LINEAR CONVERGENCE OF ACCELERATED GENERALIZED CONDITIONAL GRADIENT METHODS

KRISTIAN BREDIES, MARCELLO CARIONI, SILVIO FANZON, AND DANIEL WALTER

Abstract. We propose an *accelerated generalized conditional gradient method* (AGCG) 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 extreme points of the unit ball of the regularizer and an iterate $u_k \in \operatorname{cone}(\mathcal{A}_k)$. 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 AGCG algorithm is connected to a *Primal-Dual-Active-point Method* (PDAP) on the lifted problem for which we finally derive the desired convergence rates.

Key words: non-smooth optimization, conditional gradient method, sparsity, Choquet's theorem.

2020 Mathematics Subject Classification: 65J22, 65K05, 52A41, 46G12, 46A55.

1. Introduction

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

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

over a Banach space $\mathcal{M}$. Here, the forward operator K maps continuously from $\mathcal{M}$ into a not necessarily finite dimensional Hilbert space Y of observations 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 its 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 nonsmoothness 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 [12, 18], inverse problems in the space of measures regularized with the total variation [4, 14, 20] and dynamic inverse problems with optimal transport regularizers [8, 11, 7, 9].

1.1. **Contribution & 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 $K_*\colon Y \to \mathcal{C}$, $(K_*)^* = K$. We refer to Section 3.3 for such an example. In this case a simple approach to computing a minimizer to ($\mathcal{P}_\mathcal{M}$) is constituted by *generalized conditional gradient* (GCG) algorithms. At each iteration this method updates the iterate u_k by computing the dual variable $p_k = -K^*\nabla F(Ku_k)$ and setting

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

where $s_k \in (0,1)$ is chosen according to some stepsize rule and $M > 0$ is a suitable predetermined constant ensuring the existence of the descent direction v_k. If $\mathcal{G}(u) = I_K(u)$ is the indicator function of a compact convex set $K \subset \mathcal{M}$, then the iteration scheme reduces to the *Frank-Wolfe* (FW) algorithm for constrained minimization [26, 19, 24].

In the present paper we focus on positively homogeneous functionals $\mathcal{G}$ with compact sublevel sets. This encompasses the important case of norm regularization $\mathcal{G}(u) = \|u\|_\mathcal{M}$ and also allows for seminorm penalties if the problem is posed in a suitable quotient space [6]. In this setting, GCG methods benefit from several desirable properties. On the one hand, they only rely on the repeated solution of (partially) linearized problems that, in some interesting cases, can be solved analytically. On the other hand, the descent direction can be chosen to satisfy $v_k = M_k\hat{v}_k$ where $M_k \geq 0$ is a scaling factor and $\widehat{v}_k$ is an *extreme point* of the "unit ball" $B = \{\, u \mid \mathcal{G}(u) \leq 1 \,\}$ of the regularizer. Thus, the partially linearized problem can be solved equivalently in the set of extreme points of B, reducing 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 extreme points of B, the iterate u_k exhibits *k-sparsity*, i.e., it is contained in the conic hull of at most k extreme points of B. The connection between the structure enhancing properties of a functional $\mathcal{G}$ and the set of extreme points $\operatorname{Ext}(B)$ of B has been recently studied in [6] and [5]. 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 $\dim Y$ extreme points.

While the FW method [22, 23, 27, 30], and its many variants such as *away-step* FW [28, 31], or *fully-corrective* FW [29, 37, 38], have received a lot of attention, GCG algorithms for general nonsmooth functionals $\mathcal{G}$ are less frequently studied. We refer, e.g., to [8, 13, 36, 41] as well as [39, 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 [15]. Moreover, the absence of steps that remove extreme points from u_k instead of adding them often leads to clustering phenomena in practice. Thus, despite its various advantages, these shortcomings of GCG methods limits their practical utility.

The main contribution of the present work is the analysis of an *accelerated generalized conditional gradient method* (AGCG) for ($\mathcal{P}_\mathcal{M}$). It relies on the mutual update of a sequence of sparse iterates u_k and a sequence of finite, ordered *active sets* $\mathcal{A}_k = \left\{\, u_i^k \in \operatorname{Ext}(B) \mid i = 1,\dots,N_k \,\right\}$, $N_k \in \mathbb{N}$, of extreme points constituting 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

$$\mathcal{A}_k^+ = \mathcal{A}_k \cup \{\widehat{v}_k\}, \qquad \widehat{v}_k \in \operatorname*{arg\,max}_{v\in\operatorname{Ext}(B)} \langle p_k, v\rangle\,. \tag{1.1}$$

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

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

where we minimize over the cone spanned by $\mathcal{A}_k^+$, and $\mathcal{G}$ is replaced by the *gauge function* associated with $\mathcal{A}_k^+$ [3, 17], i.e.,

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

Finally, we prune $\mathcal{A}_k^+$ by removing unnecessary extreme points obtaining the next active set $\mathcal{A}_{k+1}$. As we will see, the proposed method combines the advantages of GCG methods, e.g., its global sublinear convergence, with an improved convergence behavior and sparser iterates. More in detail we show that $J(u_k)$ converges locally 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 particular, such assumptions imply that the unique minimizer to ($\mathcal{P}_{\mathcal{M}}$) is of the form $\bar{u} = \sum_{i=1}^{N} \bar{\lambda}_i \bar{u}_i$ for a finite set of extreme points $\bar{u}_i \in \operatorname{Ext}(B)$ and coefficients $\bar{\lambda}_i > 0$.
The convergence analysis of the proposed method relies on three key observations: First, using Choquet's theorem, [33, Page 14], we argue the existence of a surjective mapping $\mathcal{I}$ from the set of positive measures on $\mathcal{B} := \overline{\operatorname{Ext}(B)}^*$ to the domain of $\mathcal{G}$ defined by

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

Second, we consider an auxiliary problem

$$\inf_{\mu \in M^+(\mathcal{B})} \left[F(\mathcal{K}\mu) + \|\mu\|_{M(\mathcal{B})} \right], \tag{1.3}$$

where the forward operator K is replaced by a lifted model $\mathcal{K}\colon M^+(\mathcal{B}) \to Y$ satisfying $\mathcal{K}\mu = K\mathcal{I}(\mu)$, $\mu \in M^+(\mathcal{B})$. As it turns out, see Theorem 4.4, both 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 (1.3). 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 (1.3). Such methods are accelerated variants of the generalized conditional gradient method for minimization problems over spaces of Borel measures introduced in [14] and their linear convergence has been recently proven for measures supported on a compact subset of the euclidean space, see [25, 34]. Moreover, if we interpret (1.3) as a *sparse dictionary learning problem*, in which the dictionary is given by $\operatorname{Ext}(B)$, they can be also linked to a *fully corrective greedy selection* method [35], or an *accelerated gradient boosting* algorithm [40]. Combining these three observations and carefully extending the techniques and results of [34] to the present setting we are able to conclude the linear convergence of our algorithm.
In this paper we mainly focus on the theoretical aspects of the accelerated generalized conditional gradient method, proposing as a guiding numerical example, the problem of identifying the initial source of a heat equation from given temperature measurements. We remark however that our method is applicable to a wide range of optimization problems and regularizers promoting various structural properties in the solutions. We plan to investigate such examples in a forthcoming paper, complemented by a detailed numerical analysis [10].

1.2. **Outline.** The paper is structured as follows. In Section 2 we formulate the minimization problem ($\mathcal{P}_{\mathcal{M}}$) we are interested to solve, specifying a basic set of assumptions necessary to prove its well-posedness. Section 3 introduces the AGCG algorithm as well as sufficient conditions for its convergence. These are illustrated in more detail for a particular example in Section 3.3. The convergence proofs are broken into three parts: First, in Section 4, we introduce a "lifting" of the minimization problem ($\mathcal{P}_{\mathcal{M}}$) to a suitable space of Radon measures by virtue of Choquet's theorem leading to the equivalent auxiliary problem ($\mathcal{P}_{M^+}$). Second, we then propose an extension of the

PDAP algorithm to compute its solution in Section 4.3. It turns out that both, PDAP and AGCG, are equivalent given the correct interpretation. This equivalence will be extensively used in Section 5 to carry out the proofs of the main convergence statements. Finally, the Appendix contains the proofs of several auxiliary results.

2. The minimization problem

In this section we introduce the minimization problem we are concerned with solving. 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

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

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 ($\mathcal{P}_{\mathcal{M}}$).

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

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

See, for example, [14, 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 $K : \mathcal{M} \to Y$ is sequentially weak*-to-strong continuous in $\mathrm{dom}(\mathcal{G})$, namely, for every sequence $\{u_k\}_k$ in $\mathrm{dom}(\mathcal{G})$ such that $u_k \stackrel{*}{\rightharpoonup} u$, $u \in \mathcal{M}$, it holds that $\lim_{n\to\infty} Ku_k = Ku$ in Y.

The above conditions guarantee the existence of minimizers to ($\mathcal{P}_{\mathcal{M}}$).

Proposition 2.3. *Let Assumptions (A1)-(A3) hold. 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. Moreover, the function $f(u) = F(Ku)$ is Gâteaux-differentiable with

$$f'(u)(\delta u) = \langle K_* \nabla F(Ku), \delta u \rangle \quad \forall \delta u \in \mathcal{M}.$$

Hence, see e.g. [39, Proposition 6.3], $\bar{u} \in \mathcal{M}$ is a solution to ($\mathcal{P}_{\mathcal{M}}$) if and only if

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

This variational inequality holds if and only if

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

which is equivalent to (2.1), thanks to the one-homogeneity of $\mathcal{G}$. Last, $K\bar{u}_1 = K\bar{u}_2$ 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 $\bar{y} := K\bar{u} \in Y$ and $\bar{p} = -K_* \nabla F(K\bar{u}) \in \mathcal{C}$ as the (unique) optimal observation and dual variable for Problem ($\mathcal{P}_{\mathcal{M}}$), respectively. Set $B = S^-(\mathcal{G}, 1)$. In the following we further require the notion of an *extreme point* of B.

Definition 2.4. An element $u \in B$ is called an extreme point of B if there are *no* $u_1, u_2 \in B$, $s \in (0, 1)$ with $u = (1 - s)u_1 + su_2$.

The set of all extreme points of B is denoted by $\operatorname{Ext}(B)$. Since B is weak* compact, non empty and convex thanks to Assumptions (A1)-(A3), by the Krein-Milman Theorem we infer that $\operatorname{Ext}(B) \neq \emptyset$ and for every $u \in \operatorname{dom}(\mathcal{G})$ there is $\{u_k\}_k$ in $\operatorname{cone}(\operatorname{Ext}(B))$ with $u_k \overset{*}{\rightharpoonup} u$. Set $\mathcal{B} := \overline{\operatorname{Ext}(B)}^*$. 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

$$u_k \overset{*}{\rightharpoonup} u \ \text{ if and only if } \ \lim_{k \to \infty} d_{\mathcal{B}}(u_k, u) = 0 \,. \tag{2.2}$$

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

Remark 2.5. Let us mention that all results in this paper still hold true 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 the domain $\operatorname{dom}(F) = \{\, y \mid F(y) < \infty \,\}$ is open. In this case, F is required to be smooth on $\operatorname{dom}(F)$ and $\nabla F \colon \operatorname{dom}(F) \to Y$ is assumed to be Lipschitz continuous on compact subsets of $\operatorname{dom}(F)$. For more details, we refer the interested reader to [39] and [34].

3. A numerical minimization algorithm

This section puts the focus on developing an implementable and efficient solution algorithm for the minimization problem ($\mathcal{P}_{\mathcal{M}}$). For this purpose, recall that every $u \in \operatorname{dom}(\mathcal{G})$ can be approximated by a sequence of finite conic combinations of elements in $\operatorname{Ext}(B)$ by Krein-Milman theorem. The considered method exploits this observation by alternating between the update of a finite set

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

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

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

where $\mathbb{R}_+ = [0, +\infty)$. For the convenience of the reader we give a short description of the individual steps of the method in the following. Given the current active set $\mathcal{A}_k^u = \{u_i^k\}_{i=1}^{N_k}$, $u_i^k \in \operatorname{Ext}(B)$, and the iterate $u_k \in \operatorname{cone}(\mathcal{A}_k^u)$ we first compute the dual variable $p_k = -K_* \nabla F(Ku_k) \in \mathcal{C}$ and update

$$\mathcal{A}_k^{u,+} = \mathcal{A}_k^u \cup \{\widehat{v}_k^u\}, \qquad \widehat{v}_k^u \in \operatorname*{arg\,max}_{v \in \operatorname{Ext}(B)} \langle p_k, v \rangle \,. \tag{3.1}$$

This step requires the minimization of a linear functional over the not necessarily compact set of extreme points. The following lemma addresses the well-posedness of this auxiliary problem.

