FULL LENGTH PAPER

Series A

Asymptotic linear convergence of fully-corrective generalized conditional gradient methods

Kristian Bredies[1] · Marcello Carioni[2] · Silvio Fanzon[3] · Daniel Walter[4]

Abstract

We propose *a fully-corrective generalized conditional gradient method* (FC-GCG) for the minimization of the sum of a smooth, convex loss function and a convex one-homogeneous regularizer over a Banach space. The algorithm relies on the mutual update of a finite set $\mathcal{A}_k$ of extremal points of the unit ball of the regularizer and of an iterate $u_k \in \text{cone}(\mathcal{A}_k)$. Each iteration requires the solution of one linear problem to update $\mathcal{A}_k$ and of one finite dimensional convex minimization problem to update the iterate. Under standard hypotheses on the minimization problem we show that the algorithm converges sublinearly to a solution. Subsequently, imposing additional assumptions on the associated dual variables, this is improved to a linear rate of convergence. The proof of both results relies on two key observations: First, we prove the equivalence of the considered problem to the minimization of a lifted functional over a particular space of Radon measures using Choquet's theorem. Second, the FC-GCG algorithm is connected to a *Primal-Dual-Active-Point method* on the lifted problem for which we finally derive the desired convergence rates.

✉ Daniel Walter
   daniel.walter@hu-berlin.de

   Kristian Bredies
   kristian.bredies@uni-graz.at

   Marcello Carioni
   m.c.carioni@utwente.nl

   Silvio Fanzon
   S.Fanzon@hull.ac.uk

1 Department of Mathematics and Scientific Computing, University of Graz, Heinrichstraße 36, 8010 Graz, Austria

2 Department of Applied Mathematics, University of Twente, 7500 AE Enschede, The Netherlands

3 Department of Mathematics, University of Hull, Cottingham Road, Hull HU6 7RX, United Kingdom

4 Department of Mathematics, Humboldt-Universität zu Berlin, Berlin, Germany

🖄 Springer

1 Introduction

This paper is concerned with the analysis of an efficient solution algorithm for minimization problems in composite form

$$\inf_{u \in \mathcal{M}} J(u), \quad J(u) := F(Ku) + \mathcal{G}(u) \qquad (\mathcal{P}_{\mathcal{M}})$$

over a Banach space $\mathcal{M}$. Here, the forward operator K maps continuously from $\mathcal{M}$ into a Hilbert space Y of observations, not necessarily finite dimensional, and F denotes a smooth convex loss function. The second part of the objective functional is constituted by a convex but possibly non-smooth functional $\mathcal{G}$ which promotes desired structural properties. We refer, e.g., to the sparsifying property of total variation regularization or the staircasing effect of bounded variation penalties. The observation that certain structural features of minimizers can be brought forth by a suitable choice of the functional $\mathcal{G}$ and of the space $\mathcal{M}$, has made the analysis of problems of the form ($\mathcal{P}_{\mathcal{M}}$) a flourishing topic in the context of optimal control, inverse problems, compressed sensing and machine learning. As a consequence, the interest for such models has also sparked the demand for efficient solution algorithms. Since many of the structural features of ($\mathcal{P}_{\mathcal{M}}$) are tightly linked to properties of the underlying, possibly infinite dimensional Banach space, a particular focus in this context lies on *function space methods*, i.e., algorithms solving ($\mathcal{P}_{\mathcal{M}}$) without discretizing $\mathcal{M}$. This is a challenging task for a variety of reasons: On the one hand, it requires algorithms that can handle the non-smoothness of the objective functional J. On the other hand, it forces to work directly on the Banach space $\mathcal{M}$ which usually lacks "nice" properties such as reflexivity or uniform convexity. Efficient algorithms have been developed, for instance, for inverse problems regularized with ℓ^p penalties [18, 31], inverse problems in the space of measures regularized with the total variation [11, 21, 33] and dynamic inverse problems with optimal transport regularizers [14–17].

1.1 Contribution and related work

In many interesting applications it is meaningful to assume that $\mathcal{M}$ is given as the topological dual of a separable Banach space $\mathcal{C}$, and K is the adjoint of a "predual" operator $K_* : Y \to \mathcal{C}$, $(K_*)^* = K$. We refer to Sect. 4 for a few examples. In this case a simple approach to computing a minimizer to ($\mathcal{P}_{\mathcal{M}}$) is constituted by *generalized conditional gradient* (GCG) algorithms. Assuming that the solution set of ($\mathcal{P}_{\mathcal{M}}$) is bounded by some $M > 0$, this method updates the iterate u_k by computing the dual variable $p_k = -K^* \nabla F(Ku_k)$ and setting

$$v_k \in \operatorname*{arg\,min}_{\|v\|_{\mathcal{M}} \leq M} [-\langle p_k, v \rangle + \mathcal{G}(v)], \quad u_{k+1} = u_k + s_k(v_k - u_k),$$

where $s_k \in (0, 1)$ is chosen according to some stepsize rule. If $\mathcal{G}(u) = I_A(u)$ is the indicator function of a compact convex set $A \subset \mathcal{M}$, then the iteration scheme reduces to the *Frank-Wolfe* (FW) algorithm for constrained minimization [32, 37, 40].

In the present paper we focus on convex positively homogeneous functionals $\mathcal{G}$ with compact sublevel sets, where compactness is intended with respect to the weak* topology induced by $\mathcal{C}$. The compactness assumption is equivalent to $\mathcal{G}$ having weak* closed and norm bounded sublevels. In particular, these assumptions on $\mathcal{G}$ encompass the important case of norm regularization $\mathcal{G}(u) = \|u\|_{\mathcal{M}}$, and also allow for seminorm penalties if the problem is posed in a suitable quotient space [13]. More in general, our setting covers the case of all *gauge functions* of the form

$$\kappa_{\mathcal{A}}(u) := \inf \{\rho \geq 0 \mid u \in \rho A\} \quad \text{where} \quad \rho A = \{\rho v \mid v \in A\},$$

and $\mathcal{A} \subset \mathcal{M}$ is a weak* compact and convex set. This type of functional is of particular relevance in machine learning applications [28, 62].

In this general setting, GCG methods benefit from several desirable properties. For example, they only rely on the repeated solution of (partially) linearized problems that, in some interesting cases, can be solved analytically. Additionally, the descent direction can be chosen to satisfy $v_k = M_k \widehat{v}_k$ where $M_k \geq 0$ is a scaling factor and $\widehat{v}_k$ is an *extremal point* of the "unit ball" $B = \{u \mid \mathcal{G}(u) \leq 1\}$ of the regularizer. We denote by $\text{Ext}(B)$ the set of such points. Thus, the above mentioned partially linearized problems can be solved, equivalently, in the set $\text{Ext}(B)$, reducing considerably the complexity of each iteration of the algorithm. As a further consequence, initializing the algorithm by $u_0 = 0$ and selecting descent directions v_k that are extremal points of B, the iterate u_k exhibits *k-sparsity*, i.e., it is contained in the conic hull of at most k points in $\text{Ext}(B)$. The connection between the structure enhancing properties of a functional $\mathcal{G}$ and the set of extremal points of B has been recently studied in [13] and [12]. In this context, a central result is given by *convex representer theorems*. Loosely speaking, these state that problems of the form ($\mathcal{P}_{\mathcal{M}}$) admit solutions $\bar{u}$ which are contained in the conic hull of at most $\dim Y$ extremal points.

While the FW method [35, 36, 41, 47], and its many variants such as *away-step* FW [43, 49], or *fully-corrective* FW [44, 58, 59], have received a lot of attention, GCG algorithms for general non-smooth functionals $\mathcal{G}$ are less frequently studied. We refer, e.g., to [16, 20, 57, 62] as well as [60, Chapter 6] which all prove a global sublinear $\mathcal{O}(1/k)$ rate of convergence of $J(u_k)$ towards the minimum value. Note that this rate is known to be optimal [24]. Moreover, the absence of steps that remove extremal points from u_k often leads to clustering phenomena in practice. Thus, despite its various advantages, these shortcomings of GCG methods limit their practical utility.

The main contribution of the present work is the analysis of a *fully-corrective generalized conditional gradient method* (FC-GCG) for ($\mathcal{P}_{\mathcal{M}}$). This relies on the mutual update of a sequence of sparse iterates u_k and of a sequence of finite, ordered *active sets*

$$\mathcal{A}_k^u = \left\{ u_i^k \in \text{Ext}(B) \mid i = 1, \ldots, N_k \right\}, \quad N_k \in \mathbb{N},$$

of extremal points, which constitute the *atoms* representing the sparse iterate u_k. Given the current iterate u_k and active set $\mathcal{A}_k$, the proposed method first enlarges $\mathcal{A}_k$ by setting $N_k^+ = N_k + 1$ and

$$\mathcal{A}_k^{u,+} = \mathcal{A}_k^u \cup \{\widehat{v}_k\}, \quad \widehat{v}_k \in \arg\max_{v \in \text{Ext}(B)} \langle p_k, v \rangle. \tag{1.1}$$

Subsequently, the new iterate u_{k+1} is found by solving the subproblem

$$\min_{u \in \text{cone}(\mathcal{A}_k^{u,+})} \left[F(Ku) + \kappa_{\mathcal{A}_k^{u,+}}(u) \right], \tag{1.2}$$

where the minimization occurs over the cone spanned by $\mathcal{A}_k^{u,+}$, and $\mathcal{G}$ is replaced by the gauge function associated with $\mathcal{A}_k^{u,+}$, which in this case [10, 28] simplifies to

$$\kappa_{\mathcal{A}_k^{u,+}}(u) = \min_{\lambda \in \mathbb{R}_+^{N_k^+}} \left\{ \sum_{i=1}^{N_k^+} \lambda_i \;\middle|\; u = \lambda_{N_k^+} \widehat{v}_k + \sum_{i=1}^{N_k} \lambda_i u_i^k \right\}.$$

Finally, $\mathcal{A}_k^{u,+}$ is pruned by removing the extremal points for which the weight is set to zero by (1.2), obtaining the next active set $\mathcal{A}_{k+1}^u$.

As we will see, the proposed method combines the advantages of GCG methods, e.g., its global convergence, with an improved convergence behavior and sparser iterates. More in detail, we first prove a global sublinear convergence rate $\mathcal{O}(1/k)$ for $J(u_k)$ towards the minimum of $(\mathcal{P}_{\mathcal{M}})$, see Theorem 3.3. This is achieved under mild assumptions, see (A1)–(A3) discussed in Sect. 2. These are quite standard and are, in particular, sufficient for the well-posedness of $(\mathcal{P}_{\mathcal{M}})$, as shown in Proposition 2.3. Subsequently, in Theorem 3.8, we show that $J(u_k)$ converges asymptotically at a linear rate of $\mathcal{O}(\zeta^k)$, $\zeta \in (0, 1)$, provided that the optimal dual variable and the loss in $(\mathcal{P}_{\mathcal{M}})$ meet certain structural requirements, see Assumptions (B1)–(B5) in Sect. 3.3. Specifically, these assumptions imply that the minimizer to $(\mathcal{P}_{\mathcal{M}})$ is unique and sparse, i.e., of the form $\bar{u} = \sum_{i=1}^N \bar{\lambda}_i \bar{u}_i$ for a finite number of extremal points $\bar{u}_i \in \text{Ext}(B)$ and coefficients $\bar{\lambda}_i > 0$. More crucially, we also require the existence of a "distance function" g defined on $\text{Ext}(B)$ such that the forward operator K and the optimal dual variable $\bar{p} = -K_* \nabla F(K\bar{u})$ associated to $\bar{u}$ satisfy, respectively, the following Lipschitz and quadratic growth conditions

$$\|K(u - \bar{u}_i)\|_Y \lesssim g(u, \bar{u}_i), \quad 1 - \langle \bar{p}, u \rangle \gtrsim g(u, \bar{u}_i)^2, \tag{1.3}$$

for all $i = 1, \ldots, N$, and extremal points u in the weak* vicinity of $\bar{u}_i$, respectively.

The proofs of the convergence results of Theorems 3.3, 3.8 rely on a novel lifting strategy, discussed in Sect. 5, which allows us to recast $(\mathcal{P}_{\mathcal{M}})$ into an equivalent minimization problem on $M^+(\mathcal{B})$, the space of positive Radon measures on the set $\mathcal{B} := \overline{\text{Ext}(B)}^*$. Specifically, this is achieved by making three key observations: First,

using Choquet's Theorem [51, Page 14], we argue the existence of a surjective mapping $\mathcal{I}$ from $M^+(\mathcal{B})$ to the domain of $\mathcal{G}$ with

$$\langle p, \mathcal{I}(\mu) \rangle = \int_{\mathcal{B}} \langle p, v \rangle \, \mathrm{d}\mu(v), \quad \text{for all } p \in \mathcal{C}, \ \mu \in M^+(\mathcal{B}).$$

Second, we consider the auxiliary problem

$$\inf_{\mu \in M^+(\mathcal{B})} \left[F(\mathcal{K}\mu) + \|\mu\|_{M(\mathcal{B})} \right], \tag{$\mathcal{P}_{M+}$}$$

where the forward operator K is replaced by a lifted model $\mathcal{K} \colon M^+(\mathcal{B}) \to Y$ satisfying $\mathcal{K}\mu = K\mathcal{I}(\mu)$ for all $\mu \in M^+(\mathcal{B})$. As it turns out, see Theorem 5.4, the above minimization problems are equivalent, in the sense that $\bar{u} = \mathcal{I}(\bar{\mu})$ solves ($\mathcal{P}_{\mathcal{M}}$) if and only if $\bar{\mu}$ minimizes in ($\mathcal{P}_{M+}$). As a third key ingredient, such equivalence allows to interpret the iteration scheme introduced in (1.1)–(1.2) as one step of an *exchange algorithm* or a *Primal-Dual-Active-Point method* (PDAP), cf. Algorithm 2, applied to ($\mathcal{P}_{M+}$). These methods are fully-corrective variants of the generalized conditional gradient method for minimization problems over spaces of Borel measures introduced in [21], and their linear convergence has been recently proven in the finite-dimensional case, i.e., for measures supported on a compact subset of the euclidean space, see [39, 52]. Combining these three key observations, and carefully extending the techniques and results of [52] to the present setting, we are able to conclude the asymptotic linear convergence of our algorithm.

If we interpret ($\mathcal{P}_{\mathcal{M}}$) as a *sparse dictionary learning problem*, in which the dictionary is given by $\mathrm{Ext}(B)$, our FC-GCG method can also be linked to a *fully-corrective greedy selection* method, [56], or an *accelerated gradient boosting* algorithm, [61]. In this context, a similar lifting approach was implicitly used in [63] to derive a fully-corrective GCG method for matrix-regularization problems which eventually converges at a global linear rate, albeit under very restrictive conditions. More in detail, expressed in the notation of the present paper, the authors require the strong convexity of $F \circ K$, as well as that $\mathrm{Ext}(B)$ is a finite set. For a related result in settings with finitely many atoms we also point out [56]. In contrast, the present manuscript is geared towards potentially compact forward operators K, i.e. $F \circ K$ is not strongly convex, as well as regularizers whose sublevel sets admit infinitely many extremal points. These additional difficulties will be circumvented by a careful localization of $\mathcal{A}_k^{u,+}$ around the optimal extremal points, as well as by exploiting the assumed quadratic growth behavior in (1.3). While the main contribution of the present work is clearly given by the FC-GCG method, the lifting strategy of Sect. 5 may be of independent interest, and could also be applicable beyond the analysis of efficient solution algorithms. Finally, [62] presents a conditional gradient method for problems with gauge-like regularizers which relies on a "lifting" of the problem by a simultaneous update of an upper bound ρ_k on the value of the regularizer, i.e. $\mathcal{G}(u_k) \leq \rho_k$. This avoids evaluations of $\mathcal{G}$ throughout the iterations and yields a sublinear rate of convergence similar to Theorem 3.3.

In this paper, we mainly focus on the theoretical aspects of the FC-GCG method and, in particular, its convergence. Naturally, from the practical standpoint, the applicability of our method relies on the knowledge of the set $\mathrm{Ext}(B)$, as well as on the efficient solution of the constrained linear problem in (1.1). Especially the computational cost of the latter greatly varies between different instances of Problem ($\mathcal{P_M}$). To illustrate the working of our method and the role of assumptions (B1)–(B5) we present in Sect. 4 several examples of applications where our algorithm can be implemented. For the first three examples we provide a natural and easy to verify set of assumptions that imply (B1)–(B5), and for all of them we discuss the computational burden of computing a solution for the constrained linear problem in (1.1). We first consider the problem of identifying the initial source of a heat equation from given temperature measurements, see Sect. 4.1. For this example we also demonstrate numerically the expected linear convergence of Algorithm 1, discussing the stopping criterion and the advantages of our algorithm compared to classical GCG methods. Then, in Sect. 4.2, we consider the trace regularization of linear operators from a Hilbert space into itself, which favours the reconstruction of rank-one operators. In Sect. 4.3, we discuss minimum effort problems, where the regularization enforced by the supremum norm favours binary solutions. Finally, in Sect. 4.4, we briefly deal with the optimal transport regularization of dynamic inverse problems [17], showing that the algorithm introduced in [16] is, in some instances, a particular case of Algorithm 1. For this example the verification of the hypotheses (B1)–(B5), necessary to ensure fast convergence, is non-trivial and is left to future work.

1.2 Outline

The paper is structured as follows. In Sect. 2 we formulate the minimization problem ($\mathcal{P_M}$) we are interested to solve. We further make a basic set of Assumptions (A1)–(A3), under which we prove its well-posedness, see Proposition 2.3. In Sect. 3 we introduce the FC-GCG algorithm, discussing well-posedness of its steps and providing a stopping criterion. In addition we introduce sufficient conditions for fast convergence (B1)–(B5), and state the sublinear global and linear local convergence results in Theorems 3.3, 3.8 respectively. These results, together with the assumptions, are discussed for specific examples in Sect. 4. We then pass to the convergence proofs for the FC-GCG algorithm. These are broken into three parts: First, in Sect. 5, we introduce a "lifting" of the minimization problem ($\mathcal{P_M}$) to the space of Radon measures $M^+(\mathcal{B})$ by virtue of Choquet's theorem. We then prove the equivalence of ($\mathcal{P_M}$) to the auxiliary problem ($\mathcal{P}_{M+}$), see Theorem 5.4. Second, we propose an extension of the PDAP algorithm to compute solutions to ($\mathcal{P}_{M+}$) in Sect. 5.5. It turns out that PDAP and FC-GCG are equivalent given the correct interpretation, see Theorem 5.10. This equivalence will be extensively used in Sect. 6 to carry out the proofs of the main convergence statements. Finally, the appendix contains some auxiliary results, as well as of proofs omitted in the main body of the paper.

2 The minimization problem

In this section we introduce the minimization problem we are concerned with solving, and prove its well-posedness under suitable assumptions. We start by establishing some notations. Throughout the paper $\mathcal{C}$ denotes a separable Banach space with norm $\|\cdot\|_{\mathcal{C}}$ and topological dual space $\mathcal{M} \simeq \mathcal{C}^*$. We denote the duality pairing between $p \in \mathcal{C}$ and $u \in \mathcal{M}$ by $\langle p, u \rangle$. The space $\mathcal{M}$ is equipped with the canonical dual norm

$$\|u\|_{\mathcal{M}} := \sup_{\|p\|_{\mathcal{C}} \leq 1} \langle p, u \rangle \quad \text{for all} \ \ u \in \mathcal{M}.$$

Let $\mathcal{G} \colon \mathcal{M} \to [0, \infty]$ be a convex, weak* lower semi-continuous and positively one-homogeneous functional, that is $\mathcal{G}(\lambda u) = \lambda \mathcal{G}(u)$ for all $\lambda \geq 0$. Let $K \colon \mathcal{M} \to Y$ be a linear weak*-to-weak continuous operator, mapping to a given Hilbert space Y, and $F \colon Y \to \mathbb{R}$ be a convex mapping. The inner product and induced norm on Y will be denoted by $(\cdot, \cdot)_Y$ and $\|\cdot\|_Y$, respectively. Our interest lies in efficient solution algorithms for problems of the form $(\mathcal{P}_{\mathcal{M}})$, which we remind the reader is of the form

$$\inf_{u \in \mathcal{M}} J(u), \quad J(u) := F(Ku) + \mathcal{G}(u).$$

Remark 2.1 Notice that the weak*-to-weak continuity of the linear operator $K \colon \mathcal{M} \to Y$ implies existence of a linear continuous operator $K_* \colon Y \to \mathcal{C}$ that is the pre-adjoint of K, i.e.,

$$\langle K_* y, u \rangle = (Ku, y)_Y \quad \text{for all} \ \ y \in Y, \ u \in \mathcal{M}.$$

See, for example, [21, Remark 3.2]. Moreover, the existence of a continuous pre-adjoint K_* implies the strong-to-strong continuity of the operator K.

Throughout the paper we moreover require the following basic assumptions.

Assumption 2.2 Assume that:

(A1) F is bounded from below, strictly convex and Fréchet differentiable on Y. Its gradient $\nabla F \colon Y \to Y$ is Lipschitz continuous on compact sets.
(A2) The sublevel set

$$S^-(\mathcal{G}, \alpha) := \{ u \in \mathcal{M} \mid \mathcal{G}(u) \leq \alpha \}$$

is weak* compact for every $\alpha \geq 0$.
(A3) The operator $K \colon \mathcal{M} \to Y$ is sequentially weak*-to-strong continuous in

$$\mathrm{dom}(\mathcal{G}) := \{ u \in \mathcal{M} \mid \mathcal{G}(u) < \infty \},$$

namely, for every sequence $\{u_k\}_k$ in $\mathrm{dom}(\mathcal{G})$ such that $u_k \overset{*}{\rightharpoonup} u$ for some $u \in \mathcal{M}$, it holds that $Ku_k \to Ku$ in Y.

Note that Assumption (A2) is equivalent to ask that the sublevel set $S^-(\mathcal{G}, \alpha)$ is closed and norm bounded for every $\alpha \geq 0$. The above conditions guarantee the existence of minimizers to $(\mathcal{P}_{\mathcal{M}})$.

Proposition 2.3 *Assume (A1)–(A3). Then there exists at least one minimizer to $(\mathcal{P}_{\mathcal{M}})$. Moreover $\bar{u} \in \mathcal{M}$ is a solution to $(\mathcal{P}_{\mathcal{M}})$ if and only if $\bar{p} = -K_* \nabla F(K\bar{u}) \in \mathcal{C}$ satisfies*

$$\langle \bar{p}, \bar{u} \rangle = \mathcal{G}(\bar{u}), \qquad \max_{v \in S^-(\mathcal{G}, 1)} \langle \bar{p}, v \rangle \leq 1. \tag{2.1}$$

Finally, if $\bar{u}_1, \bar{u}_2 \in \mathcal{M}$ are two solutions to $(\mathcal{P}_{\mathcal{M}})$, then $K\bar{u}_1 = K\bar{u}_2$.

Proof The existence of a minimizer follows by the direct method of calculus of variations. Indeed, the sublevel sets of J are weak* compact thanks to Assumption (A1) and Assumption (A2). Moreover J is weak* lower semicontinuous, since $\mathcal{G}$ is weak* lower semicontinuous, K is weak*-to-weak continuous and F is convex and continuous (Assumption (A1)). In order to show (2.1), notice that the function $f(u) = F(Ku)$ is Gâteaux-differentiable with $f'(u)(v) = \langle K_* \nabla F(Ku), v \rangle$ for all $v \in \mathcal{M}$. Hence, see e.g. [60, Proposition 6.3], $\bar{u} \in \mathcal{M}$ is a solution to $(\mathcal{P}_{\mathcal{M}})$ if and only if

$$\langle \bar{p}, u - \bar{u} \rangle + \mathcal{G}(\bar{u}) \leq \mathcal{G}(u) \quad \text{for all } u \in \mathcal{M}.$$

This variational inequality holds if and only if

$$\langle \bar{p}, \bar{u} \rangle = \mathcal{G}(\bar{u}), \quad \langle \bar{p}, v \rangle \leq \mathcal{G}(v) \quad \text{for all } v \in \mathcal{M},$$

which is equivalent to (2.1), thanks to the one-homogeneity of $\mathcal{G}$. Last, the identity $K\bar{u}_1 = K\bar{u}_2$ for two solutions $\bar{u}_1, \bar{u}_2$ of $(\mathcal{P}_{\mathcal{M}})$ follows from the strict convexity of F. □

For the remainder of the paper we refer to

$$\bar{y} := K\bar{u} \in Y, \quad \bar{p} := -K_* \nabla F(K\bar{u}) \in \mathcal{C}, \tag{2.2}$$

as the (unique) optimal observation and dual variable for $(\mathcal{P}_{\mathcal{M}})$, respectively, where $\bar{u}$ is any minimizer to $(\mathcal{P}_{\mathcal{M}})$. Set $B = S^-(\mathcal{G}, 1)$. In the following we further require the notion of an *extremal point* of B.

Definition 2.4 An element $u \in B$ is called extremal point of B if there are *no* $u_1, u_2 \in B$, $u_1 \neq u_2$ and $s \in (0, 1)$ with $u = (1 - s)u_1 + su_2$. The set of all extremal points of B is denoted by $\text{Ext}(B)$.

Since B is weak* compact, non empty and convex thanks to Assumptions (A1)–(A3), by the Krein-Milman Theorem we infer that $\text{Ext}(B) \neq \emptyset$ and for every $u \in \text{dom}(\mathcal{G})$ there is $\{u_k\}_k$ in $\text{cone}(\text{Ext}(B))$ with $u_k \overset{*}{\rightharpoonup} u$. Set $\mathcal{B} := \overline{\text{Ext}(B)}^*$. Since the predual space $\mathcal{C}$ is separable and $\mathcal{B}$ is weak* compact, there exists a metric $d_{\mathcal{B}}$ metrizing the weak* convergence on $\mathcal{B}$, that is, for all sequences $\{u_k\}_k$ in $\mathcal{B}$ and $u \in \mathcal{B}$ we have

$$u_k \overset{*}{\rightharpoonup} u \quad \text{if and only if} \quad \lim_{k \to \infty} d_{\mathcal{B}}(u_k, u) = 0. \tag{2.3}$$

In particular, we have that $(\mathcal{B}, d_{\mathcal{B}})$ is a compact separable metric space.

Remark 2.5 Let us mention that all the results in this paper still hold under slightly weaker conditions on the functional F. In particular, the strict convexity of F can be replaced by assuming $Ku_1 = Ku_2$ for all solutions u_1, u_2 of $((\mathcal{P}_{\mathcal{M}}))$. Moreover, all results also apply to convex, weakly lower semicontinuous functionals $F \colon Y \to \mathbb{R} \cup \{+\infty\}$ for which $\mathrm{dom}(F)$ is open. In this case, F is required to be smooth on $\mathrm{dom}(F)$ and $\nabla F \colon \mathrm{dom}(F) \to Y$ is assumed to be Lipschitz continuous on compact subsets of $\mathrm{dom}(F)$. For more details, we refer the interested reader to [52, 60].

3 A numerical minimization algorithm

This section concerns the development of an implementable and efficient solution algorithm for the minimization problem $(\mathcal{P}_{\mathcal{M}})$, namely, the *fully-corrective generalized conditional gradient method* (FC-GCG). As anticipated in the introduction, FC-GCG comprises the two basic steps at (1.1) and (1.2), which we now describe in more detail. For this purpose, recall that every $u \in \mathrm{dom}(\mathcal{G})$ can be approximated, in the weak* topology of $\mathcal{M}$, by a sequence in $\mathrm{cone}(\mathrm{Ext}(B))$ thanks to Krein-Milman's Theorem. The considered method exploits this observation by alternating between the update of a finite set

$$\mathcal{A}_k^u = \{u_i^k\}_{i=1}^{N_k} \subset \mathrm{Ext}(B),$$

the so-called *active set*, and of an iterate

$$u_k \in \mathrm{cone}(\mathcal{A}_k^u) = \left\{ u = \sum_{i=1}^{N_k} \lambda_i^k u_i^k \,\middle|\, \lambda^k \in \mathbb{R}_+^{N_k} \right\},$$

where $\mathbb{R}_+ = [0, +\infty)$. Given the current active set $\mathcal{A}_k^u = \{u_i^k\}_{i=1}^{N_k}$ with $u_i^k \in \mathrm{Ext}(B)$ and the iterate $u_k \in \mathrm{cone}(\mathcal{A}_k^u)$, we compute the dual variable $p_k = -K_*\nabla F(Ku_k) \in \mathcal{C}$ and update

$$\mathcal{A}_k^{u,+} = \mathcal{A}_k^u \cup \{\widehat{v}_k^u\}, \quad \widehat{v}_k^u \in \arg\max_{v \in \mathrm{Ext}(B)} \langle p_k, v \rangle . \tag{3.1}$$

The above step requires the minimization of a linear functional over the not necessarily compact set of extremal points. Remarkably, this problem is well-posed, as shown in Lemma A.1 in Appendix A. Setting $N_k^+ = N_k + 1$ and $u_{N_k^+}^k = \widehat{v}_k^u$, we introduce the gauge function associated with $\mathcal{A}_k^{u,+}$ as

$$\kappa_{\mathcal{A}_k^{u,+}}(u) := \min_{\lambda \in \mathbb{R}_+^{N_k^+}} \left\{ \sum_{i=1}^{N_k^+} \lambda_i \,\middle|\, u = \sum_{i=1}^{N_k^+} \lambda_i u_i^k \right\}.$$

Note that $\kappa_{\mathcal{A}_k^{u,+}}(u)$ is well-defined on cone$(\mathcal{A}_k^{u,+})$, as the minimum is achieved due to lower semicontinuity of the ℓ^1 norm and closedness of $\mathbb{R}_+$. Subsequently, the next iterate u_{k+1} is found by solving the subproblem

$$\min_{u \in \text{cone}(\mathcal{A}_k^{u,+})} \left[F(Ku) + \kappa_{\mathcal{A}_k^{u,+}}(u) \right], \tag{3.2}$$

where the search for the minimizer is restricted to the cone spanned by $\mathcal{A}_k^{u,+}$ and the possibly complicated regularizer $\mathcal{G}$ is replaced by the easier-to-handle gauge function of the finite set $\mathcal{A}_k^{u,+}$. We point out that the objective functional in (3.2) constitutes an upper bound on J, i.e., there holds

$$J(u) \leq F(Ku) + \kappa_{\mathcal{A}_k^{u,+}}(u) \quad \text{for all } u \in \text{cone}(\mathcal{A}_k^{u,+})$$

due to the convexity and one-homogeneity of $\mathcal{G}$. Apart from that, Problems ($\mathcal{P}_\mathcal{M}$) and (3.2) share the same basic structure of minimizing the sum of a smooth fidelity term and a gauge-like regularizer. In particular, the existence result and the necessary first-order conditions in Proposition 2.3 also apply to (3.2). However, as shown in Proposition A.2 in Appendix A, the big advantage of (3.2) is that a solution can be computed as

$$u_{k+1} = \sum_{i=1}^{N_k^+} \lambda_i^{k+1} u_i^k, \tag{3.3}$$

where $\lambda^{k+1} \in \mathbb{R}_+^{N_k^+}$ solves the following finite dimensional optimization problem:

$$\min_{\lambda \in \mathbb{R}_+^{N_k^+}} \left[F\left(\sum_{i=1}^{N_k^+} \lambda_i K u_i^k \right) + \sum_{i=1}^{N_k^+} \lambda_i \right]. \tag{3.4}$$

Thus, in practice, we first determine a minimizer $\lambda^{k+1} \in \mathbb{R}^{N_k^+}$ to (3.4). We remark that (3.4) constitutes a finite dimensional non-smooth convex minimization problem, which can be efficiently solved by proximal methods or generalized Newton algorithms provided that F is sufficiently smooth. Once this is accomplished, the new iterate u_{k+1} is defined according to (3.3). As a final step, we truncate the active set $\mathcal{A}_k^{u,+}$ by removing all extremal points that were assigned a zero weight by the optimization procedure (3.4), i.e., we set

$$\mathcal{A}_{k+1}^u = \mathcal{A}_k^{u,+} \setminus \{ u_i^k \mid \lambda_i^{k+1} = 0, \ i = 1, \ldots, N_k^+ \}$$

and increment k by one. The method is summarized in Algorithm 1 below.

