of $(\mathcal{P}_{\beta,h}^n)$.

Arguing similarly as in Section 5.2.2 a fully discrete A-optimal design can now be obtained by solving the sensor placement problem

$$\min_{u \in \mathcal{M}^+(\Omega_o)} [\mathrm{Tr}_{\mathbb{R}^n}((\mathcal{I}_h^n(u) + \mathcal{I}_0^n)^{-1}) + \beta \|u\|_{\mathcal{M}}]. \tag{5.63}$$

Here, given a design measure $u \in \mathcal{M}^+(\Omega_o)$, the matrices $\mathcal{I}_0^n \in \mathrm{PD}(n)$ and $\mathcal{I}_h^n(u) \in \mathrm{NND}(n)$ are characterized through

$$(\mathcal{I}_0^n)_{ij} = \delta_{ij}\lambda_i, \quad \mathcal{I}_h^n(u)_{ij} = \langle \partial_i S_n^h[\hat{q}]\, \partial_j S_n^h[\hat{q}], u \rangle, \quad i, j \in \{1, \dots, n\}.$$

With a slight abuse of notation we abbreviate $\psi_h^n(u) = \mathrm{Tr}_{\mathbb{R}^n}((\mathcal{I}_h^n(u) + \mathcal{I}_0^n)^{-1})$ for $u \in \mathcal{M}^+(\Omega_o)$.

Efficient evaluation of the covariance operator

In all steps of Algorithm 7 matrix-vector products between the covariance matrix $(\mathcal{I}_h^n(u) + \mathcal{I}_0^n)^{-1}$ corresponding to a sparse design measure $u \in \mathcal{M}^+(\Omega_o)$ and a potentially large set of vectors $\{\delta q_i\}_{i \in \mathbf{I}}$, $\mathbf{I} \subset \mathbb{N}$ need to be computed. Since the dimension of the discretized parameter space is assumed to be reasonably small we comment on an efficient realization of this task based on Cholesky decompositions. Here we exploit the structure of the covariance matrix and the sparsity of the design measure. Set $N = \#\operatorname{supp} u < \infty$. We distinguish between two cases.

Case 1: Assume that $N \geq n$. In this case we compute the Cholesky decomposition of the matrix $\mathcal{I}_h^n(u) + \mathcal{I}_0^n = GG^\top$ in $O(n^3)$ operations and solve $(\mathcal{I}_h^n(u) + \mathcal{I}_0^n)z = \delta q_i$ by forward-backward substitution for all $i \in \mathbf{I}$. This can be realized in $O(\#\mathbf{I} \cdot n^2)$ operations.

Case 2: Second consider $N < n$ and let $u = \sum_{i=1}^N \mathbf{u}_i \delta_{x_i}$. The associated Fisher-information matrix can be decomposed as $\mathcal{I}_h^n(u) = X^T \Sigma^{-1} X$ where the matrices $X \in \mathbb{R}^{N \times n}$ and $\Sigma^{-1} \in \mathbb{R}^{N \times N}$ are defined as

$$X_{jk} = \partial_k S^h[\hat{q}](x_j), \quad \Sigma_{ij}^{-1} = \delta_{ij}\mathbf{u}_i, \quad i, j = 1, \dots, N, \ k = 1, \dots, n.$$

Applying the Sherman-Morrison-Woodbury formula, [130], we obtain

$$\begin{aligned}(\mathcal{I}_h^n(u) + \mathcal{I}_0^n)^{-1} &= (X^T \Sigma^{-1} X + \mathcal{I}_0^n)^{-1} \\ &= (\mathcal{I}_0^n)^{-1} - (\mathcal{I}_0^n)^{-1} X^\top \Sigma^{-1/2} (\mathrm{Id} + \Sigma^{-1/2} X (\mathcal{I}_0^n)^{-1} X^\top \Sigma^{-1/2})^{-1} \Sigma^{-1/2} X (\mathcal{I}_0^n)^{-1},\end{aligned}$$

where $\mathrm{Id} \in \mathbb{R}^{N \times N}$ is the identity matrix and $\Sigma_{ij}^{-1/2} = \delta_{ij}\sqrt{\lambda_i}$, $i, j \in \{1, \dots, N\}$. Since $\mathcal{I}_0^n$ is a diagonal matrix calculating its inverse is straightforward and can be done in $O(n)$ operations. Furthermore a Cholesky decomposition of

$$\mathrm{Id} + \Sigma^{-1/2} X (\mathcal{I}_0^n)^{-1} X^\top \Sigma^{-1/2} = \hat{G}\hat{G}^\top,$$

can be obtained in $O(N^3)$ operations. Solving the systems $(\mathcal{I}_h^n(u) + \mathcal{I}_0^n)z = \delta q_i$, $i \in \mathbf{I}$, then requires a combined effort of $O(\#\mathbf{I} \cdot (n + N^2 + n \cdot N))$.

Evaluation of the optimal design criterion

Based on the previous arguments we now discuss the efficient evaluation of the A-optimal design criterion for a given sparse design measure $u = \sum_{i=1}^{N} \mathbf{u}_i \delta_{x_i}$. As before we distinguish two cases.

Case 1: Let $N \geq n$. In this situation we compute a Cholesky decomposition of $\mathcal{I}_h^n(u) + \mathcal{I}_0^n = GG^\top$ and observe

$$\operatorname{Tr}_{\mathbb{R}^n}((\mathcal{I}_h^n(u) + \mathcal{I}_0^n)^{-1}) = \sum_{i=1}^{n} (\mathbf{e}_i, (\mathcal{I}_h^n(u) + \mathcal{I}_0^n)^{-1} \mathbf{e}_i)_{\mathbb{R}^n} = \sum_{i=1}^{n} |G^{-1}\mathbf{e}_i|^2_{\mathbb{R}^n}.$$

Hence the optimal design criterion is evaluated in $O(n^3)$ operations.

Case 2: If $N < n$ we compute a Cholesky decomposition of

$$\operatorname{Id} + \Sigma^{-1/2} X (\mathcal{I}_0^n)^{-1} X^\top \Sigma^{-1/2} = \hat{G}\hat{G}^\top.$$

Similar calculations as before show

$$\operatorname{Tr}_{\mathbb{R}^n}((\mathcal{I}_h^n(u) + \mathcal{I}_0^n)^{-1}) = \sum_{i=1}^{n} [\lambda_i - \lambda_i^2 |\hat{G}^{-1} \Sigma^{-1/2} X_{\cdot,i}|^2_{\mathbb{R}^N}],$$

where $X_{\cdot,i} \in \mathbb{R}^N$ denotes the i-th column of X. Consequently the objective functional can be evaluated in $O(n \cdot N^2)$ operations.

Evaluation of the gradient in step 1.

The new sensor location $\hat{x}^k \in \Omega_o$ in step 1. of Algorithm 7 is found as a global minimizer of $\nabla \psi_h^n(u^k)$. Due to Proposition 5.43 the search for the new point can be restricted to $\mathcal{N}_h^o$. Thus given a sparse design measure $u = \sum_{i=1}^{N} \mathbf{u}_i \delta_{x_i}$ we have to efficiently evaluate the gradient $\nabla \psi_h^n(u)$ for every $x \in \mathcal{N}_h^o$. A similar computation as on the continuous level gives

$$\nabla \psi_h^n(u)(x) = -(\partial S_n^h[\hat{q}](x), (\mathcal{I}_h^n(u) + \mathcal{I}_0^n)^{-2} \partial S_n^h[\hat{q}](x))_{\mathbb{R}^n} = -\|(\mathcal{I}_h^n(u) + \mathcal{I}_0^n)^{-1} \partial S_n^h[\hat{q}](x)\|^2_{\mathbb{R}^n}.$$

Similar to the continuous case this relates the gradient to the sensitivity vector $\partial S_n^h[\hat{q}] \in \mathcal{C}(\Omega_o, \mathbb{R}^n)$. This allows for its efficient evaluation in $O(n^3 + \#\mathcal{N}_h^o \cdot n^2)$ operations, if $N \geq n$, and

$$O(N^3 + \#\mathcal{N}_h^o(n + N^2 + n \cdot N))$$

operations if $N < n$. Moreover the representation of the gradient also facilitates a parallelization of its computation.

Simultaneous insertion of multiple points

While the insertion of a single point in every iteration of Algorithm 7 guarantees the convergence of the procedure, the number of Dirac deltas in the approximated optimal design will affect the practical performance of the algorithm. In order to accelerate the convergence of the method in practice we consider the following heuristic multi-point insertion strategy. First the gradient $\nabla \psi_h^n(u^k)$ is evaluated in the grid nodes contained in Ω_o. If u^k is not an optimal solution we have

$\nabla\psi_h^n(u^k)(x) < -\beta$ for $x \in \mathcal{N}_h^o$. In order to obtain a set of new candidate locations we first build the connectivity graph of the nodes $\mathcal{N}_h$ and compute the subset

$$\mathcal{N}_h^- = \{\, x \in \Omega_o \mid x \in \mathcal{N}_h^o,\ \nabla\psi_h^n(u^k) < -\beta \,\} \subset \mathcal{N}_h^o$$

Subsequently, we compare the value of the gradient in each node $x \in \mathcal{N}_h^-$ with those of the neighbouring ones. We call $x \in \mathcal{N}_h^-$ a local minimum of the current gradient $\nabla\psi_n^h(u^k)$ if x minimizes $\nabla\psi_h^n(u^k)$ over its adjacent grid nodes. All local minima of the gradient in this sense are assembled in a set of promising new sensor locations $\mathcal{N}_h^k$. Then the elements of this set are ordered and we add the M points $\hat{x}_i^k$, $i = 1, \dots M$, corresponding to the smallest local minima to the active set

$$\mathcal{A}_k = \operatorname{supp} u^k \cup \{\hat{x}_i^k\}_{i=1}^M.$$

Here M is an a priori chosen maximal number of new Dirac delta functions. The new iterate u^{k+1} is now found by solving the finite dimensional subproblem on $\mathcal{A}_k$. Since the global minimizer $\hat{x}^k \in \mathcal{N}_h^o$ of the gradient is an element of $\mathcal{N}_h^k$ the convergence guarantees derived in the previous section remain valid.

Sequential point insertion for $(\mathcal{P}_{\beta,h})$

We briefly discuss the numerical realization of Algorithm 7 for the FE-discretized sensor placement problem $(\mathcal{P}_{\beta,h})$ and highlight the differences to the fully discrete case. For simplification we assume that the implicitly discretized parameter space Q_h is given by V_h. Adapting the arguments in Example 5.6 this is e.g. the case for the identification of the right hand side of the Neumann Laplacian with zero order term. Under these assumptions the semi-discrete problem A-optimal design problem is equivalent to solving

$$\min_{u \in \mathcal{M}^+(\Omega_o)} [\operatorname{Tr}_{V_h}((\mathcal{I}_h(u) + \mathcal{I}_0)^{-1}) + \beta\|u\|_{\mathcal{M}}]. \tag{5.64}$$

Setting $\psi_h(u) = \operatorname{Tr}_{V_h}((\mathcal{I}_h(u) + \mathcal{I}_0)^{-1})$ our special interest lies in the numerical realization of the point insertion step (step 1. in Algorithm 7). As before we have to compute the discrete gradient on the grid points contained in Ω_o. We assume that the discrete set $\mathcal{N}_h^o$ is large and, in particular, it is not possible to pre-compute the discretized Green's function G_h^x for every possible sensor location. Given $x \in \mathcal{N}_h^o$ we derive

$$\nabla\psi_h(u)(x) = -(G_h^x, (\mathcal{I}_h(u) + \mathcal{I}_0)^{-2} G_h^x)_{L^2(\Omega)} = -\|(\mathcal{I}_h(u) + \mathcal{I}_0)^{-1} G_h^x\|_{L^2(\Omega)}^2,$$

for a sparse design measure $u \in \mathcal{M}^+(\Omega_o)$. To obtain the evaluated gradient we proceed in two steps. First we determine G_h^x. As on the continuous level we compute a function $\mathcal{G}_h^x \in W_h$ fulfilling the discrete PDE

$$a'_{h,y}(\hat{q}, S^h[\hat{q}])(\varphi_h, \mathcal{G}_h^x) = \langle \varphi_h, \delta_x \rangle \quad \forall \varphi_h \in Y_h.$$

The function $G_h^x \in V_h$ is then identified with

$$-a'_{h,q}(\hat{q}, S[\hat{q}])(\mathcal{G}_h^x, \cdot) \in V_h^* \simeq V_h.$$

Second, we need to compute the application of $(\mathcal{I}_h(u) + \mathcal{I}_0)^{-1}$ to G_h^x. To this end, we recall that $\mathcal{I}_0$ is often modeled as a differential operator. In particular, it is infeasible to pre-compute its (discretized) inverse and we can only compute the action of $(\mathcal{I}_h(u)+\mathcal{I}_0)^{-1}$ on a given function by numerically solving the associated covariance PDE. In order to find the new sensor location this procedure has to be repeated for every $x \in \mathcal{N}_h^o$ leading to a total of $2\#\mathcal{N}_h^o$ PDE solves for one evaluation of the gradient. While this seems reasonable if $\mathcal{N}_h^o$ is small we recall that sensor placement problems with a possibly infinite number of candidate sensor locations in the continuous problem are at the heart of this thesis. In this light the proposed evaluation strategy is numerically prohibitive. To illustrate this fact we again consider the identification of the right hand side of a Laplacian equation and $\bar{\Omega} = \Omega_o$, i.e. $\mathcal{N}_h^o = \mathcal{N}_h$. In this case, computing $G_h^x \in V_h$ for every $x \in \mathcal{N}_h$ is equivalent to computing the sensitivity operator $\partial S^h[\hat{q}]$. This corresponds to inverting the associated stiffness matrix which is infeasible.

We draw several conclusions from the discussions in this section. On the one hand they justify, at least to some extend, the proposed full discretization of the problem by a low-rank approximation of the parameter. This leads to sensor placement problems which are amenable for efficient and practically fast solution algorithms. On the other hand they highlight again that the results of this chapter should be understood as a first step on sensor placement problems with both, an infinite dimensional parameter and infinitely many possible sensor locations. For example they open up new questions on the efficient numerical solution of $(\mathcal{P}_{\beta,h})$. As described, a first computational bottleneck in Algorithm 7 applied to $(\mathcal{P}_{\beta,h})$ is given by the search for a global minimizer of $\nabla\psi(u^k)$ in a subset of the grid nodes. To mitigate the computational complexity of this step we may e.g. resort to randomized methods for the compuation of the minimum. Furthermore sophisticated low-rank approximations of $\partial S^h[\hat{q}]$ could be considered. Finally these results suggest the use of adaptive methods to keep both, the number of sensor locations as well as the number of parameters, as small as possible. Altogether this leaves space and need for future research.

As a last remark we point out that we did not discuss the efficient numerical solution of the finite dimensional subproblems in step 2 of Algorithm 7. One particular reason for this approach is that the overall convergence rate of the algorithm is independent of the method used for their solution. In particular, any algorithmic procedure for smooth and convex optimization problems with box constraints can be used. However we stress that the previous considerations on the efficient evaluation of the design criterion and its gradient also apply to the subproblems. Moreover we are confident that these arguments can be extended to the computation of higher order derivatives as used in Newton-like methods or interior point procedures.

5.4 Numerical examples

We close this chapter with the study of two numerical examples. First we consider the task of inferring on the distributed source term entering in a Laplacian equation. The main focus of this example lies on the influence of the cost parameter β, the truncation parameter N and the a priori covariance operator $\mathcal{I}_0^{-1}$ on the sparsity pattern of the optimal design. In a second, more practically motivated, example, we again consider optimal sensor placement problems for the estimation of a diffusion coefficient, which is modeled as a log-normal distributed random field. The primary goal in this example is to study the scalability of the primal-dual-active-point method as well as to compare it to the continuation strategy discussed in Section 4.4.6. In what follows, we consider the unit square $\bar{\Omega} = \Omega_o = [0,1]^2$ and a sequence $\mathcal{T}_{h_k}$, $k \in \{1,2,\dots,8\}$, of uniform triangulations of Ω_o with $h_k = \sqrt{2}/2^k$. The parameter q is modeled as a Gaussian random field distributed according to $\mu_0 = \mathcal{N}(0, \mathcal{I}_0^{-1})$. The prior covariance operator $\mathcal{I}_0^{-1}$ is given as the inverse of

$$\mathcal{I}_0 = c_1(-\Delta + c_2 \operatorname{Id})^s.$$

Here Δ denotes the Dirichlet Laplacian on Ω and the values of $s > 1$ and $c_1,\ c_2 > 0$ control the smoothing properties of $\mathcal{I}_0^{-1}$. We will be specify these constants for each example separately. Due to the tensor structure of the spatial domain, analytic expressions for the eigenpairs $(\lambda_{(i,j)}, \phi_{(i,j)})_{i,j\in\mathbb{N}}$ of $\mathcal{I}_0^{-1}$ can be obtained. In detail, we get

$$\lambda_{(i,j)} = (\pi^2(i^2+j^2)+c_2)^{-s}/c_1, \quad \phi_{(i,j)}(x_1,x_2) = 2\sin(\pi i x_1)\sin(\pi j x_2) \quad \forall i,j \in \mathbb{N}. \tag{5.65}$$

Consequently the random field q admits a Karhunen-Loève expansion in terms of $\phi_{(i,j)}$ as

$$q = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \sqrt{\lambda_{(i,j)}}\zeta_{(i,j)}\phi_{(i,j)} = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} q_{(i,j)}\phi_{(i,j)}, \quad \text{where} \quad q_{(i,j)} \sim \mathcal{N}(0, \lambda_{(i,j)}) \quad \forall i,j \in \mathbb{N}.$$

After truncating both series after $N \in \mathbb{N}$ terms we obtain the discretized random field

$$q^N = \sum_{i=1}^{N}\sum_{j=1}^{N} q_{(i,j)}\phi_{(i,j)} \quad q_{(i,j)} \sim \mathcal{N}(0, \lambda_{(i,j)}) \quad \forall i,j \in \mathbb{N}.$$

In the following examples, the truncated random field describes our prior uncertainty on the true value of an unknown parameter entering into a partial differential equation. For optimal inference on the parameter we take pointwise measurements of the state variable according to a solution of the A-optimal design problem (5.63). The a priori matrix $\mathcal{I}_0^N$ is chosen as a diagonal matrix with

$$\left[\mathcal{I}_0^N\right]_{kk} = 1/\lambda_{(i,j)}, \quad k = N(i-1)+j, \ \ i,j \in \{1, \dots N\}. \tag{5.66}$$

To visualize the obtained optimal design measures we stick to the post-processing procedure discussed in Section 4.6.1 and replace clusters of optimal sensors by a single Dirac delta function located at their center of mass. Its coefficient is given by the added measurement weights of all sensors in the cluster.

5.4.1 Estimation of a distributed source term

As a first example we consider the identification of the distributed source term in a diffusion process described by the Laplacian with mixed Dirichlet and Neumann boundary conditions. We define the Dirichlet part of the boundary as $\Gamma_D = \{0,1\} \times (0,1)$ corresponding to the left and right boundaries of the unit square. Given $q \in L^2(\Omega)$ the associated state $y = S[q] \in H^1(\Omega) \cap \mathcal{C}(\Omega_o)$ fulfills

$$a(q,y) = \int_\Omega \left[\nabla y \cdot \nabla \varphi - \frac{1}{4} q \varphi\right] \,\mathrm{d}x = 0 \quad \forall \varphi \in H^1(\Omega), \ \varphi = 0 \text{ on } \Gamma_D,$$

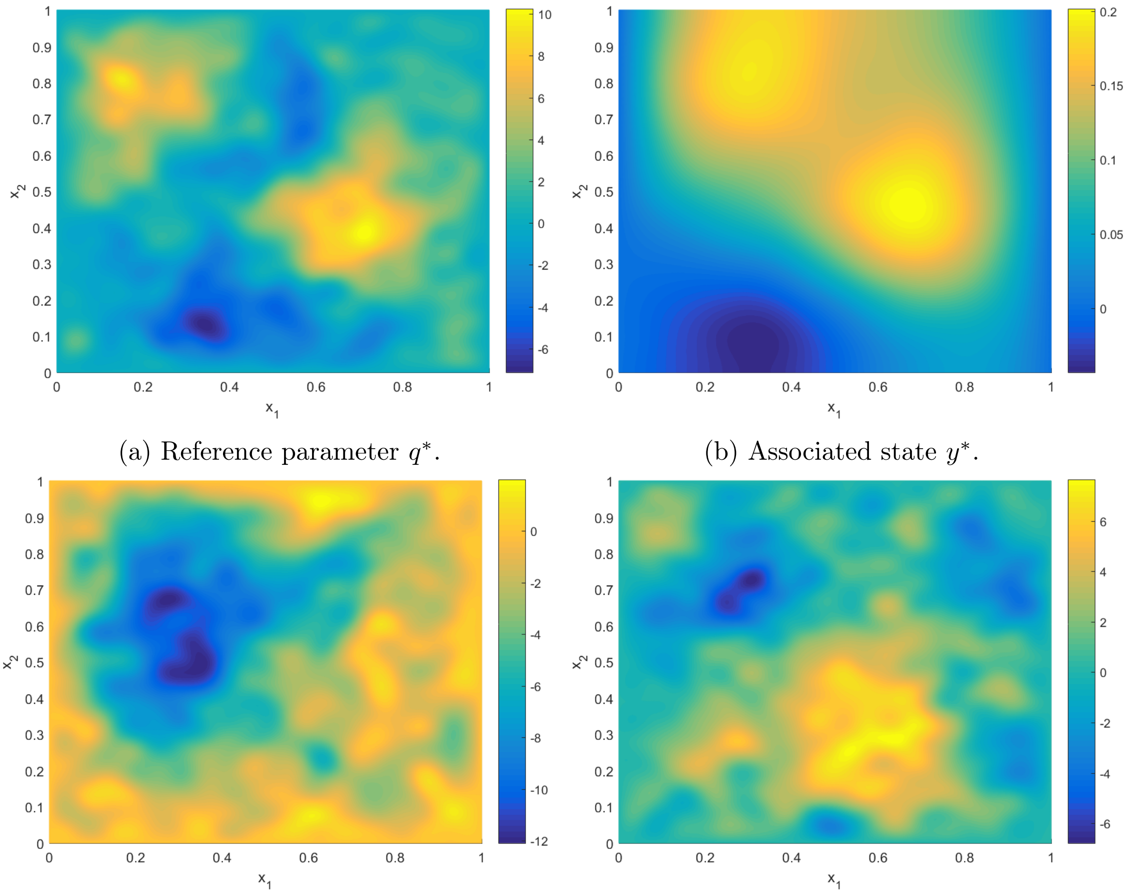
as well as $y = 0$ on Γ_D. Due to the linear dependence between state and parameter, the sensitivity $\delta y_{(i,j)}$ of the state with respect to $q_{(i,j)}$, $i,j = 1, \dots, N$, fulfills

$$a(q, \delta_{(i,j)}y)(\varphi) = \int_\Omega \left[\nabla \delta_{(i,j)} y \cdot \nabla \varphi - \frac{1}{4}\phi_{(i,j)}\varphi\right] \,\mathrm{d}x = 0 \quad \forall \varphi \in H^1(\Omega), \ \varphi = 0 \text{ on } \Gamma_D$$

and $\delta_{(i,j)}y = 0$ on Γ_D.

Estimation for different β

We fix the smoothing parameters to $c_1 = 10^{-5}$, $c_2 = 10$ and $s = 2$. In this section we tend to illustrate the influence of the cost parameter $\beta > 0$ on the optimal design measure and thus also on the estimates for the unknown parameters. For this purpose, we consider a reference parameter q^* which is obtained as a realization of the random field $q^N/4$ for $N = 20$ and $h = h_8$. It is depicted in Figure 5.1 alongside the associated state and two additional realizations of $q^N/4$. we especially point out to the different scales of the obtained realizations which indicates high variability in the prior distribution. Now, we set the truncation index to $N = 12$ and compute A-optimal designs for $\beta = 1, 10^{-3}, 10^{-5}$. According to each obtained measurement setup $\bar{u}^N_{\beta,h} = \sum_{i=1}^{N_h} \mathbf{u}_i \delta_{x_i}$, we generate a vector of measurement data $\mathbf{y}_d \in \mathbb{R}^{N_h}$ with $\mathbf{y}_d^i = y^*(x_i) + \epsilon_i$ where ϵ_i is a realization of $\varepsilon_i \sim \mathcal{N}(0, 1/\mathbf{u}_i)$. Subsequently, we compute the posterior distribution $\mu_{\text{post}}^{\mathbf{y}_d}$ of $q^N/4$ given the data $\mathbf{y}_d$. The computed optimal designs are displayed in Figures 5.2, 5.3 and 5.4. Alongside each measurement design, we plot the mean of $\mu_{\text{post}}^{\mathbf{y}_d}$ given by the MAP estimate $q_{\text{post}}^{\mathbf{y}_d}$ as well as two realizations of $q^N/4$ given $\mathbf{y}_d \in \mathbb{R}^N$.



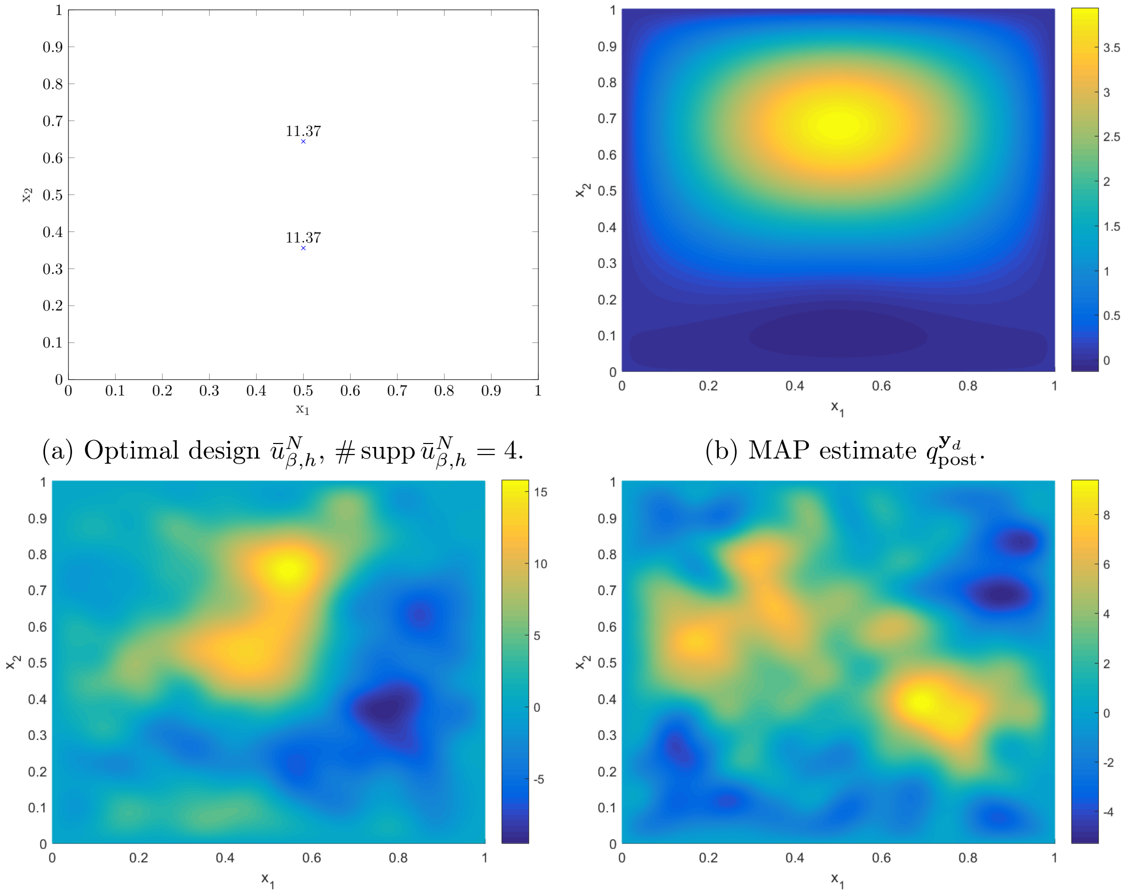
(a) Reference parameter q^*.

(b) Associated state y^*.

(c) Realizations of $q^N/4$.

Figure 5.1: Reference parameter q^*, associated state y and realizations of random field.

Let us first interpret the structure of the obtained optimal designs. We observe that the number of placed sensors grows as the cost parameter β decreases. This verifies the interpretation of β as



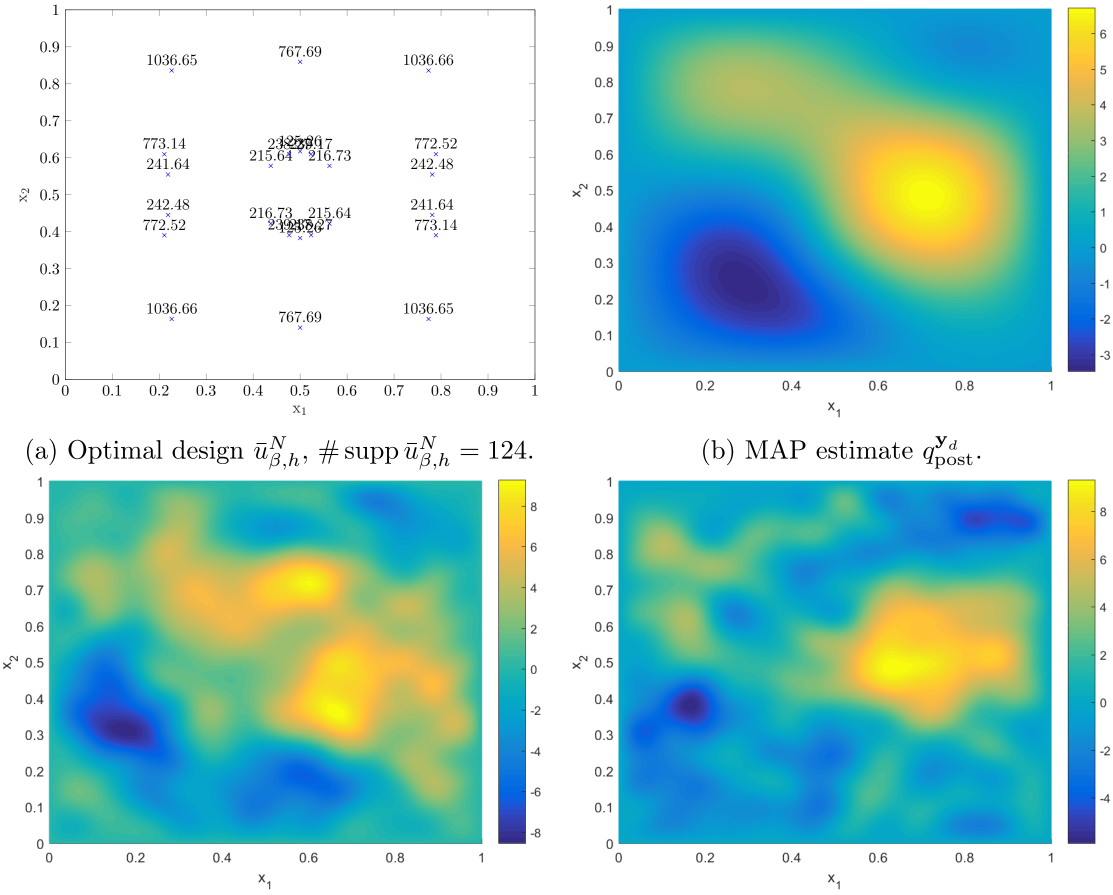
(a) Optimal design $\bar{u}^N_{\beta,h}$, $\#\,\mathrm{supp}\,\bar{u}^N_{\beta,h} = 4$.

(b) MAP estimate $q^{\mathbf{y}_d}_{\mathrm{post}}$.

(c) Random draws from $\mu^{\mathbf{y}_d}_{\mathrm{post}}$.

Figure 5.2: Optimal design, MAP estimate and draws from posterior for $\beta = 1$.

tool to provide indirect control on the sparsity optimal solutions to the sensor placement problem. Moreover, we note that all obtained designs are symmetric with respect to the x_1 and x_2 axis. For large β, sensors are exclusively placed in the center of Ω_o while for smaller values of the cost parameter we also get optimally positioned sensors towards the boundary. We recall that A-optimal designs are chosen to minimize the pointwise posterior variance of the random field. In order to explain their structure, we should therefore take a look at the pointwise prior variance field of the unknown parameter. Furthermore, the uncertainty in the parameter is also propagated into the solution of the partial differential equation. Therefore, it is also necessary to interpret the (truncated) state $\delta y = \sum_{i=1}^{N} \sum_{i=1}^{N} q_{(i,j)} \delta_{(i,j)} y$ as random field. The pointwise variance of the state variable describes our prior uncertainty on the true value of the measurement at a point $x \in \Omega_o$ if no additional measurement errors are present. Intuitively, if measurement resources are limited, i.e. β is large, it is reasonable to only take measurements at points in which the prior uncertainty of the state variable is large. We plot the pointwise prior variances of the parameter as well as the state in Figure 5.5. Note that both functions are symmetric. Furthermore their maximum is assumed in the center of Ω_o and they become smaller towards the boundary. In particular, we point out that the pointwise variance of the parameter is equal to zero on the whole boundary which is a consequence of the predescribed homogeneous Dirichlet boundary conditions in the prior covariance operator. Thus, if β is large, i.e. the cost of a single measurement is high, the



(a) Optimal design $\bar{u}^N_{\beta,h}$, $\# \operatorname{supp} \bar{u}^N_{\beta,h} = 124$.

(b) MAP estimate $q^{\mathbf{y}_d}_{\text{post}}$.

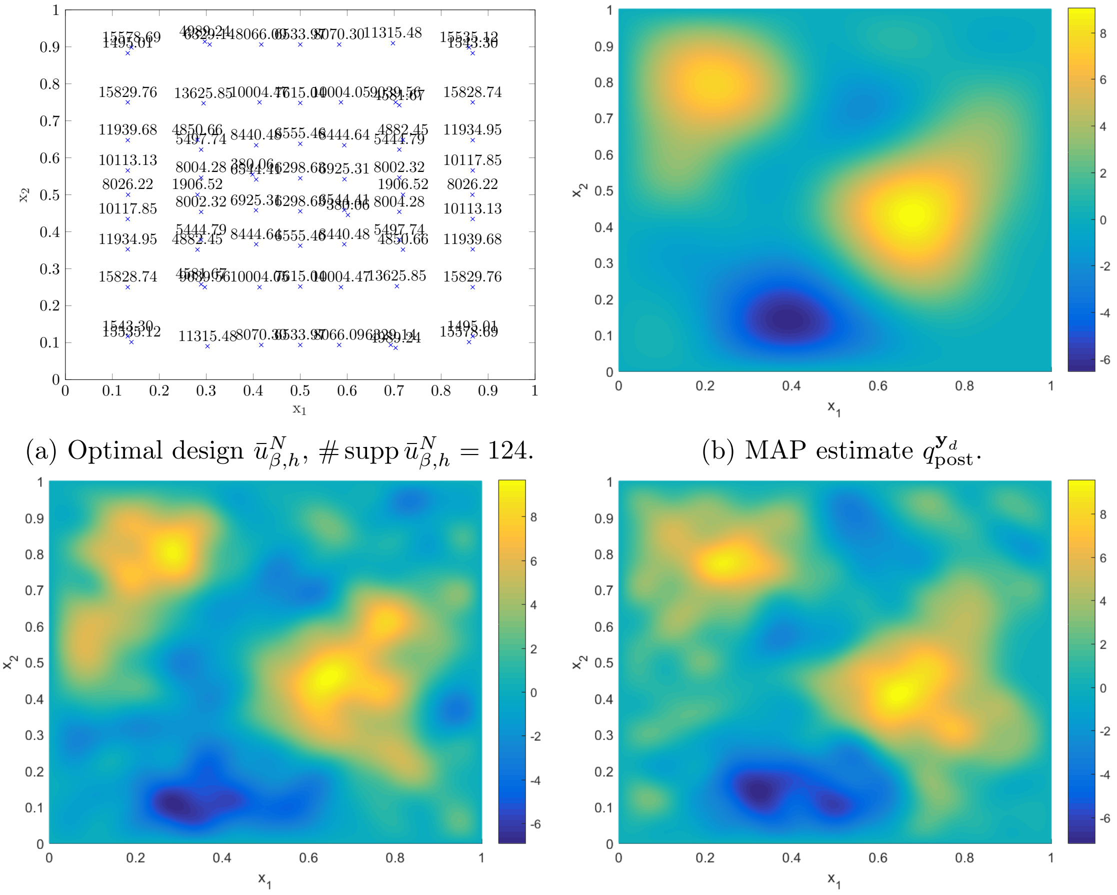
(c) Random draws from $\mu^{\mathbf{y}_d}_{\text{post}}$.

Figure 5.3: Optimal design, MAP estimate and draws from posterior for $\beta = 10^{-3}$.

optimal design proposes to only place measurements at points of highest prior uncertainty. As β gets smaller, additional measurements may also be performed at points corresponding to smaller prior uncertainty in order to further decrease the posterior variance. We also point out that the A-optimal design criterion measures the variability of the posterior measure $\mu^{\mathbf{y}_d}_{\text{post}}$ around its mean as well as the expected mean squared error of the MAP estimator. Consequently, as β decreases, we may, on the one hand, expect that random draws from the posterior distribution are close to $q^{\mathbf{y}_d}_{\text{post}}$. On the other hand, $q^{\mathbf{y}_d}_{\text{post}}$ should be close to the reference parameter q^*. This intuition is, at least visually, confirmed by our numerical results. Additionally, they again highlight the dependence of the Bayesian approach on the provided prior distribution.

Optimal designs for different truncation and smoothing parameters

Second, we study the dependence of the optimal design on the number of unknown parameters in the KL expansion of the random field as well as the exponent $s > 1$. For this purpose we choose the smoothing parameters c_1, c_2 as in the previous examples and set the cost parameter to $\beta = 0.01$. Subsequently, A-optimal designs for $N \in \{5, 10, 15, 20\}$ and $s \in \{1.6, 2\}$ are computed. The resulting design measures $\bar{u}^N_{\beta,h}$ can be found in Figures 5.6 and 5.7, respectively. We draw several conclusions based on these results. First, we note that the number of unknown



(a) Optimal design $\bar{u}^N_{\beta,h}$, $\#\operatorname{supp}\bar{u}^N_{\beta,h} = 124$.

(b) MAP estimate $q^{\mathbf{y}_d}_{\text{post}}$.

(c) Random draws from $\mu^{\mathbf{y}_d}_{\text{post}}$.

Figure 5.4: Optimal design, MAP estimate and draws from posterior for $\beta = 10^{-5}$.

parameters $q_{(i,j)}$, $i, j = 1, \dots, N$, in the truncated KL-expansion q^N grows quadratic with N. However, the support size of the computed A-optimal designs does not. Quite the contrary, the number of Dirac delta functions remains constant for N large enough. This stems back to the smoothing properties of the prior covariance operator $\mathcal{I}_0^{-1}$ i.e. the convergence of its eigenvalues $\lambda_{(i,j)}$ towards zero. Clearly, the rate of convergence at which this sequence approaches zero and thus the exponent s critically impact the obtained results. In fact, we point out that there is almost no visual difference between the optimal designs associated to $N = 5$ and $N = 20$ for $s = 2$. This indicates, at least for the present example, that we may obtain a good approximation to an optimal design for the original problem by solving the spectral discretized sensor placement problem for only a small number of modes. However, we again stress that these observations inherently depend on the choice of the prior distribution. We also recall Theorem 5.31 and Proposition 5.35 which give an upper bound of $n(n+1)/2$ for $n = N \cdot N$ on the number of Dirac delta functions in a fully discrete optimal design. The present results show that this upper bound is overly pessimistic in general. Moreover, we point out that the support sizes of the designs corresponding to $N = 20$, $\#\operatorname{supp}\bar{u}^N_{\beta,h} = 15$ and $\#\operatorname{supp}\bar{u}^N_{\beta,h} = 93$, respectively, are small in comparison to the number of possible sensor locations $\mathcal{N}_{h_8} = 66049$. In particular, the computed results may hint at the existence of a sparse solution to the continuous optimal sensor placement problem. However, this is far from being conclusive. Last, the computed results provide, to some extend,

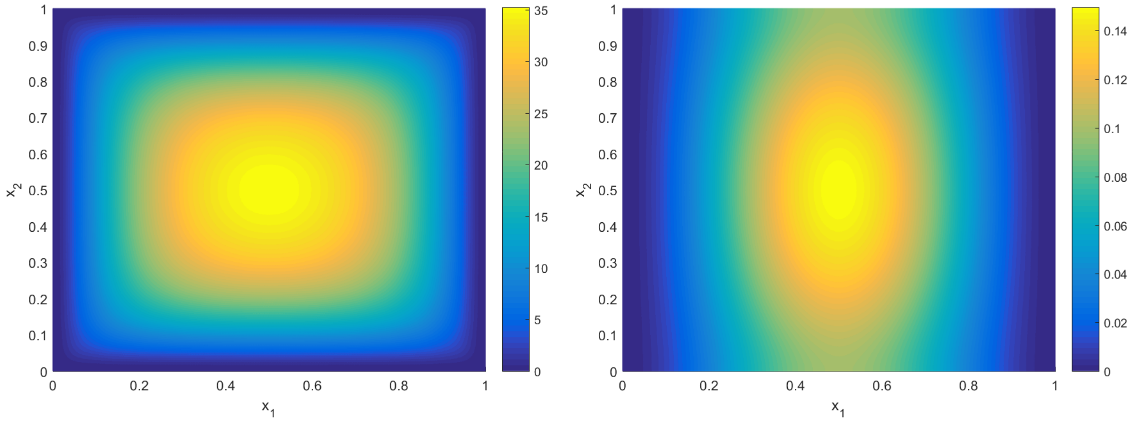


Figure 5.5: Pointwise prior variances for unknown parameter (left) and associated state (right).

a visual confirmation of the weak* convergence result on the spectral discretized optimal designs, see Proposition 5.32.

5.4.2 Estimation of a distributed diffusion coefficient

The setting in this second example is motivated by the task of estimating spatially varying diffusion parameters, which is a common problem in, e.g., geophysical applications. Therefore we consider a stationary diffusion process, where the unknown parameter is the distributed diffusion coefficient. We again set $\bar{\Omega} = [0,1]^2$ to be the unit square and define the Dirichlet boundary as $\Gamma_D = \{0,1\} \times (0,1)$. For $q \in \mathbb{R}^n \simeq \mathbb{R}^{N \times N}$ for some $N \in \mathbb{N}$ we define the parametrization

$$q^N(x) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} q_{(i,j)} \phi_{(i,j)}, \quad \text{where } \phi_{(i,j)}(x_1, x_2) = 2 \sin(\pi i x_1) \sin(\pi j x_2).$$

Given $q \in \mathbb{R}^{N \times N}$ the associated state $y = S[q]$ is the unique element of $H^1(\Omega) \cap \mathcal{C}(\Omega_o)$ satisfying

$$a(q,y)(\varphi) = \int_\Omega [\exp(q^N) \nabla y \cdot \nabla \varphi - f\varphi] \mathrm{d}x = 0 \quad \forall \varphi \in H^1(\Omega),\ \varphi = 0 \text{ on } \Gamma_D. \tag{5.67}$$

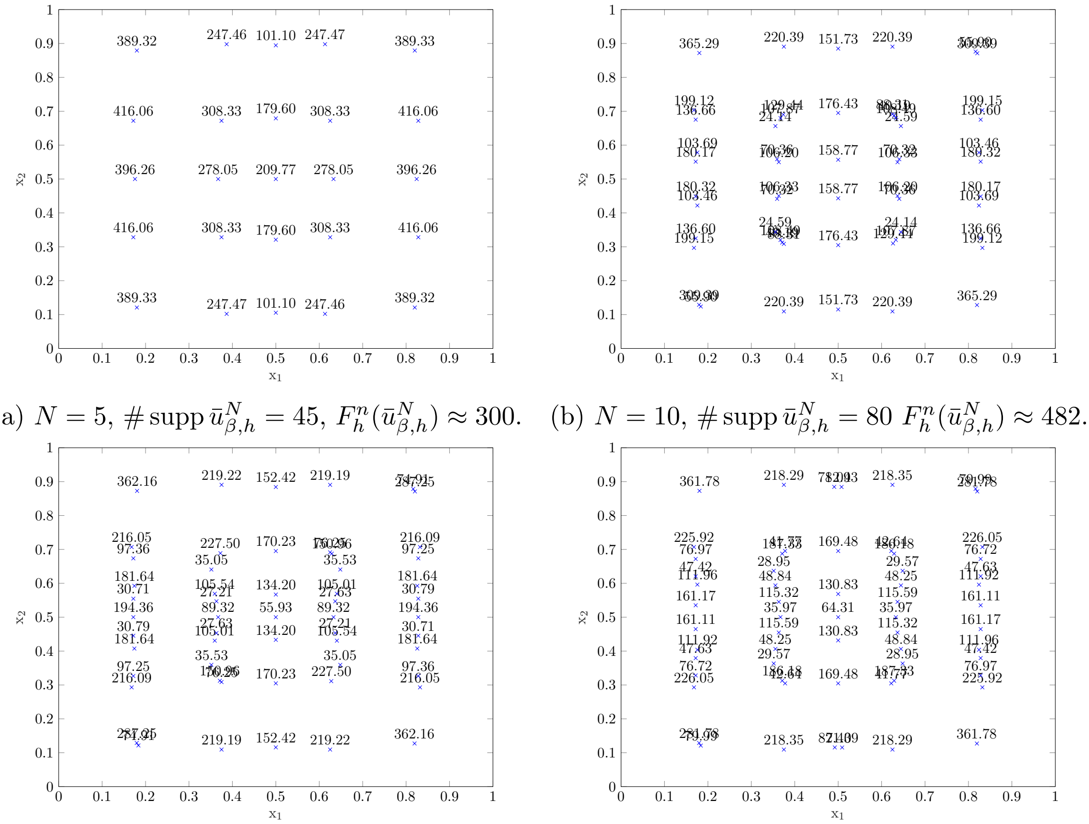
for some known source term $f \in L^2(\Omega)$ and $y = x_1$ on Γ_D. It can be easily seen that (5.67) corresponds to the linear equation

$$\begin{aligned} -\nabla \cdot \big(\exp(q^N) \nabla y\big) &= f && \text{in } \Omega, \\ y &= x_1 && \text{on } \Gamma_D, \\ \exp(q^N) \partial_n y &= 0 && \text{on } \partial\Omega \setminus \Gamma_D. \end{aligned} \tag{5.68}$$

Note that due to the linearity of the equation, the sensitivity $\delta_{(i,j)} y = \partial_{(i,j)} S[q] \in H^1(\Omega) \cap \mathcal{C}(\Omega_o)$ of the state with respect to the (i,j)-th entry of q for $i,j \in \{1,\dots,N\}$ satisfies

$$a(q, \delta_{(i,j)} y)(\varphi) = -\int_{\Omega_o} \frac{1}{2} \phi_{(i,j)} \exp(q^N) \nabla y \cdot \nabla \varphi \mathrm{d}x \quad \forall \varphi \in H^1(\Omega),\ \varphi = 0 \text{ on } \Gamma_D$$

and $\delta_{(i,j)} y = 0$ on Γ_D.



(a) $N = 5$, $\#\operatorname{supp} \bar{u}^N_{\beta,h} = 45$, $F^n_h(\bar{u}^N_{\beta,h}) \approx 300$.

(b) $N = 10$, $\#\operatorname{supp} \bar{u}^N_{\beta,h} = 80$ $F^n_h(\bar{u}^N_{\beta,h}) \approx 482$.

(c) $N = 15$, $\#\operatorname{supp} \bar{u}^N_{\beta,h} = 93$, $F^n_h(\bar{u}^N_{\beta,h}) \approx 543$.

(d) $N = 20$, $\#\operatorname{supp} \bar{u}^N_{\beta,h} = 93$, $F^n_h(\bar{u}^N_{\beta,h}) \approx 573$.

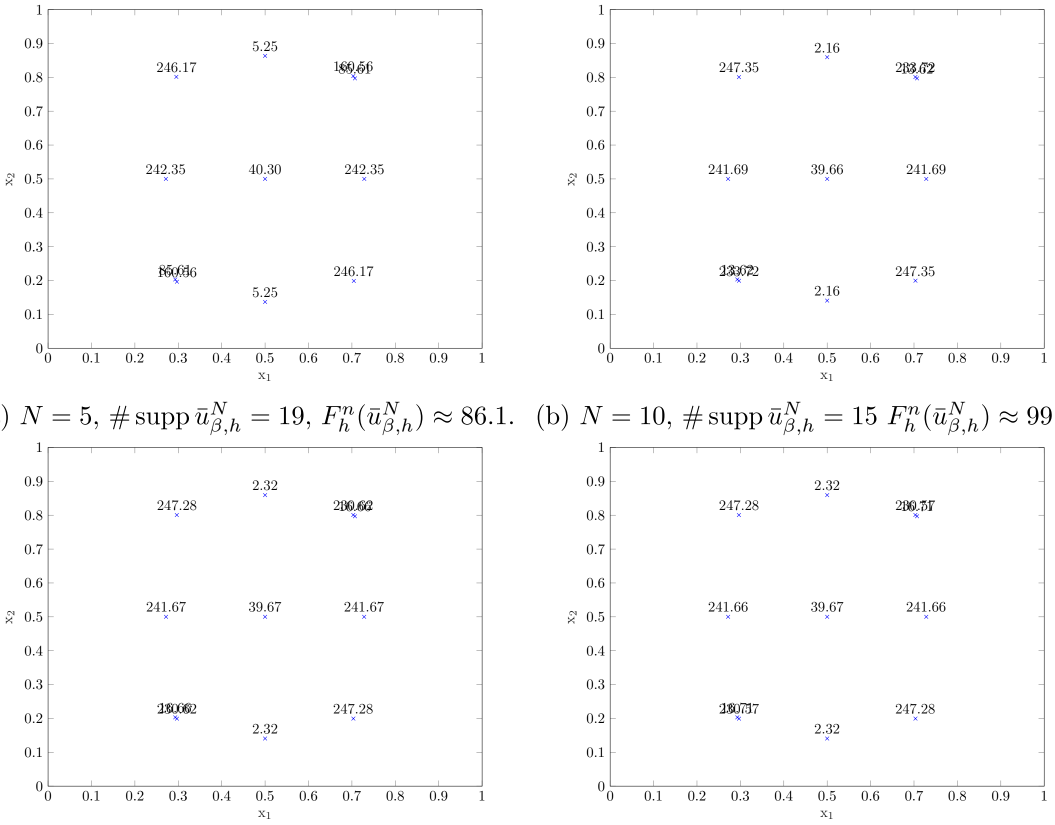
Figure 5.6: Optimal designs $\bar{u}^N_{\beta,h}$ for different N and $s = 1.6$.

Optimal designs for different right hand sides

In contrast to the previous example the dependence of the state on the unknown parameter is nonlinear for this one. Thus, the sensitivity operator $\partial S[\hat{q}]$ depends on the linearization point $\hat{q}$ as well as the associated state $y = S[\hat{q}]$. Therefore, this example is suitable to study the influence of changes in the state equation on the obtained optimal designs. For this purpose, we consider the diffusion equation in (5.68) for two different right hand sides f given by

$$f_1(x_1,x_2) = 0, \quad f_2(x_1,x_2) = \frac{5}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_1-0.75)^2+(x_2-0.5)^2}{2\sigma^2}\right), \quad (x_1,x_2) \in \Omega_o.$$

We consider a total of $n = N \cdot N = 400$ terms in the parametrization of q^N and set $\beta = 1$, $c_1 = 10^{-5}$ and $c_2 = 10$. The linearization point is chosen as $\hat{q} = 0$. We interpret the expansion coefficients $q_{i,j}$, $i,j = 1,\dots N$, as random variables with $q_{(i,j)} \sim \mathcal{N}(0, \lambda_{(i,j)})$, see (5.65), for $s = 2$. Let us briefly give some interpretation to the considered setup. As described at the beginning of this chapter equation (5.68) models the diffusion of a fluid in a porous medium Ω whose permeability is described by the unknown diffusion coefficient. The state variable y is given by the fluid pressure. A priori we do not have information on the true permeability and thus it is assumed to be constant on Ω, i.e $\hat{q} = 0$. Furthermore, choosing $f = f_1 = 0$, corresponds to the assumption that the net fluid flow into and out of Ω is zero. In a different scenario, water is continuously pumped into the ground through a pipeline or a well located at $(0.5, 2.5)$. This is



(a) $N = 5$, $\# \operatorname{supp} \bar{u}^N_{\beta,h} = 19$, $F^n_h(\bar{u}^N_{\beta,h}) \approx 86.1$.
(b) $N = 10$, $\# \operatorname{supp} \bar{u}^N_{\beta,h} = 15$ $F^n_h(\bar{u}^N_{\beta,h}) \approx 99.6$.

(c) $N = 15$, $\# \operatorname{supp} \bar{u}^N_{\beta,h} = 15$, $F^n_h(\bar{u}^N_{\beta,h}) \approx 103$.
(d) $N = 20$, $\# \operatorname{supp} \bar{u}^N_{\beta,h} = 15$, $F^n_h(\bar{u}^N_{\beta,h}) \approx 104$.

Figure 5.7: Optimal designs $\bar{u}^N_{\beta,h}$ for different N and $s = 2$.

modeled by the smoothed Dirac delta function f_2. We set $\sigma = 0.01$ in the following. For both choices of f a Bayesian A-optimal measurement design with $\mathcal{I}^N_0$ chosen according to (5.66) for $s = 2$ is computed using the Primal-Dual-Active-Point method. The resulting measures are displayed in Figures 5.8 and 5.9. In order to interpret the obtained results we plot the pointwise variance of the random field $\delta y = \sum_{i=1}^N \sum_{j=1}^N q_{(i,j)} \delta_{(i,j)} y$ alongside. Note that the variance depends on the state variable $y = S[\hat{q}]$ and thus on the right hand side f. By construction, the pointwise prior variance of the random field q^N is (up to scaling) the same as in Figure 5.5. For $f = f_1 = 0$ we make similar observations as in the first example. The pointwise prior variances of the state and the parameter are symmetric. This symmetry is also recovered in the A-optimal design. Moreover, note that some of the optimal sensors are again placed at locations in the center of Ω_o corresponding to points of highest prior uncertainty. In contrast, the optimal design corresponding to f_2 is still symmetric with respect to the x_1 axis but optimal sensors cluster towards the center of the well at $(0.75, 0.5)$ while none of them are placed close to the left part of the boundary. An, at least partial, explanation is provided by the pointwise prior variance of the (linearized) state which peaks at this point and is considerably smaller outside of a small neighborhood. Nevertheless, we point out that optimal sensors are also placed at locations in which the prior variance of the parameter is large but the uncertainty on the true value of the measurement is small. This stresses that optimization based Bayesian optimal sensor placement entangles information provided by both, the mathematical model and the prior distribution.