Lemma 3.1. *Let $p \in \mathcal{C}$ be given. Then, there exists $\bar{v} \in \operatorname{Ext}(B)$ 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

$$\max_{v \in B} \langle p, v \rangle \tag{3.2}$$

admits a solution $\bar{v} \in \operatorname{Ext}(B)$. This statement is classical, but for the reader's convenience we produce a proof. Let v^* be a maximizer of (3.2), 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 $\operatorname{Ext}(H) \neq \emptyset$ by Krein-Milman's Theorem, since $H \subset B$ is convex, non-empty, weak* closed and thus weak* compact. Thus, let $\bar{v} \in \operatorname{Ext}(H)$ be given. We claim that $\bar{v} \in \operatorname{Ext}(B)$. Indeed, assume that $\bar{v} = sv_1 + (1-s)v_2$ 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 $\bar{v} \in H$. As $\bar{v} \in \operatorname{Ext}(H)$ we then have $\bar{v} = v_1 = v_2$, showing that $\bar{v} \in \operatorname{Ext}(B)$. The claimed statement now follows from $\operatorname{Ext}(B) \subset \mathcal{B} \subset B$ ending the proof. □

Setting $N_k^+ = N_k + 1$ and $u^k_{N_+^k} = \widehat{v}_k^u$ we now introduce the gauge function associated with $\mathcal{A}_k^{u,+}$ as

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

Note that $\kappa_{\mathcal{A}_k^{u,+}}(u)$ is well-defined on $\operatorname{cone}(\mathcal{A}_k^{u,+})$ due to the lower semicontinuity of the $|\cdot|_{\ell^1}$-norm and the closedness of $\mathbb{R}_+$.

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

$$\min_{u \in \operatorname{cone}(\mathcal{A}_k^{u,+})} \left[F(Ku) + \kappa_{\mathcal{A}_k^{u,+}}(u) \right], \tag{3.3}$$

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. We now show that a solution of (3.3) can be computed as $u_{k+1} = \sum_{i=1}^{N_k^+} \lambda_i^{k+1} u_i^k$ where λ^{k+1} solves the following finite dimensional optimization problem:

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

This is the content of the following proposition.

Proposition 3.2. *Let $\mathcal{A}_k^{u,+} = \{u_i^k\}_{i=1}^{N_k^+} \subset \operatorname{Ext}(B)$ be given. If $\widehat{\lambda} \in \mathbb{R}_+^{N_k^+}$ is a minimizer to (3.4), then*

$$\widehat{u} = \sum_{i=1}^{N_k^+} \widehat{\lambda}_i u_i^k \quad \textit{solves} \quad (3.3).$$

Vice versa, if $\widehat{u}$ is a minimizer to (3.3), 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 $\widehat{\lambda} \in \mathbb{R}_+^{N_k^+}$ denote an arbitrary solution to (3.4). Set $\widehat{u} = \sum_{i=1}^{N_k^+} \widehat{\lambda}_i u_i^k$ and fix an arbitrary $u \in \operatorname{cone}(\mathcal{A}_k^{u,+})$. Then there exists $\lambda^u \in \mathbb{R}_+^{N_k^+}$ such that

$$u = \sum_{i=1}^{N_k^+} \lambda_i^u u_i^k\,, \qquad \sum_{i=1}^{N_k^+} \lambda_i^u = \kappa_{\mathcal{A}_k^+}(u)\,.$$

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 (3.3).
Vice versa, let $\widehat{u}$ be a minimizer to (3.3). Since $\widehat{u} \in \operatorname{cone}(\mathcal{A}_k^{u,+})$, there exists $\widehat{\lambda} \in \mathbb{R}_+^{N_k^+}$ with

$$\widehat{u} = \sum_{i=1}^{N_k^+} \widehat{\lambda}_i u_i^k\,, \qquad \sum_{i=1}^{N_k^+} \widehat{\lambda}_i = \kappa_{\mathcal{A}_k^{u,+}}(\widehat{u})\,.$$

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

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

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

Thus, in practice, we first compute a minimizer $\lambda^{k+1} \in \mathbb{R}^{N_k^+}$ to (3.4). This constitutes a finite dimensional nonsmooth 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, we set $u_{k+1} = \sum_{i=1}^{N_k^+} \lambda_i^{k+1} u_i^k$ that is a minimizer of (3.3) due to Proposition 3.2. As a final step, we truncate the active set $\mathcal{A}_k^{u,+}$ by removing all extreme points that were assigned a zero weight, 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.
Note that the method terminates at step $k \geq 1$ if the maximum in (3.1) is smaller or equal to one. This is justified by the following proposition.

Proposition 3.3. *Let $u_k \in \mathcal{A}_k^u$ be generated by Algorithm 1, $k \geq 1$. Set $p_k = -K_* \nabla F(Ku_k)$ and either assume $\max_{v \in \operatorname{Ext}(B)} \langle p_k, v\rangle \leq 1$ or $\widehat{v}_k^u \in \mathcal{A}_k^u$. Then u_k is a minimizer to ($\mathcal{P}_\mathcal{M}$).*

Proof. First, assume that $\max_{v \in \operatorname{Ext}(B)} \langle p_k, v\rangle \leq 1$. We recall that, thanks to Proposition 2.3 and Lemma 3.1, u_k is a solution to ($\mathcal{P}_\mathcal{M}$) if and only if

$$\langle p_k, u_k\rangle = \mathcal{G}(u_k)\,, \qquad \max_{v \in \operatorname{Ext}(B)} \langle p_k, v\rangle \leq 1\,. \tag{3.5}$$

Thus, if $u_k = 0$ the desired statement immediately follows from $\mathcal{G}(0) = 0$. Without loss of generality assume $u_k \neq 0$. By assumption and due to Lemma 3.1 we have

$$\max_{v \in B} \langle p_k, v\rangle = \max_{v \in \operatorname{Ext}(B)} \langle p_k, v\rangle \leq 1.$$

Algorithm 1 AGCG for ($\mathcal{P}_\mathcal{M}$)

1. Let $u_0 = \sum_{i=1}^{N_0} \lambda_i^0 u_i^0$, $\lambda_i^0 > 0$, $\mathcal{A}_0 = \{u_i^0\}_{i=1}^{N_0} \subset \operatorname{Ext}(B)$.

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

2. Given $\mathcal{A}_k^u = \{u_i^k\}_{i=1}^{N_k} \subset \operatorname{Ext}(B)$ and u_k, calculate $\widehat{v}_k^u \in \operatorname{Ext}(B)$ with

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

if $\langle p_k, \widehat{v}_k^u \rangle \leq 1$ or $\widehat{v}_k^u \in \mathcal{A}_k^u$ **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} from solving (3.4) and set $u_{k+1} = \sum_{i=1}^{N_k^+} \lambda_i^{k+1} u_i^k$.

6. Update

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

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

end for

Consequently, since $u_k / \mathcal{G}(u_k) \in B$ we also get

$$\langle p_k, u^k \rangle \leq \mathcal{G}(u^k). \tag{3.6}$$

By construction we further have $u_k = \sum_{i=1}^{N_k} \lambda_i^k u_i^k$, for a minimizer $\lambda^k \in \mathbb{R}_+^{N_k}$ 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 λ^k is a minimizer to this problem if and only if

$$\langle p_k, u_k \rangle = \sum_{i=1}^{N_k} \lambda_i^k \langle p_k, u_i^k \rangle = \sum_{i=1}^{N_k} \lambda_i^k, \; \langle p_k, u_i^k \rangle \leq 1 \tag{3.7}$$

for all $i = 1, \dots, N$. Using the convexity and the 1-homogeneity of $\mathcal{G}$ we thus estimate

$$\mathcal{G}(u_k) \leq \sum_{i=1}^{N_k} \lambda_i^k \mathcal{G}(u_i^k) \leq \sum_{i=1}^{N_k} \lambda_i^k \,.$$

Combining these observations with (3.6) we finally have that $\langle p_k, u_k \rangle = \mathcal{G}(u_k)$ which implies that u_k is a minimizer of ($\mathcal{P}_\mathcal{M}$) by the optimality conditions in (3.5).
Now, if $\widehat{v}_k^u \in \mathcal{A}_k^u$, we have

$$\max_{v \in \operatorname{Ext}(B)} \langle p_k, v \rangle = \langle p_k, \widehat{v}_k^u \rangle \leq 1$$

due to (3.7). Thus, the optimality of u_k follows from the previous arguments. □

3.1. **Worst-case convergence rates.** The main contribution of the present manuscript is the derivation of convergence results for the sequence of *residuals*

$$r_J(u_k) = J(u_k) - \min_{u \in \mathcal{M}} J(u) \tag{3.8}$$

associated to the iterates generated by Algorithm 1. At a first glance this is a challenging task for a variety of reasons. For example the space $\mathcal{M}$ generally lacks useful properties such as reflexivity, smoothness or strict convexity. Hence many well-known analytic tools and proof techniques do not directly translate to the presented method. For the convenience of the reader, in this section, we restrict ourselves to stating the relevant convergence results as well as detailing the additional assumptions which are needed to prove them. Their technical proofs are then postponed to Section 5.1 and 5.3. As a first step, we start with the following sublinear convergence result which holds without additional assumptions on the problem.

Theorem 3.4. *Let Assumptions (A1)-(A3) 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*

$$r_J(u_k) \leq c\,\frac{1}{k+1} \quad \text{for all} \;\; k \in \mathbb{N}\,, \tag{3.9}$$

where r_J is defined at (3.8). 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.*

3.2. **Non-degeneracy and fast convergence.** While Theorem 3.4 proves the convergence of Algorithm 1, the provided slow, sublinear rate of convergence does not match the computed results of Section 3.3. 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 extreme points $\bar{u}_1, \ldots, \bar{u}_N \in \operatorname{Ext}(B)$, where $\bar{p}$ denotes the unique dual variable of the problem (the uniqueness follows from Proposition 2.3). Moreover, we ask that the restriction of the operator K onto 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 3.6). Additionally, we ask that F is strongly convex around the unique optimal observation. This set of assumptions is stated below.

Assumption 3.5. *(Uniqueness and strong convexity)*

(B1) The map $F \colon Y \to \mathbb{R}$ is strictly convex and strongly convex around $\bar{y}$, 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 $N > 0$ as well as $\{\bar{u}_i\}_{i=1}^N \subset \operatorname{Ext}(B)$ such that the unique dual variable $\bar{p} = -K_* \nabla F(\bar{y}) \in \mathcal{C}$ satisfies

$$\operatorname*{arg\,max}_{v \in \mathcal{B}} \langle \bar{p}, v \rangle = \{\, v \in \mathcal{B} \mid \langle \bar{p}, v \rangle = 1 \,\} = \{\bar{u}_i\}_{i=1}^N\,, \tag{3.10}$$

(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 3.6. *Suppose that Assumptions (B1)-(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 (3.10).

Proof. Arguing similarly to Lemma 3.1 we get

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

and thus, in particular, $\mathcal{G}(u) = 1$ for all $u \in \operatorname{conv}(\bar{\mathcal{A}})$. Hence, recalling the first order optimality conditions from Proposition 2.3, every minimizer $\bar{u}$ to $(\mathcal{P}_{\mathcal{M}})$ satisfies $\bar{u} \in \operatorname{conv}(\bar{\mathcal{A}})$, i.e., there exist $\bar{\lambda}_i \geq 0$ such that $\bar{u} = \sum_{i=1}^N \bar{\lambda}_i \bar{u}_i$. Now assume that

$$\bar{u}_1 = \sum_{i=1}^N \bar{\lambda}_i \bar{u}_i\,, \quad \bar{u}_2 = \sum_{i=1}^N \bar{\gamma}_i \bar{u}_i\,, \quad \bar{\lambda}_i, \bar{\gamma}_i \geq 0$$

are two minimizers of $(\mathcal{P}_{\mathcal{M}})$. Since $K\bar{u}_1 = K\bar{u}_2$ we have

$$0 = K(\bar{u}_1 - \bar{u}_2) = \sum_{i=1}^N (\bar{\lambda}_i - \bar{\gamma}_i) K\bar{u}_i\,.$$

From (B3) we then conclude $\bar{\lambda}_i = \bar{\gamma}_i$ and so $\bar{u}_1 = \bar{u}_2$, implying that $(\mathcal{P}_{\mathcal{M}})$ has a unique solution and such a solution is of the form (3.11). □

In the next set of assumptions we assume *strict complementarity*, i.e.,

$$\bar{u} \notin \operatorname{cone}\left(\bar{\mathcal{A}} \setminus \{\bar{u}_i\}\right)$$

for every $i = 1, \dots, N$ or, equivalently, $\bar{\lambda}_i > 0$ for all $i = 1, \dots, N$. The final assumption concerns the existence of a "distance function" g such that $K_{|\operatorname{Ext}(B)}$ is Lipschitz continuous and the linear functional in grows quadratically with respect to g 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 Section 3.3. This set of assumptions is stated below.