Algorithm 1 FC-GCG for $(\mathcal{P}_{\mathcal{M}})$

---

1. Let $u_0 = \sum_{i=1}^{N_0} \lambda_i^0 u_i^0$, $\lambda_i^0 > 0$, $\mathcal{A}_0^u = \{u_i^0\}_{i=1}^{N_0} \subset \text{Ext}(B)$.
for $k = 0, 1, 2, \ldots$ **do**
   2. Given $\mathcal{A}_k^u = \{u_i^k\}_{i=1}^{N_k} \subset \text{Ext}(B)$ and $u_k \in \text{cone}(\mathcal{A}_k^u)$, calculate p_k and $\widehat{v}_k^u$ with

$$p_k = -K_* \nabla F(Ku_k), \quad \widehat{v}_k^u \in \arg\max_{v \in \text{Ext}(B)} \langle p_k, v \rangle.$$

   if $\langle p_k, \widehat{v}_k^u \rangle \leq 1$ and $k \geq 1$ **then**
      3. Terminate with $\bar{u} = u_k$ a minimizer to $(\mathcal{P}_{\mathcal{M}})$.
   end if
   4. Update $N_k^+ = N_k + 1$, $u_{N_k^+}^k = \widehat{v}_k^u$ and $\mathcal{A}_k^{u,+} = \mathcal{A}_k^u \cup \{\widehat{v}_k^u\}$.
   5. Determine λ^{k+1} with

$$\lambda^{k+1} \in \arg\min_{\lambda \in \mathbb{R}_+^{N_k^+}} \left[F\left(\sum_{i=1}^{N_k^+} \lambda_i K u_i^k \right) + \sum_{i=1}^{N_k^+} \lambda_i \right],$$

   and set $u_{k+1} = \sum_{i=1}^{N_k^+} \lambda_i^{k+1} u_i^k$.
   6. Update

$$\mathcal{A}_{k+1}^u = \mathcal{A}_k^{u,+} \setminus \{u_i^k \mid \lambda_i^{k+1} = 0, \ i = 1, \ldots, N_k^+\}$$

   and set $N_{k+1} = \#\mathcal{A}_{k+1}^u$.
end for

3.1 Stopping condition

Note that Algorithm 1 terminates at iteration $k \geq 1$ if

$$\max_{v \in \text{Ext}(B)} \langle p_k, v \rangle \leq 1. \tag{3.5}$$

This is justified by the fact that, in this case, the current iterate u_k is a minimizer to $(\mathcal{P}_{\mathcal{M}})$, as shown in Proposition 3.1 below. In particular, in this situation, the algorithm converges in a finite number of iterations.

Proposition 3.1 *Suppose (A1)–(A3) in Assumption 2.2 hold. Let $u_k \in \text{cone}(\mathcal{A}_k^u)$ be generated by Algorithm 1, with $k \geq 1$. Set $p_k = -K_* \nabla F(Ku_k)$ and assume (3.5). Then u_k solves $(\mathcal{P}_{\mathcal{M}})$.*

Proof Thanks to Proposition 2.3 and Lemma A.1, u_k solves $(\mathcal{P}_{\mathcal{M}})$ if and only if

$$\langle p_k, u_k \rangle = \mathcal{G}(u_k), \quad \max_{v \in B} \langle p_k, v \rangle = \max_{v \in \text{Ext}(B)} \langle p_k, v \rangle \leq 1. \tag{3.6}$$

The second condition in (3.6) is satisfied by assumption, so we are only left to check the first. If $u_k = 0$ this is trivial, since $\mathcal{G}(0) = 0$. Therefore assume $u_k \neq 0$. This implies $\mathcal{G}(u_k) \neq 0$. Indeed, if by contradiction $\mathcal{G}(u_k) = 0$, then by one-homogeneity

of $\mathcal{G}$ we would obtain $\lambda u_k \in S^-(\mathcal{G}, 0)$ for all $\lambda \geq 0$. But then $S^-(\mathcal{G}, 0)$ would not be norm bounded, contradicting (A2). Thus $\mathcal{G}(u_k) \neq 0$. Since $u_k/\mathcal{G}(u_k) \in B$, from the second condition in (3.6) we get

$$\langle p_k, u_k \rangle \leq \mathcal{G}(u_k). \tag{3.7}$$

By construction $u_k = \sum_{i=1}^{N_k} \lambda_i^k u_i^k$, with $u_i^k \in \text{Ext}(B)$ and $\lambda^k \in \mathbb{R}_+^{N_k}$ a minimizer of

$$\min_{\lambda \in \mathbb{R}_+^{N_k}} \left[F \left(\sum_{i=1}^{N_k} \lambda_i K u_i^k \right) + \sum_{i=1}^{N_k} \lambda_i \right].$$

Note that since λ^k is a minimizer to the above problem, we have

$$\langle p_k, u_k \rangle = \sum_{i=1}^{N_k} \lambda_i^k \langle p_k, u_i^k \rangle = \sum_{i=1}^{N_k} \lambda_i^k, \quad \langle p_k, u_i^k \rangle \leq 1, \tag{3.8}$$

for $i = 1, \ldots, N_k$. As $u_k = \sum_{i=1}^{N_k} \lambda_i^k u_i^k$ with $u_i^k \in \text{Ext}(B)$, by convexity and one-homogeneity of $\mathcal{G}$ we estimate $\mathcal{G}(u_k) \leq \sum_{i=1}^{N_k} \lambda_i^k$. Therefore $\langle p_k, u_k \rangle = \mathcal{G}(u_k)$ by (3.7)–(3.8), ending the proof. □

Remark 3.2 Note that Algorithm 1 terminates also in case $\widehat{v}_k^u \in \mathcal{A}_k^u$ for some $k \geq 1$, i.e., if the algorithm cannot find a new point to insert. Indeed in this situation we have $\langle p_k, \widehat{v}_k^u \rangle \leq 1$ by the optimality conditions at (3.8). Therefore the stopping condition (3.5) is satisfied.

3.2 Worst-case convergence rates

The main contribution of the present manuscript is the derivation of convergence results for the sequence of *residuals*

$$r_J(u_k) = J(u_k) - \min_{u \in \mathcal{M}} J(u) \tag{3.9}$$

associated to the iterates generated by Algorithm 1. This is a challenging task for a variety of reasons. For example the space $\mathcal{M}$ is, in general, not reflexive. Moreover the functional J lacks useful properties, such as smoothness or strict convexity. Classical approaches [5, 35, 36] provide sublinear convergence rates for GCG methods defined in Banach spaces, however very few linear convergence results are available in the literature and these results are usually limited to specific cases. In this section, we state the convergence results anticipated in the introduction, and we detail the additional assumptions which are needed to prove them. Their proofs, which are rather technical, are then postponed to Sects. 6.1 and 6.2.10. We start with the following sublinear convergence result, which holds under the basic Assumptions (A1)–(A3).

Theorem 3.3 *Let (A1)–(A3) in Assumption 2.2 hold. Then, Algorithm 1 either terminates after a finite number of steps, with $\bar{u} = u_k$ a minimizer to ($\mathcal{P}_{\mathcal{M}}$), or there is a constant $c > 0$ such that*

$$r_J(u_k) \leq c \, \frac{1}{k+1} \quad \text{for all } k \in \mathbb{N}, \tag{3.10}$$

where r_J is defined at (3.9). Moreover, in this case, the sequence $\{u_k\}_k$ admits at least one weak* accumulation point and each of such point is a solution to ($\mathcal{P}_{\mathcal{M}}$). If the solution $\bar{u}$ to ($\mathcal{P}_{\mathcal{M}}$) is unique, then we have $u_k \overset{*}{\rightharpoonup} \bar{u}$ in $\mathcal{M}$ for the whole sequence.

3.3 Non-degeneracy and fast convergence

While Theorem 3.3 proves the convergence of Algorithm 1, the provided, slow, sublinear rate of convergence does not match the computed results of Sect. 4.1. Motivated by this gap between theory and numerical observations, we argue the asymptotic linear convergence of $r_J(u_k)$ provided that certain structural assumptions on Problem ($\mathcal{P}_{\mathcal{M}}$) are satisfied. First, in addition to Assumptions (A1)–(A3), we require that the solution set of the linear problem

$$\max_{v \in \mathcal{B}} \langle \bar{p}, v \rangle$$

consists of a finite number of extremal points $\bar{u}_1, \ldots, \bar{u}_N \in \text{Ext}(B)$, where $\bar{p}$ denotes the unique dual variable of ($\mathcal{P}_{\mathcal{M}}$), see (2.2). Moreover, we ask that the restriction of the operator K into the span of $\bar{\mathcal{A}} := \{\bar{u}_i\}_{i=1}^N$ is injective. Such assumptions ensure the uniqueness of the minimizer to ($\mathcal{P}_{\mathcal{M}}$), denoted by $\bar{u}$, see Proposition 3.5 below. Additionally, we ask that F is strongly convex around the unique optimal observation $\bar{y}$, see (2.2). This set of assumptions is summarized below.

Assumption 3.4 (Uniqueness and strong convexity)

(B1) The map $F \colon Y \to \mathbb{R}$ is strictly convex and strongly convex around the unique optimal observation $\bar{y}$, i.e., there exists a neighborhood $\mathcal{N}(\bar{y})$ of $\bar{y}$ and $\theta > 0$ such that

$$(\nabla F(y_1) - \nabla F(y_2), y_1 - y_2)_Y \geq \theta \|y_1 - y_2\|_Y^2, \quad \text{for all } y_1, y_2 \in \mathcal{N}(\bar{y}).$$

(B2) There is $N > 0$ and a finite collection of extremal points $\bar{\mathcal{A}} := \{\bar{u}_i\}_{i=1}^N \subset \text{Ext}(B)$ such that the unique dual variable $\bar{p} = -K_* \nabla F(\bar{y}) \in \mathcal{C}$ satisfies

$$\arg\max_{v \in \mathcal{B}} \langle \bar{p}, v \rangle = \{ v \in \mathcal{B} \mid \langle \bar{p}, v \rangle = 1 \} = \{\bar{u}_i\}_{i=1}^N.$$

(B3) The set $\{K\bar{u}_i\}_{i=1}^N \subset Y$ is linearly independent.

We now check that the above assumptions imply uniqueness of solutions to ($\mathcal{P}_{\mathcal{M}}$).

Proposition 3.5 *Let Assumptions (A1)–(A3) and (B2)–(B3) hold. Then ($\mathcal{P}_{\mathcal{M}}$) admits a unique solution $\bar{u}$, which is of the form*

$$\bar{u} = \sum_{i=1}^{N} \bar{\lambda}_i \bar{u}_i \,, \tag{3.11}$$

for some $\bar{\lambda}_i \geq 0$, where $\{\bar{u}_i\}_{i=1}^{N}$ are the points from (B2).

Proof Note that ($\mathcal{P}_{\mathcal{M}}$) admits at least one solution $\bar{u}$ thanks to Proposition 2.3. We first claim that $\bar{u} \in \text{cone}(\bar{\mathcal{A}})$, i.e., there exist $\bar{\lambda}_i \geq 0$ such that $\bar{u} = \sum_{i=1}^{N} \bar{\lambda}_i \bar{u}_i$. If $\bar{u} = 0$ the claim is trivial. Hence, suppose $\bar{u} \neq 0$. By the assumptions on $\mathcal{G}$ we conclude $\mathcal{G}(\bar{u}) > 0$. Recalling the first order optimality conditions from Proposition 2.3, we then infer $\langle \bar{p}, \bar{v} \rangle = 1$ with $\bar{v} := \bar{u}/\mathcal{G}(\bar{u}) \in B$. Moreover, by Assumption (B2), and arguing similarly to the proof of Lemma A.1, we get

$$\arg\max_{v \in B} \langle \bar{p}, v \rangle = \text{conv}(\bar{\mathcal{A}})\,, \qquad \max_{v \in B} \langle \bar{p}, v \rangle = 1 \,.$$

Therefore $\bar{v} \in \text{conv}(\bar{\mathcal{A}})$, showing our claim that $\bar{u} \in \text{cone}(\bar{\mathcal{A}})$. Now assume that $\bar{u}_1, \bar{u}_2$ minimize in ($\mathcal{P}_{\mathcal{M}}$). Then $\bar{u}_1 = \sum_{i=1}^{N} \bar{\lambda}_i \bar{u}_i$ and $\bar{u}_2 = \sum_{i=1}^{N} \bar{\gamma}_i \bar{u}_i$ for some $\bar{\lambda}_i, \bar{\gamma}_i \geq 0$. By Proposition 2.3 we have $K\bar{u}_1 = K\bar{u}_2$. From (B3) and linearity of K we then conclude $\bar{\lambda}_i = \bar{\gamma}_i$, so that $\bar{u}_1 = \bar{u}_2$. Thus ($\mathcal{P}_{\mathcal{M}}$) has a unique solution, and such a solution is of the form (3.11). $\square$

In the next set of assumptions we suppose *strict complementarity* for the minimizer $\bar{u}$, i.e.,

$$\bar{u} \notin \text{cone}\left(\bar{\mathcal{A}} \setminus \{\bar{u}_i\}\right)\,, \quad \text{for every } i = 1, \ldots, N \,,$$

or, equivalently, $\bar{\lambda}_i > 0$ for all $i = 1, \ldots, N$. The final assumption concerns the existence of a "distance function" g such that K is Lipschitz continuous and the linear functional $u \mapsto \langle \bar{p}, u \rangle$ grows quadratically, both with respect to g and in the vicinity of $\bar{u}_i \in \bar{\mathcal{A}}$. Of course, the particular form of g depends on the space $\mathcal{M}$ and the functional $\mathcal{G}$ and thus it has to be constructed on a case-by-case basis. We give an example in Sect. 4.1. This set of assumptions is summarized below.

Assumption 3.6 *(Non-degeneracy)*

(B4) The unique minimizer $\bar{u}$ of ($\mathcal{P}_{\mathcal{M}}$) in (3.11) is such that $\bar{\lambda}_i > 0$ for all $i = 1, \ldots, N$.

(B5) There exists a function $g \colon \text{Ext}(B) \times \text{Ext}(B) \to [0, \infty)$, positive constants $\tau, \kappa > 0$ and pairwise disjoint $d_{\mathcal{B}}$-closed neighborhoods $\{\bar{U}_i\}_{i=1}^{N}$ of $\{\bar{u}_i\}_{i=1}^{N}$, such that

$$\|K(u - \bar{u}_i)\|_Y \leq \tau g(u, \bar{u}_i)\,, \qquad 1 - \langle \bar{p}, u \rangle \geq \kappa g(u, \bar{u}_i)^2 \,, \tag{3.12}$$

for all $u \in U_i$ and $i = 1, \ldots, N$, where $U_i := \bar{U}_i \cap \text{Ext}(B)$ and $d_{\mathcal{B}}$ is the metric at (2.3).

Remark 3.7 Note that, without loss of generality, we can always choose $\{\bar{U}_i\}_{i=1}^N$ for which there exists a constant $\sigma > 0$ such that

$$\langle \bar{p}, u \rangle \leq 1 - \sigma \quad \text{for all } u \in \mathcal{B} \setminus \bar{U}_i, \ i = 1, \ldots, N, \tag{3.13}$$

due to Assumption (B2) and the $d_{\mathcal{B}}$-continuity of $u \mapsto \langle \bar{p}, u \rangle$.

We can finally state the main convergence result of the paper. For its proof we refer to Sect. 6.2.10.

Theorem 3.8 *Let Assumptions (A1)–(A3) and (B1)–(B5) hold. Then Algorithm 1 either terminates after a finite number of steps, with $\bar{u} = u_k$ the unique minimizer to ($\mathcal{P}_\mathcal{M}$), or there exists a constant $c > 0$ and $\zeta \in [3/4, 1)$ such that*

$$r_J(u_k) \leq c\,\zeta^k \tag{3.14}$$

for all $k \in \mathbb{N}$ sufficiently large. In this case there holds $u_k \overset{*}{\rightharpoonup} \bar{u}$ in $\mathcal{M}$.

Remark 3.9 Before ending this section, we briefly comment on Assumptions (B1)–(B5). First, we note that Assumption (B3) is trivially fulfilled if K is injective, as seen for example in Sect. 4.1. Moreover, Assumptions (B3), (B4), as well as the uniqueness of the minimizer to ($\mathcal{P}_\mathcal{M}$), are not fully independent. Indeed, if $\{K\bar{u}_i\}_{i=1}^N$ is linearly dependent and there is a solution $\bar{u} = \sum_{i=1}^N \bar{\lambda}_i \bar{u}_i$ with $\bar{\lambda}_i > 0$, then there exists a second solution $\tilde{u} \neq \bar{u}$ with $\tilde{u} = \sum_{i=1}^N \tilde{\lambda}_i \bar{u}_i$, $\tilde{\lambda}_i \geq 0$, and equality holds for at least one index, cf. also [52, Sect. 3.2]. Second, it is worthwhile to further discuss the quadratic growth condition in Assumption (B5) and relate it to more well-known concepts in the literature. For this purpose, recall that the necessary and sufficient optimality conditions for ($\mathcal{P}_\mathcal{M}$) are given by the variational subgradient inequality

$$\langle \bar{p}, u - \bar{u} \rangle + \mathcal{G}(\bar{u}) \leq \mathcal{G}(u) \quad \text{for all } u \in \mathcal{M}.$$

Due to the one-homogeneity of $\mathcal{G}$, this can be equivalently reformulated as

$$\langle \bar{p}, \bar{u} \rangle = \mathcal{G}(\bar{u}), \quad 1 \geq \langle \bar{p}, u \rangle \quad \text{for all } u \in B,$$

see Proposition 2.3. By applying Lemma A.1, this is equivalent to

$$\langle \bar{p}, \bar{u} \rangle = \mathcal{G}(\bar{u}), \quad 1 \geq \langle \bar{p}, u \rangle \quad \text{for all } u \in \text{Ext}(B).$$

Finally, due to Assumption (B2), we arrive at

$$\langle \bar{p}, \bar{u} \rangle = \mathcal{G}(\bar{u}), \quad 0 \geq \langle \bar{p}, u - \bar{u}_i \rangle \quad \text{for all } u \in \text{Ext}(B).$$

In particular, this implies that $\bar{p} \in \partial I_{\text{Ext}(B)}(\bar{u}_i)$ where $I_{\text{Ext}(B)}(\bar{u}_i)$ denotes the subdifferential of the nonconvex indicator function of $\text{Ext}(B)$ at $\bar{u}_i$. In this context, Assumption (B5) implies the locally strengthened condition

$$\langle \bar{p}, \bar{u} \rangle = \mathcal{G}(\bar{u}), \quad 0 \geq \langle \bar{p}, u - \bar{u}_i \rangle + \kappa g(u, \bar{u}_i)^2 \quad \text{whenever } u \in U_i.$$

This is very reminiscent of the concept of *strongly metric subregular subdifferentials* in convex optimization which plays a vital role for the derivation of fast convergence rates for proximal point methods [4], and "vanilla" generalized conditional gradient methods [48].

We point out that, in the general case, we are not aware of possibly stronger, but more intuitive, structural assumptions on $\bar{p}$ as well as Ext(B) which eventually ensure Assumptions (B2) and (B5). However, for particular instances such a characterization is indeed possible, see Sect. 4 and the examples and references therein.

4 Examples

Summarizing the previous sections, we see that the practical application and the convergence analysis of Algorithm 1 rest on three pillars: First, a characterization of the set of extremal points Ext(B), second, an efficient method for Step 2 in Algorithm 1 and, third, the derivation of sufficient structural assumptions to ensure Assumptions (B1)–(B5). In this section, we outline this program for three examples, namely, sparse initial value identification, trace regularization, and minimum effort problems, see Sects. 4.1, 4.2, 4.3, respectively. In all these cases, a particular focus is put on the computation of $\widehat{v}_k^u$ and the verification of Assumption (B5). For the sake of brevity, we decided to strike a balance between practically interesting settings and problems for which the characterization of Ext(B) and the derivation of Assumption (B5) can be done in a concise manner. As a main take-away message, these examples suggest that there is no general "recipe" for the resolution of Step 2 in Algorithm 1. Quite the reverse, the method of choice for computing $\widehat{v}_k^u$ as well as the associated computational burden strongly depend on the example at hand.

We also stress that Algorithm 1 is applicable to far more complex problems, in which characterizing extremal points and deriving quadratic growth conditions could also get much more convoluted. One such problem, namely the optimal transport regularization of dynamic inverse problems, is briefly teased in Sect. 4.4, and will be the subject of a follow-up work. Other examples are given by works [25, 45], in which the authors apply the program outlined above to certain regularizers given by certain infimal convolutions.

4.1 Sparse source identification

Let us consider the inverse problem of identifying the initial source of a heat equation on a convex polygonal spatial domain $\Omega \subset \mathbb{R}^2$ from distributed temperature measurements y_d at a given final time $T > 0$. Our particular interest lies in the recovery of sparse sources

$$u^\dagger = \sum_{i=1}^{N} \lambda_i^\dagger \delta_{x_i^\dagger},$$

given as a linear combination of finitely many point measures, where the coefficients $\lambda_i^\dagger \in \mathbb{R}$, the positions $x_i^\dagger \in \Omega$, and the number $N \in \mathbb{N}$ of points are all assumed to be unknown. Taking the ill-posedness of the described inverse problem into account we follow [21, 50] and consider the convex Tikhonov-regularized problem

$$\min_{u \in M(\Omega), y} \left[\frac{1}{2} \|y(T) - y_d\|_{L^2(\Omega)}^2 + \beta \|u\|_{M(\Omega)} \right]. \tag{4.1}$$

Here $M(\Omega)$ denotes the space of Borel measures on the open set Ω, y is a scalar function defined on $[0, 1] \times \Omega$ with $y(t) := y(t)(\cdot)$, $y_d \in L^2(\Omega)$ is a given desired state, and the pair (y, u) satisfies, in the sense of distributions, the heat equation

$$\partial_t y - \triangle y = 0 \text{ in } (0, T) \times \Omega, \ y = 0 \text{ in } (0, T) \times \partial\Omega, \ y(0) = u \text{ in } \Omega. \tag{4.2}$$

The a priori assumption on the sparsity of the unknown source is encoded in the choice of the regularizer, defined as the total variation norm of u with $\beta > 0$. To fit (4.1) into the setting of ($\mathcal{P}_\mathcal{M}$) we set $\mathcal{C} = C_0(\Omega)$, the space of continuous functions vanishing on $\partial\Omega$, and we equip it with the canonical supremum norm

$$\|p\|_\mathcal{C} = \max_{x \in \Omega} |p(x)| \quad \text{for all } p \in C_0(\Omega).$$

This makes $\mathcal{C}$ a Banach space. According to the Riesz-Markov-Kakutani theorem we have $\mathcal{C}^* \simeq \mathcal{M}$ for $\mathcal{M} = M(\Omega)$. Moreover, define $Y = L^2(\Omega)$, $F = (1/2)\| \cdot -y_d\|_{L^2(\Omega)}^2$ and $\mathcal{G} = \beta\| \cdot \|_{M(\Omega)}$. Of course, these functionals satisfy (A1) and (A2) in Assumption 2.2. Finally, we replace the PDE constraint by introducing a *source-to-observation* operator $K \colon M(\Omega) \to L^2(\Omega)$ mapping a measure $u \in M(\Omega)$ to $y(T)$, where y solves (4.2). It is readily verified that K is injective, thanks to a priori estimates for weak solutions to (4.2) [26, Lemma 2.2], as well as weak*-to-strong continuous [26, Lemma 2.3]. Hence, (4.1) admits a unique solution and (A3) in Assumption 2.2 is satisfied. Moreover, K is the adjoint of the operator $K_* \colon L^2(\Omega) \to C_0(\Omega)$ defined by $K_*\varphi := z(0)$, where the pair (z, φ) satisfies, in the sense of distributions, the backwards heat equation

$$\partial_t z + \triangle z = 0 \text{ in } (0, T) \times \Omega, \ z = 0 \text{ in } (0, T) \times \partial\Omega, \ z(T) = \varphi \text{ in } \Omega. \tag{4.3}$$

For more details we refer to [26, 50]. Note that K_* is well-defined as $z(0) \in C_0(\Omega)$, due to parabolic regularity estimates. For sake of completeness, this is justified in Lemma B.1 in the appendix.

The next lemma characterizes the set of extremal points of the unit ball of the regularizer $\mathcal{G}$, that is, of the set $B = S^-(\beta\| \cdot \|_{M(\Omega)}, 1)$.

Lemma 4.1 *We have*

$$\text{Ext}(B) = \{\sigma\beta^{-1}\delta_x \mid x \in \Omega, \ \sigma \in \{-1, 1\}\}. \tag{4.4}$$

Moreover, $\overline{\text{Ext}(B)}^* = \text{Ext}(B) \cup \{0\}$.

Proof The characterization of $\mathrm{Ext}(B)$ is well-known [13, Proposition 4.1]. As for the second claim, note that $0 \in \overline{\mathrm{Ext}(B)}^*$: indeed one can take $\{x_k\}_k$ in Ω such that $x_k \to x$ with $x \in \partial\Omega$, so that $u_k := \beta^{-1}\delta_{x_k} \stackrel{*}{\rightharpoonup} 0$. Thus $\mathrm{Ext}(B) \cup \{0\} \subset \overline{\mathrm{Ext}(B)}^*$. For the opposite inclusion, assume given a sequence $u_k = \sigma_k\beta^{-1}\delta_{x_k}$ in $\mathrm{Ext}(B)$, such that $u_k \stackrel{*}{\rightharpoonup} u$. Then, up to subsequences, $\sigma_k \to \sigma \in \{-1, 1\}$ and $x_k \to x \in \overline{\Omega}$. Hence, if $x \in \partial\Omega$ then $u = 0$, while if $x \in \Omega$ then $u = \sigma\beta^{-1}\delta_x \in \mathrm{Ext}(B)$. $\qquad\square$

Moreover, applying Proposition 2.3 and the characterization of K_*, we immediately deduce the optimality conditions for (4.1).

Proposition 4.2 *Let $\bar{u} \in M(\Omega)$ be given and denote by $\bar{z}$ the solution to (4.3) for $\varphi = y_d - \bar{y}$, $\bar{y} = K\bar{u}$. Then $\bar{u}$ is a minimizer of (4.1) if and only if*

$$\langle \bar{z}(0), \bar{u} \rangle = \beta\|\bar{u}\|_{M(\Omega)}, \quad \|\bar{z}(0)\|_C \leq \beta.$$

This implies

$$\mathrm{supp}\,\bar{u} \subset \{ x \in \Omega \mid |\bar{z}(0)(x)| = \beta \}.$$

Thanks to the characterization of extremal points presented in Lemma 4.1, the FC-GCG method presented in Algorithm 1 for solving (4.1) generates a sequence of iterates

$$u_k = \sum_{i=1}^{N_k} \lambda_i^k u_i^k, \quad u_i^k = \sigma_i^k\beta^{-1}\delta_{x_i^k}, \quad \sigma_i^k \in \{-1, 1\}, \quad x_i^k \in \Omega, \quad \lambda_i^k \geq 0,$$

as well as an associated sequence of active sets $\mathcal{A}_k^u = \{u_i^k\}_{i=1}^{N_k}$. Moreover, let z_k denote the solution of (4.3) for $\varphi = y_d - y_k(T)$, $y_k(T) = Ku_k$. We now claim that the new candidate extremal point in the iteration k of Algorithm 1 can be chosen as

$$\widehat{v}_k^u = \mathrm{sign}(z_k(0)(\widehat{x}_k))\beta^{-1}\delta_{\widehat{x}_k} \text{ where } \widehat{x}_k \in \arg\max_{x\in\Omega} |z_k(0)(x)|, \tag{4.5}$$

i.e., Step 2 in Algorithm 1 is equivalent to computing a global extremum of a continuous function. This is verified in the following proposition.