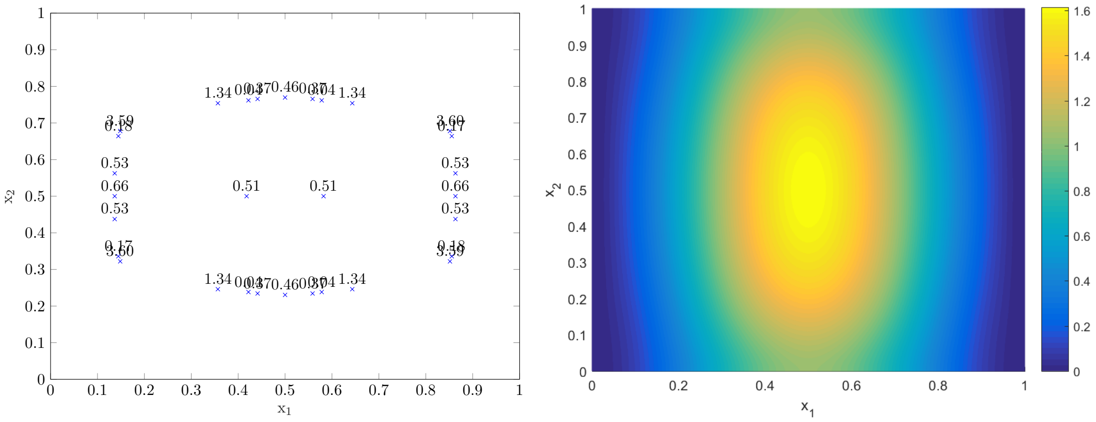


Figure 5.8: Optimal design (left) and pointwise prior variance of δy (right) for f_1.

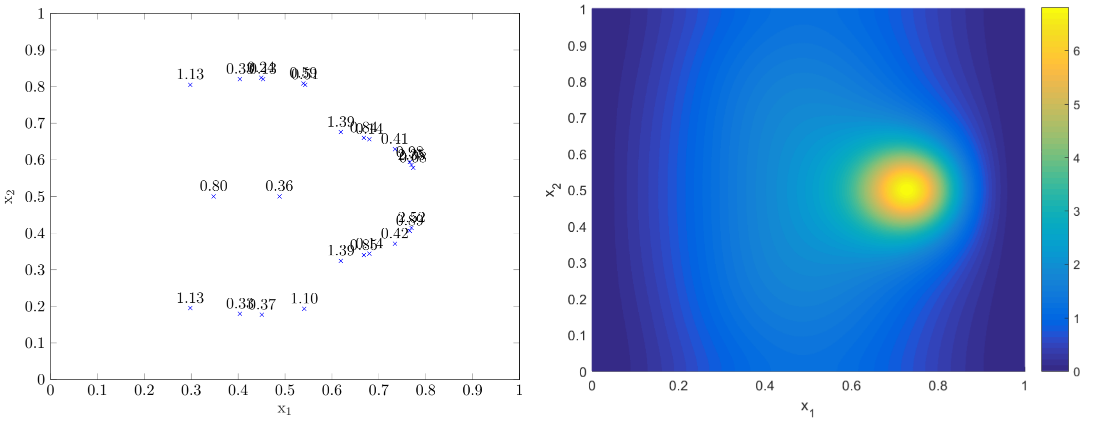


Figure 5.9: Optimal design (left) and pointwise prior variance of δy (right) for f_2.

Comparison of point-insertion and path-following: No a priori knowledge

In this section we compare the performance of the Primal-Dual-Active-Point method (Algorithm 7) and the algorithmic solution approach based on the Hilbert-space regularization (Algorithm 6) on the computation of Bayesian A-optimal designs for the diffusion coefficient example. Such a comparison was postponed until now, since the small number of parameters in the numerical examples contained in Chapter 4 aid the performance of the PDAP method. Let us briefly recap the path-following approach. For $\varepsilon > 0$ we determine the unique solution $\bar{u}^{N,\varepsilon}_{\beta,h}$ to the regularized discrete problem

$$\min_{u_h \in V_h, u_h \geq 0} \left[\psi^n_h(\Lambda_h u_h) + \beta \|u_h\|_{L^1(\Omega_o)} + \frac{\varepsilon}{2} \|u_h\|^2_{L^2(\Omega_o),h} \right], \qquad (\mathcal{P}^{n,\varepsilon}_{\beta,h})$$

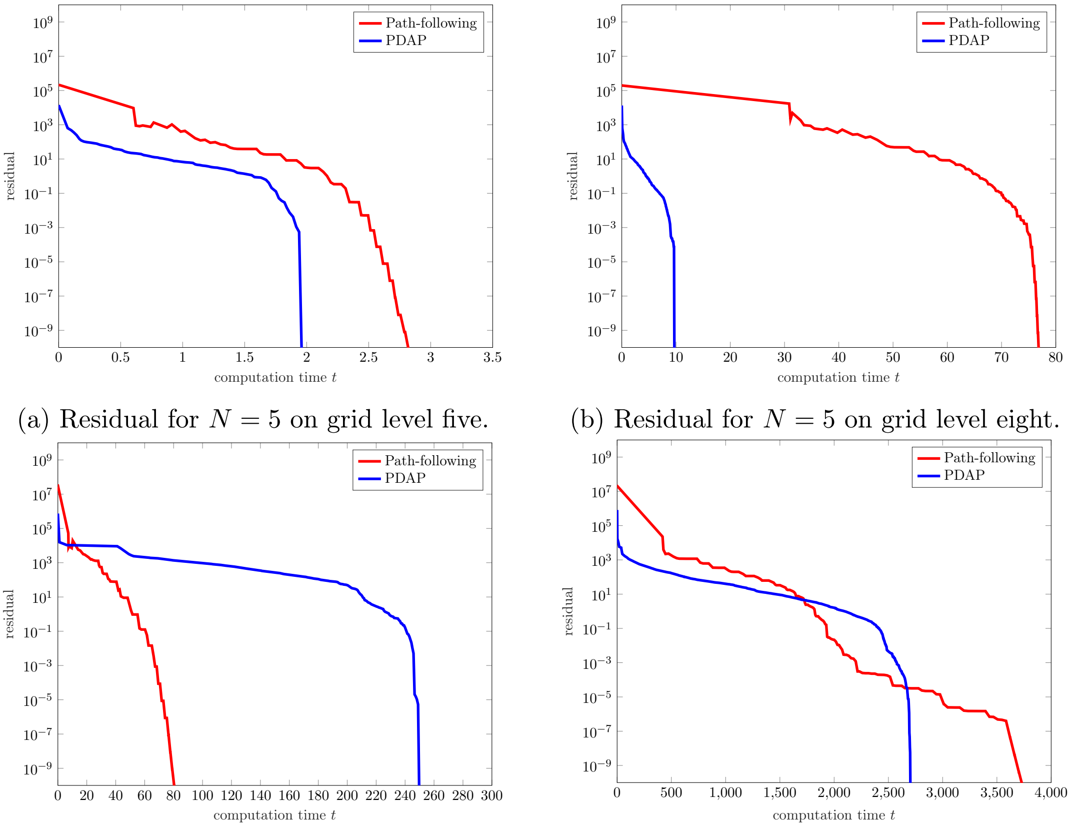
where $\|u_h\|^2_{L^2(\Omega_o),h} = (i_h(u_h^2), 1)_{\Omega_o,h}$ denotes the lumped regularization term and ψ^n_h, for $n = N \times N$, denotes the discretized Bayesian A-optimal design criterion. For a more detailed discussion of this problem we refer to Section 4.6. To compute a solution of the unregularized problem, which is recovered for $\varepsilon = 0$, we employ a continuation strategy for the regularization parameter ε as described in Section 4.4.6. Since both algorithms are fundamentally different and partly rely on different computational routines, a comparison in terms of number of steps is difficult. For this reason, we focus on the computation times in the following. We place special emphasis on the qualitative influence of the mesh width and the support size of the optimal design.

Therefore we consider the A-optimal design problem for the diffusion coefficient example with different $N \in \mathbb{N}$ and on different refinement levels of the spatial discretization. The cost parameter is chosen as $\beta = 1$. In order to provide some control over the minimum number of Dirac deltas in the optimal design measure $\bar{u}^N_{\beta,h}$ we formally assume that no a priori knowledge is present, i.e. we first set $\mathcal{I}^N_0 = 0$. The parameter-to-state mapping is linearized at $\hat{q} = 0 \in \mathbb{R}^{N \times N}$. Given a fixed $N \in \mathbb{N}$ and h small enough such that the discrete design problem (5.63) admits an optimal solution $\bar{u}^N_{\beta,h}$ we note that $\# \operatorname{supp} \bar{u}^N_{\beta,h} \geq n = N^2$; cf. Proposition 4.3. Consequently, by increasing N we also raise the number of optimal Dirac delta functions that both algorithms have to identify.

Let us briefly comment on the implementation of the two different algorithms. For PDAP we stick to the description in Algorithm 7 without an additional application of the post-processing strategy in Algorithm. The iteration is stopped at step k if the primal-dual gap fulfills $\Phi(u^k) \leq 10^{-9}$. For Algorithm 6 we set $\varepsilon_1 = 10^{-3}$ and $\varepsilon_l = \varepsilon_{l-1}/\sqrt{10}$ for $l > 1$. For each l the regularized subproblem $(\mathcal{P}^{n,\varepsilon}_{\beta,h})$ is solved by using the semi-smooth Newton method presented in [208]. We include a globalization strategy based on a damping of the Newton steps to ensure a decrease of the regularized objective function value in every iteration. The arising linear systems are solved by a cg-method up to machine precision. If the norm of the right-hand side in the Newton system is smaller than some tolerance, ε_l is decreased as described above. For a relevant comparison, we compute the residual at the end of each iteration in PDAP and at the end of each step in the semi-smooth Newton method for $(\mathcal{P}^{n,\varepsilon}_{\beta,h})$. Note that, as for the previous example, we only take the computational time for the iterations of each Algorithms into account; the state and sensitivity equations are solved beforehand.

In the following we choose $N \in \{5, 15\}$ and consider the discretized design problems $(\mathcal{P}^{n,\varepsilon}_{\beta,h})$ and (5.63) on the grid $\mathcal{T}_{h_k}$ for levels $k \in \{5, 8\}$. Since $\mathcal{I}^N_0 = 0$ there holds $0 \notin \operatorname{supp} \psi^n_h$. As a consequence, we have to construct an initial iterate different from zero. To account for the different regularities of the solutions to the corresponding continuous problems, we choose the initial iterate u^1 for the solution of (5.63) as a linear combination of $(N+1)^2$ Dirac delta functions (located in nodes of the coarse grid) while the starting point $\bar{u}^1_\varepsilon \in V_h \subset L^2(\Omega_o)$ for the solution of $(P^{N,\varepsilon_1}_{\beta,h})$ is chosen as $\bar{u}^1_\varepsilon \equiv 1$. Observe that $r_F(\bar{u}^1_\varepsilon) \neq r_F(u^1)$. However, we stress that we are interested in a qualitative comparison of both algorithms rather than a quantitative one. The results can be found in Figure 5.10. First, we note that the runtime for both algorithms is affected by the increased number of support points for larger N. In fact, on grid level eight, we obtain $\# \operatorname{supp} \bar{u}^N_{\beta,h} = 58$ for $N = 5$ and $\# \operatorname{supp} \bar{u}^N_{\beta,h} = 630$ for $N = 15$, respectively. Clustering adjacent support points as described in Section 4.6.1, we obtain 30 and 240 clusters, respectively, and the post-processed solutions (as described in section 4.6.1) are given in Figure 5.11. On both grid levels we observe that the computation time for PDAP is affected more than the one for Algorithm 6 by the increased support size of the optimal design. This is a consequence of the different update strategies for the iterates in both algorithms. In each semi-smooth Newton step in Algorithm 6 the current iterate is updated globally on Ω_o. In contrast, at most one new support point is added in each iteration of PDAP. Hence, if the support of the optimal solution is increased, so is the number of necessary iterations in PDAP, explaining the increase of the computation time.

Let us now consider the influence of the number of grid points of the spatial discretization. Here, we observe that the path-following algorithm is affected more, which can be explained as follows:



(a) Residual for $N = 5$ on grid level five.

(b) Residual for $N = 5$ on grid level eight.

(c) Residual for $N = 15$ on grid level five.

(d) Residual for $N = 15$ on grid level eight.

Figure 5.10: Residuals $r_F(\cdot)$ for various number of parameters and discretizations plotted over computation time t in seconds for $\mathcal{I}_0^N = 0$.

For each $\varepsilon > 0$ the unique optimal solution to $(\mathcal{P}_{\beta,h}^{n,\varepsilon})$ is given by the component-wise projection formula

$$\bar{u}_{\beta,h}^{N,\varepsilon}(x_i) = \max\left\{ -\frac{1}{\varepsilon}(\nabla\psi_{l,h}^n(\bar{u}_{\beta,h}^{N,\varepsilon})(x_i) + \beta), 0 \right\} \quad \forall x_i \in \mathcal{N}_{h_k},$$

where $\psi_{l,h}^n(\bar{u}_{\beta,h}^{N,\varepsilon}) = \psi_{l,h}^n(\Lambda_h \bar{u}_{\beta,h}^{N,\varepsilon})$. This indicates that the set of nodes in the support of the solution depends on the fineness of the discretization. As a consequence, the path-following method can only exploit the increased sparsity in later iterations (for smaller ε), which leads to larger computational times on finer grids. In contrast, in PDAP we only need to calculate the gradient $\nabla\psi_h^n(u^k)$ as well as its maximum on the whole domain, while the dimension of the occurring sub-problems and thus also the size of the linear systems in the semi-smooth Newton method can be bounded independent of the discretization in every iteration. Together with the mesh-independence observations for the residual and the support size from Section 4.7.1 this explains the better scaling of the successive point insertion algorithm with respect to the number of nodes in the triangulation.

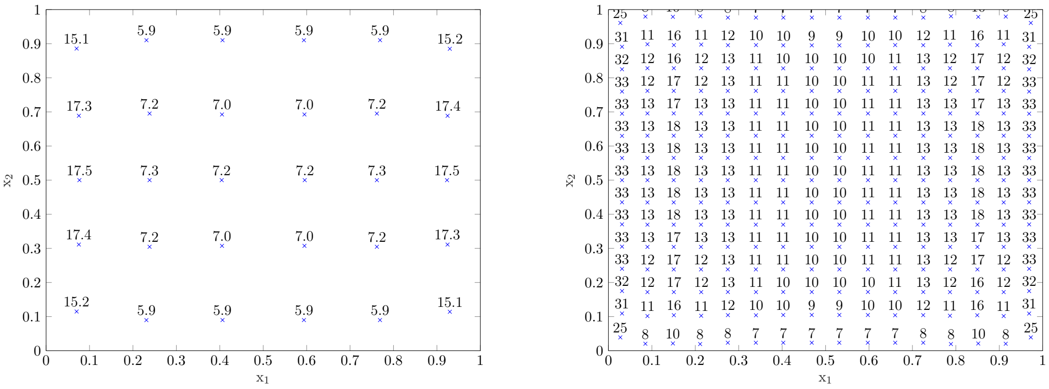


Figure 5.11: Optimal designs for $\mathcal{I}_0^N = 0$ and $N = 5$ (left) and $N = 15$ (right) on grid level eight.

Comparison of point-insertion and path-following: A priori knowledge

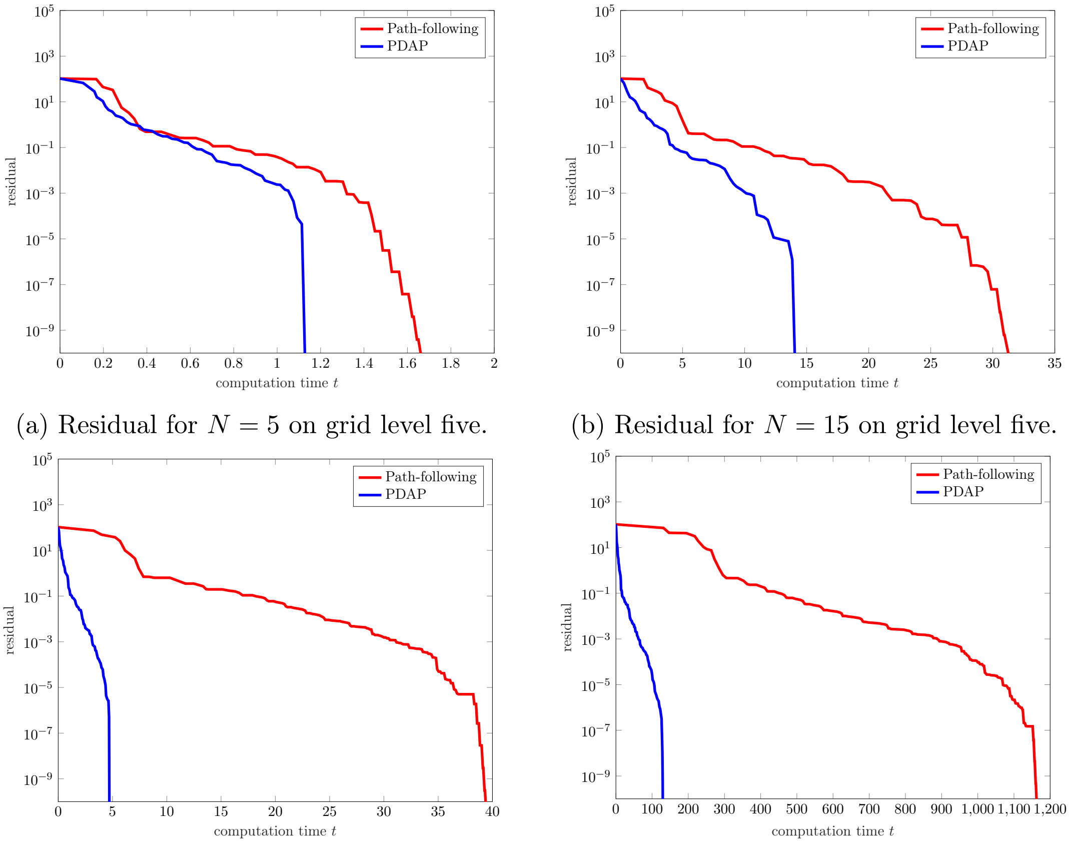
To conclude this section, we again consider the previous setup in the Bayesian setting. Concretely, we choose $\mathcal{I}_0^N \in \mathrm{PD}(N^2)$ according to (5.66) with $c_1 = 10^{-5}$, $c_2 = 10$ and $s = 2$.

Since $\mathcal{I}_0$ is positive definite we can choose the starting point for both algorithms as $u^1 = 0$. In Figure 5.12 the computed residuals for the path-following algorithm and PDAP are shown. For the path-following algorithm we again observe an increased computation time with respect to the spatial discretization in comparison to PDAP. Due to the positive definiteness of $\mathcal{I}_0$, the support of the solution is not bounded from below by $n = N^2$. Concretely, on grid level eight there holds $\# \operatorname{supp} \bar{u}_{\beta,h}^N = 26$ for $N = 5$ and $\# \operatorname{supp} \bar{u}_{\beta,h}^N = 38$ for $N = 15$, i.e. the number of optimal Dirac delta functions does not increase as significantly as in the case of $\mathcal{I}_0^N = 0$ for larger N. Consequently, we also observe a better behavior of the computation time for PDAP with respect to N. The corresponding optimal designs can be found in Figure 5.13. As in the first example, the displayed designs are obtained by the post-processing procedure described in Section 4.6.1, which leads to 10 and 18 connected clusters of the support for $N = 5$ and $N = 10$, respectively.

Accelerating Primal-Dual-Active-Point methods

In the previous sections we observed that PDAP scales well with respect to the spatial discretization while it does not scale well with respect to the support size of the optimal design. As discussed earlier, this is mainly caused by inserting only one point in every iteration. To remedy this defect, we implement the heuristic multiple point insertion strategy discussed in Section 5.3.2. More in detail, instead of only adding a global minimizer $\hat{x}^k$ of $\nabla\psi_h^n(u^k)$ in each iteration, we update the active set by adding the grid points corresponding to the $M \leq n(n+1)/2$ smallest local minimizers of the gradient. The upper bound on the number of inserted Dirac deltas ensures that the dimension of the sub-problems in PDAP stays uniformly bounded throughout the iterations. However, we note that in our numerical experiments this upper bound was never attained. The resulting algorithm will be referenced as Multi-PDAP in the following.

To compare the three algorithms we again consider the A-optimal design problem for the diffusion coefficient example on $\mathcal{T}_{h_8}$ with $N \in \{5, 15\}$. The cost parameter and a priori knowledge are chosen as $\beta = 1$ and $\mathcal{I}_0^N = 0$, respectively. The computed residuals over the computation time are plotted in Figure 5.14. We observe that the insertion of multiple points in each iteration



(a) Residual for $N = 5$ on grid level five.

(b) Residual for $N = 15$ on grid level five.

(c) Residual for $N = 5$ on grid level eight.

(d) Residual for $N = 15$ on grid level eight.

Figure 5.12: Residual $r_F(\cdot)$ for different N and discretizations plotted over computation time t in seconds for $\mathcal{I}_0^N$ given by the prior (5.66).

significantly improves the speed of convergence of the successive point insertion algorithm, which shows the practical efficiency of the proposed heuristic strategy. Finally, we again stress that all comparisons between the two implementations of PDAP and Algorithm 6 should not be understood quantitatively; the path-following algorithm may possibly be accelerated by, e.g., the inexact solution of the regularized sub-problems.

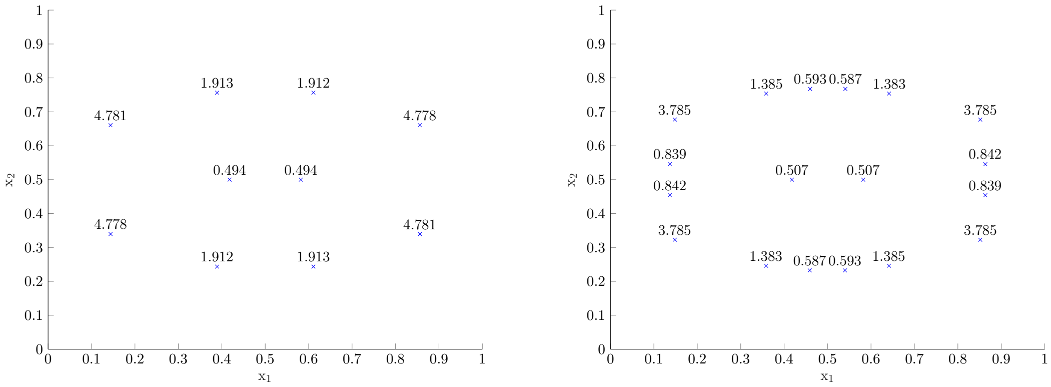
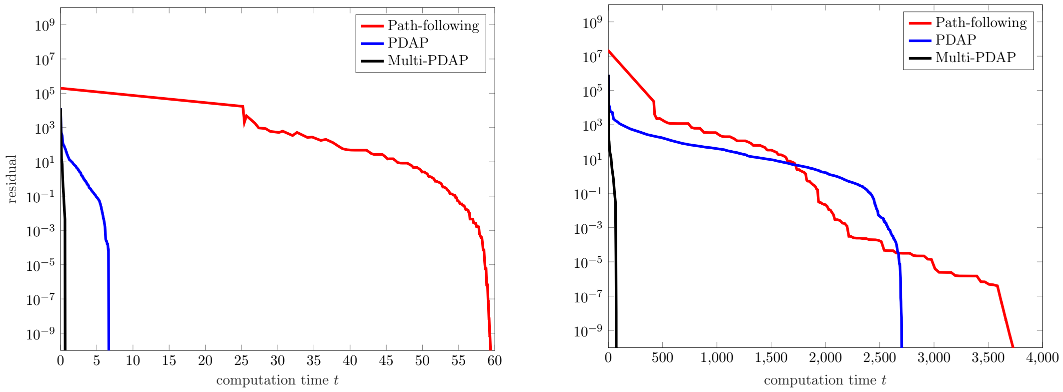


Figure 5.13: Optimal designs for $\mathcal{I}_0^N$ given by the prior (5.66) and $N = 5$ (left) and $N = 15$ (right) on grid level eight.



(a) Residual for $N = 5$ on level eight.

(b) Residual for $N = 15$ on level eight.

Figure 5.14: Evolution of the residual $r_F(\cdot)$ over the computation time t in seconds on grid level eight for different numbers of parameters.

6 Algorithmic framework

In the last part of this thesis we elaborate in greater detail on the numerical solution algorithm for the optimal sensor placement problems considered in Chapter 4 and Chapter 5. To this end we recall that the major challenges in this context are given by the non-smoothness of the objective functional and the non-reflexivity of the measure space. A first naive approach on its solution would be to consider numerical solution methods for the discretized problems i.e. we replace the space of Radon measures $\mathcal{M}(\Omega)$ by the space $\mathcal{M}_h$ of measures supported in the nodes of a grid. This reduces the problem to a finite dimensional one. While the resulting problems still remain non-smooth their efficient numerical solution can be realized by applying a large variety of well-studied algorithms. For examples we point out to semi-smooth Newton methods, [194], the fast iterative shrinkage-thresholding algorithm (FISTA), [26], and the alternating direction of multipliers method, [45]. However such reasoning harbours the danger of yielding *mesh dependent* solution methods. That is to say that while a particular algorithm may be efficient for the solution of the discrete problem associated to a fixed discretization parameter its convergence behaviour can critically depend on h.

To some extend the mesh dependent behaviour of a particular algorithm can stem back to the fact that its description may not remain meaningful on the space $\mathcal{M}(\Omega)$. For this reason we are interested in the derivation of iterative solution methods for the continuous problem on the space of Radon measures. Obviously function space based solution approaches are at first glance only of limited utility since the computation of a minimizer usually still requires a discretization of the problem. However adapting such methods to the discretized problems often yields algorithms with a *mesh independent* convergence behavior, [147, 162]. Thus while each step of the method may suffer from increasing computational complexity for decreasing h the number of iterations to fulfill a suitable convergence criterion is stable with respect to the discretization parameter.

The main goal of this chapter lies in the analysis of an efficient iterative numerical solution algorithm on the function space level. To this end the presentation is divided into two parts. Since the aforementioned difficulties are not restricted to the particular case of Radon measures we embed the considered sensor placement problems into the larger framework of composite minimization problems

$$\min_{u \in \mathcal{M}} [f(u) + g(u)].$$

Here we minimize the sum of a differentiable function f and a convex but not necessarily smooth regularizer g over a possibly non-reflexive Banach space $\mathcal{M}$. Similar problems have received tremendous attention in the context of optimal control and inverse problems over the last decades. This is owed to the fact that the right choice of the space and the nonsmooth regularizer enhances desirable structural features in its minimizers. Given a spatial domain Ω we refer e.g. to the broadly discussed topic of sparse regularization for $\mathcal{M} = \mathcal{M}(\Omega)$, [50], the bang-bang structure of minimizers in the case of $\mathcal{M} = L^\infty(\Omega)$, [72], and the staircaising effect for functions of bounded total variation $\mathcal{M} = \mathrm{BV}(\Omega)$, [227]. However this comes at the price of having to deal with spaces $\mathcal{M}$ lacking many desirable properties for the analysis of the problem. For example we stress that all of the previously stated spaces are non-reflexive, not strictly convex and nonsmooth. In particular this makes their function space based solution highly challenging since many well-known algorithms do not yield extensions to such problems. Moreover we highlight that the unit ball in those spaces is neither compact with respect to the strong nor the weak topology. For iterative solution approaches to the problem this raises the question if and in which sense convergence of the generated iterates can be expected. As a remedy we point out that all of the mentioned spaces admit an interpretation as the topological dual space of a separable Banach space $\mathcal{C}$. This allows to tackle these problems by resorting to weaker topological concepts. Thus we may restrict ourselves to this type of spaces without loosing much of the desired generality. Note that these considerations once again underline the additional care that has to be taken when discussing algorithms on infinite dimensional spaces.

In the first part of this chapter we demonstrate that the described class of composite minimization problems can be solved by an adapted version of the conditional gradient method, [112]. In more detail the method is based on the iterative solution of partially linearized subproblems

$$u^{k+1} = u^k + s^k(v^k - u^k), \quad v^k \in \operatorname*{arg\,min}_{v \in \mathcal{M}} [\langle \nabla f(u^k), v \rangle + g(v)], \quad s^k \in [0,1],$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $\mathcal{M}$ and its predual space $\mathcal{C}$. We show that this generalized conditional gradient iteration is indeed suitable for the solution of composite minimization problems and yields provable qualitative and quantitative convergence guarantees under mild assumptions. Several instructive examples highlight the simplicity of the method and point out to possible applications.

In the second part of the chapter the presented algorithm is applied to sparse minimization problems

$$\min_{u \in \mathcal{M}_{ad}} [F(Ku) + G(\|u\|_{\mathcal{M}})]. \tag{6.1}$$

The optimization variable is searched for in a subset $\mathcal{M}_{ad}$ of $\mathcal{M}(\Omega, H)$, the space of Borel measures which assume values in a Hilbert space H on a set Ω. Here K denotes a linear continuous operator. For example it may be given as the solution mapping associated to a linear equation or, as in the previous chapters, we identify it with the Fisher operator $\mathcal{I}$. The functional F is a scalar-valued, smooth and not necessarily convex function, while G is a, in general, nonsmooth but convex function acting on the total variation of u. Regularization terms of this particular form are known to favor optimal solutions which are sparse i.e they are zero outside of a Lebesgue null set. This observation makes measure-valued optimization variables appealing for a wide range of applications. Besides the sensor placement framework developed in this thesis we point out to actuator placement problems, [74], acoustic inversion, [32, 209], and super-resolution, [55, 95].

Function space based solution methods for this type of problems can be founded on *path-following* strategies in order to circumvent the non-reflexivity of the space $\mathcal{M}(\Omega, H)$. Here the original problem is replaced by a sequence of regularized ones

$$\min_{u \in \mathcal{M}_{ad}} [F(Ku) + G(\|u\|_{L^1(\Omega,H)}) + \frac{\varepsilon}{2}\|u\|^2_{L^2(\Omega,H)}], \tag{6.2}$$

over the Hilbert space $L^2(\Omega, H)$. Note that the appearance of the $L^1(\Omega, H)$ norm in the objective functional still promotes optimal solutions which are nonzero only on small subsets of Ω. Furthermore in the limiting case for $\varepsilon \to 0$ the regularized solutions approximate solutions to the original one. We point out, e.g., to [208] for a reference. For fixed $\varepsilon > 0$ those problems are amenable for efficient function space based solution methods such as semi-smooth Newton, [139, 248], or proximal-type methods, [233,234]. However the convergence behavior of these algorithms may deteriorate for small values of ε. In the practical realization it is therefore necessary to start at a large value of ε. A solution to the original problem is then obtained by alternating between decreasing the regularization parameter and a possibly inexact solution of the corresponding regularized problem using the previous iterate as a warmstart. Thus a complete analysis of path-following methods requires a rigorous convergence analysis of the method used for the solution of the regularized problem in dependence of ε , a quantification of the additional regularization error and sophisticated update strategies for the parameter.

In contrast we base the algorithmic solution of sparse minimization problems on the presented generalized conditional gradient method. Its application does not require an additional regularization of the problem. It turns out that this method computes a minimizer by sequentially adding new Dirac delta functions, i.e. measures supported on a single point, to the current iterate. Thus it yields measure-valued iterates supported on finitely many points. While its implementation is fairly easy it generally suffers from the characteristic slow convergence behavior of first-order optimization methods. This also prohibits a solution of the problem to high precision.

We emphasize that the idea of using sequential point insertion algorithms for sparse minimization is not new. For an overview of previous works in this direction we point out to Section 6.3. In this context we also refer to optimization problems with regularizers promoting sparsity of solutions in a given basis see e.g. [47,81]. A standard tool for the algorithmic solution of this type of problems are iterative shrinkage algorithms [48]. In [49] the authors identify this procedure as a special case of the generalized conditional gradient method on the problem.

The main novelty of the present work is the analysis of an accelerated version of the conditional gradient method based on alternating between point insertion and coefficient optimization steps. We show that under additional structural assumptions on the problem, cf. also the notion of non-degeneracy in [95], this improved version yields a linear rate of convergence for the objective function values as well as the iterates in a suitable dual norm. To the best of our knowledge we are not aware of any comparable results.

6.1 Problem setting

Throughout the course of this chapter we consider the following composite minimization problem

$$\min_{u\in\mathcal{M}} j(u) := [f(u) + g(u)]. \tag{$\mathfrak{P}$}$$

Here the function f will in general be non-convex but smooth in a sense made clear below while g is convex but typically non-differentiable. The optimization variable u is searched for in a Banach space $\mathcal{M}$. It is given by the topological dual space of a separable Banach space $\mathcal{C}$. We will refer to $\mathcal{C}$ as the predual space of $\mathcal{M}$. The norm on $\mathcal{C}$ is denoted by $\|\cdot\|_{\mathcal{C}}$. In general the space $\mathcal{C}$ will be non-reflexive i.e. $\mathcal{C} \subsetneq \mathcal{M}^*$. The corresponding duality pairing between $\varphi \in \mathcal{C}$ and $u \in \mathcal{M}$ is denoted by $\langle \varphi, u\rangle = \langle \varphi, u\rangle_{\mathcal{C},\mathcal{M}}$. Furthermore we recall the concept of weak*-convergence on $\mathcal{M}$.

Definition 6.1. A sequence $\{u_k\}_{k\in\mathbb{N}} \subset \mathcal{M}$ is called weak* convergent with limit $\bar{u} \in \mathcal{M}$ if

$$\langle \varphi, u_k\rangle \to \langle \varphi, \bar{u}\rangle \quad \forall \varphi \in \mathcal{C}.$$

Whenever $\{u_k\}_{k\in\mathbb{N}}$ converges weak* to $\bar{u} \in \mathcal{M}$ it is denoted by $u_k \rightharpoonup^* \bar{u}$.

The space $\mathcal{M}$ is equipped with the topology induced by the corresponding dual norm

$$\|u\|_{\mathcal{M}} = \sup_{\varphi\in\mathcal{C}, \|\varphi\|_{\mathcal{C}}\leq 1} \langle u, \varphi\rangle \quad \forall u \in \mathcal{M}.$$

In particular it is a Banach space with respect to the induced norm. Given an extended real valued functional $\phi\colon \mathcal{M} \to \mathbb{R} \cup \{+\infty\}$ and a convex weak* closed subset $M \subset \mathcal{M}$ we define the domain of ϕ in M as

$$\operatorname{dom}_M \phi = \{ u \in M \mid \phi(u) < \infty \}.$$

If $M = \mathcal{M}$ the index will be dropped.

6.1.1 Existence of minimizers

The proof for the existence of at least one minimizer to ($\mathfrak{P}$) will be based on Tonelli's direct method, see e.g. [78, Chap. 1]. Thus we require the relative sequential compactness of bounded sets in $\mathcal{M}$ with respect to a suitable topology. For the case of a general non-reflexive space this is neither true for the strong topology, i.e. the topology induced by the norm on $\mathcal{M}$, nor the weak topology. As a remedy we recall the following sequential version of the Banach-Alaoglu theorem, cf. [52, Corollary 3.30] which holds due to the separability of the predual space.

Proposition 6.1 (Banach-Alaoglu). *Let $\{u_k\}_{k\in\mathbb{N}} \subset \mathcal{M}$ denote a bounded sequence in $\mathcal{M}$. Then there exists a subsequence $\{u_{k_j}\}_{j\in\mathbb{N}}$ and an element $u \in \mathcal{M}$ with $u_{k_j} \rightharpoonup^* u$.*

Thus norm bounded sets in $\mathcal{M}$ are relative sequentially compact with respect to the weak* topology. Throughout the course of this chapter we impose the following general assumptions on the objective functional under consideration.

Assumption 6.1. The functions f and g fulfill:

A6.1 The extended real-valued functional $g\colon \mathcal{M} \to \mathbb{R} \cup \{+\infty\}$ is proper, i.e. not equal to $+\infty$, convex and (sequentially) weak* lower semi-continuous on $\mathcal{M}$.

A6.2 The extended real-valued function $f\colon \mathcal{M} \mapsto \mathbb{R} \cup \{+\infty\}$ is proper and (sequentially) weak* lower semi-continuous on $\operatorname{dom} g$. There holds

$$\operatorname{dom} j = \operatorname{dom} f \cap \operatorname{dom} g \neq \emptyset,$$

and for every sequence $\{u_k\}_{k\in\mathbb{N}} \subset \operatorname{dom} g$ we have

$$u_k \rightharpoonup^* u \Rightarrow f(u) \leq \liminf_{k\to\infty} f(u_k).$$

Furthermore, restricted to its domain, f is (sequentially) weak* continuous. Given a sequence $\{u_k\}_{k\in\mathbb{N}} \subset \operatorname{dom} f$ there holds

$$u_k \rightharpoonup^* u \in \operatorname{dom} f \Rightarrow f(u_k) \to f(u).$$

A6.3 The domain of j is (sequentially) weak* open in $\operatorname{dom} g$ in the following sense: Given $\{ u_k \}_{k\in\mathbb{N}} \subset \operatorname{dom} g$ there holds

$$u_k \rightharpoonup^* u \in \operatorname{dom} j \Rightarrow \exists \bar{k} \in K \colon u_k \in \operatorname{dom} j \quad \forall k \geq \bar{k}.$$

A6.4 On $\operatorname{dom} j$ the function f is assumed to be Gâteaux-differentiable. For every $u \in \operatorname{dom} j$ the Gâteaux derivative $f'(u)(\cdot)$ of f can be identified with $\nabla f(u) \in \mathcal{C}$, i.e. there holds

$$f'(u)(\delta u) = \langle \nabla f(u), \delta u \rangle \quad \forall \delta u \in \mathcal{M}.$$

Furthermore the mapping

$$\nabla f \colon \mathcal{M} \to \mathcal{C} \quad u \mapsto \nabla f(u)$$

is (sequentially) weak*-to-strong continuous.

A6.5 The functional $j \colon \mathcal{M} \to \mathbb{R} \cup \{+\infty\}$ is radially unbounded. For every sequence $\{u_k\}_{k\in\mathbb{N}} \subset \operatorname{dom} g$ there holds

$$\|u_k\|_{\mathcal{M}} \to \infty \Rightarrow j(u_k) \to +\infty.$$

The existence of a global minimizer to ($\mathfrak{P}$) follows by standard arguments.

Proposition 6.2. *There exists at least one optimal solution to* ($\mathfrak{P}$). *Moreover the set of optimal solutions is bounded.*

Proof. Since j is proper we have

$$\hat{\jmath} = \inf_{u\in\mathcal{M}} j(u) < +\infty.$$

Denote by $u_k \subset \operatorname{dom} j$, $k \in \mathbb{N}$, an arbitrary infimizing sequence for j. For all k large enough we have

$$\hat{\jmath} = \inf_{u\in\mathcal{M}} j(u) \leq j(u_k) \leq \hat{\jmath} + 1.$$

Due to the radial unboundedness of j there exists a constant $c > 0$ with $\|u_k\|_{\mathcal{M}} \leq c$ for all $k \in \mathbb{N}$. By Proposition 6.1 there exists a subsequence of u_k (denoted with the same index) and an element $\bar{u} \in \operatorname{dom} g$ with $u_k \rightharpoonup^* \bar{u}$. By assumption g is weak* lower semi-continuous on $\mathcal{M}$ and f is weak* lower semi-continuous on $\operatorname{dom} g$. Since $\{u_k\}_{k\in\mathbb{N}} \subset \operatorname{dom} g$ we arrive at

$$j(\bar{u}) \leq \liminf_{k\to\infty} j(u_k) = \hat{\jmath},$$

from which we conclude $\hat{\jmath} \in \mathbb{R}$ and the optimality of $\bar{u} \in \operatorname{dom} j$. It remains to show the boundedness of the set of minimizers. To do so assume the contrary i.e. there exists a sequence $\{u_k\}_{k\in\mathbb{N}}$ with

$$j(u_k) = \hat{\jmath} \quad \forall k \in \mathbb{N}, \quad \|u_k\|_{\mathcal{M}} \to \infty.$$

This however contradicts the radial unboundedness of j. Thus the set of minimizers to ($\mathfrak{P}$) is bounded. □

6.1.2 Optimality conditions

The aim of this section is to establish first-order necessary optimality conditions for $(\mathfrak{P})$. Since f is assumed to be smooth in the sense of **A6.4** and g is convex we can therefore mainly rely on well-known results from convex analysis and non-linear functional analysis. Associated to the convex functional g we introduce its subdifferential at a point $u \in \mathcal{M}$ by

$$\partial g(u) = \{ \varphi \in \mathcal{C} \mid \langle \varphi, \tilde{u} - u \rangle + g(u) \leq g(\tilde{u}) \quad \forall \tilde{u} \in \mathcal{M} \}, \tag{6.3}$$

At this point we briefly pause to point out that the convex subdifferential of g is defined as subset of the predual space $\mathcal{C}$. This is in contrast to its usual definition as a subset of the dual space $\mathcal{M}^*$ (formed with respect to the norm topology on $\mathcal{M}$). In particular the set $\partial g(u)$ may be empty for an arbitrary $u \in \mathcal{M}$. The following proposition however states that the subdifferential at a minimizer of $(\mathfrak{P})$ necessarily contains the negative gradient of f.

Proposition 6.3. *Let $\bar{u}$ be a minimizer to $(\mathfrak{P})$. Then there holds*

$$\langle -\nabla f(\bar{u}), u - \bar{u} \rangle + g(\bar{u}) \leq g(u) \quad \forall u \in \mathcal{M}. \tag{6.4}$$

Equivalently, this can be expressed by $-\nabla f(\bar{u}) \in \partial g(\bar{u})$. Vice versa, if f is convex, every $\bar{u} \in \operatorname{dom} j$ which fulfils (6.4) is a global minimizer of $(\mathfrak{P})$.

Proof. We give the proof for the sake of completeness. Let $\bar{u} \in \mathcal{M}$ be an optimal solution to $(\mathfrak{P})$. First we note that (6.4) holds trivially if $u \notin \operatorname{dom} g$. Now, given an arbitrary $u \in \operatorname{dom} g$ we have $\bar{u} + t(u - \bar{u}) \in \operatorname{dom} j$ for all positive t small enough due to **A6.3**. Using the optimality of $\bar{u}$ and the convexity of g we obtain

$$0 \leq j\left(\bar{u} + t(u - \bar{u})\right) - j(\bar{u}) \leq f\left(\bar{u} + t(u - \bar{u})\right) - f(\bar{u}) + t\left(g(u) - g(\bar{u})\right).$$

Dividing both sides of the inequality by t and letting $t \to 0$ yields

$$0 \leq f'(\bar{u})(u - \bar{u}) + g(u) - g(\bar{u}).$$

By rearranging and $f'(\bar{u})(\bar{u} - u) = \langle \nabla f(\bar{u}), \bar{u} - u \rangle$, see **A6.4**, we conclude (6.4) since $u \in \operatorname{dom} g$ was chosen arbitrary. Due to the definition of the subdifferential this is equivalent to $-\nabla f(\bar{u}) \in \partial g(\bar{u})$.

Assume now that (6.4) holds at $\bar{u}$ and f is convex. Then we have

$$0 \leq \langle \nabla f(\bar{u}), u - \bar{u} \rangle + g(u) - g(\bar{u}) \leq f(u) - f(\bar{u}) + g(u) - g(\bar{u}) = j(u) - j(\bar{u}) \quad \forall u \in \mathcal{M},$$

which yields the optimality of $\bar{u}$. □

While (6.4) provides a simple necessary condition for the optimality of $\bar{u}$ it is only of limited practical use without any further characterization of the set $\partial g(\bar{u})$. In particular it does not allow to infer on the structural properties of minimizers to $(\mathfrak{P})$. In the following proposition we provide an important result from convex analysis which allows to characterize the subdifferential of g by properties of the convex conjugate function

$$g^* \colon \mathcal{C} \to \mathbb{R} \cup \{ +\infty \}, \quad \varphi \mapsto \sup_{u \in \mathcal{M}} [\langle \varphi, u \rangle - g(u)]. \tag{6.5}$$

Proposition 6.4. *Let $\varphi \in \mathcal{C}$ and $u \in \mathcal{M}$ be given. Then there holds*

$$\varphi \in \partial g(u) \Leftrightarrow g(u) + g^*(\varphi) = \langle p, u \rangle \Leftrightarrow u \in \partial g^*(\varphi), \tag{6.6}$$

where the subdifferential of g^* at φ is given by

$$\partial g^*(\varphi) = \{\, u \in \mathcal{M} \mid \langle u, \tilde{\varphi} - \varphi \rangle + g^*(\varphi) \leq g(\tilde{\varphi}) \quad \forall \tilde{\varphi} \in \mathcal{C} \,\}.$$

Proof. The statement is obtained from Proposition 5.1 and Corollary 5.2 in [98, Chapter 1] noting that g is weak* lower semi-continuous. □

Corollary 6.5. *Let $\bar{u} \in \mathcal{M}$ be an optimal solution to $(\mathfrak{P})$. Then there holds*

$$-\nabla f(\bar{u}) \in \partial g(\bar{u}) \Leftrightarrow g(\bar{u}) + g^*(-\nabla f(\bar{u})) = \langle -\nabla f(\bar{u}), \bar{u} \rangle \Leftrightarrow \bar{u} \in \partial g^*(-\nabla f(\bar{u}))$$

Proof. The statement readily follows by combining Proposition 6.3 and Proposition 6.4. □

In the following example, we illustrate how these results allow to derive equivalent first-order optimality conditions.

Example 6.1. *Let $G\colon \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be proper, convex, lower semi-continuous and monotonically increasing on $\mathbb{R}_+$ with $\lim_{t\to\infty} G(t) = +\infty$. Further assume that $\operatorname{dom} G \subset \mathbb{R}_+$. We set $g(u) = G(\|u\|_{\mathcal{M}})$. This setting includes the case of norm regularization $g_1(u) = G_1(\|u\|_{\mathcal{M}}) = \alpha\|u\|_{\mathcal{M}}$ where we choose $G_1(m) = \alpha m + I_{[0,\infty)}(m)$ for $\alpha > 0$. The, at first sight unnecessary, indicator function of the nonnegative real axis, will allow for a simpler statement of first order optimality conditions. We stress however that its appearance does not change the optimization problem. Additionally, norm constraints can be considered by setting*

$$g_2(u) = I_{\|\cdot\|_{\mathcal{M}} \leq M_0}(u) = I_{[0,M_0]}(\|u\|_{\mathcal{M}}).$$

Here $I_{\|\cdot\|_{\mathcal{M}} \leq M_0}$ denotes the indicator function of the ball $\bar{B}_{M_0}(0)$ with radius $M_0 > 0$ in $\mathcal{M}$, i.e.

$$I_{\|\cdot\|_{\mathcal{M}} \leq M_0}(u) = \begin{cases} 0 & \|u\|_{\mathcal{M}} \leq M_0 \\ +\infty & \|u\|_{\mathcal{M}} > M_0 \end{cases}.$$

It is straightforward to verify that g is proper, convex and sequentially weak lower semi-continuous on $\mathcal{M}$. If $\bar{u} \in \mathcal{M}$ is an optimal solution to $(\mathfrak{P})$ then Proposition 6.3 yields*

$$\langle -\nabla f(\bar{u}), u - \bar{u} \rangle + G(\|\bar{u}\|_{\mathcal{M}}) \leq G(\|u\|_{\mathcal{M}}) \quad \forall u \in \mathcal{M}. \tag{6.7}$$

Let us first assume that $\bar{u} \neq 0$. Due to the monotonicity of G we immediately derive

$$\langle -\nabla f(\bar{u}), u - \bar{u} \rangle \leq 0 \quad \forall u \in \mathcal{M},\ \|u\|_{\mathcal{M}} \leq \|\bar{u}\|_{\mathcal{M}},$$

or, equivalently $-\nabla f(\bar{u}) \in \partial\left(I_{\|\cdot\|_{\mathcal{M}} \leq \|\bar{u}\|_{\mathcal{M}}}\right)(\bar{u})$. Let us calculate the convex conjugate of the indicator function as

$$\left(I_{\|\cdot\|_{\mathcal{M}} \leq \|\bar{u}\|_{\mathcal{M}}}\right)^*(\varphi) = \sup_{\|u\|_{\mathcal{M}} \leq \|\bar{u}\|_{\mathcal{M}}} \langle \varphi, u \rangle = \|\bar{u}\|_{\mathcal{M}} \|\varphi\|_{\mathcal{C}} \quad \forall \varphi \in \mathcal{C}.$$

We conclude $\|\bar{u}\|_{\mathcal{M}} \|\nabla f(\bar{u})\|_{\mathcal{C}} = \langle -\nabla f(\bar{u}), \bar{u} \rangle$. *Consequently, testing* (6.7) *with* $u_m = (m/\|\bar{u}\|_{\mathcal{M}})\bar{u}$ *for* $m \in \mathbb{R}_+$ *yields*

$$\|\nabla f(\bar{u})\|_{\mathcal{C}} \, (m - \|\bar{u}\|_{\mathcal{M}}) + G(\|\bar{u}\|_{\mathcal{M}}) \leq G(m) \quad \forall m \in \mathbb{R}_+.$$

Collecting all the previous results we get that every non-zero minimizer $\bar{u}$ *to* $(\mathfrak{P})$ *fulfills*

$$\langle -\nabla f(\bar{u}), \bar{u} \rangle = \|\nabla f(\bar{u})\|_{\mathcal{C}} \|\bar{u}\|_{\mathcal{M}}, \quad \|\nabla f(\bar{u})\|_{\mathcal{C}} \in \partial G(\|\bar{u}\|_{\mathcal{M}}),$$

where $\partial G(\|\bar{u}\|_{\mathcal{M}})$ *denotes the convex subdifferential of* G *at* $\|\bar{u}\|_{\mathcal{M}}$. *If* $\bar{u} = 0$ *the inequality in* (6.7) *simplifies to*

$$\langle -\nabla f(0), u \rangle + G(0) \leq G(\|u\|_{\mathcal{M}}) \quad \forall u \in \mathcal{M}. \tag{6.8}$$

Consider an arbitrary but fixed $u \in \mathcal{M}$, $\|u\|_{\mathcal{M}} = 1$, *and* $m \in \mathbb{R}_+$. *Testing* (6.8) *with* mu *yields*

$$m\langle -\nabla f(0), u \rangle + G(0) \leq G(m) \quad \forall m \in \mathbb{R}_+.$$

Since u *was chosen arbitrary we can take the supremum over* $u \in \mathcal{M}$, $\|u\|_{\mathcal{M}} = 1$, *on both sides of the inequality to arrive at*

$$m\|\nabla f(0)\|_{\mathcal{C}} + G(0) \leq G(m) \quad \forall m \in \mathbb{R}_+,$$

where we recall the dual representation of the norm on $\mathcal{C}$.

$$\|\varphi\|_{\mathcal{C}} = \sup_{\|u\|_{\mathcal{M}}=1} \langle \varphi, u \rangle \quad \forall \varphi \in \mathcal{C}.$$

Thus we conclude

$$0 = \langle -\nabla f(\bar{u}), \bar{u} \rangle = \|\nabla f(\bar{u})\|_{\mathcal{C}} \|\bar{u}\|_{\mathcal{M}}, \quad \|\nabla f(\bar{u})\|_{\mathcal{C}} \in \partial G(\|\bar{u}\|_{\mathcal{M}}).$$

6.2 Generalized conditional gradient methods

In this section we elaborate on the algorithmic solution of $(\mathfrak{P})$ on the function space level by generalized conditional gradient methods. To this end we first review the results on the original conditional gradient method for constrained minimization problems. Subsequently the method is adapted to general composite minimization problems. We discuss the convergence of this generalized algorithm and show that its worst-case convergence guarantees are on par with those of the original method.

6.2.1 Conditional gradient methods for smooth functions

To motivate our course of action in the following sections let us first consider the minimization of a smooth and convex function $f \colon \mathbb{R}^n \to \mathbb{R}$, $n \in \mathbb{N}$, with Lipschitz continuous gradient over a convex and compact subset $M \subset \mathbb{R}^n$. Obviously this problem can be fit into the general setting considered in the previous section by choosing the spaces as $\mathcal{M} \simeq \mathcal{C} = \mathbb{R}^n$ together with the euclidean norm and the nonsmooth function g as the indicator function I_M of the compact set. Thus we arrive at

$$\min_{u\in M} f(u) = \min_{u\in\mathbb{R}^n} j(u) = [f(u) + I_M(u)]. \tag{6.9}$$

In this case existence of a minimizer $\bar{u} \in M$ follows due to the Weierstrass theorem. The necessary and sufficient optimality condition is given by

$$(\nabla f(\bar{u}), u - \bar{u})_{\mathbb{R}^n} \geq 0 \quad \forall u \in M.$$

For the sake of simplicity we assume the uniqueness of $\bar{u}$ throughout this introductory remarks. The algorithmic solution of constrained optimization problems with smooth objective functional is a well-studied subject. We refer e.g. to the monographs [35, 203]. In particular the minimization problem in (6.9) can be solved by applying a projected gradient iteration defined by

$$u^0 \in M, \quad u^{k+1} = P_M\left(u^k - \frac{1}{L}\nabla f(u^k)\right) \quad \text{where} \quad P_M(u) = \min_{v\in M} \frac{1}{2}|u-v|^2_{\mathbb{R}^n}. \tag{6.10}$$

Here L denotes the Lipschitz constant of ∇f. Note that the sequence of iterates is feasible, i.e. $\{u^k\}_{k\in\mathbb{N}} \subset M$, by definition of the projection. It is well known that the iterates $\{u^k\}_{k\in\mathbb{N}}$ define a minimizing sequence for f on M and the objective function values converge at a *sublinear* rate

$$j(u^k) - j(\bar{u}) \leq \frac{c}{1+qk} \quad \forall k \in \mathbb{N},$$

for some constants $c,\ q > 0$ independent of the iteration number. If f is strongly convex on M the convergence is *linear* i.e.

$$j(u^k) - j(\bar{u}) \leq c\zeta^k \quad \forall k \in \mathbb{N},$$

for some $\zeta \in (0,1)$. Even for non-strongly convex f a convergence rate of $1/k^2$ can be recovered by adding additional improvement steps such as Nesterov acceleration [201, Chapter 2]. For a discussion of projected gradient methods in the broader context of general Hilbert spaces we refer to [34, 121].