Assumption 3.7. *(Non-degeneracy)*

- (B4) The unique minimizer $\bar{u}$ of $(\mathcal{P}_{\mathcal{M}})$ is such that $\bar{\lambda}_i > 0$ for all $i = 1, \dots, N$.
- (B5) There exists a function $g\colon \operatorname{Ext}(B) \times \operatorname{Ext}(B) \to [0, \infty)$, 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

$$\|K(u - \bar{u}_i)\|_Y \leq \tau g(u, \bar{u}_i), \quad 1 - \langle \bar{p}, u\rangle \geq \kappa g(u, \bar{u}_i)^2 \quad \text{for all } u \in U_i\,, \tag{3.12}$$

for all $i = 1, \dots, N$, where $U_i := \bar{U}_i \cap \operatorname{Ext}(B)$ and $d_{\mathcal{B}}$ is the metric at (2.2).

Note that, w.l.o.g., we can always choose $\{\bar{U}_i\}_{i=1}^N$ for which there exists a constant $\sigma > 0$ such that

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

due to Assumption (B2) and the $d_{\mathcal{B}}$-continuity of $u \mapsto \langle \bar{p}, u\rangle$.
We arrive at the following improved convergence result. For its proof we refer to Section 5.3.

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 minimizer to $(\mathcal{P}_{\mathcal{M}})$ or there exists a constant $c > 0$ and $\zeta \in [1/2, 1)$ such that*

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

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

3.3. **A guiding example.** We briefly outline the application of Algorithm 1 to a concrete problem formulated in our setting. For this purpose, 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. Note that the coefficients $\lambda_i^\dagger \in \mathbb{R}$, the positions $x_i^\dagger \in \Omega$ as well as the unknown number $N \in \mathbb{N}$ of points are assumed to be unknown. Taking the ill-posedness of the described inverse problem into account we follow [14, 32] and consider the convex Tikhonov-regularized problem

$$\min_{u \in M(\Omega), y} \left[\frac{1}{2} \|y(T) - y_d\|^2_{L^2(\Omega)} + \beta \|u\|_{M(\Omega)} \right] \tag{3.15}$$

where $M(\Omega)$ denotes the space of Borel measures on the open set Ω, $y_d \in L^2(\Omega)$ is a given desired state and the pair (y, u) satisfies the heat equation

$$\partial_t y - \triangle y = 0 \text{ in } (0,T) \times \Omega, \;\; y = 0 \text{ in } (0,T) \times \partial\Omega, \;\; y(0) = u \text{ in } \Omega\,. \tag{3.16}$$

Here, 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 (3.15) into the setting of ($\mathcal{P}_\mathcal{M}$) we set $\mathcal{C} = C_0(\Omega)$, the space of continuous functions which are zero on $\partial\Omega$, and equip it with the canonical supremum norm

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

which makes it a Banach space. According to the Riesz-Markov-Kakutani theorem we then have $\mathcal{C}^* \simeq \mathcal{M}$ for $\mathcal{M} = M(\Omega)$. Moreover, define $Y = L^2(\Omega)$, $F = (1/2)\| \cdot - y_d\|^2_{L^2(\Omega)}$ and $\mathcal{G} = \beta \| \cdot \|_{M(\Omega)}$. Of course, these functionals satisfy (A1) and (A2) in Assumption 2.2. Finally, we eliminate 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 (3.16). It is readily verified that K is injective, thanks to a priori estimates for weak solutions to (3.16) [16, Lemma 2.2], as well as weak*-to-strong continuous [16, Lemma 2.3]. Hence, (3.15) admits a unique solution and (A3) in Assumption 2.2 is satisfied. Moreover, K is the adjoint of the operator $K_* \colon L^2(\Omega) \to C_0(\Omega)$ with $K_* \varphi = z(0)$ where z satisfies the backwards heat equation

$$\partial_t z + \triangle z = 0 \text{ in } (0,T) \times \Omega, \;\; z = 0 \text{ in } (0,T) \times \partial\Omega, \;\; z(T) = \varphi \text{ in } \Omega \tag{3.17}$$

for $\varphi \in L^2(\Omega)$ in the weak sense. For more details we refer to [32, 16]. 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 by the following well-known result that will be used also in the remaining part of this section.

Lemma 3.9. *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 (3.17). 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 immediately from [32, Lemma 3.1] and the Sobolev embedding $H^2(\Omega) \cap H_0^1(\Omega) \hookrightarrow C_0(\Omega)$. ☐

The next lemma characterizes the set of extreme points of the ball $B = S^-(\beta \| \cdot \|_{M(\Omega)}, 1)$ in this case.

Lemma 3.10. *We have*

$$\operatorname{Ext}(B) = \{\sigma\beta^{-1}\delta_x : x \in \Omega,\, \sigma \in \{-1, 1\}\}. \tag{3.18}$$

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

Proof. The characterization of $\operatorname{Ext}(B)$ is a well-known result (see, for example, Proposition 4.1 in [6]). As for the second claim, the inclusion $\operatorname{Ext}(B) \cup \{0\} \subset \overline{\operatorname{Ext}(B)}^*$ follows immediately: if $\{x_k\}_k$ in Ω is such that $x_k \to x$ with $x \in \partial\Omega$, then $u_k := \beta^{-1}\delta_{x_k} \overset{*}{\rightharpoonup} 0$. For the opposite inclusion, assume given a sequence $u_k = \sigma_k\beta^{-1}\delta_{x_k}$ in $\operatorname{Ext}(B)$, such that $u_k \overset{*}{\rightharpoonup} u$. 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 $u = \sigma\beta^{-1}\delta_x \in \operatorname{Ext}(B)$. □

Moreover, applying Proposition 2.3 and the characterization of K_*, we immediately deduce the optimality conditions for (3.15).

Proposition 3.11. *Let $\bar{u} \in M(\Omega)$ be given and denote by $\bar{z}$ the solution to (3.17) for $\varphi = y_d - \bar{y}$, $\bar{y} = K\bar{u}$. Then $\bar{u}$ is a minimizer of (3.15) if and only if*

$$\langle \bar{z}(0), \bar{u} \rangle = \beta\|\bar{u}\|_{M(\Omega)}, \;\; \|\bar{z}(0)\|_{\mathcal{C}} \le \beta\,.$$

This implies

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

Thanks to the characterization of extreme points presented in Lemma 3.10, the method presented in Algorithm 1 for solving (3.15) generates a sequence of active sets $\mathcal{A}_k^u = \{\sigma_i^k\beta^{-1}\delta_{x_i^k}\}_{i=1}^{N_k}$, $\sigma_i^k \in \{-1, 1\}$, $x_i^k \in \Omega$ comprised of (signed) Dirac deltas as well as an associated sequence of sparse iterates $u_k = \sum_{i=1}^{N_k} \lambda_i^k\sigma_i^k\beta^{-1}\delta_{x_i^k}$, where $\lambda_i^k \in \mathbb{R}_+$. Moreover, let z_k denote the solution of (3.17) for $\varphi = y_d - y_k(T)$, $y_k(T) = Ku_k$. We now claim that the new candidate extreme point in the iteration k of Algorithm 1 can be chosen as

$$\widehat{v}_k^u = \operatorname{sign}(z_k(0)(\hat{x}_k))\beta^{-1}\delta_{\hat{x}_k} \;\text{ where }\; \hat{x}_k \in \Omega \;\text{ satisfies }\; |z_k(0)(\hat{x}_k)| = \max_{x\in\Omega} |z_k(0)(x)|, \tag{3.19}$$

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 3.12. *Let $\widehat{v}_k^u$ be defined as in (3.19). 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 $z_k(0) = K_*(y_k(T) - y_d) = -K_*\nabla F(Ku_k)$ from the characterization of K_*. The remaining statement follows directly from

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

as well as

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

Now, before presenting the numerical experiments, let us briefly discuss the non-degeneracy conditions in Assumptions (B1)-(B5) from Section 3.2 and, in particular, the choice of the distance function g in this setting. In particular, denoting by $\bar{u} \in M(\Omega)$ the minimizer of (3.15), we propose a natural and easy to verify set of assumptions for $\bar{z}(0)$ that implies our general non-degeneracy assumptions from Section 3.2 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

(3.15). We finally remind that $\bar{z}$ is the solution to (3.17) 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 [32, 34].

Assumption 3.13. Assume that:

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

$$\{\, x \in \Omega \mid |\bar{z}(0)(x)| = \beta \,\} = \{\bar{x}_i\}_{i=1}^N. \tag{3.20}$$

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

$$\operatorname{sign}(\bar{z}(0)(\bar{x}_i))(\delta x, \nabla^2 \bar{z}(0)(\bar{x}_i)\delta x)_{\mathbb{R}^2} \le -\gamma|\delta x|^2 \quad \forall \delta x \in \mathbb{R}^2.$$

Regarding (C2) recall that $\bar{z}(0)$ is at least two times continuously differentiable in the vicinity of $\bar{x}_i$, $i = 1, \ldots, N$, see Lemma 3.9. 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 show that (C1) guarantees Assumptions (B2)-(B3). For this purpose, set $\bar{u}_i = \operatorname{sign}(\bar{z}(0)(\bar{x}_i))\beta^{-1}\delta_{\bar{x}_i} \in \operatorname{Ext}(B)$ for every $i = 1, \ldots, N$.

Proposition 3.14. *Let Assumption (C1) hold. Then, we have*

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

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

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

$$\langle \bar{z}(0), \sigma\beta^{-1}\delta_x \rangle = (\sigma/\beta)\bar{z}(0)(x) \le \|\bar{z}(0)\|_{\mathcal{C}}/\beta = 1,$$

see Proposition 3.11 and (3.20), with equality if and only if $|\bar{z}(0)(x)| = \|\bar{z}(0)\|_{\mathcal{C}}$, $\sigma = \operatorname{sign}(\bar{z}(0)(x))$. Hence, the claimed statement follows from (3.20) and Lemma 3.10. Finally, the linear independence of $\{K\bar{u}_i\}_{i=1}^N$ follows from the injectivity of K. ☐

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

$$\|\psi\|_{\operatorname{Lip}(\Omega_0)} := \sup_{x,y \in \Omega_0, x \neq y} \frac{|\psi(x) - \psi(y)|}{|x - y|} \tag{3.21}$$

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

$$\|K_*\|_{Y,\operatorname{Lip}(\Omega_0)} := \sup_{\|y\|_Y \le 1} \|K_* y\|_{\operatorname{Lip}(\Omega_0)} . \tag{3.22}$$

Note that $\|K_*\|_{Y,\operatorname{Lip}(\Omega_0)} < \infty$, due to Lemma 3.9. In the next lemma we show that Assumptions (C1)-(C2) imply the quadratic growth of $\beta - |\bar{z}(0)|$ around $\{\bar{x}_i\}_{i=1}^N$.

Lemma 3.15. *Let Assumption 3.13 hold and fix an index $i = 1, \ldots, N$. Then, there exists $R > 0$ such that $\bar{B}_R(\bar{x}_i) \subset \Omega$, $\operatorname{sign}(\bar{z}(0)(x)) = \operatorname{sign}(\bar{z}(0)(\bar{x}_i))$ for all $x \in B_R(\bar{x}_i)$ and*

$$\beta - |\bar{z}(0)(x)| \ge (\gamma/4)|x - \bar{x}_i|^2 \quad \text{for all } x \in B_R(\bar{x}_i)\,, \tag{3.23}$$

$$\|K(\delta_x - \delta_{\bar{x}_i})\|_Y \le \|K_*\|_{Y,\operatorname{Lip}(B_R(\bar{x}_i))}|x - \bar{x}_i|, \quad \text{for all } x \in B_R(\bar{x}_i)\,. \tag{3.24}$$

Proof. We start by noting that $\nabla \bar{z}(0)(\bar{x}_i) = 0$ thanks to (C1) and Proposition 3.11. 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 $\nabla^2 \bar{z}(0)$ (Lemma 3.9) as well as Assumption (C2), there exists $R > 0$ such that $B_R(\bar{x}_i) \subset \Omega$ and

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

Combining both observations finally yields

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

showing (3.23). 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 $\operatorname{sign}(\bar{z}(0)(x)) = \operatorname{sign}(\bar{z}(0)(\bar{x}_i))$ for all $x \in B_R(\bar{x}_i)$ due to $\bar{z}(0)(\bar{x}_i) \neq 0$ and $\bar{z}(0) \in C_0(\Omega)$. □

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

$$g(u_1, u_2) = |\sigma_1 - \sigma_2| + |x_1 - x_2|\,. \tag{3.25}$$

Such g will be the one verifying Assumption (B3). In the next lemma we show that the weak* convergence in $M(\Omega)$ of a sequence of extreme points to $\bar{u}_i$ is equivalent to convergence w.r.t. g.

Lemma 3.16. *Consider a sequence $\{u_k\}_k$ in $\operatorname{Ext}(B)$, 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 $\sigma_k = \operatorname{sign}(\bar{z}(0)(\bar{x}_i))$ for all $k \in \mathbb{N}$ large enough.