Proposition 4.3 *Let $\widehat{v}_k^u$ be defined as in (4.5). There holds $z_k(0) = -K_*\nabla F(Ku_k)$ as well as*

$$\langle z_k(0), \widehat{v}_k^u \rangle = \max_{v\in\mathcal{B}} \langle z_k(0), v \rangle.$$

Proof We directly get $z_k(0) = K_*(y_d - y(T)) = -K_*\nabla F(Ku_k)$ from the characterization of K_*. The remaining statement follows directly from

$$\langle z_k(0), v \rangle \leq \|z_k(0)\|_C\|v\|_{\mathcal{M}} \leq \|z_k(0)\|_C/\beta \quad \text{for all } v \in \mathcal{B}$$

as well as

$$\langle z_k(0), \widehat{v}_k^u \rangle = \text{sign}(z_k(0)(\widehat{x}_k))\beta^{-1}z_k(0)(\widehat{x}_k) = \|z_k(0)\|_{\mathcal{C}}/\beta \,.$$

$\square$

Thus, for sparsity examples, solving (3.1) amounts to computing an extremum of the continuous function $z_k(0) = -K_*\nabla F(Ku_k)$, where u_k is the iterate generated by Algorithm 1. Since $|z_k(0)|$ is, in general, non-concave, this optimization task could be non-trivial. However, since the spatial domain Ω is low dimensional in this example, it is possible to resort to heuristic strategies to approximate an extremum of $|z_k(0)|$. In particular, a standard widely accepted strategy [11, 21], consists in discretizing the domain Ω using a uniform grid $\{x^h\}_{h=1,\dots N^2}$ and performing local searches around x^h using gradient descent methods. Then an extremum of $|z_k(0)|$ can be estimated as

$$\underset{y^h: h=1,\dots,N^2}{\arg\max} \ |z_k(0)(y^h)| \,, \tag{4.6}$$

where y^h is the outcome of the local search around the point of the grid x^h. Moreover, a practical implementation of Algorithm 1 for this particular problem also entails a discretization of the heat equation, e.g. by piecewise polynomial and continuous finite elements. In this case, the computation of the new point $\overline{x}_k$ becomes trivial, see Sect. 4.1.1.

We now discuss the non-degeneracy conditions. Denoting by $\bar{u} \in M(\Omega)$ the minimizer of (4.1), we propose a natural and easy to verify set of assumptions for $\bar{z}(0)$ that implies our general non-degeneracy assumptions from Sect. 3.3 for a suitable choice of g. More precisely, this new set of assumptions on $\bar{z}(0)$ will imply Assumption (B2), (B3) and (B5). We remark that Assumption (B4) still needs to be assumed to ensure the fast convergence; however we decide not to state it in the following set of assumptions as, for this specific example, it would be formulated exactly as in (B4). Moreover, its verification can be done straightforwardly looking at the structure of the unique minimizer of (4.1). We finally remind that $\bar{z}$ is the solution to (4.3) for $\varphi = y_d - \bar{y}$, $\bar{y} = K\bar{u}$ and therefore by the characterization of K_* it holds that $\bar{z}(0) = -K_*\nabla F(K\bar{u})$. The following structural assumptions are made, see also [50, 52].

(C1) There are $\bar{x}_i \in \Omega, i = 1, \dots, N, N > 0$, such that

$$\{x \in \Omega \mid |\bar{z}(0)(x)| = \beta\} = \{\bar{x}_i\}_{i=1}^N. \tag{4.7}$$

(C2) There exists $\gamma > 0$ such that for all $i = 1, \dots, N$ we have

$$\text{sign}(\bar{z}(0)(\bar{x}_i))(\delta x, \nabla^2 \bar{z}(0)(\bar{x}_i)\delta x)_{\mathbb{R}^2} \leq -\gamma|\delta x|^2 \ \text{ for all } \ \delta x \in \mathbb{R}^2.$$

Regarding (C2) recall that $\bar{z}(0)$ is at least two times continuously differentiable in the vicinity of $\bar{x}_i$, $i = 1, \dots, N$, see Lemma B.1. Loosely speaking, the additional requirements in Assumptions (C1)–(C2) state that $\bar{z}(0)$ only admits a finite number of global minima/maxima and its curvature around them does not degenerate. The

latter corresponds to a second order sufficient optimality condition for the global extrema of $\bar{z}(0)$. We now prove that (C1) and (C2) imply (B2), (B3) and (B5). First we show that (C1) guarantees Assumptions (B2)–(B3). For this purpose, set $\bar{u}_i = \text{sign}(\bar{z}(0)(\bar{x}_i))\beta^{-1}\delta_{\bar{x}_i} \in \text{Ext}(B)$ for every $i = 1, \ldots, N$.

Proposition 4.4 *Let Assumption (C1) hold. Then, we have*

$$\underset{v \in \overline{\text{Ext}(B)}^*}{\arg\max} \langle \bar{z}(0), v \rangle = \{\bar{u}_i\}_{i=1}^N .$$

Moreover, the set $\{K\bar{u}_i\}_{i=1}^N$ is linearly independent.

Proof First recall that every $v \in \text{Ext}(B)$ is of the form $v = \sigma\beta^{-1}\delta_x$ for some $\sigma \in \{-1, 1\}$, $x \in \Omega$. Moreover, we have

$$\langle \bar{z}(0), \sigma\beta^{-1}\delta_x \rangle = (\sigma/\beta)\bar{z}(0)(x) \leq \|\bar{z}(0)\|_C/\beta = 1,$$

see Proposition 4.2 and (4.7), with equality if and only if $|\bar{z}(0)(x)| = \|\bar{z}(0)\|_C$, $\sigma = \text{sign}(\bar{z}(0)(x))$. Hence, the claimed statement follows from (4.7) and Lemma 4.1. Finally, the linear independence of $\{K\bar{u}_i\}_{i=1}^N$ follows from the injectivity of K. $\qquad\square$

Next we address Assumption (B5). For every subdomain Ω_0 with $\bar{\Omega}_0 \subset \Omega$ define the quantities

$$\|\psi\|_{\text{Lip}(\Omega_0)} := \sup_{x,y \in \Omega_0, x \neq y} \frac{|\psi(x) - \psi(y)|}{|x - y|} \tag{4.8}$$

for $\psi \in C_0(\Omega)$ and

$$\|K_*\|_{Y,\text{Lip}(\Omega_0)} := \sup_{\|y\|_Y \leq 1} \|K_*y\|_{\text{Lip}(\Omega_0)} . \tag{4.9}$$

Note that $\|K_*\|_{Y,\text{Lip}(\Omega_0)} < \infty$, due to Lemma B.1. The next lemma shows that Assumptions (C1)–(C2) imply the quadratic growth of $\beta - |\bar{z}(0)|$ around $\{\bar{x}_i\}_{i=1}^N$. The proof can be found in Sect. B.1.1 in the appendix.

Lemma 4.5 *Let Assumption (C1)–(C2) hold and fix an index $i = 1, \ldots, N$. Then, there exists $R > 0$ such that $\bar{B}_R(\bar{x}_i) \subset \Omega$, $\text{sign}(\bar{z}(0)(x)) = \text{sign}(\bar{z}(0)(\bar{x}_i))$ for all $x \in B_R(\bar{x}_i)$ and*

$$\beta - |\bar{z}(0)(x)| \geq (\gamma/4)|x - \bar{x}_i|^2 \quad \text{for all } x \in B_R(\bar{x}_i), \tag{4.10}$$

$$\|K(\delta_x - \delta_{\bar{x}_i})\|_Y \leq \|K_*\|_{Y,\text{Lip}(B_R(\bar{x}_i))}|x - \bar{x}_i|, \quad \text{for all } x \in B_R(\bar{x}_i). \tag{4.11}$$

Now, given arbitrary extremal points $\sigma_1\beta^{-1}\delta_{x_1}, \sigma_2\beta^{-1}\delta_{x_2} \in \text{Ext}(B)$, with $\sigma_1, \sigma_2 \in \{-1, 1\}$ and $x_1, x_2 \in \Omega$, define the distance function $g: \text{Ext}(B) \times \text{Ext}(B) \to [0, \infty)$ by

$$g(u_1, u_2) = |\sigma_1 - \sigma_2| + |x_1 - x_2|. \tag{4.12}$$

Such g will be the one verifying Assumption (B5). In the next lemma we show that the weak* convergence in $M(\Omega)$ of a sequence of extremal points to $\bar{u}_i$ is equivalent to convergence with respect to g. The proof can be found in Sect. B.1.2 in the appendix.

Lemma 4.6 *Consider a sequence $\{u_k\}_k$ in $\mathrm{Ext}(B)$, i.e., $u_k = \sigma_k \beta^{-1} \delta_{x_k}$, for $\sigma_k \in \{-1, 1\}$ and $x_k \in \Omega$. Then, there holds*

$$u_k \overset{*}{\rightharpoonup} \bar{u}_i \text{ if and only if } \lim_{k\to\infty} g(u_k, \bar{u}_i) = 0.$$

In particular, if $u_k \overset{*}{\rightharpoonup} \bar{u}_i$, then $\sigma_k = \mathrm{sign}(\bar{z}(0)(\bar{x}_i))$ for all $k \in \mathbb{N}$ large enough.

Finally we combine the previous observations to conclude Assumption (B5).

Proposition 4.7 *Let $R > 0$ be chosen according to Lemma 4.5. Then, there is a $d_{\mathcal{B}}$-neighbourhood $\bar{U}_i$ of $\bar{u}_i$ which satisfies*

$$U_i := \bar{U}_i \cap \mathrm{Ext}(B) \subset \left\{ \sigma\beta^{-1}\delta_x \,\middle|\, \sigma = \mathrm{sign}(\bar{z}(0)(\bar{x}_i)), \ x \in B_R(\bar{x}_i) \right\}$$

as well as

$$\|K(u - \bar{u}_i)\|_{L^2} \leq (\|K_*\|_{Y,\mathrm{Lip}(B_R(\bar{x}_i))}/\beta)\, g(u, \bar{u}_i),$$
$$\langle \bar{z}(0), \bar{u}_i - u \rangle \geq (\gamma/(4\beta))\, g(u, \bar{u}_i)^2$$

for every $u \in U_i$ and for every $i = 1, \ldots, N$.

Proof The statement on the existence of a $d_{\mathcal{B}}$-neighbourhood $\bar{U}_i$ with the stated properties follows immediately from Lemma 4.6. In fact, if such a neighbourhood does not exist, then there exists a sequence $\{u_k\}_k$ in $\mathrm{Ext}(B)$, i.e., $u_k = \sigma_k \beta^{-1}\delta_{x_k}$, with $u_k \overset{*}{\rightharpoonup} \bar{u}_i$ as well as $\sigma_k \neq \mathrm{sign}(\bar{z}(0)(\bar{x}_i))$ or $x_k \notin B_R(\bar{x}_i)$. This contradicts $\lim_{k\to\infty} g(u_k, \bar{u}_i) = 0$. The remaining statements are also readily verified using Lemma 4.5 noting that for every $u = \sigma\beta^{-1}\delta_x \in U_i$, we have $g(u, \bar{u}_i) = |x - \bar{x}_i|$, $\langle \bar{z}(0), u \rangle = |\bar{z}(0)(x)|/\beta$ and $\langle \bar{z}(0), \bar{u}_i \rangle = 1$. $\qquad\square$

4.1.1 Numerical experiment

We close this section with a numerical experiment showing the effectiveness of Algorithm 1 in a concrete setting. For this purpose, set $\Omega = (0, 1)^2$ and $T = 0.1$. Moreover, fix $u^\dagger = 25\delta_{x_1} - 10\delta_{x_2}$, where $x_1 = (0.75, 0.75)$ and $x_2 = (0.25, 0.25)$, as well as $y_d = Ku^\dagger + \zeta$ where ζ is a noise term with $\|\zeta\|_{L^2(\Omega)}/\|Ku^\dagger\|_{L^2(\Omega)} \approx 0.1$. The regularization parameter is chosen as $\beta = 0.001$. The heat equation is discretized using a dg(0)cg(1) scheme on a temporal grid with stepsize $\delta = 0.001$ and a uniform triangulation of Ω with grid size $h = 1/128$ and nodes $\{x_i^h\}_{i=1}^{N_h}$. For the adjoint equation, a conforming discretization scheme is considered. All computations were carried out in Matlab 2019 on a notebook with 32 GB RAM and an Intel®Core™ i7-10870H
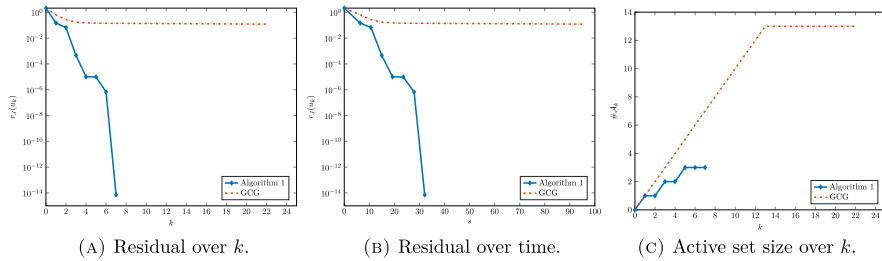
(A) Residual over k. (B) Residual over time. (C) Active set size over k.

Fig. 1 Convergence behaviour of relevant quantities

CPU@2.20 GHz. In Fig. 1(A), we report on the convergence history of the residuals $r_J(u_k) = J(u_k) - \min_{u \in M(\Omega)} J(u)$ associated with a sequence $\{u_k\}_k$ generated by Algorithm 1 starting from $u_0 = 0$ and $\mathcal{A}_0 = \emptyset$. Due to the dg(0)cg(1) scheme used in the discretization of the state equation, the dual variable $z_k(0)$ is now a piecewise linear and continuous function on the spatial grid. As a consequence, $|z_k(0)|$ achieves its global maximum in a gridpoint and the new candidate $\widehat{x}_k$ can be cheaply computed as

$$\widehat{x}_k \in \arg\max_{x \in \{x_i^h\}_{i=1}^{N_h}} |z_k(0)|.$$

In each iteration, the finite dimensional subproblem (3.2) is solved using a semismooth Newton method. Moreover, we plot the size of the support of u_k in dependence of k in Fig. 1(C). By construction, this corresponds to the number of Dirac deltas in the active set $\mathcal{A}_k^u$. In order to highlight the practical efficiency of Algorithm 1, we also include a comparison to the iterates generated by a generalized conditional gradient method (GCG) given by

$$u_{k+1} = (1 - s_k)u_k + s_k v_k \quad \text{where} \quad v_k = \begin{cases} M_0 \widehat{v}_k^u & \text{if } \|z_k(0)\|_{\mathcal{C}} \geq \beta, \\ 0 & \text{else.} \end{cases}$$

Here $\widehat{v}_k^u$ is chosen as before, $M_0 = J(0)/\beta$ and $s_k \in (0, 1)$ is an explicitly given stepsize as described in [21]. Both methods were run for a maximum of 100 iterations or until $r_J(u_k) \leq 10^{-12}$. As expected, the GCG update exhibits the typical sublinear convergence behaviour of conditional gradient methods. In particular, after 200 iterations the residual is still of magnitude $r_J(u_k) \approx 5 \times 10^{-2}$. In contrast, we observe a vastly improved rate of convergence for Algorithm 1. The stopping criterion is met after 7 iterations. Moreover, while the support size of u_k in GCG strictly increases in the first 13 iterations, Algorithm 1 removes Dirac deltas which are assigned a zero coefficient. This leads to smaller active sets and thus sparser iterates. Both observations are testament to the practical efficiency of Algorithm 1. Finally, for a fair comparison, we also plot the residual as a function of the computational time (in seconds) for $k = 1, \ldots, 20$. This is done to acknowledge the vastly different computational cost of the update steps in both methods, i.e., forming a convex combination in GCG and

the full resolution of a finite-dimensional minimization problem in Algorithm 1. As we can see, the additional computational effort of fully resolving (3.2) is outweighed by its practical utility. More in detail, Algorithm 1 converges after around 30 s while GCG fails to decrease $r_J(u_k)$ below 10^{-1} in the considered time frame.

4.2 Rank-one matrix reconstruction by trace regularization

Let H be a separable Hilbert space with norm $\|\cdot\|_H$ induced by an inner product $(\cdot, \cdot)_H$. In the following, to simplify notation, we will focus on infinite dimensional Hilbert spaces, the finite dimensional case, i.e. $H \simeq \mathbb{R}^n$, follows by the same arguments. Denote by $K(H)$ the space of bounded, linear, compact and self-adjoint operators from H into itself which we equip with the standard operator norm. An operator $\mathcal{U} \in K(H)$ is called positive, denoted by $\mathcal{U} \geq 0$, if

$$(h, \mathcal{U}h)_H \geq 0 \quad \text{for all } h \in H.$$

Moreover, let $\{h_i\}_{i \in \mathbb{N}}$ be an orthonormal basis (ONB) of H. For an operator $\mathcal{U} \in K(H)$ we formally define its trace as

$$\text{Tr}(\mathcal{U}) = \sum_{i \in \mathbb{N}} (h_i, \mathcal{U}h_i)_H.$$

An operator $\mathcal{U} \in K(H)$ is called *trace-class* if $\text{Tr}(|\mathcal{U}|) < \infty$, where $|\mathcal{U}|$ is the unique positive square root of $\mathcal{U}^2$, and *Hilbert-Schmidt* (HS) if $\text{Tr}(\mathcal{U}^2) < \infty$. The set of all Hilbert-Schmidt operators together with the norm $\|\mathcal{U}\|_{\text{HS}} = (\mathcal{U}, \mathcal{U})_{\text{HS}}^{1/2}$ induced by the inner product

$$(\mathcal{U}_1, \mathcal{U}_2)_{\text{HS}} = \text{Tr}(\mathcal{U}_1 \mathcal{U}_2) \quad \text{for all } \mathcal{U}_1, \mathcal{U}_2 \in \text{HS}(H)$$

forms a Hilbert space $\text{HS}(H)$. Moreover, the space of all trace-class operators, denoted by $T(H)$, forms a Banach space when equipped with the nuclear norm $\|\mathcal{U}\|_{\text{T}} = \text{Tr}(|\mathcal{U}|)$. In this case, we also have $K(H)^* \simeq T(H)$ where the duality pairing is realized by

$$\langle \mathcal{U}_1, \mathcal{U}_2 \rangle = \text{Tr}(\mathcal{U}_1 \mathcal{U}_2) \quad \text{for all } \mathcal{U}_1 \in K(H), \ \mathcal{U}_2 \in T(H).$$

With these prerequisites, consider

$$\min_{\mathcal{U} \in T(H), \ \mathcal{U} \geq 0} [F(K\mathcal{U}) + \beta \text{Tr}(\mathcal{U})] \tag{4.13}$$

where $\beta > 0$ and $K \colon \text{HS}(H) \to Y$ is weak-to-strong continuous. Problems of this type appear, e.g., as convex relaxations of quadratic inverse problems, see [7, 23], since they are known to favour solutions with rank one. As before, we start by computing the extremal points of

$$B = \{\mathcal{U} \in T(H) \mid \beta \text{Tr}(\mathcal{U}) \leq 1, \ \mathcal{U} \geq 0\}.$$

For this purpose, given $h \in H$, we introduce the associated rank one operator $\mathcal{U}_h := h \otimes h$ by $\mathcal{U}_h h_1 = (h, h_1)_H h$ for all $h_1 \in H$. Note that $\mathcal{U}_h \in T(H)$ with $\mathrm{Tr}(\mathcal{U}_h) = \|h\|_H^2$. Moreover, we recall that for every $\mathcal{U} \in K(H)$, $\mathcal{U} \geq 0$, there exists an ONB $\{h_i^{\mathcal{U}}\}_i$ of H as well as $\{\sigma_i^{\mathcal{U}}\}_i$ in $\mathbb{R}$ with $\sigma_i^{\mathcal{U}} \geq \sigma_{i+1}^{\mathcal{U}} \geq 0$, $i \in \mathbb{N}$, and

$$\mathcal{U} h_i^{\mathcal{U}} = \sigma_i^{\mathcal{U}} h_i^{\mathcal{U}}, \quad \mathcal{U} = \sum_{i \in \mathbb{N}} \sigma_i^{\mathcal{U}} h_i^{\mathcal{U}} \otimes h_i^{\mathcal{U}}.$$

We arrive at the following characterization.

Lemma 4.8 *We have*

$$\mathrm{Ext}(B) = \{\, \beta^{-1} h \otimes h \mid \|h\|_H = 1 \,\} \cup \{0\}, \tag{4.14}$$

as well as

$$\mathcal{B} = \{\, \beta^{-1} h \otimes h \mid \|h\|_H \leq 1 \,\}. \tag{4.15}$$

A proof of Lemma 4.8 can be found in the appendix, cf. Sect. B.2.1. A consequence of Lemma 4.8 is that minimizers to (4.13) always exhibit finite rank and share their eigenfunctions with the dual variable. Additionally, the next proposition provides optimality conditions for the Problem (4.13). The proof can be found in the appendix, cf. Sect. B.2.2.

Proposition 4.9 *Let $\bar{\mathcal{U}} \in T(H)$, $\bar{\mathcal{U}} \geq 0$, be given and set*

$$\bar{P} = -K_* \nabla F(K \bar{\mathcal{U}}) \in K(H).$$

Then $\bar{\mathcal{U}}$ is a minimizer of (4.13) if and only if

$$\sigma_1^{\bar{P}} \leq \beta, \quad \mathrm{Tr}(\bar{P} \bar{\mathcal{U}}) = \beta \, \mathrm{Tr}(\bar{\mathcal{U}}). \tag{4.16}$$

In particular, if $\sigma_1^{\bar{P}} = \beta$ and $\bar{N} \geq 1$ is the smallest index with $\sigma_{\bar{N}+1}^{\bar{P}} < \sigma_1^{\bar{P}}$, this implies that a minimizer $\bar{\mathcal{U}}$ of (4.13) is of the form

$$\bar{\mathcal{U}} = \sum_{i=1}^{\bar{N}} \bar{\lambda}_i h_i^{\bar{P}} \otimes h_i^{\bar{P}} \quad \text{for some } \bar{\lambda}_i \geq 0. \tag{4.17}$$

Now, denote by $\mathcal{U}_k$ the k-th iterate in Algorithm 2 and let P_k be the corresponding dual variable. Then we immediately obtain

$$\underset{\mathcal{V} \in \mathrm{Ext}(B)}{\arg\max} \ \mathrm{Tr}(P_k \mathcal{V}) = \left\{\, \beta^{-1} h \otimes h \,\middle|\, P_k h = \sigma_1^{P_k} h, \ \|h\|_H = 1 \,\right\}.$$

In particular, selecting the new candidate point $\widehat{\mathcal{V}}_k^U$ in Step 2 of Algorithm 1 can be realized by computing one eigenfunction for the leading eigenvalue of P_k and $\mathcal{U}_k$ has at

most rank k. Regarding the fast convergence of Algorithm 2, we recall that, in the best case, Problem (4.13) produces rank one solutions. In the following, we make a slightly stronger assumption. More in detail, denoting by $\bar{P}$ the unique optimal dual variable for (4.13), i.e., there holds $\bar{P} = -K_*\nabla F(K\bar{\mathcal{U}})$ for every minimizer $\bar{\mathcal{U}}$ of (4.13), we assume that:

(D1) The first eigenvalue of $\bar{P}$ is simple, i.e., $\beta = \sigma_1^{\bar{P}} > \sigma_2^{\bar{P}}$.

On the one hand, invoking (4.17), this implies that the solution to (4.13) has at most rank one. On the other hand, it also ensures the unique solvability of the linear problem induced by $\bar{P}$ as well as a quadratic growth behavior with respect to the Hilbert-Schmidt norm as the next theorem shows. The proof can be found in the appendix, cf. Sect. B.2.3.

Theorem 4.10 *Let Assumption (D1) hold. Then the unique optimal dual variable $\bar{P}$ of (4.13) satisfies*

$$\arg\max_{\mathcal{V}\in\mathrm{Ext}(B)} \langle \bar{P}, \mathcal{V}\rangle = \arg\max_{\mathcal{V}\in B} \langle \bar{P}, \mathcal{V}\rangle = \{\beta^{-1}h_1^{\bar{P}} \otimes h_1^{\bar{P}}\}. \tag{4.18}$$

In particular, this implies that the solution of (4.13) is unique and of the form $\bar{\mathcal{U}} = \bar{\lambda}h_1^{\bar{P}} \otimes h_1^{\bar{P}}, \bar{\lambda} \geq 0$. Moreover, setting $\bar{\mathcal{U}}_1 = \beta^{-1}h_1^{\bar{P}} \otimes h_1^{\bar{P}}$, there are $\tau, \kappa > 0$ with

$$\|K(\mathcal{U} - \bar{\mathcal{U}}_1)\|_Y \leq \tau\|\mathcal{U} - \bar{\mathcal{U}}_1\|_{\mathrm{HS}}, \tag{4.19}$$

$$1 - \langle \bar{P}, \mathcal{U}\rangle \geq \kappa\|\mathcal{U} - \bar{\mathcal{U}}_1\|_{\mathrm{HS}}^2, \tag{4.20}$$

for all $\mathcal{U} \in B$ with $\mathrm{Tr}(\mathcal{U}) = \beta^{-1}$.

Hence, Assumption (D1) implies Assumptions (B2) and (B5) for $g(\cdot, \cdot) = \|\cdot - \cdot\|_{\mathrm{HS}}$ and $U_1 = \mathrm{Ext}(B) \cap \mathcal{B}$. As a consequence, if F is strongly convex in the sense of Assumption (B1) as well as $K\bar{\mathcal{U}}_1 \neq 0$ and $\bar{\lambda} > 0$, then Algorithm 1 converges with an asymptotically linear rate.

Remark 4.11 For Problem (4.13) and $H \simeq \mathbb{R}^n$, Algorithm 1 can be interpreted as a particular instance of a totally corrective conditional gradient method proposed for general matrix problems in [63]. However, the presented linear convergence analysis in this earlier work does not cover trace-regularized problems since, even in the finite dimensional case, $\mathrm{Ext}(B)$ is uncountable. Also in the finite dimensional setting, we want to mention [42] where the author presents a conditional gradient method for the trace-constrained setting, i.e.,

$$\min_{\mathcal{U}\in\mathbb{R}^{n\times n}} F(K\mathcal{U}) \quad \text{s.t.} \quad \mathcal{U} = \mathcal{U}^\top, \ \mathrm{Tr}(\mathcal{U}) \leq \alpha, \ \mathcal{U} \geq 0,$$

which eventually converges linearly under assumptions similar to (D1). From this point of view, the results of this section can be seen as a first step to bridging the gap between the trace-regularized and the trace-constrained setting.

4.3 Minimum effort problems

Let $\Omega \subset \mathbb{R}^d$ be open and bounded, with $d \in \mathbb{N}$, $d \geq 1$. Consider the space $\mathcal{M} := L^\infty(\Omega)$ of essentially bounded Lebesgue measurable functions on Ω, which we equip with the usual essential supremum norm

$$\|u\|_\infty := \operatorname*{ess\,sup}_{x \in \Omega} |u(x)| \quad \text{for all } u \in L^\infty(\Omega).$$

We consider *minimum effort problems*, that is,

$$\min_{u \in L^\infty(\Omega)} J(u), \quad J(u) := F(Ku) + \alpha \|u\|_\infty, \tag{4.21}$$

where $\alpha > 0$ is a fixed parameter, see [29]. This type of regularizer favors *binary solutions*, i.e., functions $\bar{u}$ with $\bar{u}(x) \in \{-\bar{\lambda}, \bar{\lambda}\}$ for a.e. $x \in \Omega$ and some $\bar{\lambda} \geq 0$. Problems of this form appear, e.g., in the optimal maneuvering of spacecrafts [8]. Due to the nonsmoothness of the L^∞-norm, previous solution approaches [29, 46] were, e.g., based on a regularized semi-smooth Newton method for a bilinear reformulation of (4.21) [29] or on multiple shooting and parametric methods [46]. We now show that FC-GCG yields a simple algorithm which eventually solves (4.21) without regularizing and/or reformulating the problem. Notably, FC-GCG does not require an a priori discretization of the space domain Ω, since Step 2 in Algorithm 1 is explicitly realized by computing the sign of the dual variable $p_k = -K_* \nabla F(Ku_k)$.

For this purpose, we first note that $L^1(\Omega)^* \simeq \mathcal{M}$ where $L^1(\Omega)$ is equipped with the canonical norm $\|\cdot\|_1$ and where the duality pairing is realized by

$$\langle p, u \rangle = \int_\Omega p(x) u(x) \, \mathrm{d}x \quad \text{for all } p \in L^1(\Omega), \ u \in L^\infty(\Omega).$$

If $K : L^\infty(\Omega) \to Y$ is a linear weak*-to-weak continuous operator and $F : Y \to \mathbb{R}$ is a convex fidelity satisfying (A1), one can easily verify that the minimization problem (4.21) satisfies Assumptions (A1)–(A3). As in the previous examples, we now characterize the extremal points of the unit ball of the regularizer, i.e., the set

$$B := \{ u \in L^\infty(\Omega) \mid \alpha \|u\|_\infty \leq 1 \}.$$

Subsequently, we use this characterization to describe minimizers of (4.21). These are the subjects of the following Propositions 4.12, 4.13, the proofs of which can be found in Sects. B.3.1, B.3.2 of the appendix, respectively.

Proposition 4.12 *For any $\alpha > 0$ we have*

$$\mathrm{Ext}(B) = \left\{ u \in L^\infty(\Omega) \,\middle|\, \alpha|u| = 1 \ \text{a.e. in } \Omega \right\}. \tag{4.22}$$

Moreover there holds $\mathcal{B} = B$.

Proposition 4.13 *An element $\bar{u} \in L^{\infty}(\Omega)$ is a minimizer of* (4.21) *if and only if the associated dual variable $\bar{p} := -K_* \nabla F(K\bar{u}) \in L^1(\Omega)$ satisfies*

$$\int_{\Omega} |\bar{p}(x)| \, \mathrm{d}x \leq \alpha \,, \quad \int_{\Omega} \bar{p}(x) \, \bar{u}(x) \, \mathrm{d}x = \alpha \|\bar{u}\|_{\infty} \,. \tag{4.23}$$

In particular, if $\bar{u} \neq 0$ is a solution to (4.21)*, then we have*

$$\int_{\Omega} |\bar{p}(x)| \, \mathrm{d}x = \alpha, \quad \bar{u}(x) \in \begin{cases} \{\bar{\lambda}\} & \text{if } \bar{p}(x) > 0 \,, \\ \{-\bar{\lambda}\} & \text{if } \bar{p}(x) < 0 \,, \\ [-\bar{\lambda}, \bar{\lambda}] & \text{else} \,, \end{cases} \tag{4.24}$$

for some $\bar{\lambda} > 0$ and a.e. $x \in \Omega$.

