HAL open science

A nonsmooth Frank–Wolfe algorithm through a dual cutting-plane approach

Guilherme Mazanti, Thibault Moquet, Laurent Pfeiffer

▶ **To cite this version:**

Guilherme Mazanti, Thibault Moquet, Laurent Pfeiffer. A nonsmooth Frank–Wolfe algorithm through a dual cutting-plane approach. 2024. hal-04543088

HAL Id: hal-04543088

https://hal.science/hal-04543088v1

Preprint submitted on 11 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A nonsmooth Frank–Wolfe algorithm through a dual cutting-plane approach

Guilherme Mazanti*† Thibault Moquet*† Laurent Pfeiffer*†

March 30, 2024

Abstract

An extension of the Frank–Wolfe Algorithm (FWA), also known as Conditional Gradient algorithm, is proposed. In its standard form, the FWA allows to solve constrained optimization problems involving β-smooth cost functions, calling at each iteration a Linear Minimization Oracle. More specifically, the oracle solves a problem obtained by linearization of the original cost function. The algorithm designed and investigated in this article, named Dualized Level-Set (DLS) algorithm, extends the FWA and allows to address a class of nonsmooth costs, involving in particular support functions. The key idea behind the construction of the DLS method is a general interpretation of the FWA as a cutting-plane algorithm, from the dual point of view. The DLS algorithm essentially results from a dualization of a specific cutting-plane algorithm, based on projections on some level sets. The DLS algorithm generates a sequence of primal-dual candidates, and we prove that the corresponding primal-dual gap converges with a rate of $O(1/\sqrt{t})$.

Keywords: Frank–Wolfe algorithm, Conditional Gradient algorithm, cutting-plane algorithms, simplicial algorithms, duality in convex analysis, nonsmooth optimization.

Mathematics Subject Classification (2020): 90C25 · 90C30 · 90C46.

1 Introduction

The Frank–Wolfe Algorithm (FWA), also known as Conditional Gradient Algorithm, is an iterative minimization algorithm which was first introduced in [9]. It aims at solving numerically problems of the form

$$\underset{x\in K}{\text{minimize}}\ f(x), \tag{1}$$

where f is a convex function with Lipschitz-continuous gradient and K is a closed convex bounded set of some Banach space $\mathcal{X}$. This method relies on a Linear Minimization Oracle (LMO) of the form

$$\underset{x\in K}{\text{minimize}}\ \langle \mu, x\rangle, \tag{LMO$_\mu$}$$

for some well-chosen $\mu \in \mathcal{X}^*$. The result of this oracle is used to update the candidate to optimality through convex combinations.

One simple choice consists in taking, at iteration t,

$$\mu^t = \nabla f\big(x^t\big), \quad \gamma_t = 2/(t+2), \quad \text{and} \quad x^{t+1} = \gamma_t v^t + (1-\gamma_t)x^t,$$

where v^t is a solution to (LMO$_\mu$) with $\mu = \mu^t$. We will refer to that method as *agnostic FWA*. It is well-known, (see for instance [11]) that the agnostic FWA converges (in value) to a minimizer of f over K with a speed of order $1/t$.

There are many other possible choices for iteration updates. We mention the FWA with line-search and the fully-corrective FWA, which consist in taking

$$x^{t+1} \in \underset{x\in \tilde{K}^t}{\text{argmin}}\ f(x), \tag{2}$$

*Université Paris-Saclay, CNRS, CentraleSupélec, Inria, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France.

†Fédération de Mathématiques de CentraleSupélec, 91190, Gif-sur-Yvette, France.

with $\tilde{K}^t$ respectively defined by $\operatorname{conv}\{x^t, v^t\}$ and $\operatorname{conv}\{v^0, \dots, v^t\}$. In other words, we replace in Problem (1) the feasible set K by an inner polyhedric approximation, the set $\tilde{K}^t$. These variants enjoy in general the same sublinear convergence properties; yet improved rates of convergence can be obtained in many situations, see [21].

A great number of applications of the FWA can be found in [11, 19] and references therein. We mention, among others, machine learning (see [11, 15]), optimal transport (see [8]), image processing (see [12]), and potential mean-field games (see [16, 18]).