Proof. First assume that $\lim_{k\to\infty} g(u_k, \bar{u}_i) = 0$. Then, we have $x_k \to \bar{x}_i$ in Ω as well as $\sigma_k \to \operatorname{sign}(\bar{z}(0)(\bar{x}_i))$ and consequently

$$\langle p, u_k\rangle = \sigma_k \beta^{-1} p(x_k) \to \operatorname{sign}(\bar{z}(0)(\bar{x}_i))\beta^{-1}p(\bar{x}_i) = \langle p, \bar{u}_i\rangle \quad \forall p \in C_0(\Omega)\,.$$

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

$$\bar{\sigma}\beta^{-1}p(\bar{x}) = \operatorname{sign}(\bar{z}(0)(\bar{x}_i))\beta^{-1}p(\bar{x}_i) \quad \text{for all } p \in C_0(\Omega)$$

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

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

Proposition 3.17. *Let $R > 0$ be chosen according to Lemma 3.15. Then, there is a $d_{\mathcal{B}}$-neighbourhood $\bar{U}_i$ of $\bar{u}_i$ which satisfies*

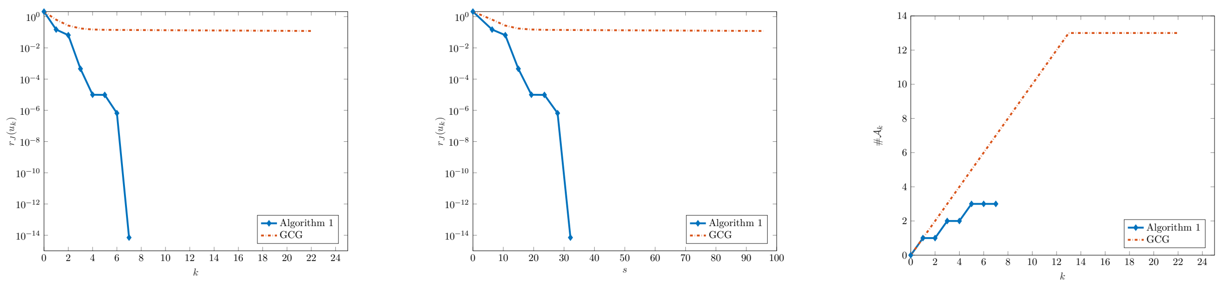
$$U_i := \bar{U}_i \cap \operatorname{Ext}(B) \subset \left\{ \sigma\beta^{-1}\delta_x \mid \sigma = \operatorname{sign}(\bar{z}(0)(\bar{x}_i)),\ x \in B_R(\bar{x}_i) \right\}$$

as well as

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

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

Proof. The statement on the existence of a $d_{\mathcal{B}}$-neighbourhood $\bar{U}_i$ with the stated properties follows immediately from Lemma 3.16. In fact, if such a neighbourhood does not exist, then there exists a sequence $\{u_k\}_k$ in $\operatorname{Ext}(B)$, i.e., $u_k = \sigma_k\beta^{-1}\delta_{x_k}$, with $u_k \xrightarrow{*} \bar{u}_i$ as well as $\sigma_k \ne \operatorname{sign}(\bar{z}(0)(\bar{x}_i))$ 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 3.15 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 $\langle \bar{z}(0), \bar{u}_i\rangle = 1$. □



(A) Residual over k. (B) Residual over time. (C) Active set size over k.

FIGURE 1. Convergence behaviour of relevant quantities.

We close this section by a numerical experiment. 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 $h = 1/128$. 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 CPU@2.20 GHz.
Now, in Figure 1a, 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 $\mathcal{A}_0 = \{\}$. In each iteration, the finite dimensional subproblem (3.3) is solved using a semismooth Newton method. Moreover, we plot the size of the support of u_k in dependence of k in Figure 1c. 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

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

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 [14]. Both methods were run for a maximum of 100 iterations or until $r_J(u_k) \le 10^{-12}$. 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 $k = 1, \dots, 20$. 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 (3.3) is outweighed by its practical utility. More in detail, Algorithm 1 converges after around 30s while GCG fails to decrease $r_J(u_k)$ below 10^{-1} in the considered time frame.

4. A lifting to a space of measures

Our approach to proving Theorem 3.4 and Theorem 3.8 relies on the observation of *Choquet's Theorem*. This classical result allows us to prove that for every $u \in \operatorname{dom}(\mathcal{G})$ there exists a positive measure μ supported on $\operatorname{Ext}(B)$ such that $\mathcal{G}(u) = \|\mu\|_{M(\mathcal{B})}$ and

$$\langle p, u \rangle = \int_{\mathcal{B}} \langle p, v \rangle \, \mathrm{d}\mu(v) \quad \text{for all} \quad p \in \mathcal{C} \tag{4.1}$$

hold (see Proposition 4.2). Motivated by this observation, we study an auxiliary problem

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

where the infimum is taken over $M^+(\mathcal{B})$, the cone of positive measures on $\mathcal{B}$, instead of $\mathcal{M}$. Note that the forward operator is replaced by a "lifted" mapping $\mathcal{K}$ which satisfies $\mathcal{K}\mu = Ku$ whenever (4.1) holds for $\mu \in M^+(\mathcal{B})$ and $u \in \mathcal{M}$. Moreover, the role of the nonsmooth term is now played by the total variation norm. It turns out (see Section 4.1) 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 Section 4.3, we propose an extension of the Primal-Dual-Active-Point method from [34] to compute the solution of ($\mathcal{P}_{M^+}$) and we describe it in Algorithm 2. Finally, using the results of Section 4.1 we show that Algorithm 2 and Algorithm 1 are equivalent, setting the ground to prove Theorem 3.4 and Theorem 3.8 by using convergence results for Algorithm 2.

4.1. **An equivalent problem.** To start, let $C(\mathcal{B})$ denote the vector space of real valued bounded continuous functions over $\mathcal{B}$, which we equip with the supremum norm

$$\|P\|_{C(\mathcal{B})} := \max_{u \in \mathcal{B}} |P(u)|, \tag{4.2}$$

making it a Banach space. Following the definitions of [2], we denote by $\varPi$ 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 $\mu \colon \varPi \to \mathbb{R}$. We call μ positive if $\mu(\varPi) \subset [0, \infty)$. Given a finite Radon measure μ its total variation measure $|\mu|$ is defined by

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

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 the norm (4.2): 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. Moreover, we require the following definition.

Definition 4.1. We say that a measure $\mu \in M^+(\mathcal{B})$ *represents* $u \in \mathcal{M}$ if

$$\langle p, u\rangle = \int_{\mathcal{B}} \langle p, v\rangle \,\mathrm{d}\mu(v)\,, \quad \text{for all} \quad p \in \mathcal{C}\,. \tag{4.3}$$

An element $u \in \mathcal{M}$ such that (4.3) holds is also called the *(weak) barycenter* of μ in $\mathcal{B}$.

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

Proposition 4.2. *There exists a surjective linear map $\mathcal{I}\colon M^+(\mathcal{B}) \to \mathrm{dom}(\mathcal{G})$ such that*

$$\langle p, \mathcal{I}(\mu)\rangle = \int_{\mathcal{B}} \langle p, v\rangle \,\mathrm{d}\mu(v)\,, \quad \textit{for all} \quad p \in \mathcal{C}\,. \tag{4.4}$$

In particular, μ represents $\mathcal{I}(\mu)$ in the sense of (4.3) and it holds:

i) For any $\mu \in M^+(\mathcal{B})$ we have $\mathcal{G}(\mathcal{I}(\mu)) \le \|\mu\|_{M(\mathcal{B})}$.

ii) For any $u \in \mathrm{dom}(\mathcal{G})$ there exists $\mu \in M^+(\mathcal{B})$ concentrated on $\mathrm{Ext}(B)$ such that

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

For a proof of the above statement we refer the reader to Appendix A.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 A.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 A.1.

Proposition 4.3. *There exists a linear continuous operator $\mathcal{K}\colon M(\mathcal{B}) \to Y$ such that*

$$(\mathcal{K}\mu, y)_Y = \int_{\mathcal{B}} (Kv, y)_Y \,\mathrm{d}\mu(v)\,, \quad \textit{for all} \quad y \in Y\,, \ \mu \in M(\mathcal{B})\,. \tag{4.5}$$

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

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

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

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

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

If in addition Y is separable, then (4.5) holds in the strong sense, that is,

$$\mathcal{K}\mu = \int_{\mathcal{B}} Kv \,\mathrm{d}\mu(v)\,, \tag{4.6}$$

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 4.3. 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 4.2.

Theorem 4.4. *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

$$\min_{u\in\mathcal{M}} J(u) = \min_{\mu\in M^+(\mathcal{B})} j(\mu)\,. \tag{4.7}$$

Proof. First recall that $\mathcal{K}$ is weak*-to-strong continuous thanks to Proposition 4.3. 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 $\bar{u} \in \operatorname{dom}(\mathcal{G})$. According to ii) in Proposition 4.2 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 $u \in \operatorname{dom}(\mathcal{G})$, μ represents u and $\mathcal{G}(u) \leq \|\mu\|_{M(\mathcal{B})}$. Finally, point ii) in Proposition 4.3 guarantees $\mathcal{K}\mu = Ku$ and $\mathcal{K}\bar{\mu} = K\bar{u}$. Thus, we infer

$$j(\bar{\mu}) = J(\bar{u}) \leq J(u) \leq j(\mu)\,, \tag{4.8}$$

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 4.2, we have that $\bar{u} \in \operatorname{dom}(\mathcal{G})$, $\bar{\mu}$ represents $\bar{u}$ and $\mathcal{G}(\bar{u}) \leq \|\bar{\mu}\|_{M(\mathcal{B})}$. Moreover, let $u \in \operatorname{dom}(\mathcal{G})$. By point ii) in Proposition 4.2, 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

$$J(\bar{u}) \leq j(\bar{\mu}) \leq j(\mu) = J(u)\,. \tag{4.9}$$

The final part of the statement follows from (4.8)-(4.9). □

4.2. **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 4.5. *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*

$$P(v) = \langle p, v\rangle \quad \textit{for all} \quad v \in \mathcal{B}\,. \tag{4.10}$$

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

$$P(\bar{v}) = \max_{v\in\mathcal{B}} P(v) = \max_{v\in\mathcal{B}} \langle p, v\rangle\,. \tag{4.11}$$

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 (4.3). Moreover μ represents u by Proposition 4.2. Applying point ii) in Proposition 4.3 yields $\mathcal{K}\delta_v = Kv$ and $\mathcal{K}\mu = Ku$. Thus, by iv) in Proposition 4.3,

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

yielding (4.10). The statement of (4.11) now follows from Lemma 3.1. □

Theorem 4.6. *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*

$$\bar P(v) \le 1 \quad \text{for all} \quad v \in \mathcal{B}\,, \quad \langle\!\langle \bar P, \bar\mu \rangle\!\rangle = \|\bar\mu\|_{M(\mathcal{B})}\,. \tag{4.12}$$

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 $M^+(\mathcal{B})$. Then $\bar\mu$ 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 (4.12). ☐

4.3. **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 [34] for the solution of ($\mathcal{P}_{M^+}$). This method, is an accelerated 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 the procedure to ($\mathcal{P}_{M^+}$) and discuss its connection to Algorithm 1. Similar to our proposed method, PDAP alternates between the update of an active set $\mathcal{A}_k^\mu = \{u_i^k\}_{k=1}^{N_k} \subset \mathrm{Ext}(B)$ and a sparse iterate μ_k which is supported on $\mathcal{A}_k^\mu$. As in Section 3, we provide a short description of the individual steps of this method and we summarize them in Algorithm 2.

Given the current iterate μ_k, 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 a global maximizer $\{\widehat v_k^\mu\}$ of P_k over $\mathrm{Ext}(B)$ to it, i.e., we set

$$\mathcal{A}_k^{\mu,+} = \{u_i^k\}_{i=1}^{N_k} \cup \{\widehat v_k^\mu\}\,, \qquad \widehat v_k^\mu \in \operatorname*{arg\,max}_{v\in \mathrm{Ext}(B)} P_k(v)\,.$$

Using Proposition 4.5 we note that this update step is equivalent to maximizing a linear functional over $\mathrm{Ext}(B)$. This is the content of the next lemma whose proof is immediate using Proposition 4.5.

Lemma 4.7. *Let μ_k be generated by Algorithm 2 and set $u_k = \mathcal{I}(\mu_k)$. Moreover, define $P_k = -\mathcal{K}_*\nabla F(\mathcal{K}\mu_k) \in C(\mathcal{B})$ and $p_k = -K_*\nabla F(Ku_k) \in \mathcal{C}$. Then, there holds*

$$\operatorname*{arg\,max}_{v\in \mathrm{Ext}(B)} P_k(v) = \operatorname*{arg\,max}_{v\in \mathrm{Ext}(B)} \langle p_k, v\rangle\,.$$

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

$$\min_{\mu\in M^+(\mathcal{A}_k^{\mu,+})} \left[F(\mathcal{K}\mu) + \|\mu\|_{M(\mathcal{B})}\right], \tag{4.13}$$

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 (4.13) to the finite dimensional minimization problem (3.4) in Step 5 of Algorithm 1.