Now, let u_k denote the k-th iterate in Algorithm 1 and denote by $p_k = -K_* \nabla F$ (Ku_k) the current dual variable. As in Proposition 4.13, we readily verify

$$\max_{v \in \mathrm{Ext}(B)} \langle p_k, v \rangle = \max_{v \in B} \langle p_k, v \rangle = \alpha^{-1} \int_{\Omega} |p_k(x)| \, \mathrm{d}x \,,$$

where the maximum is realized for $v_k^u = \mathrm{sign}(p_k)$. Hence, realizing Step 2 in Algorithm 1 only requires the computation of the sign of p_k. Next, assume that there is $C > 0$ with:

(E1) For all $y \in Y$ we have $\|K_* y\|_{\infty} \leq C \|y\|_Y$.
(E2) Denoting by $\mathcal{L}$ the d-dimensional Lebesgue measure, there holds

$$\mathcal{L}\left(\{|\bar{p}| \leq \varepsilon\}\right) := \mathcal{L}\left(\{x \in \Omega \mid -\varepsilon \leq \bar{p}(x) \leq \varepsilon\}\right) \leq C\varepsilon, \quad \text{for all } \varepsilon > 0 \,.$$

We remark that Assumption (E2) has already been considered in the context of bang-bang optimal control, i.e. the L^{∞}-norm constrained setting, see, e.g., [27]. Such assumption in particular implies that $\mathcal{L}(\{|\bar{p}| = 0\}) = 0$, and thus

$$\bar{u} := \frac{\bar{\lambda}}{\alpha} \left(\chi_{\{\bar{p}>0\}} - \chi_{\{\bar{p}<0\}}\right) \,, \tag{4.25}$$

for some $\bar{\lambda} \geq 0$, thanks to Proposition 4.13. Under these prerequisites, we show in the next theorem that Assumptions (B2) and (B5) are satisfied for the choice of $g(u, v) := \|u - v\|_1$ and $\bar{U}_1 := B$.

Theorem 4.14 *Let Assumptions (E1) and (E2) hold. Then we have*

$$\arg\max_{v \in \mathrm{Ext}(B)} \langle \bar{p}, v \rangle = \arg\max_{v \in B} \langle \bar{p}, v \rangle = \{\alpha^{-1} \mathrm{sign}(\bar{p})\} \,. \tag{4.26}$$

Moreover, setting $\bar{u} = \alpha^{-1} \mathrm{sign}(\bar{p})$, there holds

$$\|K(u - \bar{u})\|_Y \leq \tau \|u - \bar{u}\|_1, \quad 1 - \langle \bar{p}, u \rangle \geq \kappa \|u - \bar{u}\|_1^2 \,,$$

for all $u \in B$, and some $\tau, \kappa > 0$.

Proof The statement in (4.26) follows immediately from $\mathcal{L}(\{|\bar{p}| = 0\}) = 0$. The Lipschitz estimate on K is immediately deduced from Assumption (E1), while the quadratic growth estimate can be established as in [27, Proposition 2.7]. □

Therefore, Assumptions (E1)–(E2) guarantee fast convergence for Algorithm 1, provided the operators F and K are chosen to satisfy (B1),(B3), and (B4) holds for the unique minimizer of (4.21).

4.4 Dynamic inverse problems regularized with the Benamou-Brenier energy

Consider the space $M(\Omega)$ of Radon measures on the open bounded domain $\Omega \subset \mathbb{R}^n$, $n \geq 1$, and denote by $X := [0, 1] \times \Omega$ the time-space domain. The aim of dynamic inverse problems is to reconstruct a time-dependent curve of measures $t \mapsto \rho_t$ with $\rho_t \in M(\Omega)$ and $t \in [0, 1]$, starting from noisy ill-posed measurements $K\rho$, where $\rho = dt \otimes \rho_t \in M(X)$ and $K : M(X) \to Y$ is a linear operator, with Y a Hilbert space. In [17] it has been proposed to reconstruct the dynamic data ρ by solving a variational inverse problem regularized with a coercive version of the Benamou-Brenier energy $J_{\alpha,\beta}$, allowing for a correlation between the measurements at different time instants (see also [22]). Such variational models have been applied, for instance, to dynamic cell imaging in PET [55], to 4d image reconstruction in nanoscopy [22] and to particle image velocimetry methods [54]. A general variant of the problem considered in [17] can be formulated in our setting by considering

$$\min_{(\rho,m)\in\mathcal{M}} F(K\rho) + J_{\alpha,\beta}(\rho, m), \tag{4.27}$$

where $\mathcal{M} := M(X) \times M(X; \mathbb{R}^d)$, $K : M(X) \to Y$ is weak*-to-weak continuous and such that (A3) holds, while $F : Y \to \mathbb{R}$ is a convex fidelity term satisfying (A1). The Benamou-Brenier energy [9] can be regarded as a dynamic version of optimal transport, and is defined as follows. Setting $\Phi : \mathbb{R} \times \mathbb{R}^d \to [0, \infty]$ as

$$\Phi(t, x) := \frac{|x|^2}{2t} \quad \text{if } t > 0, \quad \Phi(t, x) := 0 \quad \text{if } t = |x| = 0, \tag{4.28}$$

and $\Phi := \infty$ otherwise, the Benamou-Brenier energy is defined for $(\rho, m) \in \mathcal{M}$ by the formula

$$J(\rho, m) := \int_X \Phi\left(\frac{d\rho}{d\sigma}, \frac{dm}{d\sigma}\right) d\sigma, \tag{4.29}$$

where $\sigma \in M^+(X)$ is an arbitrary measure such that $\rho, |m| \ll \sigma$. We refer the reader to [3, 53] for more details. Following [17], the regularizer $J_{\alpha,\beta}$ is then defined by

$$J_{\alpha,\beta}(\rho, m) := \beta J(\rho, m) + \alpha \|\rho\|_{M(X)} + I_{\mathcal{D}}(\rho, m), \tag{4.30}$$

where $\alpha, \beta > 0$, and $\mathcal{D}$ is the set of pairs $(\rho, m) \in \mathcal{M}$ satisfying the continuity equation $\partial_t \rho + \operatorname{div} m = 0$ in X in the weak sense.

The functional $J_{\alpha,\beta}$ is convex, weak* lower semicontinuous, positively one-homogeneous and satisfies (A2), see [15, Lemma 4]. Denote by B the unit ball of $J_{\alpha,\beta}$. In [15, Theorem 6] it has been shown that

$$\mathrm{Ext}(B) = \{(0,0)\} \cup \mathfrak{C}_{\alpha,\beta},$$

where $\mathfrak{C}_{\alpha,\beta}$ is the set of measures concentrated on absolutely continuous curves in Ω, i.e., pairs (ρ_γ, m_γ) satisfying

$$\rho_\gamma = a_\gamma \, \mathrm{d}t \otimes \delta_{\gamma(t)}, \quad m_\gamma = \dot{\gamma}(t)\rho_\gamma, \quad a_\gamma = \left(\frac{\beta}{2}\int_0^1 |\dot{\gamma}(t)|^2 \, \mathrm{d}t + \alpha\right)^{-1}, \quad (4.31)$$

for some $\gamma : [0,1] \to \Omega$ an absolutely continuous curve with weak derivative in L^2. Therefore, it is possible to apply the FC-GCG method of Algorithm 1 for computing solutions to (4.27). In this setting the iterates are of the form $u_k = (\rho_k, m_k)$ with

$$\rho_k = \sum_{i=1}^{N_k} \lambda_i^k \rho_{\gamma_i^k}, \quad m_k = \sum_{i=1}^{N_k} \lambda_i^k m_{\gamma_i^k},$$

with $\lambda_i^k \geq 0$ and $\gamma_i^k : [0,1] \to \Omega$ absolutely continuous. Moreover it can be shown, see [16], that Step 2 in Algorithm 1 is equivalent to solving

$$\widehat{\gamma}_k \in \underset{\gamma \, : \, [0,1] \to \Omega}{\arg\max} \; a_\gamma \int_0^1 p_k(t, \gamma(t)) \, \mathrm{d}t, \quad (4.32)$$

where $p_k = -K_* \nabla F(K\rho_k) \in C(X)$ is the dual variable at the k-th iteration. The new point inserted is then $\widehat{v}_k = (\rho_{\widehat{\gamma}_k}, m_{\widehat{\gamma}_k})$. Thanks to the validity of Assumption (A1)–(A3), Theorem 3.3 guarantees the sublinear convergence of Algorithm 1. On the other hand, the verification of hypotheses (B1)–(B5), necessary for ensuring fast convergence of Algorithm 1, is non-trivial and is left to future work. We remark that an implementable version of Algorithm 1 for solving (4.27) under specific choices of the fidelity term F and the operator K has been recently proposed in [16] (see also [38]). We refer to these papers for more details about the practical implementation and the modifications needed to deal with time dependent measurement operators. Similarly to [16], solving (3.1) amounts to computing an absolutely continuous curve γ by solving (4.32) at every iteration of the algorithm. This is a challenging non-concave variational problem in the space of curves. In [16], the authors proposed to solve (4.32) by a multistart gradient descent approach in the space of curves. The initialization curves are chosen to be a linear interpolation of randomly generated points $\{x^h\}$ in Ω. Moreover several heuristic rules are employed to optimally select $\{x^h\}$ and to reduce the computational time of the multistart gradient descent routine. Together with further acceleration strategies, this method is shown to be computationally feasible and very accurate for the task of tracking several dynamic sources in presence of high noise and severe spatial undersampling. In [38], the authors proposed to speed up the

algorithm in [16] by considering inexact subproblems (3.1) and solving them using known algorithms for computing shortest paths on directed acyclic graphs.

5 A lifting to a space of measures

Our approach to proving Theorem 3.3 and Theorem 3.8 relies on the observation of *Choquet's Theorem*. This classical result allows us to prove that for every $u \in \mathrm{dom}(\mathcal{G})$ there exists a positive measure μ concentrated on $\mathrm{Ext}(B)$ such that $\mathcal{G}(u) = \|\mu\|_{M(\mathcal{B})}$ and

$$\langle p, u \rangle = \int_{\mathcal{B}} \langle p, v \rangle \, \mathrm{d}\mu(v) \quad \text{for all} \quad p \in \mathcal{C}, \tag{5.1}$$

see Proposition 5.2. Motivated by this observation, we study the auxiliary problem $(\mathcal{P}_{M+})$, which we remind the reader is of the form

$$\inf_{\mu \in M^+(\mathcal{B})} \left[F(\mathcal{K}\mu) + \|\mu\|_{M(\mathcal{B})} \right],$$

where the infimum is taken over $M^+(\mathcal{B})$, the cone of positive measures on $\mathcal{B}$. Note that the forward operator is replaced by a "lifted" mapping $\mathcal{K} \colon M(\mathcal{B}) \to Y$ which satisfies $\mathcal{K}\mu = Ku$ whenever (5.1) holds for $\mu \in M^+(\mathcal{B})$ and $u \in \mathcal{M}$, see Proposition 5.3. Moreover, the role of the non-smooth regularizer $\mathcal{G}$ is now played by the total variation norm. It turns out (see Sect. 5.2) that Problems $(\mathcal{P}_{\mathcal{M}})$ and $(\mathcal{P}_{M+})$ are equivalent in the sense that every minimizer to $(\mathcal{P}_{\mathcal{M}})$ can be converted to a solution of $(\mathcal{P}_{M+})$ (and vice versa). Subsequently, in Sect. 5.5, we propose an extension of the Primal-Dual-Active-Point method from [52] to compute a solution of $(\mathcal{P}_{M+})$. This is described in Algorithm 2. Finally, using the results of Sect. 5.2, we show that Algorithm 2 and Algorithm 1 are equivalent, setting the ground to prove Theorem 3.3 and Theorem 3.8 by means of convergence results for Algorithm 2.

5.1 Measure theory notation

Let $C(\mathcal{B})$ denote the vector space of real valued bounded continuous functions over $\mathcal{B}$, which we equip with the supremum norm

$$\|P\|_{C(\mathcal{B})} := \max_{u \in \mathcal{B}} |P(u)|,$$

making it a Banach space. Following the definitions of [2], we denote by Π the σ-algebra of Borel sets on $\mathcal{B}$ formed with respect to the topology induced by $d_{\mathcal{B}}$. A finite Radon measure on $\mathcal{B}$ is a σ-additive map $\mu \colon \Pi \to \mathbb{R}$. We call μ positive if $\mu(\Pi) \subset [0, \infty)$. Given a finite Radon measure μ its total variation measure $|\mu|$ is defined by

$$|\mu|(E) := \sup \left\{ \sum_{i=1}^{\infty} |\mu(E_i)| \;\middle|\; \bigcup_{i=1}^{\infty} E_i = E, \; E_i \in \Pi \text{ pairwise disjoint} \right\},$$

for all $E \in \Pi$. Note that $|\mu|$ is always a positive finite Radon measure. The set of finite Radon measures over $\mathcal{B}$ is a vector space denoted by $M(\mathcal{B})$, which becomes a Banach space when endowed with the total variation norm

$$\|\mu\|_{M(\mathcal{B})} := |\mu|(\mathcal{B}).$$

We recall that $M(\mathcal{B})$ is the dual space of $C(\mathcal{B})$ with respect to $\|\cdot\|_{C(\mathcal{B})}$. The duality pairing will be denoted by $\langle\!\langle \cdot, \cdot \rangle\!\rangle$. The set $M^+(\mathcal{B})$ of all positive finite Radon measures on $\mathcal{B}$ forms a weak* closed cone. For $\mu \in M(\mathcal{B})$ and $E \in \Pi$, we say that μ is concentrated on E if $\mu(\mathcal{B} \setminus E) = 0$. For a set $E' \in \Pi$ we define the restriction of μ to E' as the measure $\mu \llcorner E'$ such that $\mu \llcorner E'(E) = \mu(E' \cap E)$ for all $E \in \Pi$. Finally, the support of a measure $\mu \in M(\mathcal{B})$, denoted by $\operatorname{supp} \mu$, is the closure of the set of all points $v \in \mathcal{B}$ such that $|\mu|(U) > 0$ for all neighbourhoods U of v.

5.2 An equivalent problem

We require the following definition.

Definition 5.1 We say that a measure $\mu \in M^+(\mathcal{B})$ *represents* $u \in \mathcal{M}$ if

$$\langle p, u \rangle = \int_{\mathcal{B}} \langle p, v \rangle \, d\mu(v), \quad \text{for all} \quad p \in \mathcal{C}. \tag{5.2}$$

An element $u \in \mathcal{M}$ such that (5.2) holds is also called the *(weak) barycenter* of μ in $\mathcal{B}$.

It turns out that each $u \in \operatorname{dom}(\mathcal{G})$ is the barycenter of at least one measure $\mu \in M^+(\mathcal{B})$.

Proposition 5.2 *There exists a surjective linear map* $\mathcal{I} \colon M^+(\mathcal{B}) \to \operatorname{dom}(\mathcal{G})$ *such that*

$$\langle p, \mathcal{I}(\mu) \rangle = \int_{\mathcal{B}} \langle p, v \rangle \, d\mu(v), \quad \text{for all} \quad p \in \mathcal{C}. \tag{5.3}$$

In particular, μ *represents* $\mathcal{I}(\mu)$ *in the sense of* (5.2) *and it holds:*

(i) For any $\mu \in M^+(\mathcal{B})$ *we have* $\mathcal{G}(\mathcal{I}(\mu)) \leq \|\mu\|_{M(\mathcal{B})}$.
(ii) For any $u \in \operatorname{dom}(\mathcal{G})$ *there exists* $\mu \in M^+(\mathcal{B})$ *concentrated on* $\operatorname{Ext}(B)$ *such that*

$$u = \mathcal{I}(\mu) \quad \text{and} \quad \mathcal{G}(u) = \|\mu\|_{M(\mathcal{B})}.$$

(iii) $\mathcal{I}$ *is weak*-to-weak* continuous.*

For a proof of the above statement we refer the reader to Appendix C.1. Here we only mention that existence of the map $\mathcal{I}$ is easily obtained by the theory of weak* integration, while surjectivity is a consequence of the classic Choquet's Theorem C.1.

Thus, instead of solving ($\mathcal{P}_{\mathcal{M}}$) directly we can equivalently determine a measure $\bar{\mu} \in M^+(\mathcal{B})$ which represents any of its minimizers. Again, this can be done by solving a suitable minimization problem. To make this idea more rigorous we argue the existence of a linear continuous operator $\mathcal{K} \colon M(\mathcal{B}) \to Y$ which agrees with K on measures representing points of $\mathcal{M}$. Again, we postpone the proof of such statement to Appendix C.1.

Proposition 5.3 *There exists a linear continuous operator $\mathcal{K} \colon M(\mathcal{B}) \to Y$ such that*

$$(\mathcal{K}\mu, y)_Y = \int_{\mathcal{B}} (Kv, y)_Y \, d\mu(v), \quad \text{for all} \quad y \in Y, \ \mu \in M(\mathcal{B}). \tag{5.4}$$

Moreover $\mathcal{K}$ satisfies the following properties:

(i) *The norm of $\mathcal{K}$ is such that*

$$\|\mathcal{K}\|_{\mathcal{L}(M(\mathcal{B}),Y)} \leq C \|K\|_{\mathcal{L}(\mathcal{M},Y)}, \quad C := \sup_{v \in \mathcal{B}} \|v\|_{\mathcal{M}}.$$

(ii) *For every $\mu \in M^+(\mathcal{B})$ representing $u \in \mathcal{M}$, there holds $\mathcal{K}\mu = Ku$.*
(iii) *$\mathcal{K}$ is weak*-to-strong continuous.*
(iv) *$\mathcal{K}$ is the adjoint operator of $\mathcal{K}_* \colon Y \to C(\mathcal{B})$, where*

$$[\mathcal{K}_* y](v) := \langle K_* y, v \rangle \quad \text{for all} \quad v \in \mathcal{B}.$$

If in addition Y is separable, then (5.4) holds in the strong sense, that is,

$$\mathcal{K}\mu = \int_{\mathcal{B}} Kv \, d\mu(v), \tag{5.5}$$

for all $\mu \in M(\mathcal{B})$, where the right hand side integral is in the Bochner sense.

We are now in position to investigate the announced equivalence of ($\mathcal{P}_{\mathcal{M}}$) and the sparse minimization problem ($\mathcal{P}_{M+}$). To this end, define

$$j(\mu) := F(\mathcal{K}\mu) + \|\mu\|_{M(\mathcal{B})},$$

where $\mathcal{K} \colon M(\mathcal{B}) \to Y$ is the operator from Proposition 5.3. The following theorem establishes existence of minimizers for ($\mathcal{P}_{M+}$) and clarifies the connection between ($\mathcal{P}_{\mathcal{M}}$) and ($\mathcal{P}_{M+}$), which is given in terms of the map $\mathcal{I}$ introduced in Proposition 5.2.

Theorem 5.4 *The functional j is weak* lower semicontinuous, has weak* compact sublevels and ($\mathcal{P}_{M+}$) admits at least a solution. In addition, minimizers of ($\mathcal{P}_{\mathcal{M}}$) and ($\mathcal{P}_{M+}$) enjoy the following relationship:*

(i) If $\bar{u} \in \mathcal{M}$ is a minimizer to $(\mathcal{P}_{\mathcal{M}})$, there exists a minimizer $\bar{\mu} \in M^+(\mathcal{B})$ to $(\mathcal{P}_{M^+})$ such that $\bar{u} = \mathcal{I}(\bar{\mu})$.

(ii) Vice versa, if $\bar{\mu} \in M^+(\mathcal{B})$ is an optimal solution to $(\mathcal{P}_{M^+})$ then $\bar{u} := \mathcal{I}(\bar{\mu})$ minimizes in $(\mathcal{P}_{\mathcal{M}})$.

In particular we have that

$$\min_{u \in \mathcal{M}} J(u) = \min_{\mu \in M^+(\mathcal{B})} j(\mu). \tag{5.6}$$

Proof First recall that $\mathcal{K}$ is weak*-to-strong continuous thanks to Proposition 5.3. Given that F is continuous, we then infer weak* lower semicontinuity of j. As F is bounded from below, we immediately have that j is bounded from below and its sublevels are weak* compact, concluding the existence of minimizers to $(\mathcal{P}_{M^+})$ by the direct method.

We pass to the proof of (i). Assume that $\bar{u}$ is a minimizer to $(\mathcal{P}_{\mathcal{M}})$, so that $\bar{u} \in \text{dom}(\mathcal{G})$. According to (ii) in Proposition 5.2 there exists $\bar{\mu} \in M^+(\mathcal{B})$ such that $\mathcal{I}(\bar{\mu}) = \bar{u}$ and $\|\bar{\mu}\|_{M(\mathcal{B})} = \mathcal{G}(\bar{u})$. Let $\mu \in M^+(\mathcal{B})$ be arbitrary and set $u := \mathcal{I}(\mu)$. The same proposition yields that $u \in \text{dom}(\mathcal{G})$, μ represents u and $\mathcal{G}(u) \leq \|\mu\|_{M(\mathcal{B})}$. Finally, point (ii) in Proposition 5.3 guarantees $\mathcal{K}\mu = Ku$ and $\mathcal{K}\bar{\mu} = K\bar{u}$. Thus

$$j(\bar{\mu}) = J(\bar{u}) \leq J(u) \leq j(\mu), \tag{5.7}$$

where the first inequality follows from the optimality of $\bar{u}$. This proves (i). Second, let $\bar{\mu} \in M^+(\mathcal{B})$ be a solution to $(\mathcal{P}_{M^+})$ and $\bar{u} := \mathcal{I}(\bar{\mu})$. Thus, by Proposition 5.2, we have that $\bar{u} \in \text{dom}(\mathcal{G})$, $\bar{\mu}$ represents $\bar{u}$ and $\mathcal{G}(\bar{u}) \leq \|\bar{\mu}\|_{M(\mathcal{B})}$. Moreover, let $u \in \text{dom}(\mathcal{G})$. By point (ii) in Proposition 5.2, we get $\mu \in M^+(\mathcal{B})$ representing u and such that $\|\mu\|_{M(\mathcal{B})} = \mathcal{G}(u)$. Again, $\mathcal{K}\mu = Ku$, $\mathcal{K}\bar{\mu} = K\bar{u}$ and we conclude the proof of (ii) noting that

$$J(\bar{u}) \leq j(\bar{\mu}) \leq j(\mu) = J(u). \tag{5.8}$$

The final part of the statement follows from (5.7)–(5.8). $\qquad\square$

5.3 Optimality conditions

In this section we establish the relationship between the dual variables of the problems $(\mathcal{P}_{\mathcal{M}})$ and $(\mathcal{P}_{M^+})$. Moreover we characterize optimality conditions for $(\mathcal{P}_{M^+})$.

Proposition 5.5 *Let $\mu \in M^+(\mathcal{B})$ be given and set $u := \mathcal{I}(\mu)$. Define the corresponding dual variables $P := -\mathcal{K}_* \nabla F(\mathcal{K}\mu) \in C(\mathcal{B})$ and $p := -K_* \nabla F(Ku) \in \mathcal{C}$. Then*

$$P(v) = \langle p, v \rangle \quad \text{for all} \quad v \in \mathcal{B}. \tag{5.9}$$

Moreover, there exists $\bar{v} \in \text{Ext}(B)$ such that

$$P(\bar{v}) = \max_{v \in \mathcal{B}} P(v) = \max_{v \in \mathcal{B}} \langle p, v \rangle. \tag{5.10}$$

Proof Let $v \in \mathcal{B}$ be arbitrary. It is immediate to check that $\delta_v \in M^+(\mathcal{B})$ represents v in the sense of (5.2). Moreover μ represents u by Proposition 5.2. Applying point (ii) in Proposition 5.3 yields $\mathcal{K}\delta_v = Kv$ and $\mathcal{K}\mu = Ku$. Thus, by (iv) in Proposition 5.3,

$$
\begin{aligned}
\langle p, v \rangle &= -\langle K_* \nabla F(Ku), v \rangle = -(\nabla F(Ku), Kv)_Y \\
&= -(\nabla F(\mathcal{K}\mu), \mathcal{K}\delta_v)_Y = -\langle\!\langle \mathcal{K}_* \nabla F(\mathcal{K}\mu), \delta_v \rangle\!\rangle = P(v),
\end{aligned}
$$

yielding (5.9). The statement of (5.10) now follows from Lemma A.1. $\qquad\square$

Theorem 5.6 *A measure $\bar{\mu} \in M^+(\mathcal{B})$ is a solution of $(\mathcal{P}_{M^+})$ if and only if the dual variable $\bar{P} := -\mathcal{K}_* \nabla F(\mathcal{K}\bar{\mu}) \in C(\mathcal{B})$ satisfies*

$$
\bar{P}(v) \leq 1 \quad \text{for all} \ \ v \in \mathcal{B}, \quad \langle\!\langle \bar{P}, \bar{\mu} \rangle\!\rangle = \|\bar{\mu}\|_{M(\mathcal{B})} . \tag{5.11}
$$

Proof Define the mapping $\Phi \colon M(\mathcal{B}) \to [0, +\infty]$ by $\Phi(\mu) = \|\mu\|_{M(\mathcal{B})} + I_{M^+(\mathcal{B})}(\mu)$ where $I_{M^+(\mathcal{B})}$ denotes the convex indicator function of $M^+(\mathcal{B})$. Note that the map $f(\mu) := F(\mathcal{K}\mu)$ is Gâteaux-differentiable with $f'(\mu)(v) = \langle \mathcal{K}_* \nabla F(\mathcal{K}\mu), v \rangle$ for all $\mu, v \in \mathcal{M}(\mathcal{B})$. Then, by standard subdifferential calculus [19], we have that $\bar{\mu}$ is a solution to $(\mathcal{P}_{M^+})$ if and only if the dual variable $\bar{P}$ satisfies $\bar{P} \in \partial \Phi(\bar{\mu})$ where $\partial \Phi(\bar{\mu})$ is the convex subdifferential of Φ at $\bar{\mu}$. Since Φ is positively 1-homogeneous, it can be easily checked that this inclusion is equivalent to (5.11). $\qquad\square$

5.4 Uniqueness of solutions

As a consequence of Theorem 5.4 and Proposition 3.5 we have the following uniqueness result for solutions to $(\mathcal{P}_{M^+})$, under the stronger Assumptions (B2)–(B3).

Proposition 5.7 *Let Assumptions (A1)–(A3) and (B2)–(B3) hold. Then $(\mathcal{P}_{\mathcal{M}})$ and $(\mathcal{P}_{M^+})$ admit unique solutions given by, respectively,*

$$
\bar{u} = \sum_{i=1}^{N} \bar{\lambda}_i \bar{u}_i, \qquad \bar{\mu} = \sum_{i=1}^{N} \bar{\lambda}_i \delta_{\bar{u}_i},
$$

where the points $\{\bar{u}_i\}_{i=1}^{N} \subset \mathrm{Ext}(B)$ are as in (B2) and $\bar{\lambda}_i \geq 0$.

Proof By Proposition 3.5 we have that $(\mathcal{P}_{\mathcal{M}})$ admits a unique solution $\bar{u}$, which is of the form $\bar{u} = \sum_{i=1}^{N} \bar{\lambda}_i \bar{u}_i$ for some $\bar{\lambda}_i \geq 0$. Let $\bar{\mu} \in M^+(\mathcal{B})$ be a solution to $(\mathcal{P}_{M^+})$, which exists by Theorem 5.4. From the optimality conditions (5.11) in Theorem 5.6, one can easily verify that

$$
\mathrm{supp}\, \bar{\mu} \subset \{v \in \mathcal{B} \mid \bar{P}(v) = 1\}.
$$

Hence, by Assumption (B2) and Proposition 5.5, there exist $\bar{\sigma}_i \geq 0$ such that $\bar{\mu} = \sum_{i=1}^{N} \bar{\sigma}_i \delta_{\bar{u}_i}$. By (5.3) we have $\mathcal{I}(\delta_u) = u$ for all $u \in \mathcal{B}$. As $\mathcal{I}$ is linear, we then conclude $\mathcal{I}(\bar{\mu}) = \sum_{i=1}^{N} \bar{\sigma}_i \bar{u}_i$. On the other hand, by Theorem 5.4 (ii), we know that

$\mathcal{I}(\bar{\mu})$ minimizes in ($\mathcal{P}_{\mathcal{M}}$). Thus $\mathcal{I}(\bar{\mu}) = \bar{u}$, given that $\bar{u}$ is the unique solution of ($\mathcal{P}_{\mathcal{M}}$). We have then shown $\sum_{i=1}^{N}(\bar{\lambda}_i - \bar{\sigma}_i)\bar{u}_i = 0$. By applying the linear operator K to such identity, and invoking (B3), we infer $\bar{\lambda}_i = \bar{\sigma}_i$. Therefore $\bar{\mu} = \sum_{i=1}^{N} \bar{\lambda}_i \delta_{\bar{u}_i}$. As $\bar{\mu}$ is an arbitrary minimizer of ($\mathcal{P}_{M+}$), the thesis is achieved. $\qquad\square$

5.5 A Primal-Dual-Active-Point method for ($\mathcal{P}_{M+}$)

In the following we describe a variant of the Primal-Dual-Active-Point strategy (PDAP) from [52] for the solution of ($\mathcal{P}_{M+}$). The latter is a fully-corrective version of a generalized conditional gradient method (also known as Frank-Wolfe algorithm) for solving convex minimization problems over spaces of measures supported on subsets of the euclidean space. In this section we generalize such procedure to ($\mathcal{P}_{M+}$) and discuss its connection to Algorithm 1. Similarly to [52], our proposed PDAP method alternates between the update of an active set $\mathcal{A}_k^\mu = \{u_i^k\}_{i=1}^{N_k}$ contained in Ext(B), and of a sparse iterate $\mu_k \in M^+(\mathcal{B})$ supported on $\mathcal{A}_k^\mu$, i.e.,

$$\mu_k = \sum_{i=1}^{N_k} \lambda_i^k \, \delta_{u_i^k} \,, \tag{5.12}$$

for some $\lambda_i^k \geq 0$. We now provide a short description of the individual steps of this method and we summarize them in Algorithm 2 below. Given the current iterate μ_k of the form (5.12), we first compute the corresponding dual variable $P_k = -\mathcal{K}_* \nabla F(\mathcal{K}\mu_k) \in C(\mathcal{B})$ and enrich the active set $\mathcal{A}_k^\mu$ by adding to it a global maximizer $\{\widehat{v}_k^\mu\}$ of P_k over Ext(B), i.e., we set

$$\mathcal{A}_k^{\mu,+} = \{u_i^k\}_{i=1}^{N_k} \cup \{\widehat{v}_k^\mu\}, \quad \widehat{v}_k^\mu \in \underset{v \in \text{Ext}(B)}{\arg\max} \, P_k(v) \,.$$

Using Proposition 5.5 we note that this update step is equivalent to maximizing a linear functional over Ext(B). This is the content of the next lemma, whose proof is an immediate consequence of Proposition 5.5, and is hence omitted.