As an alternative to projected gradient methods we consider a conditional gradient iteration defined by

$$u^0 \in M, \quad u^{k+1} = u^k + s^k(v^k - u^k), \quad v^k \in \operatorname*{arg\,min}_{v\in\mathcal{M}}\ (\nabla f(u^k), v)_{\mathbb{R}^d}, \quad s^k \in [0,1]. \tag{6.11}$$

This method was originally proposed in a paper by Frank and Wolfe, [112], for the minimization of a quadratic function over a polytope. The term conditional gradient method was coined in [185]. Feasibility of the iterates is ensured by taking u^{k+1} as a convex combination between the previous iterate u^k and an auxiliary variable v^k given by a minimizer to a linear program over M. A sublinear rate for the convergence of the sequence $\{f(u^k)\}_{k\in\mathbb{N}}$ towards the global minimum of f on M can be proven for various choices of the step size s^k. We mention for example the closed loop step sizes rule of [91]

$$s^k = \begin{cases} 0 & (\nabla f(u^k), u^k - v^k)_{\mathbb{R}^n} = 0 \\ \frac{(\nabla f(u^k), u^k - v^k)_{\mathbb{R}^n}}{L\|u^k - v^k\|^2} & 0 < \frac{(\nabla f(u^k), u^k - v^k)_{\mathbb{R}^n}}{L\|u^k - v^k\|^2} < 1 \\ 1 & 1 \leq \frac{(\nabla f(u^k), u^k - v^k)_{\mathbb{R}^n}}{L\|u^k - v^k\|^2} \end{cases}, \tag{6.12}$$

and implicit step sizes based on line minimization

$$s^k \in \operatorname*{arg\,min}_{s\in[0,1]} f(u_s^k) \quad s.t. \quad u_s^k = u^k + s(v^k - u^k), \tag{6.13}$$

or an Armijo-Goldstein backtracking on the objective functional, [92]. The method is also known to converge for open loop step size sequences, [93, 159, 274], fulfilling

$$s^k \to 0, \quad \sum_{i=1}^{\infty} s^k = +\infty, \tag{6.14}$$

whose determination may neither require evaluations of the objective functional f nor the Lipschitz constant of its gradient ∇f on M. In particular this covers the choice of $s^k = 2/(k+2)$.

However in contrast to projected gradient methods the sublinear rate is tight even for strongly convex f. For a reference we point out to the example in [56]. Stronger convergence results can only be expected under more restrictive assumptions on the geometry of the admissible set, the function f and/or the location of the minimizer. We give a brief overview in the following. In [125] the authors establish linear convergence if f is strongly convex and $\bar{u}$ lies in the interior of M. A similar result is derived in [25] for a conditional gradient method applied to convex feasibility problems over a general convex and compact set M. Moreover a linear rate of convergence is provided in [83, 185] for convex f on strongly convex sets if the norm of $\nabla f(u)$ is uniformly bounded away from zero for $u \in M$. Here M is called strongly convex with respect to $|\cdot|_{\mathbb{R}^n}$ if there exists $\theta > 0$ with

$$u_1,\ u_2 \in M,\ u_3 \in \mathbb{R}^n,\ |u_3|_{\mathbb{R}^n} = 1,\ s \in [0,1] \Rightarrow su_1 + (1-s)u_2 + s(1-s) + \frac{\theta}{2}|u_1 - u_2|^2_{\mathbb{R}^n} u_3 \in M.$$

For example the unit ball with respect to the euclidean norm is strongly convex. In several papers, [91, 92], the assumptions on f and the admissible set are replaced by a growth condition on the linear functional induced by the optimal gradient

$$\operatorname*{arg\,min}_{v\in M} (\nabla f(\bar{u}), v)_{\mathbb{R}^n} = \{\bar{u}\}, \quad (\nabla f(\bar{u}), u - \bar{u})_{\mathbb{R}^n} \geq \theta |u - \bar{u}|^2_{\mathbb{R}^n}. \tag{6.15}$$

Note that these works do neither require strong convexity of the function f nor of the set M. If M is a polytope the first condition together with the fundamental theorem of linear programming implies that $\bar{u}$ is a vertex. One interesting and relevant application of those results is illustrated in Example 6.2. More recently a $1/k^2$ rate independent of the location of the minimizer was obtained in [115] for strongly convex f and M. We emphasize that all of these improved results are either based on closed loop step size choices, (6.12), or implicit step sizes, (6.13), which take into account information on the current iterate. For the general open loop step size rule (6.14) no improved results can be expected even in the outlined restrictive settings.

A second line of work puts the focus on acceleration schemes in order to improve the convergence behavior of conditional gradient methods. We also give a brief sketch of these approaches. To give some geometric interpretation for the following discussion let us assume that the admissible set is a polytope and $0 \in M$. Then M can be represented as the convex hull of the finite set $\mathcal{A}$ containing its vertices. Thus we can restate the minimization problem as

$$\min_{u\in\mathbb{R}^n} f(u) \quad s.t. \quad u \in \operatorname{conv}(\mathcal{A}) = \left\{ \sum_{v\in\mathcal{A}} \lambda_v v \mid \sum_{v\in\mathcal{A}} \lambda_v = 1,\ \lambda_v \geq 0 \right\}.$$

From linear programming theory it is well-known that a linear functional attains its minimum on a compact and convex polytope at a vertex. As a consequence we can choose

$$v^k \in \operatorname*{arg\,min}_{v \in \mathbb{R}^n} (\nabla f(u^k), v)_{\mathbb{R}^n} \cap \mathcal{A}.$$

In particular if we set $u^0 = 0$ there exists a set of vertices $\mathcal{A}_k$ with

$$\mathcal{A}_k \subset \{v^i\}_{i=1}^k \subset \mathcal{A}, \quad u^{k+1} \in \operatorname{conv}\{\mathcal{A}_k\}.$$

From this perspective the conditional gradient method can be interpreted as follows. In each iteration we select one vertex v^k of the polytope M. The new iterate is then found by moving from u^k towards v^k along the connecting line. By construction u^{k+1} lies in the simplex spanned by a subset $\mathcal{A}_k$ of the previously determined vertices $\{v^i\}_{i=1}^k$. Acceleration schemes can now be based on the vertex representation of u^k. In [268] Wolfe proposed to add the possibility of performing an alternative step

$$v^k \in \operatorname*{arg\,max}_{v \in \mathcal{A}_{k-1}} (\nabla f(u^k), v), \quad u^{k+1} = u^k + s^k(u^k - v^k), \quad s^k \geq 0,$$

instead of the conditional gradient update which allows the iterate to move away from previously considered vertices. Linear convergence of conditional gradient methods based on Wolfe's away step is discussed in [2,114,179]. Obviously the new iterate can also be determined by minimizing f over the smaller simplex

$$u^{k+1} \in \operatorname*{arg\,min}_{u \in M} f(u) \quad s.t. \quad u \in \operatorname{conv}\{\mathcal{A}_k\}.$$

This version of the algorithm is known as fully corrective conditional gradient, [151], or simplicial decomposition, [260]. Since the number of vertices is bounded the method converges in finitely many steps if $v^k \in \mathcal{A}_k$ for all $k \in \mathbb{N}$ is ensured. In the context of machine learning similar methods are known by the name of boosting or forward greedy selection. Linear convergence of such a method on a finite dimensional predictor problem with sparsity constraints is proven in [236]. Note that both accelerated versions can be also applied for general convex and compact M by identifying them with the closure of the convex hull of their extremal points. This is a consequence of the Krein-Milman theorem. However we are not aware of any improved convergence results in the case that the number of extremal points is infinite as for e.g. the euclidean unit ball.

Certainly these results raise the question in which situations a conditional gradient method should be given preference over the apparently better behaving projected gradient iteration. A first argument in favour of the conditional gradient algorithm lies in the complexity of the occurring subproblems. Computing the projection in (6.10) corresponds to minimizing a quadratic approximation of f at the current iterate over M:

$$u^{k+1} \in \operatorname*{arg\,min}_{v \in M} \frac{1}{2}|v - u^k + \frac{1}{L}\nabla f(u^k)|^2_{\mathbb{R}^d} = \operatorname*{arg\,min}_{v \in M} [f(u^k) + (\nabla f(u^k), v - u^k)_{\mathbb{R}^n} + \frac{L}{2}|v - u^k|^2_{\mathbb{R}^n}].$$

In contrast the conditional gradient step only requires the minimization of a linear model:

$$v^k \in \operatorname*{arg\,min}_{v \in M} (\nabla f(u^k), v)_{\mathbb{R}^n} = \operatorname*{arg\,min}_{v \in M} [f(u^k) + (\nabla f(u^k), v - u^k)_{\mathbb{R}^n}].$$

While computing projections can be cheaply realized for many sets there are certainly relevant problems in which this step represents a computational bottleneck. As an example we point out to minimization problems over subsets of the positive semi-definite matrices. Calculating the projection of a given u with respect to the Frobenius norm onto such admissible sets requires its full singular value decomposition, [46, Section 8.1]. On the contrary to solve the linear subproblems in (6.11) only one leading eigenpair of $\nabla f(u^k)$ needs to be computed, [133].

Secondly conditional gradient iterates often exhibit certain desired structural properties depending on the geometry of the feasible set. For example if f is minimized over the l^1 unit ball the element v^k can be chosen as a multiple of a canonic basis vector in $\mathbb{R}^n$. Thus, assuming that $u^0 = 0$, the iterate u^k is a sparse vector containing at most k non-zero entries. Similarly if M is a bounded subset of the positive semi-definite matrices the method can be realized to yield iterates u^k which are low-rank.

Most importantly however we point out that a straightforward extension of the projected gradient method to infinite dimensional spaces requires reflexivity and strict convexity of $\mathcal{M}$. We point out to [207] for a reference. In contrast the conditional gradient method generalizes naturally to minimization problems in non-reflexive Banach spaces. As a matter of fact the aforementioned improved convergence results in [83, 91, 92, 185] all hold in this general setting when we replace the euclidean norm in the previous considerations by the corresponding Banach space norm. The following two examples highlight this advantage of the conditional gradient method and the associated flexibility. In the first one we consider a bang-bang control problem on $L^2(\Omega)$. Here the objective functional is not strongly convex and a projected gradient iteration only yields a provable sublinear rate of convergence which is however also observed in practice. In comparison, linear convergence of the conditional gradient method can be obtained by interpreting the admissible set as a subset of the space of Radon measures. The second example deals with constrained minimization problems in spaces of measures. One of the main results of this thesis establishes a linear rate of convergence for an accelerated conditional gradient method for this type of problems. For more details we direct the reader to Section 6.3.

Example 6.2. *Consider a bounded domain $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, a desired state $y_d \in L^2(\Omega)$ and a linear operator $K\colon L^1(\Omega) \to L^2(\Omega)$ which we assume to be injective and compact. In the following we aim to compute the unique minimizer $\bar{u}$ of*

$$\min_{u \in U_{ad}} f(u) := \frac{1}{2}\|Ku - y_d\|^2_{L^2(\Omega)} \quad \textit{where} \quad U_{ad} = \left\{\, u \in L^2(\Omega) \mid |u(x)| \leq 1 \ \textit{a.e. } x \in \Omega \,\right\}.$$

It is well-known that $\bar{u}$ is characterized by the bang-bang condition

$$\bar{u}(x) \in \begin{cases} \{1\} & [K^*(K\bar{u} - y_d)](x) < 0 \\ [-1, 1] & [K^*(K\bar{u} - y_d)](x) = 0 \\ \{-1\} & [K^*(K\bar{u} - y_d)](x) > 0. \end{cases} \quad \textit{for a.e. } x \in \Omega. \tag{6.16}$$

Interpreting U_{ad} as a subset of $L^2(\Omega)$ a realization of the projected gradient method for its computation is given by

$$u^0 = 0, \quad u^{k+1}(x) = \min\left\{1,\ \max\{\, u^k(x) - s^k[K^*(Ku^k - y_d)](x), -1 \,\}\right\} \quad \textit{for a.e. } x \in \Omega,$$

where $s^k \geq 0$ is a suitably chosen step size see e.g. [214]. Note that the Hessian of the objective functional $\nabla^2 f(u) = K^*K$ is compact for every $u \in U_{ad}$. In particular this implies that f is not strongly convex on the admissible set and the projected gradient method only guarantees a sublinear rate of convergence on this problem.

Alternatively we consider $U_{ad} \subset L^1(\Omega) \subset \mathcal{M}(\Omega)$ and compute $\bar{u}$ by applying the following conditional gradient iteration

$$u^0 = 0, \quad u^{k+1} = u^k + s^k(v^k - u^k), \quad v^k(x) = \begin{cases} 1 & [K^*(Ku^k - y_d)](x) \leq 0 \\ -1 & [K^*(Ku^k - y_d)](x) > 0 \end{cases} \quad \textit{for a.e. } x \in \Omega.$$

Choosing the step size $s^k \in [0,1]$ according to the line minimization rule (6.13) *leads to*

$$s^k = \min\left\{ 1,\ \max\{(K^*(Ku^k - y_d), u^k - v^k)_{L^2(\Omega)} / \|K(u^k - v^k)\|^2_{L^2(\Omega)}, 0\} \right\},$$

where division by zero results in $+\infty$. While this method also only guarantees a sublinear rate in general a better convergence behavior can be expected if additional structural assumptions on the adjoint state $K^*(K\bar{u} - y_d)$ in the vicinity of its roots are imposed. More in detail we require the existence of a constant $c > 0$ such that for all $\varepsilon > 0$ there holds

$$\mu_L\left(\{\, x \in \Omega \mid [K^*(K\bar{u} - y_d)](x) = 0 \,\}\right) = 0, \quad \mu_L\left(\{\, x \in \Omega \mid -\varepsilon \leq [K^*(K\bar{u} - y_d)](x) \leq \varepsilon \,\}\right) \leq c\varepsilon. \tag{6.17}$$

Note that the first condition together with (6.16) *imply that $\bar{u}$ is strictly bang-bang i.e. it achieves the upper or the lower bound almost everywhere in Ω. We point out that the assumptions in* (6.17) *are well-established in the context of bang-bang optimal control problems see e.g. [66, 82]. From the bang-bang condition* (6.16) *and [66, Proposition 2.7] we now infer*

$$\operatorname*{arg\,min}_{v \in U_{ad}} (K^*(K\bar{u} - y_d), v)_{L^2(\Omega)} = \{\bar{u}\}, \quad (K^*(K\bar{u} - y_d), u - \bar{u})_{L^2(\Omega)} \geq \theta \|u - \bar{u}\|^2_{L^1(\Omega)}$$

for all $u \in U_{ad}$ and some positive constant $\theta > 0$. Thus Theorem 3.1 *in [92] yields the existence of $\zeta \in (0,1)$ such that the conditional gradient sequence $\{u^k\}_{k \in \mathbb{N}}$ satisfies*

$$f(u^k) - f(\bar{u}) + \|u^k - \bar{u}\|_{L^1(\Omega)} \leq c\zeta^k,$$

for some $c > 0$ independent of the iteration number k. To the best of our knowledge there are no comparable results for the projected gradient method under the additional assumptions of (6.17).

Example 6.3. *Let $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, be a compact set and denote by $\mathcal{M}(\Omega)$ the space of Radon measures on Ω. In this example we consider minimization problems of the form*

$$\min_{u \in \mathcal{M}^+(\Omega)} f(u) := F(Ku) \quad s.t. \quad \|u\|_{\mathcal{M}} \leq M_0, \tag{6.18}$$

where $K \colon \mathcal{M}(\Omega) \to Y$ is a linear and continuous operator taking values in a Hilbert space Y and the functional $F \colon Y \to \mathbb{R}$ is convex and continuously differentiable. The optimization variable u is searched for in the set of positive Radon measures $\mathcal{M}^+(\Omega)$. An additional upper bound $M_0 > 0$ on the total variation of the measure is enforced. Moreover we assume that the adjoint operator of K satisfies $K^* \colon Y \to \mathcal{C}(\Omega)$. Following the discussions in Section 6.3.2 a given admissiblie measure $\bar{u} \in \mathcal{M}^+(\Omega)$ is a minimizer of (6.18) *if and only if*

$$\|(K^*\nabla F(K\bar{u}))^-\|_{\mathcal{C}} \in \begin{cases} \{0\} & \|\bar{u}\|_{\mathcal{M}} \in [0, M_0) \\ [0, +\infty) & \|\bar{u}\|_{\mathcal{M}} = M_0 \end{cases},$$

and $\bar{u}$ fulfills the sparsity condition

$$\operatorname{supp} \bar{u} \subset \left\{\, x \in \Omega \mid [(K^*\nabla F(K\bar{u}))^-](x) = \|(K^*\nabla F(K\bar{u}))^-\|_{\mathcal{C}} \right\}, \quad \langle (K^*\nabla F(K\bar{u}))^+, \bar{u} \rangle = 0.$$

Here $(\varphi)^- = -\min\{0,\varphi\}$ *and* $(\varphi)^+ = \max\{0,\varphi\}$ *denote the negative and positive part of a function* $\varphi \in \mathcal{C}(\Omega)$ *respectively.*

To compute such a minimizer we apply a conditional gradient iteration which is defined as

$$u^0 = 0, \quad u^{k+1} = u^k + s^k(v^k - u^k),$$

where $v^k \in \mathcal{M}^+(\Omega)$ *is given by*

$$v^k = \begin{cases} M_0\delta_{\hat{x}^k} & \min_{x\in\Omega}[K^*\nabla F(Ku^k)](x) \leq 0 \\ 0 & \min_{x\in\Omega}[K^*\nabla F(Ku^k)](x) > 0 \end{cases}, \quad \hat{x}^k \in \operatorname*{arg\,min}_{x\in\Omega}[K^*\nabla F(Ku^k)](x).$$

While sublinear rates of convergence for conditional gradient methods on similar problems have been established in several recent works, [44, 50, 97, 200, 209], we are not aware of any results on conditions or acceleration schemes that guarantee an improved convergence behaviour. In particular there holds

$$u_1, u_2 \in \mathcal{M}^+(\Omega), \quad \|u_1\|_{\mathcal{M}} = \|u_2\|_{\mathcal{M}} = M_0 \Rightarrow \|u_1 + s(u_2-u_1)\|_{\mathcal{M}} = M_0 \quad \forall s \in [0,1].$$

Thus the admissible set in (6.18) *is not strongly convex and the arguments of Levitin, [185], and Demyanov, [83], cannot be applied to obtain improved convergence rates. Furthermore, denoting by* $\langle\cdot,\cdot\rangle$ *the duality pairing between* $\mathcal{C}(\Omega)$ *and* $\mathcal{M}(\Omega)$*, we readily obtain*

$$\operatorname*{arg\,min}_{\substack{v\in\mathcal{M}^+(\Omega)\\ \|v\|_{\mathcal{M}}\leq M_0}} \langle K^*\nabla F(K\bar{u}), v\rangle = \left\{ v \in \mathcal{M}^+(\Omega) \mid \operatorname{supp} v \subset \operatorname*{arg\,min}_{x\in\Omega}[K^*\nabla F(K\bar{u})](x), \quad \|v\|_{\mathcal{M}} = M_0 \right\},$$

if $\min_{x\in\Omega}[K^*\nabla F(K\bar{u})] < 0$*. As a consequence, assuming the extremality condition*

$$\operatorname*{arg\,min}_{\substack{v\in\mathcal{M}^+(\Omega)\\ \|v\|_{\mathcal{M}}\leq M_0}} \langle K^*\nabla F(K\bar{u}), v\rangle = \{\bar{u}\}$$

from Dunn's papers, [91, 92], implies that the optimal solution to (6.18) *is unique and given by a single Dirac delta function* $\bar{u} = M_0\delta_{\bar{x}}$ *for some* $\bar{x} \in \Omega$*. Moreover, even in this case, quadratic growth conditions of the form*

$$\langle K^*\nabla F(K\bar{u}), u - \bar{u}\rangle \geq \theta\|u-\bar{u}\|^2_{\mathcal{M}} \quad \forall u \in \mathcal{M}^+(\Omega), \quad \|u\|_{\mathcal{M}} \leq M_0,$$

cannot be fulfilled for any $\theta > 0$*. To see this take a sequence of points* $\{x_k\}_{k\in\mathbb{N}} \subset \Omega$*, with* $x_k \neq \bar{x}$, $x_k \to \bar{x}$*. Then it is readily verified that the sequence of Dirac delta functions* $\{u_k\}_{k\in\mathbb{N}}$ *defined by* $u_k = M_0\delta_{x_k}$*, fulfills*

$$\langle K^*\nabla F(K\bar{u}), u_k - \bar{u}\rangle \to 0 \quad \textit{but} \quad \|u_k - \bar{u}\|_{\mathcal{M}} = 2M_0 \quad \forall k \in \mathbb{N}.$$

In Section 6.3.5 we close this gap by providing an accelerated version of the conditional gradient method which achieves a linear rate of convergence on problem (6.18) *if certain structural requirements are met. Amongst other things we assume uniqueness and sparsity of the minimizer to* (6.18) *i.e.* $\bar{u}$ *consists of finitely many Dirac delta functions. In summary we obtain*

$$f(u^k) - f(\bar{u}) + \|u^k - \bar{u}\|_{\mathcal{C}^{0,1}(\Omega)^*} \leq c\zeta^k,$$

for some constants $c > 0$, $\zeta \in (0, 1)$ and all $k \in \mathbb{N}$ large enough. Here $\| \cdot \|_{\mathcal{C}^{0,1}(\Omega)^*}$ denotes the canonical norm on the dual space of the Lipschitz continuous functions. Note that we have

$$\mathcal{M}(\Omega) \hookrightarrow \mathcal{C}^{0,1}(\Omega)^*$$

i.e. we obtain quantitative convergence statements for the iterates in a weaker norm. We point out the similarity of this result to the improved convergence statement for the bang-bang optimization problem in the previous example. There the admissible set U_{ad} is given by the unit ball in $L^\infty(\Omega)$. However the convergence of the iterates $\{u^k\}_{k\in\mathbb{N}}$ is quantified with respect to the norm on the weaker space $L^1(\Omega)$.

6.2.2 Conditional gradient methods for composite minimization

In the following we present a generalization of the conditional gradient method for the solution of the composite minimization problem ($\mathfrak{P}$). To this end we first provide some preparatory results. Recall that the set of minimizers to ($\mathfrak{P}$) is bounded by some $M_0 > 0$, see Theorem 6.2. Associated to this constant define the auxiliary problem

$$\min_{\|u\|_{\mathcal{M}} \le M_0} [f(u) + g(u)]. \tag{$\mathfrak{P}_{M_0}$}$$

The following proposition states the equivalence of ($\mathfrak{P}_{M_0}$) and the original problem ($\mathfrak{P}$).

Proposition 6.6. *The set of minimizers to ($\mathfrak{P}$) and ($\mathfrak{P}_{M_0}$) coincide. If $\bar{u} \in \mathcal{M}$ is a minimizer of ($\mathfrak{P}$) then there holds*

$$\langle -\nabla f(\bar{u}), u - \bar{u} \rangle + g(\bar{u}) \le g(u) \quad \forall u \in \mathcal{M},\ \|u\|_{\mathcal{M}} \le M_0. \tag{6.19}$$

If f is convex, this condition is sufficient for optimality.

Proof. The equivalence of the set of minimizers to both problems follows immediately by comparing objective function values. The variational inquality in (6.19) can be deduced from Proposition 6.3. □

As a consequence we may consider a minimization algorithm for the constrained problem ($\mathfrak{P}_{M_0}$) in order to compute a minimzer of ($\mathfrak{P}$). We stress however that j is in general non-convex and thus the variational inquality in the last proposition is only necessary and not sufficient for optimality. Elements $\bar{u} \in \operatorname{dom} j$ fulfilling (6.19) will be called stationary points. In the following proposition these points are related to the roots of the non-negative primal-dual gap $\Phi \colon \mathcal{M} \to \mathbb{R}_+ \cup \{+\infty\}$ given by

$$\Phi(u) = \begin{cases} \max_{\|v\|_{\mathcal{M}} \le M_0} [\langle \nabla f(u), u - v \rangle + g(u) - g(v)], & u \in \operatorname{dom} j \\ +\infty, & \text{else} \end{cases} \tag{6.20}$$

Proposition 6.7. *Let $\bar{u} \in \operatorname{dom} j$ be given. Then $\bar{u}$ fulfills (6.19) if and only if $\Phi(\bar{u}) = 0$.*

Proof. Assume that $\bar{u}$ fulfills (6.19). Reordering yields

$$\langle \nabla f(\bar{u}), \bar{u} - v \rangle + g(\bar{u}) - g(v) \le 0 \quad \forall v \in \mathcal{M},\ \|v\|_{\mathcal{M}} \le M_0.$$

Maximizing with respect to v, $\|v\|_{\mathcal{M}} \leq M_0$, on both sides we conclude $\Phi(\bar{u}) \leq 0$. Since Φ only assumes non-negative values the statement follows. Conversely if $\bar{u}$ fulfills $\Phi(\bar{u}) = 0$ we readily obtain

$$\langle \nabla f(\bar{u}), \bar{u} \rangle + g(\bar{u}) \leq \langle \nabla f(\bar{u}), v \rangle + g(v) \quad \forall v \in \mathcal{M},\ \|v\|_{\mathcal{M}} \leq M_0.$$

By rearranging both sides we arrive at (6.19). □

Remark 6.1. Let $\{\varphi_i\}_{i\in\mathbb{N}}$ denote a countable dense subset of $\mathcal{C}$. Since $\mathcal{C}$ is separable the weak* topology on every closed ball $B \subset \mathcal{M}$ is metrizable. A suitable metric is given by

$$d\colon B \times B \to R, \quad (u_1, u_2) \mapsto \sum_{i=1}^{\infty} \frac{1}{2^i} |\langle \varphi_i, u_1 - u_2 \rangle|.$$

For a reference we point out to [52, Theorem 3.28]. In metric spaces sequential openness and opennes with respect to the corresponding metric are equivalent. We draw several conclusion from this discussion. Due to Assumption **A6.3** given an arbitrary $\bar{u} \in \operatorname{dom} j$, $\|\bar{u}\|_{\mathcal{M}} \leq M_0$ there exists $\varepsilon > 0$ with

$$B_{d,\varepsilon}(\bar{u}) = \{\, u \in \operatorname{dom} g \mid d(u, \bar{u}) < \varepsilon,\ \|u\|_{\mathcal{M}} \leq M_0 \} \subset \operatorname{dom} j.$$

Moreover, since the mapping

$$d(\cdot, \bar{u})\colon \mathcal{M} \to \mathbb{R}, \quad u \mapsto d(u, \bar{u})$$

is convex, the set $B_{d,\varepsilon}(\bar{u})$ is convex as well. In particular, given sequences $\{u_1^k\}_{k\in\mathbb{N}}$, $\{u_2^k\}_{k\in\mathbb{N}} \subset \operatorname{dom} g$ we have

$$u_1^k \rightharpoonup^* \bar{u}, \quad u_2^k \rightharpoonup^* \bar{u}, \quad \max\{\|u_1^k\|_{\mathcal{M}}, \|u_2^k\|_{\mathcal{M}}\} \leq M_0 \quad \forall k \in \mathbb{N} \Rightarrow u_1^k + s(u_2^k - u_1^k) \in B_{d,\varepsilon}(\bar{u}),$$

for all $s \in [0,1]$ and all $k \in \mathbb{N}$ large enough.

Clearly every measure $\bar{u}$ fulfilling the variational inequality in (6.4) and in particular every minimizer of ($\mathfrak{P}$) fulfills $\Phi(\bar{u}) = 0$. As a last preliminary step we consider well-posedness results and first-order optimality conditions for partial linearizations of ($\mathfrak{P}_{M_0}$). For brevity we define the convex function

$$g_{M_0}\colon \mathcal{M} \to \mathbb{R} \cup \{+\infty\}, \quad u \mapsto g(u) + I_{\|u\|_{\mathcal{M}} \leq M_0}(u).$$

Lemma 6.8. *Given an arbitrary $u \in \operatorname{dom} j$ there exists at least one minimizer $\bar{v} \in \mathcal{M}$ of the partially linearized problem*

$$\min_{\|v\|_{\mathcal{M}} \leq M_0} \left[\langle \nabla f(u), v \rangle + g(v) \right]. \tag{$\mathfrak{P}_{\text{lin}}$}$$

Furthermore $\bar{v} \in \mathcal{M}$, $\|\bar{v}\|_{\mathcal{M}} \leq M_0$, is a minimizer to ($\mathfrak{P}_{\text{lin}}$) if and only if

$$\bar{v} \in g^*_{M_0}(-\nabla f(u)). \tag{6.21}$$

Proof. The linearized objective functional

$$j_{\text{lin}}\colon \mathcal{M} \to \mathbb{R} \cup \{+\infty\}, \quad v \mapsto \langle \nabla f(u), v \rangle + g(v) + I_{\|v\|_{\mathcal{M}} \leq M_0}(v),$$

is proper, convex and weak* lower semi-continuous on $\mathcal{M}$. Furthermore the norm of every infimizing sequences for j_{lin} is bounded by M_0. Existence of a minimizer to ($\mathfrak{P}_{\text{lin}}$) can now be concluded as in Proposition 6.2. The necessary and sufficient optimality condition in (6.21) are obtained by applying Proposition 6.3 and Proposition 6.4. □

Algorithm 8 Generalized conditional gradient method (GCG) for ($\mathfrak{P}$)

```
1. Let $u^0 \in \operatorname{dom} j$, $\|u^0\|_{\mathcal{M}} \le M_0$.
while $\Phi(u^k) \ge \mathrm{TOL}$ do
    2. Determine $v^k \in \mathcal{M}$ such that
           $v^k \in \operatorname*{arg\,min}_{\|v\|_{\mathcal{M}} \le M_0} [\langle \nabla f(u^k), v \rangle + g(v)].$
    3. Choose $s^k \in [0,1]$. Set $u^{k+1/2} = u^k + s^k(v^k - u^k)$.
    4. Choose $u^{k+1} \in \mathcal{M}$ with $j(u^{k+1}) \le j(u^{k+1/2})$ and $\|u^{k+1}\|_{\mathcal{M}} \le M_0$.
end while
```

The generalized conditional gradient method (GCG) for the solution of ($\mathfrak{P}$) is summarized in Algorithm 8. In the k-th step of the method an intermediate iterate is obtained as convex combination $u^{k+1/2} = u^k + s^k(v^k - u^k)$ for some $s^k \in [0,1]$ between the current iterate $u^k \in \operatorname{dom} j$ and a minimizer $v^k \in \operatorname{dom} g$ of the partially linearized problem

$$\min_{\|v\|_{\mathcal{M}} \le M_0} [\langle \nabla f(u^k), v \rangle + g(v)]. \tag{6.22}$$

Clearly if $M \subset \mathbb{R}^n$ is a compact and convex set we recover the conditional gradient iteration described in (6.11) by setting $g = I_M$. Note that the auxiliary problem in (6.22) corresponds to the minimization of a composite functional comprising a linear approximation of f at the current iterate u^k and the non-smooth term g. The new iterate u^{k+1} is now chosen from the sublevel set associated to $u^{k+1/2}$

$$u^{k+1} \in \left\{ u \in \mathcal{M} \mid j(u) \le j(u^{k+1/2}), \quad \|u\|_{\mathcal{M}} \le M_0 \right\}.$$

In particular the choice of $u^{k+1} = u^{k+1/2}$ is possible. From this point of view step 4. in Algorithm 8 should be interpreted as a black box improvement step which allows e.g. to accelerate the algorithm or to exploit structural properties of the iterates. However it is not necessary to ensure convergence of the method in the following. Possible realizations of this step for a concrete problem are discussed in Sections 6.3.3 and 6.3.4.

Remark 6.2. We point out that the additional norm constraint in ($\mathfrak{P}_{\mathrm{lin}}$) is crucial in order to ensure the well-posedness of the conditional gradient step. In fact the partially linearized problem without additional norm constraints

$$\min_{v \in \mathcal{M}} [\langle \nabla f(u^k), v \rangle + g(v)]$$

may be unbounded if e.g. g is positive homogeneous.

Let us briefly summarize previous approaches in this direction. Generalized conditional gradient methods for minimization problems on finite dimensional spaces are considered in [14,131,195,224]. The general Hilbert space case is covered in [49]. In particular this last work provides convergence results for the sequence of iterates for general f. Additionally a sublinear rate of convergence for the objective function values is shown assuming convexity of f and Lipschitz continuity of the gradient ∇f. We stress however that composite minimization problems in Hilbert spaces can be solved by proximal gradient methods if the computation of the prox-operator, [199], associated to g is inexpensive. These methods rely on a fix-point reformulation of the subdifferential inclusion $-\nabla f(\bar{u}) \in \partial g(\bar{u})$ and generalize the projected gradient iteration from (6.10). Proximal gradient methods yield linear convergence of the objective functional values for strongly convex objective functionals. For general convex j a $1/k^2$ rate can be ensured by adding suitable acceleration steps. As a reference we point out to [27, 202] for a discussion of these methods on finite dimensional spaces and to [48, 75, 233] for the general Hilbert space case. Moreover if f and g admit additional regularity the fix-point formulation of the optimality condition may be amenable for efficient solution methods such as generalized Newton-type algorithms, [223,257]. In contrast improved convergence results for GCG methods on problems incorporating nonsmooth g other than convex indicator functions are scarce. We are only aware of [24]. There the authors consider the important case of norm-regularized problems, i.e. $g = \beta\|\cdot\|_{\mathcal{M}}$ for some $\beta > 0$, in finite dimensions. Linear convergence of an accelerated conditional gradient scheme based on a smooth reformulation of the problem is proven if the associated norm balls are polytopes. The constants appearing in the convergence estimate heavily depend on the geometry of the norm ball and, possibly, the number of its vertices. This makes a straightforward extension of this result to infinite dimensional spaces impossible.

The main motivation for the application of generalized conditional gradient methods in the context of this thesis are minimization problems where the space $\mathcal{M}$ is given by the dual space of an infinite dimensional separable Banach space. Important examples include the space of essentially bounded functions $\mathcal{M} = L^\infty(\Omega)$, the space of Radon measures $\mathcal{M} = \mathcal{M}(\Omega)$ and the space of functions with bounded total variation $\mathcal{M} = \mathrm{BV}(\Omega)$ on a subset $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$. The algorithmic solution of these type of problems on the function space level is challenging since $\mathcal{M}$ generally lacks desirable properties such as reflexivity, strict convexity and smoothness. The aim of this section is to show that the generalized conditional gradient method from Algorithm 8 is able to cope with both, the composite structure of the objective functional as well as complicated spaces $\mathcal{M}$. The simple structure of the resulting algorithm is highlighted on several instructive examples.

To the best of our knowledge generalized conditional methods on general Banach spaces have only been considered recently in [274, 277]. There the authors assume that $\mathcal{M}$ is a complete normed space. The analysis in the present work distinguishes itself from those two papers in several points. First the authors limit their discussion to the case of convex functions f and only consider quantitative convergence results. Qualitative results on the convergence of the iterates are only provided in the case of reflexive $\mathcal{M}$. Note that this is in part a consequence of the assumed generality in these papers. In particular since no further assumptions beyond completeness of the space $\mathcal{M}$ are made it is unclear in which topology the unit ball is compact. In contrast we exploit the duality relation $\mathcal{M} = \mathcal{C}^*$ and the implied weak* compactness of the unit ball to provide qualitative convergence guarantees for the objective function values as well as the iterates even if f is nonconvex. In the convex case these results are strengthened and quantitative statements are derived. In this context additional effort has to be paid due to the potential openness of the domain of the smooth part which is a topic that is also not covered by these prior works. While this may seem as a minor technical difference we recall that this additional assumption on the domain is indeed crucial to deal with the sensor placement problems of the previous chapters. A second key difference lies in the choice of the step size. We comment on the details at a later point of this section. Last we point out that there has been considerable work on GCG methods on the space of Radon measures. For a discussion on known results in this case we refer to Section 6.3.

Let us now return to the analysis of the generalized conditional gradient method in Algorithm 8. Since j is nonconvex in the general case only (subsequential) convergence of the iterates $\{u^k\}_{k\in\mathbb{N}}$

towards stationary points can be expected. To monitor the convergence of the GCG method we thus consider the primal-dual gap Φ of the iterates as termination criterion. By construction there holds

$$\Phi(u^k) = \langle \nabla f(u^k), u^k\rangle + g(u^k) - \langle \nabla f(u^k), v^k\rangle - g(v^k). \tag{6.23}$$

As a consequence the termination criterion can be evaluated cheaply once a solution of the partially linearized problem is obtained.

Remark 6.3. Obviously the presented algorithm implicitly assumes that the linearized subproblems in step 2. can be solved efficiently and their computational cost is neglectable in comparison to a solution of the original problem ($\mathfrak{P}_{M_0}$) by a different method.

Remark 6.4. At this point let us justify the term *primal-dual gap* for the functional Φ in (6.20). To avoid unnecessary additional notation we restrict the following discussion to convex functions f. However similar arguments are also valid in the nonconvex case. Define the constrained dual objective functional

$$d_{M_0} \colon \mathcal{M} \to \mathbb{R} \cup \{+\infty\}, \quad u \mapsto f(u) + g_{M_0}(u)$$

as well as its predual or primal counterpart

$$p_{M_0} \colon \mathcal{C} \to \mathbb{R} \cup \{+\infty\}, \quad \varphi \mapsto -f^*(\varphi) - g^*_{M_0}(-\varphi).$$

Let $\bar{u} \in \mathcal{M}$ denote an optimal solution of ($\mathfrak{P}_{M_0}$). This implies

$$\bar{u} \in \partial g^*_{M_0}(-\nabla f(\bar{u})).$$

Moreover we get $\bar{u} \in \partial f^*(\nabla f(\bar{u}))$ from the differentiability of f at $\bar{u}$. We conclude that $\nabla f(\bar{u}) \in \mathcal{C}$ is a maximizer of p_{M_0} since

$$0 = \bar{u} - \bar{u} \in \partial f^*(\nabla f(\bar{u})) - \partial g^*_{M_0}(-\nabla f(\bar{u})) \subset \partial(f^*(\cdot) + g^*(-\cdot))(\nabla f(\bar{u})) = -\partial p_{M_0}(\nabla f(\bar{u})),$$

where we used the inclusion rule for the subdifferential of a sum. Furthermore due to the continuity of f on its domain strong duality holds

$$j(\bar{u}) = \min_{u \in \mathcal{M}} d_{M_0}(u) = \max_{\varphi \in \mathcal{C}} p_{M_0}(\varphi) = -\min_{\varphi \in \mathcal{C}} -p_{M_0}(\varphi) = p_{M_0}(\nabla f(\bar{u})).$$

Now denote by $\{u^k\}_{k\in\mathbb{N}}$ the sequence generated by Algorithm 8. We obtain

$$\begin{aligned} d_{M_0}(u^k) - p_{M_0}(\nabla f(u^k)) &= f(u^k) + f^*(\nabla f(u^k)) + g_{M_0}(u^k) + g^*_{M_0}(-\nabla f(u^k)) \\ &= \langle \nabla f(u^k), u^k\rangle + g_{M_0}(u^k) + g^*_{M_0}(-\nabla f(u^k)) \\ &= \langle \nabla f(u^k), u^k\rangle + g_{M_0}(u^k) - \langle \nabla f(u^k), v^k\rangle - g_{M_0}(v^k) \\ &= \Phi(u^k). \end{aligned}$$

Here we used $v^k \in \partial g^*_{M_0}(-\nabla f(u^k))$, $\{\nabla f(u^k)\} = \partial f(u^k)$ and Proposition 6.4. From this perspective the functional Φ gives the gap between primal and dual objective function values associated to $(\nabla f(u^k), u^k)$ in each iteration.

Naturally the convergence of the proposed method will depend on the choice of the step size s^k. As already mentioned in the preliminary discussions convergence of the classical conditional gradient scheme is provable for a large variety of step sizes. For abbreviation we set

$$u_s^k = u^k + s(v^k - u^k) \quad \forall s \in [0,1].$$

In [274, 277] the authors establish convergence of a generalized conditional gradient scheme in Banach spaces for two particular choices of the step size. More concretely open loop and minimization step sizes similar to those in (6.14) and (6.13) are studied. Here we point out that the first choice does not yield a descent method while the second choice amounts to the solution of a one dimensional minimization problem in every iteration.

In the present work we base our algorithm on a generalization of the well known Armijo-Goldstein condition cf. [49]. This particular choice of the step size guarantees descend in every iteration and its determination only requires a backtracking line search on the objective functional.

Definition 6.2. Let $\gamma \in (0,1)$, $\alpha \in (0,1/2]$. The step size s^k is chosen according to the Quasi-Armijo-Goldstein condition if $s^k = \gamma^{n_k}$ where $n_k \in \mathbb{N}$ is the smallest integer with

$$\alpha\gamma^{n_k}\Phi(u^k) \leq j(u^k) - j(u^k_{\gamma^{n_k}}). \tag{6.24}$$

The following lemma illustrates that this choice of the step size is always possible if u^k is not a stationary point.

Lemma 6.9. *Let an arbitrary measure $u \in \operatorname{dom} j$ be given. Assume that $\Phi(u) > 0$ and denote by $v \in \operatorname{dom} g$ the solution of the associated partially linearized problem $(\mathfrak{P}_{\text{lin}})$. Define $u_s = u + s(v-u)$ and the extended real-valued function*

$$W\colon\; [0,1] \to \mathbb{R} \cup \{-\infty\} \quad W(s) = \frac{j(u) - j(u_s)}{s\Phi(u)}.$$

The function W is upper semi-continuous on $(0,1]$ and there holds $\liminf_{s\to 0} W(s) = 1$.

Proof. Since the domain of j is sequentially weak* open in $\operatorname{dom} g$ there holds $u_s \in \operatorname{dom} j$ for all s small enough. Due to the definition of v we have

$$W(s) = \frac{j(u) - j(u_s)}{s\Phi(u)} = \frac{j(u) - j(u_s)}{s\left(\langle \nabla f(u), u - v\rangle + g(u) - g(v)\right)}.$$

From the mean value theorem we get the existence of $\zeta_s \in [0,1]$ and $\tilde{u}_s = u + \zeta_s(u_s - u) \in \operatorname{dom} j$ with

$$W(s) = \frac{s\langle \nabla f(\tilde{u}_s), u - v\rangle + g(u) - g(u_s)}{s\left(\langle \nabla f(u), u - v\rangle + g(u) - g(v)\right)}$$

Using the convexity of g, we estimate

$$\frac{s\langle \nabla f(\tilde{u}_s), u - v\rangle + g(u) - g(u_s)}{s\left(\langle \nabla f(u), u - v\rangle + g(u) - g(v)\right)} \geq \frac{s\left(\langle \nabla f(\tilde{u}_s), u - v\rangle + g(u) - g(v)\right)}{s\left(\langle \nabla f(u), u - v\rangle + g(u) - g(v)\right)}.$$

Since ζ_s is bounded independently of s, there holds $\tilde{u}_s \rightharpoonup^* u$ for $s \to 0$. Due to the weak*-to-strong continuity of ∇f, the right-hand side of the inequality tends to 1 yielding $\liminf_{s\to 0} W(s) \geq 1$. The upper semi-continuity of W on $(0,1)$ follows directly from $u_s \in \operatorname{dom} g$ for all $s \in (0,1]$ and from the lower weak* semi-continuity of j on $\operatorname{dom} g$. □

Before proceeding to the proof of convergence results for Algorithm 8 we discuss the abstract generalized conditional gradient method for the special case of norm regularization

$$\min_{\|u\|_{\mathcal{M}} \leq M_0} f(u) + \beta \|u\|_{\mathcal{M}},$$

where $\beta > 0$ denotes a given regularization parameter. In this case it is readily verified that a solution to the linearized problem ($\mathfrak{P}_{\text{lin}}$) in the k-th iteration is given by any $v^k \in \mathcal{M}$ fulfilling

$$\langle \nabla f(u^k), v^k \rangle = -\|\nabla f(u^k)\|_{\mathcal{C}} \|v^k\|_{\mathcal{M}}, \quad \|v^k\|_{\mathcal{M}} = \begin{cases} M_0 & \|\nabla f(u^k)\|_{\mathcal{C}} \geq \beta \\ 0 & \|\nabla f(u^k)\|_{\mathcal{C}} < \beta \end{cases}.$$

In particular if $\|\nabla f(u^k)\|_{\mathcal{C}} \geq \beta$ the algorithmic solution of the linearized subproblem requires the computation of an element $\tilde{v}^k \in \partial \| - \nabla f(u^k)\|_{\mathcal{C}}$. The following examples describe the generalized conditional gradient iterations for the norm regularized problem and two choices of $\mathcal{M}$. While the space of optimization variables is non-reflexive and not strictly convex in both cases the computation of an element in the subdifferential can be done analytically. This underlines the simple structure of the presented method.

Example 6.4. *Let $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, be a bounded domain. Set $\mathcal{C} = L^1(\Omega)$ and $\mathcal{M} = L^\infty(\Omega)$ with the usual norms*

$$\|\varphi\|_{L^1(\Omega)} = \int_\Omega |v| \, \mathrm{d}x, \quad \|u\|_{L^\infty} = \operatorname*{ess\,sup}_{x \in \Omega} |u(x)|,$$

for $\varphi \in L^1(\Omega)$ and $u \in L^\infty(\Omega)$. In this case the optimization problem ($\mathfrak{P}$) can be related to so called minimum effort control problems, [72]. Denote by u^k the k-th iterate generated by the GCG method and set $p^k = -\nabla f(u^k)$. A solution v^k to the partially linearized problem is obtained by scaling the sign of p^k:

$$v^k = \begin{cases} M_0 \operatorname{sgn}(p^k) & \|p^k\|_{L^1(\Omega)} \geq \beta \\ 0 & \|p^k\|_{L^1(\Omega)} < \beta \end{cases}, \quad \operatorname{sgn}(p^k)(x) = \begin{cases} 1 & p^k(x) \geq 0 \\ -1 & p^k(x) < 0 \end{cases} \quad \text{for a.e. } x \in \Omega.$$

In particular this implies that v^k admits a strict bang-bang structure i.e. its image only contains two values.

Example 6.5. *As a second example consider a bounded domain $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$ and a time interval $I = [0, T]$, $T > 0$. By $\mathcal{C}_0(\Omega)$ we denote the space of continous functions on $\bar{\Omega}$ which are zero at the boundary. Its dual space is given by the space of Radon measures on Ω which we identify by restricting elements of $\mathcal{M}(\bar{\Omega})$ to the interior*

$$\mathcal{M}(\Omega) \simeq \{\, u|_\Omega \mid u \in \mathcal{M}(\bar{\Omega}) \,\}.$$

We consider $\mathcal{C} = L^2(I, \mathcal{C}_0(\Omega))$, the space of all strongly measurable functions $\varphi \colon I \to \mathcal{C}_0(\Omega)$ for which the associated norm

$$\|\varphi\|_{\mathcal{C}} = \sqrt{\int_I \|\varphi(t)\|^2_{\mathcal{C}(\Omega)} \, \mathrm{d}t}$$

is finite. This space is a separable Banach space due to the separability of $\mathcal{C}_0(\Omega)$ see e.g. [263, Theorem I.5.18]. Its topological dual space is given by $\mathcal{M} = L^2_{w*}(I, \mathcal{M}(\Omega))$, the space of weak measurable functions $u \colon I \to \mathcal{M}(\Omega)$ with finite dual norm*

$$\|u\|_{\mathcal{M}} = \sqrt{\int_I \|u(t)\|^2_{\mathcal{M}(\Omega)} \mathrm{d}t}.$$

The associated duality pairing between $\mathcal{C}$ and $\mathcal{M}$ is given by

$$\langle \varphi, u \rangle = \int_I \langle \varphi(t), u(t) \rangle_{\mathcal{C}_0(\Omega),\mathcal{M}(\Omega)} \, \mathrm{d}t, \quad \varphi \in \mathcal{C},\ u \in \mathcal{M}.$$

For a reference on this dual identification we point out to [96, 8.20.3]. Optimal control problems on the space $L^2_{w^*}(I, \mathcal{M}(\Omega))$ are analyzed in [60, 254].

Let u^k denote the GCG iterate in the k-th iteration and set $p^k = -\nabla f(u^k) \in \mathcal{C}$. Then p^k interpreted as a scalar-valued function on $I \times \bar{\Omega}$ is carathéodory. Thus there exists a measurable selection $\hat{x}^k_t \colon I \to \bar{\Omega}$ with $|p^k(\hat{x}^k_t, t)| = \|p^k(\cdot, t)\|_{\mathcal{C}(\Omega)}$ for a.e. $t \in I$. If $p^k \neq 0$ we define the function $\tilde{v}^k(t) = (p^k(x_t, t)/\|p^k\|_{\mathcal{C}})\delta_{\hat{x}^k_t}$ for a.e. $t \in I$. As in [60, Theorem 3.3] we now argue that

$$\tilde{v}^k \in L^2_{w^*}(I, \mathcal{M}(\Omega)) \quad \textit{with} \quad \langle p^k, \tilde{v}^k \rangle = \frac{1}{\|p^k\|_{\mathcal{C}}} \int_I p^k(\hat{x}^k_t) \langle p^k(t), \delta_{\hat{x}^k_t} \rangle_{\mathcal{C}_0(\Omega),\mathcal{M}(\Omega)} \, \mathrm{d}t = \|p^k\|_{\mathcal{C}}.$$

As a consequence a solution v^k to the linearized subproblem is given by a time-dependent Dirac delta function moving along a measurable trajectory:

$$v^k = \begin{cases} \frac{p^k(\hat{x}^k_t)}{\|p^k\|_{\mathcal{C}}} \delta_{\hat{x}^k_t} & \|p^k\|_{\mathcal{C}} \geq \beta \\ 0 & \|p^k\|_{\mathcal{C}} < \beta \end{cases}, \quad \hat{x}^k_t \in \operatorname*{arg\,max}_{x \in \Omega} |p^k(x, t)| \quad \textit{for a.e. } t \in I.$$

6.2.3 Convergence analysis

This section is devoted to the derivation of convergence results for the generalized conditional gradient method. The following presentation is divided into two parts. First we prove subsequential weak* convergence of $\{u^k\}_{k\in\mathbb{N}}$ towards stationary points of j under no additional assumptions on f. Second, convexity of f and additional smoothness of ∇f is assumed. In this case $\{u^k\}_{k\in\mathbb{N}}$ defines a minimizing sequence for j and the objective function values converge sublinearly. Since the general problem $(\mathfrak{P})$ encompasses minimization of smooth functions over convex and compact sets in $\mathbb{R}^n$ this result is sharp, [56]. Furthermore every weak* accumulation point of $\{u^k\}_{k\in\mathbb{N}}$ is a global minimizer of j.

Convergence in the general case

As a preparational step we establish semi-continuity properties of the primal-dual gap Φ.

Lemma 6.10. *Given a sequence $\{u_k\}_{k\in\mathbb{N}} \subset \operatorname{dom} j$ with weak* limit $\bar{u} \in \operatorname{dom} j$ there holds $\liminf_{k\to\infty} \Phi(u_k) \geq \Phi(\bar{u})$.*

Proof. For an arbitrary $v \in \mathcal{M}$, $\|v\|_{\mathcal{M}} \leq M_0$, we obtain

$$\Phi(u_k) \geq \langle \nabla f(u_k), u_k - v \rangle + g(u_k) - g(v).$$

Taking the limes inferior for $k \to \infty$ on both sides of the inequality yields

$$\liminf_{k\to\infty} \Phi(u_k) \geq \langle \nabla f(\bar{u}), \bar{u} - v \rangle + g(\bar{u}) - g(v),$$

due to the weak* convergence of $\{u_k\}_{k\in\mathbb{N}}$ and the continuity properties of ∇f and g. Since v was chosen arbitrary, we can maximize over all $v \in \mathcal{M}$, $\|v\|_{\mathcal{M}} \leq M_0$ from which we conclude $\liminf_{k\to\infty} \Phi(u_k) \geq \Phi(\bar{u})$. □

By construction the norms of the GCG iterates $\{u^k\}_{k\in\mathbb{N}}$ are uniformly bounded by M_0. Applying the Banach-Alaoglu Theorem we can thus extract at least one weak* convergent subsequence. The following theorem characterizes the weak* accumulation points of $\{u^k\}_{k\in\mathbb{N}}$.

Theorem 6.11. *Assume that the sequences $\{u^k\}_{k\in\mathbb{N}}$, $\{u^{k+1/2}\}_{k\in\mathbb{N}}$ and $\{v^k\}_{k\in\mathbb{N}}$ are generated by Algorithm 8 and let f and g fulfil the requirements of Assumption 6.1. Then there exists at least one subsequence of $\{u^k\}_{k\in\mathbb{N}}$ converging in the weak* sense. Every weak* accumulation point $\bar{u}$ of $\{u^k\}_{k\in\mathbb{N}}$ fulfills $\Phi(\bar{u}) = 0$.*

Proof. Without loss of generality assume that $\Phi(u^k) > 0$ for all k. By construction we have $\max\left\{\|u^k\|_{\mathcal{M}}, \|v^k\|_{\mathcal{M}}\right\} \leq M_0$ for all $k \in \mathbb{N}$. Consequently we can extract subsequences (denoted by the same index in the following) such that $u^k \rightharpoonup^* \bar{u}$ and $v^k \rightharpoonup^* \bar{v}$ for some $\bar{u}$, $\bar{v} \in \operatorname{dom} g$. Due to the choice of the step size s^k there holds $\{u^k\}_{k\in\mathbb{N}} \subset \operatorname{dom} j$ and

$$j(\bar{u}) \leq \liminf_{k\to\infty} j(u^k) \leq j(u^0) < \infty$$

Thus we get $\bar{u} \in \operatorname{dom} j$. From

$$\sum_{k=0}^{\infty} \Big[j(u^k) - j(u^{k+1}) \Big] \leq j(u^0) - j(\bar{u}), < \infty,$$

we additionally conclude $\lim_{k\to\infty}[j(u^k) - j(u^{k+1})] = 0$. We will prove the claimed result by contradiction. For this purpose, assume that $0 < \Phi(\bar{u}) \leq \liminf_{k\to\infty} \Phi(u^k)$. From the definition of the Quasi-Armijo rule, see (6.24), we obtain

$$0 \leq \alpha s^k \leq \frac{j(u^k) - j(u^{k+1/2})}{\Phi(u^k)} \leq \frac{j(u^k) - j(u^{k+1})}{\Phi(u^k)}.$$

Taking the limit superior yields

$$0 \leq \alpha \limsup_{k\to\infty} s^k \leq \frac{\lim_{k\to\infty}\left[j(u^k) - j(u^{k+1})\right]}{\liminf_{k\to\infty} \Phi(u^k)},$$

from which we conclude that $s^k \to 0$ as $k \to \infty$, since $\liminf_{k\to\infty} \Phi(u^k) > 0$ by assumption. From the convergence of the step sizes we get $s^k/\gamma < 1$ for all k large enough as well as

$$u^{k+1/2} \rightharpoonup^* \bar{u}, \quad u^k + \frac{s^k}{\gamma}(v^k - u^k) \rightharpoonup^* \bar{u}.$$

Again using (6.24) we obtain for k large enough that:

$$\frac{\alpha s^k \Phi(u^k)}{\gamma} > j(u^k) - j\left((u^k + (s^k/\gamma)(v^k - u^k))\right).$$

Moreover, see Remark 6.1, for every $s \in [0, s^k/\gamma]$ we have $u^k + s(v^k - u^k) \in \operatorname{dom} j$ if k is chosen large enough. Again, by possibly passing to a subsequence, there exists $\hat{s}^k \in [0, s^k/\gamma]$ and $\hat{u}^k = u^k + \hat{s}^k(v^k - u^k) \in \operatorname{dom} j$ with $\hat{u}^k \rightharpoonup^* \bar{u}$ and

$$\langle \nabla f(\hat{u}^k), u^k - v^k \rangle + g(u^k) - g(v^k) \leq \frac{j(u^k) - j\left((u^k + (s^k/\gamma)(v^k - u^k))\right)}{s^k/\gamma} < \alpha \Phi(u^k).$$

due to the mean value theorem and the convexity of g. Considering the term on the left-hand side, there holds

$$\langle \nabla f(\hat{u}^k), u^k - v^k\rangle + g(u^k) - g(v^k) = \langle \nabla f(\hat{u}^k) - \nabla f(u^k), u^k - v^k\rangle + \Phi(u^k).$$

Combining the previous arguments and rearranging we obtain

$$\langle \nabla f(\hat{u}^k) - \nabla f(u^k), u^k - v^k\rangle \leq (\alpha - 1)\Phi(u^k),$$

and consequently

$$0 = \lim_{k\to\infty} [\langle \nabla f(\hat{u}^k) - \nabla f(u^k), u^k - v^k\rangle] \leq (\alpha - 1) \liminf_{k\to\infty} \Phi(u^k),$$

where we used the weak* convergence of $\hat{u}^k$, u^k as well as the continuity properties of ∇f. Dividing both sides by $\alpha - 1 < 0$, we conclude $\Phi(\bar{u}) \leq \liminf_{k\to\infty} \Phi(u^k) \leq 0$, which gives a contradiction. □

Rates of convergence for convex f

Throughout this section we make the following additional assumptions on the smooth part f.