Lemma 4.8. *A measure $\widehat\mu \in M^+(\mathcal{A}_k^{\mu,+})$ is a solution to (4.13) 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 $\mu \in M^+(\mathcal{A}_k^{\mu,+})$ be arbitrary. By definition of $M^+(\mathcal{A}_k^{\mu,+})$, there exists at least one $\lambda^\mu \in \mathbb{R}_+^{N_k^+}$, such that $\mu = \sum_{i=1}^{N_k^+} \lambda_i^\mu \delta_{u_i^k}$. Now, noting that $\lambda_i^\mu u_i^k = \mathcal{I}(\lambda_i^\mu \delta_{u_i^k})$, we get

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

from which the characterization of minimizers $\widehat{\mu}$ to (4.13) readily follows. ☐

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}:

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

The method is again summarized in Algorithm 2. As for Algorithm 1 we define the residuals

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

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

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

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

$$P_k = -\mathcal{K}_* \nabla F(\mathcal{K}\mu_k),\; P_k(\widehat{v}_k^\mu) = \max_{v \in \mathcal{B}} P_k(v).$$

if $P_k(\widehat{v}_k^\mu) \le 1$ or $\widehat{v}_k^\mu \in \mathcal{A}_k^\mu$ **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} by solving (4.13).
6. Update

$$\mathcal{A}_{k+1}^\mu = \operatorname{supp} \mu_{k+1}.$$

and set $k = k + 1$.

end for

associated with the iterates μ_k of Algorithm 2 by

$$r_j(\mu_k) := j(\mu_k) - \min_{\mu \in M^+(\mathcal{B})} j(\mu)\,. \tag{4.14}$$

Note that due to Theorem 4.4, 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 Algorithm 1 and Algorithm 2 as stated in the next theorem.

Theorem 4.9. *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, the update steps from 2 to 6 in Algorithm 1 and Algorithm 2 can be realized such that $\widehat{v}_k^u = \widehat{v}_k^\mu$, $u_{k+1} = \mathcal{I}(\mu_{k+1})$ and $\mathcal{A}_{k+1}^u = \mathcal{A}_{k+1}^\mu$.*
In particular, if $\{u_k\}_k$ and $\{\mathcal{A}_k^u\}_k$ are sequences of iterates and active sets generated by Algorithm 1, then there exist sequences $\{\mu_k\}_k$ and $\{\mathcal{A}_k^\mu\}_k$ generated by Algorithm 2 such that

$$u_k = \mathcal{I}(\mu_k)\,, \quad \mathcal{A}_k^u = \mathcal{A}_k^\mu\,, \quad \widehat{v}_k^u = \widehat{v}_k^\mu \tag{4.15}$$

and it holds

$$0 \le r_J(u_k) \le r_j(\mu_k) \quad \text{for all } k \in \mathbb{N}\,, \tag{4.16}$$

where the residuals $r_J(u_k)$ and $r_j(\mu_k)$ are defined in (3.8) and (4.14), respectively.

Proof. According to Lemma 4.7 we can choose $\widehat{v}_k^u = \widehat{v}_k^\mu$ and 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

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

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 (4.13) by Lemma 4.8. Moreover, we note that

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

Finally, the claims in (4.15) follow by an induction argument, while (4.16) is a consequence of Theorem 4.4 and Proposition 4.2 *i*). □

5. Convergence analysis

According to Theorem 4.9, Algorithm 1 converges at least as fast as the PDAP method from Algorithm 2. In the following, we use this observation to prove Theorems 3.4 and 3.8. 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 $\{\mu_k\}_k$. Dropping superscripts, we denote by $\mathcal{A}_k$ and $\widehat{v}_k$ the associated active sets and the new candidate points, respectively.

5.1. Worst-case convergence rate.

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

$$E_{\mu_0} := \left\{ \mu \in M^+(\mathcal{B}) \mid j(\mu) \le j(\mu_0) \right\} . \tag{5.1}$$

By Theorem 4.4, 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

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

Clearly, the additional norm constraint does not change the set of global minimizers. The following proposition relates $\widehat{v}_k \in \operatorname{Ext}(B)$ from Step 2 in Algorithm 2 to a particular conditional gradient descent direction η_k for ($\widehat{P}_{M^+}$).

Proposition 5.1. *Let $\mu_k \in M^+(\mathcal{B})$, $\widehat{v}_k \in \operatorname{Ext}(B)$ and $P_k = -\mathcal{K}_* \nabla F(\mathcal{K}\mu_k) \in C(\mathcal{B})$ be the iterate generated by Algorithm 2 at step k. Set $\eta_k \in M^+(\mathcal{B})$ 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

$$\min_{\substack{\eta \in M^+(\mathcal{B}),\\ \|\eta\|_{M(\mathcal{B})} \le M_0}} [-\langle\!\langle P_k, \eta \rangle\!\rangle + \|\eta\|_{M(\mathcal{B})}] \,. \tag{5.2}$$

Moreover, we have

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

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

$$\begin{aligned}\min_{\substack{\eta\in M^+(\mathcal{B}),\\ \|\eta\|_{M(\mathcal{B})}\le M_0}} [-\langle\!\langle P_k,\eta\rangle\!\rangle + \|\eta\|_{M(\mathcal{B})}] &\ge \min_{\substack{\eta\in M^+(\mathcal{B}),\\ \|\eta\|_{M(\mathcal{B})}\le M_0}} [(1-\max_{v\in\mathcal{B}} P(v))\|\eta\|_{M(\mathcal{B})}] \\ &= \begin{cases} 0 & \text{if } P(\widehat{v}_k) < 1\,, \\ (1-\max_{v\in\mathcal{B}} P(v))M_0 & \text{otherwise,} \end{cases} \\ &= [-\langle\!\langle P_k,\eta_k\rangle\!\rangle + \|\eta_k\|_{M(\mathcal{B})}]\,. \end{aligned}$$

The proof is finished noting that $(1-s)\mu_k + s\eta_k \in M^+(\mathcal{A}_k \cup \{\widehat{v}_k\})$ and that μ_{k+1} is a solution to (4.13). □

In particular, Proposition 5.1 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 5.2 below). Finally, the combination of Theorem 4.9 with Theorem 5.2 yields the convergence of Algorithm 1.

Theorem 5.2. *Let Assumptions (A1)-(A3) hold. Let $\{\mu_k\}_k$ be a sequence generated by Algorithm 2. 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*

$$r_j(\mu_k) \le c\,\frac{1}{k+1} \quad \text{for all } \; k\in\mathbb{N}\,, \tag{5.3}$$

where r_j is defined at (4.14). 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. First note that for every $k\ge 1$ we have $\mu_k \in M^+(\mathcal{A}_k^+)$. Since μ_{k+1} is a solution to (4.13) we conclude

$$j(\mu_{k+1}) \le j(\mu_k) \le j(\mu_0).$$

Consequently, $\{j(\mu_k)\}_k$ is monotone decreasing and $\mu_k \in E_{\mu_0}$. Now, we show that the gradient of the composite mapping $F\circ\mathcal{K}$ is Lipschitz continuous on E_{μ_0}. Since E_{μ_0} is weak* compact and $\mathcal{K}$ is weak*-to-strong continuous, the image set

$$\mathcal{K}E_{\mu_0} = \{\, \mathcal{K}\mu \mid \mu\in E_{\mu_0}\,\}$$

is compact. Thus, using Assumptions (A1) and Proposition 4.3, there exists a constant $L_{\mu_0} > 0$ such that

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

for all $\mu_1,\mu_2 \in E_{\mu_0}$, where C is the constant defined in Proposition 4.3 i). In particular, this estimate proves the Lipschitz continuity of the map $\nabla(F\circ\mathcal{K})$. Thus, the claimed convergence statement follows from [39, Theorem 6.14]. □

Proof of Theorem 3.4. Assume that Algorithm 1 does not converge after finitely many steps and generates a sequence $\{u_k\}_k$. According to Theorem 4.9 there exists a sequence $\{\mu_k\}_k$ generated by Algorithm 2 with $u_k = \mathcal{I}(\mu_k)$. Invoking Theorem 4.9 as well as Theorem 5.2 yields

$$0 \le r_J(u_k) \le r_j(\mu_k) \le c\,\frac{1}{k+1} \quad \text{for all } \; k\in\mathbb{N}\,.$$

In particular, this implies $\lim_{k\to\infty} r_J(u_k) = 0$. Since J is weak* lower semicontinuous and its sublevel sets are bounded, $\{u_k\}_k$ admits at least one weak* accumulation point in $\mathcal{M}$. Each such point $\bar{u}$ satisfies $r_J(\bar{u}) = 0$, i.e., it is a solution to ($\mathcal{P}_\mathcal{M}$). Finally, if the solution $\bar{u}$ to ($\mathcal{P}_\mathcal{M}$) is unique, the convergence $u_k \overset{*}{\rightharpoonup} \bar{u}$ for the whole sequence follows immediately. □

5.2. **Auxiliary results in the non-degenerate case.** In this section, we further investigate the convergence behaviour of the iterates $\{\mu_k\}_k$ but now under the premise of Assumptions (B1)-(B5). In particular, we show that the active sets $\mathcal{A}_k = \{u_i^k\}_{i=1}^{N_k}$ as well as the new candidate points $\widehat{v}_k$ are asymptotically close to the set of optimal extreme points $\bar{\mathcal{A}} = \{\bar{u}_i\}_{i=1}^N$. Moreover, we quantify this closeness in terms of the distance function g from Assumption (B5). Subsequently, in Section 5.3, these results are used in a refined analysis of the energy descent in each iteration of Algorithm 2. We first collect some immediate properties of the solutions and optimal dual variables of ($\mathcal{P}_{M^+}$).

Proposition 5.3. *Let Assumptions (A1)-(A3) and (B1)-(B5) hold. Denote by $\bar{u} = \sum_{i=1}^N \bar{\lambda}_i \bar{u}_i$ the unique solution to ($\mathcal{P}_\mathcal{M}$). Then, the solution to ($\mathcal{P}_{M^+}$) is also unique and given by $\bar{\mu} = \sum_{i=1}^{N_k} \bar{\lambda}_i \delta_{\bar{u}_i}$. Moreover assume that $\{\mu_k\}_k$ is generated by Algorithm 2. Then, there exists $M \in \mathbb{N}$ such that $\mu_k \neq 0$, $\mathcal{A}_k \neq \emptyset$ for all $k \geq M$.*

Proof. Notice first that by Theorem 4.4, the measure $\bar{\mu} = \sum_{i=1}^{N_k} \bar{\lambda}_i \delta_{\bar{u}_i}$ is a solution to ($\mathcal{P}_{M^+}$). We now prove its uniqueness. Assume that $\bar{\mu}, \bar{\eta}$ minimize in ($\mathcal{P}_{M^+}$). Since F is strictly convex by Assumption (A1), we have $\mathcal{K}\bar{\mu} = \mathcal{K}\bar{\eta}$. From Theorem 4.6 one can readily verify that $\operatorname{supp}\bar{\mu}$ and $\operatorname{supp}\bar{\eta}$ are both contained in the set $\{\, v \in \mathcal{B} \mid \bar{P}(v) = 1 \,\}$. Hence, by Assumption (B2) and Proposition 4.5, there exist $\bar{\lambda}_i, \bar{\gamma}_i \geq 0$ such that $\bar{\mu} = \sum_{i=1}^N \bar{\lambda}_i \delta_{\bar{u}_i}$, $\bar{\eta} = \sum_{i=1}^N \bar{\gamma}_i \delta_{\bar{u}_i}$. Therefore,

$$0 = \mathcal{K}(\bar{\mu} - \bar{\eta}) = \sum_{i=1}^N (\bar{\lambda}_i - \bar{\gamma}_i) K \bar{u}_i \,.$$

From (B3) we then conclude $\bar{\lambda}_i = \bar{\gamma}_i$ and so $\bar{\mu} = \bar{\eta}$, implying that ($\mathcal{P}_{M^+}$) has a unique solution. Finally, notice that the prerequisites of Theorem 5.2 are all fulfilled and thus, we have $\mu_k \overset{*}{\rightharpoonup} \bar{\mu}$. Since $\bar{\mu} \neq 0$, we readily conclude that there exists $M \in \mathbb{N}$ such that $\mu_k \neq 0$, $\mathcal{A}_k \neq \emptyset$ for all $k \geq M$. □

Proposition 5.4. *Let $\bar{\mu} = \sum_{i=1}^{N_k} \bar{\lambda}_i \delta_{\bar{u}_i}$ be the unique solution to ($\mathcal{P}_{M^+}$) and set $\bar{P} = -\mathcal{K}_* \nabla F(\mathcal{K}\bar{\mu}) \in C(\mathcal{B})$. Moreover, let $\sigma, \kappa > 0$ and $\{\bar{U}_i\}_{i=1}^N$ denote the constants and $d_\mathcal{B}$-closed neighbourhoods from Assumption (B5). Then, there holds $\bar{P}(\bar{u}_i) = 1$,*

$$\bar{P}(v) \leq 1 - \sigma \quad \text{for all} \;\; v \in \mathcal{B} \setminus \bigcup_{i=1}^N \bar{U}_i \tag{5.4}$$

as well as

$$1 - \bar{P}(u) \geq \kappa g(u, \bar{u}_i)^2 \quad \text{for all} \;\; u \in U_i \tag{5.5}$$

where $U_i = \bar{U}_i \cap \operatorname{Ext}(B)$, $i = 1, \ldots, N$.