Lemma 5.8 *Let μ_k be as in* (5.12) *and set $u_k = \mathcal{I}(\mu_k)$. Define $P_k := -\mathcal{K}_* \nabla F(\mathcal{K}\mu_k) \in C(\mathcal{B})$ and $p_k := -K_* \nabla F(K u_k) \in \mathcal{C}$. Then, there holds*

$$\underset{v \in \text{Ext}(B)}{\arg\max} \, P_k(v) = \underset{v \in \text{Ext}(B)}{\arg\max} \, \langle p_k, v \rangle \,.$$

Subsequently, setting $N_k^+ := N_k + 1$ and $u_{N_k^+}^k := \widehat{v}_k^\mu$, we find the next iterate μ_{k+1} by solving

$$\min_{\mu \in M^+(\mathcal{A}_k^{\mu,+})} \left[F(\mathcal{K}\mu) + \|\mu\|_{M(\mathcal{B})} \right], \tag{5.13}$$

where the whole cone $M^+(\mathcal{B})$ is replaced by the restricted subset

$$M^+(\mathcal{A}_k^{\mu,+}) := \left\{ \sum_{i=1}^{N_k^+} \lambda_i \delta_{u_i^k} \, \middle| \, \lambda \in \mathbb{R}_+^{N_k^+} \right\} \subset M^+(\mathcal{B}),$$

see Step 5 of Algorithm 2. The following lemma compares the update obtained by solving (5.13) to the finite dimensional minimization problem (3.4) in Step 5 of Algorithm 1.

Lemma 5.9 *A measure $\widehat{\mu} \in M^+(\mathcal{A}_k^{\mu,+})$ is a solution to (5.13) if and only if*

$$\widehat{\mu} = \sum_{i=1}^{N_k^+} \widehat{\lambda}_i \delta_{u_i^k},$$

where $\widehat{\lambda} \in \mathbb{R}_+^{N_k^+}$ is a minimizer of the finite dimensional minimization problem (3.4).

Proof Let $\mu \in M^+(\mathcal{A}_k^{\mu,+})$, so that there exists at least one $\lambda^\mu \in \mathbb{R}_+^{N_k^+}$ such that $\mu = \sum_{i=1}^{N_k^+} \lambda_i^\mu \delta_{u_i^k}$. Noting that $u_i^k = \mathcal{I}(\delta_{u_i^k})$, from Proposition 5.3 we get that $\mathcal{K}\delta_{u_i^k} = Ku_i^k$. Therefore

$$j(\mu) = F(\mathcal{K}\mu) + \|\mu\|_{M(\mathcal{B})} = F\left(\sum_{i=1}^{N_k^+} \lambda_i^\mu Ku_i^k\right) + \sum_{i=1}^{N_k^+} \lambda_i^\mu,$$

from which the characterization of minimizers $\widehat{\mu}$ to (5.13) readily follows. $\qquad\square$

Finally, see Step 6 of Algorithm 2, the active set is truncated by choosing $\mathcal{A}_{k+1}^\mu$ as the *support* of μ_{k+1}, that is,

$$\mathcal{A}_{k+1}^\mu := \operatorname{supp} \mu_{k+1} = \mathcal{A}_k^{\mu,+} \setminus \left\{ u_i^k \in \mathcal{A}_k^{\mu,+} \, \middle| \, \lambda_i^{k+1} = 0 \right\}.$$

The method is summarized in Algorithm 2.

As for Algorithm 1, we define the residuals associated with the iterates μ_k of Algorithm 2 by

$$r_j(\mu_k) := j(\mu_k) - \min_{\mu \in M^+(\mathcal{B})} j(\mu). \tag{5.14}$$

Note that due to Theorem 5.4, such residuals can be written as

$$r_j(\mu_k) = j(\mu_k) - \min_{u \in \mathcal{M}} J(u).$$

Summarizing the previous observations, we conclude the equivalence between Algorithms 1 and 2 as stated in the next theorem.

Algorithm 2 PDAP for $(\mathcal{P}_{M+})$

---

1. Let $\mu_0 = \sum_{i=1}^{N_0} \lambda_i^0 \delta_{u_i^0}$, $\lambda_i^0 > 0$, $\mathcal{A}_0^\mu = \{u_i^0\}_{i=1}^{N_0} \subset \text{Ext}(B)$.

for $k = 0, 1, 2, \ldots$ **do**

  2. Given $\mathcal{A}_k^\mu = \{u_i^k\}_{i=1}^{N_k} \subset \text{Ext}(B)$ and $\mu_k \in M^+(\mathcal{A}_k^\mu)$, calculate P_k and $\widehat{v}_k^\mu$ with

$$P_k = -\mathcal{K}_* \nabla F(\mathcal{K}\mu_k), \quad \widehat{v}_k^\mu \in \underset{v \in \text{Ext}(B)}{\arg\max} \; P_k(v).$$

  if $P_k(\widehat{v}_k^\mu) \leq 1$ and $k \geq 1$ **then**

    3. Terminate with $\bar{\mu} = \mu_k$ a minimizer to $(\mathcal{P}_{M+})$.

  end if

  4. Update $N_k^+ = N_k + 1$, $u_{N_k^+}^k = \widehat{v}_k^\mu$ and $\mathcal{A}_k^{\mu,+} = \mathcal{A}_k^\mu \cup \{\widehat{v}_k^\mu\}$.

  5. Determine μ_{k+1} with

$$\mu_{k+1} \in \underset{\mu \in M^+(\mathcal{A}_k^{\mu,+})}{\arg\min} \; \left[F(\mathcal{K}\mu) + \|\mu\|_{M(\mathcal{B})} \right].$$

  6. Update

$$\mathcal{A}_{k+1}^\mu = \text{supp}\,\mu_{k+1}$$

  and set $N_{k+1} = \#\mathcal{A}_{k+1}^\mu$.

end for

Theorem 5.10 *Let* $\mathcal{A}_k^\mu = \{u_i^k\}_{i=1}^{N_k}$ *and* $\mu_k \in M^+(\mathcal{A}_k^\mu)$ *be given. Set* $\mathcal{A}_k^u := \mathcal{A}_k^\mu$ *and* $u_k := \mathcal{I}(\mu_k)$. *Then, Steps 2 to 6 in Algorithms* 1 *and* 2 *can be realized such that*

$$\widehat{v}_k^u = \widehat{v}_k^\mu, \quad u_{k+1} = \mathcal{I}(\mu_{k+1}), \quad \mathcal{A}_{k+1}^u = \mathcal{A}_{k+1}^\mu.$$

In particular, if $\{u_k\}_k$ *and* $\{\mathcal{A}_k^u\}_k$ *are sequences of iterates and active sets generated by Algorithm* 1 *and we have*

$$u_0 = \mathcal{I}(\mu_0), \quad \mathcal{A}_0^u = \mathcal{A}_0^\mu,$$

then there exist sequences $\{\mu_k\}_k$ *and* $\{\mathcal{A}_k^\mu\}_k$ *generated by Algorithm* 2 *such that*

$$u_k = \mathcal{I}(\mu_k), \quad \mathcal{A}_k^u = \mathcal{A}_k^\mu, \quad \widehat{v}_k^u = \widehat{v}_k^\mu \tag{5.15}$$

and it holds

$$0 \leq r_J(u_k) \leq r_j(\mu_k) \quad \textit{for all } k \in \mathbb{N}, \tag{5.16}$$

where the residuals $r_J(u_k)$ *and* $r_j(\mu_k)$ *are defined in* (3.9) *and* (5.14), *respectively.*

Proof By Lemma 5.8 we can choose $\widehat{v}_k^u = \widehat{v}_k^\mu$. Thus $\mathcal{A}_k^{u,+} = \mathcal{A}_k^{\mu,+} = \{u_i^k\}_{i=1}^{N_k^+}$. Next, in Step 5 of Algorithm 1 we compute a solution $\lambda^{k+1} \in \mathbb{R}_+^{N_k^+}$ to

$$\min_{\lambda \in \mathbb{R}_+^{N_k^+}} \left[F\left(\sum_{i=1}^{N_k^+} \lambda_i K u_i^k \right) + \sum_{i=1}^{N_k^+} \lambda_i \right]$$

and we update u_k as $u_{k+1} = \sum_{i=1}^{N_k^+} \lambda_i^{k+1} u_i^k$. Therefore, defining $\mu_{k+1} = \sum_{i=1}^{N_k^+} \lambda_i^{k+1} \delta_{u_i^k}$ it holds that $u_{k+1} = \mathcal{I}(\mu_{k+1})$ and μ_{k+1} is a solution to (5.13) by Lemma 5.9. Moreover, we note that

$$\mathcal{A}_{k+1}^\mu = \operatorname{supp} \mu_{k+1} = \mathcal{A}_k^{u,+} \setminus \{u_i^k \mid \lambda_i^{k+1} = 0\} = \mathcal{A}_{k+1}^u .$$

Concerning the claim in (5.15), given $u_k = \sum_{i=1}^{N_k^+} \lambda_i^k u_i^k$ and $\mathcal{A}_k^u$, we set $\mu_k = \sum_{i=1}^{N_k^+} \lambda_i^k \delta_{u_i^k}$ and $\mathcal{A}_k^\mu = \mathcal{A}_k^u$. Therefore (5.15) follows by the first part of the statement and an induction argument. Finally, as $\mathcal{I}(\mu_k) = u_k$, by Proposition 5.2 (i) and Proposition 5.3 (ii) we have $\mathcal{G}(u_k) \le \|\mu_k\|_{M(\mathcal{B})}$ and $\mathcal{K}\mu_k = K u_k$. Therefore $J(u_k) \le j(\mu_k)$, and (5.16) follows by (5.6). $\qquad\square$

6 Convergence analysis

We are now prepared to prove Theorems 3.3 and 3.8. For this purpose we rely on Theorem 5.10, which states that the FC-GCG method from Algorithm 1 converges at least as fast as the PDAP method in Algorithm 2, thanks to the estimate (5.16). In particular, by Step 5 in Algorithm 2, we have that the iterates of FC-GCG and PDAP satisfy

$$r_J(u_k) \le r_j(\mu_k) \le r_j(\mu) \quad \text{for all } \mu \in M^+(\mathcal{A}_k^\mu), \ k \in \mathbb{N}. \tag{6.1}$$

In the following we use (6.1), as well as specific choices of the measure μ in the upper bound, to prove Theorem 3.3 and Theorem 3.8. Specifically, the proof of Theorem 3.3 is carried out in Sect. 6.1. The proof of Theorem 3.8, which is more technical, is conducted in Sect. 6.2.10, after establishing some preliminary results in Sect. 6.2.

For the remainder of the paper we silently assume that Algorithm 2 does not stop after a finite number of iterations, and generates a sequence $\{\mu_k\}_k$ in $M^+(\mathcal{B})$. Dropping superscripts, we denote the k-th active set, iterate, dual variable, candidate point computed in Step 2, and enlarged active set by, respectively,

$$\mathcal{A}_k = \{u_i^k\}_{i=1}^{N_k}, \quad \mu_k = \sum_{i=1}^{N_k} \lambda_i^k \delta_{u_i^k},$$

$$P_k = -\mathcal{K}_* \nabla F(\mathcal{K}\mu_k) \in C(\mathcal{B}), \quad \widehat{v}_k \in \arg\max_{v \in \mathcal{B}} P_k(v), \quad \mathcal{A}_k^+ := \mathcal{A}_k \cup \{\widehat{v}_k\},$$

(6.2)

where we recall that $\lambda_i^k > 0$ and $u_i^k, \widehat{v}_k \in \text{Ext}(B)$.

6.1 Worst-case convergence rate

We first argue that Algorithm 2 converges at least sublinearly. To start, define the sublevel set

$$E_{\mu_0} := \left\{ \mu \in M^+(\mathcal{B}) \mid j(\mu) \leq j(\mu_0) \right\}.$$

(6.3)

By Theorem 5.4, we have that E_{μ_0} is weak* compact. Let $M_0 > 0$ be an arbitrary but fixed upper bound on the norm of elements in E_{μ_0} and consider the norm constrained problem

$$\min_{\mu \in M^+(\mathcal{B})} j(\mu) \quad \text{s.t.} \quad \|\mu\|_{M(\mathcal{B})} \leq M_0. \qquad (\widehat{P}_{M^+})$$

Clearly, by definition of M_0, the additional norm constraint does not change the set of global minimizers. The following proposition relates $\widehat{v}_k$ to a particular conditional gradient descent direction η_k for $(\widehat{P}_{M^+})$.

Proposition 6.1 *Define $\eta_k \in M^+(\mathcal{B})$ as*

$$\eta_k := \begin{cases} 0 & \text{if } P_k(\widehat{v}_k) < 1, \\ M_0 \, \delta_{\widehat{v}_k} & \text{otherwise.} \end{cases}$$

Then, η_k is a minimizer of the partially linearized problem

$$\min_{\substack{\eta \in M^+(\mathcal{B}), \\ \|\eta\|_{M(\mathcal{B})} \leq M_0}} [- \langle\!\langle P_k, \eta \rangle\!\rangle + \|\eta\|_{M(\mathcal{B})}].$$

(6.4)

Moreover, we have

$$r_j(\mu_{k+1}) \leq r_j((1-s)\mu_k + s\eta_k) \quad \text{for all } s \in [0, 1].$$

Proof Since we are testing against positive measures, we can estimate

$$\min_{\substack{\eta \in M^+(\mathcal{B}), \\ \|\eta\|_{M(\mathcal{B})} \leq M_0}} [- \langle\!\langle P_k, \eta \rangle\!\rangle + \|\eta\|_{M(\mathcal{B})}] \geq \min_{\substack{\eta \in M^+(\mathcal{B}), \\ \|\eta\|_{M(\mathcal{B})} \leq M_0}} [(1 - \max_{v \in \mathcal{B}} P_k(v)) \|\eta\|_{M(\mathcal{B})}]$$

$$= \begin{cases} 0 & \text{if } P_k(\widehat{v}_k) < 1, \\ (1 - \max_{v \in \mathcal{B}} P_k(v)) M_0 & \text{otherwise,} \end{cases}$$

$$= - \langle\!\langle P_k, \eta_k \rangle\!\rangle + \|\eta_k\|_{M(\mathcal{B})}.$$

The proof is finished noting that $(1 - s)\mu_k + s\eta_k$ belongs to $M^+(\mathcal{A}_k^+)$, and that μ_{k+1} solves (5.13). $\qquad\square$

In particular, Proposition 6.1 shows that, in each iteration, Algorithm 2 achieves at least as much descent as a conditional gradient update. Such observation allows us to prove sublinear convergence for Algorithm 2 using known convergence results for conditional gradient methods in general Banach space (see Theorem 6.2 below). Finally, the combination of Theorem 5.10 with Theorem 6.2 yields the convergence of Algorithm 1.

Theorem 6.2 *Let (A1)–(A3) in Assumption 2.2 hold. Then, the sequence $\{j(\mu_k)\}_k$ is monotone decreasing, $\mu_k \in E_{\mu_0}$, and there exists a constant $c > 0$ such that*

$$r_j(\mu_k) \le c\,\frac{1}{k+1} \quad \text{for all } k \in \mathbb{N}\,, \tag{6.5}$$

where r_j is defined at (5.14). The sequence $\{\mu_k\}_k$ admits at least one weak* accumulation point and each such point is a solution to $(\mathcal{P}_{M+})$. If the solution $\bar\mu$ to $(\mathcal{P}_{M+})$ is unique, we have $\mu_k \overset{*}{\rightharpoonup} \bar\mu$ for the whole sequence.

Proof Since μ_{k+1} is a solution to (5.13) we clearly have $j(\mu_{k+1}) \le j(\mu_k) \le j(\mu_0)$. Thus $\{j(\mu_k)\}_k$ is monotone decreasing and $\mu_k \in E_{\mu_0}$. Now, we show that $\nabla(F \circ \mathcal{K})$ is Lipschitz continuous on E_{μ_0}. Indeed, since E_{μ_0} is weak* compact and $\mathcal{K}$ is weak*-to-strong continuous, see Proposition 5.3 (iii), the image set $\mathcal{K}E_{\mu_0} = \{ \mathcal{K}\mu \mid \mu \in E_{\mu_0} \}$ is compact in Y. Using Assumption (A1), we have that ∇F is Lipschitz continuous on $\mathcal{K}E_{\mu_0}$ for some constant $L_{\mu_0} > 0$. Hence

$$
\begin{aligned}
\|\mathcal{K}_*(\nabla F(\mathcal{K}\mu_1) - \nabla F(\mathcal{K}\mu_2))\|_{C(\mathcal{B})} &\le C\|K\|_{\mathcal{L}(\mathcal{M},Y)}\|\nabla F(\mathcal{K}\mu_1) - \nabla F(\mathcal{K}\mu_2)\|_Y \\
&\le L_{\mu_0}C\|K\|_{\mathcal{L}(\mathcal{M},Y)}\|\mathcal{K}(\mu_1 - \mu_2)\|_Y \\
&\le L_{\mu_0}C^2\|K\|_{\mathcal{L}(\mathcal{M},Y)}^2\|\mu_1 - \mu_2\|_{M(\mathcal{B})}
\end{aligned}
$$

for all $\mu_1, \mu_2 \in E_{\mu_0}$, where C is the constant from Proposition 5.3 (i). The claimed convergence statement now follows from [60, Theorem 6.14]. $\qquad\square$

Proof of Theorem 3.3 Assume that Algorithm 1 does not converge after finitely many steps and generates a sequence $\{u_k\}_k$. According to Theorem 5.10 there exists a sequence $\{\mu_k\}_k$ generated by Algorithm 2 with $u_k = \mathcal{I}(\mu_k)$. Invoking Theorem 5.10 as well as Theorem 6.2 yields

$$r_J(u_k) \le r_j(\mu_k) \le c\,\frac{1}{k+1} \quad \text{for all } k \in \mathbb{N}\,.$$

In particular Theorem 6.2 implies that, up to subsequences, $\mu_k \overset{*}{\rightharpoonup} \bar\mu$ with $\bar\mu$ solution of $(\mathcal{P}_{M+})$. Since $\mathcal{I}$ is weak*-to-weak* continuous by Proposition 5.2 (iii), and since $u_k = \mathcal{I}(\mu_k)$, we infer $u_k \overset{*}{\rightharpoonup} \bar u$ with $\bar u := \mathcal{I}(\bar\mu)$. As $\bar\mu$ minimizes in $(\mathcal{P}_{M+})$, by Theorem 5.4 (ii), we infer that $\bar u$ minimizes in $(\mathcal{P}_{\mathcal{M}})$. The rest of the statement follows by a similar argument. $\qquad\square$

Remark 6.3 In the next section, devoted to the proof of the linear convergence of Algorithm 1, the global sublinear convergence provided by Theorem 3.3 plays an important role. This is because the claimed linear convergence in Theorem 3.8 is only asymptotic, i.e., it relies on the iterates u_k being sufficiently close, in the weak* sense, to a solution $\bar{u}$ of ($\mathcal{P}_\mathcal{M}$). Hence the application of Theorem 3.3 is a necessary starting point.

The proof of Theorem 3.8 relies on the equivalence of the FC-GCG Algorithm to a PDAP method for the lifted problem and follows, conceptually, similar steps as the fast convergence proof in the euclidean setting for the latter [52]. It is then natural to conclude Theorem 3.3 by adopting the interpretation of FC-GCG as an accelerated generalized conditional gradient method for the constrained nonsmooth problem ($\widehat{P}_{M+}$). We point out that the sublinear convergence of Algorithm 1 could be also derived directly, i.e., without relying on the lifting strategy, by adapting the results of [62] for gauge-like regularizers.

6.2 Fast convergence and proof of Theorem 3.8

In this section we further investigate the convergence behaviour of the iterates $\{\mu_k\}_k$ of Algorithm 2, but now under the premise of Assumptions (B1)–(B5). The goal is to show an improved linear local convergence rate for Algorithm 2, see Theorem 6.10 below. Thanks to this result, and to Theorem 5.10, we will then be able to prove linear convergence for the FC-GCG method of Algorithm 1, as stated in Theorem 3.8. As the proofs are quite technical, after establishing some notations we give a detailed summary of the results.

6.2.1 Notation

We employ the notation at (6.2) for μ_k, $\mathcal{A}_k$, $\mathcal{A}_k^+$, $\widehat{v}_k$, P_k. Further, denote by

$$\bar{u} = \sum_{i=1}^{N} \bar{\lambda}_i \bar{u}_i, \quad \bar{\mu} = \sum_{i=1}^{N} \bar{\lambda}_i \delta_{\bar{u}_i}, \quad \bar{\lambda}_i > 0, \quad \bar{u}_i \in \mathrm{Ext}(B),$$

the unique solutions to ($\mathcal{P}_\mathcal{M}$) and ($\mathcal{P}_{M+}$), respectively. The existence and uniqueness are guaranteed by Proposition 5.7, which holds since we are assuming (B2)–(B3), while $\bar{\lambda}_i > 0$ by the non-degeneracy Assumption (B4). Since ($\mathcal{P}_{M+}$) has a unique solution and the prerequisites of Theorem 6.2 are fulfilled, we have the convergences

$$\mu_k \overset{*}{\rightharpoonup} \bar{\mu} \quad \text{weakly* in} \quad M(\mathcal{B}), \quad r_j(\mu_k) \to 0, \tag{6.6}$$

as $k \to \infty$, along the whole sequence. Also, we denote by

$$\bar{\mathcal{A}} := \{\bar{u}_i\}_{i=1}^{N}, \quad \bar{p} := -K_* \nabla F(K\bar{u}) \in \mathcal{C}, \quad \bar{P} := -\mathcal{K}_* \nabla F(\mathcal{K}\bar{\mu}) \in C(\mathcal{B}),$$

the set of optimal extremal points and the optimal dual variables associated to $\bar{u}$ and $\bar{\mu}$, respectively. Moreover, let $\{\bar{U}_i\}_{i=1}^{N}$ be the pairwise disjoint $d_\mathcal{B}$-closed neighborhoods

of $\{\bar{u}_i\}_{i=1}^{N}$, as in Assumption (B5), and set $U_i := \bar{U}_i \cap \mathrm{Ext}(B)$. Finally, the observations are denoted by

$$y_k := \mathcal{K}\mu_k, \quad \bar{y} := K\bar{u} = \mathcal{K}\bar{\mu},$$

where the equality $K\bar{u} = \mathcal{K}\bar{\mu}$ follows from Proposition 5.3 (ii).

6.2.2 Summary of results

Our aim is to show the existence of $\zeta \in [3/4, 1)$ and $c > 0$ such that

$$r_j(\mu_k) \le c\,\zeta^k, \tag{6.7}$$

for all k sufficiently large, see Theorem 6.10 below. To obtain (6.7) it will be sufficient, for fixed k, to construct a measure $\mu^+ \in M(\mathcal{A}_k^+)$ such that

$$r_j(\mu_{k+1}) \le r_j(\mu^+) \le \zeta r_j(\mu_k).$$

We will choose $\mu^+ := \mu_k^s$ as the convex combination

$$\mu_k^s := \mu_k + s(\widehat{\mu}_k - \mu_k), \tag{6.8}$$

for some $s \in [0, 1]$, where the surrogate sequence $\widehat{\mu}_k$ is obtained by suitably modifying μ_k in a neighbourhood of the inserted point $\widehat{v}_k$. More in detail, we proceed as follows:

(i) We start with some preparatory results in Sects. 6.2.3, 6.2.4, 6.2.5. Afterwards, in Sect. 6.2.6, we show that the active sets $\mathcal{A}_k$ cluster around $\bar{\mathcal{A}}$, in the sense that $\mathcal{A}_k \subset \cup_{i=1}^{N} U_i$ for k sufficiently large. Moreover, every point in $\bar{\mathcal{A}}$ is approximated by at least one in $\mathcal{A}_k$. This is a consequence of the uniform convergence of the dual variables P_k to the optimal dual variable $\bar{P}$, see Sect. 6.2.5, as well as the isolation of the maximizers of $\bar{P}$, see Assumption (B2).

(ii) In Sect. 6.2.7, we quantify the distance between $\mathcal{A}_k^+$ and $\bar{\mathcal{A}}$ in terms of the function g in Assumption (B5), see Proposition 6.8. This is a crucial part of our analysis, and also the first point where the growth estimates in (B5) come into play. Precisely, we show that $\mathcal{A}_k \cap U_i$ approaches $\bar{u}_i$ in the sense that

$$\int_{U_i} g(v, \bar{u}_i)\,\mathrm{d}\mu_k(v) \lesssim r_j(\mu_k)^{1/2}, \quad \text{for all } i = 1, \ldots, N. \tag{6.9}$$

In addition, we prove that the inserted point $\widehat{v}_k$ at Step 2 is close to $\bar{\mathcal{A}}$, in the sense that there exists an index $\hat{i}_k \in \{1, \ldots, N\}$ such that $\widehat{v}_k \in U_{\hat{i}_k}$ and

$$g(\widehat{v}_k, \bar{u}_{\hat{i}_k}) \lesssim r_j(\mu_k)^{1/2}. \tag{6.10}$$

(iii) In Sect. 6.2.8 we introduce the surrogate sequence

$$\widehat{\mu}_k := \mu_k(\bar{U}_{\hat{i}_k})\delta_{\widehat{v}_k} + \sum_{i=1,\ i\neq\hat{i}_k}^{N} \mu_k \llcorner \bar{U}_i,$$

obtained by modifying μ_k in the neighbourhood $\bar{U}_{\hat{i}_k}$ of the inserted point $\widehat{v}_k$. In Lemma 6.9 we prove that $\widehat{\mu}_k \overset{*}{\rightharpoonup} \bar{\mu}$, as well as the following key estimate

$$\|\mathcal{K}(\widehat{\mu}_k - \mu_k)\|_Y \lesssim r_j(\mu_k)^{1/2}. \tag{6.11}$$

This estimate relies on (6.9)–(6.10), and thus inherently requires Assumption (B5).

(iv) In Sect. 6.2.9 we employ (6.11) to prove the estimate

$$r_j(\mu_{k+1}) \leq r_j(\mu_k^s) \leq \zeta r_j(\mu_k), \tag{6.12}$$

for some $\zeta \in [3/4, 1)$ and a suitable step size $s \in (0, 1)$, where μ_k^s is defined in (6.8). From (6.12) we then obtain the local linear convergence rate for PDAP, see Theorem 6.10.

(v) In Sect. 6.2.10 we finally prove Theorem 3.8. Thanks to the above analysis, the proof is a simple consequence of the linear convergence rate for PDAP established in Theorem 6.10, and of the link between PDAP and FC-GCG granted by the lifting strategy, see Theorem 5.10.

Remark 6.4 The proof strategy described above is inspired by [52], where the authors propose a method in the spirit of Algorithm 2 for solving a minimization problem in the space of measures supported on a compact set in $\mathbb{R}^n$. However, the proofs in our setting often require novel techniques compared to the ones in [52]. Loosely speaking, this can be attributed to the fact that the metric space $(\mathcal{B}, d_{\mathcal{B}})$ does not posses an obvious geometric structure. For example, in the euclidean setting of [52], the estimates for the distance between $\bar{A}$ and $\mathcal{A}_k$, or $\widehat{v}_k$, respectively, rely on the convexity of euclidean balls, as well as on the higher order differentiability of the dual variable. Such strategy does not extend to our setting, where $d_{\mathcal{B}}$-neighborhoods are generally non-convex and no apparent differentiable structure is available. The difference between both approaches shows most prominently in the key estimate for $\mathcal{K}(\widehat{\mu}_k - \mu_k)$ anticipated in (6.11). In contrast, the fast convergence result in [52] relies on

$$\|\mathcal{K}(\widehat{\mu}_k - \mu_k)\|_Y \lesssim (P_k(\widehat{v}_k) - 1)^{1/2}, \tag{6.13}$$

see [52, Lemma 5.15]. With the notation of the current paper, estimate (6.13) is obtained from a *perturbed quadratic growth condition* of the form

$$\langle\!\langle P_k, \widehat{v}_k - u \rangle\!\rangle \geq \frac{\kappa}{2} g(u, \widehat{v}_k)^2$$

in the vicinity of $\widehat{v}_k$, the derivation of which relies on higher order differentiability of the dual variable. In general though, the quadratic growth condition in Assumption (B5) is *not* stable with respect to perturbations of the dual variable. Thus, such arguments cannot be applied in our setting.

6.2.3 Properties of iterates

Since $\bar{\mu} \neq 0$ and $\|\cdot\|_{M(\mathcal{B})}$ is weak* lower semicontinuous, from (6.6) we conclude the existence of $M \in \mathbb{N}$ such that

$$\mu_k \neq 0, \quad \mathcal{A}_k \neq \emptyset, \quad \text{for all } k \geq M. \tag{6.14}$$

Also, as a consequence of Theorem 6.2

$$\{\mu_k\}_k \subset E_{\mu_0}, \quad \sup_{k \in \mathbb{N}} \|\mu_k\|_{M(\mathcal{B})} \leq M_0, \tag{6.15}$$

where E_{μ_0} is the weak* compact set in (6.3), and $M_0 > 0$ is an upper bound on the norm of elements in E_{μ_0}. In particular $\bar{\mu} \in E_{\mu_0}$ and $\|\bar{\mu}\|_{M(\mathcal{B})} \leq M_0$.

6.2.4 Properties of dual variables

Recall that $\bar{P}$ and $\bar{p}$ are related by $\bar{P}(v) = \langle \bar{p}, v \rangle$ for all $v \in \mathcal{B}$, thanks to Proposition 5.5. In particular, Assumption (B2) reads

$$\arg\max_{v \in \mathcal{B}} \bar{P}(v) = \{v \in \mathcal{B} \mid \bar{P}(v) = \langle \bar{p}, v \rangle = 1\} = \bar{\mathcal{A}}. \tag{6.16}$$

Similarly, we can reformulate Assumption (B5) and Remark 3.7 in terms of $\bar{P}$ to obtain

$$\bar{P}(v) \leq 1 - \sigma \quad \text{for all } v \in \mathcal{B} \setminus \bigcup_{i=1}^{N} \bar{U}_i, \tag{6.17}$$

$$\bar{P}(v) \leq 1 - \kappa g(v, \bar{u}_i)^2 \quad \text{for all } v \in U_i, \tag{6.18}$$

where we remind the reader that $g \colon \text{Ext}(B) \times \text{Ext}(B) \to [0, \infty)$, and $\kappa, \sigma > 0$ are constants. The dual variables P_k satisfy the following optimality conditions. For a proof, see Appendix C.2.1.