Assumption 6.2. For an arbitrary $u_0 \in \operatorname{dom} j$ define the sublevel set

$$E_j(u_0) = \{\, u \in \mathcal{M} \mid j(u) \leq j(u_0)\,\}.$$

Let the following additional assumptions hold

A6.6 Let f be convex on $E_j(u_0)$ for every $u_0 \in \operatorname{dom} j$.

A6.7 The gradient ∇f is Lipschitz-continuous on sublevel sets, i.e. for $u_0 \in \operatorname{dom} j$ there exists a constant $L_{u_0} > 0$ only depending on $j(u_0)$ with

$$\|\nabla f(u_1) - \nabla f(u_2)\|_{\mathcal{C}} \leq L_{u_0} \|u_1 - u_2\|_{\mathcal{M}} \quad \forall u_1, u_2 \in E_j(u_0).$$

Since j is now convex every of its stationary points $\bar{u} \in \mathcal{M}$ is a global minimizer. We define the residual of j as

$$r_j \colon \mathcal{M} \to \mathbb{R} \cup \{+\infty\}, \quad u \mapsto j(u) - \min_{\tilde{u}\in\mathcal{M}} j(\tilde{u}).$$

By convexity of f the residual can be bounded by the primal-dual gap Φ. Furthermore, the following growth estimate for j at u^k in the search direction is obtained.

Lemma 6.12. *For every $u \in \operatorname{dom} j$ there holds*

$$r_j(u) \leq \Phi(u). \tag{6.25}$$

Fix an index $k \in \mathbb{N}$. Let u^k, v^k be generated by Algorithm 8. Further let a step size $s \in [0,1]$ with $u^k_s = u^k + s(v^k - u^k) \in E_j(u^0)$ be given. Then there holds

$$j(u^k_s) - j(u^k) \leq -s\Phi(u^k) + \frac{L_{u^0}}{2}\left(s\|u^k - v^k\|_{\mathcal{M}}\right)^2. \tag{6.26}$$

Proof. We first proof (6.25). This clearly holds for $u \notin \operatorname{dom} j$. Let $u \in \operatorname{dom} j$ be given. From the convexity of f on sublevel sets we readily obtain

$$j(u) - j(\bar{u}) \leq \langle \nabla f(u), u - \bar{u} \rangle + g(u) - g(\bar{u}).$$

The right hand side is estimated by

$$\langle \nabla f(u), u - \bar{u} \rangle + g(u) - g(\bar{u}) \leq \max_{\|v\|_{\mathcal{M}} \leq M_0} \langle \nabla f(u), u - \bar{u} \rangle + g(u) - g(\bar{u})] = \Phi(u),$$

using $\|\bar{u}\|_{\mathcal{M}} \leq M_0$. This yields the claimed result. We proceed to the second claim. Due to the convexity of the sublevel set $E_j(u^0)$ we obtain

$$j(u^k_s) - j(u^k) = -\, s\langle \nabla f(u^k), u^k - v^k \rangle + g(u^k_s) - g(u^k) + \int_0^s \langle \nabla f(u_\sigma) - \nabla f(u^k), v^k - u^k \rangle \mathrm{d}\sigma,$$

with $u_\sigma = u^k + \sigma(v^k - u^k) \in E_j(u^0)$ for $\sigma \in [0, s]$. Using the convexity of g, $\|\bar{u}\|_{\mathcal{M}} \leq M_0$ and the definition of v^k we obtain

$$-s\langle \nabla f(u^k), u^k - v^k \rangle + g(u^k_s) - g(u^k) \leq -s\left(\langle \nabla f(u^k), u^k - v^k \rangle + g(u^k) - g(v^k) \right),$$

where the right-hand side simplifies to $-s\Phi(u^k)$. Due to the Lipschitz continuity of $\nabla f(u^k)$ on $E_j(u^0)$ we get

$$\begin{aligned} \int_0^s \langle \nabla f(u_\sigma) - \nabla f(u^k), v^k - u^k \rangle \mathrm{d}\sigma &\leq \|v^k - u^k\|_{\mathcal{M}} \int_0^s \|\nabla f(u_\sigma) - \nabla f(u^k)\|_{\mathcal{C}} \mathrm{d}\sigma \\ &\leq L_{u^0} \|v^k - u^k\|^2_{\mathcal{M}} \int_0^s \sigma \mathrm{d}\sigma \\ &= \frac{L_{u^0}}{2} (s\|v^k - u^k\|_{\mathcal{M}})^2. \end{aligned}$$

Combining both estimates yields the proof. □

Due to the possibly open domain of f in $\mathcal{M}$ we also need the following technical lemma concerning the continuity properties of the function W which was introduced in Lemma 6.9.

Lemma 6.13. *Let $u \in \operatorname{dom} j$ with $\Phi(u) > 0$ be given and denote by $v \in \operatorname{dom} g$ the solution to the associated linearized problem $(\mathfrak{P}_{\text{lin}})$. If $v \in \operatorname{dom} j$ we have $W \in \mathcal{C}((0,1))$. Otherwise there exists $\hat{s} \in (0,1]$ with $W \in \mathcal{C}((0,\hat{s}))$ and $\lim_{s \to^- \hat{s}} W(s) = -\infty$.*

Proof. Since u is not optimal the function W is proper. Set $u_s = u + s(v - u)$ and define the convex auxiliary function

$$\hat{\jmath}\colon [0,1] \to \mathbb{R} \quad s \mapsto j(u_s),$$

Since $\Phi(u) > 0$ there exists $s \in (0,1]$ with $\hat{\jmath}(s) \in \mathbb{R}$. We further conclude

$$(0, \hat{s}) \subset \operatorname{dom}_{(0,1]} \hat{\jmath}, \quad \hat{s} = \sup \operatorname{dom}_{[0,1]} \hat{\jmath} \in (0,1].$$

Note that $\hat{\jmath}$ is continuous on $(0, \hat{s})$, see [98, Proposition 2.5]. Let us distinguish two cases. If $v \in \operatorname{dom} j$ there holds $\hat{s} = 1$. From its definition we thus get $W \in \mathcal{C}((0,1))$. In the second case if $v \notin \operatorname{dom} j$ there holds

$$\hat{s} \notin \operatorname{dom}_{[0,1]} \hat{\jmath}, \quad \lim_{s \to^- \hat{s}} j(u_s) = +\infty,$$

due to the openness assumption on the domain of f. Hence we conclude

$$W \in \mathcal{C}((0,\hat{s})), \quad \lim_{s \to^- \hat{s}} W(s) = -\infty,$$

which finishes the proof. $\square$

Collecting all the previous results we can prove a sublinear rate of convergence for the residuals of the iterates generated by Algorithm 8.

Theorem 6.14. *Assume that the sequences $\{u^k\}$, $\{u^{k+1/2}\}$ and $\{v^k\}$, $k \in \mathbb{N}$, are generated by Algorithm 8 with $\{s^k\}_{k\in\mathbb{N}}$ chosen according to the Quasi-Armijo-Goldstein condition with parameters $\alpha \in (0,1/2]$, $\gamma \in (0,1)$. Let Assumption 6.1 and Assumption 6.2 hold. Furthermore let $\Phi(u^k) > 0$ for all $k \in \mathbb{N}$. Then $\{u^k\}_{k\in\mathbb{N}}$ is a minimizing sequence for j and there holds*

$$r_j(u^k) \le \frac{r_j(u^0)}{1+qk}, \quad q = \alpha \min\left\{\frac{c}{4L_{u^0}M_0^2}, 1\right\} \tag{6.27}$$

where $c = 2\gamma(1-\alpha)r_j(u^0)$. Moreover there exists a weak accumulation point $\bar{u}$ of $\{u^k\}_{k\in\mathbb{N}}$ and every such point is a global minimum of j.*

Proof. By the definition of the step size s^k as well as (6.25) there holds

$$\alpha s^k r_j(u^k) \le \alpha s^k \Phi(u^k) \le r_j(u^k) - r_j(u^{k+1/2}),$$

which yields

$$r_j(u^{k+1/2}) \le (1-\alpha s^k) r_j(u^k). \tag{6.28}$$

Since $\Phi(u^k) > 0$ 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

$$r_j(u^{k+1}) \le r_j(u^{k+1/2}) \le (1-\alpha) r_j(u^k) \le r_j(u^k) - \alpha \frac{r_j(u^k)^2}{r_j(u^0)}.$$

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

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

using Lemma 6.13 and applying the intermediate value theorem to W. Consequently, $u^k + s(v^k - u^k) \in E_j(u^0)$ for all $0 \le s \le \hat{s}^k$ due to the convexity of j. Because of the Lipschitz-continuity of ∇f on $E_j(u^0)$, Lemma 6.12 can be applied and, defining $\delta u^k = v^k - u^k$, there holds

$$\alpha = \frac{j(u^k) - j(u^k + \hat{s}^k \delta u^k)}{\hat{s}^k \Phi(u^k)} \ge 1 - \frac{L_{u^0}\hat{s}^k}{2} \frac{\|\delta u^k\|_{\mathcal{M}}^2}{\Phi(u^k)} \ge 1 - \frac{L_{u^0} s^k}{2\gamma} \frac{\|\delta u^k\|_{\mathcal{M}}^2}{\Phi(u^k)}.$$

The last estimate is true because of $\hat{s}^k \le s^k/\gamma$. Note that $\delta u^k \neq 0$ since $\Phi(u^k) > 0$. Reordering and using (6.25) yields

$$1 \ge s^k \ge 2\gamma(1-\alpha) \frac{\Phi(u^k)}{L_{u^0}\|v^k - u^k\|_{\mathcal{M}}^2} \ge 2\gamma(1-\alpha) \frac{r_j(u^k)}{L_{u^0}\|v^k - u^k\|_{\mathcal{M}}^2}.$$

Combining the estimates in both cases and using $r_j(u^{k+1}) \leq r_j(u^{k+1/2})$, the inequality

$$0 \leq \frac{r_j(u^{k+1})}{r_j(u^0)} \leq \frac{r_j(u^{k+1/2})}{r_j(u^0)} \leq \frac{r_j(u^k)}{r_j(u^0)} - q_k \left(\frac{r_j(u^k)}{r_j(u^0)} \right)^2 \quad \forall k \in \mathbb{N} \tag{6.29}$$

holds, where the constant q_k is given by

$$q_k = r_j(u^0)\alpha \min \left\{ \frac{2\gamma(1-\alpha)}{L_{u^0}\|v^k - u^k\|^2_{\mathcal{M}}}, \frac{1}{r_j(u^k)} \right\} \geq \alpha \min \left\{ \frac{2\gamma(1-\alpha) r_j(u^0)}{4L_{u^0}(M_0)^2}, 1 \right\} := q,$$

if $s^k < 1$ and $q_k = \alpha$ otherwise. The claimed convergence rate (6.27) now follows directly from the recursion formula (6.29), see [92, Lemma 3.1]. Following Theorem 6.11, there exists at least one subsequence (denoted by the same index) of u^k with weak* limit $\bar{u}$ and $\Phi(\bar{u}) = 0$. Since j is convex every stationary point is an optimal solution of $(\mathfrak{P}_{M_0})$ and thus a global minimizer of j. □

To close on the discussion of generalized conditional gradient methods for the solution of $(\mathfrak{P})$ we point out to quantitative convergence statements if the smoothness assumptions on the gradient are relaxed. In particular the assumption on the Lipschitz continuity of ∇f can be replaced by

A6.7 The gradient ∇f is Hölder-continuous of order $\kappa \in (0,1]$ on sublevel sets i.e. for $u_0 \in \operatorname{dom} j$ there exists a constant $L_{u_0} > 0$ only depending on $j(u_0)$ with

$$\|\nabla f(u_1) - \nabla f(u_2)\|_{\mathcal{C}} \leq L_{u_0} \|u_1 - u_2\|^{\kappa}_{\mathcal{M}} \quad \forall u_1, u_2 \in E_j(u_0),$$

which yields a reduced rate of convergence

$$r_j(u^k) \leq \frac{r_j(u^0)}{(1+\kappa q k)^{\kappa}}, \quad q = \alpha \min \left\{ \sqrt[\kappa]{\frac{(1+\kappa)\gamma(1-\alpha) r_j(u^0)}{L_{u^0}(2M_0)^{1+\kappa}}}, 1 \right\}. \tag{6.30}$$

This result can be established along the same lines as in the Lipschitz-continuous case. We outline the necessary steps for the sake of completeness. If ∇f is Hölder-continuous of order $\kappa \in (0,1]$ on the sublevel sets of j the estimate in (6.26) generalizes to

$$j(u^k_s) - j(u^k) \leq -s\Phi(u^k) + \frac{L_{u^0}}{1+\kappa}(s\|u^k - v^k\|_{\mathcal{M}})^{1+\kappa}.$$

If the step size s^k in the k-th iteration of the GCG method is equal to one we immediately get

$$\frac{r_j(u^{k+1})}{r_j(u^0)} \leq \frac{r_j(u^{k+1/2})}{r_j(u^0)} \leq \frac{r_j(u^k)}{r_j(u^0)} - \alpha \left(\frac{r_j(u^k)}{r_j(u^0)} \right)^{1+1/\kappa}.$$

For $s^k < 1$ we conclude

$$s^k \geq \sqrt[\kappa]{\frac{(1+\kappa)\gamma(1-\alpha)}{L_{u^0}\|v^k - u^k\|^{1+\kappa}_{\mathcal{M}}}}\, r_j(u^k)^{1/\kappa}.$$

Combining these observations we obtain

$$\frac{r_j(u^{k+1})}{r_j(u^0)} \leq \frac{r_j(u^{k+1/2})}{r_j(u^0)} \leq \frac{r_j(u^k)}{r_j(u^0)} - q_k \left(\frac{r_j(u^k)}{r_j(u^0)} \right)^{1+1/\kappa}$$

where the constant q_k is equal to α if $s^k = 1$ and

$$q_k = \alpha r_j(u^0)^{1/\kappa} \min\left\{ \sqrt[\kappa]{\frac{(1+\kappa)\gamma(1-\alpha)}{L_{u^0}\|v^k - u^k\|_{\mathcal{M}}^{1+\kappa}}}, \frac{1}{r_j(u^k)^{1/\kappa}} \right\} \geq \alpha \min\left\{ \sqrt[\kappa]{\frac{(1+\kappa)\gamma(1-\alpha) r_j(u^0)}{L_{u^0}(2M_0)^{1+\kappa}}}, 1 \right\}$$

if $s^k < 1$. The convergence rate in (6.30) now follows from [212, Lemma 6]. To our knowledge the only paper providing similar results for conditional gradient methods on functions with Hölder continuous gradient is the recent preprint [274].

Remark 6.5. Consider the minimization of a differentiable function on a ball of radius $M_0 > 0$ i.e.

$$\min_{\|u\|_{\mathcal{M}} \leq M_0} f(u).$$

Recently, see e.g [159], sublinear convergence rates for the classical conditional gradient method based on the assumption of bounded curvature

$$C_f = \sup_{\substack{\|u\|_{\mathcal{M}}, \|v\|_{\mathcal{M}} \leq M_0 \\ s \in (0,1] \\ u_s = u + s(v-u)}} [\frac{2}{s^2}(f(u_s) - f(u) - s\langle \nabla f(u), v - u\rangle)] < \infty,$$

were established. This assumption is weaker than requiring Lipschitz continuity of the gradient in the sense that the Lipschitz constant of ∇f on the ball gives an upper bound for the curvature constant. In particular there holds

$$f(u_s) - f(u) \leq s\langle \nabla f(u), v^k - u^k\rangle + \frac{s^2}{2} C_f,$$

which is an analogue of the estimate in (6.26). For an extension of this approach to the Hölder-continuous case and composite minimization problems we refer to [274]. We emphasize that a straightforward adaption of this concept to the setting considered in this chapter is not possible since $\operatorname{dom} f$ will be a proper subset of the admissible set in general. Thus the definition of the curvature constant is not meaningful in the present setting. In our previous considerations we circumvented such problems by assuming the Lipschitz continuity of ∇f only on the sublevel sets of f rather than the whole ball. Second we point out that the major difficulties in the problems considered in this thesis lie in the non-reflexivity of the space $\mathcal{M}$ and not in a lack of regularity for the function f. Therefore further discussions on the topic of curvature constants are postponed to future work.

6.3 The Primal-Dual-Active-Point method

This section is devoted to the discussion of generalized conditional gradient methods for minimization problems with measure valued optimization variables. To this end we consider sparse minimization problems of the form

$$\min_{u \in \mathcal{M}_{ad}} [F(Ku) + G(\|u\|_{\mathcal{M}})],$$

where $\mathcal{M}_{ad}$ is a subset of the space of vector measures $\mathcal{M}(\Omega, H)$ on the spatial domain Ω which take values in a Hilbert space H. The operator K is assumed to be linear and continuous and F

is a differentiable function. In contrast the second term $G(\|u\|_{\mathcal{M}})$ is in general non-smooth. For the precise assumptions on the appearing functionals and operators we refer to Section 6.3.2.

The rest of this section is structured as follows. In Section 6.3.1 a brief introduction to the necessary theory on Hilbert space valued vector measures is given. Section 6.3.2 focuses on the application of Algorithm 8 to sparse minimization problems. It turns out that the method can be based on a sequence of finitely supported measures. In each step of the algorithm a single new Dirac delta function is added to the current iterate based on the solution of a partially linearized problem. Worst case convergence results for these kind of methods are derived. The remainder of the chapter is devoted to a discussion and rigorous analysis of additional improvement steps to augment the procedure. If the operator K has finite rank we prove existence of a minimizer comprising only a finite number of Dirac delta functions in Section 6.3.3. In this case the method is implemented with a sparsification step which ensures a uniform bound on the number of support points in the iterates. This guarantees convergence to a finitely supported stationary point. Second we propose an accelerated variant of the GCG method which is based on alternating between adding new Dirac delta functions and optimizing their coefficients in Section 6.3.4. Imposing additional structural assumptions on the problem this new *Primal-Dual-Active-Point method* yields linear convergence of the objective function values see Section 6.3.5. Moreover we also quantify the convergence of the iterates through several criteria such as the distance of their support points to the optimal ones.

To close this introductory part we briefly reflect on comparable results from the literature and the major novelties of the present work. Generalized conditional gradient methods for concrete realizations of the presented setting have recently received considerable attention. We refer e.g. to [44, 50, 97, 200, 209]. In all of these papers a sublinear convergence rate for the objective function values is proven. We also mention the early work of Fedorov and Wynn, [105, 272] and subsequent papers, e.g. [197, 269, 270, 276], on comparable algorithms in the context of approximate design theory in statistics. Most of these prior works consider scalar valued measures and convex objective functionals. This section aims to extend conditional gradient methods to the case of general vector-valued measures and provide convergence results for convex and nonconvex objective functionals. In this context the main contributions of the present work lie in the improved convergence statements contained in Sections 6.3.3 and 6.3.4. In particular we emphasize that acceleration steps for the GCG method similar to those in Section 6.3.4 were already proposed in [44, 50, 97, 209, 270]. However to the best of our knowledge this work is the first to improve on the usual sublinear worst-case convergence rate of the objective function values for conditional gradient methods in this case. Additionally we are not aware of any approaches to quantify the convergence of the iterates or to guarantee the uniform boundedness of their support size.

6.3.1 Vector measures

For the rest of this chapter let $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, be compact and denote by H a separable Hilbert space with respect to the norm $\|\cdot\|_H$ induced by the scalar product $(\cdot,\cdot)_H$. In the following H is identified with its topological dual space. A countably additive mapping $u\colon \mathcal{B}(\Omega) \to H$ is called a vector measure. Associated to u we define its total variation measure as

$$|u|\colon \mathcal{B}(\Omega) \to \mathbb{R}_+, \quad |u|(O) = \sup\left\{ \sum_{i=1}^{\infty} \|u(O_i)\|_H \mid O_i \in \mathcal{B}(\Omega),\ \text{disjoint partition of } O \right\}.$$

The space of vector measures with finite total variation is now denoted by

$$\mathcal{M}(\Omega, H) = \{\, u \colon \mathcal{B}(\Omega) \to H \mid u \text{ countably additive}, \quad |u| < \infty \,\}.$$

For each vector measure $u \in \mathcal{M}(\Omega, H)$ we thus clearly have $|u| \in \mathcal{M}^+(\Omega)$. The support of u is defined as the support of the corresponding total variation measure, see Section 3.1.2,

$$\operatorname{supp} u = \operatorname{supp} |u|.$$

The space $\mathcal{M}(\Omega, H)$ is a Banach space with respect to the norm

$$\|u\|_{\mathcal{M}} := |u|(\Omega) = \||u|\|_{\mathcal{M}(\Omega)} = \int_\Omega \,\mathrm{d}|u|.$$

For a reference see the discussion in [183, Chapter 12.3]. Furthermore for $u \in \mathcal{M}(\Omega, H)$ it is easy to see that

$$\|u(O)\|_H \leq |u|(O) \quad \forall O \in \mathcal{B}(\Omega).$$

In particular this implies that u is absolutely continuous with respect to $|u|$, i.e. there holds

$$|u|(O) = 0 \Rightarrow \|u(O)\|_H = 0 \quad \forall O \in \mathcal{B}(O).$$

Moreover there exists a unique function

$$u' \in L^\infty(\Omega, |u|; H) \quad \text{with} \quad \|u(x)\|_H = 1 \quad |u| - a.e.\ x \in \Omega,$$

such that u can be decomposed as

$$u(O) = \int_O \,\mathrm{d}u = \int_O u' \,\mathrm{d}|u| \quad \forall O \in \mathcal{B}(\Omega).$$

We point out to [182, Chapter 12.4] for a reference. The function u' is called the Radon-Nikodým derivative of u with respect to $|u|$, see [87]. We refer to this splitting of u in terms of its Radon-Nikodým derivative u' and its total variation measure $|u|$ as its polar decomposition. For abbreviation we write $\mathrm{d}u = u'\mathrm{d}|u|$ in the following.

By $\mathcal{C}(\Omega, H)$ we further denote the space of bounded and continuous functions on Ω which assume values in H. It is a separable Banach space when endowed with the usual supremum norm

$$\|\varphi\|_{\mathcal{C}} = \max_{x \in \Omega} \|\varphi(x)\|_H \quad \forall \varphi \in \mathcal{C}(\Omega, H),$$

see e.g. [7, Lemma 3.85]. Following Singer's representation theorem, [135], its topological dual space is identified with $\mathcal{M}(\Omega, H)$ where the associated duality paring is given by

$$\langle \varphi, u \rangle = \int_\Omega (\varphi(x), u'(x))_H \,\mathrm{d}|u|(x) \quad \forall \varphi \in \mathcal{C}(\Omega, H),\ u \in \mathcal{M}(\Omega, H).$$

As a consequence we conclude

$$\|u\|_{\mathcal{M}} = \sup_{\substack{\varphi \in \mathcal{C}(\Omega, H) \\ \|\varphi\|_{\mathcal{C}} \leq 1}} \langle \varphi, u \rangle = \sup_{\substack{\varphi \in \mathcal{C}(\Omega, H) \\ \|\varphi\|_{\mathcal{C}} \leq 1}} \int_\Omega (\varphi(x), u'(x))_H \,\mathrm{d}|u|(x).$$

The duality relation between the space of vector measures and the space of H-valued continuous functions allows to consider minimization problems over $\mathcal{M}(\Omega, H)$ in the general framework presented in the previous sections. In this context we emphasize that from an application point of view, see Example 6.6 and the presentations in the previous chapters, it is also necessary to treat situations in which the objective functional is minimized over a proper subset $\mathcal{M}_{ad} \subset \mathcal{M}(\Omega, H)$ instead of the whole space. The remainder of this section is therefore devoted to the study of admissible sets given by

$$\mathcal{M}_{ad} = \mathcal{M}(\Omega, C) = \{\, u \in \mathcal{M}(\Omega, H) \mid u(O) \in C \quad \forall O \in \mathcal{B}(\Omega) \,\} . \tag{6.31}$$

Here C denotes a closed and convex cone, see Definition 6.3, in the Hilbert space H. While we already discussed the special case of positive scalar-valued measures, i.e. $H = \mathbb{R}$ and $C = \mathbb{R}_+$, in the previous chapters the following motivational example justifies a discussion of this matter in the presented generality.

Example 6.6. *Set $I = [0, T]$, $T > 0$, and $H = L^2(I)$. In this situation our special interest lies on vector measures u given by a finite sum of Dirac delta functions on fixed points of the spatial domain with time dependent coefficients:*

$$u = \sum_{i=1}^{N} \mathbf{u}_i \delta_{x_i}, \quad N \in \mathbb{N},\ \mathbf{u}_i \in L^2(I),\ x_i \in \Omega,\ i = 1, \dots, N.$$

For example we might think of u as an ensemble of heat sources located at the positions $\{x_i\}_{i=1}^N$. The functions $\{\mathbf{u}_i\}_{i=1}^N$ represent the intensities of the individual sources. From a modeling point of view it is reasonable to choose the coefficient functions from the cone of almost everywhere non-negative functions

$$L^2_+(I) = \left\{ \mathbf{u} \in L^2(I) \mid u(t) \geq 0 \quad a.e.\ t \in I \right\}.$$

It is straightforward to see that the set of finitely supported vector measures with nonnegative coefficient functions

$$\mathcal{M}_{\mathbb{N}}(\Omega, L^2_+(I)) = \left\{ u \in \mathcal{M}(\Omega, L^2_+(I)) \mid u = \sum_{i=1}^{N} \mathbf{u}_i \delta_{x_i},\ N \in \mathbb{N},\ \mathbf{u}_i \in L^2_+(I),\ x_i \in \Omega,\ i = 1, \dots, N \right\}$$

is embedded in the larger set

$$\left\{ u \in \mathcal{M}(\Omega, L^2(I)) \mid u'(x) \in L^2_+(I) \quad |u| - a.e.\ x \in \Omega \right\}. \tag{6.32}$$

The results in this section answer some important questions regarding these sets. First we establish the equivalence between the set in (6.32) and $\mathcal{M}(\Omega, L^2_+(I))$ defined according to (6.31). Second we characterize it as the polar cone of a closed cone in $\mathcal{C}(\Omega, L^2(I))$. In particular this implies its weak closedness. Last, the embedding of $\mathcal{M}_{\mathbb{N}}(\Omega, L^2_+(I))$ into $\mathcal{M}(\Omega, L^2_+(\Omega))$ turns out to be weak* dense i.e. every vector measure $u \in \mathcal{M}(\Omega, L^2_+(I))$ can be weak*-approximated by a sequence of finitely supported ones with nonnegative coefficient functions.*

Let us first fix the notion of a convex cone in a general Banach space.

Definition 6.3. Let X be a Banach space with topological dual space X^* and duality pairing $\langle \cdot, \cdot \rangle_*$.

- A nonempty set $C \subset X$ is called a convex cone if

$$0 \in C, \quad \lambda \mathbf{u}_1 + \mathbf{u}_2 \in C, \quad \lambda \in \mathbb{R}_+ \setminus \{0\},\ \mathbf{u}_1,\ \mathbf{u}_2 \in C.$$

- The polar cone $C^o \subset X^*$ of C is defined as

$$C^o = \{\, \mathbf{x}^* \in X^* \mid \langle \mathbf{x}, \mathbf{x}^* \rangle_* \leq 0 \quad \forall \mathbf{x} \in X \,\}\,.$$

We emphasize that the polar cone of a convex cone C is closed with respect to the weak* topology on X^*. In the following $C \subset H$ will always denote a nonempty closed and convex cone. The associated H-projection onto C is defined as

$$P_C \colon H \to C, \quad \mathbf{u} \mapsto \operatorname*{arg\,min}_{v \in C} \frac{1}{2} \|v - \mathbf{u}\|_H^2.$$

The next proposition summarizes some key properties of the projection P_C.

Proposition 6.15. *Let $C \subset H$ be a convex and closed cone. Then there holds $(C^o)^o = C$ and*

$$\mathbf{u} = P_C(\mathbf{u}) + P_{C^o}(\mathbf{u}), \quad (P_{C^o}(\mathbf{u}), P_C(\mathbf{u}))_H = 0 \quad \forall \mathbf{u} \in H,$$

as well as

$$\|P_C(\mathbf{u}_1) - P_C(\mathbf{u}_2)\|_H \leq \|\mathbf{u}_1 - \mathbf{u}_2\|_H \quad \forall \mathbf{u}_1,\ \mathbf{u}_2 \in H. \tag{6.33}$$

Proof. The first statement can be found on [46, p. 53]. The remaining claims follow from the discussions in [154, Section II]. □

In particular the non-expansiveness of the projection, (6.33), implies the continuity of the function

$$P_C(\varphi) \colon \Omega \to C, \quad x \mapsto P_C(\varphi(x)), \quad \|P_C(\varphi)\|_{\mathcal{C}} \leq \|\varphi\|_{\mathcal{C}} \quad \forall \varphi \in \mathcal{C}(\Omega, H).$$

The set of continuous C- valued functions on Ω is denoted by

$$\mathcal{C}(\Omega, C) = \{\, \varphi \in \mathcal{C}(\Omega, H) \mid \varphi(x) \in C \quad \forall x \in \Omega \,\}$$

Obviously $\mathcal{C}(\Omega, C)$ is a convex cone which is closed with respect to the norm on $\mathcal{C}(\Omega, H)$.

We now turn to the study of the set $\mathcal{M}(\Omega, C)$ as defined in (6.31). Again it is straightforward to verify that $\mathcal{M}(\Omega, C)$ is a convex cone. In the following we aim at a characterization of its elements $u \in \mathcal{M}(\Omega, C)$ in terms of the Radon-Nikodým derivative u' with respect to $|u|$. To this end consider an arbitrary $u \in \mathcal{M}(\Omega, H)$ with $\mathrm{d}u = u'\ \mathrm{d}|u|$. Given $\varepsilon > 0$ define the averaged integral of u' by

$$D(u', x, \varepsilon) = \frac{1}{|u|(B_\varepsilon(x))} u(B_\varepsilon(x)) = \frac{1}{|u|(B_\varepsilon(x))} \int_{B_\varepsilon(x)} u'\ \mathrm{d}|u|(x) \quad |u| - a.e.\ x \in \Omega. \tag{6.34}$$

Remark 6.6. We briefly point out that the integral in (6.34) is indeed well-defined in a $|u|$ almost everywhere sense since

$$|u|\,(\{\, x \in \Omega \mid \exists \varepsilon > 0 \colon |u|(B_\varepsilon(x)) = 0 \,\}) = 0.$$

This statement is implicitly contained in the proof of Theorem 1.29 in [104].

We arrive at the following vector-valued version of the Lebesgue differentiation theorem.

Proposition 6.16. *Let $u \in \mathcal{M}(\Omega, H)$ with polar decomposition $\mathrm{d}u = u' \, \mathrm{d}|u|$ be given. Then there holds*

$$\lim_{\varepsilon \to^+ 0} \|u'(x) - D(u', x, \varepsilon)\|_H = 0, \quad |u| - a.e.\ x \in \Omega \tag{6.35}$$

Proof. Let $\{h_k\}_{k \in \mathbb{N}}$ denote a dense subset of H. For every $k \in \mathbb{N}$ there exists a set $O_k \in \mathcal{B}(\Omega)$ with $|u|(O_k) = 0$ as well as

$$\lim_{\varepsilon \to^+ 0} \frac{1}{|u|(B_\varepsilon(x))} \int_{B_\varepsilon(x)} \|u'(y) - h_k\|_H \; \mathrm{d}|u|(y) = \|u'(x) - h_k\|_H \quad \forall x \in \Omega \setminus O_k,$$

following the scalar version of the Lebesgue differentiation theorem c.f. [104, Theorem 1.33]. Define the set $O = \bigcup_{i \in \mathbb{N}} O_n$ and let $\delta > 0$ as well as $x \in \Omega \setminus O$ be given. Choose $k \in \mathbb{N}$ such that $\|u'(x) - h_k\|_H < \delta/2$. Then $|u|(O) = 0$ and there holds

$$\begin{aligned}\lim_{\varepsilon \to^+ 0} \|u'(x) - D(u', x, \varepsilon)\|_H &\leq \lim_{\varepsilon \to^+ 0} \frac{1}{|u|(B_\varepsilon(x))} \int_{B_\varepsilon(x)} \|u'(y) - u'(x)\|_H \; \mathrm{d}|u|(y) \\ &\leq \lim_{\varepsilon \to^+ 0} \frac{1}{|u|(B_\varepsilon(x))} \int_{B_\varepsilon(x)} [\|u'(y) - h_k\|_H + \|u'(x) - h_k\|_H] \; \mathrm{d}|u|(y) \\ &= 2\|u'(x) - h_k\|_H < \delta.\end{aligned}$$

Since O is a $|u|$ null set and $\delta > 0$ was chosen arbitrary the statement follows. □

The following theorem is a direct consequence.

Theorem 6.17. *Let $u \in \mathcal{M}(\Omega, C)$ with polar decomposition $\mathrm{d}u = u' \; \mathrm{d}|u|$ be given. Then there holds*

$$u'(x) \in C \quad |u| - a.e.\ x \in \Omega.$$

Vice versa if we have $u \in \mathcal{M}(\Omega, H)$ with $u'(x) \in C$ for $|u|$-a.e $x \in \Omega$ then $u \in \mathcal{M}(\Omega, C)$.

Proof. Let $u \in \mathcal{M}(\Omega, C)$ and $\varepsilon > 0$ be given. Then there holds

$$D(u', x, \varepsilon) = \frac{1}{|u|(B_\varepsilon(x))} u(B_\varepsilon(x)) \in C \quad |u| - a.e.\ x \in \Omega$$

since $u(O) \in C$ for all $O \in \mathcal{B}(\Omega)$ and $|u|(B_\varepsilon(x)) > 0$ for $|u|$-a.e $x \in \Omega$. In perspective of the results in the previous proposition we thus conclude

$$u'(x) = \lim_{\varepsilon \to^+ 0} D(u', x, \varepsilon) \in C, \quad |u| - a.e.\ x \in \Omega,$$

since C is closed.

Conversely assume that $u \in \mathcal{M}(\Omega, H)$ with $u'(x) \in C$ for $|u|$-a.e $x \in \Omega$. Let an arbitrary $\mathbf{u} \in C^o$ be given. By definition of the Bochner integral we obtain

$$(\mathbf{u}, u(O))_H = \int_O \underbrace{(\mathbf{u}, u'(x))_H}_{\leq 0} \mathrm{d}|u|(x) \leq 0$$

for $O \in \mathcal{B}(\Omega)$. Since $\mathbf{u} \in C^o$ was chosen arbitrary we conclude $u(O) \in (C^o)^o = C$ for every set $O \in \mathcal{B}(\Omega)$. This yields $u \in \mathcal{M}(\Omega, C)$. □

Combining both statements we arrive at

$$\mathcal{M}(\Omega, C) = \left\{ u \in \mathcal{M}(\Omega, C) \mid u'(x) \in C \quad |u| - a.e. \ x \in \Omega \right\}.$$

The following theorem establishes the weak* closedness of $\mathcal{M}(\Omega, C)$.

Theorem 6.18. *Let $C \subset H$ be a nonempty closed and convex cone. Then there holds*

$$\mathcal{C}(\Omega, C^o)^o = \mathcal{M}(\Omega, C). \tag{6.36}$$

In particular $\mathcal{M}(\Omega, C)$ is weak closed.*

Proof. As already discussed before $\mathcal{C}(\Omega, C^o)^o$ is a closed and convex cone in $\mathcal{C}(\Omega, H)$. Thus we obtain the second statement directly by proving the first one. Given an arbitrary $u \in \mathcal{M}(\Omega, C)$ and $\varphi \in \mathcal{C}(\Omega, C^o)$ expanding the duality pairing yields

$$\langle \varphi, u \rangle = \int_\Omega (\varphi(x), u'(x))_H \ \mathrm{d}|u|(x) \leq 0,$$

since the integrand is nonpositive for $|u|$-a.e $x \in \Omega$. Therefore there holds $\mathcal{M}(\Omega, C) \subset \mathcal{C}(\Omega, C^o)^o$. Now let an arbitrary $u \in \mathcal{C}(\Omega, C^o)^o$ and a compact set $O \in \mathcal{B}(\Omega)$ be given. If $u(O) \notin C$ there exists $\mathbf{u} \in C^o$ with

$$\alpha := (\mathbf{u}, u(O))_H > 0.$$

Since $|u|$ is regular there are an open set $O_2 \in \mathcal{B}(\Omega)$ and a bump function $\chi \in \mathcal{C}(\Omega)$ with

$$O \subset O_2, \quad |u|(O_2 \setminus O) \leq \frac{\alpha}{2}, \quad \chi(x) \begin{cases} = 1 & x \in O \\ = 0 & x \in \Omega \setminus O_2 \\ \in [0,1] & \text{else} \end{cases}.$$

By construction the function $\varphi_{\mathbf{u}} = \mathbf{u}\chi$ is an element of $\mathcal{C}(\Omega, C^o)$ and

$$\langle \varphi_{\mathbf{u}}, u \rangle \geq \alpha - |u|(O_2 \setminus O) \geq \frac{\alpha}{2} > 0.$$

Thus we conclude

$$u(O) \in C, \quad \forall O \in \mathcal{B}(\Omega), \ O \text{ compact}.$$

If $O \in \mathcal{B}(\Omega)$ is an arbitrary Borel set the regularity of $|u|$ yields the existence of a compact set O_1^ε and an open set O_2^ε with

$$O_1^\varepsilon \subset O \subset O_2^\varepsilon \subset \Omega, \quad |u|(O_2^\varepsilon \setminus O_1^\varepsilon) < \varepsilon,$$

for every $\varepsilon > 0$. This implies

$$\|u(O) - u(O_1^\varepsilon)\|_H = \|u(O \setminus O_1^\varepsilon)\|_H \leq |u|(O \setminus O_1^\varepsilon) \leq |u|(O_2^\varepsilon \setminus O_1^\varepsilon) < \varepsilon.$$

Since $u(O_1^\varepsilon) \in C$, $\varepsilon > 0$, and C is closed this implies $u(O) \in C$. Therefore $u \in \mathcal{M}(\Omega, C)$ has to hold finishing the proof. □

As a consequence the cone $\mathcal{M}(\Omega, C)$ can be identified with the weak* closure of the cone of finitely supported C-valued vector measures.

Proposition 6.19. *Define the set*

$$\mathcal{M}_{\mathbb{N}}(\Omega, C) = \left\{ u \in \mathcal{M}(\Omega, C) \mid u = \sum_{i=1}^{N} \mathbf{u}_i \delta_{x_i},\ N \in \mathbb{N},\ \mathbf{u}_i \in C,\ x_i \in \Omega,\ i = 1, \dots, N \right\}.$$

Let $u \in \mathcal{M}(\Omega, C)$ be given. Then there exists $\{u_k\}_{k\in\mathbb{N}} \subset \mathcal{M}_{\mathbb{N}}(\Omega, C)$ fulfilling $\|u_k\|_{\mathcal{M}} \leq \|u\|_{\mathcal{M}}$ and $u_k \rightharpoonup^* u$. In particular we have

$$\overline{\mathcal{M}_{\mathbb{N}}(\Omega, C)}^* = \mathcal{M}(\Omega, C).$$

Proof. The proof is a slight adaptation of the corresponding one for the case of $C = \mathbb{R}^m$, $m \in \mathbb{N}$, presented in [50, Appendix A]. Let $u \in \mathcal{M}(\Omega, C)$, $u \neq 0$, be given. Fix an arbitrary index $k \in \mathbb{N}$. For $x \in \Omega$ we define $Q^k_x = 2^{-k}] - 1/2, 1/2]^d$ and set

$$u_k = \sum_{x \in 2^{-k}\mathbb{Z}^d} u(Q^k_x \cap \Omega)\delta_x \in \mathcal{M}_{\mathbb{N}}(\Omega, C).$$

It is straightforward to see that

$$\bigcup_{x \in 2^{-k}\mathbb{Z}^d} (Q^k_x \cap \Omega) = \Omega, \quad (Q^k_{x_1} \cap \Omega) \cap (Q^k_{x_2} \cap \Omega) = \emptyset \quad \forall x_1,\ x_2 \in 2^{-k}\mathbb{Z}^d,$$

and thus

$$\|u_k\|_{\mathcal{M}} = \sum_{x \in 2^{-k}\mathbb{Z}} \|u(Q^k_x \cap \Omega)\|_H \leq |u|(\Omega) = \|u\|_{\mathcal{M}}.$$

Consider an arbitrary $\varphi \in \mathcal{C}(\Omega, H)$ and $\varepsilon > 0$. Since Ω is compact there holds

$$|\varphi(y) - \varphi(x)| \leq \frac{\varepsilon}{\|u\|_{\mathcal{M}}} \quad \forall x \in 2^{-k}\mathbb{Z}^d,\ y \in Q^k_x \cap \Omega,$$

and all $k \in \mathbb{N}$ large enough. We estimate

$$|\langle \varphi, u - u_k \rangle| \leq \sum_{x \in 2^{-k}\mathbb{Z}^d} \int_{Q^k_x \cap \Omega} |(\varphi(y) - \varphi(x), u'(y))_H| \mathrm{d}|u|(y) \leq \varepsilon.$$

Since $\varepsilon > 0$ and $\varphi \in \mathcal{C}(\Omega, H)$ were chosen arbitrary we conclude the first claimed statement. From the weak* closedness of $\mathcal{M}(\Omega, C)$ we further obtain

$$\mathcal{M}(\Omega, C) \subset \overline{\mathcal{M}_{\mathbb{N}}(\Omega, C)}^* \subset \mathcal{M}(\Omega, C).$$

This finishes the proof. □

6.3.2 Generalized conditional gradient methods for vector measures

We now turn to the study of the sparse minimization problem

$$\min_{u\in\mathcal{M}(\Omega,H)} j(u) := [F(Ku) + G(\|u\|_{\mathcal{M}}) + I_{\mathcal{M}(\Omega,C)}(u)] \tag{$\mathfrak{P}^{\mathcal{M}}$}$$

Here $I_{\mathcal{M}(\Omega,C)}$ denotes the indicator function of the set $\mathcal{M}(\Omega,C)$. In order to ensure well-posedness of this problem the following standing assumptions are made.

Assumption 6.3. Let the following assumptions hold.

A6.8 Let Y be a Hilbert space and let $K\colon \mathcal{M}(\Omega,H) \to Y$ be a linear and weak*-to-strong continuous operator with adjoint $K^*\colon Y \to \mathcal{C}(\Omega,H)$.

A6.9 The function $G\colon \mathbb{R} \to \mathbb{R}$ is proper, convex, lower semi-continuous, and monotonically increasing on $\mathbb{R}_+$ with $G(t) \to \infty$ for $t \to \infty$. There holds $\operatorname{dom} G \subset \mathbb{R}_+$.

A6.10 The set $C \subset H$ is a nonempty, closed and convex cone. Furthermore the domain of the functional j is nonempty and j is radially unbounded.

A6.11 The function $F\colon Y \to \mathbb{R} \cup \{\,+\infty\,\}$ is lower semi-continuous on

$$Y_{ad} := \{\, Ku \mid u \in \operatorname{dom} G(\|\cdot\|_{\mathcal{M}}) \cap \mathcal{M}(\Omega,C)\,\}\,.$$

Moreover, F is continuously Fréchet differentiable on

$$\widehat{Y}_{ad} := \{\, Ku \mid u \in \operatorname{dom} j\,\}.$$

The set $\widehat{Y}_{ad}$ is open in Y_{ad}. The Fréchet derivative of F at $y \in \widehat{Y}_{ad}$ will be denoted by $\nabla F(y)$.

Remark 6.7. Note that these general assumptions on the convex function G in particular allow for a unified treatment of norm penalized problems $G_1(\|u\|_{\mathcal{M}}) = \beta\|u\|_{\mathcal{M}} + I_{[0,\infty)}(\|u\|_{\mathcal{M}})$ for $\beta > 0$ and norm constraint problems $G_2(\|u\|_{\mathcal{M}}) = I_{[0,M_0]}(\|u\|_{\mathcal{M}})$, $M_0 > 0$.

Corollary 6.20. *The functions $f = F \circ K$ and $g = G \circ \|\cdot\|_{\mathcal{M}} + I_{\mathcal{M}(\Omega,C)}$ fulfill Assumption 6.1.*

Proof. The claimed statement follows immediately noting that the weak* closedness and convexity of $\mathcal{M}(\Omega,C)$ imply the weak* lower semi-continuity of $I_{\mathcal{M}(\Omega,C)}$ on $\mathcal{M}(\Omega,H)$ and applying the chain rule yields

$$f'(u)(\delta u) = \langle K^*\nabla F(Ku), \delta u\rangle \quad \forall \delta u \in \mathcal{M}(\Omega,H),$$

for $u \in \operatorname{dom} j$. □

As a consequence existence of minimizers as well as first order necessary optimality conditions can be obtained from the general results in Propositions 6.2 and 6.3.

Proposition 6.21. *Let Assumption 6.3 hold. There exists at least one optimal solution $\bar{u} \in \mathcal{M}(\Omega,C)$ to ($\mathfrak{P}^{\mathcal{M}}$). Set $\bar{p} = -K^*\nabla F(K\bar{u}) \in \mathcal{C}(\Omega,H)$. Then there holds*

$$\langle \bar{p}, u - \bar{u}\rangle + G(\|\bar{u}\|_{\mathcal{M}}) \le G(\|u\|_{\mathcal{M}}) \quad \forall u \in \mathcal{M}(\Omega,C). \tag{6.37}$$

Throughout the rest of this chapter we will refer to $\bar{y} = K\bar{u}$ as the optimal state and, with a slight abuse of notation, to the continuous function $\bar{p} = -K^*\nabla F(K\bar{u})$ as the adjoint state associated to $\bar{u}$. Let us turn to a structural characterization of minimizers obtained from $(\mathfrak{P}^{\mathcal{M}})$.

Theorem 6.22. *Let $\bar{u} \in \operatorname{dom} j \subset \mathcal{M}(\Omega, C)$ be given. Then* (6.37) *holds if and only if*

$$\langle \bar{p}, \bar{u} \rangle = \|P_C(\bar{p})\|_{\mathcal{C}} \|\bar{u}\|_{\mathcal{M}}, \quad \|P_C(\bar{p})\|_{\mathcal{C}} \in \partial G(\|\bar{u}\|_{\mathcal{M}}) \tag{6.38}$$

Proof. First assume that (6.38) holds for $\bar{u} \in \operatorname{dom} j$. Let an arbitrary $u \in \mathcal{M}(\Omega, C)$ be given. We estimate

$$\langle \bar{p}, u \rangle = \int_\Omega (\bar{p}(x), u'(x))_H \, \mathrm{d}|u|(x) \leq \int_\Omega (P_C(\bar{p}(x)), u'(x))_H \, \mathrm{d}|u|(x) \leq \|P_C(\bar{p})\|_{\mathcal{C}} \|u\|_{\mathcal{M}}.$$

Putting everything together yields

$$\begin{aligned}\langle \bar{p}, u - \bar{u} \rangle + G(\|\bar{u}\|_{\mathcal{M}}) &= -\|P_C(\bar{p})\|_{\mathcal{C}} \|\bar{u}\|_{\mathcal{M}} + \langle \bar{p}, u \rangle + G(\|\bar{u}\|_{\mathcal{M}}) \\ &\leq \|P_C(\bar{p})\|_{\mathcal{C}} (\|u\|_{\mathcal{M}} - \|\bar{u}\|_{\mathcal{M}}) + G(\|\bar{u}\|_{\mathcal{M}}) \\ &\leq G(\|u\|_{\mathcal{M}}).\end{aligned}$$

Since $u \in \mathcal{M}(\Omega, C)$ was chosen arbitrary the variational inequality (6.37) follows.

Conversely assume that (6.37) holds. First let $\bar{u} \neq 0$ hold. From the monotonicity of G we infer

$$\langle \bar{p}, u - \bar{u} \rangle \leq 0 \quad \forall u \in \mathcal{M}(\Omega, C), \ \|u\|_{\mathcal{M}} \leq \|\bar{u}\|_{\mathcal{M}}$$

or, equivalently,

$$\bar{p} \in \partial (I_{\mathcal{M}(\Omega,C)}(\cdot) + I_{\|\cdot\|_{\mathcal{M}} \leq \|\bar{u}\|_{\mathcal{M}}}(\cdot))(\bar{u}).$$

Applying Proposition 6.4 yields

$$(I_{\mathcal{M}(\Omega,C)} + I_{\|\cdot\|_{\mathcal{M}} \leq \|\bar{u}\|_{\mathcal{M}}})^*(\bar{p}) = \sup_{\substack{u \in \mathcal{M}(\Omega,C) \\ \|u\|_{\mathcal{M}} \leq \|\bar{u}\|_{\mathcal{M}}}} \langle \bar{p}, u \rangle = \langle \bar{p}, \bar{u} \rangle.$$

For an arbitrary measure $u \in \mathcal{M}(\Omega, C), \ \|u\|_{\mathcal{M}} \leq \|\bar{u}\|_{\mathcal{M}}$, we readily obtain

$$\langle \bar{p}, u \rangle \leq \int_\Omega (P_C(\bar{p}(x)), u'(x))_H \, \mathrm{d}|u|(x) \leq \|P_C(\bar{p}(x))\|_{\mathcal{C}} \|\bar{u}\|_{\mathcal{M}}. \tag{6.39}$$

Let $\hat{x} \in \Omega$ with $\|P_C(\bar{p}(\hat{x}))\|_H = \|P_C(\bar{p})\|_{\mathcal{C}}$ be given and define

$$\tilde{u} = \|\bar{u}\|_{\mathcal{M}} \begin{cases} 0 & P_C(\bar{p}) = 0 \\ \frac{P_C(\bar{p}(\hat{x}))}{\|P_C(\bar{p})\|_{\mathcal{C}}} \delta_{\hat{x}} & P_C(\bar{u}) \neq 0 \end{cases} \in \mathcal{M}(\Omega, C).$$

We claim that $\tilde{u}$ achieves equality in (6.39). If $P_C(\bar{p}) = 0$ this trivially holds. In the second case we compute

$$\langle \bar{p}, \tilde{u} \rangle = \|\bar{u}\|_{\mathcal{M}} \frac{(P_C(\bar{p}(\hat{x})) + P_{C^\circ}(\bar{p}(\hat{x})), P_C(\bar{p}(\hat{x})))_H}{\|P_C(\bar{p})\|_{\mathcal{C}}} = \|\bar{u}\|_{\mathcal{M}} \|P_C(\bar{p})\|_{\mathcal{C}},$$

where we used $(P_{C^o}(\bar{p}(x)), P_C(\bar{p}(x)))_H = 0$, $x \in \Omega$. Consequently we conclude

$$\langle \bar{p}, \bar{u} \rangle = \|\bar{u}\|_{\mathcal{M}} \|P_C(\bar{p})\|_{\mathcal{C}}.$$

In a similar way we get

$$\sup_{\substack{u \in \mathcal{M}(\Omega, C) \\ \|u\|_{\mathcal{M}} \leq m}} \langle \bar{p}, u \rangle = m \|P_C(\bar{p})\|_{\mathcal{C}} \quad \forall m \in \mathbb{R}_+.$$

Combining these results the variational inequality (6.37) can be reformulated as

$$\|P_C(\bar{p})\|_{\mathcal{C}}(m - \|\bar{u}\|_{\mathcal{M}}) + G(\|\bar{u}\|_{\mathcal{M}}) \leq G(m) \quad \forall m \in \mathbb{R}_+$$

By definition of the subdifferential and $\operatorname{dom} G \subset \mathbb{R}_+$ this yields the second condition in (6.38). The case $\bar{u} = 0$ follows by similar arguments finishing the proof. □

Example 6.7. *For the the examples of norm regularization $G_1(\|u\|_{\mathcal{M}}) = \beta \|u\|_{\mathcal{M}}$ and norm constraints $G_2(\|u\|_{\mathcal{M}}) = I_{[0,M_0]}(\|u\|_{\mathcal{M}})$ the subdifferential inclusions in* (6.38) *are given by*

$$\|P_C(\bar{p})\|_{\mathcal{C}} \in \partial G_1(\|\bar{u}\|_{\mathcal{M}}) = \begin{cases} \{\beta\} & \|\bar{u}\|_{\mathcal{M}} \neq 0 \\ [0, \beta] & \|\bar{u}\|_{\mathcal{M}} = 0 \end{cases},$$

$$\|P_C(\bar{p})\|_{\mathcal{C}} \in \partial G_2(\|\bar{u}\|_{\mathcal{M}}) = \begin{cases} \{0\} & \|\bar{u}\|_{\mathcal{M}} \in [0, M_0) \\ [0, +\infty) & \|\bar{u}\|_{\mathcal{M}} = M_0 \end{cases}.$$

The first condition in (6.38) can be equivalently expressed through a sparsity condition on the total variation measure $|\bar{u}|$ and a projection formula for the Radon-Nikodým derivative $\bar{u}'$.