Proof. This follows immediately from Assumptions (B2)-(B5) noting that $\bar{P}(v) = \langle \bar{p}, v \rangle$ by Proposition 4.5. □

We further require the following observation on the dual variables P_k.

Proposition 5.5. *Let $\mu_k \in M^+(\mathcal{B})$, $\mathcal{A}_k = \{u_i^k\}_{i=1}^{N_k}$ and $P_k = -\mathcal{K}_* \nabla F(\mathcal{K}\mu_k) \in C(\mathcal{B})$ be generated by Algorithm 2. For all $k \geq 1$ we have*

$$P_k(u_i^k) = 1 \quad \text{for all} \;\; u_i^k \in \mathcal{A}_k \,, \qquad \langle\!\langle P_k, \mu_k \rangle\!\rangle = \|\mu_k\|_{M(\mathcal{B})} \,.$$

In particular, there exists $M \in \mathbb{N}$ such that $\max_{v\in\mathcal{B}} P_k(v) \geq 1$, for all $k \geq M$.

Proof. By construction, we have $\mu_k = \sum_{i=1}^{N_k} \lambda_i^k \delta_{u_i^k}$ where $\lambda^k \in \mathbb{R}^{N_k}$, $\lambda_i^k > 0$, is a solution to

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

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

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

Since $\lambda_i^k > 0$, $P_k(u_i^k) = 1$ has to hold for $i = 1, \dots, N$. Note that

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

Finally, we point out that $\mathcal{A}_k \neq \emptyset$ if k is large, see Proposition 5.3. Hence, there exists $M \in \mathbb{N}$ such that

$$\max_{v \in \mathcal{B}} P_k(v) \geq \max_{v \in \mathcal{A}_k} P_k(v) \geq 1$$

for all $k \geq M$. □

Now, due to the strong convexity of F around $\bar{y}$, see Assumption (B1), the worst-case convergence guarantee of Theorem 5.2 also carries over to the observations and the dual variables.

Proposition 5.6. *Let $y_k = \mathcal{K}\mu_k$, $P_k = -\mathcal{K}_* \nabla F(\mathcal{K}\mu_k)$, $\bar{y} = \mathcal{K}\bar{\mu}$ and $\bar{P} = -\mathcal{K}_* \nabla F(\mathcal{K}\bar{\mu})$. Then, there exist $M \in \mathbb{N}$ and $c > 0$ such that, for all $k \geq M$, there holds*

$$\|y_k - \bar{y}\|_Y + \|\nabla F(y_k) - \nabla F(\bar{y})\|_Y + \|P_k - \bar{P}\|_{C(\mathcal{B})} \leq c\sqrt{r_j(\mu_k)}\,. \tag{5.6}$$

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

Proof. First note again that $\mu_k \in E_{\mu_0}$ by construction and Theorem 5.2. As in the proof of Theorem 5.2, we estimate, using Assumptions (A1)-(A3),

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

where L_{μ_0} is a positive constant depending on the weak* compact set E_{μ_0} according to (5.1) and C is the constant defined in Proposition 4.3. Hence, it suffices to prove the estimate for $\|y_k - \bar{y}\|_Y$. Let $\mathcal{N}(\bar{y}) \subset Y$ and $\theta > 0$ denote the neighbourhood and constant from Assumption (B1). Recall that $\mu_k \overset{*}{\rightharpoonup} \bar{\mu}$ in $M(\mathcal{B})$ along the whole sequence by Theorem 5.2, using the uniqueness result in Proposition 5.3. By weak*-to-strong continuity of $\mathcal{K}$, see Proposition 4.3, we conclude that $y_k \to \bar{y}$ strongly 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 (4.12) in the final inequality. Rearranging the terms in the previous estimate we finally obtain

$$\|y_k - \bar{y}\|_Y \leq \sqrt{\frac{2}{\theta}}\sqrt{r_j(\mu_k)}, \tag{5.7}$$

concluding the proof. □

Combining Propositions 5.5 and 5.6, we deduce that the active set clusters around $\{\bar{u}_i\}_{i=1}^N$, as stated in the following corollary.

Corollary 5.7. *Let $\sigma > 0$ and $\{\bar{U}_i\}_{i=1}^N$ denote the constant and the sets from Assumption (B5). There exists $M \in \mathbb{N}$ such that for all $k \geq M$ we have*

$$P_k(v) \leq 1 - \sigma/2 \quad \text{for all} \quad v \in \mathcal{B} \setminus \bigcup_{i=1}^N \bar{U}_i \,, \tag{5.8}$$

$$\mathcal{A}_k \subset \bigcup_{i=1}^N U_i \,, \tag{5.9}$$

where $U_i = \bar{U}_i \cap \operatorname{Ext}(B)$, $i = 1, \dots, N$.

Proof. According to Proposition 5.6 there exists $M \in \mathbb{N}$ with $\|P_k - \bar{P}\|_{C(\mathcal{B})} \leq \sigma/2$ for all $k \geq M$. As a consequence, for all $v \in \mathcal{B} \setminus \bigcup_{i=1}^N \bar{U}_i$ and k sufficiently large we have

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

using (5.4). This concludes (5.8). Concerning (5.9), note that $P_k(v) = 1$, $v \in \mathcal{A}_k$ by Proposition 5.5. Since by construction $\mathcal{A}_k \subset \operatorname{Ext}(B)$, from (5.8) we conclude (5.9). □

Additionally, each point in $\{\bar{u}_i\}_{i=1}^N$ is approximated by at least one point in $\mathcal{A}_k$.

Proposition 5.8. *Let $\{\bar{U}_i\}_{i=1}^N$ denote the sets from Assumption (B5) and let $\bar{\mu} = \sum_{i=1}^N \bar{\lambda}_i \delta_{\bar{u}_i}$ be the unique solution of ($\mathcal{P}_{M+}$). Then, there holds $\lim_k \mu_k(\bar{U}_i) = \bar{\lambda}_i$ for $i = 1, \dots, N$. In particular, there exists $M \in \mathbb{N}$ such that for all $k \geq M$ there holds*

$$\mathcal{A}_k \cap U_i \neq \emptyset \quad \text{for all} \quad i = 1, \dots, N \,,$$

where $U_i = \bar{U}_i \cap \operatorname{Ext}(B)$, $i = 1, \dots, N$.

Proof. By (B5), the sets $\bar{U}_i$ are pairwise disjoint and $d_{\mathcal{B}}$-closed. Consider an arbitrary but fixed index $l \in \{1, \dots, N\}$. Urysohn's lemma yields the existence of a $d_{\mathcal{B}}$-continuous function $\varphi_l \colon \mathcal{B} \to [0,1]$ such that $\varphi_l(v) = 1$ for all $v \in \bar{U}_l$, and $\varphi_l(v) = 0$ for all $v \in \bar{U}_i$, $i \neq l$. Recall that $\mu_k \overset{*}{\rightharpoonup} \mu$ along the whole sequence by Proposition 5.3. 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 (5.9) in Corollary 5.7. Since $\bar{\lambda}_l > 0$. we conclude $\mu_k(\bar{U}_l) > 0$ for k sufficiently large. Recalling that $\mathcal{A}_k \subset \operatorname{Ext}(B)$ and that $U_l := \bar{U}_l \cap \operatorname{Ext}(B)$, we then infer $\mu_k(U_l) > 0$ for k sufficiently large, concluding $\mathcal{A}_k \cap U_l \neq 0$. □

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 extreme points.

Proposition 5.9. *Let $U_i = \bar{U}_i \cap \operatorname{Ext}(B)$, $i = 1, \dots, N$, where $\{\bar{U}_i\}_{i=1}^N$ are the sets from Assumption (B5). There exist $M \in \mathbb{N}$ and a constant $c > 0$ such that*

$$\sum_{v \in \mathcal{A}_k \cap U_i} \mu_k(\{v\}) g(v, \bar{u}_i) \leq c \sqrt{r_j(\mu_k)} \,, \tag{5.10}$$

for all $k \geq M$ and $i = 1, \dots, N$.

Proof. By the minimality of $\bar{\mu}$ and the convexity of F we obtain

$$\begin{aligned} r_j(\mu_k) &= F(\mathcal{K}\mu_k) - F(\mathcal{K}\bar{\mu}) + \|\mu_k\|_{M(\mathcal{B})} - \|\bar{\mu}\|_{M(\mathcal{B})} \\ &\geq (\nabla F(\mathcal{K}\bar{\mu}), \mathcal{K}\mu_k - \mathcal{K}\bar{\mu})_Y + \|\mu_k\|_{M(\mathcal{B})} - \|\bar{\mu}\|_{M(\mathcal{B})} \\ &= \langle\!\langle \bar{P}, \bar{\mu} - \mu_k \rangle\!\rangle + \|\mu_k\|_{M(\mathcal{B})} - \|\bar{\mu}\|_{M(\mathcal{B})} = \|\mu_k\|_{M(\mathcal{B})} - \langle\!\langle \bar{P}, \mu_k \rangle\!\rangle \,, \end{aligned} \tag{5.11}$$

where we used the optimality condition (4.12) 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 (5.10) for the index i_0. To this end, define the measure $\tilde{\mu}_k \in M^+(\mathcal{B})$ by setting

$$\tilde{\mu}_k := \sum_{i=1}^{N} \mu_k(\bar{U}_i)\delta_{\bar{u}_i} \,.$$

By (5.9) and $\bar{P}(\bar{u}_i) = 1$, see Proposition 5.4, we obtain

$$\langle\!\langle \bar{P}, \tilde{\mu}_k \rangle\!\rangle = \sum_{i=1}^{N} \bar{P}(\bar{u}_i)\mu_k(\bar{U}_i) = \sum_{i=1}^{N} \mu_k(\bar{U}_i) = \|\mu_k\|_{M(\mathcal{B})} \,. \tag{5.12}$$

Again, invoking (5.9), it is immediate to check that

$$\langle\!\langle \bar{P}, \tilde{\mu}_k - \mu_k \rangle\!\rangle = \sum_{i=1}^{N} \sum_{v \in \mathcal{A}_k \cap U_i} \mu_k(\{v\})(1 - \bar{P}(v)) \,. \tag{5.13}$$

Putting together (5.11)-(5.13), we then have

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

where in the second line we used (5.5) from Proposition 5.4. By the sparsity of μ_k and Proposition 5.8, 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

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

Again recall that $\mu_k \in E_{\mu_0}$, and thus, $\mu_k(\bar{U}_{i_0}) \leq M_0$ where M_0 is the constant from the beginning of Section 5.1. The thesis now follows by taking the square root of the above inequality and by setting $c := \sqrt{M_0/\kappa}$. ☐

Next, we derive a similar result for the distance of the point $\widehat{v}_k$ from Step 2 in Algorithm 2 to the closest element in $\{\bar{u}_i\}_{i=1}^N$.

Proposition 5.10. *Let $v_k \in \operatorname{Ext}(B)$ be calculated as in Step 2 in Algorithm 2. There exists $M \in \mathbb{N}$ such that for each $k \geq M$ there is an index $\hat{\imath}_k \in \{1, \ldots, N\}$ satisfying $v_k \in U_{\hat{\imath}_k}$ and*

$$g(\widehat{v}_k, \bar{u}_{\hat{\imath}_k}) \leq c\sqrt{r_j(\mu_k)} \,, \tag{5.14}$$

where c is a positive constant independent on k.