Proposition 6.5 *Let M be as in (6.14). Then, for all $k \geq M$ we have*

$$P_k = 1 \text{ on } \mathcal{A}_k, \quad \langle\!\langle P_k, \mu_k \rangle\!\rangle = \|\mu_k\|_{M(\mathcal{B})}, \quad \max_{v \in \mathcal{B}} P_k(v) \geq 1.$$

6.2.5 Convergence of dual variables and observations

Due to the strong convexity of F around $\bar{y}$, see Assumption (B1), the worst-case convergence guarantee of Theorem 6.2 also carries over to the observations y_k and the dual variables P_k, as stated in the following proposition. For a proof, we refer the reader to Appendix C.2.2.

Proposition 6.6 *There exist $M \in \mathbb{N}$ and $c > 0$ such that, for all $k \geq M$, there holds*

$$\|y_k - \bar{y}\|_Y + \|\nabla F(y_k) - \nabla F(\bar{y})\|_Y + \|P_k - \bar{P}\|_{C(\mathcal{B})} \leq c\sqrt{r_j(\mu_k)}. \qquad (6.19)$$

In particular, we have $y_k \to \bar{y}$ in Y and $P_k \to \bar{P}$ in $C(\mathcal{B})$.

6.2.6 Asymptotic behavior of $\mathcal{A}_k$

In the following we show that the active sets $\mathcal{A}_k$ cluster around $\bar{\mathcal{A}}$. Specifically, for k sufficiently large, we have $\mathcal{A}_k \subset \cup_{i=1}^N U_i$, with $\mathcal{A}_k \cap U_i \neq \emptyset$ for all $i = 1, \ldots, N$.

Proposition 6.7 *There exists $M \in \mathbb{N}$ such that for all $k \geq M$ we have*

$$P_k(v) \leq 1 - \sigma/2 \quad \text{for all } v \in \mathcal{B} \setminus \bigcup_{i=1}^N \bar{U}_i, \qquad (6.20)$$

$$\mathcal{A}_k \subset \bigcup_{i=1}^N U_i, \qquad (6.21)$$

where $\sigma > 0$ is the constant in (6.17). Moreover, for all $k \geq M$ and $i = 1, \ldots, N$,

$$\lim_{k \to \infty} \mu_k(\bar{U}_i) = \bar{\lambda}_i, \qquad \mathcal{A}_k \cap U_i \neq \emptyset. \qquad (6.22)$$

Proof As $P_k \to \bar{P}$ uniformly by Proposition 6.6, we deduce the existence of $M \in \mathbb{N}$ such that, for all $k \geq M$, it holds $\|P_k - \bar{P}\|_{C(\mathcal{B})} \leq \sigma/2$. As a consequence, for all $v \in \mathcal{B} \setminus \bigcup_{i=1}^N \bar{U}_i$ and k sufficiently large we have

$$P_k(v) = \bar{P}(v) + P_k(v) - \bar{P}(v)$$
$$\leq 1 - \sigma + \|P_k - \bar{P}\|_{C(\mathcal{B})} \leq 1 - \sigma/2,$$

where we used (6.17). This proves (6.20). Now, recall that $P_k = 1$ on $\mathcal{A}_k$ by Proposition 6.5. Therefore $\mathcal{A}_k \subset \cup_{i=1}^N \bar{U}_i$, otherwise (6.20) would yield a contradiction. Since by construction $\mathcal{A}_k \subset \text{Ext}(B)$ and $U_i := \bar{U}_i \cap \text{Ext}(B)$, we conclude (6.21). Consider now an arbitrary but fixed index l in $\{1, \ldots, N\}$. Recall that the sets $\bar{U}_i$ are pairwise disjoint and $d_\mathcal{B}$-closed. Therefore we can apply Urysohn's lemma to obtain a $d_\mathcal{B}$-continuous function $\varphi_l \colon \mathcal{B} \to [0, 1]$ such that $\varphi_l = 1$ in $\bar{U}_l$, and $\varphi_l = 0$ in

$\bar{U}_i$, $i \neq l$. Recall that $\mu_k \overset{*}{\rightharpoonup} \mu$ along the whole sequence by (6.6). Therefore, since $\bar{\mu} = \sum_{i=1}^{N} \bar{\lambda}_i \delta_{\bar{u}_i}$, we get

$$\bar{\lambda}_l = \langle\!\langle \varphi_l, \bar{\mu} \rangle\!\rangle = \lim_{k \to \infty} \langle\!\langle \varphi_l, \mu_k \rangle\!\rangle = \lim_{k \to \infty} \mu_k(\bar{U}_l),$$

where in the last equality we used (6.21). Finally, since $\bar{\lambda}_l > 0$, from the above convergence we get $\mu_k(\bar{U}_l) > 0$ for k sufficiently large. Recalling that $\mathcal{A}_k \subset \text{Ext}(B)$ and that $U_l := \bar{U}_l \cap \text{Ext}(B)$, we then infer $\mu_k(U_l) > 0$ for k sufficiently large, concluding $\mathcal{A}_k \cap U_l \neq 0$. □

6.2.7 Distance between $\mathcal{A}_k$ and $\bar{\mathcal{A}}$

In the following proposition we use the previous results to quantify the distance between the active set $\mathcal{A}_k$ and the set of optimal extremal points $\bar{\mathcal{A}}$, see (6.23) below. We also provide an estimate for the distance of $\widehat{v}_k$ to the closest element in $\bar{\mathcal{A}}$, see (6.24). We remark that this is the first point in which the estimates of Assumption (B5) are employed.

Proposition 6.8 *There exist $M \in \mathbb{N}$ and a constant $c > 0$ such that for all $k \geq M$ and $i = 1, \ldots, N$ it holds*

$$\sum_{v \in \mathcal{A}_k \cap U_i} \mu_k(\{v\}) g(v, \bar{u}_i) \leq c \sqrt{r_j(\mu_k)}. \tag{6.23}$$

Moreover, for each $k \geq M$ there is an index $\hat{\imath}_k \in \{1, \ldots, N\}$ such that $\widehat{v}_k \in U_{\hat{\imath}_k}$ and

$$g(\widehat{v}_k, \bar{u}_{\hat{\imath}_k}) \leq c \sqrt{r_j(\mu_k)}. \tag{6.24}$$

Proof By minimality of $\bar{\mu}$ in ($\mathcal{P}_{M+}$) and convexity of F we obtain

$$\begin{aligned}
r_j(\mu_k) &= F(\mathcal{K}\mu_k) - F(\mathcal{K}\bar{\mu}) + \|\mu_k\|_{M(\mathcal{B})} - \|\bar{\mu}\|_{M(\mathcal{B})} \\
&\geq (\nabla F(\mathcal{K}\bar{\mu}), \mathcal{K}\mu_k - \mathcal{K}\bar{\mu})_Y + \|\mu_k\|_{M(\mathcal{B})} - \|\bar{\mu}\|_{M(\mathcal{B})} \\
&= \langle\!\langle \bar{P}, \bar{\mu} - \mu_k \rangle\!\rangle + \|\mu_k\|_{M(\mathcal{B})} - \|\bar{\mu}\|_{M(\mathcal{B})} = \|\mu_k\|_{M(\mathcal{B})} - \langle\!\langle \bar{P}, \mu_k \rangle\!\rangle,
\end{aligned} \tag{6.25}$$

where we used the optimality condition (5.11) in the last line. Fix $i_0 \in \{1, \ldots, N\}$ and an arbitrary $k \in \mathbb{N}$ large enough such that all previous results in this section hold. We will show (6.23) for the index i_0. By (6.21) and definition of μ_k we obtain

$$\|\mu_k\|_{M(\mathcal{B})} = \sum_{i=1}^{N} \sum_{v \in \mathcal{A}_k \cap U_i} \mu_k(\{v\}), \tag{6.26}$$

$$\langle\!\langle \bar{P}, \mu_k \rangle\!\rangle = \sum_{i=1}^{N} \sum_{v \in \mathcal{A}_k \cap U_i} \mu_k(\{v\}) \, \bar{P}(v). \tag{6.27}$$

Putting together (6.25)–(6.27), and using (6.18), that is, (B5), we estimate

$$
\begin{aligned}
r_j(\mu_k) &\geq \sum_{i=1}^{N} \sum_{v \in \mathcal{A}_k \cap U_i} \mu_k(\{v\})(1 - \bar{P}(v)) \\
&\geq \kappa \sum_{i=1}^{N} \sum_{v \in \mathcal{A}_k \cap U_i} \mu_k(\{v\})g(v, \bar{u}_i)^2 \geq \kappa \sum_{v \in \mathcal{A}_k \cap U_{i_0}} \mu_k(\{v\})g(v, \bar{u}_{i_0})^2 .
\end{aligned}
$$

Since $\mathcal{A}_k \cap U_{i_0} \neq \emptyset$ by Proposition 6.7, we have $\mu_k(\bar{U}_{i_0}) = \sum_{v \in \mathcal{A}_k \cap U_{i_0}} \mu_k(\{v\}) > 0$. Due to the convexity of $(\cdot)^2$, we conclude

$$
\begin{aligned}
r_j(\mu_k) &\geq \kappa \sum_{v \in \mathcal{A}_k \cap U_{i_0}} \mu_k(\{v\})g(v, \bar{u}_{i_0})^2 \\
&\geq \frac{\kappa}{\mu_k(\bar{U}_{i_0})} \left(\sum_{v \in \mathcal{A}_k \cap U_{i_0}} \mu_k(\{v\})g(v, \bar{u}_{i_0}) \right)^2 .
\end{aligned} \tag{6.28}
$$

Recall $\mu_k(\bar{U}_{i_0}) \leq M_0$ by (6.15). Estimate (6.23) now follows by (6.28) with $c := \sqrt{M_0/\kappa}$.

We now show (6.24). Note that by Proposition 6.5 we have $P_k(\widehat{v}_k) = \max_{v \in \mathcal{B}} P_k(v) \geq 1$ for all $k \in \mathbb{N}$ large enough. Therefore $\widehat{v}_k \in \cup_{i=1}^{N} \bar{U}_i$ by (6.20). Recalling that $\widehat{v}_k \in \text{Ext}(B)$ and that $U_i := \bar{U}_i \cap \text{Ext}(B)$ are pairwise disjoint, we deduce that $\widehat{v}_k \in U_{\hat{\imath}_k}$ for some unique index $\hat{\imath}_k$ in $\{1, \ldots, N\}$. Utilizing (6.18), i.e. (B5), and the fact that $\bar{P}(\bar{u}_{\hat{\imath}_k}) = 1$ by (6.16), we estimate

$$
\begin{aligned}
\kappa\, g(\widehat{v}_k, \bar{u}_{\hat{\imath}_k})^2 &\leq 1 - \bar{P}(\widehat{v}_k) = \bar{P}(\bar{u}_{\hat{\imath}_k}) - \bar{P}(\widehat{v}_k) \\
&\leq \bar{P}(\bar{u}_{\hat{\imath}_k}) - \bar{P}(\widehat{v}_k) + P_k(\widehat{v}_k) - P_k(\bar{u}_{\hat{\imath}_k}) ,
\end{aligned} \tag{6.29}
$$

where in the last line we used that $P_k(\widehat{v}_k) = \max_{v \in \mathcal{B}} P_k(v)$. Using Assumption (B5) as well as Proposition 5.5, the right hand side in (6.29) is further bounded by

$$
\begin{aligned}
\bar{P}(\bar{u}_{\hat{\imath}_k}) &- \bar{P}(\widehat{v}_k) + P_k(v_k) - P_k(\bar{u}_{\hat{\imath}_k}) \\
&= \langle \bar{p} - p_k, \bar{u}_{\hat{\imath}_k} - \widehat{v}_k \rangle \\
&= (\nabla F(\bar{y}) - \nabla F(y_k), K(\widehat{v}_k - \bar{u}_{\hat{\imath}_k}))_Y \\
&\leq \|\nabla F(\bar{y}) - \nabla F(y_k)\|_Y \|K(\widehat{v}_k - \bar{u}_{\hat{\imath}_k})\|_Y \\
&\leq \tau \|\nabla F(\bar{y}) - \nabla F(y_k)\|_Y\, g(\widehat{v}_k, \bar{u}_{\hat{\imath}_k}) ,
\end{aligned} \tag{6.30}
$$

where $\tau > 0$ is the constant from (3.12), which does not depend on k. Finally, using (6.29) and (6.30) together with Proposition 6.6, we conclude (6.24). $\qquad \square$

6.2.8 Surrogate sequence

Let $M \in \mathbb{N}$ be sufficiently large so that all the above results hold. For $k \geq M$ we denote by $\hat{\imath}_k \in \{1, \ldots, N\}$ the index from Proposition 6.8. Starting from the sequence $\{\mu_k\}_k$ generated by Algorithm 2, we define the *surrogate sequence* $\{\widehat{\mu}_k\}_k$ in $M^+(\mathcal{B})$ by setting

$$\widehat{\mu}_k := \mu_k(\bar{U}_{\hat{\imath}_k})\delta_{\widehat{v}_k} + \sum_{i=1,\ i \neq \hat{\imath}_k}^{N} \mu_k \, \llcorner \, \bar{U}_i \,, \tag{6.31}$$

where $\mu_k \, \llcorner \, \bar{U}_i$ is defined as in Sect. 5.1. Notice that $\widehat{\mu}_k$ is just a local modification of μ_k around $\widehat{v}_k$. In the following lemma we investigate the properties of $\widehat{\mu}_k$. Most importantly, we establish the weak* convergence of $\widehat{\mu}_k$ towards $\bar{\mu}$, and the crucial estimate (6.11).

Lemma 6.9 *For all $k \geq M$ there holds*

$$\|\mu_k\|_{M(\mathcal{B})} = \|\widehat{\mu}_k\|_{M(\mathcal{B})}, \tag{6.32}$$

$$\langle\!\langle P_k, \widehat{\mu}_k - \mu_k \rangle\!\rangle = \mu_k(\bar{U}_{\hat{\imath}_k})(P_k(\widehat{v}_k) - 1), \tag{6.33}$$

$$\|\mathcal{K}(\widehat{\mu}_k - \mu_k)\|_Y \leq c\sqrt{r_j(\mu_k)}, \tag{6.34}$$

where $c > 0$ does not depend on k. Moreover, as $k \to \infty$, we have

$$\mathcal{K}\widehat{\mu}_k \to \mathcal{K}\bar{\mu} \quad \text{strongly in } Y, \quad j(\widehat{\mu}_k) \to j(\bar{\mu})\,, \quad \widehat{\mu}_k \overset{*}{\rightharpoonup} \bar{\mu}.$$

In particular, $\widehat{\mu}_k \in E_{\mu_0}$ for all $k \in \mathbb{N}$ large enough, where E_{μ_0} is the set in (6.3).

Proof Recall that $\widehat{v}_k \in U_{\hat{\imath}_k}$ by Proposition 6.8. Using (6.21), the definition of $\widehat{\mu}_k$, and the fact that the sets U_i are pairwise disjoint, it is immediate to check that (6.32) holds. Noting that

$$\widehat{\mu}_k - \mu_k = \mu_k(\bar{U}_{\hat{\imath}_k})\delta_{\widehat{v}_k} - \mu_k \, \llcorner \, \bar{U}_{\hat{\imath}_k}\,, \tag{6.35}$$

we also obtain (6.33), since

$$\langle\!\langle P_k, \widehat{\mu}_k - \mu_k \rangle\!\rangle = \mu_k(\bar{U}_{\hat{\imath}_k})P_k(\widehat{v}_k) - \int_{\bar{U}_{\hat{\imath}_k}} P_k(v)\,\mathrm{d}\mu_k(v)$$

$$= \mu_k(\bar{U}_{\hat{\imath}_k})(P_k(\widehat{v}_k) - 1)\,,$$

where in the last equality we used that $P_k = 1$ on $\mathcal{A}_k = \operatorname{supp}\mu_k$ by Proposition 6.5. We now show (6.34). By (6.21) and (6.35) we compute

$$\widehat{\mu}_k - \mu_k = \sum_{v \in \mathcal{A}_k \cap U_{\hat{\imath}_k}} \mu_k(\{v\})(\delta_{\widehat{v}_k} - \delta_v)\,.$$

Note that $\mathcal{I}(\delta_{\widehat{v}_k} - \delta_v) = \widehat{v}_k - v$. Thus linearity of $\mathcal{K}$, Proposition 5.3 (ii), and triangle inequality imply

$$\|\mathcal{K}(\widehat{\mu}_k - \mu_k)\|_Y \leq \sum_{v \in \mathcal{A}_k \cap U_{\hat{\imath}_k}} \mu_k(\{v\}) \|K(\widehat{v}_k - v)\|_Y .$$

Recalling that $\widehat{v}_k \in U_{\hat{\imath}_k}$, by (B5) and (6.23), (6.24) we estimate

$$\sum_{v \in \mathcal{A}_k \cap U_{\hat{\imath}_k}} \mu_k(\{v\}) \|K(\widehat{v}_k - v)\|_Y$$
$$\leq \sum_{v \in \mathcal{A}_k \cap U_{\hat{\imath}_k}} \mu_k(\{v\})(\|K(\widehat{v}_k - \bar{u}_{\hat{\imath}_k})\|_Y + \|K(v - \bar{u}_{\hat{\imath}_k})\|_Y)$$
$$\leq \tau \sum_{v \in \mathcal{A}_k \cap U_{\hat{\imath}_k}} \mu_k(\{v\})(g(\widehat{v}_k, \bar{u}_{\hat{\imath}_k}) + g(v, \bar{u}_{\hat{\imath}_k}))$$
$$\leq (\mu_k(U_{\hat{\imath}_k}) + 1)\tau c \sqrt{r_j(\mu_k)} \leq (M_0 + 1)\tau c \sqrt{r_j(\mu_k)} ,$$

where in the last inequality we used (6.15). This establishes (6.34).

As for the remaining part of the statement, first recall that $r_j(\mu_k) \to 0$ and $\mu_k \overset{*}{\rightharpoonup} \bar{\mu}$ along the whole sequence by (6.6). Since $\mathcal{K}$ is weak*-to-strong continuous, see Proposition 5.3 (iii), we conclude that $\mathcal{K}\mu_k \to \mathcal{K}\bar{\mu}$ strongly in Y. Thus, $\mathcal{K}\widehat{\mu}_k \to \mathcal{K}\bar{\mu}$ from (6.34). We now show that $j(\widehat{\mu}_k) \to j(\bar{\mu})$. Recalling that $\|\widehat{\mu}_k\|_{M(\mathcal{B})} = \|\mu_k\|_{M(\mathcal{B})}$ by (6.32), we obtain

$$0 \leq r_j(\widehat{\mu}_k) = F(\mathcal{K}\widehat{\mu}_k) + \|\mu_k\|_{M(\mathcal{B})} - j(\bar{\mu})$$
$$\leq r_j(\mu_k) + |F(\mathcal{K}\widehat{\mu}_k) - F(\mathcal{K}\mu_k)| . \tag{6.36}$$

We have $r_j(\mu_k) \to 0$ by Theorem 6.2, while $|F(\mathcal{K}\widehat{\mu}_k) - F(\mathcal{K}\mu_k)| \to 0$ since $\mathcal{K}\widehat{\mu}_k, \mathcal{K}\mu_k \to \mathcal{K}\bar{\mu}$ and F is continuous. Therefore the right hand side of (6.36) converges to zero, which implies $r_j(\widehat{\mu}_k) \to 0$, that is, $j(\widehat{\mu}_k) \to j(\bar{\mu})$. We are left to show that $\widehat{\mu}_k \overset{*}{\rightharpoonup} \bar{\mu}$. Indeed, we have shown that $\{\widehat{\mu}_k\}_k$ is a minimizing sequence for $(\mathcal{P}_{M+})$. Since j is weak* lower semicontinuous and has weak* compact sublevels (Theorem 5.4), we infer the existence of a subsequence $\{\mu_{n_k}\}_k$ and of $\hat{\mu} \in M^+(\mathcal{B})$ such that $\mu_{n_k} \overset{*}{\rightharpoonup} \hat{\mu}$, with $\hat{\mu}$ minimizer of j. By uniqueness of the minimizer, see Proposition 5.7, we conclude that $\mu_{n_k} \overset{*}{\rightharpoonup} \bar{\mu}$. Moreover, since the weak* limit does not depend on the subsequence $\{\mu_{n_k}\}_k$, we also infer $\mu_k \overset{*}{\rightharpoonup} \bar{\mu}$. Finally, since μ_0 is not a minimizer of j, we have $\widehat{\mu}_k \in E_{\mu_0}$ for k sufficiently large. $\qquad\square$

6.2.9 Linear convergence of PDAP

Finally, we are in position to prove the linear convergence of Algorithm 2.

Theorem 6.10 *Assume (A1)–(A3) and (B1)–(B5). Let $\{\mu_k\}_k$ be generated by Algorithm 2. Then, we have $\mu_k \overset{*}{\rightharpoonup} \bar{\mu}$ in $M(\mathcal{B})$. Moreover, there is $\bar{k} \in \mathbb{N}$ and $\zeta \in [3/4, 1)$ such that*

$$r_j(\mu_{k+1}) \leq \zeta\, r_j(\mu_k) \tag{6.37}$$

for all $k \geq \bar{k}$. In particular, there is $c > 0$ with

$$r_j(\mu_k) \leq c\, \zeta^k \tag{6.38}$$

for all $k \in \mathbb{N}$ sufficiently large.

Proof The fact that $\mu_k \overset{*}{\rightharpoonup} \bar{\mu}$ in $M(\mathcal{B})$ along the whole sequence is already established in (6.6). We have to show the improved convergence rate at (6.37). To this end, for a fixed $s \in [0, 1]$ define

$$\mu_k^s := \mu_k + s(\widehat{\mu}_k - \mu_k),$$

with $\widehat{\mu}_k$ as in Definition (6.31). We will obtain (6.37) by estimating the residual of μ_k^s and choosing an optimal value of s. We start by noting that $\mathcal{A}_k \subset \cup_{i=1}^N U_i$ by (6.21), with the sets U_i pairwise disjoint. Therefore $\mu_k = \sum_{i=1}^N \mu_k \llcorner \bar{U}_i$ and

$$\mu_k^s = \left[\sum_{i=1, i \neq \hat{\imath}_k}^N \mu_k \llcorner \bar{U}_i \right] + s\mu_k(\bar{U}_{\hat{\imath}_k})\delta_{\widehat{v}_k} + (1 - s)\mu_k \llcorner \bar{U}_{\hat{\imath}_k}.$$

Since $\widehat{v}_k \in \bar{U}_{\hat{\imath}_k}$, from the above we deduce

$$\|\mu_k^s\|_{M(\mathcal{B})} = \|\mu_k\|_{M(\mathcal{B})} . \tag{6.39}$$

Recall that μ_{k+1} is optimal in (5.13) by definition of Algorithm 2. As $\operatorname{supp} \mu_k^s \subset \mathcal{A}_k^+$, we infer

$$j(\mu_{k+1}) \leq j(\mu_k^s) \quad \text{for all } s \in [0, 1]. \tag{6.40}$$

Next, we estimate the residual of μ_k^s. By (6.39) and regularity of F we infer

$$\begin{aligned}
j(\mu_k^s) - j(\mu_k) &= F(\mathcal{K}\mu_k + s(\mathcal{K}\widehat{\mu}_k - \mathcal{K}\mu_k)) - F(\mathcal{K}\mu_k) \\
&\leq s\langle \nabla F(\mathcal{K}\mu_k), \mathcal{K}\widehat{\mu}_k - \mathcal{K}\mu_k \rangle_Y + R(\mu_k) \\
&= -s\, \langle\!\langle P_k, \widehat{\mu}_k - \mu_k \rangle\!\rangle + R_s(\mu_k),
\end{aligned} \tag{6.41}$$

where $R_s(\mu_k)$ is a remainder defined by

$$R_s(\mu_k) := \int_0^s (\nabla F(\mathcal{K}\mu_k^t) - \nabla F(\mathcal{K}\mu_k), K\widehat{\mu}_k - K\mu_k)_Y \, dt.$$

In order to estimate $R_s(\mu_k)$, first recall that $\{\mu_k\}_k \subset E_{\mu_0}$ by (6.15). Moreover, by Lemma 6.9, we have $\widehat{\mu}_k \in E_{\mu_0}$ for k sufficiently large. Therefore, as E_{μ_0} is convex,

we also have $\mu_k^s \in E_{\mu_0}$ for all $s \in [0, 1]$ and k sufficiently large. Note that the set $\mathcal{K}E_{\mu_0} := \{\mathcal{K}\mu \mid \mu \in E_{\mu_0}\}$ is compact, given that $\mathcal{K}$ is weak*-to-strong continuous, see Proposition 5.3 (iii). Denote by L_{μ_0} the Lipschitz constant of ∇F on the set $\mathcal{K}E_{\mu_0}$, which exists by Assumption (A1). By Cauchy-Schwarz we then estimate

$$
\begin{aligned}
|R_s(\mu_k)| &\le \|\mathcal{K}\widehat{\mu}_k - \mathcal{K}\mu_k\|_Y \int_0^s \|\nabla F(\mathcal{K}\mu_k^t) - \nabla F(\mathcal{K}\mu_k)\|_Y \, dt \\
&\le L_{\mu_0} \|\mathcal{K}\widehat{\mu}_k - \mathcal{K}\mu_k\|_Y \int_0^s \|\mathcal{K}\mu_k^t - \mathcal{K}\mu_k\|_Y \, dt \\
&= s^2 \frac{L_{\mu_0}}{2} \|\mathcal{K}(\widehat{\mu}_k - \mu_k)\|_Y^2 .
\end{aligned}
\tag{6.42}
$$

By minimality of $\bar{\mu}$ and (6.41)–(6.42), we infer

$$
\begin{aligned}
r_j(\mu_k^s) &= j(\mu_k^s) - j(\bar{\mu}) \\
&= j(\mu_k) - j(\bar{\mu}) - s \langle\!\langle P_k, \widehat{\mu}_k - \mu_k \rangle\!\rangle + R_s(\mu_k) \\
&\le r_j(\mu_k) - s \langle\!\langle P_k, \widehat{\mu}_k - \mu_k \rangle\!\rangle + s^2 \frac{L_{\mu_0}}{2} \|\mathcal{K}(\widehat{\mu}_k - \mu_k)\|_Y^2.
\end{aligned}
\tag{6.43}
$$

Recall that by construction $P_k(\widehat{v}_k) = \max_{v \in \mathcal{B}} P_k(v)$. Moreover $\max_{v \in \mathcal{B}} P_k(v) \ge 1$ by Proposition 6.5. Thus $P_k(\widehat{v}_k) \ge 1$ and we can apply Proposition 6.1 to infer that $M_0 \delta_{\widehat{v}_k}$ minimizes in (6.4), that is,

$$
\langle\!\langle P_k, \eta \rangle\!\rangle - \|\eta\|_{M(\mathcal{B})} \le M_0 (P_k(\widehat{v}_k) - 1)
\tag{6.44}
$$

for all $\eta \in M^+(\mathcal{B})$ with $\|\eta\|_{M(\mathcal{B})} \le M_0$. We can now use convexity of F to estimate

$$
\begin{aligned}
r_j(\mu_k) &= F(\mathcal{K}\mu_k) - F(\mathcal{K}\bar{\mu}) + \|\mu_k\|_{M(\mathcal{B})} - \|\bar{\mu}\|_{M(\mathcal{B})} \\
&\le (\nabla F(\mathcal{K}\mu_k), \mathcal{K}(\mu_k - \bar{\mu}))_Y + \|\mu_k\|_{M(\mathcal{B})} - \|\bar{\mu}\|_{M(\mathcal{B})} \\
&= \langle\!\langle P_k, \bar{\mu} - \mu_k \rangle\!\rangle + \|\mu_k\|_{M(\mathcal{B})} - \|\bar{\mu}\|_{M(\mathcal{B})} \\
&= \langle\!\langle P_k, \bar{\mu} \rangle\!\rangle - \|\bar{\mu}\|_{M(\mathcal{B})},
\end{aligned}
$$

where in the last equality we used Proposition 6.5. Since $\|\bar{\mu}\|_{M(\mathcal{B})} \le M_0$, see Sect. 6.2.3, we can apply (6.44) with $\eta = \bar{\mu}$ to obtain

$$
r_j(\mu_k) \le M_0 (P_k(\widehat{v}_k) - 1) .
$$

By (6.33) we then infer

$$
\langle\!\langle P_k, \widehat{\mu}_k - \mu_k \rangle\!\rangle = \mu_k(\bar{U}_{\widehat{i}_k})(P_k(\widehat{v}_k) - 1) \ge \frac{\mu_k(\bar{U}_{\widehat{i}_k})}{M_0} r_j(\mu_k) .
\tag{6.45}
$$

Using (6.45) and (6.34) in (6.43) yields

$$
r_j(\mu_k^s) \le r_j(\mu_k) - s \frac{\mu_k(\bar{U}_{\widehat{i}_k})}{M_0} r_j(\mu_k) + s^2 \frac{L_{\mu_0} c_2}{2} r_j(\mu_k) ,
\tag{6.46}
$$

where $c_2 > 0$ is the square of the constant in (6.34). Define the constant

$$c_1 := \min_{i=1,\ldots,N} \frac{\bar{\mu}(\bar{U}_i)}{2M_0} = \min_{i=1,\ldots,N} \frac{\bar{\lambda}_i}{2M_0}.$$

Notice that $c_1 > 0$, since (B4) holds. Moreover, $c_1 \leq 1/2$ given that $\|\bar{\mu}\|_{M(\mathcal{B})} \leq M_0$. Invoking (6.22) we conclude that

$$\frac{\mu_k(\bar{U}_{\hat{i}_k})}{M_0} \geq \min_{i=1,\ldots,N} \frac{\mu_k(\bar{U}_i)}{M_0} \to \min_{i=1,\ldots,N} \frac{\bar{\lambda}_i}{M_0} = 2c_1, \quad \text{as } k \to \infty,$$

which, together with (6.46) yields

$$r_j(\mu_k^s) \leq \left(1 - sc_1 + s^2 \frac{L_{\mu_0} c_2}{2}\right) r_j(\mu_k),$$

for all k sufficiently large and all $s \in [0, 1]$. Subtracting $j(\bar{\mu})$ from both sides of (6.40) yields

$$r_j(\mu_{k+1}) \leq \varphi(s) r_j(\mu_k), \qquad \varphi(s) := 1 - sc_1 + s^2 \frac{L_{\mu_0} c_2}{2}, \qquad (6.47)$$

for all k sufficiently large and all $s \in [0, 1]$. It is immediate to check that

$$\min_{s \in [0,1]} \varphi(s) = \begin{cases} 1 - c_1 + \dfrac{L_{\mu_0} c_2}{2} & \text{if } c_1 > L_{\mu_0} c_2, \\ 1 - \dfrac{c_1^2}{2L_{\mu_0} c_2} & \text{if } c_1 \leq L_{\mu_0} c_2. \end{cases}$$

Notice that $\min_{s \in [0,1]} \varphi(s) \leq \zeta$, where

$$\zeta := 1 - \frac{c_1}{2} \min\left\{1, \frac{c_1}{L_{\mu_0} c_2}\right\}, \quad \frac{3}{4} \leq \zeta < 1.$$

Thus, from (6.47) we obtain an integer $\bar{k} \in \mathbb{N}$ such that $r_j(\mu_{k+1}) \leq \zeta \, r_j(\mu_k)$ for all $k \geq \bar{k}$, establishing (6.37). In particular, $r_j(\mu_k) \leq r_j(\mu_{\bar{k}}) \zeta^{k-\bar{k}}$ for all $k \geq \bar{k}$, and the final thesis (6.38) follows by setting $c := r_j(\mu_{\bar{k}}) \zeta^{-\bar{k}} > 0$. $\qquad \square$

6.2.10 Proof of Theorem 3.8

Assume that Algorithm 1 does not converge after finitely many steps and generates a sequence $\{u_k\}_k$ in $\mathcal{M}$. By Proposition 5.7 both Problems $(\mathcal{P}_{\mathcal{M}})$ and $(\mathcal{P}_{M+})$ admit a unique solution, given by $\bar{u} = \sum_{i=1}^{N} \bar{\lambda}_i \bar{u}_i$ and $\bar{\mu} = \sum_{i=1}^{N} \bar{\lambda}_i \delta_{\bar{u}_i}$, respectively. Thanks to Theorem 5.10, there exists a sequence $\{\mu_k\}_k$ generated by Algorithm 2 with $u_k = \mathcal{I}(\mu_k)$. Invoking Theorem 5.10 as well as Theorem 6.10 yields $r_J(u_k) \leq r_j(\mu_k) \leq c \, \zeta^k$, for all k sufficiently large, where $c > 0$ and $\zeta \in [3/4, 1)$. This shows (3.14). In addition, Theorem 6.10 ensures that $\mu_k \overset{*}{\rightharpoonup} \bar{\mu}$ along the whole sequence. Recalling

that $\mathcal{I}$ is weak*-to-weak* continuous by Proposition 5.2 (iii), we infer $u_k \overset{*}{\rightharpoonup} \mathcal{I}(\bar{\mu})$. Since $\bar{\mu}$ minimizes in $(\mathcal{P}_{M+})$, by Theorem 5.4 (ii), we infer that $\mathcal{I}(\bar{\mu})$ minimizes in $(\mathcal{P}_\mathcal{M})$. As the minimizer is unique, we conclude $\bar{u} = \mathcal{I}(\bar{\mu})$.