Proposition 6.23. *Let $\varphi \in \mathcal{C}(\Omega, H)$ and $u \in \mathcal{M}(\Omega, C)$ with polar decomposition $\mathrm{d}u = u' \mathrm{d}|u|$ be given. Then the following two statements are equivalent:*

- *There holds*

$$\langle \varphi, u \rangle = \|P_C(\varphi)\|_{\mathcal{C}} \|u\|_{\mathcal{M}}. \tag{6.40}$$

- *There holds*

$$\operatorname{supp} |u| \subset \{\, x \in \Omega \mid \|P_C(\varphi(x))\|_H = \|P_C(\varphi)\|_{\mathcal{C}} \,\}, \tag{6.41}$$

  as well as

$$\left.\begin{array}{ll} u'(x) = \frac{1}{\|P_C(\varphi)\|_{\mathcal{C}}} P_C(\varphi(x)) & \textit{if } \|P_C(\varphi)\|_{\mathcal{C}} \neq 0 \\ (P_{C^o}(\varphi(x)), u'(x))_H = 0 & \textit{if } \|P_C(\varphi)\|_{\mathcal{C}} = 0 \end{array}\right\} \quad |u| - a.e.\ x \in \Omega. \tag{6.42}$$

Proof. Assume that (6.40) holds. If $\|P_C(\varphi)\|_{\mathcal{C}} = 0$ the support condition in (6.41) becomes trivial and

$$\langle \varphi, u \rangle = \langle P_{C^o}(\varphi), u \rangle = \int_\Omega (P_{C^o}(\varphi(x)), u'(x))_H \mathrm{d}|u|(x) = 0$$

Since the integrand is non-positive it vanishes $|u|$-almost everywhere. This yields (6.42) in this case. Let $\|P_c(\varphi)\|_{\mathcal{C}} \neq 0$. We readily observe that

$$\|P_C(\varphi)\|_{\mathcal{C}}\|u\|_{\mathcal{M}} = \langle \varphi, u\rangle \leq \langle P_C(\varphi), u\rangle \leq \|P_C(\varphi)\|_{\mathcal{C}}\|u\|_{\mathcal{M}}.$$

Therefore there holds

$$\langle P_C(\varphi), u\rangle = \|P_C(\varphi)\|_{\mathcal{C}}\|u\|_{\mathcal{M}}.$$

Rearranging this equality and writing out the duality paring yields

$$\int_\Omega [(P_C(\varphi(x)), u'(x))_H - \|P_C(\varphi)\|_{\mathcal{C}}] \,\mathrm{d}|u|(x) = 0. \tag{6.43}$$

By estimating

$$(P_C(\varphi(x)), u'(x))_H \leq \|P_C(\varphi(x))\|_H \|u'(x)\|_H \leq \|P_C(\varphi)\|_{\mathcal{C}}, \tag{6.44}$$

it follows that the integrand in (6.43) is non-positive and thus vanishes for $|u|$-a.e. $x \in \Omega$. Accordingly there holds

$$(P_C(\varphi)(x), u'(x))_H = \|P_C(\varphi)\|_{\mathcal{C}} \quad |u| - a.e.\ x \in \Omega.$$

In perspective of (6.44) this can only be valid if

$$\|P_C(\varphi)(x)\|_H = \|P_C(\varphi)\|_{\mathcal{C}}, \quad u'(x) = \frac{1}{\|P_C(\varphi)\|_{\mathcal{C}}} P_C(\varphi)(x),$$

for $|u|$-almost all $x \in \Omega$. Therefore (6.42) holds. It remains to show the inclusion for $\operatorname{supp}|u|$ in (6.41). W.l.o.g assume $u \neq 0$. To this end we note that the function

$$h\colon \Omega \to \mathbb{R}_-, \quad h(x) = \|P_C(\varphi(x))\|_H - \|P_C(\varphi)\|_{\mathcal{C}},$$

is continuous, non-negative and its integral with respect to $|u|$ vanishes. Let an arbitrary point $\hat{x} \in \Omega$ with $h(\hat{x}) < 0$ be given. Since h is continuous this holds in a whole neighborhood $B_\delta(\hat{x})$. Let an arbitrary nonnegative function $y \in \mathcal{C}_0(B_\delta(\hat{x}))$ be given. Then there exists $t > 0$ small enough such that $h + ty \leq 0$ on Ω. We conclude

$$0 \geq \langle h + ty, u\rangle = t\langle y, u\rangle \geq 0.$$

Due to the arbitrary choice of y this implies $|u|_{|B_\delta(\hat{x})} = 0$ and $B_\delta(\hat{x}) \subset \Omega \setminus \operatorname{supp}|u|$.

Conversely let (6.41) and (6.42) hold. If $\|P_C(\varphi)\|_{\mathcal{C}} = 0$ we immediately get

$$\langle \varphi, u\rangle = \int_\Omega (P_{C^\circ}(\varphi(x)), u'(x))_H \,\mathrm{d}|u|(x) = 0 = \|P_C(\varphi)\|_{\mathcal{C}}\|u\|_{\mathcal{M}}.$$

In the second case, for $\|P_C(\varphi)\|_{\mathcal{C}} \neq 0$, we split the integral to obtain

$$\begin{aligned}\langle \varphi, u\rangle &= \int_\Omega (P_C(\varphi(x)), u'(x))_H \,\mathrm{d}|u|(x) + \int_\Omega (P_{C^\circ}(\varphi(x)), u'(x))_H \,\mathrm{d}|u|(x) \\ &= \frac{1}{\|P_C(\varphi)\|_{\mathcal{C}}} \int_\Omega (P_C(\varphi(x)), P_C(\varphi(x)))_H \,\mathrm{d}|u|(x) \\ &= \|P_C(\varphi)\|_{\mathcal{C}}\|u\|_{\mathcal{M}}.\end{aligned}$$

Here we again used that $(P_{C^\circ}(\varphi(x)), P_C(\varphi(x)))_H = 0$ for $|u|$-almost every $x \in \Omega$. This finishes the proof. □

Throughout the following discussions we will restrict ourselves to optimal vector measures $\bar{u} \neq 0$ with non-degenerate adjoint state $\bar{p}$, i.e $\|P_C(\bar{p})\|_{\mathcal{C}} \neq 0$. As a consequence of the previous proposition the optimality of $\bar{u} \in \mathcal{M}(\Omega, C)$ is characterized by necessary conditions on its polar decomposition.

Theorem 6.24. *Let $\bar{u}$ be an optimal solution to $(\mathfrak{P}^{\mathcal{M}})$ with polar decomposition $\mathrm{d}\bar{u} = \bar{u}'\mathrm{d}|\bar{u}|$ and $P_C(\bar{u}) \neq 0$. Then we have*

$$\|P_C(\bar{p})\|_{\mathcal{C}} \in \partial G(\|\bar{u}\|_{\mathcal{M}}),$$

as well as

$$\operatorname{supp}|\bar{u}| \subset \{\, x \in \Omega \mid \|P_C(\bar{p}(x))\|_H = \|P_C(\bar{p})\|_{\mathcal{C}} \,\}, \quad \bar{u}'(x) = \frac{1}{\|P_C(\bar{p})\|_{\mathcal{C}}} P_C(\bar{p}(x)) \quad |\bar{u}| - a.e.\ x \in \Omega.$$

These conditions are sufficient for optimality if F is convex on its domain.

Proof. The statement follows immediately by combining Propositions 6.21 and 6.23. □

The following two corollaries highlight how these necessary first order optimality conditions allow to draw further conclusions on structural properties of minimizers to $(\mathfrak{P}^{\mathcal{M}})$.

Corollary 6.25. *Let a minimizer $\bar{u}$ to $(\mathfrak{P}^{\mathcal{M}})$ be given and assume that $\|P_C(\bar{p}(x))\|_H$ achieves its maximum in a finite collection of points:*

$$\{\, x \in \Omega \mid \|P_C(\bar{p}(x))\|_H = \|P_C(\bar{p})\|_{\mathcal{C}} \,\} = \{\bar{x}_i\}_{i=1}^N. \tag{6.45}$$

Then $\bar{u}$ is given as a sum of Dirac delta functions, i.e. there holds

$$\bar{u} = \frac{1}{\|P_C(\bar{p})\|_{\mathcal{C}}} \sum_{i=1}^N \bar{c}_i P_C(\bar{p}(\bar{x}_i))\delta_{\bar{x}_i},$$

for some $\bar{c}_i \in \mathbb{R}_+$, $i = 1, \ldots, N$.

Proof. From the inclusion condition on $\operatorname{supp}|\bar{u}|$ we infer $|\bar{u}| = \sum_{i=1}^N \bar{c}_i \delta_{\bar{x}_i}$ for some $\bar{c}_i \in \mathbb{R}_+$, $i = 1, \ldots, N$. The claim now directly follows from the characterization of the Radon-Nikodým derivative yielding

$$\bar{u} = \sum_{i=1}^N \bar{c}_i \bar{u}_i'(\bar{x}_i)\delta_{\bar{x}_i}, \quad \bar{u}'(\bar{x}_i) = \frac{1}{\|P_C(\bar{p})\|_{\mathcal{C}}} P_C(\bar{p}(\bar{x}_i)).$$

□

Corollary 6.26. *Assume that F is strictly convex on its domain. Then the optimal state $\bar{y}$ and adjoint state $\bar{p}$ are the same for every minimizer to $(\mathfrak{P}^{\mathcal{M}})$. Furthermore assume that* (6.45) *holds and that the set*

$$\{\, K(P_C(\bar{p}(\bar{x}_i))\delta_{\bar{x}_i}) \mid i = 1, \ldots, N \,\} \subset Y, \tag{6.46}$$

is linearly independent. Then $(\mathfrak{P}^{\mathcal{M}})$ admits a unique minimizer $\bar{u} \in \mathcal{M}(\Omega, C)$.

Proof. The prove for the uniqueness of the optimal state is standard: assume that there are two optimal solutions $\bar{u}_1, \bar{u}_2$ to $(\mathfrak{P}^{\mathcal{M}})$ with $K\bar{u}_1 \neq K\bar{u}_2$. Set $u_s = u_1 + s(u_2 + u_1)$ for $s \in (0,1)$. Then u_s is also a minimizer of $(\mathfrak{P}^{\mathcal{M}})$. Since F is strictly convex we conclude

$$\min_{u \in \mathcal{M}(\Omega,H)} j(u) = j(u_s) < (1-s)j(u_1) + sj(u_2) = j(u_s).$$

This gives a contradiction. The uniqueness of the adjoint state follows now due to $\bar{p} = -K^*\nabla F(\bar{y})$. Assume that (6.45) holds and that the set in (6.46) is linear independent. Moreover define the operator

$$\hat{K} \colon \mathbb{R}^N \to Y, \quad v \mapsto \frac{1}{\|P_C(\bar{p})\|_C} \sum_{i=1}^N v_i K(P_C(\bar{p}(\bar{x}_i))\delta_{\bar{x}_i}).$$

Following Corollary 6.25 every minimizer $\bar{u}$ to $(\mathfrak{P}^{\mathcal{M}})$ is of the form

$$\bar{u} = \frac{1}{\|P_C(\bar{p})\|_C} \sum_{i=1}^N \|\bar{\mathbf{u}}_i\|_H \bar{p}(\bar{x}_i)\delta_{\bar{x}_i}, \quad \|\bar{\mathbf{u}}_i\|_{\mathcal{M}} \in \mathbb{R}_+.$$

Obviously the vector $(\|\bar{\mathbf{u}}_1\|_H, \dots, \|\bar{\mathbf{u}}_N\|_H)^\top$ is given by an optimal solution to

$$\min_{v \in \mathbb{R}^N_+} F(\hat{K}v) + G(\|v\|_1). \tag{6.47}$$

Since the set in (6.46) is linearly independent we conclude that the operator $\hat{K}$ is injective. Thus the composite functional $F \circ \hat{K}$ is stricly convex on its domain in $\mathbb{R}^N_+$ and (6.47) admits a unique solution. Combining all previous considerations yields the uniqueness of the minimizer to $(\mathfrak{P}^{\mathcal{M}})$. □

The remainder of this section is devoted to the algorithmic solution of $(\mathfrak{P}^{\mathcal{M}})$ by applying the generalized conditional gradient method described in Algorithm 8. To this end we make the following observation.

Lemma 6.27. *There exists* $u \in \operatorname{dom} j \cap \mathcal{M}_{\mathbb{N}}(\Omega, C)$.

Proof. Let an arbitrary $u \in \operatorname{dom} j$ be given. Following Proposition 6.19 there exists a sequence $\{u_k\}_{k\in\mathbb{N}} \subset \mathcal{M}_{\mathbb{N}}(\Omega, C)$ with $u^k \rightharpoonup^* u$, $\|u^k\|_{\mathcal{M}} \leq \|u\|_{\mathcal{M}}$. Since the domain of F in Y_{ad} is open and G is monotonically increasing on $\mathbb{R}_+$ we conclude $u_k \in \operatorname{dom} j$ for all k large enough. This completes the proof. □

As a consequence Algorithm 8 can be started from a finitely supported iterate $u^0 \in \mathcal{M}_{\mathbb{N}}(\Omega, C)$ in the domain of j. Furthermore let the constant $M_0 > 0$ be chosen to bound the norms of the elements in the sublevel set

$$E_j(u^0) = \left\{ u \in \mathcal{M}(\Omega, C) \mid j(u) \leq j(u^0) \right\}.$$

This choice is possible due to the radial unboundedness of j. Denote by u^k the iterate in the k-th step of Algorithm 8 and by $p^k = -K^*\nabla F(Ku^k)$ the associated adjoint state. The new intermediate iterate is determined as convex combination between u^k and a solution v^k to the partially linearized problem

$$\min_{\substack{v \in \mathcal{M}(\Omega,C) \\ \|v\|_{\mathcal{M}} \leq M_0}} \langle -p^k, v \rangle + G(\|v\|_{\mathcal{M}}) \tag{6.48}$$

The following proposition states that this problem admits at least one solution supported on a single point in Ω.

Proposition 6.28. *Let $u^k \in \operatorname{dom} j$ be given and set $p^k = -K^* \nabla F(Ku^k) \in \mathcal{C}(\Omega, H)$. Choose a point $\hat{x}^k \in \Omega$ with $\|P_C(p^k(\hat{x}^k))\|_H = \|P_C(p^k)\|_{\mathcal{C}}$ and*

$$\|v^k\|_{\mathcal{M}} \leq M_0, \quad \|v^k\|_{\mathcal{M}} \in \begin{cases} \{0\} & \|P_C(p^k)\|_{\mathcal{C}} < \inf \partial G(0) \\ \partial G^*(\|P_C(p^k)\|_{\mathcal{C}}) & \|P_C(p^k)\|_{\mathcal{C}} \in \bigcup_{m \in [0,M_0]} \partial G(m) \\ \{M_0\} & \|P_C(p^k)\|_{\mathcal{C}} > \sup \partial G(M_0). \end{cases} . \tag{6.49}$$

Then the measure

$$v^k = \|v^k\|_{\mathcal{M}} \begin{cases} 0 & P_C(p^k) = 0 \\ \frac{P_C(p^k(\hat{x}^k))}{\|P_C(p^k)\|_{\mathcal{C}}} \delta_{\hat{x}^k} & P_C(p^k) \neq 0 \end{cases} . \tag{6.50}$$

is a minimizer of (6.48).

Proof. We note that with the substitution $v = m\tilde{v}$ for $m \in [0, M_0]$ and $\tilde{v} \in \mathcal{M}(\Omega, H)$, $\|\tilde{v}\|_{\mathcal{M}} \leq 1$, the problem (6.48) can be decomposed into

$$\min_{m \in [0,M_0]} \min_{\substack{\tilde{v} \in \mathcal{M}(\Omega,C) \\ \|\tilde{v}\|_{\mathcal{M}} \leq 1}} [-m\langle p^k, \tilde{v} \rangle + G(m)].$$

Due to the non-negativity of m we estimate

$$m\langle -p^k, \tilde{v} \rangle = -m \int_\Omega (p^k(x), \tilde{v}'(x))_H \, \mathrm{d}|\tilde{v}|(x) \geq -m \int_\Omega (P_C(p^k(x)), \tilde{v}'(x))_H \, \mathrm{d}|\tilde{v}|(x) \geq -m\|P_C(p^k)\|_{\mathcal{C}}.$$

for every $\tilde{v} \in \mathcal{M}(\Omega, C)$, $\|\tilde{v}\|_{\mathcal{M}} \leq 1$. Accordingly a solution to the inner problem is given by

$$\hat{v} = \begin{cases} 0 & P_C(p^k) = 0 \\ \frac{P_C(p^k(\hat{x}))}{\|P_C(p^k)\|_{\mathcal{C}}} \delta_{\hat{x}} & P_C(p^k) \neq 0 \end{cases} , \quad \hat{x} \in \operatorname*{arg\,max}_{x \in \Omega} \|P_C(p^k(x))\|_H.$$

To solve the outer problem it thus suffices to consider

$$\min_{m \in [0,M_0]} [-m\|P_C(p^k)\|_{\mathcal{C}} + G(m)].$$

By standard arguments, $\bar{m} \in [0, M_0]$ is optimal if and only if

$$\|P_C(p^k)\|_{\mathcal{C}} \in \partial(G(\cdot) + I_{[0,M_0]}(\cdot))(\bar{m}).$$

Since $I_{[0,M_0]}$ is continuous on the interior of its domain we can split the subdifferential to obtain

$$\|P_C(p^k)\|_{\mathcal{C}} \in \partial G(\bar{m}) + \partial I_{[0,M_0]}(\bar{m}).$$

Distinguishing between the three different cases in (6.49) completes the proof. □

Example 6.8. *For a better illustration of the previous proposition we derive the condition in* (6.49) *for the case of norm regularization* $G_1(\|u\|_{\mathcal{M}}) = \beta\|u\|_{\mathcal{M}} + I_{[0,\infty)}(\|u\|_{\mathcal{M}})$ *and norm constraints* $G_2(\|u\|_{\mathcal{M}}) = I_{[0,M_1]}(\|u\|_{\mathcal{M}})$*. In the second case we can clearly assume that* $M_1 = M_0$*. For* G_1 *we obtain*

$$\|v^k\|_{\mathcal{M}} \in \begin{cases} \{0\} & \|P_C(p^k)\|_{\mathcal{C}} \in [0,\beta) \\ [0, M_0] & \|P_C(p^k)\|_{\mathcal{C}} = \beta \\ \{M_0\} & \|P_C(p^k)\|_{\mathcal{C}} > \beta \end{cases}.$$

In the norm constrained case we analogously conclude

$$\|v^k\|_{\mathcal{M}} \in \begin{cases} [0, M_0] & \|P_C(p^k)\|_{\mathcal{C}} = 0 \\ \{M_0\} & \|P_C(p^k)\|_{\mathcal{C}} > 0 \end{cases}.$$

We summarize the resulting generalized conditional gradient method in Algorithm 9. As a consequence of the previous proposition we may compute a minimizer to $(\mathfrak{P}^{\mathcal{M}})$ based on the sequential insertion of a single Dirac delta function into the current iterated vector measure u^k. Thus, since $u^0 \in \mathcal{M}_{\mathbb{N}}(\Omega, C)$, there holds $u^k \in \mathcal{M}_{\mathbb{N}}(\Omega, C)$ for all $k \in \mathbb{N}$. It is however important to note that the GCG step only allows for a removal of points in the unlikely case of $s^k = 1$, i.e. u^k is replaced by the solution v^k to the linearized problem. In particular if $(\mathfrak{P}^{\mathcal{M}})$ admits a unique sparse minimizer $\bar{u}$ each of its Dirac delta functions may be approximated by an ever growing number of point measures in the iterate u^k. This leads to undesired clustering of Dirac delta functions around the optimal positions. To mitigate these effects we include a black box point removal step into the method, see step 4. In order to discuss these additional optimization steps we consider an ordered set of distinct points $\mathcal{A}$ and the associated parametrization $U_{\mathcal{A}}$ defined by

Algorithm 9 Generalized conditional gradient method for vector measures

```
while $\phi(u^k) \geq \mathrm{TOL}$ do
    1. Compute $p^k = -K^*\nabla F(Ku^k)$. Determine $\hat{x}^k \in \arg\max_{x\in\Omega} \|P_C(p^k(x))\|_{\mathcal{C}}$ and $\|v^k\|_{\mathcal{M}}$ according to (6.49).
    2. Set $v^k = \|v^k\|_{\mathcal{M}} \begin{cases} 0 & P_C(p^k) = 0 \\ \frac{P_C(p^k(\hat{x}^k))}{\|P_C(p^k)\|_{\mathcal{C}}}\delta_{\hat{x}^k} & P_C(p^k) \neq 0 \end{cases}$
    3. Select stepsize $s^k \in [0,1]$ and set $u^{k+1/2} = u^k + s^k(v^k - u^k)$.
    4. Set $\mathcal{A}_k = \operatorname{supp}|u^k| \cup \{\hat{x}^k\}$ and find $\mathbf{u}^{k+1} \in C^{\#\mathcal{A}_k}$ such that $u^{k+1} = U_{\mathcal{A}_k}(\mathbf{u}^{k+1})$ with $j(u^{k+1}) \leq j(u^{k+1/2})$.
end while
```

$$\mathcal{A} = \{\, x_i \in \Omega \mid i = 1, \ldots, N \,\}, \quad U_{\mathcal{A}} \colon H^N \to \mathcal{M}(\Omega, C), \quad \mathbf{u} \mapsto \sum_{i=1}^{N} \mathbf{u}_i \delta_{x_i}. \tag{6.51}$$

The point removal procedure in step 4. of Algorithm 9 is now based on the approximate solution of an auxiliary problem on the Hilbert space $H^{\#\mathcal{A}}$

$$\min_{\mathbf{u}\in C^{\#\mathcal{A}}} j(U_{\mathcal{A}}(\mathbf{u})) = F(KU_{\mathcal{A}}(\mathbf{u})) + G(\|U_{\mathcal{A}}(\mathbf{u})\|_{\mathcal{M}}), \quad \text{with} \quad \|U_{\mathcal{A}}(\mathbf{u})\|_{\mathcal{M}} = \sum_{i=1}^{\#\mathcal{A}} \|\mathbf{u}_i\|_H, \tag{$\mathfrak{P}^{\mathcal{M}}(\mathcal{A})$}$$

where the set $\mathcal{A}$ is chosen as $\mathcal{A} = \operatorname{supp}|u^k| \cup \{\hat{x}^k\}$. Thus, loosely speaking, we fix the positions of the Dirac delta functions in the current iterate u^k and approximately optimize their coefficient functions while ensuring $j(u^{k+1}) \leq j(u^{k+1/2})$. In particular this choice implies $\|u^k\|_{\mathcal{M}} \leq M_0$ for all $k \in \mathbb{N}$ and all Dirac delta functions whose coefficient functions are zero get removed from the iterate due to the choice of the set $\mathcal{A}_k$. Similar approaches were already considered for the sparse minimization problems in [44] and [50]. The present work delimits itself from these previous instances by deriving improved convergence statements for the GCG method when augmented with two particular realizations of this additional step, see Section 6.3.3 and 6.3.4. As in the general case we emphasize that the improvement step is not necessary to ensure convergence of the algorithm, i.e. we can choose $u^{k+1} = u^{k+1/2}$. However, as we will see in the following considerations, a sophisticated choice of the point removal step greatly benefits the sparsity of the iterates as well as the overall convergence of the method.

Remark 6.8. For completeness we also mention the possibility to include improvement steps based on a parametrization of the vector measure by its support points. Therefore given $N \in \mathbb{N}$, and a coefficient vector $\mathbf{u} \in C^N$ let us define

$$U(x, \mathbf{u}) \colon \Omega^N \to \mathcal{M}(\Omega, C), \quad x \mapsto \sum_{i=1}^{N} \mathbf{u}_i \delta_{x_i}.$$

In contrast to ($\mathfrak{P}^{\mathcal{M}}(\mathcal{A})$) we now fix the coefficients of the measure and approximately minimize for the optimal positions

$$\min_{x \in \Omega^N} j(U(x, \mathbf{u})) = F(KU(x, \mathbf{u})) + G(\|U(x, \mathbf{u})\|_{\mathcal{M}}). \tag{6.52}$$

For example the authors in [50] propose to move the Dirac delta functions according to the gradient flow of the smooth part $F(KU(x, \mathbf{u}))$ with respect to the positions. This bears similarity to the particle gradient flow method discussed in [69]. The authors in [44] advocate a solution of (6.52) by first order methods. Note that these suggestions presuppose that the adjoint operator K^* maps continuously to $\mathcal{C}^1(\Omega, H)$ and, given $y \in Y$, the gradient $\nabla[K^*y]$ is readily available. In the practical parts of this thesis however we apply the presented optimization algorithm to problems that stem from a finite element discretization K_h of the operator K with $K_h y \notin \mathcal{C}^1(\Omega, H)$. Moreover, even if F is convex, the position problem (6.52) is in general nonconvex. Thus it may admit a large number of stationary points and the computation of a global minimizer may be infeasible. In contrast, if F is convex so is the coefficient problem ($\mathfrak{P}^{\mathcal{M}}(\mathcal{A})$). For these reasons improvement steps based on point moving are out of the scope of this thesis and will not be discussed in more detail.

As in the general case the termination criterion for Algorithm 9 is based on the primal-dual-gap of the iterates $\Phi(u^k)$. From the definition of v^k and (6.23) the primal-dual-gap is readily calculated as

$$\Phi(u^k) = \langle -p^k, u^k \rangle + G(\|u^k\|_{\mathcal{M}}) + \|P_C(p^k)\|_{\mathcal{C}} \|v^k\|_{\mathcal{M}} - G(\|v^k\|_{\mathcal{M}}).$$

The following worst-case convergence results are a direct consequence of Theorem 6.11 and 6.14,

Theorem 6.29. *Let F, K and G fulfill Assumption 6.3. Let the sequence $\{u^k\}_{k\in\mathbb{N}}$ be generated by Algorithm 9 where the stepsize is chosen according to the Quasi-Armijo-Goldstein condition with parameters $\gamma \in (0,1)$, $\alpha \in (0, 1/2]$. Then the following convergence results hold true:*

- *There exists at least one weak* convergent subsequence of $\{u^k\}_{k\in\mathbb{N}}$. Every weak* accumulation point $\bar{u}$ of $\{u^k\}_{k\in\mathbb{N}}$ is a stationary point, i.e. $\Phi(\bar{u}) = 0$.*

- *If F is convex on its domain and ∇F is Lipschitz continuous on*

$$KE_j(u^0) = \left\{ Ku \mid j(u) \leq j(u^0) \right\}, \tag{6.53}$$

with Lipschitz constant L_{Ku^0} then $F \circ K$ is Lipschitz continuous on $E_j(u^0)$ and $\{u^k\}_{k\in\mathbb{N}}$ is a minimizing sequence for j. Each of its weak accumulation points is a global minimizer of j and there holds*

$$r_j(u^k) \leq \frac{r_j(u^0)}{1+qk}, \quad q = \alpha \min\left\{ \frac{c}{4L_{u^0}M_0^2}, 1 \right\},$$

with $c = 2\gamma(1-\alpha)r(u^0)$ and $L_{u^0} = L_{Ku^0}\|K^*\|^2_{\mathcal{L}(Y,\mathcal{C}(\Omega,H))}$.

Proof. The first result is readily obtained from Theorem 6.11. Second assume that F is convex and its gradient ∇F is Lipschitz on $KE_j(u^0)$ with Lipschitz constant L_{Ku^0}. Define the reduced functional $f = F \circ K$. Obviously f is convex on its domain and we have

$$\begin{aligned}\sup_{u_1,u_2\in E_j(u^0)} \|\nabla f(u_1) - \nabla f(u_2)\|_{\mathcal{C}} &\leq \sup_{u_1,u_2\in E_j(u^0)} \|K^*\|_{\mathcal{L}(Y,\mathcal{C}(\Omega,H))}\|\nabla F(Ku_1) - \nabla F(Ku_2)\|_Y \\ &\leq \sup_{u_1,u_2\in E_j(u^0)} L_{Ku^0}\|K^*\|^2_{\mathcal{L}(Y,\mathcal{C}(\Omega,H))}\|u_1 - u_2\|_{\mathcal{M}}.\end{aligned}$$

Thus ∇f is Lipschitz continuous on $E_j(u^0)$ with constant $L_{u^0} = L_{Ku^0}\|K^*\|^2_{\mathcal{L}(Y,\mathcal{C}(\Omega,H))}$. The remaining statements now follow by applying Theorem 6.14. □

Remark 6.9. Let us briefly summarize some previous convergence results for generalized conditional gradient methods in spaces of vector measures:

[50]: Here the authors provide a sublinear rate of convergence for the special case of

$$C = \mathbb{R}^n, \quad G(\|u\|_{\mathcal{M}}) = \beta\|u\|_{\mathcal{M}} + I_{[0,\infty)}(\|u\|_{\mathcal{M}}), \quad F = 1/2\|\cdot - y_d\|^2_Y.$$

The step size $s^k \in [0,1]$ is chosen to maximize a lower bound on the expected descend in the k-th iteration

$$s^k \in \operatorname*{arg\,min}_{s\in[0,1]}[-s\Phi(u^k) + \frac{s^2}{2}\|K(u^k - v^k)\|^2_Y].$$

[44]: This work considers a general smooth and convex function F and

$$Y = \mathbb{R}^n, \quad C = \mathbb{R}_+, \quad G(\|u\|_{\mathcal{M}}) = I_{[0,M_0]}(\|u\|_{\mathcal{M}}).$$

A fixed step size $s^k = 2/(k+2)$ is used in the proof of the sublinear convergence rate. We point out that the authors do not assume Lipschitz continuity of the gradient ∇F but suppose that the curvature constant of F, see e.g. [159], on $\{ Ku \ |\|u\|_{\mathcal{M}} \leq M_0 \}$ is bounded.

While both of these works focus on different problems it is worthwhile to discuss the differences in the proofs of these results. Similar to our approach the authors in [50] describe and analyze the conditional gradient method directly on the non-reflexive space $\mathcal{M}(\Omega, \mathbb{R}^n)$. In contrast the second paper relies on an equivalent reformulation of the problem as minimization problem for a smooth function over a finite dimensional compact set. As a matter of fact we might proceed along the same path for the discussion of Algorithm 9. Let us outline these ideas for the nonsmooth norm regularized problem

$$\min_{\substack{u\in\mathcal{M}(\Omega,C)\\ \|u\|_{\mathcal{M}}\leq M_0}} j(u)=[F(Ku)+\beta\|u\|_{\mathcal{M}}], \tag{6.54}$$

where we assume $\operatorname{dom} F=Y$ and Lipschitz continuity of ∇F on the whole domain for simplicity. To this end define the compact and convex admissible set

$$W_{ad}:=\{\,(y,m)\mid m\in[0,M_0],\quad \exists u\in\mathcal{M}(\Omega,C),\ \|u\|_{\mathcal{M}}\leq m\colon y=Ku\,\}\subset Y\times\mathbb{R}_+.$$

It is straightforward to see that $\bar{u}\in\mathcal{M}(\Omega,C)$ minimizes in (6.54) iff $\bar{m}=\|\bar{u}\|_{\mathcal{M}}$ and $\bar{y}=K\bar{u}$ give a minimizing pair for

$$\min_{(y,m)\in W_{ad}} h(y,m):=[F(y)+\beta m]. \tag{6.55}$$

Since the function h is smooth and W_{ad} is convex and compact a classical conditional gradient method can be applied to compute a minimizing pair $(\bar{y},\bar{m})$. We claim that such a method is (almost) equivalent to the application of Algorithm 9 to (6.54) with $u^{k+1}=u^{k+1/2}$. More precisely, the algorithms may be realized to ensure $Ku^k=y^k$ and $\|u^k\|_{\mathcal{M}}\leq m^k$ for $k\in\mathbb{N}$. Set the initial iterate to $(y^0,m^0)=(Ku^0,\|u^0\|_{\mathcal{M}})$. The proof is done by induction. Given an iterate (y^k,m^k) with $Ku^k=y^k$ and $\|u^k\|_{\mathcal{M}}\leq m^k$ the new descent direction $(\delta y^k,\delta m^k)$ in the conditional gradient method for (6.55) is found by solving the linearized problem

$$\min_{(\delta y,\delta m)\in W_{ad}}[(\partial_y h(y^k,m^k),\delta y)+\partial_m h(y^k,m^k)\delta m]=\min_{(\delta y,\delta m)\in W_{ad}}[(\nabla F(y^k),\delta y)_Y+\beta\delta m].$$

Obviously one minimizer to this problem is given by $(\delta y^k,\delta m^k)=(Kv^k,\|v^k\|_{\mathcal{M}})$ where v^k is chosen according to Algorithm 9. Choosing the same stepsize s^k in both algorithms we get

$$y^{k+1}=K(u^k+s^k(v^k-u^k))=Ku^{k+1},\quad \|u^{k+1}\|_{\mathcal{M}}\leq m^k+s^k\|v^k\|_{\mathcal{M}}=m^{k+1}.$$

In particular this implies $j(u^k)\leq h(y^k,m^k)$ for all $k\in\mathbb{N}$. Since $\nabla h=(\partial_y h,\partial_m h)$ is Lipschitz continuous the classical convergence results for the conditional gradient, see e.g. [92], can be applied to conclude the sublinear convergence of $h(y^k,m^k)$ towards its minimum value on W_{ad}

$$\min_{(y,m)\in W_{ad}} h(y,m)=\min_{u\in\mathcal{M}(\Omega,C)} j(u)=j(\bar{u}).$$

As a consequence the sublinear convergence of $j(u^k)$ towards $j(\bar{u})$ also follows.

There are several reasons why we decided to stick to a discussion of generalized conditional gradient methods on the measure space i.e. without reformulating the problem. First such a reformulation clearly requires the linearity of the operator K. Thus conditional gradient methods for sparse optimal control problems with nonlinear state equation, see e.g. [63, 64], cannot be discussed in this way. In contrast we based our convergence analysis on the general results of Section 6.2 which obviously allow to consider far more general problems. In particular the discussions on the structure of solutions to the partially linearized problems and the subsequential weak* convergence of the sequence $\{u^k\}_{k\in\mathbb{N}}$ extend naturally to the case of smooth but nonlinear control-to-state mappings K.

Second we aim to improve on the convergence results of Theorem 6.29 in the following two sections. More precisely we prove linear convergence of the residual $r_j(u^k)$ under additional assumptions on the minimizers of ($\mathfrak{P}^{\mathcal{M}}$) for a particular choice of the point removal step in Algorithm 9. As already mentioned in the introductory part of Section 6.2 there are several works on improved convergence rates for the classical conditional gradient method with and without additional acceleration steps. However these results usually require uniform convexity of the objective functional or uniform lower bounds on its gradient. Moreover additional geometric properties of the admissible set such as polyhedricity or strong convexity are needed. In this context note that $h(\cdot,\cdot)$ is not uniformly convex due to the linear dependence on m and the structure of the set W_{ad} can be fairly complicated. As a consequence, to the best of our knowledge, none of these pre-existing works allow to obtain the improved convergence results for the solution of (6.54) based on the reformulated problem (6.55). Moreover we also provide convergence rates for the measure valued iterates $\{u^k\}_{k\in\mathbb{N}}$ which requires to exploit certain structural properties. These considerations make a direct analysis of Algorithm 9 on $\mathcal{M}(\Omega, H)$ indispensable.

6.3.3 Sparsification for finite rank operators

This section is devoted to generalized conditional gradient methods in the important special case of K being a finite rank operator i.e. $\dim \operatorname{Im} K < \infty$. For better illustration we may pick up on Example 6.6. In this case $K\colon \mathcal{M}(\Omega, L^2(I)) \to \mathbb{R}^N$ gives, for example, averaged values of the temperature field induced by the heat source u on a finite number N of observational patches. The main result of this section comes in two parts. First we give a constructive proof for the existence of a finitely supported optimal solution to ($\mathfrak{P}^{\mathcal{M}}$) provided that K has finite rank. In a second step we augment Algorithm 9 by an additional sparsification step which ensures subsequential convergence towards sparse stationary points of j. To this end let an arbitrary measure $u_1 \in \mathcal{M}(\Omega, C)$ be given. Associated to it we consider the minimum norm problem

$$\min_{u\in\mathcal{M}(\Omega,C)} \|u\|_{\mathcal{M}} \quad s.t. \quad Ku = Ku_1 \tag{$\mathfrak{P}(u_1)$}$$

Since the operator K is weak*-to-strong continuous the solution set

$$U_1 = \{\, u \in \mathcal{M}(\Omega, C) \mid u \text{ solves } (\mathfrak{P}(u_1)) \,\},$$

is nonempty, convex and weak* closed. We recall the notion of an extremal point of the solution set as well as the Krein-Milman theorem c.f. [43, Theorem 2.19].

Definition 6.4. An element $u \in U_1$ is called an extremal point of U_1 if for all $v_1,\ v_2 \in U_1$ and $s \in [0,1]$ there holds

$$u = (1-s)v_1 + sv_2 \Rightarrow v_1 = v_2 = u.$$

Theorem 6.30 (Krein-Milman)**.** *The set U_1 is the weak* closure of the convex hull of its extremal points:*

$$U_1 = \overline{\operatorname{conv}\{\, u \in U_1 \mid u \ \textit{extremal}\}}^{*}.$$

Proof. Since U_1 is convex, nonempty and weak* closed the set of its extremal points is nonempty. Taking the weak* closure of their convex hull we obtain U_1 following the Krein-Milman Theorem, [43, Theorem 2.19]. □

In the following theorem we show that every extremal point of U_1 is supported on at most N points.

Theorem 6.31. *Suppose that* $\dim \operatorname{Im} K = N < \infty$. *The extremal points of* U_1 *can be written as a linear combinations of no more than N Dirac delta functions:*

$$\{\, u \in U_1 \ \mid \ u \text{ extremal } \} \subset \left\{ \sum_{i=1}^{N} \mathbf{u}_i \delta_{x_i} \ \middle|\ \mathbf{u}_i \in C,\ x_i \in \Omega, \quad i = 1, \ldots, N \right\}$$

Proof. Let $u \in U_1$ be extremal. The proof will be done by contradiction. Assume, therefore, that $\operatorname{supp} |u|$ consists of more than N points. Then, there exists a disjoint partition $\{\, \Omega_i \,\}_{i=1,\ldots,N}$ of the set Ω with

$$|u|(\Omega_i) > 0 \quad \text{for all } i = 1, \ldots, N+1.$$

Define for $i = 1, \ldots, N+1$ the restrictions

$$u_i = u|_{\Omega_i} \in \mathcal{M}(\Omega, C).$$

It is clear that $\|u_i\|_{\mathcal{M}} = |u|(\Omega_i) > 0$ and $\|u\|_{\mathcal{M}} = \sum_{i=1}^{N+1} \|u_i\|_{\mathcal{M}}$. Now, we consider the renormalized measures and their image under K, i.e.

$$v_i = \frac{u_i}{\|u_i\|_{\mathcal{M}}}, \quad w_i = K v_i \in \operatorname{Im} K \subset Y,$$

and look for a nontrivial solution $\lambda \in \mathbb{R}^{N+1} \setminus \{0\}$ of the system of linear equations

$$\sum_{i=1}^{N+1} \lambda_i K v_i = \sum_{i=1}^{N+1} \lambda_i w_i = 0 \in \operatorname{Im} K.$$

Since the number of equations is one smaller than the number of variables, such a solution exists. Without restriction, we may assume $\sum_{i=1,\ldots,N+1} \lambda_i \geq 0$ (otherwise, we take the negative of λ). We define

$$\tau = \max_{i=1,\ldots,N+1} \frac{|\lambda_i|}{\|u_i\|_{\mathcal{M}}} > 0$$

and u_+ and u_- as

$$u_\pm = u \pm \frac{1}{\tau} \sum_{i=1}^{N+1} \lambda_i v_i = \sum_{i=1}^{N+1} \left(1 \pm \frac{\lambda_i}{\tau \|u_i\|_{\mathcal{M}}} \right) u_i.$$

Clearly, $u_+ \neq u_- \neq u$. By construction and linearity of K we have $K u_\pm = K u = K u_1$. Furthermore, we directly verify that

$$\|u_\pm\|_{\mathcal{M}} = \int_\Omega \mathrm{d}|u_\pm| = \sum_{i=1}^{N+1} \int_{\Omega_i} \mathrm{d}|u_\pm| = \sum_{i=1}^{N+1} \left(\|u_i\|_{\mathcal{M}} \pm \frac{\lambda_i}{\tau} \right) = \|u\|_{\mathcal{M}} \pm \frac{1}{\tau} \sum_{i=1}^{N+1} \lambda_i$$

as well as $u_\pm \in \mathcal{M}(\Omega, C)$ since $|\lambda_i|/\tau \leq \|u_i\|_{\mathcal{M}}$. Since $\sum_{i=1,\ldots,N+1} \lambda_i \geq 0$ we have $\|u_-\|_{\mathcal{M}} \leq \|u\|_{\mathcal{M}}$, and u_- is an optimal solution of ($\mathfrak{P}(u_1)$), i.e., $u_- \in U_1$. Moreover, we see that it must hold $\sum_{i=1}^{N+1} \lambda_i = 0$, since the norm cannot be strictly smaller. It follows that also $u_+ \in U_1$. We conclude the proof with the observation that

$$u = \frac{1}{2} u_+ + \frac{1}{2} u_-,$$

which contradicts the assumption that u is extremal in U_1. □

As an immediate consequence of the previous theorem we conclude the existence of finitely supported minimizers to $(\mathfrak{P}^{\mathcal{M}})$.

Proposition 6.32. *Let $u \in \mathcal{M}(\Omega, C)$ be given. Then there exists a measure $\tilde{u} \in \mathcal{M}(\Omega, C)$ with*

$$Ku = K\tilde{u}, \quad \|\tilde{u}\|_{\mathcal{M}} \leq \|u\|_{\mathcal{M}}, \quad \tilde{u} \in \mathcal{M}_N(\Omega, C)$$

In particular there exists a minimizer $\bar{u} \in \mathcal{M}(\Omega, C)$ to $(\mathfrak{P}^{\mathcal{M}})$ with $\# \operatorname{supp} |\bar{u}| \leq N$.

Proof. Following the previous theorem the minimum norm problem $(\mathfrak{P}(u))$ associated to a measure $u \in \mathcal{M}(\Omega, C)$ admits at least one optimal solution $\tilde{u} \in \mathcal{M}(\Omega, C)$ with $\operatorname{supp} |\tilde{u}| \leq N$. By construction we further have

$$Ku = K\tilde{u}, \quad \|\tilde{u}\|_{\mathcal{M}} \leq \|u\|_{\mathcal{M}}.$$

Since $u \in \mathcal{M}(\Omega, C)$ was chosen arbitrary the same reasoning particularly applies to any minimizer of $(\mathfrak{P}^{\mathcal{M}})$. Due to the monotonicity of G on $\mathbb{R}_+$ the statement follows. □

Obviously the previous proposition does not only yield the existence of a sparse minimizer. More precisely, given any measure $u \in \mathcal{M}(\Omega, C)$ we get at least one sparse measure $\tilde{u}$, $\# \operatorname{supp} |\tilde{u}| \leq N$, yielding the same image under K without increasing the objective function value. From an algorithmic point of view it is desirable to exploit this sparse representation property for the iterates $\{u^k\}_{k \in \mathbb{N}}$ generated by the generalized conditional gradient method. This would bound the number of support points in the iterates and thus mitigates clustering effects. By slightly altering the proof of Theorem 6.31 we arrive at a constructive sparsifying procedure to remove excess points from a given sparse measure. The method is summarized in Algorithm 10.

Algorithm 10 Support-point removal for vector measures

```
  1. Let $u = \sum_{i=1}^{\mathbf{N}} \mathbf{u}_i \delta_{x_i} \in \mathcal{M}(\Omega, C)$ $\mathbf{u}_i \neq 0$, be given.
  while $\{K(\mathbf{u}_i \delta_{x_i})\}_{i=1}^{\mathbf{N}}$ linearly dependent do
    2. Set $\mathbf{v}_i = \mathbf{u}_i / \|\mathbf{u}_i\|_H$
    3. Find $0 \neq \lambda$ with $0 = \sum_{i=1}^{\mathbf{N}} \lambda_i K(\mathbf{v}_i \delta_{x_i})$.
    4. Set $\mu = \max_i \{ \lambda_i / \|\mathbf{u}_i\|_H \}$, $\mathbf{u}_{\text{new},i} = (1 - \lambda_i / (\mu \|\mathbf{u}_i\|_H)) \mathbf{u}_i$.
    5. Update $u = u_{\text{new}} = \sum_{\{ i \mid \mathbf{u}_{\text{new},i} > 0 \}} \mathbf{u}_{\text{new},i} \delta_{x_i}$.
  end while
```

Proposition 6.33. *Suppose that* $\dim \operatorname{Im} K = N < \infty$. *Let $u = \sum_{i=1,\ldots,\mathbf{N}} \mathbf{u}_i \delta_{x_i}$ be an arbitrary sparse measure with $\mathbf{N} \in \mathbb{N}$, $\mathbf{u}_i \in C$, $\mathbf{u}_i \neq 0$, $x_i \in \Omega$ (pairwise distinct). Furthermore assume that the set $\{K(\mathbf{u}_i \delta_{x_i})\}_{i=1}^{\mathbf{N}}$ is linearly dependent and $u^{new} \in \mathcal{M}(\Omega, H)$ is obtained after one iteration of Algorithm 10 applied to u. Then there holds $u^{new} \in \mathcal{M}(\Omega, C)$. Moreover, the new measure u^{new} satisfies*

$$Ku^{new} = Ku, \quad \|u^{new}\|_{\mathcal{M}} \leq \|u\|_{\mathcal{M}}, \quad \operatorname{supp} |u^{new}| \subset \operatorname{supp} |u|, \quad \# \operatorname{supp} |u^{new}| \leq \mathbf{N} - 1.$$

Proof. As in the previous proof, we define

$$u_i = u|_{\{x_i\}} = \mathbf{u}_i \delta_{x_i}, \text{ and } w_i = K(\mathbf{v}_i \delta_{x_i}), \text{ where } \mathbf{v}_i = \frac{\mathbf{u}_i}{\|\mathbf{u_i}\|_H}.$$

By assumption the set $\{w_i\}_{i=1}^{\mathbf{N}}$ is linearly dependent. We find a nontrivial solution of the system of equations $\sum_{i=1,\dots,\mathbf{N}} \lambda_i w_i = 0$ with $\sum_{i=1,\dots,\mathbf{N}} \lambda_i \geq 0$. Now, in contrast to the previous proof, we set

$$\mu = \max_{n=1,\dots,\mathbf{N}} \frac{\lambda_i}{\|\mathbf{u}_i\|_H} > 0.$$

We set

$$u_{new} = u - \frac{1}{\mu} \sum_{i=1}^{\mathbf{N}} \lambda_i \mathbf{v}_i \delta_{x_i} = \sum_{i=1}^{\mathbf{N}} \left(1 - \frac{\lambda_i}{\mu \|\mathbf{u}_i\|_H}\right) \mathbf{u}_i \delta_{x_i}$$

Thus the coefficients of the new measure u^{new} are given as $\mathbf{u}_i^{new} = [1 - \lambda_i/(\mu\|\mathbf{u}_i\|_H)]\mathbf{u}_i \in C$ since $\lambda_i/\mu \leq \|\mathbf{u}_i\|_H$ and $Ku = Ku^{new}$. Moreover it holds that $\|u^{new}\|_{\mathcal{M}} = \|u\|_{\mathcal{M}} - \sum_{i=1,\dots,\mathbf{N}} \lambda_i/\mu \leq \|u\|_{\mathcal{M}}$. The proof is finished with the observation that

$$\mathbf{u}_{\hat{i}}^{new} = 0 \quad \text{for } \hat{\imath} \in \operatorname*{arg\,max}_{i=1,\dots,\mathbf{N}} \frac{\lambda_i}{\|\mathbf{u}_i\|_H}.$$

□

The remainder of this section is devoted to the analysis of an augmented generalized conditional gradient method in which we choose the new iterate u^{k+1} by applying Algorithm 10 to the intermediate iterate $u^{k+1/2}$. To this end we first prove the weak* closedness of sets comprising vector measures supported on a uniformly bounded number of support points.

Proposition 6.34. *Let Ω be compact. For any $N \in \mathbb{N}$ the set*

$$\mathcal{M}_N(\Omega, C) = \left\{ \sum_{i=1}^{N} \mathbf{u}_i \delta_{x_i} \;\middle|\; \mathbf{u}_i \in C,\ x_i \in \Omega, \quad i = 1, \dots, N \right\}$$

is weak closed.*

Proof. Let an arbitrary weak* convergent sequence $\{u_k\}_{k\in\mathbb{N}} \subset \mathcal{M}_N(\Omega, C)$ with limit $\bar{u} \in \mathcal{M}(\Omega, C)$ be given. For each $k \in \mathbb{N}$ there exist $\mathbf{u}_i^k \in C$, $x_i^k \in \Omega$, $i = 1, \dots, N$ with

$$u_k = \sum_{i=1}^{N} \mathbf{u}_i^k \delta_{x_i^k} \quad \text{and} \quad \|u_k\|_{\mathcal{M}} = \sum_{i=1,\dots,N} \|\mathbf{u}_i^k\|_H \leq c,$$

for some $c > 0$. Introducing $\mathbf{u}^k = (\mathbf{u}_1^k, \dots, \mathbf{u}_N^k)^\top \in C^N$ and $x^k = (x_1^k, \dots, x_N^k)^\top \in \Omega^N$ there exist a subsequence of $(\mathbf{u}^k, x^k) \in C^N \times \Omega^N$ denoted by the same symbol and $(\mathbf{u}, x) \in C^N \times \Omega^N$ with $\mathbf{u}^k \rightharpoonup \mathbf{u}$ and $x^k \to x$. This follows from the compactness of Ω, the boundedness of $\mathbf{u}^k$ and the weak closedness of C. Defining

$$u = \sum_{i=1,\dots,N} \mathbf{u}_i \delta_{x_i} \in \mathcal{M}_N(\Omega, C),$$

we arrive at

$$\langle \varphi, u \rangle = \lim_{k\to\infty} \sum_{i=1,\dots,N} (\mathbf{u}_i^k, \varphi(x_i^k))_H = \lim_{k\to\infty} \langle \varphi, u_k \rangle = \langle \varphi, \bar{u} \rangle$$

for all $\varphi \in \mathcal{C}(\Omega, H)$ since $\mathbf{u}_i^k \rightharpoonup \mathbf{u}_j$ and $\|\varphi(x_i^k) - \varphi(x_i)\|_H \to 0$. Due to the uniqueness of the weak* limit we get $\bar{u} = u \in \mathcal{M}_N(\Omega, C)$ yielding the weak* closedness of $\mathcal{M}_N(\Omega, C)$. □

As a corollary each accumulation point of a sequence of measures with uniformly bounded support size is also finitely supported.

Corollary 6.35. *Let Ω be compact. Consider a sequence $\{u_k\}_{k\in\mathbb{N}} \subset \mathcal{M}(\Omega, C)$ which fulfils $\#\operatorname{supp}|u_k| \leq N$ for some $N \in \mathbb{N}$ and all $k \in \mathbb{N}$. Then every accumulation point $\bar{u}$ of $\{u_k\}_{k\in\mathbb{N}}$ also satisfies $\#\operatorname{supp}|\bar{u}| \leq N$.*

Proof. By assumption there holds $\#\operatorname{supp}|u^k| \leq N$, $k \in \mathbb{N}$, and thus $\{u^k\}_{k\in\mathbb{N}} \subset \mathcal{M}_N(\Omega, C)$. Since $\mathcal{M}_N(\Omega, C)$ is weak* closed, see Proposition 6.34, the statement follows. □

Finally, combining the GCG method with the sparsifying procedure from Algorithm 10 we obtain a convergent solution algorithm for ($\mathfrak{P}^{\mathcal{M}}$) which additionally ensures the uniform boundedness of the support size in each iteration. As a consequence the resulting algorithm guarantees (subsequential) weak* convergence towards sparse stationary points of j.

Theorem 6.36. *Assume that $\dim \operatorname{Im} K = N < \infty$ and $\#\operatorname{supp}|u^0| \leq N$. Let F, K and G fulfill Assumption 6.3. Let the sequence $\{u^k\}_{k\in\mathbb{N}}$ be generated by Algorithm 9 where u^{k+1} is obtained by applying Algorithm 10 to $u^{k+1/2}$ in each iteration. Then the results of Theorem 6.29 apply to $\{u^k\}_{k\in\mathbb{N}}$. Additionally there holds $u^k \in \mathcal{M}_N(\Omega, C)$, $k \in N$, and consequently $\#\operatorname{supp}|\bar{u}| \leq N$ for every weak* accumulation point $\bar{u}$ of $\{u^k\}_{k\in\mathbb{N}}$.*

Proof. Let $k \in \mathbb{N}$ be given. Denote by $u^{k+1/2}$ the intermediate iterate obtained in step 3. of Algorithm 9 and assume that u^{k+1} is obtained by application of Algorithm 10 to $u^{k+1/2}$. By construction we have

$$u^{k+1} \in \mathcal{M}(\Omega, C), \quad Ku^{k+1} = Ku^{k+1/2}, \quad \|u^{k+1}\|_{\mathcal{M}} \leq \|u^{k+1}\|_{\mathcal{M}},$$

and consequently $j(u^{k+1}) \leq j(u^{k+1/2})$ due to the monotonicity of G on $\mathbb{R}_+$. Thus Theorem 6.29 applies to $\{u^k\}_{k\in\mathbb{N}}$. It remains to prove the uniform bound on the number of support points. By assumption we have $\#\operatorname{supp}|u^0| \leq N$. Moreover note that the set

$$\left\{ K(u(\{x\})\delta_x) \mid x \in \operatorname{supp}|u^{k+1/2}| \right\} \subset \operatorname{Im} K,$$

is linearly dependent if $\#\operatorname{supp}|u^{k+1/2}| > N$. Inductively applying Proposition 6.33 we thus conclude $\#\operatorname{supp}|u^{k+1}| \leq N$. The sparsity statement on the weak* accumulation points of $\{u^k\}_{k\in\mathbb{N}}$ now directly follows from the weak* closedness of $\mathcal{M}_N(\Omega, C)$. □

6.3.4 Acceleration strategy

The remainder of this thesis puts the focus on a fully corrective variant of Algorithm 9 where the new coefficient vector $\mathbf{u}^{k+1}$ is chosen as a minimizer of the coefficient optimization problem

$$\mathbf{u}^k \in \operatorname*{arg\,min}_{\mathbf{u}\in C^{\#\mathcal{A}_k}} [F(KU_{\mathcal{A}_k}(\mathbf{u})) + G(\|U_{\mathcal{A}_k}(\mathbf{u})\|_{\mathcal{M}})],$$

on the point set $\mathcal{A}_k = \operatorname{supp}|u^k| \cup \{\hat{x}^k\}$. The resulting method is described in Algorithm 11. In comparison to Algorithm 9 we may drop the intermediate conditional gradient step since we have $\operatorname{supp}|u^{k+1/2}| \subset \mathcal{A}_k$ and all subproblems are solved up to optimality. However the computation of the solution $v^k \in \mathcal{M}(\Omega, C)$ to the linearized problem is still necessary for the exact evaluation of the termination criterion $\Phi(u^k)$.

From this perspective the resulting algorithm can be also interpreted as a method acting on a sequence of active sets $\mathcal{A}_k$ containing a finite number of points. Recall that the support points of an optimal measure $\bar{u}$ align themselves with global maximizers of the *dual certificate*

$$\|P_C(\bar{p})\|_H \colon \Omega \to \mathbb{R}_+, \quad x \mapsto \|P_C(\bar{p}(x))\|_H.$$

In the k-th step of Algorithm 11 we greedily add a new point $\hat{x}^k$ to the active set which maximizes the violation of this constraint by the current dual certificate $\|P_C(p^k)\|_H$

$$\hat{x}^k \in \underset{x\in\Omega}{\arg\max}[\|P_C(p^k(x))\|_H - \max_{\tilde{x}\in\operatorname{supp}|u^k|} \|P_C(p^k(\tilde{x}))\|_H] = \underset{x\in\Omega}{\arg\max}\, \|P_C(p^k(x))\|_H.$$

The coefficient optimization problem $\mathfrak{P}^{\mathcal{M}}(\mathcal{A}_k)$ can then be seen as a solution of the original problem ($\mathfrak{P}^{\mathcal{M}}$) on the reduced cone $\mathcal{M}(\mathcal{A}_k, C)$. Again we emphasize that the iterates are pruned in each iteration by removing all Dirac delta functions with zero coefficient function.