Proof. According to Proposition 5.5, we deduce that $P_k(\widehat{v}_k) = \max_{v\in\mathcal{B}} P_k(v) \geq 1$ for all $k \in \mathbb{N}$ large enough. From (5.8) in Corollary 5.7, we then deduce the existence of indices $\hat{\imath}_k \in \{1,\dots,N\}$ such that $\widehat{v}_k \in U_{\hat{\imath}_k}$ for all $k \in \mathbb{N}$ sufficiently large, as we wanted to show. We are left to prove (5.14). Utilizing the quadratic growth of $\bar{P}$, cf. (5.5), we estimate

$$\kappa\, g(\widehat{v}_k, \bar{u}_{\hat{\imath}_k})^2 \leq \bar{P}(\bar{u}_{\hat{\imath}_k}) - \bar{P}(\widehat{v}_k) \leq \bar{P}(\bar{u}_{\hat{\imath}_k}) - \bar{P}(\widehat{v}_k) + P_k(\widehat{v}_k) - P_k(\bar{u}_{\hat{\imath}_k})\,, \tag{5.15}$$

where we used that $P_k(\widehat{v}_k) = \max_{v\in\mathcal{B}} P_k(v)$ by construction. Using Assumption (B5) as well as Proposition 4.5, the right-hand side in (5.15) is further bounded by

$$\begin{aligned}\bar{P}(\bar{u}_{\hat{\imath}_k}) - \bar{P}(\widehat{v}_k) + P_k(v_k) - P_k(\bar{u}_{\hat{\imath}_k}) &= \langle \bar{p} - p_k, \bar{u}_{\hat{\imath}_k} - \widehat{v}_k\rangle \\ &= (\nabla F(\bar{y}) - \nabla F(y_k), K(\widehat{v}_k - \bar{u}_{\hat{\imath}_k}))_Y \\ &\leq \|\nabla F(\bar{y}) - \nabla F(y_k)\|_Y\, \|K(\widehat{v}_k - \bar{u}_{\hat{\imath}_k})\|_Y \\ &\leq \tau \|\nabla F(\bar{y}) - \nabla F(y_k)\|_Y\, g(\widehat{v}_k, \bar{u}_{\hat{\imath}_k})\,, \end{aligned} \tag{5.16}$$

where $\tau > 0$ from (3.12) does not depend on k. Finally, using (5.15) and (5.16) together with Proposition 5.6, we conclude (5.14). ☐

5.3. **Fast convergence and a proof of Theorem 3.8.** The rest of the section is devoted to the proof of Theorem 3.8 using the results of Section 5.2. For this purpose, let $M \in \mathbb{N}$ be such that Propositions 5.6, 5.8, 5.9, 5.10 and Corollary 5.7 hold for all $k \geq M$. For $k \geq M$ denote by $\hat{\imath}_k \in \{1,\dots,N\}$ the index from Proposition 5.10. Starting from the sequence $\{\mu_k\}_k$ generated by Algorithm 2, we define the following surrogate sequence.

Definition 5.11. Introduce the sequence $\{\widehat{\mu}_k\}_k$ in $M^+(\mathcal{B})$ by setting

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

where $\widehat{v}_k \in \operatorname{Ext}(B)$ is the point computed in Step 2 of Algorithm 2 and $\{\bar{U}_i\}_{i=1}^N$ are the $d_{\mathcal{B}}$-closed neighborhoods of $\{\bar{u}_i\}_{i=1}^N$ in (B5).

We first collect several results which establish the weak* convergence of $\widehat{\mu}_k$ towards $\bar{\mu}$. This will be used to show that $\widehat{\mu}_k \in E_{\mu_0}$ for k large enough.

Lemma 5.12. *For all $k \geq M$ there holds*

$$\|\mu_k\|_{M(\mathcal{B})} = \|\widehat{\mu}_k\|_{M(\mathcal{B})} \quad \textit{and} \quad \langle\!\langle P_k, \widehat{\mu}_k - \mu_k\rangle\!\rangle = \mu_k(\bar{U}_{\hat{\imath}_k})(P_k(\widehat{v}_k) - 1)\,.$$

Proof. Set $k \geq M$. Recall that $\widehat{v}_k \in U_{\hat{\imath}_k}$ by Proposition 5.10. Using (5.9) in Corollary 5.7 and the definition of $\widehat{\mu}_k$, we obtain

$$\|\mu_k\|_{M(\mathcal{B})} = \sum_{i=1}^{N} \mu_k(\bar{U}_i) = \mu_k(\bar{U}_{\hat{\imath}_k}) + \sum_{i=1,\ i\neq\hat{\imath}_k}^{N} \int_{\bar{U}_i} 1 \,\mathrm{d}\mu_k(v) = \langle\!\langle 1, \widehat{\mu}_k\rangle\!\rangle = \|\widehat{\mu}_k\|_{M(\mathcal{B})}\,,$$

as well as

$$\begin{aligned}\langle\!\langle P_k, \widehat{\mu}_k - \mu_k\rangle\!\rangle &= \mu_k(\bar{U}_{\hat{\imath}_k})P_k(\widehat{v}_k) + \sum_{i=1,\ i\neq\hat{\imath}_k}^{N} \int_{\bar{U}_i} P_k(v)\,\mathrm{d}\mu_k(v) - \int_{\mathcal{B}} P_k(v)\,\mathrm{d}\mu_k(v) \\ &= \mu_k(\bar{U}_{\hat{\imath}_k})P_k(\widehat{v}_k) - \int_{\bar{U}_{\hat{\imath}_k}} P_k(v)\,\mathrm{d}\mu_k(v) = \mu_k(\bar{U}_{\hat{\imath}_k})(P_k(\widehat{v}_k) - 1)\,,\end{aligned}$$

where in the last equality we employed Proposition 5.5. ☐

Lemma 5.13. *For all $k \geq M$ there holds*

$$\|\mathcal{K}(\widehat{\mu}_k - \mu_k)\|_Y \leq c\sqrt{r_j(\mu_k)}\,, \tag{5.17}$$

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

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

In particular, $\widehat{\mu}_k \in E_{\mu_0}$ for all $k \in \mathbb{N}$ large enough.

Proof. By (5.9) in Corollary 5.7 and definition of $\widehat{\mu}_k$ we compute

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

where we used that μ_k is a conic combination of Dirac deltas. Hence, by linearity, point *ii)* in Proposition 4.3 and triangle inequality we obtain

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

Recalling that $\widehat{v}_k \in U_{\hat{\imath}_k}$, by (B5) and Propositions 5.9, 5.10 we estimate

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

where in the last line we also use $\mu_k \in E_{\mu_0}$. This establishes (5.17). As for the remaining part of the statement, first recall that $r_j(\mu_k) \to 0$ and $\mu_k \stackrel{*}{\rightharpoonup} \bar{\mu}$ along the whole sequence by Theorem 5.2 and the uniqueness of the minimizer, see Proposition 5.3. Since $\mathcal{K}$ is weak*-to-strong continuous (see Proposition 4.3), 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 (5.17). 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 Lemma 5.12, we obtain

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

The first term on the right hand side converges to zero by Theorem 5.2. Concerning the second term, we have shown that $\mathcal{K}\widehat{\mu}_k, \mathcal{K}\mu_k \to \mathcal{K}\bar{\mu}$, so that, recalling that F is continuous, we conclude $r_j(\widehat{\mu}_k) \to 0$. The latter implies 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 4.4), and since $\bar{\mu}$ is the unique solution to $(\mathcal{P}_{M+})$, we also conclude that $\widehat{\mu}_k \stackrel{*}{\rightharpoonup} \bar{\mu}$. Finally, since μ_0 is not a minimizer of j, we also have that $\widehat{\mu}_k \in E_{\mu_0}$ for k sufficiently large. □

Finally, we are in position to prove the linear convergence of Algorithm 2 and thus, in view of Theorem 4.9, also for Algorithm 1.

Theorem 5.14. *Let Assumptions (A1)-(A3) and (B1)-(B5) hold. Let $\{\mu_k\}_k$ be generated by Algorithm 2. Then, we have $u_k \stackrel{*}{\rightharpoonup} \bar{u}$ in $\mathcal{M}$. Moreover, there is $\bar{k} \in \mathbb{N}$ and $\zeta \in [3/4, 1)$ such that*

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

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

$$r_j(\mu_k) \leq c\,\zeta^k$$

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

Proof. For a fixed $s \in [0,1]$ define

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

Note that by construction $\mu_k^s \in M^+(\mathcal{A}_k \cup \{\widehat{v}_k\})$ as well as

$$\|\mu_k^s\|_{M(\mathcal{B})} = s\mu_k(\bar{U}_{\hat{\imath}_k}) \left\|\delta_{\widehat{v}_k}\right\|_{M(\mathcal{B})} + (1-s)\mu_k(\bar{U}_{\hat{\imath}_k}) + \sum_{i=1,\ i\neq \hat{\imath}_k}^{N} \mu_k(\bar{U}_i) = \|\mu_k\|_{M(\mathcal{B})}$$

due to $\widehat{v}_k \notin \mathcal{A}_k$. Since μ_{k+1} is optimal in (4.13) by definition of Algorithm 2, we infer

$$j(\mu_{k+1}) \leq j(\mu_k^s) \quad \text{ for all } \ s \in [0,1]\,. \tag{5.18}$$

Next, we estimate the residual of μ_k^s. Since F is regular, one can readily check the identity

$$\begin{aligned} j(\mu_k^s) - j(\mu_k) &= s(\nabla F(\mathcal{K}\mu_k), \mathcal{K}\widehat{\mu}_k - \mathcal{K}\mu_k)_Y + \|\mu_k^s\|_{M(\mathcal{B})} - \|\mu_k\|_{M(\mathcal{B})} + R(\mu_k) \\ &= -s\langle\!\langle P_k, \widehat{\mu}_k - \mu_k \rangle\!\rangle + R(\mu_k)\,, \end{aligned} \tag{5.19}$$

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

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

In order to estimate the remainder, first recall that $\{\mu_k\}_k$ is contained in E_{μ_0}, see Theorem 5.2, where E_{μ_0} is the weak* compact set defined in (5.1). Since $\bar{\mu} \in E_{\mu_0}$, $j(\widehat{\mu}_k) \to j(\bar{\mu})$ (by Lemma 5.13), and μ_0 is not a minimizer of j, we also have that $\widehat{\mu}_k \in E_{\mu_0}$ for k sufficiently large. As j is convex, we infer $\mu_k^s \in E_{\mu_0}$ 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 (Proposition 4.3). Denote by L_{μ_0} the Lipschitz constant of ∇F on the set $\mathcal{K}E_{\mu_0}$, which exists by Assumptions (A1)-(A3). We can then estimate

$$|R(\mu_k)| \leq L_{\mu_0} \|\mathcal{K}\widehat{\mu}_k - \mathcal{K}\mu_k\|_Y \int_0^s \|\mathcal{K}\mu_k^t - \mathcal{K}\mu_k\|_Y \,\mathrm{d}t = s^2 \,\frac{L_{\mu_0}}{2}\, \|\mathcal{K}(\widehat{\mu}_k - \mu_k)\|_Y^2\,. \tag{5.20}$$

By minimality of $\bar{\mu}$ and (5.19)-(5.20), we infer

$$\begin{aligned} r_j(\mu_k^s) &= j(\mu_k^s) - j(\bar{\mu}) \\ &= j(\mu_k) - j(\bar{\mu}) - s\langle\!\langle P_k, \widehat{\mu}_k - \mu_k \rangle\!\rangle + R(\mu_k) \\ &\leq r_j(\mu_k) - s\langle\!\langle P_k, \widehat{\mu}_k - \mu_k \rangle\!\rangle + s^2\,\frac{L_{\mu_0}}{2}\|\mathcal{K}(\widehat{\mu}_k - \mu_k)\|_Y^2. \end{aligned} \tag{5.21}$$

Now, using the convexity of F together with Proposition 5.1 and Proposition 5.5 we get

$$r_j(\mu_k) \leq \langle\!\langle P_k, \bar{\mu} - \mu_k \rangle\!\rangle + \|\mu_k\|_{M(\mathcal{B})} - \|\bar{\mu}\|_{M(\mathcal{B})} \leq \langle\!\langle P_k, M_0\delta_{\widehat{v}_k} - \mu_k \rangle\!\rangle + \|\mu_k\|_{M(\mathcal{B})} - \|M_0\delta_{\widehat{v}_k}\|_{M(\mathcal{B})}\,,$$

where we remind that $\|\bar{\mu}\|_{M(\mathcal{B})} \leq M_0$ and M_0 is a fixed upper bound on the norm of elements in E_{μ_0}. Invoking Proposition 5.5 yields

$$\langle\!\langle P_k, M_0\delta_{\widehat{v}_k} - \mu_k \rangle\!\rangle + \|\mu_k\|_{M(\mathcal{B})} - \|M_0\delta_{\widehat{v}_k}\|_{M(\mathcal{B})} = M_0(P_k(\widehat{v}_k) - 1)$$

and thus, using Lemma 5.12, we get

$$\langle\!\langle P_k, \widehat{\mu}_k - \mu_k \rangle\!\rangle = \mu_k(\bar{U}_{\hat{\imath}_k})(P_k(\widehat{v}_k) - 1) \geq \frac{\mu_k(\bar{U}_{\hat{\imath}_k})}{M_0}\, r_j(\mu_k)\,. \tag{5.22}$$

Using (5.22) and (5.17) (see Lemma 5.13) in (5.21) yields

$$r_j(\mu_k^s) \leq r_j(\mu_k) - s\,\frac{\mu_k(\bar{U}_{\hat{\imath}_k})}{M_0}\, r_j(\mu_k) + s^2\,\frac{L_{\mu_0} c_2}{2}\, r_j(\mu_k)\,, \tag{5.23}$$

where $c_2 > 0$ is the square of the constant in (5.17). Define the constant

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

Notice that $c_1 > 0$, since (B4) holds. Moreover, $c_1 \leq 1/2$ given that $\bar{\mu} \in E_{\mu_0}$. Invoking Proposition 5.8, we conclude that