7 Conclusions

We have introduced a fully-corrective generalized conditional gradient method (FC-GCG) to solve a class of non-smooth minimization problems in Banach spaces. For such algorithm we provided a global sublinear rate of convergence under the mild Assumptions (A1)–(A3), as well as a local linear convergence rate under the Assumptions (B1)–(B5). Several examples of applications were considered, showing that it is possible to formulate a problem-dependent natural set of assumptions which is easy to verify and implies Assumptions (B1)–(B5), thus ensuring linear convergence of our method. We demonstrate numerically the fast convergence for our first example, showing that, compared to standard conditional gradient methods (GCG), our algorithm exhibits a vastly improved rate of convergence. Additionally, we have discussed in details the computational burden of each presented example, and we have described viable strategies to solve the (partially) linearized problem (1.1).

As in the euclidean case of [52], we believe that all results in the present paper are transferable to non-smooth regularizers of the form $\Phi(\mathcal{G}(\cdot))$ where $\mathcal{G}$ is as in Assumptions (A1)–(A3), and Φ is a suitable convex, monotone increasing function, e.g., $\Phi(\mathcal{G}(u)) = (1/2)\mathcal{G}(u)^2$. This generalization was omitted in the present paper for the sake of readability.

Appendix A: Complements to Sect. 3

In this section we state and prove two results. The first concerns well-posedness for the minimization problem at (3.1), while the second shows equivalence, in a suitable sense, of (3.2) and (3.4).

Lemma A.1 *Let $p \in C$ be given. Then, there exists $\bar{v} \in \mathrm{Ext}(B)$ with*

$$\langle p, \bar{v} \rangle = \max_{v \in \mathcal{B}} \langle p, v \rangle = \max_{v \in B} \langle p, v \rangle .$$

Proof In order to show the claimed result we first prove that the maximization problem

$$\max_{v \in B} \langle p, v \rangle \tag{A.1}$$

admits a solution $\bar{v} \in \mathrm{Ext}(B)$. This statement is classical, but for the reader's convenience we produce a proof. Let v^* be a maximizer of (A.1), which exists since $v \mapsto \langle p, v \rangle$ is weak* continuous and B is weak* compact. Consider the set $H := \{ v \in B \mid \langle p, v \rangle = \langle p, v^* \rangle \}$. Then $\mathrm{Ext}(H) \neq \emptyset$ by Krein-Milman's Theorem, since $H \subset B$ is convex, non-empty, weak* closed and thus weak* compact. Thus, let $\bar{v} \in \mathrm{Ext}(H)$ be given. We claim that $\bar{v} \in \mathrm{Ext}(B)$. Indeed, assume that $\bar{v} = s v_1 + (1 - s) v_2$ for $v_1, v_2 \in B$, $s \in (0, 1)$. Notice that $v_1, v_2 \in H$, otherwise we would have $\langle p, \bar{v} \rangle < \langle p, v^* \rangle$, contradicting the maximality of $\bar{v}$. As $\bar{v} \in \mathrm{Ext}(H)$ we then have $\bar{v} = v_1 = v_2$, showing that $\bar{v} \in \mathrm{Ext}(B)$. The claimed statement now follows from $\mathrm{Ext}(B) \subset \mathcal{B} \subset B$ ending the proof. $\qquad \square$

Proposition A.2 *Let $\mathcal{A}_k^{u,+} = \{u_i^k\}_{i=1}^{N_k^+} \subset \mathrm{Ext}(B)$ be given. If $\widehat{\lambda} \in \mathbb{R}_+^{N_k^+}$ solves (3.4), then $\widehat{u} = \sum_{i=1}^{N_k^+} \widehat{\lambda}_i u_i^k$ solves (3.2). Conversely, if $\widehat{u}$ solves (3.2), then there exists a solution $\widehat{\lambda} \in \mathbb{R}_+^{N_k^+}$ to (3.4) such that $\widehat{u} = \sum_{i=1}^{N_k^+} \widehat{\lambda}_i u_i^k$.*

Proof Let $\widehat{\lambda} \in \mathbb{R}_+^{N_k^+}$ be a solution to (3.4). Set $\widehat{u} = \sum_{i=1}^{N_k^+} \widehat{\lambda}_i u_i^k$ and fix an arbitrary $u \in \mathrm{cone}(\mathcal{A}_k^{u,+})$. Then there exists $\lambda^u \in \mathbb{R}_+^{N_k^+}$ such that $u = \sum_{i=1}^{N_k^+} \lambda_i^u u_i^k$ with $\sum_{i=1}^{N_k^+} \lambda_i^u = \kappa_{\mathcal{A}_k^+}(u)$. We estimate

$$F(K\widehat{u}) + \kappa_{\mathcal{A}_k^+}(\widehat{u}) \leq F\left(\sum_{i=1}^{N_k^+} \widehat{\lambda}_i K u_i^k\right) + \sum_{i=1}^{N_k^+} \widehat{\lambda}_i$$

$$\leq F\left(\sum_{i=1}^{N_k^+} \lambda_i^u K u_i^k\right) + \sum_{i=1}^{N_k^+} \lambda_i^u = F(Ku) + \kappa_{\mathcal{A}_k^+}(u) ,$$

showing that $\widehat{u}$ is a solution to (3.2). Conversely, let $\widehat{u}$ be a minimizer to (3.2). Since $\widehat{u} \in \mathrm{cone}(\mathcal{A}_k^{u,+})$, there exists $\widehat{\lambda} \in \mathbb{R}_+^{N_k^+}$ with $\widehat{u} = \sum_{i=1}^{N_k^+} \widehat{\lambda}_i u_i^k$, $\sum_{i=1}^{N_k^+} \widehat{\lambda}_i = \kappa_{\mathcal{A}_k^{u,+}}(\widehat{u})$. Now

let $\lambda \in \mathbb{R}_+^{N_k^+}$ be given and set $u^\lambda = \sum_{i=1}^{N_k^+} \lambda_i u_i^k$. From the optimality of $\widehat{u}$ and the definition of the gauge function we get

$$F\left(\sum_{i=1}^{N_k^+} \widehat{\lambda}_i K u_i^k\right) + \sum_{i=1}^{N_k^+} \widehat{\lambda}_i \leq F(Ku^\lambda) + \kappa_{\mathcal{A}_k^{u,+}}(u^\lambda) \leq F\left(\sum_{i=1}^{N_k^+} \lambda_i K u_i^k\right) + \sum_{i=1}^{N_k^+} \lambda_i .$$

Thus, $\widehat{\lambda}$ is a minimizer of (3.4). $\qquad\square$

Appendix B: Complements to Sect. 4

Here we collect the technical statements and the proofs of Sect. 4. For sake of clarity we place the proofs of each example in different subsections.

B.1. Sparse source identification (Sect. 4.1)

The next lemma justifies the well-posedness of the definition of K_*.

Lemma B.1 *Let $\varphi \in L^2(\Omega)$ be given and let $z \in L^2(0, T; H_0^1(\Omega)) \cap H^1(0, T; H^{-1}(\Omega))$ the associated unique solution of (4.3). Moreover, let Ω_0 be a subdomain with $\bar{\Omega}_0 \subset \Omega$. Then, we have $z(0) \in C_0(\Omega) \cap C^2(\Omega_0)$ and*

$$\|z(0)\|_{\mathcal{C}} + \|z(0)\|_{C^2(\Omega_0)} \leq c\|\varphi\|_{L^2(\Omega)}$$

for some $c > 0$ independent of $\varphi \in L^2(\Omega)$.

Proof This follows from [50, Lemma 3.1] and the Sobolev embedding $H^2(\Omega) \cap H_0^1(\Omega) \hookrightarrow C_0(\Omega)$. $\qquad\square$

B.1.1. Proof of Lemma 4.5

We start by noting that $\nabla \bar{z}(0)(\bar{x}_i) = 0$ thanks to (C1) and Proposition 4.2. Now fix $x \in \Omega$. By Taylor's expansion we infer that

$$\begin{aligned}\beta - |\bar{z}(0)(x)| &= |\bar{z}(0)(\bar{x}_i)| - |\bar{z}(0)(x)| \\ &= -(1/2)\operatorname{sign}(\bar{z}(0)(\bar{x}_i))((x - \bar{x}_i), \nabla^2 \bar{z}(0)(x_s)(x - \bar{x}_i))_{\mathbb{R}^2} ,\end{aligned}$$

for some $x_s = (1 - s)\bar{x}_i + sx$, $s \in (0, 1)$. Due to the continuity of $\nabla^2 \bar{z}(0)$, see Lemma B.1, as well as Assumption (C2), there exists $R > 0$ such that $B_R(\bar{x}_i) \subset \Omega$ and

$$\operatorname{sign}(\bar{z}(0)(\bar{x}_i))((x - \bar{x}_i), \nabla^2 \bar{z}(0)(x_s)(x - \bar{x}_i))_{\mathbb{R}^2} \leq -(\gamma/2)|x - \bar{x}_i|^2 ,$$

for all $x \in B_R(\bar{x}_i)$. Combining both observations finally yields

$$\beta - |\bar{z}(0)(x)| \geq (\gamma/4)|x - \bar{x}_i|^2,$$

for all $x \in B_R(\bar{x}_i)$, showing (4.10). Moreover for every $x \in B_R(\bar{x}_i)$ we readily verify

$$
\begin{aligned}
\|K(\delta_x - \delta_{\bar{x}_i})\|_Y &= \sup_{\|y\|_Y \leq 1} (y, K(\delta_x - \delta_{\bar{x}_i}))_Y \\
&= \sup_{\|y\|_Y \leq 1} [K_* y](x) - [K_* y](\bar{x}_i) \\
&\leq \sup_{\|y\|_Y \leq 1} \|K_* y\|_{\mathrm{Lip}(B_R(\bar{x}_i))} |x - \bar{x}_i| \\
&\leq \|K_*\|_{Y,\mathrm{Lip}(B_R(\bar{x}_i))} |x - \bar{x}_i|.
\end{aligned}
$$

To finish, $R > 0$ can be chosen small enough to ensure

$$\mathrm{sign}(\bar{z}(0)(x)) = \mathrm{sign}(\bar{z}(0)(\bar{x}_i)),$$

for all $x \in B_R(\bar{x}_i)$, due to $\bar{z}(0)(\bar{x}_i) \neq 0$ and $\bar{z}(0) \in C_0(\Omega)$. □

B.1.2. Proof of Lemma 4.6

First assume that $g(u_k, \bar{u}_i) \to 0$. Then, we have $x_k \to \bar{x}_i$ in Ω as well as $\sigma_k \to \mathrm{sign}(\bar{z}(0)(\bar{x}_i))$. Consequently, for all $p \in C_0(\Omega)$ we have

$$
\begin{aligned}
\lim_{k\to\infty} \langle p, u_k \rangle &= \lim_{k\to\infty} \sigma_k \beta^{-1} p(x_k) \\
&= \mathrm{sign}(\bar{z}(0)(\bar{x}_i))\beta^{-1} p(\bar{x}_i) = \langle p, \bar{u}_i \rangle.
\end{aligned}
$$

This implies $u_k \overset{*}{\rightharpoonup} \bar{u}_i$. For the other direction assume that $u_k \overset{*}{\rightharpoonup} \bar{u}_i$. Note that from every subsequence of (σ_k, x_k) we can extract a further convergent subsequence still denoted by (σ_k, x_k), relabeling the indices. Let $(\bar{\sigma}, \bar{x}) \in \{-1, 1\} \times \bar{\Omega}$ denote its limit. As $\bar{x} \in \partial\Omega$ would imply $u_k \overset{*}{\rightharpoonup} 0 \neq \bar{u}_i$, we conclude $\bar{x} \in \Omega$. Then, using again the weak* convergence of u_k, we get

$$\bar{\sigma}\beta^{-1} p(\bar{x}) = \mathrm{sign}(\bar{z}(0)(\bar{x}_i))\beta^{-1} p(\bar{x}_i),$$

for all $p \in C_0(\Omega)$, from which we immediately conclude $\bar{\sigma} = \mathrm{sign}(\bar{z}(0)(\bar{x}_i))$ and $\bar{x} = \bar{x}_i$. Since the initial subsequence was chosen arbitrary we get $(\sigma_k, x_k) \to (\mathrm{sign}(\bar{z}(0)(\bar{x}_i)), \bar{x}_i)$ for the whole sequence and thus, $g(u_k, \bar{u}_i) \to 0$. Finally, since $\sigma_k \in \{-1, 1\}$ for all $k \in \mathbb{N}$ and $\sigma_k \to \mathrm{sign}(\bar{z}(0)(\bar{x}_i))$, we necessarily have $\sigma_k = \mathrm{sign}(\bar{z}(0)(\bar{x}_i))$ for all $k \in \mathbb{N}$ sufficiently large. □

B.2 Rank-one matrix reconstruction by trace regularization (Sect. 4.2)

B.2.1. Proof of Lemma 4.8

Recall that every extremal point $\mathcal{U}$ of B necessarily satisfies $\mathrm{Tr}(\mathcal{U}) = \beta^{-1}$ or $\mathrm{Tr}(\mathcal{U}) = 0$. We first focus on those extremal points with $\mathrm{Tr}(\mathcal{U}) = \beta^{-1}$. According to Lidskii's theorem, for every $\mathcal{U} \in B$ there holds $\mathrm{Tr}(\mathcal{U}) = \sum_{i \in \mathbb{N}} \sigma_i^{\mathcal{U}}$. From these observations, we already see that every extremal point of B has at most rank one. Indeed, if $\mathcal{U} \in B, \mathrm{Tr}(\mathcal{U}) = \beta^{-1}$, is at least of rank two, i.e. $\sigma_1^{\mathcal{U}} \geq \sigma_2^{\mathcal{U}} > 0$, then we can decompose $\mathcal{U}$ as

$$
\mathcal{U} = (\beta \sigma_1^{\mathcal{U}})\beta^{-1} h_1^{\mathcal{U}} \otimes h_1^{\mathcal{U}} + \left(\beta \sum_{j \in \mathbb{N}, j \neq 1} \sigma_j^{\mathcal{U}} \right) \sum_{i \in \mathbb{N}, i \neq 1} \left(\beta \sum_{j \in \mathbb{N}, j \neq 1} \sigma_j^{\mathcal{U}} \right)^{-1} \sigma_i^{\mathcal{U}} h_i^{\mathcal{U}} \otimes h_i^{\mathcal{U}}
$$

$$
= (\beta \sigma_1^{\mathcal{U}})\beta^{-1} h_1^{\mathcal{U}} \otimes h_1^{\mathcal{U}} + (1 - \beta \sigma_1^{\mathcal{U}}) \sum_{i \in \mathbb{N}, i \neq 1} \left(\beta \sum_{j \in \mathbb{N}, j \neq 1} \sigma_j^{\mathcal{U}} \right)^{-1} \sigma_i^{\mathcal{U}} h_i^{\mathcal{U}} \otimes h_i^{\mathcal{U}}
$$

$$
:= (\beta \sigma_1^{\mathcal{U}})\mathcal{U}_1 + (1 - \beta \sigma_1^{\mathcal{U}})\mathcal{U}_2
$$

where $(\beta \sigma_1^{\mathcal{U}}) \in (0, 1)$ and the second equality follows from

$$
\beta^{-1} = \mathrm{Tr}(\mathcal{U}) = \sum_{i \in \mathbb{N}} \sigma_i^{\mathcal{U}}.
$$

Noting that $\mathcal{U}_1, \mathcal{U}_2 \in B$, as well as $\mathcal{U}_1 \neq \mathcal{U}_2$, we conclude that $\mathcal{U}$ is not extremal. Hence, if $\mathcal{U} \in B, \mathcal{G}(\mathcal{U}) = 1$, is extremal then we have $\mathcal{U} = \beta^{-1} h \otimes h, \|h\|_H = 1$. It remains to show that every operator of this form is actually an extremal point. For this purpose assume that there is $\mathcal{U} = \beta^{-1} h \otimes h, \|h\|_H = 1, s \in (0, 1)$ as well as $\mathcal{U}_1, \mathcal{U}_2 \in B$, with $\mathcal{U} = (1 - s)\mathcal{U}_1 + s\mathcal{U}_2$. Then we also have

$$
1 = (1 - s) \sum_{i \in \mathbb{N}} \beta \sigma_i^{\mathcal{U}_1} (h_i^{\mathcal{U}_1}, h)_H^2 + s \sum_{i \in \mathbb{N}} \beta \sigma_i^{\mathcal{U}_2} (h_i^{\mathcal{U}_2}, h)_H^2 , \tag{B.1}
$$

as well as

$$
\max \left\{ \sum_{i \in \mathbb{N}} \beta \sigma_i^{\mathcal{U}_1}, \sum_{i \in \mathbb{N}} \beta \sigma_i^{\mathcal{U}_2} \right\} \leq 1. \tag{B.2}
$$

Moreover, we have $(h_i^{\mathcal{U}_j}, h)_H^2 \leq 1$ for all $i \in \mathbb{N}, j = 1, 2$. Together with (B.1) and (B.2), this yields

$$
1 = \sum_{i \in \mathbb{N}} \beta \sigma_i^{\mathcal{U}_1} (h_i^{\mathcal{U}_1}, h)_H^2 = \sum_{i \in \mathbb{N}} \beta \sigma_i^{\mathcal{U}_2} (h_i^{\mathcal{U}_2}, h)_H^2. \tag{B.3}
$$

We will now prove that these observations imply $\mathcal{U}_1 = \beta^{-1} h \otimes h$, the proof for $\mathcal{U}_2$ follows by the same steps. Note that $(h_i^{\mathcal{U}_1}, h)_H^2 = 1$ holds if and only if we have $h_i^{\mathcal{U}_1} = h$ or $h_i^{\mathcal{U}_1} = -h$. Recalling that $\{h_i^{\mathcal{U}_1}\}_{i \in \mathbb{N}}$ is an orthonormal set, (B.3) implies the existence of a unique index $\bar{\imath}$ with

$$(h_{\bar{\imath}}^{\mathcal{U}_1}, h)_H^2 = 1 \quad \text{as well as } (h_i^{\mathcal{U}_1}, h)_H^2 = 0, \ i \neq \bar{\imath}.$$

Together with (B.2) and (B.3), we finally get

$$1 = \sum_{i \in \mathbb{N}} \beta \sigma_i^{\mathcal{U}_1} (h_i^{\mathcal{U}_1}, h)_H^2 = \beta \sigma_{\bar{\imath}}^{\mathcal{U}_1}$$

as well as $\sigma_i^{\mathcal{U}_1} = 0, i \neq \bar{\imath}$. Consequently, there holds

$$\mathcal{U}_1 = \sigma_{\bar{\imath}}^{\mathcal{U}_1} h_{\bar{\imath}}^{\mathcal{U}_1} \otimes h_{\bar{\imath}}^{\mathcal{U}_1} = \beta^{-1} h \otimes h.$$

Repeating these arguments for $\mathcal{U}_2$ implies that $\mathcal{U}$ is an extremal point.

Finally, $\mathrm{Tr}(\mathcal{U}) = 0$ holds if and only if $\mathcal{U} = 0$. Assume that $\mathcal{U} = 0$ is not extremal, i.e., there are operators $\mathcal{U}_1, \mathcal{U}_2 \in B, \mathcal{U}_1 \neq \mathcal{U}_2$, as well as $s \in (0, 1)$ with $0 = (1 - s)\mathcal{U}_1 + s\mathcal{U}_2$. Then, without loss of generality, we have $\sigma_1^{\mathcal{U}_1} > 0$ and thus

$$0 = (h_1^{\mathcal{U}_1}, ((1 - s)\mathcal{U}_1 + s\mathcal{U}_2)h_1^{\mathcal{U}_1})_H \geq \sigma_1^{\mathcal{U}_1} > 0$$

yielding a contradiction. Hence 0 is an extremal point and the proof is finished.

It remains to characterize $\mathcal{B}$. Since H is infinite dimensional, for every $h \in H, \|h\|_H \leq 1$, there exists a weakly convergent sequence $\{h_k\}_k$ with limit h such that $\|h_k\|_H = 1$ for all k. Setting $\mathcal{U}_k = \beta^{-1} h_k \otimes h_k$ and $\mathcal{U} = \beta^{-1} h \otimes h$, we then also conclude

$$\langle P, \mathcal{U}_k \rangle = \beta^{-1}(h_k, P h_k) \rightarrow \beta^{-1}(h, P h) = \langle P, \mathcal{U} \rangle$$

for all $P \in K(H)$, using compactness of P. Hence, $\mathcal{U}_k \overset{*}{\rightharpoonup} U$ and thus $\mathcal{U} \in \mathcal{B}$. For the converse inclusion, let $\mathcal{U}_k = \beta^{-1} h_k \otimes h_k \in B$ with $\|h_k\|_H = 1$ denote a weak* convergent sequence with limit $\mathcal{U}$. Then there exists a subsequence of $\{h_k\}_k$, denoted by the same index, which converges weakly to some $h \in H, \|h\|_H \leq 1$. Repeating previous arguments and noting the uniqueness of weak* limits, we then arrive at $\mathcal{U} = \beta^{-1} h \otimes h$. □

B.2.2. Proof of Proposition 4.9

The optimality condition in (4.16) follows immediately from Proposition 2.3, the characterization of the extremal points as well as $\sigma_1^{\bar{P}} = \sup_{\|h\|_H=1}(h, \bar{P}h)_H$. Now,

let $\bar{\mathcal{U}}$ be a minimizer of (4.13) and assume that $\sigma_1^{\bar{P}} = \beta$. Then we have

$$\operatorname{Tr}(\bar{P}\bar{\mathcal{U}}) = \sum_{i \in \mathbb{N}} \sigma_i^{\bar{\mathcal{U}}} (h_i^{\bar{\mathcal{U}}}, \bar{P}h_i^{\bar{\mathcal{U}}})_H = \beta \operatorname{Tr}(\bar{\mathcal{U}}) = \beta \sum_{i \in \mathbb{N}} \sigma_i^{\bar{\mathcal{U}}},$$

as well as $(h_i^{\bar{\mathcal{U}}}, \bar{P}h_i^{\bar{\mathcal{U}}})_H \leq \beta$, $i \in \mathbb{N}$. Consequently, we conclude that $\sigma_i^{\bar{\mathcal{U}}} > 0$ implies $\bar{P}h_i^{\bar{\mathcal{U}}} = \beta h_i^{\bar{\mathcal{U}}}$, i.e., $h_i^{\bar{\mathcal{U}}}$ is an eigenfunction for the leading eigenvalue of $\bar{P}$. Thus, since $\{h_i^{\bar{\mathcal{U}}}\}_i$ form an ONB, and possibly after a change of basis, $\bar{\mathcal{U}}$ is of the form (4.17). □

B.2.3. Proof of Theorem 4.10

The statement in (4.18), the structure of $\bar{\mathcal{U}}$ and its uniqueness as well as the Lipschitz result on K follow immediately from Assumption (D1), the assumptions on K and the strict convexity of F, see Assumption 2.2. Let $\mathcal{U} \in B$ with $\operatorname{Tr}(\mathcal{U}) = \beta^{-1}$ and $\mathcal{U} \geq 0$ be arbitrary but fixed. For the proof of the quadratic growth behavior we follow similar steps as in the finite dimensional setting, see [42, Lemma 4]. Setting $\delta = \sigma_1^{\bar{P}} - \sigma_2^{\bar{P}} > 0$, we estimate

$$
\begin{aligned}
1 - \langle \bar{P}, \mathcal{U} \rangle &= 1 - \beta (h_1^{\bar{P}}, \mathcal{U}h_1^{\bar{P}})_H - \sum_{i \in \mathbb{N}, i > 1} \sigma_i^{\bar{P}} (h_i^{\bar{P}}, \mathcal{U}h_i^{\bar{P}})_H \\
&\geq 1 - \beta \sum_{i \in \mathbb{N}} (h_i^{\bar{P}}, \mathcal{U}h_i^{\bar{P}})_H + \delta \sum_{i \in \mathbb{N}, i > 1} (h_i^{\bar{P}}, \mathcal{U}h_i^{\bar{P}})_H \\
&= \delta \left(\beta^{-1} - (h_1^{\bar{P}}, \mathcal{U}h_1^{\bar{P}})_H \right) = \delta \beta \left(\beta^{-2} - (\mathcal{U}, \bar{\mathcal{U}}_1)_{\mathrm{HS}} \right) \\
&= \delta \beta \left(\frac{1}{2} \operatorname{Tr}(\mathcal{U})^2 + \frac{1}{2} \operatorname{Tr}(\bar{\mathcal{U}}_1)^2 - (\mathcal{U}, \bar{\mathcal{U}}_1)_{\mathrm{HS}} \right) \\
&\geq \delta \beta \left(\frac{1}{2} \|\mathcal{U}\|_{\mathrm{HS}}^2 + \frac{1}{2} \|\bar{\mathcal{U}}_1\|_{\mathrm{HS}}^2 - (\mathcal{U}, \bar{\mathcal{U}}_1)_{\mathrm{HS}} \right) \\
&= (\delta \beta / 2) \|\mathcal{U} - \bar{\mathcal{U}}_1\|_{\mathrm{HS}}^2,
\end{aligned}
$$

where we use $\operatorname{Tr}(\mathcal{U}) \geq \|\mathcal{U}\|_{\mathrm{HS}}$ for all $\mathcal{U} \in B$ in the last inequality. This finishes the proof. □

B.3. Minimum effort problems (Sect. 4.3)

B.3.1. Proof of Lemma 4.12

Without loss of generality assume $\alpha = 1$. The characterization (4.22) is classical [30, page 144]. Moreover the inclusion $\overline{\operatorname{Ext}(B)}^* \subset B$ follows from weak* lower semi-continuity of the norm. Let us now prove that $B \subset \overline{\operatorname{Ext}(B)}^*$. To this end, set $R := (0, 1)^d$ and $R_\lambda := (0, \sqrt[d]{\lambda})^d$ for any $\lambda \in [0, 1]$. Define the map h^λ as $h^\lambda := 1$

in $R \backslash R_\lambda$, $h^\lambda := -1$ in R_λ, and extend it R-periodically to the whole $\mathbb{R}^d$. For $k \in \mathbb{N}$, $x \in \mathbb{R}^d$, set $h_k^\lambda(x) := h^\lambda(kx)$. Then

$$h_k^\lambda \overset{*}{\rightharpoonup} \int_R h^\lambda \, dx = 1 - 2\lambda \quad \text{weakly* in } L^\infty(A) \tag{B.4}$$

as $k \to \infty$, for all open, bounded and non-empty sets $A \subset \mathbb{R}^d$, see [6, Lemma A.1]. Let $\{Q_{n,j}\}_{j \in \mathbb{N}}$ be a family of pairwise disjoint and open d-cubes of side length $1/n$ covering $(\mathbb{R} \backslash (\mathbb{Z}/n))^d$. Define $\tilde{Q}_{n,j} := Q_{n,j} \cap \Omega$ and let J_n be the collection of indices $j \in \mathbb{N}$ such that $\tilde{Q}_{n,j} \neq \emptyset$. Let $\mathcal{P}_n$ be the set of maps $u \in B$ such that $u = c_{n,j}$ in $\tilde{Q}_{n,j}$ for some $c_{n,j} \in [-1, 1]$ and for all $j \in J_n$. Fix $u \in \mathcal{P}_n$ and $n \in \mathbb{N}$. For every $j \in J_n$, $k \in \mathbb{N}$ define $u_k(x) := h_k^{\lambda_{n,j}}(x)$ for all $x \in \tilde{Q}_{n,j}$, where $\lambda_{n,j} := (1 - c_{n,j})/2$. Therefore $\{u_k\}_k \subset \text{Ext}(B)$. By (B.4) we have that $u_k \overset{*}{\rightharpoonup} c_{n,j}$ weakly* in $L^\infty(\tilde{Q}_{n,j})$ as $k \to \infty$, for all $j \in J_n$. Thus $u_k \overset{*}{\rightharpoonup} u$ in $L^\infty(\Omega)$. In particular we have shown that $\mathcal{P}_n \subset \overline{\text{Ext}(B)}^*$. Therefore $\mathcal{P} \subset \overline{\text{Ext}(B)}^*$, where $\mathcal{P} := \cup_{n \in \mathbb{N}} \mathcal{P}_n$. As clearly $\overline{\mathcal{P}}^* = B$, the proof is concluded. □

B.3.2. Proof of Lemma 4.13

The optimality condition in (4.21) follows immediately from Proposition 2.3 by pointing out

$$\int_\Omega |\bar{p}(x)| \, dx = \alpha \max_{v \in B} \int_\Omega \bar{p}(x) v(x) \, dx$$
$$= \alpha \max_{v \in \text{Ext}(B)} \langle \bar{p}, v \rangle \leq \alpha.$$

This also implies

$$\alpha \|\bar{u}\|_\infty = \int_\Omega \bar{p}(x) \bar{u}(x) \, dx \leq \|\bar{u}\|_\infty \int_\Omega |\bar{p}(x)| \, dx \leq \alpha \|\bar{u}\|_\infty,$$

hence, equality holds everywhere. If $\bar{u} \neq 0$, this can only be the case if

$$\int_\Omega |\bar{p}(x)| \, dx = \alpha \quad \text{and} \quad \frac{\bar{u}(x)}{\|\bar{u}\|_\infty} \in \begin{cases} \{1\} & \text{if } \bar{p}(x) > 0, \\ \{-1\} & \text{if } \bar{p}(x) < 0, \\ [-1, 1] & \text{else,} \end{cases}$$

for a.e. $x \in \Omega$, yielding (4.24). □

Appendix C: Complements to Sect. 5 and Sect. 6

C.1. Proofs of Sect. 5

In this section we exhibit proofs for Propositions 5.2, 5.3. In order to define the map $\mathcal{I}$ at Proposition 5.2 we employ the classical Choquet's Theorem [51, Page 14], which we recall for reader's convenience.