In particular the description of Algorithm 11 as alternation between updating a set of active points $\mathcal{A}_k$ and solving the original problem on the reduced cone suggests a connection of the proposed procedure to the well-known Primal-Dual-Active-Set method for constrained optimization problems, [143, 177]. Before proceeding to a more detailed analysis of Algorithm 11 we highlight this similarity by a simple instructive example.

Example 6.9. *Consider the sparse minimization problem*

$$\min_{u\in\mathcal{M}^+(\Omega)} j(u) := [\frac{1}{2}\|Ku - y_d\|_Y^2 + \beta\|u\|_{\mathcal{M}}] \tag{6.56}$$

for some positive regularization parameter $\beta > 0$ and a desired state $y_d \in Y$. For simplicity we assume that K either maps to $Y = \mathbb{R}^n$, $n \in \mathbb{N}$, or $Y = L^2(\Omega_o)$ where Ω_o is a bounded domain in $\mathbb{R}^d$. Obviously this problem fits into the general framework of this section by setting $C = \mathbb{R}_+$, $F = 1/2\|\cdot - y_d\|_Y^2$ and $G(\|u\|_{\mathcal{M}}) = \beta\|u\|_{\mathcal{M}}$. By applying duality theory, see [73], we identify (6.56) *as the Fenchel dual to the state constrained problem*

$$\min_{y\in Y} j^*(y) := \frac{1}{2}\|y - y_d\|_Y^2 \quad s.t. \quad [K^*y](x) \le \beta \quad \forall x \in \Omega. \tag{6.57}$$

Since Slater's condition is satisfied in (6.57) *strong duality holds. Given a pair of minimizers $(\bar{u}, \bar{y}) \in \mathcal{M}^+(\Omega) \times Y$ to* (6.56) *and* (6.57)*, respectively, we thus conclude*

$$\bar{y} = -(K\bar{u} - y_d), \tag{6.58}$$

$$[K^*\bar{y}](x) \le \beta, \; x \in \Omega, \quad \bar{u} \in \mathcal{M}^+(\Omega), \tag{6.59}$$

$$\langle \bar{u}, K^*\bar{y} - \beta \rangle = 0. \tag{6.60}$$

Therefore the measure-valued solution $\bar{u}$ of (6.56) *can be interpreted as the Lagrange multiplier associated to the pointwise constraint in* (6.57)*. It is related to the uniquer minimizer $\bar{y}$ of* (6.57) *by the extremality conditions in* (6.59) *and* (6.60)*, respectively.*

In the following we discuss the algorithmic solution of (6.57)*. To this end it is tempting to apply a Primal-Dual-Active-Set strategy since the objective functional in* (6.57) *is quadratic and the* *admissible set is closed and convex. Formally these methods iteratively generate a sequence of active and inactive sets* $(\mathcal{A}_k, \mathcal{I}_k)_{k\in\mathbb{N}}$ *with*

$$\Omega = \mathcal{A}_k \cup \mathcal{I}_k, \quad \mathcal{A}_k \cup \mathcal{I}_k = \emptyset \quad \forall k \in \mathbb{N}$$

as well as a sequence of primal-dual variables $(u^k, y^k)_{k\in\mathbb{N}} \subset \mathcal{M}(\Omega) \times Y$ *defined by*

$$\begin{aligned} y^{k+1} &= -(Ku^{k+1} - y_d), \\ [K^* y^{k+1}](x) = \beta,\ x \in \mathcal{A}_k, \quad u^{k+1} &= 0 \quad \textit{on } \mathcal{I}_k. \end{aligned}$$

However, as already remarked in [31], such reasoning fails for state constrained problems since the choice of the active and inactive sets requires an equivalent pointwise reformulation of the extremality conditions (6.59) *and* (6.60)*. In the present case this is obviously not possible since the optimal Lagrange multipliers are only positive Radon measures. Previous approaches on the algorithmic solution of* (6.57) *are usually based on the introduction of a family of regularized problems in which the pointwise state constraint is relaxed. We refer e.g. to the well-known concepts of Lavrentiev, [193, 220], and Moreau-Yosida regularization, [144, 158], as well as barrier methods [173, 232]. All of these methods induce a path of regularized optimal solutions which can be efficiently computed and approximate* $\bar{y}$ *for vanishing regularization parameter.*

In contrast we propose a primal-dual method relying on Algorithm 11 to solve (6.57)*. Let an arbitrary primal-dual pair* $(u^k, y^k) \in \mathcal{M}(\Omega) \times Y$ *be given where* u^k *is assumed to be supported on finitely many points. We emphasize that problem* (6.57) *will be neither discretized nor regularized in the following. Our considerations are based on the particular choice of the active set as*

$$\mathcal{A}_k = \begin{cases} \operatorname{supp}|u^k| \cup \{\hat{x}^k\} & \max_{x\in\Omega}[K^* y^k](x) \geq \beta \\ \operatorname{supp}|u^k| & \textit{else} \end{cases}.$$

Here $\hat{x}^k \in \Omega$ *corresponds to a point that maximizes the violation of the state constraint by* $K^* y^k$,

$$\hat{x}^k \in \arg\max_{x\in\Omega}[[K^* y^k](x) - \beta] = \arg\max_{x\in\Omega}[K^* y^k](x).$$

The new primal-dual variables $y^{k+1} \in Y$ *and* $u^{k+1} \in \mathcal{M}^+(\Omega)$ *are then chosen to fulfill*

$$y^{k+1} = -(Ku^{k+1} - y_d), \tag{6.61}$$

$$[K^* y^{k+1}](x) \leq \beta,\ x \in \mathcal{A}_k, \quad u^{k+1}|_{\mathcal{A}_k} \in \mathcal{M}^+(\mathcal{A}_k), \quad u^{k+1}(\Omega \setminus \mathcal{A}_k) = 0, \tag{6.62}$$

$$\langle u^{k+1}, K^* y^{k+1} - \beta \rangle = 0. \tag{6.63}$$

Note that this definition ensures $u^{k+1} \in \mathcal{M}^+(\Omega)$ *i.e. the dual variables* $\{u^k\}_{k\in\mathbb{N}}$ *are feasible, as well as*

$$\operatorname{supp} u^{k+1} \subset \left\{ x \in \Omega \mid [K^* y^{k+1}](x) = \beta \right\}.$$

In contrast, the primal variables $\{y^k\}_{k\in\mathbb{N}}$ *are in general infeasible for* (6.57)*. In fact the iteration terminates at a pair of minimizers* $(\bar{y}, \bar{u}) = (y^k, u^k)$ *if* y^k *is admissible and strict complementarity holds*

$$[K^* y](x) \leq \beta,\ x \in \Omega, \quad \operatorname{supp} u^k = \arg\max_{x\in\Omega}[K^* y](x).$$

Since the active set $\mathcal{A}_k$ is finite we introduce a vector $\mathbf{u}^{k+1} \in \mathbb{R}_+^{\#\mathcal{A}_k}$ and consider the following equivalent system of nonsmooth equations

$$y^{k+1} + \sum_{x_i \in \mathcal{A}_k} \mathbf{u}_i^{k+1} K\delta_{x_i} - y_d = 0.$$
$$\mathbf{u}_i^{k+1} - \max\Big\{0, \mathbf{u}_i^{k+1} + ([K^* y^{k+1}](x_i) - \beta)\Big\} = 0,$$

for $i = 1, \dots, \#\mathcal{A}_k$. Again invoking Fenchel-Rockafellar duality theory we conclude that the pair

$$(u^{k+1}, y^{k+1}) \in \mathcal{M}^+(\Omega) \times Y,$$

fulfills (6.61)–(6.63) *if and only if*

$$y^{k+1} \in \operatorname*{arg\,min}_{y} j^*(y) \quad s.t. \quad [K^* y](x) \le \beta \quad \forall x \in \mathcal{A}_k \tag{6.64}$$

$$u^{k+1} = \sum_{x_i \in \mathcal{A}_k} \mathbf{u}_i^{k+1} \delta_{x_i}, \quad \mathbf{u}^{k+1} \in \operatorname*{arg\,min}_{\mathbf{u} \in \mathbb{R}_+^{\#\mathcal{A}_k}} j(U_{\mathcal{A}_k}(\mathbf{u})). \tag{6.65}$$

In particular the iteration can be started at $(u^0, y^0) = (0, y_d)$. Given $k \in \mathbb{N}$ the next iterate (u^{k+1}, y^{k+1}) can be computed by first eliminating the equality constraint in (6.61). *Then the vector $\mathbf{u}^{k+1}$ is determined from*

$$\mathbf{u}_i^{k+1} - \max\left\{0, \mathbf{u}_i^{k+1} - \sum_{x_j \in \mathcal{A}_k} \mathbf{u}_i^{k+1} [K^* K \delta_{x_j}](x_i) + [K^* y_d](x_i) - \beta\right\} = 0, \quad i = 1, \dots, \#\mathcal{A}_k$$

This corresponds to a solution of the finite dimensional optimization problem in (6.65) *which can be efficiently realized by e.g. semi-smooth Newton algorithms. Moreover since $\operatorname{supp} u^k \subset \mathcal{A}_k$ we can warmstart such methods by using the values of the previous coefficient vector $\mathbf{u}^k$ to construct a feasible starting point. The new primal variable is then recovered as $y^{k+1} = -(KU_{\mathcal{A}_k}(\mathbf{u}^{k+1}) - y_d)$. Following this construction we also conclude*

$$\hat{x}^k \in \operatorname*{arg\,max}_{x \in \Omega} [K^* y^k](x) = \operatorname*{arg\,max}_{x \in \Omega} -[K^*(Ku^k - y_d)](x) = \operatorname*{arg\,max}_{x \in \Omega} P_{\mathbb{R}_+}(p^k(x))$$

if $\max_{x \in \Omega} [K^* y^k](x) \ge \beta$. As a consequence one iteration of the proposed method for the solution of the state constrained problem (6.57) *is equivalent to one step of Algorithm 11 on the sparse minimization problem* (6.56) *with an additional update of the primal variable y^k.*

Anticipating the upcoming convergence results for Algorithm 11 we get (subsequential) weak convergence of the dual variables $\{u^k\}_{k \in \mathbb{N}}$ towards minimizers of* (6.56)*. Since K is weak*-to-strong continuous the primal variables thus converge strongly,*

$$y^k = -(Ku^k - y_d) \to -(K\bar{u} - y_d) = \bar{y},$$

towards the unique minimizer $\bar{y}$ of (6.57)*. Moreover from strong duality for the subproblems and the infeasibility of the primal variables we get*

$$j^*(y^k) = j(U_{\mathcal{A}_k}(\mathbf{u}^k)) = j(u^k), \quad \operatorname*{arg\,max}_{x \in \Omega} [K^* y^k](x) - \beta \ge 0.$$

Since $F = 1/2\|\cdot - y_d\|^2_Y$ on Y we conclude the following convergence results

$$r_{j^*}(y^k) + \|y^k - \bar{y}\|^2_Y + (\arg\max_{x\in\Omega}[K^* y^k](x) - \beta)^2 \leq c r_j(u^k) \leq \frac{c_1}{1+qk} \quad \forall k \in \mathbb{N}, \tag{6.66}$$

for some positive constants c_1, $q > 0$ depending on $r_{j^*}(y^0)$. This is a consequence of the interpretation of Algorithm 11 as accelerated GCG method, see Theorem 6.37, and an obvious adaption of Lemma 6.46. In particular this implies that the primal variables y^k gradually become more feasible since the maximum constraint violation tends to zero. If the number of constraints in (6.57) *is finite, i.e. Ω consist only of finitely many points, the proposed method terminates after finitely many steps at the global minimizer see Corollary 6.40.*

In the light of the results in Section 6.3.5 improved convergence rates can be expected if additional structural assumptions hold. To this end assume that the state constraint is only active in a finite collection of points at $\bar{y}$, the associated Lagrange multiplier $\bar{u}$ is unique and strict complementarity holds,

$$\operatorname{supp} \bar{u} = \{\, x \in \Omega \mid [K^*\bar{y}](x) = \beta \,\} = \{\bar{x}_i\}_{i=1}^N \subset \operatorname{int} \Omega.$$

Furthermore assume that K^* maps to (locally) smooth functions and the Hessian $\nabla^2[K^*\bar{y}](\bar{x}_i)$ at the global maximizers is negative definite. Then the improved convergence result

$$r_{j^*}(y^k) + \|y^k - \bar{y}\|_Y + (\arg\max_{x\in\Omega}[K^* y^k](x) - \beta) \leq c_2 \zeta^k \tag{6.67}$$

holds for some constants $c_2 > 0$, $\zeta \in (0,1)$ and all $k \in \mathbb{N}$ large enough. We comment on these sufficient conditions for fast convergence rates at a later point of this chapter.

To close on this instructive example we briefly discuss similar approaches from the literature. In the context of semi-infinite problems, $Y = \mathbb{R}^n$, the proposed algorithm closely resembles the so-called exchange method see e.g. [278]. While convergence of this procedure is well understood, c.f. [140, Theorem 7.2.], quantitative convergence results similar to those in (6.66) *were only provided recently in [97]. We are not aware of improved convergence results for this method comparable to those in* (6.67). *If in addition Ω contains only finitely points we recover a version of the primal-dual Goldfarb-Idnani method, [120]. Despite their similarity to the presented algorithm we point out that these methods are based on the solution of the primal subproblem* (6.64). *By construction y^k will be infeasible for* (6.64) *in general. As a consequence, in contrast to the dual subproblem, its direct numerical solution using the current primal variable y^k as a starting point is not possible.*

Algorithm 11 Primal-Dual-Active-Point strategy

```
while $\Phi(u^k) \geq \mathrm{TOL}$ do
    1. Calculate $p^k = -K^*\nabla F(Ku^k)$. Determine the new point $\hat{x}^k \in \arg\max_{x\in\Omega} \|P_C(p^k(x))\|_H$.
    2. Set $\mathcal{A}_k = \operatorname{supp}|u^k| \cup \{\, \hat{x}^k \,\}$, compute a solution $\mathbf{u}^{k+1} \in C^{\#\mathcal{A}_k}$ of $(\mathfrak{P}^{\mathcal{M}}(\mathcal{A}_k))$ . Determine the new iterate as $u^{k+1} = U_{\mathcal{A}_k}(\mathbf{u}^{k+1})$.
end while
```

From this perspective Algorithm 11 can be interpreted as a *Primal-Dual-Active-Point* method. Following the naming convention for the Primal-Dual-Active-Set strategy (PDAS) we shall refer to it as PDAP in the upcoming discussions. Due to the choice of the position $\hat{x}^k$ of the new Dirac delta function the PDAP method can be interpreted as a particular instance of the generalized conditional gradient method described in Algorithm 9. Therefore the following worst-case convergence results hold.

Theorem 6.37. *Let $\{u^k\}_{k\in\mathbb{N}}$ be generated by Algorithm 11. Then the results of Theorem 6.29 apply to $\{u^k\}_{k\in\mathbb{N}}$ with $\gamma \in (0,1)$ and $\alpha \in (0,1/2]$ chosen arbitrary.*

Proof. Observe that the first step in Algorithm 9 and 11 as well as the choice of the set

$$\mathcal{A}_k = \operatorname{supp}|u^k| \cup \{\hat{x}^k\},$$

coincide for both algorithms. The claim follows since $\bar{\mathbf{u}} \in C^N$ is chosen as a global minimizer of $j(U_{\mathcal{A}_k}(\cdot))$. □

In the following proposition first order necessary optimality conditions for solutions $\bar{\mathbf{u}} \in C^{\#\mathcal{A}}$ to the coefficient optimization problem $(\mathfrak{P}^{\mathcal{M}}(\mathcal{A}))$ are presented. To motivate the following results we point out that the nonsmooth term $G(\|u\|_{\mathcal{M}})$ in the original problem $(\mathfrak{P}^{\mathcal{M}})$ leads to a penalization of the vector $\mathbf{u} \in C^{\#\mathcal{A}}$ in the coefficient optimization problem based on its $l^1(H)$ norm

$$\|\mathbf{u}\|_{l^1(H)} = \sum_{i=1}^{\#\mathcal{A}} \|\mathbf{u}_i\|_H.$$

This type of joint or group sparse regularization is known to promote sparsity on the vector of optimal norms $(\|\bar{\mathbf{u}}_1\|_H, \dots, \|\bar{\mathbf{u}}_{\#\mathcal{A}}\|_H)^\top$.

Proposition 6.38. *Let $\mathcal{A} = \{\, x_i \in \Omega \mid i = 1, \dots, N \,\}$ be given and denote by $\bar{\mathbf{u}} \in C^N$ an optimal solution to $(\mathfrak{P}^{\mathcal{M}}(\mathcal{A}))$. Set $u = U_{\mathcal{A}}(\bar{\mathbf{u}})$ and $p = -K^*\nabla F(Ku)$. Then there holds*

$$\max_{x\in\mathcal{A}} \|P_C(p(x))\|_H \in \partial G(\|u\|_{\mathcal{M}}), \quad \langle p, u\rangle = \max_{x\in\mathcal{A}} \|P_C(p(x))\|_H \|u\|_{\mathcal{M}}.$$

If $\max_{x\in\mathcal{A}} \|P_C(p(x))\|_H \neq 0$ this is equivalent to

$$\max_{x\in\mathcal{A}} \|P_C(p(x))\|_H \in \partial G(\|u\|_{\mathcal{M}}),$$

as well as

$$\bar{\mathbf{u}}_i \neq 0 \Rightarrow \|P_C(p(x_i))\|_H = \max_{x\in\mathcal{A}} \|P_C(p(x))\|_H, \quad \frac{\bar{\mathbf{u}}_i}{\|\bar{\mathbf{u}}_i\|_H} = \frac{P_C(p(x_i))}{\max_{x\in\mathcal{A}} \|P_C(p(x))\|_H}.$$

If F is convex these conditions are sufficient for optimality.

Proof. These statements are obtained from the results in Theorem 6.22 and Proposition 6.23. To this end note that

$$\mathcal{M}(\mathcal{A}, H) \simeq (H^{\#\mathcal{A}}, \|\cdot\|_{l^1(H)}) \simeq (H^{\#\mathcal{A}}, \|\cdot\|_{l^\infty(H)})^* \simeq \mathcal{C}(\mathcal{A}, H)^*,$$

where the $l^\infty(H)$ norm of $\mathbf{u} \in H^{\#\mathcal{A}}$ is given by $\|\mathbf{u}_i\|_{l^\infty(H)} = \max_{i=1,\dots,\#\mathcal{A}} \|\mathbf{u}_i\|_H$. The cone $\mathcal{M}(\mathcal{A}, C)$ is readily identified with $C^{\#\mathcal{A}}$. Moreover the operator K can be restricted to a linear continuous operator

$$K|_{\mathcal{A}} \colon \mathcal{M}(\mathcal{A}, H) \to Y, \quad U_{\mathcal{A}}(\mathbf{u}) \mapsto \sum_{i=1}^{\#\mathcal{A}} K(\mathbf{u}_i \delta_{x_i}),$$

whose adjoint operator is given by

$$(K|_{\mathcal{A}})^* \colon Y \to \mathcal{C}(\mathcal{A}, H), \quad [(K|_{\mathcal{A}})^* y](x) = [K^* y](x),$$

for $y \in Y$ and $x \in \mathcal{A}$. □

Similar to PDAS the PDAP method terminates if the active sets in two subsequent iterations coincide. This is shown in the next corollary. Additionally, this implies convergence in finitely many steps if Ω is discrete.

Corollary 6.39. *Let $\{u^k\}_{k\in\mathbb{N}}$ be generated by PDAP. Assume that $\mathcal{A}_k = \mathcal{A}_{k+1}$ for some $k > 1$. Then $u^{k+1} \in \mathcal{M}(\Omega, C)$ is a stationary point of j, i.e. $\Phi(u^k) = 0$.*

Proof. Let $k > 1$ with $\mathcal{A}_k = \mathcal{A}_{k+1}$ be given. Then there holds

$$\hat{x}^{k+1} \in \mathcal{A}_k, \quad \|P_C(p^{k+1}(\hat{x}^k))\|_H = \|P_C(p^{k+1})\|_{\mathcal{C}} = \max_{x\in\mathcal{A}_k} \|P_C(p^k(x))\|_H.$$

Since $u^{k+1} = U_{A_k}(\mathbf{u}^{k+1})$ we conclude

$$\|p^{k+1}\|_{\mathcal{C}} \in \partial G(\|u^{k+1}\|_{\mathcal{M}}), \quad \langle p^{k+1}, u^{k+1}\rangle = \max_{x\in\mathcal{A}_k} \|P_C(p^k(x))\|_H \|u^{k+1}\|_{\mathcal{M}} = \|P_C(p^{k+1})\|_{\mathcal{C}} \|u^{k+1}\|_{\mathcal{M}}.$$

from Proposition 6.38. Invoking Theorem 6.22 it follows that u^{k+1} fulfills the variational inequality (6.37) which implies $\Phi(u^k) = 0$. □

Corollary 6.40. *Assume that $\Omega = \{\, x_i \in \mathbb{R}^d \mid i = 1, \dots, N \,\}$ for some $N \in \mathbb{N}$. Then there exists $k \in \mathbb{N}$ such that $\Phi(u^k) = 0$.*

Proof. Since the subproblems in step 2. of PDAP are solved up to optimality and $j(u^{k+1}) < j(u^k)$ if $\Phi(u^k) > 0$ we have

$$\operatorname{supp}|u^{k+1}| \in \mathcal{P}(\Omega) \setminus \bigcup_{i=1}^{k} \{\operatorname{supp}|u^k|\}.$$

Here $\mathcal{P}(\Omega)$ denotes the power sets of Ω. Since Ω only contains finitely many points Algorithm 11 will thus converge after at most $k = \#\mathcal{P}(\Omega)$ steps. □

We further derive the following estimates for the primal-dual gap $\Phi(u^k)$.

Lemma 6.41. *Assume that the sequence $\{u^k\}_{k\in\mathbb{N}}$ is generated by Algorithm 11. Set $p^k = K^*\nabla F(Ku^k)$ and $\lambda^k = \max_{x\in\operatorname{supp}|u^k|} \|P_C(p^k(x))\|_H$. Then there holds*

$$\|u^k\|_{\mathcal{M}}(\|P_C(p^k)\|_{\mathcal{C}} - \lambda^k) \leq \Phi(u^k) \leq \|v^k\|_{\mathcal{M}}(\|P_C(p^k)\|_{\mathcal{C}} - \lambda^k), \tag{6.68}$$

where v^k is determined according to Proposition 6.28. In particular, we have

$$\Phi(u^k) \leq M_0 \left(\|p^k\|_{\mathcal{C}} - \lambda^k\right).$$

Proof. By construction of v^k and u^k there holds

$$\begin{aligned}\Phi(u^k) &= \langle -p^k, u^k\rangle + G(\|u^k\|_{\mathcal{M}}) + \langle p^k, v^k\rangle - G(\|v^k\|_{\mathcal{M}}) \\ &= -\lambda^k\|u^k\|_{\mathcal{M}} + G(\|u^k\|_{\mathcal{M}}) + \|P_C(p^k)\|_{\mathcal{C}}\|v^k\|_{\mathcal{M}} - G(\|v^k\|_{\mathcal{M}}).\end{aligned}$$

Since v^k is a solution of the partially linearized problem and $\|u^k\|_{\mathcal{M}} \leq M_0$ we further obtain

$$-\|P_C(p^k)\|_{\mathcal{C}}\,\|v^k\|_{\mathcal{M}} + G(\|v^k\|_{\mathcal{M}}) \leq -\|P_C(p^k)\|_{\mathcal{C}}\,\|u^k\|_{\mathcal{M}} + G(\|u^k\|_{\mathcal{M}}),$$

which gives the first inequality. Using $\lambda^k \in \partial G(\|u^k\|_{\mathcal{M}})$, see Proposition 6.38, we estimate

$$G(\|v^k\|_{\mathcal{M}}) \geq G(\|u^k\|_{\mathcal{M}}) + \lambda^k(\|v^k\|_{\mathcal{M}} - \|u^k\|_{\mathcal{M}}),$$

which provides the second inequality. The last inequality is a consequence of $\|v^k\|_{\mathcal{M}} \leq M_0$. □

Remark 6.10. Similar to the Primal-Dual-Active-Set strategy it is also possible to base the termination criterion of PDAP on the condititiion that the active sets coincide in two consecutive iterations see Corollary 6.39. However this criterion only indicates whether a given iterate is a stationary point or not. In contrast the primal-dual gap provides a natural measure on the non-stationarity of the iterate u^k. Furthermore in the convex case it constitutes a computable upper bound on the current residual $r_j(u^k)$. Therefore we prefer to compute $\Phi(u^k)$ in practice.

6.3.5 Improved convergence analysis for PDAP

This part of the thesis is devoted to an improved convergence analysis for the Primal-Dual-Active-Point method under additional structural assumptions on the sparse minimization problem ($\mathfrak{P}^{\mathcal{M}}$). To this end we first fix some additional notation and function spaces. Associated to the sequence $\{u^k\}_{k\in\mathbb{N}}$ of iterates generated by Algorithm 11 we consider the sequences of states $\{y^k\}_{k\in\mathbb{N}} \subset Y$, $y^k = Ku^k$, adjoint states $\{p^k\}_{k\in\mathbb{N}} \subset \mathcal{C}(\Omega, H)$, $p^k = -K^*\nabla F(Ku^k)$ and dual certificates $\{P^k\}_{k\in\mathbb{N}} \subset \mathcal{C}(\Omega)$, $P^k = \|P_C(p^k)\|_H$. Furthermore we define $\lambda^k = \max_{x\in\operatorname{supp}|u^k|} P^k(x)$ for all $k \in \mathbb{N}$. If $\bar{u}$ is a weak* accumulation point of $\{u^k\}_{k\in\mathbb{N}}$ we set

$$\bar{y} = K\bar{u}, \quad \bar{p} = -K^*\nabla F(K\bar{u}), \quad \bar{P} = \|P_C(\bar{p})\|_H, \quad \bar{\lambda} = \max_{x\in\operatorname{supp}|u^k|} \bar{P}(x).$$

Moreover given an open set $\Omega_R \subset \Omega$ we denote by $\mathcal{C}^2(\bar{\Omega}_R, H)$ ($\mathcal{C}^2(\bar{\Omega}_R)$) the spaces of H-valued (scalar-valued) two times continuously differentiable functions on Ω_R whose derivatives can be continuously extended up to the boundary of Ω_R. Analogously we define the space of Lipschitz continuous functions on its closure as

$$\mathcal{C}^{0,1}(\bar{\Omega}_R, H) = \left\{ \varphi \in \mathcal{C}(\bar{\Omega}_R, H) \mid \|\varphi\|_{\mathrm{Lip}} = \sup_{\substack{x_1, x_2 \in \bar{\Omega}_R \\ x_1 \neq x_2}} \frac{\|\varphi(x_1) - \varphi(x_2)\|_H}{|x_1 - x_2|_{\mathbb{R}^d}} < \infty \right\},$$

which is a Banach space with respect to the norm

$$\|\varphi\|_{\mathcal{C}^{0,1}(\bar{\Omega}_R,H)} = \|\varphi\|_{\mathcal{C}(\bar{\Omega}_R,H)} + \|\varphi\|_{\mathrm{Lip}} \quad \forall \varphi \in \mathcal{C}^{0,1}(\bar{\Omega}_R, H).$$

Throughout this last part of the thesis we make the following additional assumptions on the smooth part $f = F \circ K$ of j and the set of admissible controls. We restrict the following considerations to the special case of $C = H$. A discussion of the derived results in the presence of additional constraints on the vector measures is given in Section 6.3.8.

Assumption 6.4. The functional $F \colon Y \to \mathbb{R} \cup \{+\infty\}$ is strictly convex and two times continuously Fréchet differentiable on

$$\widehat{Y}_{ad} := \{\, Ku \mid u \in \operatorname{dom} j \,\}.$$

Moreover it is uniformly convex around the optimal state $\bar{y} \in \operatorname{dom} F$, i.e. there exists a neighbourhood $N(\bar{y}) \subset \operatorname{dom} F$ of $\bar{y}$ in Y and a constant $\gamma_0 > 0$ with

$$(\nabla F(y_1) - \nabla F(y_2), y_1 - y_2)_Y \geq \gamma_0 \|y_1 - y_2\|_Y^2 \quad \forall y_1,\ y_2 \in N(\bar{y}).$$

Note that the smoothness assumption on F implies Lipschitz continuity of its gradient ∇F on the image of the sublevel set $E_j(u_0)$, see (6.53), for an arbitrary $u_0 \in \operatorname{dom} j$.

Proposition 6.42. *Let $u_0 \in \operatorname{dom} f$ be given. Then $\nabla F \colon \operatorname{dom} F \to Y$ is Lipschitz continuous on $KE_j(u_0)$: there exists $L_{u_0} > 0$ with*

$$\|\nabla F(y_1) - \nabla F(y_2)\|_Y \leq L_{u_0} \|y_1 - y_2\|_Y \quad \forall y_1, y_2 \in KE_j(u_0).$$

Proof. Due to the weak*-to-strong continuity of K the set $KE_j(u_0)$ is compact in Y. Thus the statement follows from the continuous differentiability of ∇F. □

In the following we derive improved local convergence results for Algorithm 11 provided that several structural assumptions on the unique adjoint state $\bar{p} \in \mathcal{C}(\Omega, H)$ as well as the dual certificate $\bar{P} \in \mathcal{C}(\Omega)$ are fulfilled. For a better illustration of the intuition behind these additional requirements we split them in two parts. First recall that the support points of the total variation measure $|\bar{u}|$ associated to a minimizer $\bar{u} \in \mathcal{M}(\Omega, H)$ align themselves with global maximizers of the dual certificate $\bar{P}$. Moreover the Radon-Nikodým derivative $\bar{u}'$ is completely characterized by the adjoint state $\bar{p}$, see Theorem 6.22.

Assumption 6.5. The dual certificate $\bar{P} \in \mathcal{C}(\Omega)$ fulfills

$$\|\bar{P}\|_{\mathcal{C}(\Omega)} > 0, \quad \left\{ x \in \Omega \mid \bar{P}(x) = \lambda \right\} = \{\bar{x}_i\}_{i=1}^N \subset \operatorname{int} \Omega.$$

Moreover the set

$$\{\, K(\bar{p}(\bar{x}_i)\delta_{\bar{x}_i}) \mid i = 1, \ldots, N \,\} \subset Y,$$

is linearly independent and there exists a radius $R > 0$ with

$$\Omega_R := \bigcup_{i=1}^{N} B_R(\bar{x}_i) \subset \operatorname{int} \Omega, \quad \bar{B}_R(\bar{x}_i) \cap \bar{B}_R(\bar{x}_j) = \emptyset,\ i \neq j, \quad K^* \colon Y \to \mathcal{C}^2(\bar{\Omega}_R, H) \cap \mathcal{C}(\Omega, H).$$

Remark 6.11. In view of Remark 6.8 on acceleration based on point moving steps we emphasize that the additional regularity assumptions on K^* are a purely analytical tool. In particular given $y \in Y$ we never have to compute derivatives of $K^* y$ in the practical implementation of the algorithm.

This assumption has two important implications. On the one hand the minimizer $\bar{u}$ to $(\mathfrak{P}^{\mathcal{M}})$ is unique and given by a finite sum of Dirac delta functions

$$\bar{u} = \sum_{i=1}^{N} \bar{\mathbf{u}}_i \delta_{\bar{x}_i}, \quad \bar{\mathbf{u}}_i = \|\bar{\mathbf{u}}_i\|_H \frac{\bar{p}(\bar{x}_i)}{\bar{\lambda}}, \quad \bar{\lambda} \in G(\|\bar{u}\|_{\mathcal{M}}),$$

where $\|\bar{\mathbf{u}}_i\|_H \in \mathbb{R}_+$, $i = 1, \ldots, N$, see Corollary 6.26. On the other hand this implies $\bar{p} \in \mathcal{C}^2(\bar{\Omega}_R, H)$ and, since we have $\bar{\lambda} > 0$, R may be chosen small enough to ensure $\bar{P} \in \mathcal{C}^2(\bar{\Omega}_R)$, see Lemma 6.66, and $P^k \in \mathcal{C}^2(\bar{\Omega}_R)$ for all $k \in \mathbb{N}$ large enough following Lemma 6.68. In particular this yields

$$\nabla \bar{P}(\bar{x}_i) = 0, \quad i = 1, \ldots, N.$$

Secondly we now assume that the curvature of $\bar{P}$ around its global maximizers does not degenerate.

Assumption 6.6. There holds $\operatorname{supp} |\bar{u}| = \{\bar{x}_i\}_{i=1}^N$, i.e. $\|\bar{\mathbf{u}}_i\|_H > 0$ for $i = 1, \ldots, N$. Furthermore we have

$$-(\zeta, \nabla^2 \bar{P}(\bar{x}_i) \zeta)_{\mathbb{R}^d} \geq \theta_0 |\zeta|^2_{\mathbb{R}^d} \quad \forall \zeta \in \mathbb{R}^d,$$

for some $\theta_0 > 0$ and all $i \in \{1, \ldots, N\}$.

Remark 6.12. In the context of super-resolution the conditions in this last assumption (for the case of $H = \mathbb{R}$) are referred to as non-degenerate source condition for the measure $\bar{u}$, see [94, 95]. Furthermore we recall the connection of sparse minimization problems to state constrained optimization, cf. Example 6.9. From this point of view the equality condition on $\operatorname{supp} |u^k|$ corresponds to a strict complementarity assumption on the Lagrange multiplier associated to the state constraint. Moreover in this case the definiteness assumption on the Hessian of $\bar{P}$ can be interpreted as a condition on the curvature of the optimal state around those points in which it touches the constraint. Both of these conditions are well-established in the field of semi-infinite optimization. We refer e.g. to [191] where similar assumptions are used to derive finite element error estimates. In [237] the author imposes comparable conditions to derive second order optimality conditions for semi-infinite optimization problems.

In order to make the following presentation more transparent we state the main result of this section beforehand. The following theorem yields improved local convergence rates for the residual $r_j(u^k)$ associated to the sequence $\{u^k\}_{k \in \mathbb{N}}$ generated by the Primal-Dual-Active-Point method. Moreover since both, the iterates u^k as well as the minimizer $\bar{u}$, are sparse we may quantify the convergence of $\{u^k\}_{k \in \mathbb{N}}$ through convergence rates for the support points of the iterates as well as their coefficient functions.

Theorem 6.43. *Let the sequence $\{u^k\}_{k\in\mathbb{N}}$ be generated by Algorithm 11 started at u^0. Assume that Assumptions 6.4, 6.5 and 6.6 hold. Then $\{u^k\}_{k\in\mathbb{N}}$ is a minimizing sequence for j and there holds*

$$u^k \rightharpoonup^* \bar{u}, \quad r_j(u^k) \leq \frac{c_1}{1+qk}, \tag{6.69}$$

for all $k \in \mathbb{N}$ and some constants $c_1,\ q>0$ which only depend on the initial residual $r_j(u^0)$ and problem dependent quantities but are otherwise independent of $\{u^k\}_{k\in\mathbb{N}}$ and $\bar{u}$. Moreover there exist $R_1>0$, $\bar{k}\in\mathbb{N}$ and $\zeta\in(0,1)$ with

$$\operatorname{supp}|u^k| \subset \bigcup_{i=1}^{N} \bar{B}_{R_1}(\bar{x}_i), \quad \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_i) \neq \emptyset,\ i=1,\dots,N,$$

as well as

$$r_j(u^k) + \max_{i=1,\dots,N} \max_{x\in\operatorname{supp}|u^k|\cap\bar{B}_{R_1}(\bar{x}_i)} |x-\bar{x}_i|_{\mathbb{R}^d} + \max_{i=1,\dots,N} \|\bar{\mathbf{u}}_i - u^k(\bar{B}_{R_1}(\bar{x}_i))\|_H \leq c_2\zeta^k, \tag{6.70}$$

for all $k \geq \bar{k}$.

Proof. For the convergence rate in (6.69) we refer to Theorem 6.29. Moreover this yields subsequential weak* convergence of $\{u^k\}_{k\in\mathbb{N}}$ towards minimizers of $(\mathfrak{P}^{\mathcal{M}})$. Since the minimizer $\bar{u}$ is unique this implies weak* convergence of the whole sequence. The claim on the localization of the support points will follow from Corollary 6.51. The improved convergence results of (6.70) are found in Theorem 6.57, Proposition 6.59 and Theorem 6.64. □

In the following $c>0$ always denotes a constant which is independent of the iteration index k. As an immediate consequence of Assumption 6.3 we obtain the following estimates.

Lemma 6.44. *Given $u_1,\ u_2 \in \mathcal{M}(\Omega,H)$ with $Ku_1,\ Ku_2 \in N(\bar{y})$, there holds*

$$j(u_1) - j(u_2) \geq \gamma_0\|K(u_1-u_2)\|_Y^2 - \Phi(u_2).$$

Proof. Due to Assumption 6.4 there holds

$$\begin{aligned} j(u_1) &= F(Ku_1) + G(\|u_1\|_{\mathcal{M}}) \\ &\geq F(Ku_2) + \gamma_0\|K(u_1-u_2)\|_Y^2 + (\nabla F(Ku_2), K(u_1-u_2))_Y + G(\|u_1\|_{\mathcal{M}}) \\ &= j(u_2) + \gamma_0\|K(u_1-u_2)\|_Y^2 - \langle\nabla f(u_2), u_2-u_1\rangle - G(\|u_2\|_{\mathcal{M}}) + G(\|u_1\|_{\mathcal{M}}) \\ &\geq j(u_2) + \gamma_0\|K(u_1-u_2)\|_Y^2 - \Phi(u_2). \end{aligned}$$

□

Corollary 6.45. *Given $u\in\mathcal{M}(\Omega,H)$ with $Ku\in N(\bar{y})$ we have*

$$\gamma_0\|K(u-\bar{u})\|_Y^2 \leq j(u) - j(\bar{u}) = r_j(u) \tag{6.71}$$

Proof. By optimality of $\bar{u}$ there holds $\Phi(\bar{u})=0$. The statement now follows directly from the previous Lemma. □

In particular the quadratic growth of j implies the following convergence rates for the states $y^k = Ku^k \in Y$ and adjoint states $p^k = K^* \nabla F(Ku^k) \in \mathcal{C}(\Omega, H)$.

Lemma 6.46. *For all $k \in \mathbb{N}$ large enough there holds*

$$\|y^k - \bar{y}\|_Y + \|p^k - \bar{p}\|_{\mathcal{C}} \leq c\sqrt{r_j(u^k)}.$$

Proof. Let us first proof the claimed estimated for the iterated states y^k. Due to the weak* convergence of $\{u^k\}_{k\in\mathbb{N}}$ towards $\bar{u}$ and the weak*-to-strong continuity of K there holds $y^k \in N(\bar{y})$ for all $k \in \mathbb{N}$ large enough. Thus we have

$$\gamma_0 \|y^k - \bar{y}\|_Y^2 \leq j(u^k) - j(\bar{u}) = r_j(u^k).$$

Taking the square root yields the first estimate. The estimates for the adjoint states can be concluded by the same arguments since

$$\|p^k - \bar{p}\|_{\mathcal{C}} = \|K^*(\nabla F(Ku^k) - \nabla F(K\bar{u}))\|_{\mathcal{C}} \leq L_{u^0} \|K^*\|_{\mathcal{L}(Y,\mathcal{C}(\Omega,H))} \|y^k - \bar{y}\|_Y.$$

This finishes the proof. □

Since the subproblems in step 2. of Algorithm 11 are solved up to optimality we conclude the following characterization of the iterates u^k.

Corollary 6.47. *For all k large enough there holds $u^k \neq 0$. Let the k-th iterate in Algorithm 11 be supported on $\{x_i^k\}_{i=1}^{N_k}$. Then we have*

$$\langle p^k, u^k \rangle = \lambda^k \|u^k\|_{\mathcal{M}}, \quad \lambda^k = \max_{x \in \operatorname{supp}|u^k|} P^k(x) \ \in \partial G(\|u^k\|_{\mathcal{M}}).$$

For all k large enough there holds $\lambda^k > 0$ and thus

$$u^k = \sum_{i=1}^{N_k} \mathbf{u}_i^k \delta_{x_i^k} = \frac{1}{\lambda^k} \sum_{i=1}^{N_k} \|\mathbf{u}_i^k\|_H p^k(x_i^k) \delta_{x_i^k}. \tag{6.72}$$

Proof. We only prove the statement on the positivity of λ_k. The remaining claims follow from Proposition 6.38 and $\operatorname{supp}|u^k| \subset \mathcal{A}_{k-1}$. From the weak* convergence of $\{u^k\}_{k\in\mathbb{N}}$, the strong convergence of p^k and the weak* lower semicontinuity of the norm we readily obtain

$$\lambda^k \|u^k\|_{\mathcal{M}} = \langle p^k, u^k \rangle \to \langle \bar{p}, \bar{u} \rangle = \bar{\lambda} \|\bar{u}\|_{\mathcal{M}}, \quad \|u^k\|_{\mathcal{M}} \geq \|\bar{u}\|_{\mathcal{M}}/2,$$

for all $k \in \mathbb{N}$ large enough. This yields $\lambda_k > 0$ for all k large enough. □

Corollary 6.48. *There holds*

$$\lim_{k\to\infty} |\bar{\lambda} - \|p^k\|_{\mathcal{C}}| + |\lambda^k - \|p^k\|_{\mathcal{C}}| = 0.$$

Proof. Observe that

$$|\bar{\lambda} - \|p^k\|_{\mathcal{C}}| = |\|\bar{p}\|_{\mathcal{C}} - \|p^k\|_{\mathcal{C}}| \leq \|\bar{p} - p^k\|_{\mathcal{C}} \leq c\sqrt{r_j(u^k)} \to 0,$$

for k going to infinity. Since $\|\bar{u}\|_{\mathcal{M}} > 0$ there exists $c > 0$ such that $\|u^k\|_{\mathcal{M}} > c$ for all k large enough. We consequently obtain

$$0 \leq c(\|p^k\|_{\mathcal{C}} - \lambda^k) \leq \Phi(u^k),$$

from Lemma 6.41. The statement now directly follows due to $\liminf_{k \to 0} \Phi(u^k) = 0$. □

Following Lemma 6.67 quadratic growth of the optimal dual certificate $\bar{P}$ in a vicinity of its global maximizers can be concluded based on Assumption 6.6. The next perturbation result states that a similar behaviour also holds true for the iterated dual certificates P^k.

Lemma 6.49. *There exists $R_1 > 0$ such that for all k large enough and all $i \in \{1, \ldots, N\}$ the function P^k assumes a unique local maximum $\hat{x}_i^k$ on $B_{R_1}(\bar{x}_i)$. Furthermore there holds*

$$|\hat{x}_i^k - \bar{x}_i|_{\mathbb{R}^d} \leq c\sqrt{r_j(u^k)}, \quad i = 1, \ldots, N. \tag{6.73}$$

Additionally there exists $R_2 > 0$ with

$$P^k(x) + \frac{\theta_0}{8}|x - \hat{x}_i^k|^2_{\mathbb{R}^d} \leq P^k(\hat{x}_i^k) \quad \forall x \in \bar{B}_{R_2}(\hat{x}_i^k), \tag{6.74}$$

for all $i = 1, \ldots, N$.

Proof. Following Lemma 6.68, $R > 0$ and $\delta > 0$ may be chosen small enough such that the mapping

$$\mathcal{F} \colon \Omega_R \times B_\delta(\bar{y}) \to \mathbb{R}^d, \quad (x, y) \mapsto \frac{\partial}{\partial x}\|[K^*\nabla F(y)](x)\|_H.$$

is well-defined and continuously Fréchet differentiable. Moreover, there holds

$$\mathcal{F}(\bar{x}_i, \bar{y}) = \nabla \bar{P}(\bar{x}_i) = 0, \quad \frac{\partial}{\partial x}\mathcal{F}(\bar{x}_i, \bar{y}) = \nabla^2 \bar{P}(\bar{x}_i) \geq \theta_0 \operatorname{Id}, \quad i = 1, \ldots, N.$$

Thus we can apply the implicit function theorem to get the existence of $0 < R_1 < R$ and $0 < \tilde{\delta} \leq \delta$ such that for all $y \in Y$ with $\|y - \bar{y}\|_Y < \tilde{\delta}$ and each $i \in \{1, \ldots N\}$ there exists a unique $\hat{x}_i(y) \in B_{R_1}(\bar{x}_i)$ with

$$\mathcal{F}(\hat{x}_i(y), y) = 0, \quad |\hat{x}_i(y) - \bar{x}_i|_{\mathbb{R}^d} \leq c\|y - \bar{y}\|_Y,$$

for some $c > 0$. Note that $y^k = Ku^k \in B_{\tilde{\delta}}(\bar{y})$ for all k large enough due to $u^k \rightharpoonup^* \bar{u}$. Setting $\hat{x}_i^k = \hat{x}_i(y^k)$ and applying Lemma 6.46 we obtain

$$|\hat{x}_i^k - \bar{x}_i|_{\mathbb{R}^d} \leq c\|y - \bar{y}\|_Y \leq c\sqrt{r_j(u^k)}.$$

Next we prove that $\hat{x}_i^k$ is a local maximum of P^k. Let an arbitrary but fixed $i \in \{1, \ldots, N\}$ be given. Note that there holds

$$-\nabla^2 P^k(\hat{x}_i^k) \geq \left(-\|\nabla^2 P^k - \nabla^2 \bar{P}\|_{\mathcal{C}(\bar{\Omega}_R, \mathbb{R}^{d \times d})} - \|\nabla^2 \bar{P}(\bar{x}_i) - \nabla^2 \bar{P}(\hat{x}_i^k)\|_{\mathbb{R}^{d \times d}} + \theta_0\right) \operatorname{Id}_{\mathbb{R}^d}$$

Due to the continuity of $\nabla^2 \bar{P}$, the uniform convergence of P^k in $\mathcal{C}^2(\Omega_R)$ and (6.73) there holds

$$\|\nabla^2 P^k - \nabla^2 \bar{P}\|_{\mathcal{C}(\bar{\Omega}_R, \mathbb{R}^{d\times d})} + \|\nabla^2 \bar{P}(\bar{x}_i) - \nabla^2 \bar{P}(\hat{x}_i^k)\|_{\mathbb{R}^{d\times d}} \leq \frac{\theta_0}{2},$$

for all k large enough. Thus for every i, $\hat{x}_i^k$ is a strict local maximum of P^k. The growth estimate for P^k in the vicinity of its maxima can be derived analogously to Lemma 6.67. This concludes the proof. □

Following these preceding results the support points of u^k are located in a vicinity of the optimal positions $\{\bar{x}_i\}_{i=1}^N$ if $k \in \mathbb{N}$ is large enough. Moreover the new support point $\hat{x}^k$ determined in step 1. of Algorithm 11 is chosen from $\{\hat{x}_i^k\}_{i=1}^N$.

Corollary 6.50. *There exists $\sigma > 0$ with*

$$\bar{P}(x) \leq \bar{\lambda} - \sigma \quad \forall x \in \Omega \backslash \bigcup_{i=1}^{N_d} B_{R_1}(\bar{x}_i) \tag{6.75}$$

and, for all k large enough, there holds

$$P^k(x) \leq \lambda^k - \frac{\sigma}{2} \quad \forall x \in \Omega \backslash \bigcup_{i=1}^{N_d} B_{R_1}(\bar{x}_i). \tag{6.76}$$

Proof. By assumption the function $\bar{P}$ does not achieve its maximum outside of $\bigcup_{i=1}^N B_{R_1}(\bar{x}_i)$. The existence of $\sigma > 0$ fulfilling (6.75) follows by a continuity argument. Let an arbitrary point $x \in \Omega \backslash \bigcup_{i=1}^N B_{R_1}(\bar{x}_i)$ be given. We estimate

$$P^k(x) \leq \bar{P}(x) + \|\bar{p} - p^k\|_{\mathcal{C}} \leq \bar{\lambda} - \sigma + \|\bar{p} - p^k\|_{\mathcal{C}} \leq \lambda^k + |\lambda^k - \bar{\lambda}| + \|\bar{p} - p^k\|_{\mathcal{C}} - \sigma.$$

Choosing k large enough such that

$$|\lambda^k - \bar{\lambda}| + \|\bar{p} - p^k\|_{\mathcal{C}} \leq \frac{\sigma}{2}$$

yields (6.76) and finishes the proof. □

Corollary 6.51. *For all k large enough there holds*

$$\operatorname{supp} |u^k| \subset \bigcup_{i=1}^N \bar{B}_{R_1}(\bar{x}_i) \quad \operatorname{supp} |u^k| \cap \bar{B}_{R_1}(\bar{x}_i) \neq \emptyset$$

for all $i = 1, \ldots, N$. Furthermore the new support point $\hat{x}^k$ determined in step 1. of Algorithm 11 fulfills

$$\hat{x}^k \in \left\{\hat{x}_i^k\right\}_{i=1}^N \subset \bigcup_{i=1}^N \bar{B}_{R_1}(\bar{x}_i).$$

Proof. Let $x \in \operatorname{supp}|u^k|$ be arbitrary. Then there holds $P^k(x) = \lambda^k$. Consequently we have $x \in \bigcup_{i=1}^N B_{R_1}(\bar{x}_i)$, see (6.76). Fix now an arbitrary index $i \in \{1, \dots, N\}$ and denote by u_i^k the restriction of u^k to $\bar{B}_{R_1}(\bar{x}_i)$. Invoking Urysohn's lemma there exists a cut-off function $\chi_i \in \mathcal{C}(\Omega)$ with $\chi_i = 1$ on $\bar{B}_{R_1}(\bar{x}_i)$ and $\chi_i = 0$ on $\bar{B}_{R_1}(\bar{x}_j)$ for $j \neq i$. The weak* convergence of the iterates and the strong convergence of the adjoint states yield

$$\lambda_k \|u_i^k\|_{\mathcal{M}} = \langle \chi_i p^k, u^k \rangle \to \langle \chi_i \bar{p}, \bar{u} \rangle = \bar{\lambda} \|\mathbf{u}_i\|_H > 0.$$

Since $\lambda^k \to \bar{\lambda}$ we conclude $\|u_i^k\|_{\mathcal{M}} = \||u_i^k|\|_{\mathcal{M}(\Omega)} \neq 0$ for all k large enough. The statement on the position of the new Dirac delta function follows directly since $P^k < \lambda^k$ outside of $\bigcup_{i=1}^N \bar{B}_{R_1}(\bar{x}_i)$ and

$$\operatorname*{arg\,max}_{x \in \bigcup_{i=1}^N \bar{B}_{R_1}(\bar{x}_i)} P^k(x) \subset \left\{ \hat{x}_i^k \right\}_{i=1}^n.$$

□

In the following corollary we show, loosely speaking, that the newly added support point $\hat{x}^k$ is also contained in the support of u^{k+1}.

Corollary 6.52. *Denote by $\hat{x}^k$ the new support point determined in step 1. of Algorithm 11. Then there holds $\hat{x}^k \in \operatorname{supp}|u^{k+1}|$ for all $k \in \mathbb{N}$.*

Proof. Since the algorithm does not converge after finitely many steps we have $j(u^{k+1}) < j(u^k)$ and

$$\operatorname{supp}|u^{k+1}| \subset \operatorname{supp}|u^k| \cup \left\{ \hat{x}^k \right\}$$

for all $k \in \mathbb{N}$. Assume now that $\hat{x}^k \notin \operatorname{supp}|u^{k+1}|$. Then there holds $\operatorname{supp} u^{k+1} \subset \operatorname{supp} u^k$ and $j(u^{k+1}) = j(u^k)$ since the subproblems in step 2. are solved up to optimality. This gives a contradiciton. □

We obtain the following estimates for the support points of $|u^k|$.

Lemma 6.53. *Let an arbitrary index $i \in \{ 1, \dots, N \}$ be given. For all k large enough there holds*

$$\max_{x \in \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_i)} |x - \bar{x}_i|_{\mathbb{R}^d} \leq c \left(\sqrt{|\lambda^k - \bar{\lambda}|} + \sqrt[4]{r_j(u^k)} \right). \tag{6.77}$$

Furthermore for k large enough there holds $\operatorname{supp} u^k \subset \bigcup_{i=1}^{N_d} \bar{B}_{R_2}(\hat{x}_i^k)$ and

$$\max_{x \in \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_i)} |x - \hat{x}_i^k|_{\mathbb{R}^d} \leq c \sqrt{P^k(\hat{x}_i^k) - \lambda^k}.$$

Proof. Given an arbitrary $i \in \{1, \dots, N\}$ we first observe that $\operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_i) \neq \emptyset$, see Corollary 6.51. Let $x \in \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_i)$. Using (6.82) we obtain

$$|x - \bar{x}_i|_{\mathbb{R}^d} \leq c \sqrt{\bar{\lambda} - \bar{P}(x)} \leq c \left(\sqrt{|\bar{\lambda} - P^k(x)|} + \sqrt{\|p^k - \bar{p}\|_{\mathcal{C}}} \right) \leq c \left(\sqrt{|\bar{\lambda} - \lambda^k|} + \sqrt[4]{r_j(u^k)} \right),$$

for some constant $c > 0$ independent of x. Here we used $P^k(x) = \lambda^k$ for all $x \in \operatorname{supp}|u^k|$ as well as Lemma 6.46. Taking the maximum over all $x \in \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_i)$ yields the first statement. For the second estimate we observe that for every $x \in \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_i)$ there holds

$$|x - \hat{x}_i^k|_{\mathbb{R}^d} \leq |x - \bar{x}_i|_{\mathbb{R}^d} + |\bar{x}_i - \hat{x}_i^k|_{\mathbb{R}^d} \leq \max_{x \in \operatorname{supp}|u^k| \cap \bar{B}_{R_2}(\bar{x}_i)} |x - \bar{x}_i|_{\mathbb{R}^d} + \sqrt{r_j(u^k)}.$$

Due to (6.77) and $\lambda^k \to \bar{\lambda}$ we get $\operatorname{supp}|u^k| \subset \bigcup_{i=1}^N \bar{B}_{R_2}(\hat{x}_i^k)$ for all k large enough. Consequently we obtain for all $i \in \{1, \ldots, N_d\}$ and $x \in \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_i)$ that there holds

$$|x - \hat{x}_i^k|_{\mathbb{R}^d} \leq c\sqrt{P^k(\hat{x}_i^k) - \lambda^k}$$

using (6.74). Since the constant $c > 0$ is again independent of x we finish the proof by maximizing on both sides. □

With these auxiliary estimates at hand we now proceed to improve on the sublinear convergence rate for the residual $r_j(u^k)$. To this end fix an arbitrary index $k \in \mathbb{N}$ large enough such that all previous results hold and recall the definition of the intermediated iterate $u^{k+1/2}$ in the generalized conditional gradient method, see Algorithm 9,

$$u_s^{k+1/2} = u^k + s\Delta_1^k, \quad \Delta_1^k = v^k - u^k, \quad v^k = \|v^k\|_{\mathcal{M}} \frac{p^k(\hat{x}^k)}{\|p^k\|_{\mathcal{C}}} \delta \hat{x}^k$$

for an appropriate choice of the stepsize $s \in [0,1]$ and $\|v^k\|_{\mathcal{M}}$ chosen according to (6.49). Obviously we have $j(u^{k+1}) \leq j(u_s^{k+1/2})$ for all $s \in [0,1]$. In fact this observation for the intermediate iterates $u_s^{k+1/2}$ remains true if we allow for more general descent directions Δ^k:

$$j(u^{k+1}) \leq j(u_s^{k+1/2}), \quad u_s^{k+1/2} = u^k + s\Delta^k, \quad \operatorname{supp}|\Delta^k| \subset \operatorname{supp}|u^k| \cup \{\hat{x}^k\}, \quad s \in [0,1],$$

since the subproblems in the PDAP method are solved up to optimality.