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

which, together with (5.23) 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 $s \in [0,1]$. By subtracting $j(\bar{\mu})$ from both sides of (5.18) we then obtain

$$r_j(\mu_{k+1}) \leq \varphi(s) r_j(\mu_k)\,, \qquad \varphi(s) := 1 - sc_1 + s^2\,\frac{L_{\mu_0}c_2}{2}\,, \tag{5.24}$$

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 (5.24) we obtain an integer $\bar{k} \in \mathbb{N}$ such that

$$r_j(\mu_{k+1}) \leq \zeta\, r_j(\mu_k) \ \text{ for all } \ k \geq \bar{k}\,.$$

We thus infer $r_j(\mu_k) \leq r_j(\mu_{\bar{k}})\zeta^{k-\bar{k}}$ for all $k \geq \bar{k}$, and the final thesis follows by setting $c := r_j(\mu_{\bar{k}})\zeta^{-\bar{k}} > 0$. ☐

6. Conclusions

We have introduced an accelerated conditional gradient method (AGCG) to solve a class of non-smooth minimization problems in Banach spaces. As a guiding example, the inverse problem of identifying the initial source of a heat equation from distributed temperature measurements is considered. Compared to standard conditional gradient methods (GCG), our algorithm shows a vastly improved rate of convergence. The observed improved convergence behaviour is backed up by our theoretical analysis. More in detail, under non-degeneracy assumptions on the dual variable, we derive the linear convergence rate of our method in an abstract problem setting. We remark that, aside the sparsity example considered in Section 3.3, our algorithm can be applied to a wide range of other relevant minimization problems; however, in this paper we decided to focus on the theoretical aspects of the algorithm, leaving a detailed analysis of different examples and the related numerics to a follow-up paper. Moreover, as in the euclidean case of [34], we are positive that all results in the present paper are transferable to nonsmooth 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 ommited in the present paper for the sake of readability.

Acknowledgements

KB and SF gratefully acknowledge support by the Christian Doppler Research Association (CDG) and Austrian Science Fund (FWF) through the Partnership in Research project PIR-27 "Mathematical methods for motion-aware medical imaging" and project P 29192 "Regularization graphs for variational imaging". MC is supported by the Royal Society (Newton International Fellowship NIF\R1\192048 Minimal partitions as a robustness boost for neural network classifiers). The Institute of Mathematics and Scientific Computing, to which KB and SF are affiliated, is a member of NAWI Graz (`https://www.nawigraz.at/en/`). The authors KB and SF are further members of/associated with BioTechMed Graz (`https://biotechmedgraz.at/en/`).

Appendix A.

A.1. **Proofs of Section 4.** In this section we exhibit proofs for Propositions 4.2, 4.3. In order to define the map $\mathcal{I}$ at Proposition 4.2 we employ the classical Choquet's Theorem [33, Page 14], which we recall for reader's convenience.

Theorem A.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 4.2.

Proof of Proposition 4.2. 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

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

which 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 (4.3) 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

$$\mathcal{I}(\lambda_1\mu_1 + \lambda_2\mu_2) = \lambda_1\mathcal{I}(\mu_1) + \lambda_2\mathcal{I}(\mu_2) \quad \text{for all } \lambda_1, \lambda_2 \in \mathbb{R}_+,\ \mu_1, \mu_2 \in M^+(\mathcal{B})\,.$$

We now show i). Notice that $\mathcal{I}(0) = 0$ and the estimate holds trivially. If $\mu \neq 0$, 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 $\overline{\mathrm{conv}(\mathcal{B})}^* \subseteq B$. 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 $u \in \mathrm{dom}\,\mathcal{G}$ 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 A.1 with $X = (\mathcal{M}, \mathit{weak}^*)$, $K = B$, $v_0 = u/\mathcal{G}(u)$ to obtain a probability measure $\tilde{\mu}$ over X concentrated on $\mathrm{Ext}(B)$ and such that

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

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

Finally, we prove Proposition 4.3.

Proof of Proposition 4.3. For $\mu \in M(\mathcal{B})$ define the functional $T_\mu \colon Y \to \mathbb{R}$ by setting

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

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

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

where we recalled that the constant C is defined at i) and is finite, since $\mathcal{B}$ is norm bounded. Therefore, by Riesz's Theorem, there exists a unique element in Y representing T_μ, thus defining a map $\mathcal{K} \colon M(\mathcal{B}) \to Y$. In particular, $\mathcal{K}$ is linear and satisfies (4.5). Moreover, by (4.5) we have

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

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 (4.3) and (4.5) 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 (4.5),

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

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 (4.5) twice we see that

$$\|\mathcal{K}\mu_k\|_Y^2 = \int_{\mathcal{B}} (Kv, \mathcal{K}\mu)_Y \ \mathrm{d}\mu_k(v) = \int_{\mathcal{B}} \int_{\mathcal{B}} (Kv, Kw)_Y \ \mathrm{d}\mu_k(w) \mathrm{d}\mu_k(v)\,. \tag{A.2}$$

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 (A.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 (4.5). This concludes the proof of iv).
Assume now that Y is separable and fix $\mu \in M(\mathcal{B})$. Notice that $f \colon \mathcal{B} \to Y$ 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 ([21, 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 (4.5) and Hille's Theorem ([21, 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 (4.6) and the proof. □

References

[1] C. D. Aliprantis and K. Border. *Infinite Dimensional Analysis*. Springer-Verlag Berlin Heidelberg, 2006.
[2] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford Science Publications, 2000.
[3] F. F. Bonsall. A general atomic decomposition theorem and Banach's closed range theorem. *The Quarterly Journal of Mathematics*, 42(1):9–14, 1991.
[4] N. Boyd, S. Geoffrey, and R. Benjamin. The Alternating Descent Conditional Gradient Method for Sparse Inverse Problems. *SIAM Journal on Optimization*, 27(2):616–639, 2017.
[5] C. Boyer, A. Chambolle, Y. De Castro, V. Duval, F. de Gournay, and P. Weiss. On Representer Theorems and Convex Regularization. *SIAM Journal on Optimization*, 29(2):1260–1281, 2019.
[6] K. Bredies and M. Carioni. Sparsity of solutions for variational inverse problems with finite-dimensional data. *Calculus of Variations and Partial Differential Equations*, 59(1):14, 2020.
[7] K. Bredies, M. Carioni, and S. Fanzon. A superposition principle for the inhomogeneous continuity equation with Hellinger-Kantorovich-regular coefficients. *arXiv preprint arXiv:2007.06964*, 2020.
[8] K. Bredies, M. Carioni, S. Fanzon, and F. Romero. A generalized conditional gradient method for dynamic inverse problems with optimal-transport regularization. *arXiv preprint arXiv:2012.11706*, 2020.
[9] K. Bredies, M. Carioni, S. Fanzon, and F. Romero. On the extremal points of the ball of the Benamou–Brenier energy. *Bulletin of the London Mathematical Society*, 2021. doi: https://doi.org/10.1112/blms.12509.
[10] K. Bredies, M. Carioni, S. Fanzon, and D. Walter. Examples and numerical analysis of accelerated generalized conditional gradient methods. In preparation.
[11] K. Bredies and S. Fanzon. An optimal transport approach for solving dynamic inverse problems in spaces of measures. *ESAIM: Mathematical Modelling and Numerical Analysis*, 54(6):2351–2382, 2020.
[12] K. Bredies and D. A. Lorenz. Iterated hard shrinkage for minimization problems with sparsity constraints. *SIAM J. Sci. Comput.*, 30(2):657–683, 2008.
[13] K. Bredies, D. A. Lorenz, and P. Maass. A generalized conditional gradient method and its connection to an iterative shrinkage method. *Computational Optimization and Applications*, 42(2):173–193, 2009.
[14] K. Bredies and H.K. Pikkarainen. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(1):190–218, 2013.
[15] M. D. Canon and C. D. Cullum. A Tight Upper Bound on the Rate of Convergence of Frank-Wolfe Algorithm. *SIAM Journal on Control*, 6(4):509–516, 1968.
[16] E. Casas, B. Vexler, and E. Zuazua. Sparse initial data identification for parabolic PDE and its finite element approximations. *Mathematical Control & Related Fields*, 5(3):377–399, 2015.
[17] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The Convex Geometry of Linear Inverse Problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.
[18] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.
[19] V. F. Demianov and A. M. Rubinov. *Approximate methods in optimization problems*. Number 32. Elsevier Publishing Company, 1970.
[20] Q. Denoyelle, V. Duval, G. Peyré, and E. Soubies. The sliding Frank–Wolfe algorithm and its application to super-resolution microscopy. *Inverse Problems*, 36(1):014001, 2019.
[21] J. Diestel and J.J. Uhl. *Vector Measures*. Mathematical surveys and monographs. American Mathematical Society, 1977.
[22] J. C. Dunn. Rates of Convergence for Conditional Gradient Algorithms Near Singular and Nonsingular Extremals. *SIAM Journal on Control and Optimization*, 17(2):187–211, 1979.

[23] J. C. Dunn. Convergence Rates for Conditional Gradient Sequences Generated by Implicit Step Length Rules. *SIAM Journal on Control and Optimization*, 18(5):473–487, 1980.
[24] J. C. Dunn and S. Harshbarger. Conditional gradient algorithms with open loop step size rules. *Journal of Mathematical Analysis and Applications*, 62(2):432–444, 1978.
[25] A. Flinth, F. de Gournay, and P. Weiss. On the linear convergence rates of exchange and continuous methods for total variation minimization. *Mathematical Programming*, 190(1):221–257, 2021.
[26] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.
[27] R. M. Freund and P. Grigas. New analysis and results for the Frank–Wolfe method. *Mathematical Programming*, 155(1-2):199–230, 2016.
[28] J. GuéLat and P. Marcotte. Some comments on Wolfe's 'away step'. *Mathematical Programming*, 35(1):110–119, 1986.
[29] C. A. Holloway. An extension of the Frank and Wolfe method of feasible directions. *Mathematical Programming*, 6(1):14–27, 1974.
[30] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435. PMLR, 2013.
[31] S. Lacoste-Julien and M. Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. *Proceedings of the Annual Conference on Neural Information Processing Systems 2015*, pages 496–504, 2015.
[32] D. Leykekhman, B. Vexler, and D.l Walter. Numerical analysis of sparse initial data identification for parabolic problems. *ESAIM, Math. Model. Numer. Anal.*, 54(4):1139–1180, 2020.
[33] R. R. Phelps. *Lectures on Choquet's Theorem*, volume LN1757. Berlin: Springer, 2001.
[34] K. Pieper and D. Walter. Linear convergence of accelerated conditional gradient algorithms in spaces of measures. *ESAIM: COCV*, 27, 2021. doi: https://doi.org/10.1051/cocv/2021042.
[35] S. Shalev-Shwartz, N. Srebro, and T. Zhang. Trading Accuracy for Sparsity in Optimization Problems with Sparsity Constraints. *SIAM Journal on Optimization*, 20(6):2807–2832, 2010.
[36] Yifan Sun and Francis Bach. Safe screening for the generalized conditional gradient method. *arXiv preprint arXiv:2002.09718*, 2020.
[37] J. A. Ventura and D. W. Hearn. Restricted simplicial decomposition for convex constrained problems. *Mathematical Programming*, 59(1-3):71–85, 1993.
[38] B. Von Hohenbalken. Simplicial decomposition in nonlinear programming algorithms. *Mathematical Programming*, 13(1):49–68, 1977.
[39] D. Walter. *On sparse sensor placement for parameter identification problems with partial differential equations*. PhD thesis, Technische Universität München, 2019. https://nbn-resolving.org/urn/resolver.pl?urn:nbn:de:bvb:91-diss-20190604-1456274-1-8%20.
[40] Chu Wang, Yingfei Wang, Robert Schapire, et al. Functional frank-wolfe boosting for general loss functions. *arXiv preprint arXiv:1510.02558*, 2015.
[41] Y. Yu, X. Zhang, and D. Schuurmans. Generalized conditional gradient for sparse estimation. *Journal of Machine Learning Research*, 18(144):1–46, 2017.

(Kristian Bredies) University of Graz, Institute of Mathematics and Scientific Computing, Heinrichstrasse 36, 8010 Graz, Austria
Email address: Kristian.Bredies@uni-graz.at

(Marcello Carioni) University of Cambridge, Department of Applied Mathematics and Theoretical Physics, Wilberforce Road, Cambridge CB3 0WA, UK
Email address: mc2250@maths.cam.ac.uk

(Silvio Fanzon) University of Graz, Institute of Mathematics and Scientific Computing, Heinrichstrasse 36, 8010 Graz, Austria
Email address: Silvio.Fanzon@uni-graz.at

(Daniel Walter) Johann Radon Institute for Compuational and Applied Mathematics, Altenberger Strasse 69, 4040 Linz
Email address: daniel.walter@oeaw.ac.at