Theorem C.1 (Choquet) *Let X be a locally convex space and $K \subset X$ be a metrizable compact convex subset. For each $v_0 \in K$ there exists a probability measure μ over X concentrated on $\mathrm{Ext}(K)$ which represents v_0, that is,*

$$T(v_0) = \int_X T(v)\, \mathrm{d}\mu(v),$$

for all T in the topological dual of X.

We are now ready to prove Proposition 5.2.

C.1.1. Proof of Proposition 5.2

Let $\mu \in M^+(\mathcal{B})$ and consider the embedding $f \colon \mathcal{B} \to \mathcal{M}$, $f(v) := v$. We have that f is weak* μ-measurable, since $v \mapsto \langle p, v \rangle$ is weak* continuous and hence, μ-measurable for all $p \in \mathcal{C}$. Moreover, $v \mapsto \langle p, v \rangle$ is μ-integrable for each $p \in \mathcal{C}$, since

$$\int_{\mathcal{B}} |\langle p, v \rangle|\, \mathrm{d}\mu(v) \le \|p\|_{\mathcal{C}} \int_{\mathcal{B}} \|v\|_{\mathcal{M}}\, \mathrm{d}\mu(v)$$
$$\le \|p\|_{\mathcal{C}} \|\mu\|_{M(\mathcal{B})} \max_{v \in \mathcal{B}} \|v\|_{\mathcal{M}},$$

and the last term is finite, since $\mathcal{B}$ is norm bounded. Therefore, f is Gelfand integrable over $\mathcal{B}$ (see [1, Thm 11.52]), that is, there exists a unique $u \in \mathcal{M}$ such that (5.2) holds. We thus set $\mathcal{I}(\mu) := u$. Recalling that the weak* topology separates points, we immediately conclude that $\mathcal{I}$ is well defined and

$$\mathcal{I}(\lambda_1 \mu_1 + \lambda_2 \mu_2) = \lambda_1 \mathcal{I}(\mu_1) + \lambda_2 \mathcal{I}(\mu_2),$$

for all $\lambda_1, \lambda_2 \in \mathbb{R}_+$, $\mu_1, \mu_2 \in M^+(\mathcal{B})$. We now show (i). Notice that $\mathcal{I}(0) = 0$ and the estimate holds trivially. If $\mu \ne 0$, we can apply Theorem 11.54 in [1] to obtain

$$\frac{1}{\|\mu\|_{M(\mathcal{B})}} \mathcal{I}(\mu) \in \overline{\mathrm{conv}(\mathcal{B})}^*.$$

By recalling that B is weak* compact and convex, we conclude that $\overline{\mathrm{conv}(\mathcal{B})}^* \subset B$. Thus, by 1-homogeneity of $\mathcal{G}$ we conclude $\mathcal{G}(\mathcal{I}(\mu)) \le \|\mu\|_{M(\mathcal{B})}$ and, in particular, $\mathcal{I}(\mu) \in \mathrm{dom}\,\mathcal{G}$.

For the proof of (ii), let $u \in \operatorname{dom} \mathcal{G}$ be fixed. If $\mathcal{G}(u) = 0$ then $u = 0$, and the measure $\mu = 0$ satisfies the statement. Assume now $\mathcal{G}(u) > 0$. Apply Theorem C.1 with $X = (\mathcal{M}, weak^*)$, $K = B$, $v_0 = u/\mathcal{G}(u)$ to obtain a probability measure $\tilde{\mu}$ over X concentrated on $\operatorname{Ext}(B)$ and such that

$$\langle p, u \rangle = \int_{\mathcal{M}} \langle p, v \rangle \mathcal{G}(u) \, \mathrm{d}\tilde{\mu}(v), \tag{C.1}$$

for all $p \in \mathcal{C}$, since $(\mathcal{M}, weak^*)^* = \mathcal{C}$. Set $\mu := \mathcal{G}(u)\tilde{\mu} \llcorner B$. Since B is weak* closed, we have that $\mu \in M^+(B)$. Moreover, $\|\mu\|_{M(B)} = \mathcal{G}(u)$, μ is concentrated on $\operatorname{Ext}(B)$ and it represents u. As the weak* topology separates points, from (C.1) we conclude that $\mathcal{I}(\mu) = u$. This shows in particular that $\mathcal{I}$ is surjective, achieving the proof of (ii).

Finally, $\mathcal{I}$ is weak*-to-weak* continuous: indeed if $\mu_k \stackrel{*}{\rightharpoonup} \mu$ weakly* in $M(B)$, then by definition of weak* convergence we have $\langle\!\langle P, \mu_k \rangle\!\rangle \to \langle\!\langle P, \mu \rangle\!\rangle$ for all $P \in C(B)$. Note that the map $P(v) = \langle p, v \rangle$, $v \in B$, $p \in \mathcal{C}$ belongs to $C(B)$. Thus, for all $p \in \mathcal{C}$,

$$\int_{B} \langle p, v \rangle \, \mathrm{d}\mu_k(v) \to \int_{B} \langle p, v \rangle \, \mathrm{d}\mu(v).$$

Thanks to (5.3) the above reads $\mathcal{I}(\mu_k) \stackrel{*}{\rightharpoonup} \mathcal{I}(\mu)$ weakly* in $\mathcal{M}$, concluding the proof. $\qquad\square$

Finally, we prove Proposition 5.3.

C.1.2. Proof of Proposition 5.3

For $\mu \in M(B)$ define the functional $T_\mu : Y \to \mathbb{R}$ by setting

$$T_\mu(y) := \int_{B} (Kx, y)_Y \, \mathrm{d}\mu(x),$$

for all $y \in Y$. Notice that T_μ is well defined, since the map $v \mapsto (Kv, y)_Y = \langle K_* y, v \rangle$ is weak* continuous over $\mathcal{M}$, and hence μ-measurable. It is clear that T_μ is linear. Moreover, T_μ is continuous. Indeed, it holds

$$\|T_\mu(y)\| \leq \int_{B} |(Kv, y)_Y| \, \mathrm{d}|\mu|(v)$$

$$\leq \|K\| \|y\|_Y \sup_{v \in B} \|v\|_{\mathcal{M}} \int_{B} 1 \, \mathrm{d}|\mu|(v)$$

$$= C \|K\| \|\mu\|_{M(B)} \|y\|_Y,$$

where we recalled that the constant C is defined at (i) and is finite, since B is norm bounded. Therefore, by Riesz's Theorem, there exists a unique element in Y representing T_μ, thus defining a map $\mathcal{K} : M(B) \to Y$. In particular, $\mathcal{K}$ is linear and satisfies

(5.4). Moreover, by (5.4) we have

$$\|\mathcal{K}\mu\|_Y = \sup_{\|y\|_Y \leq 1} \left\| T_\mu(y) \right\| \leq C \, \|K\| \, \|\mu\|_{M(\mathcal{B})} \, ,$$

by the above, so $\mathcal{K}$ is bounded and (i) follows.

As for (ii), assume that $\mu \in M^+(\mathcal{B})$ represents $u \in \mathcal{M}$. By (5.2) and (5.4) we have

$$(Ku, y)_Y = \langle K_* y, u \rangle = \int_\mathcal{B} \langle K_* y, v \rangle \, \mathrm{d}\mu(v)$$

$$= \int_\mathcal{B} (Kv, y)_Y \, \mathrm{d}\mu(v) = (\mathcal{K}\mu, y)_Y \, ,$$

which holds for all $y \in Y$, thus showing that $\mathcal{K}\mu = Ku$.

In order to show (iii), assume that $\mu_k \overset{*}{\rightharpoonup} \mu$ in $M(\mathcal{B})$. First notice that by (5.4),

$$\lim_{k\to\infty} (\mathcal{K}\mu_k, y)_Y = \lim_{k\to\infty} \int_\mathcal{B} (Kv, y)_Y \, \mathrm{d}\mu_k(v)$$

$$= \int_\mathcal{B} (Kv, y) \, \mathrm{d}\mu(v) = (\mathcal{K}\mu, y)_Y \, ,$$

for every $y \in Y$, where the second equality follows because the map $v \mapsto (Kv, y)_Y$ is weak* continuous, as K is weak*-to-weak continuous in $\mathcal{M}$, and hence, belongs to $C(\mathcal{B})$. Thus, $\mathcal{K}\mu_k \rightharpoonup \mathcal{K}\mu$ in Y. Moreover, by applying (5.4) twice we see that

$$\|\mathcal{K}\mu_k\|_Y^2 = \int_\mathcal{B} (Kv, \mathcal{K}\mu_k)_Y \, \mathrm{d}\mu_k(v)$$

$$= \int_\mathcal{B} \int_\mathcal{B} (Kv, Kw)_Y \, \mathrm{d}\mu_k(w) \mathrm{d}\mu_k(v) \, . \tag{C.2}$$

Note that the function $(v, w) \mapsto (Kv, Kw)_Y$ is an element of $C(\mathcal{B} \times \mathcal{B})$. Indeed, given a sequence $(v_k, w_k) \in \mathcal{B} \times \mathcal{B}$ such that $(v_k, w_k) \overset{*}{\rightharpoonup} (v, w)$ for $(v, w) \in \mathcal{B} \times \mathcal{B}$ we estimate

$$|(Kv_k, Kw_k)_Y - (Kv, Kw)_Y|$$
$$\leq C\|K\|_{\mathcal{L}(\mathcal{M},Y)}(\|Kw_k - Kw\|_Y + \|Kv_k - Kv\|_Y) \, ,$$

where C is defined at (i). As, $v_k, w_k, v, w \in \mathcal{B}$, using the weak*-to-strong continuity of K on $\mathrm{dom}(\mathcal{G})$, cf. Assumptions (A1)–(A3), we conclude that $(Kv_k, Kw_k)_Y \to (Kv, Kw)_Y$ proving that $(v, w) \mapsto (Kv, Kw)_Y$ is an element of $C(\mathcal{B} \times \mathcal{B})$. Moreover, we have $\mu_k \otimes \mu_k \overset{*}{\rightharpoonup} \mu \otimes \mu$ in $M(\mathcal{B} \times \mathcal{B})$, which, combined with (C.2) and the fact that $(v, w) \mapsto (Kv, Kw)_Y \in C(\mathcal{B} \times \mathcal{B})$ implies the convergence $\|\mathcal{K}\mu_k\|_Y \to \|\mathcal{K}\mu\|_Y$. Combined with the weak convergence $\mathcal{K}\mu_k \rightharpoonup \mathcal{K}\mu$ we finally conclude $\mathcal{K}\mu_k \to \mathcal{K}\mu$ in Y, so that $\mathcal{K}$ is weak*-to-strong continuous.

Last, note that the operator $\mathcal{K}_*$ is well-defined, linear and continuous. For any $y \in Y$ and $\mu \in M(\mathcal{B})$ we obtain

$$\langle\langle \mathcal{K}_* y, \mu \rangle\rangle = \int_{\mathcal{B}} \langle K_* y, v \rangle \, \mathrm{d}\mu(v)$$
$$= \int_{\mathcal{B}} (Kv, y)_Y \, \mathrm{d}\mu(v) = (\mathcal{K}\mu, y)_Y,$$

where again, we used (5.4). This concludes the proof of (iv).

Assume now that Y is separable and fix $\mu \in M(\mathcal{B})$. Notice that $f : \mathcal{B} \to Y$ defined by $f(v) := Kv$ is weakly μ-measurable, since the map $v \mapsto (Kv, y)_Y = \langle K_* y, v \rangle$ is weak* continuous and hence, μ-measurable, for each $y \in Y$ fixed. As Y is separable, we also have that f is essentially separably valued. Therefore, Pettis' Theorem ([34, Sec II.1, Thm 2]) implies that f is strongly measurable with respect to μ. Moreover

$$\int_{\mathcal{B}} \|Kv\|_Y \, \mathrm{d}\mu(v) \leq C \, \|K\|_{\mathcal{L}(\mathcal{M}, Y)} \|\mu\|_{M(\mathcal{B})} < \infty,$$

where C is defined at (i), showing that f is Bochner integrable with respect to μ. Hence, by (5.4) and Hille's Theorem ([34, Sec II.1, Thm 6]), we infer

$$(\mathcal{K}\mu, y)_Y = \int_{\mathcal{B}} (Kv, y)_Y \, \mathrm{d}\mu(v) = \left(\int_{\mathcal{B}} Kv \, \mathrm{d}\mu(v), y \right)_Y$$

for all $y \in Y$, concluding (5.5) and the proof. □

C.2. Proofs of Sect. 6

C.2.1. Proof of Proposition 6.5

Let $k \geq 1$. By construction, $\lambda^k \in \mathbb{R}^{N_k}$ solves

$$\min_{\lambda \in \mathbb{R}_+^{N_k}} \left[F\left(\sum_{i=1}^{N_k} \lambda_i K u_i^k \right) + \sum_{i=1}^{N_k} \lambda_i \right].$$

Deriving the first order necessary optimality conditions for this problem, we obtain that

$$P_k(u_i^k) \leq 1 \quad \text{for all} \ \ u_i^k \in \mathcal{A}_k, \quad \sum_{i=1}^{N_k} \lambda_i^k P_k(u_i^k) = \sum_{i=1}^{N_k} \lambda_i^k.$$

Since $\lambda_i^k > 0$ by construction, we deduce that $P_k(u_i^k) = 1$ for $i = 1, \ldots, N_k$. In particular,

$$\langle\!\langle P_k, \mu_k \rangle\!\rangle = \sum_{i=1}^{N_k} \lambda_i^k P_k(u_i^k) = \sum_{i=1}^{N_k} \lambda_i^k = \|\mu_k\|_{M(\mathcal{B})}.$$

Finally, $\max_{v \in \mathcal{B}} P_k(v) \geq \max_{v \in \mathcal{A}_k} P_k(v) = 1$, concluding. $\qquad\square$

C.2.2. Proof of Proposition 6.6

Note that $\{\mu_k\}_k \subset E_{\mu_0}$ by (6.15). Arguing as in the proof of Theorem 6.2, we get

$$\|\nabla F(y_k) - \nabla F(\bar{y})\|_Y + \|P_k - \bar{P}\|_{C(\mathcal{B})} \leq L_{\mu_0}(1 + C\|K\|_{\mathcal{L}(\mathcal{M}, Y)})\|y_k - \bar{y}\|_Y,$$

where we recall that $L_{\mu_0} > 0$ is the Lipschitz constant of ∇F in $\mathcal{K}E_{\mu_0}$, and $C > 0$ is the constant in Proposition 5.3 (i). Hence, it suffices to prove (6.19) for $\|y_k - \bar{y}\|_Y$. Let $\mathcal{N}(\bar{y}) \subset Y$ and $\theta > 0$ denote the neighbourhood and constant from Assumption (B1), respectively. Recall that $\mu_k \stackrel{*}{\rightharpoonup} \bar{\mu}$ along the whole sequence by (6.6). By weak*-to-strong continuity of $\mathcal{K}$, see Proposition 5.3 (iii), we then conclude $y_k \to \bar{y}$ in Y. Thus, there exist $M \in \mathbb{N}$ such that $y_k \in \mathcal{N}(\bar{y})$ for all $k \geq M$. Using the strong convexity of F in $\mathcal{N}(\bar{y})$, we estimate

$$\begin{aligned}
j(\mu_k) &= F(y_k) + \|\mu_k\|_{M(\mathcal{B})} \\
&\geq F(\bar{y}) + \theta\|y_k - \bar{y}\|_Y^2/2 + (\nabla F(\bar{y}), \mathcal{K}\mu_k - \mathcal{K}\bar{\mu})_Y + \|\mu_k\|_{M(\mathcal{B})} \\
&= j(\bar{\mu}) + \theta\|y_k - \bar{\mu}\|_Y^2/2 + \langle\!\langle \bar{P}, \bar{\mu} - \mu_k \rangle\!\rangle + \|\mu_k\|_{M(\mathcal{B})} - \|\bar{\mu}\|_{M(\mathcal{B})} \\
&\geq j(\bar{\mu}) + \theta\|y_k - \bar{y}\|_Y^2/2,
\end{aligned}$$

where we used (5.11) in the final inequality. Thus (6.19) follows by rearranging the terms in the above estimate, recalling that $r_j(\mu_k) = j(\mu_k) - j(\bar{\mu})$. The final part of the statement holds since $r_j(\mu_k) \to 0$ by Theorem 6.2. $\qquad\square$

References

1. Aliprantis, C.D., Border, K.: Infinite Dimensional Analysis. Springer, Berlin (2006)
2. Ambrosio, L., Fusco, N., Pallara, D.: Functions of Bounded Variation and Free Discontinuity Problems. Oxford Science Publications, Oxford (2000)
3. Ambrosio, L., Gigli, N., Savaré, G.: Gradient flows. In: Metric Spaces and in the Space of Probability Measures, Birkhäuser Basel (2005)
4. Aragón Artacho, F.J., Geoffroy, M.H.: Metric subregularity of the convex subdifferential in Banach spaces. J. Nonlinear Convex Anal. **15**(1), 35–47 (2014)
5. Auslender, A.: Algorithme de recherche des points stationnaires d'une fonctionnelle dans un espace vectoriel topologique. application a un probleme de controle a evolution non lineaire. Comptes Rendus de l'Acad. des Sci. **266**, 226–229 (1968)
6. Ball, J.M., Murat, F.: $W^{1,p}$-quasiconvexity and variational problems for multiple integrals. J. Funct. Anal. **58**, 225–253 (1984)

7. Beinert, R., Bredies, K.: Tensor-free proximal methods for lifted bilinear/quadratic inverse problems with applications to phase retrieval. Found. Comput. Math. **21**(5), 1181–1232 (2021)
8. Ben-Asher, J.Z., Cliff, E.M., Burns, J.A.: Computational methods for the minimum effort problem with applications to spacecraft rotational maneuvers. In: Proceedings. ICCON IEEE International Conference on Control and Applications, pp. 472–478 (1989)
9. Benamou, J.-D., Brenier, Y.: A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. Numer. Math. **84**(3), 375–393 (2000)
10. Bonsall, F.F.: A general atomic decomposition theorem and Banach's closed range theorem. Q. J. Math. **42**(1), 9–14 (1991)
11. Boyd, N., Geoffrey, S., Benjamin, R.: The alternating descent conditional gradient method for sparse inverse problems. SIAM J. Optim. **27**(2), 616–639 (2017)
12. Boyer, C., Chambolle, A., De Castro, Y., Duval, V., de Gournay, F., Weiss, P.: On representer theorems and convex regularization. SIAM J. Optim. **29**(2), 1260–1281 (2019)
13. Bredies, K., Carioni, M.: Sparsity of solutions for variational inverse problems with finite-dimensional data. Calc. Var. Partial. Differ. Equ. **59**(1), 14 (2020)
14. Bredies, K., Carioni, M., Fanzon, S.: A superposition principle for the inhomogeneous continuity equation with Hellinger-Kantorovich-regular coefficients. Commun. Partial Differ. Equ. **47**(10), 2023–2069 (2022). https://doi.org/10.1080/03605302.2022.2109172
15. Bredies, K., Carioni, M., Fanzon, S., Romero, F.: On the extremal points of the ball of the Benamou–Brenier energy. Bull. Lond. Math. Soc. **53**(5), 1436–1452 (2021)
16. Bredies, K., Carioni, M., Fanzon, S., Romero, F.: A generalized conditional gradient method for dynamic inverse problems with optimal transport regularization. Found. Comput. Math. **23**, 833–898 (2023). https://doi.org/10.1007/s10208-022-09561-z
17. Bredies, K., Fanzon, S.: An optimal transport approach for solving dynamic inverse problems in spaces of measures. ESAIM: Math. Model. Numer. Anal. **54**(6), 2351–2382 (2020)
18. Bredies, K., Lorenz, D.A.: Iterated hard shrinkage for minimization problems with sparsity constraints. SIAM J. Sci. Comput. **30**(2), 657–683 (2008)
19. Bredies, K., Lorenz, D.A.: Mathematical Image Processing. Birkhäuser, Basel (2018)
20. Bredies, K., Lorenz, D.A., Maass, P.: A generalized conditional gradient method and its connection to an iterative shrinkage method. Comput. Optim. Appl. **42**(2), 173–193 (2009)
21. Bredies, K., Pikkarainen, H.K.: Inverse problems in spaces of measures. ESAIM: Control Optim. Calculus Var. **19**(1), 190–218 (2013)
22. Brune, C.: 4D imaging in tomography and optical nanoscopy. Ph.D. thesis, Münster (Westfalen) (2010)
23. Candès, E.J., Strohmer, T., Voroninski, V.: Phaselift: exact and stable signal recovery from magnitude measurements via convex programming. Commun. Pure Appl. Math. **66**(8), 1241–1274 (2013)
24. Canon, M.D., Cullum, C.D.: A tight upper bound on the rate of convergence of Frank–Wolfe Algorithm. SIAM J. Control **6**(4), 509–516 (1968)
25. Carioni, M., Iglesias, J.A., Walter, D.: Extremal points and sparse optimization for generalized Kantorovich-Rubinstein norms. arXiv preprint arXiv:2209.09167 (2022)
26. Casas, E., Vexler, B., Zuazua, E.: Sparse initial data identification for parabolic PDE and its finite element approximations. Math. Control Related Fields **5**(3), 377–399 (2015)
27. Casas, E., Wachsmuth, D., Wachsmuth, G.: Sufficient second-order conditions for bang-bang control problems. SIAM J. Control. Optim. **55**(5), 3066–3090 (2017)
28. Chandrasekaran, V., Recht, B., Parrilo, P.A., Willsky, A.S.: The convex geometry of linear inverse problems. Found. Comput. Math. **12**(6), 805–849 (2012)
29. Clason, C., Ito, K., Kunisch, K.: A minimum effort optimal control problem for elliptic PDEs. ESAIM: Math. Model. Numer. Anal. **46**(4), 911–927 (2012)
30. Conway, J.B.: A Course in Functional Analysis. Springer, New York (2007)
31. Daubechies, I., Defrise, M., De Mol, C.: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Commun. Pure Appl. Math. **57**(11), 1413–1457 (2004)
32. Demianov, V.F., Rubinov, A.M.: Approximate Methods in Optimization Problems. Number 32. Elsevier Publishing Company (1970)
33. Denoyelle, Q., Duval, V., Peyré, G., Soubies, E.: The sliding Frank–Wolfe algorithm and its application to super-resolution microscopy. Inverse Prob. **36**(1), 014001 (2019)
34. Diestel, J., Uhl, J.J.: Vector Measures. American Mathematical Society, Mathematical Surveys and Monographs (1977)

35. Dunn, J.C.: Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals. SIAM J. Control. Optim. **17**(2), 187–211 (1979)
36. Dunn, J.C.: Convergence rates for conditional gradient sequences generated by implicit step length rules. SIAM J. Control. Optim. **18**(5), 473–487 (1980)
37. Dunn, J.C., Harshbarger, S.: Conditional gradient algorithms with open loop step size rules. J. Math. Anal. Appl. **62**(2), 432–444 (1978)
38. Duval, V., Tovey, R.: Dynamical programming for off-the-grid dynamic inverse problems. arXiv preprint arXiv:2112.11378 (2021)
39. Flinth, A., de Gournay, F., Weiss, P.: On the linear convergence rates of exchange and continuous methods for total variation minimization. Math. Program. **190**(1), 221–257 (2021)
40. Frank, M., Wolfe, P.: An algorithm for quadratic programming. Naval Res. Logist. Q. **3**(1–2), 95–110 (1956)
41. Freund, R.M., Grigas, P.: New analysis and results for the Frank–Wolfe method. Math. Program. **155**(1–2), 199–230 (2016)
42. Garber, D.: Linear convergence of Frank-Wolfe for rank-one matrix recovery without strong convexity. Math. Program. **199**, 87–121 (2023). https://doi.org/10.1007/s10107-022-01821-8
43. GuéLat, J., Marcotte, P.: Some comments on Wolfe's 'away step'. Math. Program. **35**(1), 110–119 (1986)
44. Holloway, C.A.: An extension of the Frank and Wolfe method of feasible directions. Math. Program. **6**(1), 14–27 (1974)
45. Iglesias, J.A., Walter, D.: Extremal points of total generalized variation balls in 1D: characterization and applications. J. Convex Anal. **29**(4), 1xxx–1xxx (2022)
46. Ito, K., Kunisch, K.: Minimal effort problems and their treatment by semismooth Newton methods. SIAM J. Control. Optim. **49**(5), 2083–2100 (2011)
47. Jaggi, M.: Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In: International Conference on Machine Learning, pp. 427–435 (2013)
48. Kunisch, K., Walter, D.: On fast convergence rates for generalized conditional gradient methods with backtracking stepsize. Numerical Algebra Control and Optimization (2022). https://doi.org/10.3934/naco.2022026
49. Lacoste-Julien, S., Jaggi, M.: On the global linear convergence of Frank–Wolfe optimization variants. Proc. Annu. Conf. Neural Inf. Process. Syst. **2015**, 496–504 (2015)
50. Leykekhman, D., Vexler, B., Walter, D.: Numerical analysis of sparse initial data identification for parabolic problems. ESAIM: Math. Model. Numer. Anal. **54**(4), 1139–1180 (2020)
51. Phelps, R.R.: Lectures on Choquet's Theorem, vol. LN1757. Springer, Berlin (2001)
52. Pieper, K., Walter, D.: Linear convergence of accelerated conditional gradient algorithms in spaces of measures. ESAIM: Control Optim. Calc. Var. **27** (2021)
53. Santambrogio, F.: Optimal Transport for Applied Mathematicians. Birkhäuser, Basel (2015)
54. Saumier, L.P., Khouider, B., Agueh, M.: Optimal transport for particle image velocimetry: real data and postprocessing algorithms. SIAM J. Appl. Math. **75**(6), 2495–2514 (2015)
55. Schmitzer, B., Schäfers, K.P., Wirth, B.: Dynamic cell imaging in PET with optimal transport regularization. IEEE Trans. Med. Imaging **39**(5), 1626–1635 (2019)
56. Shalev-Shwartz, S., Srebro, N., Zhang, T.: Trading accuracy for sparsity in optimization problems with sparsity constraints. SIAM J. Optim. **20**(6), 2807–2832 (2010)
57. Sun, Y., Bach, F.: Safe screening for the generalized conditional gradient method. arXiv preprint arXiv:2002.09718 (2020)
58. Ventura, J.A., Hearn, D.W.: Restricted simplicial decomposition for convex constrained problems. Math. Program. **59**(1–3), 71–85 (1993)
59. Von Hohenbalken, B.: Simplicial decomposition in nonlinear programming algorithms. Math. Program. **13**(1), 49–68 (1977)
60. Walter, D.: On sparse sensor placement for parameter identification problems with partial differential equations. Ph.D. thesis, Technische Universität München (2019). http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:bvb:91-diss-20190604-1456274-1-8
61. Wang, C., Wang, Y., W.E, Schapire, R.: Functional Frank-Wolfe boosting for general loss functions. arXiv preprint arXiv:1510.02558 (2015)
62. Yu, Y., Zhang, X., Schuurmans, D.: Generalized conditional gradient for sparse estimation. J. Mach. Learn. Res. **18**(144), 1–46 (2017)

63. Zhang, X., Yu, Y., Schuurmans, D.: Accelerated training for matrix-norm regularization: a boosting approach. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2, NIPS'12, pp. 2906–2914, Red Hook, NY, USA (2012). Curran Associates Inc

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