In the following we will construct a descent direction Δ^k and a stepsize s^k such that the residuals $r_j(u_{s^k}^{k+1/2})$, $u_{s^k}^{k+1/2} = u^k + s^k\Delta^k$, converge linearly for all $k \in \mathbb{N}$ large enough. From Corollary 6.51 we conclude the existence of an index $\hat{\imath} \in \{1, \ldots, N\}$ with $\hat{x}^k = \hat{x}_{\hat{\imath}}^k \in \bar{B}_{R_1}(\bar{x}_{\hat{\imath}})$. Define the locally lumped measure $\hat{u}_{\hat{\imath}}^k \in \mathcal{M}(\Omega, H)$ by

$$\hat{u}_{\hat{\imath}}^k = u^k_{|\bar{B}^c_{R_1}(\bar{x}_{\hat{\imath}})} + \|u^k_{|\bar{B}_{R_1}(\bar{x}_{\hat{\imath}})}\|_{\mathcal{M}} \frac{p^k(\hat{x}^k)}{\|p^k\|_{\mathcal{C}}} \delta_{\hat{x}^k},$$

where $\bar{B}^c_{R_1}(\bar{x}_{\hat{\imath}}) = \Omega \setminus \bar{B}_{R_1}(\bar{x}_{\hat{\imath}})$. The following statements establish the weak* convergence of $\hat{u}_{\hat{\imath}}^k$ towards $\bar{u}$.

Proposition 6.54. *For all $k \in \mathbb{N}$ large enough there holds*

$$G(\|\hat{u}_{\hat{\imath}}^k\|_{\mathcal{M}}) = G(\|u^k\|_{\mathcal{M}}), \quad \langle p^k, \hat{u}_{\hat{\imath}}^k - u^k \rangle = \|u^k_{|B_{R_1}(\bar{x}_{\hat{\imath}})}\|_{\mathcal{M}}(\|p^k\|_{\mathcal{C}} - \lambda^k).$$

Proof. Since the sets $\bar{B}_{R_1}(\bar{x}_i)$ are disjoint we note that

$$\begin{aligned} \|u^k\|_{\mathcal{M}} = \sum_{i=1}^N \|u^k_{|\bar{B}_{R_1}(\bar{x}_i)}\|_{\mathcal{M}} &= \sum_{i \in \{1,\ldots,N\} \setminus \{\hat{\imath}\}} \|u^k_{|\bar{B}_{R_1}(\bar{x}_i)}\|_{\mathcal{M}} + \|u^k_{|\bar{B}_{R_1}(\bar{x}_{\hat{\imath}})}\|_{\mathcal{M}} \\ &= \|u^k_{|\bar{B}^c_{R_1}(\bar{x}_{\hat{\imath}})}\|_{\mathcal{M}} + \|u^k_{|\bar{B}_{R_1}(\bar{x}_{\hat{\imath}})}\|_{\mathcal{M}} = \|\hat{u}_{\hat{\imath}}^k\|_{\mathcal{M}}, \end{aligned}$$

and consequently $G(\|\hat{u}^k_{\hat{\imath}}\|_{\mathcal{M}}) = G(\|u^k\|_{\mathcal{M}})$. Furthermore by construction there holds

$$\begin{aligned}\langle p^k, \hat{u}^k_{\hat{\imath}} - u^k\rangle &= \|u^k_{|\bar{B}_{R_1}(\bar{x}_{\hat{\imath}})}\|_{\mathcal{M}}\|p^k\|_{\mathcal{C}} - \|u^k_{|\bar{B}_{R_1}(\bar{x}_{\hat{\imath}})}\|_{\mathcal{M}}\lambda^k \\ &= \|u^k_{|\bar{B}_{R_1}(\bar{x}_{\hat{\imath}})}\|_{\mathcal{M}}(\|p^k\|_{\mathcal{C}} - \lambda^k),\end{aligned}$$

yielding the result. □

Lemma 6.55. *For k large enough there holds*

$$\|K(\hat{u}^k_{\hat{\imath}} - u^k)\|_Y \leq c\|u^k_{|B_{R_1}(\bar{x}_{\hat{\imath}})}\|_{\mathcal{M}}\sqrt{\|p^k\|_{\mathcal{C}} - \lambda^k}.$$

Proof. Let an arbitrary $x \in \operatorname{supp} u^k \cap \bar{B}_{R_1}(\bar{x}_{\hat{\imath}})$ be given and denote by $\mathbf{u} \in H,\ \mathbf{u} \neq 0$ the coefficient of the associated Dirac delta function. Given $\varphi \in Y$ there holds

$$\begin{aligned}\left(K\left(\frac{p^k(\hat{x}^k)}{\|p^k\|_{\mathcal{C}}}\delta_{\hat{x}^k} - \frac{\mathbf{u}}{\|\mathbf{u}\|_H}\delta_x\right),\ \varphi\right)_Y &= \left\langle K^*\varphi,\ \frac{p^k(\hat{x}^k)}{\|p^k\|_{\mathcal{C}}}\delta_{\hat{x}^k} - \frac{p^k(x)}{\lambda^k}\delta_x\right\rangle \\ &= \left([K^*\varphi]\,(\hat{x}^k),\ \frac{p^k(\hat{x}^k)}{\|p^k\|_{\mathcal{C}}}\right)_H - \left([K^*\varphi]\,(x),\ \frac{p^k(x)}{\lambda^k}\right)_H \\ &\leq \|K^*\varphi\|_{\mathcal{C}^{0,1}(\bar{\Omega}_R,H)}|\hat{x}^k - x|_{\mathbb{R}^d} + \|K^*\varphi\|_{\mathcal{C}}\left\|\frac{p^k(\hat{x}^k)}{\|p^k\|_{\mathcal{C}}} - \frac{p^k(x)}{\lambda^k}\right\|_H.\end{aligned}$$

Using the properties of K^* and Lemma 6.53 the first term is estimated by

$$\|K^*\varphi\|_{\mathcal{C}^{0,1}(\bar{\Omega}_R,H)}|\hat{x}^k - x|_{\mathbb{R}^d} \leq c\|\varphi\|_Y\sqrt{\|p^k\|_{\mathcal{C}} - \lambda^k},$$

with a constant $c > 0$ independent of x. For the second term we use $\|p^k(\hat{x}^k)\|_H = \|p^k\|_{\mathcal{C}}$ to estimate

$$\begin{aligned}\left\|\frac{p^k(\hat{x}^k)}{\|p^k\|_{\mathcal{C}}} - \frac{p^k(x)}{\lambda^k}\right\|_H &\leq \left|\frac{1}{\|p^k\|_{\mathcal{C}}} - \frac{1}{\lambda^k}\right|\|p^k(\hat{x}^k)\|_H + \frac{1}{\lambda^k}\|p^k(\hat{x}^k) - p^k(x)\|_H \\ &= \frac{\|p^k\|_{\mathcal{C}} - \lambda^k}{\lambda^k} + \frac{1}{\lambda^k}\|p^k(\hat{x}^k) - p^k(x)\|_H \\ &\leq \frac{1}{\lambda^k}\left[(\|p^k\|_{\mathcal{C}} - \lambda^k) + \|p^k\|_{\mathcal{C}^{0,1}(\bar{\Omega}_R,H)}|\hat{x}^k - x|_{\mathbb{R}^d}\right] \\ &\leq \frac{1}{\lambda^k}\left[\sqrt{\|p^k\|_{\mathcal{C}} - \lambda^k} + c\right]\sqrt{\|p^k\|_{\mathcal{C}} - \lambda^k},\end{aligned}$$

with c as before. Here we used $\|p^k(\hat{x}^k)\|_H = \|p^k\|_{\mathcal{C}}$ as well as $\lambda^k \leq \|p^k\|_{\mathcal{C}}$ in the first equality Since $\lambda^k \to \bar{\lambda} > 0$ and $\|p^k\|_{\mathcal{C}^{0,1}(\bar{\Omega}_R,H)} \to \|\bar{p}\|_{\mathcal{C}^{0,1}(\bar{\Omega}_R,H)} > 0$ there holds for sufficiently large k that

$$\left(K\left(\frac{p^k(\hat{x}^k)}{\|p^k\|_{\mathcal{C}}}\delta_{\hat{x}^k} - \frac{\mathbf{u}}{\|\mathbf{u}\|_H}\delta_x\right),\ \varphi\right)_Y \leq c\sqrt{\|p^k\|_{\mathcal{C}} - \lambda^k}\|\varphi\|_Y,$$

and consequently

$$\left\|K\left(\frac{p^k(\hat{x}^k)}{\|p^k\|_{\mathcal{C}}}\delta_{\hat{x}^k} - \frac{\mathbf{u}}{\|\mathbf{u}\|_H}\delta_x\right)\right\|_Y \leq c\sqrt{\|p^k\|_{\mathcal{C}} - \lambda^k}.$$

Using $\|u^k_{|\bar{B}_{R_1}(\bar{x}_{\hat{\imath}})}\|_{\mathcal{M}} = \sum_{x_i^k \in \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_{\hat{\imath}})} \|\mathbf{u}_i\|_H$, we rewrite

$$K(\hat{u}^k_{\hat{\imath}} - u^k) = \sum_{x_i^k \in \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_{\hat{\imath}})} \|\mathbf{u}_i\|_H K \left(\frac{p^k(\hat{x}^k)}{\|p^k\|_{\mathcal{C}}} \delta_{\hat{x}^k} - \frac{\mathbf{u}_i}{\|\mathbf{u}_i\|_H} \delta_{x_i^k} \right).$$

Applying the estimate for all $x_i^k \in \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_{\hat{\imath}})$ we arrive at

$$\|K(\hat{u}^k_{\hat{\imath}} - u^k)\|_Y \leq c \|u^k_{|B_{R_1}(\bar{x}_{\hat{\imath}})}\|_{\mathcal{M}} \sqrt{\|p^k\|_{\mathcal{C}} - \lambda^k},$$

completing the proof. □

Corollary 6.56. *There holds*

$$\hat{u}^k_{\hat{\imath}} \rightharpoonup^* \bar{u}, \quad j(\hat{u}^k_{\hat{\imath}}) \to j(\bar{u}).$$

Proof. We readily obtain

$$0 \leq j(\hat{u}^k_{\hat{\imath}}) - j(\bar{u}) \leq |j(u^k) - j(\bar{u})| + |F(K\hat{u}^k_{\hat{\imath}}) - F(Ku^k)|.$$

The first term tends to 0 since $\{u^k\}_{k\in\mathbb{N}}$ is a minimizing sequence for j and the second vanishes due to Lemma 6.55. Thus $\hat{u}^k_{\hat{\imath}}$ gives a minimizing sequence for j. Since $\bar{u}$ is the unique minimizer of j the claim on the weak* convergence follows. □

Finally, we show that $\Delta^k = \hat{u}^k_{\hat{\imath}} - u^k$ yields a search direction that achieves a linear decrease in the objective functional.

Theorem 6.57. *There exists an index $\bar{k} \in \mathbb{N}$, a constant $c_{\bar{k}} > 0$ and $\zeta_1 \in (0,1)$ with*

$$r_j(u^k) \leq c_{\bar{k}} \zeta_1^k \quad \forall k \geq \bar{k}.$$

Proof. For $s \in [0,1]$ define

$$u^k_s = u^k + s(\hat{u}^k_{\hat{\imath}} - u^k) = (1-s)u^k + s\hat{u}^k_{\hat{\imath}}.$$

Since $j(\hat{u}^k_{\hat{\imath}}) \to j(\bar{u})$ we conclude $u^k_s \in E_j(u^0)$ for all s and all k large enough. Let in the following k be big enough. Along the lines of proof in Lemma 6.12 it follows that

$$\begin{aligned} j(u^k_s) &= F(Ku^k_s) + G(\|u^k_s\|_{\mathcal{M}}) \\ &\leq F(Ku^k) + s(\nabla F(Ku^k), K(\hat{u}^k_{\hat{\imath}} - u^k))_Y + \frac{s^2 L_{u^0}}{2} \|K(\hat{u}^k_{\hat{\imath}} - u^k)\|_Y^2 + G(\|u^k_s\|_{\mathcal{M}}) \\ &\leq j(u^k) + s\left[\langle -p^k, \hat{u}^k_{\hat{\imath}} - u^k \rangle + G(\|\hat{u}^k_{\hat{\imath}}\|_{\mathcal{M}}) - G(\|u^k\|_{\mathcal{M}})\right] + \frac{s^2 L_{u^0}}{2} \|K(\hat{u}^k_{\hat{\imath}} - u^k)\|_Y^2, \end{aligned}$$

where L_{u^0} denotes the Lipschitz constant of ∇F on $KE_j(u^0)$. Now, by Proposition 6.54 and Lemma 6.55, we derive the estimate

$$j(u^k_s) \leq j(u^k) - s\|u^k_{|\bar{B}_{R_1}(\bar{x}_{\hat{\imath}})}\| \left(\|p^k\|_{\mathcal{C}} - \lambda^k \right) + \frac{s^2 c_1}{2} \|u^k_{|\bar{B}_{R_1}(\bar{x}_{\hat{\imath}})}\|^2 \left(\|p^k\|_{\mathcal{C}} - \lambda^k \right).$$

Minimizing for $s \in [0,1]$, we obtain

$$j(u^k_{\hat{s}^k}) \leq j(u^k) - \frac{1}{2}\min\left\{\|u^k_{|\bar{B}_{R_1}(\bar{x}_{\hat{\imath}})}\|,\ 1/c_1\right\}\left(\|p^k\|_{\mathcal{C}} - \lambda^k\right),$$

where $\hat{s}^k = \min\{\,1,\ 1/(c_1\|u^k_{|\bar{B}_{R_1}(\bar{x}_{\hat{\imath}})}\|)\,\}$ and $c_1 > 0$ is the square of the constant from Lemma 6.55. Defining the constant $c_2 > 0$ by

$$c_2 = (1/(2M_0)) \min_{i=1,\ldots,N} \min\{\,\|\bar{u}_{|\bar{B}_{R_1}(\bar{x}_i)}\|,\ 1/c_1\,\} < 1/2,$$

we have with Lemma 6.41 that

$$j(u^k_{\hat{s}}) \leq j(u^k) - c_2 M_0\left(\|p^k\|_{\mathcal{C}} - \lambda^k\right) \leq j(u^k) - c_2\Phi(u^k) \leq j(u^k) - c_2 r_j(u^k).$$

Subtracting $j(\bar{u})$ from both sides, it follows

$$r_j(u^{k+1}) \leq r_j(u^k_{\hat{s}^k}) \leq (1-c_2)r_j(u^k).$$

Denote by $\bar{k} \in \mathbb{N}$ an arbitrary but fixed index such that all previous results hold for all k greater than $\bar{k}$. By induction we get

$$r_j(u^k) \leq (1-c_2)^{k-\bar{k}} r_j(u^{\bar{k}}).$$

Setting $\zeta_1 = (1-c_2)$ and $c_{\bar{k}} = r(u^{\bar{k}})/\zeta_1^{\bar{k}}$ yields the result. □

To close this section we elaborate on the geometric intuition behind the construction of the new search direction $\varDelta^k_2 = \hat{u}^k_{\hat{\imath}} - u^k$ and the differences to the GCG direction $\varDelta^k_1 = v^k - u^k$. We consider the special case of $G(\|u\|_{\mathcal{M}}) = \beta\|u\|_{\mathcal{M}}$ for $\beta > 0$. A schematic comparison between both is given in Figure 6.1. Let us recall that by Corollary 6.51 the support of u^k can be divided into N nonempty and disjoint clusters around the optimal positions $\{\bar{x}_i\}_{i=1}^N$ for k large enough. First we consider the intermediate iterate $u^{k+1/2}_s$ given by $\varDelta^k_1$. This yields

$$u^{k+1/2}_s = u^k + s\varDelta^k_1 = (1-s)u^k + sv^k = (1-s)u^k + sM_0\frac{p^k(\hat{x}^k)}{\|p^k\|_{\mathcal{C}}}\delta_{\hat{x}^k}.$$

Thus the GCG search direction adds a single point source in one of the clusters but, by forming the convex combination, the values of u^k are changed globally. Additionally it is readily verified that every weak* accumulation point $\bar{v}$ of $\{v^k\}_{k\in\mathbb{N}}$ is given by $\bar{v} = M_0\bar{p}(\bar{x}_i)/\bar{\lambda}\delta_{\bar{x}_i}$ for some $i = 1,\ldots,N$. In particular for every sequence of stepsizes $\{s^k\}_{k\in\mathbb{N}}$ we necessarily have

$$u^{k+1}_{s^k} = (1-s^k)u^k + s^k v^k \rightharpoonup^* \bar{u} \Rightarrow s^k \to 0.$$

as $k \to \infty$ if $\bar{u}$ consists of more than one Dirac delta function. This results in the sublinear convergence of the residual. In contrast, choosing $\varDelta^k_2$ gives

$$\begin{aligned} u^{k+1/2}_s &= u^k + s\varDelta^k_2 = (1-s)u^k + s\hat{u}^k_{\hat{\imath}} \\ &= u^k_{|\bar{B}^c_{R_1}(\bar{x}_{\hat{\imath}})} + u^k_{|\bar{B}_{R_1}(\bar{x}_{\hat{\imath}})} + s\left(\|u^k_{|\bar{B}_{R_1}(\bar{x}_{\hat{\imath}})}\|_{\mathcal{M}}\frac{p^k(\hat{x}^k)}{\|p^k\|_{\mathcal{C}}}\delta_{\hat{x}^k} - u^k_{|\bar{B}_{R_1}(\bar{x}_{\hat{\imath}})}\right). \end{aligned}$$

Here we still add a single Dirac delta function to one of the clusters. However, in contrast to the GCG search direction, the norm of its coefficient is determined by moving mass from the neighbouring Dirac delta functions in the same cluster to the new one. The values of u^k on the remaining clusters remain unchanged. Moreover note that if $s = 1$ the new search direction replaces all Dirac delta functions in the cluster by the new one. Differently from the sequence $\{v^k\}_{k\in\mathbb{N}}$, the locally lumped measures $\hat{u}^k_{\hat{\imath}}$ weak* converge to the minimizer $\bar{u}$. This allows to choose a sequence of stepsizes $\{\hat{s}^k\}_{k\in\mathbb{N}}$ which is uniformly bounded from below and thus yields the improved linear convergence rate for the residual.

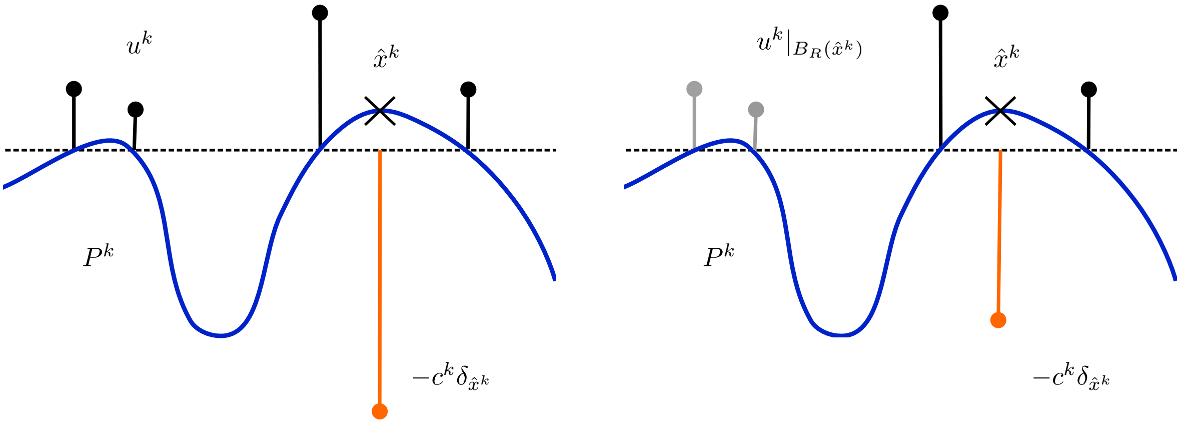


Figure 6.1: Comparison between the GCG descent direction, $c^k = M_0$, (left) and the locally lumped descent direction, $c^k = \|u^k|_{B_R(\hat{x}^k)}\|_{\mathcal{M}}$, (right) for $G(\|\cdot\|_{\mathcal{M}}) = \beta\|\cdot\|_{\mathcal{M}}$.

6.3.6 Convergence rates for the iterates

This section is devoted to quantitative convergence results for the sequence of iterates $\{u^k\}_{k\in\mathbb{N}}$. While norm convergence towards the minimizer cannot be expected in general the weak* convergence of the iterates implies convergence of the support points of u^k towards those of $\bar{u}$ as well as convergence of the coefficient functions.

Rates for the support points

We first provide an estimate for the difference between the maximum value of $\bar{P}$ and λ^k.

Lemma 6.58. *For all k large enough there exists $c > 0$ with*

$$|\bar{\lambda} - \lambda^k| \leq c\sqrt{r_j(u^{k-1})}.$$

Proof. If we choose k large enough there exists $\tilde{x}^k \in \operatorname{supp}|u^k|$ and an index $\hat{\imath}_k$ with

$$\tilde{x}^k \in \arg\max_{x\in\Omega} P^{k-1}(x), \quad |\tilde{x}^k - \bar{x}_{\hat{\imath}_k}|_{\mathbb{R}^d} \leq c\sqrt{r_j(u^{k-1})},$$

for some $c > 0$, see Corollary 6.52 and Lemma 6.49. Consequently we have

$$\begin{aligned}|\bar{\lambda} - \lambda^k| &= |\bar{P}(\bar{x}_{\hat{\imath}_k}) - P^k(\tilde{x}^k)| \leq |\bar{P}(\bar{x}_{\hat{\imath}_k}) - \bar{P}(\tilde{x}^k)| + \|\bar{p} - p^k\|_{\mathcal{C}} \\ &\leq c\left(\|\bar{p}\|_{\mathcal{C}^{0,1}(\overline{\Omega}_R,H)}|\bar{x}_{\hat{\imath}_k} - \tilde{x}^k|_{\mathbb{R}^d} + \sqrt{r_j(u^k)}\right) \\ &\leq c\sqrt{r_j(u^{k-1})},\end{aligned}$$

due to the monotonicity of $r_j(u^k)$ and Lemma 6.46. □

Putting everything together we obtain the following convergence results for the support points of the iterate u^k.

Proposition 6.59. *There exists a constant* $c > 0$ *with*

$$\max_{i=1,\dots,N} \max_{x \in \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_i)} |x - \bar{x}_i|_{\mathbb{R}^d} \le c\zeta_2^k, \tag{6.78}$$

for some $0 < \zeta_2 < 1$ *and for all* k *large enough.*

Proof. From Lemma 6.53 we get

$$\max_{i=1,\dots,N} \max_{x \in \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_i)} |x - \bar{x}_i|_{\mathbb{R}^d} \le c\left(\sqrt{|\lambda^k - \bar{\lambda}|} + \sqrt[4]{r_j(u^k)}\right).$$

Due to the monotonicity of $r_j(u^k)$, Lemma 6.57 and 6.58 there exists $0 < \zeta_1 < 1$ with

$$\sqrt{|\lambda^k - \bar{\lambda}|} + \sqrt[4]{r_j(u^k)} \le c\sqrt[4]{r_j(u^{k-1})} \le c\zeta_1^{\frac{k}{4}}. \tag{6.79}$$

By setting $\zeta_2 = \sqrt[4]{\zeta_1}$ we conclude (6.78). □

Rates for the coefficients

Let k be large enough such that all previous results hold. For $i \in \{1, \dots, N\}$ denote by u_i^k the restriction of u^k to $\bar{B}_{R_1}(\bar{x}_i)$. Due to the optimality conditions for $\bar{u}$ and u^k respectively we get

$$\bar{u} = \frac{1}{\bar{\lambda}} \sum_{i=1}^N \|\bar{\mathbf{u}}_i\|_H \bar{p}(\bar{x}_i)\delta_{\bar{x}_i}, \quad u_i^k = \frac{1}{\lambda^k} \sum_{x_i \in \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_i)} |u^k|(\{x_i\})p^k(x_i)\delta_{x_i}.$$

Recall that the iterates $\{u^k\}_{k\in\mathbb{N}}$ only converge with respect to the weak* topology on $\mathcal{M}(\Omega, H)$. Therefore a single Dirac delta function in the optimal solution $\bar{u}$ is in general approximated by several spikes in the iterate u^k, i.e. $\#\operatorname{supp}|u_i^k| > 1$ for $i = 1, \dots, N$. In particular this implies that the optimal coefficient function $\bar{\mathbf{u}}_i$ of the Dirac delta at $\bar{x}_i$ should be approximated by

$$u^k(\bar{B}_{R_1}) = \frac{1}{\lambda^k} \sum_{x_i \in \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_i)} |u^k|(\{x_i\})p^k(x_i).$$

The aim of this section is to provide a quantitative confirmation of this intuition. In detail we will prove

$$\max_{i=1,\dots,N} \|\bar{\mathbf{u}}_i - u^k(\bar{B}_{R_1}(\bar{x}_i))\|_H + \max_{i=1,\dots,N} |\|\bar{\mathbf{u}}_i\|_H - |u^k|(\bar{B}_{R_1}(\bar{x}_i))| \le c\zeta_2^k,$$

with $\zeta_2 \in (0, 1)$ as in the previous section. In the following the generic constant $c > 0$ may depend on the number of Dirac delta functions N in the minimizer $\bar{u}$. We start by providing several auxiliary results.

Lemma 6.60. *Let* x *in* $\operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_i)$ *be given. Then there holds*

$$\left\| \frac{\bar{p}(\bar{x}_i)}{\bar{\lambda}} - \frac{p^k(x)}{\lambda^k} \right\|_H \le c\zeta_2^k,$$

for some constant $c > 0$ *independent of* i *and* x.

Proof. We split the error into parts

$$\left\|\frac{\bar{p}(\bar{x}_i)}{\bar{\lambda}}-\frac{p^k(x)}{\lambda^k}\right\|_H \leq \left\|\frac{\bar{p}(\bar{x}_i)}{\bar{\lambda}}-\frac{\bar{p}(\bar{x}_i)}{\lambda^k}\right\|_H+\left\|\frac{\bar{p}(\bar{x}_i)}{\lambda^k}-\frac{\bar{p}(x)}{\lambda^k}\right\|_H+\left\|\frac{\bar{p}(x)}{\lambda^k}-\frac{p^k(x)}{\lambda^k}\right\|_H.$$

For the first term we use Lemma 6.58 to obtain

$$\left\|\frac{\bar{p}(\bar{x}_i)}{\bar{\lambda}}-\frac{\bar{p}(\bar{x}_i)}{\lambda^k}\right\|_H \leq \|\bar{p}\|_{\mathcal{C}}\frac{|\bar{\lambda}-\lambda^k|}{\bar{\lambda}\lambda^k} \leq c\zeta_2^k,$$

due to (6.79) and since $\lambda^k\bar{\lambda}$ is bounded away from zero. From the Lipschitz continuity of $\bar{p}$ and the uniform convergence of p^k the remaining terms are estimated by

$$\left\|\frac{\bar{p}(\bar{x}_i)}{\lambda^k}-\frac{\bar{p}(x)}{\lambda^k}\right\|_H+\left\|\frac{\bar{p}(x)}{\lambda^k}-\frac{p^k(x)}{\lambda^k}\right\|_H \leq \frac{c}{\lambda^k}\left(|\bar{x}_i-x|_{\mathbb{R}^d}+\|\bar{p}-p^k\|_{\mathcal{C}}\right).$$

Using (6.78) and $\|\bar{p}-p^k\|_{\mathcal{C}} \leq \sqrt[4]{r(u^{k-1})}$ for all k large enough we obtain

$$|\bar{x}_i-x|_{\mathbb{R}^d}+\|\bar{p}-p^k\|_{\mathcal{C}} \leq c\zeta_2^k,$$

independent of x, see again (6.79). Adding both estimates yields the proof. □

First we provide the convergence rate for the norms of the localized measures u_i^k, $i=1,\ldots,N$. Therefore define the auxiliary operator

$$\hat{K}\colon \mathbb{R}^N \to Y \quad v \mapsto \frac{1}{\bar{\lambda}}\sum_{i=1}^N v_i K(\bar{p}(\bar{x}_i)\delta_{\bar{x}_i}). \tag{6.80}$$

Due to the linear independence assumption in Assumption 6.4 the operator $\hat{K}$ is injective. Thus the matrix $\hat{K}^*\hat{K} \in \mathbb{R}^{N\times N}$ is invertible. We arrive at the following corollary.

Corollary 6.61. *For $v_1, v_2 \in \mathbb{R}^N$ there exists $c>0$ with*

$$|v_1-v_2|_{\mathbb{R}^N} \leq c\|\hat{K}(v_1-v_2)\|_Y.$$

Proof. There holds

$$\begin{aligned} |v_1-v_2|_{\mathbb{R}^N} &\leq \|(\hat{K}^*\hat{K})^{-1}\|_{R^{N\times N}}\|\hat{K}^*\hat{K}(v_1-v_2)\|_{R^N} \\ &\leq \|(\hat{K}^*\hat{K})^{-1}\|_{R^{N\times N}}\|\hat{K}^*\|_{\mathcal{L}(Y,\mathbb{R}^N)}\|\hat{K}(v_1-v_2)\|_Y. \end{aligned}$$

□

Lemma 6.62. *Let an arbitrary but fixed index $i \in \{1,\ldots,N\}$ be given. Then there exists $c>0$, independent of i with*

$$\left\|K\left(\|u_i^k\|_{\mathcal{M}}\frac{\bar{p}(\bar{x}_i)}{\bar{\lambda}}\delta_{\bar{x}_i}-u_i^k\right)\right\|_Y \leq c\zeta_2^k,$$

for all k large enough.

Proof. The proof follows similar steps as in Lemma 6.55. Let $x \in \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_i)$ with coefficient function $\mathbf{u} \in H$, $\mathbf{u} \neq 0$ be given. For $\varphi \in Y$ we obtain

$$\begin{aligned}\left(K\left(\frac{\bar{p}(\bar{x}_i)}{\bar{\lambda}}\delta_{\bar{x}_i} - \frac{\mathbf{u}}{\|\mathbf{u}\|_H}\delta_x\right),\, \varphi\right)_Y &= \left\langle K^*\varphi,\, \frac{\bar{p}(\bar{x}_i)}{\bar{\lambda}}\delta_{\bar{x}_i} - \frac{p^k(x)}{\lambda^k}\delta_x\right\rangle \\ &= \left([K^*\varphi](\bar{x}_i),\, \frac{\bar{p}(\bar{x}_i)}{\bar{\lambda}}\right) - \left([K^*\varphi](x),\, \frac{p^k(x)}{\lambda^k}\right)_H \\ &\leq \|K^*\varphi\|_{\mathcal{C}^{0,1}(\bar{\Omega}_R,H)}|\bar{x}_i - x|_{\mathbb{R}^d} + \|K^*\varphi\|_{\mathcal{C}}\left\|\frac{p^k(\bar{x}_i)}{\bar{\lambda}} - \frac{p^k(x)}{\lambda^k}\right\|_H \\ &\leq c\|\varphi\|_Y\zeta_2^k,\end{aligned}$$

for some constant $c > 0$ independent of x and i, see Proposition 6.59 and Lemma 6.62. Thus we conclude

$$\left\|K\left(\frac{\bar{p}(\bar{x}_i)}{\bar{\lambda}}\delta_{\bar{x}_i} - \frac{\mathbf{u}}{\|\mathbf{u}\|_H}\delta_x\right)\right\|_Y \leq c\zeta_2^k.$$

By observing that $\|u_i^k\|_{\mathcal{M}} = \sum_{x_j \in \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_j)} \|\mathbf{u}_j\|_H$ there holds

$$\begin{aligned}\left\|K\left(\|u_i^k\|_{\mathcal{M}}\frac{\bar{p}(\bar{x}_i)}{\bar{\lambda}}\delta_{\bar{x}_i} - u_i^k\right)\right\|_Y &\leq \sum_{x_j \in \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_i)} \|\mathbf{u}_j\|_H \left\|K\left(\frac{\bar{p}(\bar{x}_i)}{\bar{\lambda}}\delta_{\bar{x}_i} - \frac{\mathbf{u}_j}{\|\mathbf{u}_j\|_H}\delta_{x_j}\right)\right\|_Y \\ &\leq c\|u_i^k\|_{\mathcal{M}}\zeta_2^k \leq cM_0\zeta_2^k\end{aligned}$$

□

The following proposition characterizes the convergence behavior of $|u^k|(\bar{B}_{R_1}(\bar{x}_i)) = \|u_i^k\|_{\mathcal{M}}$.

Proposition 6.63. *There exists a constant $c > 0$ with*

$$\max_{i=1,\dots,N} |\|\bar{\mathbf{u}}_i\|_H - \|u_i^k\|_{\mathcal{M}}| \leq c\zeta_2^k,$$

for all k large enough.

Proof. Define the vectors $\bar{v}, v^k \in \mathbb{R}^N$ with $\bar{v}_i = \|\bar{\mathbf{u}}_i\|_{\mathcal{M}}$ and $v_i^k = |u^k|(\bar{B}_{R_1}(\bar{x}_i)) = \|u_i^k\|_{\mathcal{M}}$. Using Corollary 6.61 we obtain

$$\max_{i=1,\dots,N} |\|\bar{\mathbf{u}}_i\|_H - \|u_i^k\|_{\mathcal{M}}| \leq |\bar{v} - v^k| \leq c\left\|\hat{K}\left(\bar{v} - v^k\right)\right\|_Y.$$

We further estimate

$$\left\|\hat{K}\left(\bar{v} - v^k\right)\right\|_Y \leq \left\|K\left(\bar{u} - u^k\right)\right\|_Y + \sum_{i=1}^N \left\|K\left(\|u_i^k\|_{\mathcal{M}}\frac{\bar{\mathbf{u}}_i}{\|\bar{\mathbf{u}}_i\|_H}\delta_{\bar{x}_i} - u_i^k\right)\right\|_Y.$$

For the first term we get

$$\left\|K\left(\bar{u} - u^k\right)\right\|_Y \leq \sqrt{r(u^k)} \leq c\zeta_2^k,$$

for all k large enough, see Lemma 6.46. Due to Lemma 6.62 we conclude

$$\sum_{i=1}^N \left\|K\left(\|u_i^k\|_{\mathcal{M}}\frac{\bar{\mathbf{u}}_i}{\|\bar{\mathbf{u}}_i\|_H}\delta_{\bar{x}_i} - u^k\right)\right\|_Y = \sum_{i=1}^N \left\|K\left(\|u_i^k\|_{\mathcal{M}}\frac{\bar{p}(\bar{x}_i)}{\bar{\lambda}}\delta_{\bar{x}_i} - u_i^k\right)\right\|_Y \leq cN\zeta_2^k.$$

□

Summarizing all previous estimates we arrive at the following theorem.

Theorem 6.64. *There exists a constant* $c > 0$ *with*

$$\max_{i=1,\dots,N} \left\| \bar{\mathbf{u}}_i - u^k(\bar{B}_{R_1}(\bar{x}_i)) \right\|_H \leq c\zeta_2^k,$$

for all k *large enough,*

Proof. Let an arbitrary but fixed index $i \in \{1,\dots,N\}$ be given. By decomposing the norm as $\|u_i^k\|_{\mathcal{M}} = \sum_{x_j \in \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_j)} \|\mathbf{u}_j\|_H$ and using Lemma 6.60 as well as Proposition 6.63 we get

$$\begin{aligned}
\left\| \bar{\mathbf{u}}_i - u^k(\bar{B}_{R_1}(\bar{x}_i)) \right\|_H &= \left\| \int_{\bar{B}_{R_1}(\bar{x}_i)} \frac{\bar{p}(\bar{x}_i)}{\bar{\lambda}} \,\mathrm{d}|\bar{u}|(x) - \int_{\bar{B}_{R_1}(\bar{x}_i)} \frac{p^k(x)}{\lambda^k} \,\mathrm{d}|u^k|(x) \right\|_H \\
&\leq \left| |\bar{u}|(\bar{B}_{R_1}(\bar{x}_i)) - |u^k|(\bar{B}_{R_1}(\bar{x}_i)) \right| + \left\| \int_{\bar{B}_{R_1}(\bar{x}_i)} \left[\frac{\bar{p}(\bar{x}_i)}{\bar{\lambda}} - \frac{p^k(x)}{\lambda^k} \right] \mathrm{d}|u^k|(x) \right\|_H \\
&\leq \left| \|\bar{\mathbf{u}}_i\|_H - \|u_i^k\|_{\mathcal{M}} \right| + \sum_{x_j \in \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_i)} \|\mathbf{u}_j\|_H \left\| \frac{\bar{p}(\bar{x}_i)}{\bar{\lambda}} - \frac{p^k(x_j)}{\lambda^k} \right\|_H \\
&\leq cM_0\zeta_2^k,
\end{aligned}$$

with a constant $c > 0$ independent of i. Maximizing with respect to $i = 1,\dots,N$ on both sides of the inequality finishes the proof. □

Convergence rates in weaker norms

As already pointed out the norm convergence of $\{u^k\}_{k\in\mathbb{N}}$ towards the unique minimizer $\bar{u}$ in $\mathcal{M}(\Omega, H)$ cannot be expected in general. However norm convergence results can still be obtained by resorting to weaker spaces. In particular since the space of Lipschitz continuous functions embeds compactly into $\mathcal{C}(\Omega, H)$ weak* convergence on $\mathcal{M}(\Omega, H)$ implies strong convergence with respect to the canonical norm on the topological dual space of $\mathcal{C}^{0,1}(\Omega, H)$. To this end we note that

$$\|u\|_{\mathcal{C}^{0,1}(\Omega,H)^*} = \sup_{\|\varphi\|_{\mathcal{C}^{0,1}(\Omega,H)} \leq 1} \langle \varphi, u \rangle,$$

for all $u \in \mathcal{M}(\Omega, H)$. The results of the following theorem give a quantitative description of this observation.

Theorem 6.65. *There exists a constant* $c > 0$ *with*

$$\|u^k - \bar{u}\|_{\mathcal{C}^{0,1}(\Omega,H)^*} \leq c\zeta_2^k, \tag{6.81}$$

for all k *large enough.*

Proof. Let $\varphi \in \mathcal{C}^{0,1}(\Omega, H)$, $\|\varphi\|_{\mathcal{C}^{0,1}(\Omega,H)} \leq 1$ be given. We estimate

$$|\langle \varphi, u^k - \bar{u} \rangle| \leq \sum_{i=1}^{N} \left| \int_{\bar{B}_{R_1}(\bar{x}_i)} \varphi \,\mathrm{d}\bar{u}(x) - \int_{\bar{B}_{R_1}(\bar{x}_i)} \varphi \,\mathrm{d}u^k(x) \right|.$$

Fix an arbitrary index $i \in \{1, \ldots, N\}$ and split the error on the right hand side of the last inequality as

$$\left| \int_{\bar{B}_{R_1}(\bar{x}_i)} \varphi \, \mathrm{d}\bar{u}(x) - \int_{\bar{B}_{R_1}(\bar{x}_i)} \varphi \, \mathrm{d}u^k(x) \right| \\ = |(\varphi(\bar{x}_i), \bar{\mathbf{u}}_i - u^k(\bar{B}_{R_1}(\bar{x}_i)))_H| + \left| (\varphi(\bar{x}_i), u^k(\bar{B}_{R_1}(\bar{x}_i)))_H - \int_{\bar{B}_{R_1}(\bar{x}_i)} \varphi \, \mathrm{d}u^k(x) \right|.$$

The first term is bounded by

$$|(\varphi(\bar{x}_i), \bar{\mathbf{u}}_i - u^k(\bar{B}_{R_1}(\bar{x}_i)))_H| \leq \|\varphi(\bar{x}_i)\|_H \|\bar{\mathbf{u}}_i - u^k(\bar{B}_{R_1}(\bar{x}_i))\|_H \leq c\|\varphi\|_{\mathcal{C}^{0,1}(\Omega,H)} \zeta_2^k$$

for some constant $c > 0$ independent of i following Theorem 6.64. For the second term we use the Lipschitz continuity of φ to obtain

$$\left| (\varphi(\bar{x}_i), u^k(\bar{B}_{R_1}(\bar{x}_i)))_H - \int_{\bar{B}_{R_1}(\bar{x}_i)} \varphi \, \mathrm{d}u^k(x) \right| \leq \|\varphi\|_{\mathrm{Lip}} \max_{x \in \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_i)} |x - \bar{x}_i|_{\mathbb{R}^d} \|u_i^k\|_{\mathcal{M}} \leq c\zeta_2^k,$$

from the convergence results on the support points in Proposition 6.59. Again, the constant $c > 0$ can be chosen independent of the index i. Combining all previous observations we conclude

$$|\langle \varphi, u^k - \bar{u} \rangle| \leq c\zeta_2^k,$$

for some constant $c > 0$ independent of φ. Taking the supremum overall Lipschitz continuous functions $\varphi \in \mathcal{C}^{0,1}(\Omega, H)$, $\|\varphi\|_{\mathcal{C}^{0,1}(\Omega,H)} \leq 1$, on both sides of the inequality yields the claimed statement. □

Remark 6.13. We point out that, in contrast to the picturesque convergence statements for the support points and the coefficient functions, the results of the last theorem are of pure academic interest since the dual norm cannot be evaluated in general. In the case of positive measures however the same convergence rate holds true for the distance between the iterates and the minimizer with respect to a modified version of the well-known Wasserstein 1 metric see also the discussion in Section 4.4. In particular this quantity constitutes a computable upper bound on the dual norm in (6.81). While an extension of this concept to signed scalar-valued measures follows immediately we are not aware of a similar concept for the case of general vector measures.

Multiple point insertion

To close on the discussions of this section we emphasize that all of the presented results remain valid for more general choices of the active set $\mathcal{A}_k$ provided that

$$\operatorname{supp}|u^k| \cup \{\hat{x}^k\} \subset \mathcal{A}_k, \quad \#\mathcal{A}_k < \infty,$$

for all $k \in \mathbb{N}$. To this end recall that under the stated assumptions and for all $k \in \mathbb{N}$ large enough, the new Dirac delta position $\hat{x}^k$ in Algorithm 11 is taken from a finite set $\{\hat{x}_i^k\}_{i=1}^N$ where each point $\hat{x}_i^k \in B_{R_1}(\bar{x}_i)$ is given by the unique local minimizer of P^k in a vicinity of the optimal point $\bar{x}_i$. Points outside of these neighborhoods should be not considered as new positions since P^k is strictly smaller than λ^k on $\Omega \setminus \bigcup_{i=1}^N B_{R_1}(\bar{x}_i)$.

If $k \in \mathbb{N}$ is sufficiently large these considerations suggest to update the active set as

$$\mathcal{A}_k = \operatorname{supp}|u^k| \cup \left\{ x \in \Omega \mid x \in \{\hat{x}_i^k\}_{i=1}^N, \quad P^k(x) \geq \lambda^k \right\}.$$

Thus instead of only adding one global maximizer of the dual certificate to the active set we now put in all points corresponding to sufficiently large local maxima of P^k. Due to the localization of $\operatorname{supp}|u^k|$ around the optimal positions this can also be interpreted as adding up to one new Dirac delta function to each cluster in the current iterate. Intuitively this new update rule should lower the number of iterations to reduce the residual below a given threshold and improve the scalability of the method with respect to the support size of the minimizer $\bar{u}$. This intuition is backed up by the following formal reasoning. Let the active set be updated by adding the global minimizer $\hat{x}^k$ in each iteration. Assume that $\operatorname{supp}|\bar{u}| \cap \operatorname{supp}|u^k| = \emptyset$ for all $k \in \mathbb{N}$ i.e. none of the optimal positions is contained in any of the iterated supports. Fix an arbitrary index $i = 1, \dots, N$. By assumption there holds

$$\min_{x \in \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_i)} |x - \bar{x}_i| > 0, \ k \in \mathbb{N}, \qquad \max_{x \in \operatorname{supp}|u^k| \cap \bar{B}_{R_1}(\bar{x}_i)} |x - \bar{x}_i| \to 0.$$

As the movement of Dirac delta functions in u^k is not possible this means that at some point a new Dirac delta function will be inserted in the vicinity of $\bar{x}_i$. Since the index i was arbitrary and only a single point is inserted we conclude that the PDAP method eventually visits each of the N Dirac delta clusters in a separate iteration. The new definition of the active set now aims to mitigate this cycling behavior of the point insertion step by inserting new points simultaneously in all clusters. In this context we also recall that a point insertion step is always connected to one solution of $(\mathfrak{P}(\mathcal{A}_k))$. From this perspective we may also reduce the overall number of necessary solves for the coefficient optimization problems by inserting multiple points.

However these considerations are far from being conclusive and we have not been able to provide additional improved convergence results for this choice of $\mathcal{A}_k$. Moreover note that these observations are of limited practical use since all arguments are only valid in the asymptotic regime i.e. for all $k \in \mathbb{N}$ large enough and if the structural assumptions from the beginning of this section hold. In the numerical implementation of multiple point insertion strategies for the Primal-Dual-Active-Point method we resort to a heuristic procedure based on adding several local minimizers to the active set in each iteration. For a more detailed discussion we refer to Section 5.3.2.

6.3.7 Auxiliary results

In this section we summarize some technical auxiliary results that we needed in this section but were postponed until now to avoid distraction.

Lemma 6.66. *Assume that Assumption 6.5 holds. Let $\bar{p} = K^* \nabla F(K\bar{u}) \in \mathcal{C}(\Omega, H)$ be given. Define the function*

$$\bar{P} \colon \Omega \to \mathbb{R}_+ \quad x \mapsto \|\bar{p}(x)\|_H.$$

Then $R > 0$ may be chosen small enough such that $\bar{P} \in \mathcal{C}^2(\bar{\Omega}_R)$.

Proof. By Assumption 6.5 we have $\bar{p} \in \mathcal{C}^2(\bar{\Omega}_R, H)$ and $\bar{P}(\bar{x}_i) = \|\bar{p}(\bar{x}_i)\|_H = \bar{\lambda} > 0$, $i = 1, \dots, N$. In the following we denote by $\partial_{x_i}\bar{p}$, $\partial_{x_i x_j}\bar{p} \in \mathcal{C}(\bar{\Omega}_R, H)$, i, $j \in \{1, \dots, d\}$, the first and second order partial derivatives of $\bar{p}$. Note that $\bar{P} \in \mathcal{C}(\Omega)$ due to the continuity of $\bar{p}$. By continuity we may assume that $R > 0$ is chosen small enough such that $\bar{P}(x) > \bar{\lambda}/2$ for all $x \in \bigcup_{i=1}^N \bar{B}_R(\bar{x}_i)$. Using the chain rule we conclude that $\bar{P}$ is two times continuously differentiable in each $x \in \bigcup_{i=1}^N B_R(\bar{x}_i)$ with

$$\nabla \bar{P}(x)_i = \frac{(\bar{p}(x), \partial_{x_i}\bar{p}(x))_H}{\bar{P}(x)}$$
$$\nabla^2 \bar{P}(x)_{ij} = \frac{\big(\partial_{x_j}\bar{p}(x), \partial_{x_i}\bar{p}(x)\big)_H + \big(\bar{p}(x), \partial_{x_i x_j}\bar{p}(x)\big)_H}{\bar{P}(x)} - \frac{(\bar{p}(x), \partial_{x_i}\bar{p}(x))_H \big(\bar{p}(x), \partial_{x_j}\bar{p}(x)\big)_H}{\bar{P}(x)^2}$$

for all $i, j \in \{1, \ldots, d\}$. Obviously these derivatives can be continuously extended up to the boundary yielding $\bar{P} \in \mathcal{C}^2(\bar{\Omega}_R)$. □

Lemma 6.67. *There exists $R_1 > 0$ such that for all $i \in \{1, \ldots, N\}$ the quadratic growth condition*

$$\bar{P}(x) + \frac{\theta_0}{4}|x - \bar{x}_i|^2 \leq \bar{P}(\bar{x}_i) \quad \forall x \in \bar{B}_{R_1}(\bar{x}_i) \tag{6.82}$$

is satisfied.

Proof. Let an arbitrary but fixed $i \in \{1, \ldots, N\}$ be given. By Taylor expansion we obtain for $x \in \bar{B}_R(\bar{x}_i)$,

$$\bar{P}(x) = \bar{P}(\bar{x}_i) + \big(\nabla \bar{P}(\bar{x}_i), x - \bar{x}_i\big)_{\mathbb{R}^d} + \frac{1}{2}\big(x - \bar{x}_i, \nabla^2 \bar{P}(x_\zeta)(x - \bar{x}_i)\big)_{\mathbb{R}^d}$$

where $x_\zeta = (1-\zeta)x + \zeta \bar{x}_i \in \Omega_R$ for some $\zeta \in (0,1)$. Note that $\nabla \bar{P}(\bar{x}_i) = 0$ by Assumption 6.6. Using the coercivity of $\nabla^2 \bar{P}(\bar{x}_i)$ the second order term is estimated by

$$\begin{aligned}(x - \bar{x}_i, &\nabla^2 \bar{P}(x_\zeta)(x - \bar{x}_i))_{\mathbb{R}^d} \\ &\leq \big(x - \bar{x}_i, \nabla^2 \bar{P}(\bar{x}_i)(x - \bar{x}_i)\big)_{\mathbb{R}^d} + \big(x - \bar{x}_i, \nabla^2 \bar{P}(x_\zeta) - \nabla^2 \bar{P}(\bar{x}_i)(x - \bar{x}_i)\big)_{\mathbb{R}^d} \\ &\leq \big(\|\nabla^2 \bar{P}(x_\zeta) - \nabla^2 \bar{P}(\bar{x}_i)\|_{\mathbb{R}^{d\times d}} - \theta_0\big)\,|x - \bar{x}_i|^2\end{aligned}$$

Since $\nabla^2 \bar{P}$ is uniformly continuous on $\bar{\Omega}_R$ there exists $R_1 \leq R$, independent of $i \in \{1, \ldots, N_d\}$ such that

$$|x - \bar{x}_i|_{\mathbb{R}^d} \leq R_1 \Rightarrow \|\nabla^2 \bar{P}(x) - \nabla^2 \bar{P}(\bar{x}_i)\|_{\mathbb{R}^{d\times d}} \leq \frac{\theta_0}{2}.$$

Consequently, for every $x \in \bar{B}_{R_1}(\bar{x}_i)$ we obtain

$$\bar{P}(x) \leq \bar{P}(\bar{x}_i) - \frac{\theta_0}{4}|x - \bar{x}_i|^2_{\mathbb{R}^d},$$

proving (6.82) since i was arbitrary. □

Lemma 6.68. *Define the mapping*

$$P\colon \operatorname{dom} F \to \mathcal{C}(\Omega) \quad y \mapsto \|[K^*\nabla F(Ky)](\cdot)\|_H$$

Furthermore let $\bar{y} = K\bar{u}$. Then there exists $\delta > 0$ such that $P \in \mathcal{C}^1(B_\delta(\bar{y}), \mathcal{C}^2(\bar{\Omega}_R))$. In particular the mapping

$$\mathcal{F}\colon \Omega_R \times B_\delta(\bar{y}) \to \mathbb{R}^d, \quad (x, y) \mapsto \frac{\partial}{\partial x}\|[K^*\nabla F(y)](x)\|_H, \tag{6.83}$$

is continuously Fréchet differentiable.

Proof. Due to the continuity of K^*, ∇F and the norm there exists $\delta > 0$ such that

$$[P(y)](x) > \frac{\bar{\lambda}}{4} \quad \forall x \in \bigcup_{i=1}^{N_d} \bar{B}_R(\bar{x}_i)$$

for all y with $\|y - \bar{y}\|_Y \leq \delta$. Arguing as in Lemma 6.66 we conclude $P(y) \in \mathcal{C}^2(\Omega_R)$. As for $\bar{P}$ we can derive formulas for the gradient $[\nabla P(y)]$ and the Hessian $[\nabla^2 P(y)]$ which depend differentiable on y since F is two times continuously Fréchet differentiable and K^* maps continuously into $\mathcal{C}^2(\bar{\Omega}_R)$. In particular we obtain

$$\nabla[P(y)](x)_i = \mathcal{F}(x,y)_i = \frac{\Big([K^*\nabla F(y)](x), \frac{\partial}{\partial x_i}[K^*\nabla F(y)](x)\Big)_H}{P(x,y)} \quad i = 1, \ldots, d.$$

Thus the partial derivatives of $\mathcal{F}$ with respect to x and y exist on $\Omega_R \times B_\delta(\bar{y})$ and are continuous. Continuous Fréchet differentiability of the mapping in (6.83) now follows from Proposition 3.2.18 and Remark 3.2.19 in [89]. □

Lemma 6.69. *Let a compact set $\Omega \subset \mathbb{R}^d$ be given and assume that $K^* \colon Y \mapsto \mathcal{C}^{0,1}(\Omega, H)$ is linearly and continuous. Let $\mathbf{u}_1, \mathbf{u}_2 \in H$, $x_1, x_2 \in \Omega$ be given. Then there exists $c > 0$ only depending on K with*

$$\|K(\mathbf{u}_1\delta_{x_1}) - K(\mathbf{u}_1\delta_{x_2})\|_Y \leq c\|\mathbf{u}_1\|_H |x_1 - x_2|_{\mathbb{R}^d}$$
$$\|K(\mathbf{u}_1\delta_{x_1}) - K(\mathbf{u}_2\delta_{x_1})\|_Y \leq c\|\mathbf{u}_1 - \mathbf{u}_2\|_H.$$

Proof. For $\varphi \in Y \backslash \{0\}$ we obtain

$$\begin{aligned}(K(\mathbf{u}_1\delta_{x_1}) - K(\mathbf{u}_1\delta_{x_2}), \varphi)_Y &= \langle \mathbf{u}_1(\delta_{x_1} - \delta_{x_2}), [K^*\varphi]\rangle \leq \|\mathbf{u}_1\|_H \|[K^*\varphi](x_1) - K^*\varphi(x_2)\|_H \\ &\leq \|\mathbf{u}_1\|_H \|K^*\varphi\|_{\mathcal{C}^{0,1}(\Omega,H)} |x_1 - x_2|_{\mathbb{R}^d} \\ &\leq \|\mathbf{u}_1\|_H \|K^*\|_{\mathcal{L}(Y,\mathcal{C}^{0,1}(\Omega,H))} \|\varphi\|_Y |x_1 - x_2|_{\mathbb{R}^d}.\end{aligned}$$

Analogously we get

$$(K(\mathbf{u}_1\delta_{x_1}) - K(\mathbf{u}_2\delta_{x_1}), \varphi)_Y \leq \|K^*\varphi\|_{\mathcal{C}} \|\mathbf{u}_1 - \mathbf{u}_2\|_H \leq \|K^*\|_{\mathcal{L}(Y,\mathcal{C}(\Omega,H))} \|\varphi\|_Y \|\mathbf{u}_1 - \mathbf{u}_2\|_H$$

Dividing both sides of the inequalities by $\|\varphi\|_Y$ and taking the supremum over all $\varphi \in Y\backslash\{0\}$ we conclude both estimates. □

6.3.8 A note on conic constraints

In this last section we comment on improved convergence results for the Primal-Dual-Active-Point method in the case of conic constraints i.e. $C \neq H$. Let Assumptions 6.4, 6.5, 6.6 hold and denote by $\bar{u} = \sum_{i=1}^N \bar{\mathbf{u}}_i \delta_{\bar{x}_i}$ the unique minimizer to $(\mathfrak{P}^{\mathcal{M}})$. By $\bar{p}$, $\bar{P}$ and p^k, P^k we refer to the adjoint states and dual certificates associated to $\bar{u}$ and u^k, respectively. Let us first recall the unconstrained case, i.e. $C = H$. In this situation we based our proof on the local smoothness of the adjoint states around the optimal support points. Moreover, since $\bar{p}(\bar{x}_i) \neq 0$, this regularity also transfers to the dual certificates which, together with Assumption 6.6, allowed to establish the perturbation results of Lemma 6.49. Obviously such reasoning fails in the constrained situation $C \neq H$ since

$$\varphi \in \mathcal{C}^2(\bar{\Omega}_R, H) \not\Rightarrow P_C(\varphi) \in \mathcal{C}^2(\bar{\Omega}_R, H),$$

in general. This is for example the case if there exists an index $i \in \{1, \dots, N\}$ such that $P_C(\varphi(\bar{x}_i))$ lies at the boundary of C.

While this observation prevents a direct adaptation of the presented results to the general constrained case the aforementioned difficulty can be bypassed if the optimal adjoint state $\bar{p}$ maps locally into the interior of C in H. To this end let us assume that $\operatorname{int} C \neq \emptyset$. In particular this encompasses the important case of positive scalar-valued measures. Furthermore assume that $\bar{p}(\bar{x}_i) \in \operatorname{int} C$ for $i = 1, \dots, N$. Due to the projection formula for the optimal coefficient functions $\bar{p}(\bar{x}_i)/\|\bar{p}\|_{\mathcal{C}} = \bar{\mathbf{u}}_i/\|\bar{\mathbf{u}}_i\|_H$ this is equivalent to $\bar{\mathbf{u}}_i \in \operatorname{int} C$. Since $\bar{p}$ is continuous the set Ω_R can be chosen small enough such that $\bar{p}(x) \in \operatorname{int} C$ for all $x \in \bar{\Omega}_R$. Thus we obtain $\bar{P}(x) = \|P_C(\bar{p}(x))\|_H = \|\bar{p}(x)\|_H$ on $\bar{\Omega}_R$. This yields $\bar{P} \in \mathcal{C}^2(\bar{\Omega}_R)$ following Lemma 6.66. Furthermore arguing as in Lemma 6.68 gives $\|[K^*\nabla F(y)](\cdot)\|_H \in \mathcal{C}^2(\bar{\Omega}_R)$ for all y in a neighborhood of $\bar{y}$ and, in particular, $P^k \in \mathcal{C}^2(\bar{\Omega}_R)$ for all $k \in \mathbb{N}$ large enough.

The remaining results of Section 6.3.5 are now obtained by repeating the presented arguments. In particular note that the intermediate iterates $u_s^{k+1/2}$, $s \in [0,1]$ in the proof of Theorem 6.57 are admissible since $P_C(p^k(\hat{x}^k)) = p^k(\hat{x}^k)$ for all $k \in \mathbb{N}$ large enough and

$$\begin{aligned} u_s^{k+1/2} &= u^k + s\Delta_2^k = (1-s)u^k + s\hat{u}_{\hat{\imath}}^k \\ &= u^k_{|\bar{B}^c_{R_1}(\bar{x}_{\hat{\imath}})} + u^k_{|\bar{B}_{R_1}(\bar{x}_{\hat{\imath}})} + s\left(\|u^k_{|\bar{B}_{R_1}(\bar{x}_{\hat{\imath}})}\|_{\mathcal{M}} \frac{P_C(p^k(\hat{x}^k))}{\|P_C(p^k)\|_{\mathcal{C}}}\delta_{\hat{x}^k} - u^k_{|\bar{B}_{R_1}(\bar{x}_{\hat{\imath}})}\right) \in \mathcal{M}(\Omega, C), \end{aligned}$$

due to $u^k_{|\bar{B}^c_{R_1}(\bar{x}_{\hat{\imath}})}$, $u^k_{|\bar{B}_{R_1}(\bar{x}_{\hat{\imath}})}$, $P_C(p^k(\hat{x}^k))\delta_{\hat{x}^k} \in \mathcal{M}(\Omega, C)$.

As a consequence of these considerations we conclude the following convergence result in the case of additional conic constraints.

Theorem 6.70. *Let $C \subset H$ be a closed and convex cone with nonempty interior in H. Let Assumptions 6.4, 6.5, 6.6 hold and denote by $\bar{u} \in \mathcal{M}(\Omega, C)$ the unique minimizer to $(\mathfrak{P}^{\mathcal{M}})$. Further assume that $\bar{p}(x) \in \operatorname{int} C$ for all $x \in \operatorname{supp}|\bar{u}|$. Then Theorem 6.43 applies to $\{u^k\}_{k\in\mathbb{N}}$.*

Remark 6.14. Please note that for the important case of scalar measures with positivity constraints, i.e. $C = \mathbb{R}_+$, the additional condition $\bar{p}(\bar{x}_i) \in \operatorname{int} \mathbb{R}_+$ is redundant since we assume that strict complementarity, $\operatorname{supp}|\bar{u}| = \{\bar{x}_i\}_{i=1}^N$, holds.

Let us briefly discuss the limits of this extension. While these additional assumptions allow to extend the improved convergence results to, e.g., the case of positivity constraints in $\mathcal{M}(\Omega, \mathbb{R}^n)$ it obviously does not cover all interesting and practical relevant settings. For example the interior of the cone $L^2_+(I) \subset L^2(I)$ from Example 6.6 is empty. We postpone a deeper discussion of improved convergence rates in this case to future work. Another interesting point that should be addressed is to derive rigorous mesh-independence results, following e.g. the concepts presented in [162], for the proposed method. Moreover, the observed practical efficiency of the Primal-Dual-Active-Point method as well as the improved convergence results of this chapter should serve as a motivation to study accelerated conditional gradient algorithms for different problems posed in nonreflexive Banach spaces. For interesting and practically relevant examples, we again point out minimization problems involving spaces of functions with bounded total variation or the time-dependent measure-valued controls in Example 6.5.

Acknowledgement

First and foremost, I would like to thank my supervisor Prof. Dr. Boris Vexler for introducing me to the concept of sparse controls as well as his advice and guidance throughout the past years. Moreover, I am grateful to him for giving me the opportunity and encouraging me to participate in international conferences. Many results in this thesis benefit greatly from discussions at such occasions. I would also like to thank my second supervisor Prof. Dr. Karl Kunisch for his valuable remarks, our always enjoyable discussions as well as for sharing his expertise with me, in particular during my stays in Graz.

Moreover, my sincere thanks go to Prof. Dr. Ira Neitzel for mentoring me since I was a Bachelor's student. In the same way, I want to express my deepest gratitude towards Dr. Konstantin Pieper for our fruitful collaborations, hour-long Skype discussions and two very productive stays at Florida State University. I would also like to thank all of my colleagues at the institutes in Munich and Graz for providing a friendly and vibrant working atmosphere.

I gratefully acknowledge support from the International Research Training Group IGDK *Optimization and Numerical Analysis for Partial Differential Equations with Nonsmooth Structures*, funded by the German Science Foundation (DFG) and the Austrian Science Fund (FWF). My deepest gratitude goes to all members of the IGDK. I would especially like to thank Bao, Konstantin and Philip for our collaboration during the IGDK summer workshop in 2015 which led to our paper on acoustic inversion with the Helmholtz equation. My special thanks also go to Diane Clayton-Winter, the good soul of the IGDK in Munich, who has, more than once, saved me from missing important deadlines and provided indispensable help with the organization of our workshop on optimal design of experiments in 2017. Furthermore, this dissertation was also written within the scope of the Elite Graduate Program TopMath supported by the Elite Network of Bavaria (ENB).

I would further like to thank my family. Without your continuous support none of this would have been possible. Last, but certainly not least, I want to express my deepest gratitude towards all of my friends, especially those at home in Upper Franconia. Whenever necessary, you reminded me that there is also something beyond mathematics. In particular, I would like to thank Benjamin and Tobias for being the best friends one could wish for.

Bibliography

[1] S. D. Ahipasaoglu, *A first-order algorithm for the A-optimal experimental design problem: a mathematical programming approach*, Stat. Comput., 25 (2015), pp. 1113–1127.

[2] S. D. Ahipasaoglu, P. Sun, and M. J. Todd, *Linear convergence of a modified Frank-Wolfe algorithm for computing minimum-volume enclosing ellipsoids*, Optim. Methods Softw., 23 (2008), pp. 5–19.

[3] A. Alexanderian, P. J. Gloor, and O. Ghattas, *On Bayesian A- and D-optimal experimental designs in infinite dimensions*, Bayesian Anal., 11 (2016), pp. 671–695.

[4] A. Alexanderian, N. Petra, G. Stadler, and O. Ghattas, *A-optimal design of experiments for infinite-dimensional Bayesian linear inverse problems with regularized ℓ_0-sparsification*, SIAM J. Sci. Comput., 36 (2014), pp. A2122–A2148.

[5] ——, *A fast and scalable method for A-optimal design of experiments for infinite-dimensional Bayesian nonlinear inverse problems*, SIAM J. Sci. Comput., 38 (2016), pp. A243–A272.

[6] A. Alexanderian and A. K. Saibaba, *Efficient D-optimal design of experiments for infinite-dimensional Bayesian linear inverse problems*, ArXiv e-prints, (2017).

[7] C. D. Aliprantis and K. C. Border, *Infinite dimensional analysis*, Springer, Berlin, third ed., 2006. A hitchhiker's guide.

[8] H. Amann and J. Escher, *Analysis. III*, Birkhäuser Verlag, Basel, 2009. Translated from the 2001 German original by Silvio Levy and Matthew Cargo.

[9] A. C. Atkinson, A. N. Donev, and R. D. Tobias, *Optimum experimental designs, with SAS*, vol. 34 of Oxford Statistical Science Series, Oxford University Press, Oxford, 2007.

[10] A. Attia, A. Alexanderian, and A. K. Saibaba, *Goal-oriented optimal design of experiments for large-scale Bayesian linear inverse problems*, Inverse Problems, 34 (2018), p. 095009.

[11] C. L. Atwood, *Sequences converging to D-optimal designs of experiments*, Ann. Statist., 1 (1973), pp. 342–352.

[12] J.-P. Aubin, *Applied functional analysis*, Pure and Applied Mathematics (New York), Wiley-Interscience, New York, second ed., 2000. With exercises by Bernard Cornet and Jean-Michel Lasry, Translated from the French by Carole Labrousse.

[13] M. Avery, H. T. Banks, K. Basu, Y. Cheng, E. Eager, S. Khasawinah, L. Potter, and K. L. Rehm, *Experimental design and inverse problems in plant biological modeling*, J. Inverse Ill-Posed Probl., 20 (2012), pp. 169–191.

[14] F. Bach, *Duality between subgradient and conditional gradient methods*, SIAM J. Optim., 25 (2015), pp. 115–129.

[15] M. Bambach, M. Heinkenschloss, and M. Herty, *A method for model identification and parameter estimation*, Inverse Problems, 29 (2013), pp. 025009, 19.

[16] H. T. Banks and K. Kunisch, *Estimation techniques for distributed parameter systems*, vol. 1 of Systems & Control: Foundations & Applications, Birkhäuser Boston, Inc., Boston, MA, 1989.

[17] H. T. Banks and K. L. Rehm, *Experimental design for vector output systems*, Inverse Probl. Sci. Eng., 22 (2014), pp. 557–590.

[18] A. Bardow, *Optimal experimental design of ill-posed problems: The meter approach*, Computers & Chemical Engineering, 32 (2008), pp. 115–124.

[19] D. M. Bates and D. G. Watts, *Nonlinear regression analysis and its applications*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, Inc., New York, 1988.

[20] I. Bauer, H. G. Bock, S. Körkel, and J. P. Schlöder, *Numerical methods for optimum experimental design in DAE systems*, J. Comput. Appl. Math., 120 (2000), pp. 1–25. SQP-based direct discretization methods for practical optimal control problems.

[21] J. Baumeister, *Stable solution of inverse problems*, Advanced Lectures in Mathematics, Friedr. Vieweg & Sohn, Braunschweig, 1987.

[22] H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*, CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, Springer, New York, 2011. With a foreword by Hédy Attouch.

[23] E. M. L. Beale, *Confidence regions in non-linear estimation*, J. Roy. Statist. Soc. Ser. B, 22 (1960), pp. 41–88.

[24] A. Beck and S. Shtern, *Linearly convergent away-step conditional gradient for non-strongly convex functions*, Math. Program., 164 (2017), pp. 1–27.

[25] A. Beck and M. Teboulle, *A conditional gradient method with linear rate of convergence for solving convex linear systems*, Math. Methods Oper. Res., 59 (2004), pp. 235–247.

[26] ——, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.

[27] ——, *Gradient-based algorithms with applications to signal-recovery problems*, in Convex optimization in signal processing and communications, Cambridge Univ. Press, Cambridge, 2010, pp. 42–88.

[28] R. Becker, M. Braack, and B. Vexler, *Parameter identification for chemical models in combustion problems*, Appl. Numer. Math., 54 (2005), pp. 519–536.

[29] O. Benedix and B. Vexler, *A posteriori error estimation and adaptivity for elliptic optimal control problems with state constraints*, Comput. Optim. Appl., 44 (2009), pp. 3–25.

[30] P. Benner, R. Herzog, N. Lang, I. Riedel, and J. Saak, *Comparison of model order reduction methods for optimal sensor placement for thermo-elastic models*, Engineering Optimization, 0 (2018), pp. 1–19.

[31] M. Bergounioux and K. Kunisch, *Primal-dual strategy for state-constrained optimal control problems*, Comput. Optim. Appl., 22 (2002), pp. 193–224.

[32] A. Bermúdez, P. Gamallo, and R. Rodrí guez, *Finite element methods in local active control of sound*, SIAM J. Control Optim., 43 (2004), pp. 437–465.

[33] S. J. Bernau, *The square root of a positive self-adjoint operator*, J. Austral. Math. Soc., 8 (1968), pp. 17–36.

[34] D. P. Bertsekas, *On the Goldstein-Levitin-Polyak gradient projection method*, IEEE Trans. Automatic Control, AC-21 (1976), pp. 174–184.

[35] ——, *Nonlinear programming*, Athena Scientific Optimization and Computation Series, Athena Scientific, Belmont, MA, third ed., 2016.

[36] H. G. Bock, *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.

[37] H. G. Bock, S. Körkel, E. Kostina, and J. P. Schlöder, *Robustness aspects in parameter estimation, optimal design of experiments and optimal control*, in Reactive flows, diffusion and transport, Springer, Berlin, 2007, pp. 117–146.

[38] H. G. Bock, S. Körkel, and J. P. Schlöder, *Parameter estimation and optimum experimental design for differential equation models*, in Model based parameter estimation, vol. 4 of Contrib. Math. Comput. Sci., Springer, Heidelberg, 2013, pp. 1–30.

[39] D. Boffi, *Finite element approximation of eigenvalue problems*, Acta Numer., 19 (2010), pp. 1–120.

[40] V. I. Bogachev, *Gaussian measures*, vol. 62 of Mathematical Surveys and Monographs, American Mathematical Society, Providence, RI, 1998.

[41] D. Böhning, *A vertex-exchange-method in D-optimal design theory*, Metrika, 33 (1986), pp. 337–347.

[42] R. Bonic, *Some properties of Hilbert scales*, Proc. Amer. Math. Soc., 18 (1967), pp. 1000–1003.

[43] J. F. Bonnans and A. Shapiro, *Perturbation analysis of optimization problems*, Springer Series in Operations Research, Springer-Verlag, New York, 2000.

[44] N. Boyd, G. Schiebinger, and B. Recht, *The alternating descent conditional gradient method for sparse inverse problems*, SIAM J. Optim., 27 (2017), pp. 616–639.

[45] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends Mach. Learn., 3 (2011), pp. 1–122.

[46] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, Cambridge, 2004.

[47] K. Bredies and D. A. Lorenz, *Iterated hard shrinkage for minimization problems with sparsity constraints*, SIAM J. Sci. Comput., 30 (2008), pp. 657–683.

[48] ——, *Linear convergence of iterative soft-thresholding*, J. Fourier Anal. Appl., 14 (2008), pp. 813–837.

[49] K. Bredies, D. A. Lorenz, and P. Maass, *A generalized conditional gradient method and its connection to an iterative shrinkage method*, Comput. Optim. Appl., 42 (2009), pp. 173–193.

[50] K. Bredies and H. K. Pikkarainen, *Inverse problems in spaces of measures*, ESAIM Control Optim. Calc. Var., 19 (2013), pp. 190–218.

[51] S. C. Brenner and L. R. Scott, *The mathematical theory of finite element methods*, vol. 15 of Texts in Applied Mathematics, Springer, New York, third ed., 2008.

[52] H. Brezis, *Functional analysis, Sobolev spaces and partial differential equations*, Universitext, Springer, New York, 2011.

[53] C. Brislawn, *Kernels of trace class operators*, Proc. Amer. Math. Soc., 104 (1988), pp. 1181–1190.

[54] P. Brunner, C. Clason, M. Freiberger, and H. Scharfetter, *A deterministic approach to the adapted optode placement for illumination of highly scattering tissue*, Biomed. Opt. Express, 3 (2012), pp. 1732–1743.

[55] E. J. Candès and C. Fernandez-Granda, *Towards a mathematical theory of super-resolution*, Comm. Pure Appl. Math., 67 (2014), pp. 906–956.

[56] M. D. Canon and C. D. Cullum, *A tight upper bound on the rate of convergence of Frank-Wolfe algorithm*, SIAM J. Control, 6 (1968), pp. 509–516.

[57] T. Carraro, *Parameter estimation and optimal experimental design in flow reactors*, PhD Dissertation, Ruprecht-Karls-Universität Heidelberg, 2005. `https://archiv.ub.uni-heidelberg.de/volltextserver/6114/`.

[58] E. Casas, *Optimal control in coefficients of elliptic equations with state constraints*, Appl. Math. Optim., 26 (1992), pp. 21–37.

[59] E. Casas, C. Clason, and K. Kunisch, *Approximation of elliptic control problems in measure spaces with sparse solutions*, SIAM J. Control Optim., 50 (2012), pp. 1735–1752.

[60] ——, *Parabolic control problems in measure spaces with sparse solutions*, SIAM J. Control Optim., 51 (2013), pp. 28–63.

[61] E. Casas, R. Herzog, and G. Wachsmuth, *Approximation of sparse controls in semilinear equations by piecewise linear functions*, Numer. Math., 122 (2012), pp. 645–669.

[62] ——, *Optimality conditions and error analysis of semilinear elliptic control problems with L^1 cost functional*, SIAM J. Optim., 22 (2012), pp. 795–820.

[63] E. Casas and K. Kunisch, *Optimal control of the 2d stationary navier-stokes equations with measure-valued controls.* Preprint.

[64] ——, *Optimal control of semilinear elliptic equations in measure spaces*, SIAM J. Control Optim., 52 (2014), pp. 339–364.

[65] E. Casas and F. Tröltzsch, *Second order analysis for optimal control problems: improving results expected from abstract theory*, SIAM J. Optim., 22 (2012), pp. 261–279.

[66] E. Casas, D. Wachsmuth, and G. Wachsmuth, *Sufficient second-order conditions for bang-bang control problems*, SIAM J. Control Optim., 55 (2017), pp. 3066–3090.

[67] E. Casas and E. Zuazua, *Spike controls for elliptic and parabolic PDEs*, Systems Control Lett., 62 (2013), pp. 311–318.

[68] K. Chaloner and I. Verdinelli, *Bayesian experimental design: a review*, Statist. Sci., 10 (1995), pp. 273–304.

[69] L. Chizat and F. Bach, *On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport*, ArXiv e-prints, (2018).

[70] C. Christof and G. Wachsmuth, *Differential Sensitivity Analysis of Variational Inequalities with Locally Lipschitz Continuous Solution Operators*, ArXiv e-prints, (2017).

[71] M. Chung and E. Haber, *Experimental design for biological systems*, SIAM J. Control Optim., 50 (2012), pp. 471–489.

[72] C. Clason, K. Ito, and K. Kunisch, *A minimum effort optimal control problem for elliptic PDEs*, ESAIM Math. Model. Numer. Anal., 46 (2012), pp. 911–927.

[73] C. Clason and K. Kunisch, *A duality-based approach to elliptic control problems in non-reflexive Banach spaces*, ESAIM Control Optim. Calc. Var., 17 (2011), pp. 243–266.

[74] ——, *A measure space approach to optimal source placement*, Comput. Optim. Appl., 53 (2012), pp. 155–171.

[75] P. L. Combettes and V. R. Wajs, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul., 4 (2005), pp. 1168–1200.

[76] R. D. Cook and C. J. Nachtsheim, *A comparison of algorithms for constructing exact d-optimal designs*, Technometrics, 22 (1980), pp. 315–324.

[77] G. Da Prato, *An introduction to infinite-dimensional analysis*, Universitext, Springer-Verlag, Berlin, 2006. Revised and extended from the 2001 original by Da Prato.

[78] G. Dal Maso, *An introduction to Γ-convergence*, vol. 8 of Progress in Nonlinear Differential Equations and their Applications, Birkhäuser Boston, Inc., Boston, MA, 1993.

[79] M. Dashti, K. J. H. Law, A. M. Stuart, and J. Voss, *MAP estimators and their consistency in Bayesian nonparametric inverse problems*, Inverse Problems, 29 (2013), pp. 095017, 27.

[80] M. Dashti and A. M. Stuart, *The Bayesian Approach to Inverse Problems*, Springer International Publishing, Cham, 2017, pp. 311–428.

[81] I. Daubechies, M. Defrise, and C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Comm. Pure Appl. Math., 57 (2004), pp. 1413–1457.

[82] K. Deckelnick and M. Hinze, *A note on the approximation of elliptic control problems with bang-bang controls*, Comput. Optim. Appl., 51 (2012), pp. 931–939.

[83] V. F. Demyanov and A. M. Rubinov, *Approximate methods in optimization problems*, Translated from the Russian by Scripta Technica, Inc. Translation edited by George M. Kranc. Modern Analytic and Computational Methods in Science and Mathematics, No. 32, American Elsevier Publishing Co., Inc., New York, 1970.

[84] H. Dette and W. J. Studden, *Geometry of E-optimality*, Ann. Statist., 21 (1993), pp. 416–433.

[85] E. Di Nezza, G. Palatucci, and E. Valdinoci, *Hitchhiker's guide to the fractional Sobolev spaces*, Bull. Sci. Math., 136 (2012), pp. 521–573.

[86] J. Dieudonné, *Foundations of modern analysis*, Academic Press, New York-London, 1969. Enlarged and corrected printing, Pure and Applied Mathematics, Vol. 10-I.

[87] N. Dinculeanu and J. J. Uhl, Jr., *A unifying Radon-Nikodym theorem for vector measures*, J. Multivariate Anal., 3 (1973), pp. 184–203.

[88] A. L. Dontchev and R. T. Rockafellar, *Parametrically robust optimality in nonlinear programming*, Appl. Comput. Math., 5 (2006), pp. 59–65.

[89] P. Drábek and J. Milota, *Methods of nonlinear analysis*, Birkhäuser Advanced Texts: Basler Lehrbücher. [Birkhäuser Advanced Texts: Basel Textbooks], Birkhäuser/Springer Basel AG, Basel, second ed., 2013. Applications to differential equations.

[90] N. Dunford and J. T. Schwartz, *Linear operators. Part I*, Wiley Classics Library, John Wiley & Sons, Inc., New York, 1988. General theory, With the assistance of William G. Bade and Robert G. Bartle, Reprint of the 1958 original, A Wiley-Interscience Publication.

[91] J. C. Dunn, *Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals*, SIAM J. Control Optim., 17 (1979), pp. 187–211.

[92] ——, *Convergence rates for conditional gradient sequences generated by implicit step length rules*, SIAM J. Control Optim., 18 (1980), pp. 473–487.

[93] J. C. Dunn and S. Harshbarger, *Conditional gradient algorithms with open loop step size rules*, J. Math. Anal. Appl., 62 (1978), pp. 432–444.

[94] V. Duval, *A characterization of the Non-Degenerate Source Condition in Super-Resolution*, ArXiv e-prints, (2017).

[95] V. Duval and G. Peyré, *Exact support recovery for sparse spikes deconvolution*, Found. Comput. Math., 15 (2015), pp. 1315–1355.

[96] R. E. Edwards, *Functional analysis. Theory and applications*, Holt, Rinehart and Winston, New York-Toronto-London, 1965.

[97] A. Eftekhari and A. Thompson, *A Bridge Between Past and Present: Exchange and Conditional Gradient Methods are Equivalent*, ArXiv e-prints, (2018).

[98] I. Ekeland and R. Témam, *Convex analysis and variational problems*, vol. 28 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, english ed., 1999. Translated from the French.

[99] N. Eldredge, *Analysis and Probability on Infinite-Dimensional Spaces*, ArXiv e-prints, (2016).

[100] J. Elstrodt, *Maß- und Integrationstheorie*, Springer-Lehrbuch. [Springer Textbook], Springer-Verlag, Berlin, fourth ed., 2005. Grundwissen Mathematik. [Basic Knowledge in Mathematics].

[101] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of inverse problems*, vol. 375 of Mathematics and its Applications, Kluwer Academic Publishers Group, Dordrecht, 1996.

[102] S. M. Ermakov and V. B. Melas, *A duality theorem and an iteration method for finding h-optimal experimental designs*, Vestnik Leningrad. Univ. Mat. Mekh. Astronom., (1982), pp. 38–43, 134.

[103] T. Etling and R. Herzog, *Optimum experimental design by shape optimization of specimens in linear elasticity*, SIAM J. Appl. Math., 78 (2018), pp. 1553–1576.

[104] L. C. Evans and R. F. Gariepy, *Measure theory and fine properties of functions*, Textbooks in Mathematics, CRC Press, Boca Raton, FL, revised ed., 2015.

[105] V. V. Fedorov, *Theory of optimal experiments*, Academic Press, New York-London, 1972. Translated from the Russian and edited by W. J. Studden and E. M. Klimko, Probability and Mathematical Statistics, No. 12.

[106] V. V. Fedorov and P. Hackl, *Model-oriented design of experiments*, vol. 125 of Lecture Notes in Statistics, Springer-Verlag, New York, 1997.

[107] V. V. Fedorov and S. L. Leonov, *Optimal design for nonlinear response models*, Chapman & Hall/CRC Biostatistics Series, CRC Press, Boca Raton, FL, 2014.

[108] A. V. Fiacco and Y. Ishizuka, *Sensitivity and stability analysis for nonlinear programming*, Ann. Oper. Res., 27 (1990), pp. 215–235.

[109] T. Fischer, *Existence, uniqueness, and minimality of the Jordan measure decomposition*, ArXiv e-prints, (2012).

[110] R. A. Fisher, *On an absolute criterion for fitting frequency curves*, Statistical Science, 12 (1997), pp. 39–41.

[111] B. G. Fitzpatrick, *Bayesian analysis in inverse problems*, Inverse Problems, 7 (1991), pp. 675–702.

[112] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95–110.

[113] G. Friesecke, F. Henneke, and K. Kunisch, *Sparse Control of Quantum Systems*, ArXiv e-prints, (2015), p. arXiv:1507.00768.

[114] D. Garber and E. Hazan, *Playing non-linear games with linear oracles*, in 2013 IEEE 54th Annual Symposium on Foundations of Computer Science—FOCS 2013, IEEE Computer Soc., Los Alamitos, CA, 2013, pp. 420–428.

[115] ——, *Faster rates for the frank-wolfe method over strongly-convex sets*, in Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, JMLR.org, 2015, pp. 541–549.

[116] I. Y. Gejadze and V. Shutyaev, *On computation of the design function gradient for the sensor-location problem in variational data assimilation*, SIAM J. Sci. Comput., 34 (2012), pp. B127–B147.

[117] A. L. Gibbs and F. E. Su, *On choosing and bounding probability metrics*, International Statistical Review / Revue Internationale de Statistique, 70 (2002), pp. 419–435.

[118] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, vol. 224, Springer Science & Business Media, 2001.

[119] I. C. Gohberg and M. G. Kreĭn, *Introduction to the theory of linear nonselfadjoint operators*, Translated from the Russian by A. Feinstein. Translations of Mathematical Monographs, Vol. 18, American Mathematical Society, Providence, R.I., 1969.

[120] D. Goldfarb and A. Idnani, *Dual and primal-dual methods for solving strictly convex quadratic programs*, in Numerical Analysis, J. P. Hennart, ed., Berlin, Heidelberg, 1982, Springer Berlin Heidelberg, pp. 226–239.

[121] A. A. Goldstein, *Convex programming in Hilbert space*, Bull. Amer. Math. Soc., 70 (1964), pp. 709–710.

[122] R. Griesse, M. Hintermüller, and M. Hinze, *Differential stability of control-constrained optimal control problems for the Navier-Stokes equations*, Numer. Funct. Anal. Optim., 26 (2005), pp. 829–850.

[123] R. Griesse and B. Vexler, *Numerical sensitivity analysis for the quantity of interest in PDE-constrained optimization*, SIAM J. Sci. Comput., 29 (2007), pp. 22–48.

[124] P. Grisvard, *Elliptic problems in nonsmooth domains*, vol. 69 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011. Reprint of the 1985 original [MR0775683], With a foreword by Susanne C. Brenner.

[125] J. Guélat and P. Marcotte, *Some comments on Wolfe's "away step"*, Math. Programming, 35 (1986), pp. 110–119.

[126] A. Günther and M. Hinze, *A posteriori error control of a state constrained elliptic control problem*, J. Numer. Math., 16 (2008), pp. 307–322.

[127] E. Haber, L. Horesh, and L. Tenorio, *Numerical methods for experimental design of large-scale linear ill-posed inverse problems*, Inverse Problems, 24 (2008), pp. 055012, 17.

[128] ——, *Numerical methods for the design of large-scale nonlinear discrete ill-posed inverse problems*, Inverse Problems, 26 (2010), pp. 025002, 14.

[129] J. Hadamard, *Sur les problèmes aux dérivés partielles et leur signification physique*, Princeton University Bulletin, 13 (1902), pp. 49–52.

[130] W. W. Hager, *Updating the inverse of a matrix*, SIAM Rev., 31 (1989), pp. 221–239.

[131] Z. Harchaoui, A. Juditsky, and A. Nemirovski, *Conditional gradient algorithms for norm-regularized smooth convex optimization*, Math. Program., 152 (2015), pp. 75–112.

[132] R. Harman and L. Pronzato, *Improvements on removing nonoptimal support points in D-optimum design algorithms*, Statist. Probab. Lett., 77 (2007), pp. 90–94.

[133] E. Hazan, *Sparse approximate solutions to semidefinite programs*, in LATIN 2008: Theoretical informatics, vol. 4957 of Lecture Notes in Comput. Sci., Springer, Berlin, 2008, pp. 306–316.

[134] J. Heinonen, *Lectures on Lipschitz analysis*, vol. 100 of Report. University of Jyväskylä Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, 2005.

[135] W. Hensgen, *A simple proof of Singer's representation theorem*, Proc. Amer. Math. Soc., 124 (1996), pp. 3211–3212.

[136] E. Herberg, M. Hinze, and H. Schumacher, *Maximal discrete sparsity in parabolic optimal control with measures*, ArXiv e-prints, (2018).

[137] R. Herzog and I. Riedel, *Sequentially optimal sensor placement in thermoelastic models for real time applications*, Optim. Eng., 16 (2015), pp. 737–766.

[138] R. Herzog, I. Riedel, and D. Uciński, *Optimal sensor placement for joint parameter and state estimation problems in large-scale dynamical systems with applications to thermomechanics*, Optim. Eng., 19 (2018), pp. 591–627.

[139] R. Herzog, G. Stadler, and G. Wachsmuth, *Directional sparsity in optimal control of partial differential equations*, SIAM J. Control Optim., 50 (2012), pp. 943–963.

[140] R. Hettich and K. O. Kortanek, *Semi-infinite programming: theory, methods, and applications*, SIAM Rev., 35 (1993), pp. 380–429.

[141] R. Hettich and P. Zencke, *Numerische Methoden der Approximation und semi-infiniten Optimierung*, B. G. Teubner, Stuttgart, 1982. Teubner Studienbücher Mathematik. [Teubner Mathematical Textbooks].

[142] R. P. Hettich and H. T. Jongen, *Semi-infinite programming: conditions of optimality and applications*, Optimization techniques (Proc. 8th IFIP Conf., Würzburg, 1977), Part 2, (1978), pp. 1–11. Lecture Notes in Control and Information Sci., Vol. 7.

[143] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2002), pp. 865–888 (2003).

[144] M. Hintermüller and K. Kunisch, *Feasible and noninterior path-following in constrained minimization with low multiplier regularity*, SIAM J. Control Optim., 45 (2006), pp. 1198–1221.

[145] M. Hintermüller and K. Kunisch, *Path-following methods for a class of constrained minimization problems in function space*, SIAM J. Optim., 17 (2006), pp. 159–187.

[146] M. Hintermüller, A. Schiela, and W. Wollner, *The length of the primal-dual path in Moreau-Yosida-based path-following methods for state constrained optimal control*, SIAM J. Optim., 24 (2014), pp. 108–126.

[147] M. Hintermüller and M. Ulbrich, *A mesh-independence result for semismooth Newton methods*, Math. Program., 101 (2004), pp. 151–184.

[148] M. Hinze, *A variational discretization concept in control constrained optimization: the linear-quadratic case*, Comput. Optim. Appl., 30 (2005), pp. 45–61.

[149] A. Hofinger and H. K. Pikkarainen, *Convergence rate for the Bayesian approach to linear inverse problems*, Inverse Problems, 23 (2007), pp. 2469–2484.

[150] ——, *Convergence rates for linear inverse problems in the presence of an additive normal noise*, Stoch. Anal. Appl., 27 (2009), pp. 240–257.

[151] C. A. Holloway, *An extension of the Frank and Wolfe method of feasible directions*, Math. Programming, 6 (1974), pp. 14–27.

[152] L. Horesh, E. Haber, and L. Tenorio, *Optimal experimental design for the large-scale nonlinear ill-posed problem of impedance imaging*, in Large-scale inverse problems and quantification of uncertainty, Wiley Ser. Comput. Stat., Wiley, Chichester, 2011, pp. 273–290.

[153] J. Idier, ed., *Bayesian approach to inverse problems*, Digital Signal and Image Processing Series, ISTE, London; John Wiley & Sons, Inc., Hoboken, NJ, 2008.

[154] J. M. Ingram and M. M. Marsh, *Projections onto convex cones in Hilbert space*, J. Approx. Theory, 64 (1991), pp. 343–350.

[155] V. Isakov, *Inverse problems for partial differential equations*, vol. 127 of Applied Mathematical Sciences, Springer, Cham, third ed., 2017.

[156] K. Ishihara, *Convergence of the finite element method applied to the eigenvalue problem $\Delta u + \lambda u = 0$*, Publ. Res. Inst. Math. Sci., 13 (1977/78), pp. 47–60.

[157] K. Ito and K. Kunisch, *Maximizing robustness in nonlinear ill-posed inverse problems*, SIAM J. Control Optim., 33 (1995), pp. 643–666.

[158] ——, *Semi-smooth Newton methods for state-constrained optimal control problems*, Systems Control Lett., 50 (2003), pp. 221–228.

[159] M. Jaggi, *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, 2013, pp. I–427–I–435.

[160] J. Kaipio and E. Somersalo, *Statistical and computational inverse problems*, vol. 160 of Applied Mathematical Sciences, Springer-Verlag, New York, 2005.

[161] L. V. Kantorovic and G. v. S. Rubinstein, *On a space of completely additive functions*, Vestnik Leningrad. Univ., 13 (1958), pp. 52–59.

[162] C. T. Kelley and E. W. Sachs, *Mesh independence of the gradient projection method for optimal control problems*, SIAM J. Control Optim., 30 (1992), pp. 477–493.

[163] J. Kiefer, *Optimum designs in regression problems, ii*, Ann. Math. Statist., 32 (1961), pp. 298–325.

[164] ——, *General equivalence theory for optimum designs (approximate theory)*, Ann. Statist., 2 (1974), pp. 849–879.

[165] J. Kiefer and J. Wolfowitz, *Optimum designs in regression problems*, Ann. Math. Statist., 30 (1959), pp. 271–294.

[166] ——, *The equivalence of two extremum problems*, Canad. J. Math., 12 (1960), pp. 363–366.

[167] D. Klatte, *Stable local minimizers in semi-infinite optimization: regularity and second-order conditions*, J. Comput. Appl. Math., 56 (1994), pp. 137–157. Stochastic programming: stability, numerical methods and applications (Gosen, 1992).

[168] D. Klatte and B. Kummer, *Nonsmooth equations in optimization*, vol. 60 of Nonconvex Optimization and its Applications, Kluwer Academic Publishers, Dordrecht, 2002. Regularity, calculus, methods and applications.

[169] S. Körkel, *Numerische Methoden für Optimale Versuchsplanungsprobleme bei nichtlinearen DAE-Modellen*, PhD Dissertation, Ruprecht-Karls-Universität Heidelberg, 2002. `https://archiv.ub.uni-heidelberg.de/volltextserver/6114/`.

[170] S. Körkel, I. Bauer, H. G. Bock, and J. Schlöder, *A sequential approach for nonlinear optimum experimental design in DAE systems*, Scientific Computing in Chemical Engineering II, 2 (1999), pp. 338–345.

[171] S. Körkel, E. Kostina, H. G. Bock, and J. P. Schlöder, *Numerical methods for optimal control problems in design of robust optimal experiments for nonlinear dynamic processes*, Optim. Methods Softw., 19 (2004), pp. 327–338. The First International Conference on Optimization Methods and Software. Part II.

[172] C. Kreutz and J. Timmer, *Systems biology: experimental design*, The FEBS journal, 276 (2009), pp. 923–942.

[173] F. Kruse and M. Ulbrich, *A self-concordant interior point approach for optimal control with state constraints*, SIAM J. Optim., 25 (2015), pp. 770–806.

[174] S. Kullback and R. A. Leibler, *On information and sufficiency*, Ann. Math. Statistics, 22 (1951), pp. 79–86.

[175] P. Kumar and E. A. Yildirim, *Minimum-volume enclosing ellipsoids and core sets*, J. Optim. Theory Appl., 126 (2005), pp. 1–21.

[176] K. Kunisch, K. Pieper, and B. Vexler, *Measure valued directional sparsity for parabolic optimal control problems*, SIAM J. Control Optim., 52 (2014), pp. 3078–3108.

[177] K. Kunisch and A. Rösch, *Primal-dual active set strategy for a general class of constrained optimal control problems*, SIAM J. Optim., 13 (2002), pp. 321–334.

[178] K. Kunisch, P. Trautmann, and B. Vexler, *Optimal control of the undamped linear wave equation with measure valued controls*, SIAM J. Control Optim., 54 (2016), pp. 1212–1244.

[179] S. Lacoste-Julien and M. Jaggi, *On the global linear convergence of frank-wolfe optimization variants*, in Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS'15, Cambridge, MA, USA, 2015, MIT Press, pp. 496–504.

[180] T. Lahmer, *Optimal experimental design for nonlinear ill-posed problems applied to gravity dams*, Inverse Problems, 27 (2011), p. 125005.

[181] T. Lahmer, B. Kaltenbacher, and V. Schulz, *Optimal measurement selection for piezoelectric material tensor identification*, Inverse Probl. Sci. Eng., 16 (2008), pp. 369–387.

[182] S. Lang, *Real analysis*, Addison-Wesley Publishing Company, Advanced Book Program, Reading, MA, second ed., 1983.

[183] ——, *Real and functional analysis*, vol. 142 of Graduate Texts in Mathematics, Springer-Verlag, New York, third ed., 1993.

[184] C. L. Lawson and R. J. Hanson, *Solving least squares problems*, vol. 15 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1995. Revised reprint of the 1974 original.

[185] E. Levitin and B. Polyak, *Constrained minimization methods*, USSR Computational Mathematics and Mathematical Physics, 6 (1966), pp. 1 – 50.

[186] D. Leykekhman, D. Meidner, and B. Vexler, *Optimal error estimates for finite element discretization of elliptic optimal control problems with finitely many pointwise state constraints*, Comput. Optim. Appl., 55 (2013), pp. 769–802.

[187] M. López and G. Still, *Semi-infinite programming*, European J. Oper. Res., 180 (2007), pp. 491–518.

[188] K. Malanowski, *Sensitivity analysis for parametric optimal control of semilinear parabolic equations*, J. Convex Anal., 9 (2002), pp. 543–561. Special issue on optimization (Montpellier, 2000).

[189] I. W. McKeague, G. Nicholls, K. Speer, and R. Herbei, *Statistical inversion of south atlantic circulation in an abyssal neutral density layer*, Journal of Marine Research, 63 (2005), pp. 683–704.

[190] P. Merino, I. Neitzel, and F. Tröltzsch, *Error estimates for the finite element discretization of semi-infinite elliptic optimal control problems*, Discuss. Math. Differ. Incl. Control Optim., 30 (2010), pp. 221–236.

[191] ——, *On linear-quadratic elliptic control problems of semi-infinite type*, Appl. Anal., 90 (2011), pp. 1047–1074.

[192] P. Merino, I. Neitzel, and F. Tröltzsch, *An adaptive numerical method for semi-infinite elliptic control problems based on error estimates*, Optim. Methods Softw., 30 (2015), pp. 492–515.

[193] C. Meyer, U. Prüfert, and F. Tröltzsch, *On two numerical methods for state-constrained elliptic control problems*, Optim. Methods Softw., 22 (2007), pp. 871–899.

[194] A. Milzarek and M. Ulbrich, *A semismooth Newton method with multidimensional filter globalization for l_1-optimization*, SIAM J. Optim., 24 (2014), pp. 298–333.

[195] H. Mine and M. Fukushima, *A minimization method for the sum of a convex function and a continuously differentiable function*, J. Optim. Theory Appl., 33 (1981), pp. 9–23.

[196] B. Mityagin, *The Zero Set of a Real Analytic Function*, ArXiv e-prints, (2015).

[197] I. Molchanov and S. Zuyev, *Steepest descent algorithms in a space of measures*, Stat. Comput., 12 (2002), pp. 115–123.

[198] ——, *Optimisation in space of measures and optimal design*, ESAIM Probab. Stat., 8 (2004), pp. 12–24.

[199] J.-J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.

[200] I. Neitzel, K. Pieper, B. Vexler, , and D. Walter, *A sparse control approach to optimal sensor placement in pde-constrained parameter estimation problems*, submitted.

[201] Y. Nesterov, *Introductory lectures on convex optimization*, vol. 87 of Applied Optimization, Kluwer Academic Publishers, Boston, MA, 2004. A basic course.

[202] Y. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program., 140 (2013), pp. 125–161.

[203] J. Nocedal and S. J. Wright, *Numerical optimization*, Springer Series in Operations Research and Financial Engineering, Springer, New York, second ed., 2006.

[204] M. Patriksson, *Simplicial decomposition algorithms*, Springer US, Boston, MA, 2009, pp. 3579–3585.

[205] A. Pázman, *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.

[206] G. C. Pflug and A. Pichler, *Multistage stochastic optimization*, Springer Series in Operations Research and Financial Engineering, Springer, Cham, 2014.

[207] R. R. Phelps, *Metric projections and the gradient projection method in Banach spaces*, SIAM J. Control Optim., 23 (1985), pp. 973–977.

[208] K. Pieper, *Finite element discretization and efficient numerical solution of elliptic and parabolic sparse control problems*, PhD Dissertation, Technische Universität München, 2015. `http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:bvb:91-diss-20150420-1241413-1-4`.

[209] K. Pieper, B. Q. Tang, P. Trautmann, and D. Walter, *Inverse point source location for the Helmholtz equation*, submitted.

[210] K. Pieper and B. Vexler, *A priori error analysis for discretization of sparse elliptic optimal control problems in measure space*, SIAM J. Control Optim., 51 (2013), pp. 2788–2808.

[211] K. Pieper and D. Walter, *Linear convergence of accelerated conditional gradient algorithms in spaces of measures*, in preparation.

[212] B. T. Polyak, *Introduction to optimization*, Translations Series in Mathematics and Engineering, Optimization Software, Inc., Publications Division, New York, 1987. Translated from the Russian, With a foreword by Dimitri P. Bertsekas.

[213] S. C. Power, *Another proof of Lidskiĭ's theorem on the trace*, Bull. London Math. Soc., 15 (1983), pp. 146–148.

[214] J. Preininger and P. T. Vuong, *On the convergence of the gradient projection method for convex optimal control problems with bang-bang solutions*, Comput. Optim. Appl., 70 (2018), pp. 221–238.

[215] C. Prévôt and M. Röckner, *A concise course on stochastic partial differential equations*, vol. 1905 of Lecture Notes in Mathematics, Springer, Berlin, 2007.

[216] L. Pronzato, *Removing non-optimal support points in D-optimum design algorithms*, Statist. Probab. Lett., 63 (2003), pp. 223–228.

[217] L. Pronzato and A. Pazman, *Design of experiments in nonlinear models: asymptotic normality, optimality criteria and small-sample properties*, Lecture notes in statistics, Springer, New York, NY, 2013.

[218] L. Pronzato and E. Walter, *Robust experiment design via maximin optimization*, Math. Biosci., 89 (1988), pp. 161–176.

[219] ——, *Minimum-volume ellipsoids containing compact sets: application to parameter bounding*, Automatica J. IFAC, 30 (1994), pp. 1731–1739.

[220] U. Prüfert, F. Tröltzsch, and M. Weiser, *The convergence of an interior point method for an elliptic control problem with mixed control-state constraints*, Comput. Optim. Appl., 39 (2008), pp. 183–218.

[221] F. Pukelsheim, *On linear regression designs which maximize information*, J. Statist. Plann. Inference, 4 (1980), pp. 339–364.

[222] ——, *Optimal design of experiments*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons, Inc., New York, 1993. A Wiley-Interscience Publication.

[223] L. Q. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–367.

[224] A. Rakotomamonjy, R. Flamary, and N. Courty, *Generalized conditional gradient: analysis of convergence and applications*, ArXiv e-prints, (2015).

[225] R. Rannacher and B. Vexler, *A priori error estimates for the finite element discretization of elliptic parameter identification problems with pointwise measurements*, SIAM J. Control Optim., 44 (2005), pp. 1844–1863.

[226] R. Rao Chivukula and I. R. Sarma, *On tensor products of spaces of continuous functions*, Studia Math., 75 (1983), pp. 335–339.

[227] W. Ring, *Structural properties of solutions to total variation regularization problems*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 799–810.

[228] C. P. Robert, *The Bayesian choice*, Springer Texts in Statistics, Springer, New York, second ed., 2007. From decision-theoretic foundations to computational implementation.

[229] R. T. Rockafellar, *Convex analysis*, Princeton Mathematical Series, No. 28, Princeton University Press, Princeton, N.J., 1970.

[230] W. Rudin, *Real and complex analysis*, McGraw-Hill Book Co., New York, third ed., 1987.

[231] A. K. Saibaba, A. Alexanderian, and I. C. F. Ipsen, *Randomized matrix-free trace and log-determinant estimators*, Numer. Math., 137 (2017), pp. 353–395.

[232] A. Schiela, *Barrier methods for optimal control problems with state constraints*, SIAM J. Optim., 20 (2009), pp. 1002–1031.

[233] A. Schindele and A. Borzì, *Proximal methods for elliptic optimal control problems with sparsity cost functional*, Applied Mathematics, 7 (2016), p. 967.

[234] A. Schindele and A. Borzì, *Proximal schemes for parabolic optimal control problems with sparsity promoting cost functionals*, Internat. J. Control, 90 (2017), pp. 2349–2367.

[235] B. Schweizer, *Darcy's law and groundwater flow modelling*, Snapshots of modern mathematics from Oberwolfach, (2015).

[236] S. Shalev-Shwartz, N. Srebro, and T. Zhang, *Trading accuracy for sparsity in optimization problems with sparsity constraints*, SIAM J. Optim., 20 (2010), pp. 2807–2832.

[237] A. Shapiro, *Second-order derivatives of extremal-value functions and optimality conditions for semi-infinite programs*, Math. Oper. Res., 10 (1985), pp. 207–219.

[238] ——, *On Lipschitzian stability of optimal solutions of parametrized semi-infinite programs*, Math. Oper. Res., 19 (1994), pp. 743–752.

[239] ——, *Sensitivity analysis of parametrized programs via generalized equations*, SIAM J. Control Optim., 32 (1994), pp. 553–571.

[240] ——, *Directional differentiability of the optimal value function in convex semi-infinite programming*, Math. Programming, 70 (1995), pp. 149–157.

[241] S. Silvey, D. Titterington, and B. Torsney, *An algorithm for optimal designs on a design space*, Communications in Statistics - Theory and Methods, 7 (1978), pp. 1379–1389.

[242] B. Simon, *Notes on infinite determinants of Hilbert space operators*, Advances in Math., 24 (1977), pp. 244–273.

[243] ——, *Trace ideals and their applications*, vol. 120 of Mathematical Surveys and Monographs, American Mathematical Society, Providence, RI, second ed., 2005.

[244] A. Sinkoe and J. Hahn, *Optimal experimental design for parameter estimation of an il-6 signaling model*, Processes, 5 (2017), p. 49.

[245] K. SMITH, *On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations*, Biometrika, 12 (1918), pp. 1–85.

[246] A. Spantini, T. Cui, K. Willcox, L. Tenorio, and Y. Marzouk, *Goal-oriented optimal approximations of Bayesian linear inverse problems*, SIAM J. Sci. Comput., 39 (2017), pp. S167–S196.

[247] R. C. St. John and N. R. Draper, *D-optimality for regression designs: a review*, Technometrics, 17 (1975), pp. 15–23.

[248] G. Stadler, *Elliptic optimal control problems with L^1-control cost and applications for the placement of control devices*, Comput. Optim. Appl., 44 (2009), pp. 159–181.

[249] M. Stone, *Application of a measure of information to the design and comparison of regression experiments*, Ann. Math. Statist., 30 (1959), pp. 55–70.

[250] A. M. Stuart, *Inverse problems: a Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.

[251] A. Tarantola, *Inverse problem theory and methods for model parameter estimation*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2005.

[252] A. N. Tikhonov and V. Y. Arsenin, *Solutions of ill-posed problems*, V. H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York-Toronto, Ont.-London, 1977. Translated from the Russian, Preface by translation editor Fritz John, Scripta Series in Mathematics.

[253] D. M. Titterington, *Optimal design: some geometrical aspects of D-optimality*, Biometrika, 62 (1975), pp. 313–320.

[254] P. Trautmann, B. Vexler, and A. Zlotnik, *Finite element error analysis for measure-valued optimal control problems governed by a 1D wave equation with variable coefficients*, Math. Control Relat. Fields, 8 (2018), pp. 411–449.

[255] F. Tröltzsch, *Optimal control of partial differential equations*, vol. 112 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, 2010. Theory, methods and applications, Translated from the 2005 German original by Jürgen Sprekels.

[256] D. Uciński, *Optimal measurement methods for distributed parameter system identification*, Systems and Control Series, CRC Press, Boca Raton, FL, 2005.

[257] M. Ulbrich, *Semismooth Newton methods for operator equations in function spaces*, SIAM J. Optim., 13 (2002), pp. 805–842 (2003).

[258] B. Vexler, *Adaptive finite element methods for parameter identification problems*, PhD Dissertation, Ruprecht-Karls-Universität Heidelberg, 2004. `http://www.ub.uni-heidelberg.de/archiv/4603`.

[259] C. Villani, *Optimal transport*, vol. 338 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Springer-Verlag, Berlin, 2009. Old and new.

[260] B. von Hohenbalken, *Simplicial decomposition in nonlinear programming algorithms*, Math. Programming, 13 (1977), pp. 49–68.

[261] G. Wachsmuth and D. Wachsmuth, *Convergence and regularization results for optimal control problems with sparsity functional*, ESAIM Control Optim. Calc. Var., 17 (2011), pp. 858–886.

[262] D. Walter, *Gradient based optimization algorithms for sensor placement problems with sparsity*, Master Thesis, Technische Universität München, 2016.

[263] J. Warga, *Optimal control of differential and functional equations*, Academic Press, New York-London, 1972.

[264] D. Werner, *Funktionalanalysis*, Springer-Verlag, Berlin, extended ed., 2000.

[265] C. K. Wikle, R. F. Milliff, R. Herbei, and W. B. Leeds, *Modern statistical methods in oceanography: A hierarchical perspective*, Statistical Science, (2013), pp. 466–486.

[266] S. Willard, *General topology*, Dover Publications, Inc., Mineola, NY, 2004. Reprint of the 1970 original [Addison-Wesley, Reading, MA; MR0264581].

[267] J. Wloka, *Partial differential equations*, Cambridge University Press, Cambridge, 1987. Translated from the German by C. B. Thomas and M. J. Thomas.

[268] P. Wolfe, *Convergence theory in nonlinear programming*, North-Holland, Amsterdam, 1970.

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

[270] ——, *Some iterative procedures for generating nonsingular optimal designs*, Communications in Statistics-Theory and Methods, 7 (1978), pp. 1399–1412.

[271] C.-F. Wu and H. P. Wynn, *The convergence of general step-length algorithms for regular optimum design criteria*, Ann. Statist., 6 (1978), pp. 1273–1285.

[272] H. P. Wynn, *The sequential generation of D-optimum experimental designs*, Ann. Math. Statist., 41 (1970), pp. 1655–1664.

[273] ——, *Results in the theory and construction of D-optimum experimental designs*, J. Roy. Statist. Soc. Ser. B, 34 (1972), pp. 133–147, 170–186.

[274] H.-K. Xu, *Convergence Analysis of the Frank-Wolfe Algorithm and Its Generalization in Banach Spaces*, ArXiv e-prints, (2017).

[275] M. Yang, S. Biedermann, and E. Tang, *On optimal designs for nonlinear models: a general and efficient algorithm*, J. Amer. Statist. Assoc., 108 (2013), pp. 1411–1420.

[276] Y. Yu, *D-optimal designs via a cocktail algorithm*, Stat. Comput., 21 (2011), pp. 475–481.

[277] Y. Yu, X. Zhang, and D. Schuurmans, *Generalized conditional gradient for sparse estimation*, Journal of Machine Learning Research, 18 (2017), pp. 1–46.

[278] L. Zhang, S.-Y. Wu, and M. A. López, *A new exchange method for convex semi-infinite programming*, SIAM J. Optim., 20 (2010), pp. 2959–2977.

[279] O. C. Zienkiewicz and J. Z. Zhu, *The superconvergent patch recovery and a posteriori error estimates. I. The recovery technique*, Internat. J. Numer. Methods Engrg., 33 (1992), pp. 1331–1364.