The article is dedicated to the design of an extension of the FWA which can handle nonsmooth cost functions, and which we call Dualized Level-Set (DLS) method. It allows to solve problems of the form

$$\underset{x \in \mathcal{E}_1}{\text{minimize}}\ f(x) + \sigma_Q(Ax - b) + \iota_K(x),$$

where $\mathcal{E}_1$ and $\mathcal{E}_2$ are Hilbert spaces, $A\colon \mathcal{E}_1 \to \mathcal{E}_2$ is a bounded linear operator, $b \in \mathcal{E}_2$, f is convex and has Lipschitz-continuous gradient, $K \subset \mathcal{E}_1$ is a set for which we have an LMO, and $Q \subset \mathcal{E}_2$ is of the form $Q = Q_1 + Q_2$, where Q_1 is a closed convex bounded set and Q_2 is a closed convex cone. The term $\sigma_Q(Ax - b)$ is in general non-differentiable with respect to x. It typically describes equality constraints of the form $Ax = b$ (if $Q = \mathcal{E}_2$) or finitely many inequality constraints of the form $Ax \leq b$ (if Q is the closed positive orthant of $\mathcal{E}_2 = \mathbb{R}^m$).

There already exist some extensions of the FWA to a nonsmooth framework. We mention the Frank-Wolfe Augmented Lagrangian (FW-AL) method from [10], the Conditional-Gradient-based Augmented Lagrangian (CGAL) algorithm from [22] and the Conditional Gradient with Augmented Lagrangian and Proximal-step (CGALP) from [20]. All rely on a Moreau regularization of the nonsmooth term (σ_Q in our framework), which amounts to the augmented Lagrangian method when $\sigma_Q(Ax-b)$ models inequality or equality constraints. It is proved in [20, Theorems 4.1 and 4.2] that the sequence of candidates generated by the CGALP method is asymptotically feasible and that the associated Lagrangian values converge at a rate of $o(1/t^b)$, where $b \in [0, 1/3)$ is a parameter of the algorithm (see also the discussion provided in [20, Examples 3.4 and 4.4]).

We follow a different approach, based on an interpretation of the fully-corrective FWA as a cutting-plane algorithm, from the point of view of the dual problem. This interpretation is not new and can be found in [1, Section 7.7] and in [24]. In [24], the authors use a specific implementation of the Frank–Wolfe algorithm to construct a new variant of cutting-plane algorithm that enjoys a linear convergence. As we will explain below, we follow the reverse path. Let us mention that [2] gives a dual interpretation of the agnostic FWA as a mirror descent algorithm. We will explain more in detail the dual interpretation of the fully-corrective FWA in Section 3, we only give a rough description of it in this introduction. Observe that the dual of Problem (1) is given by

$$\underset{\mu \in \mathcal{X}^*}{\text{minimize}}\ f^*(\mu) + \sigma_K(-\mu).$$

The dual of Problem (2) is the same problem, with K replaced by $\tilde{K}^t$. Since $\tilde{K}^t \subset K$, $\sigma_{\tilde{K}^t}$ is a lower approximation of σ_K; moreover, since $\tilde{K}^t$ is polyhedral, $\sigma_{\tilde{K}^t}$ is piecewise affine. So the FWA algorithm (with line search or the fully corrective variant) amounts to solving at each iteration a simplified version of the dual problem obtained through a piecewise-affine lower approximation of the dual cost: this is the basic principle of every cutting-plane-type method. To be rigorous, let us note that cutting-plane methods usually approximate the whole cost function while here, only the second term is approximated. Such methods are referred to as simplicial methods in [1, 24]. We find it convenient to keep the terminology cutting-plane in this article.

In a nutshell, our general strategy for the design of the desired extension of the FWA consists in "dualizing" a cutting-plane-type algorithm. This strategy brings two main difficulties. First, we need convergence guaranties for the dualized algorithm (and not just for the chosen cutting-plane-type algorithm). Second, the dualized algorithm must be implementable. Our attention has focused on the method introduced in [17, Section 2.2.1], which we will refer to as the Level-Set method. While the simplest cutting-plane algorithms simply consists in minimizing a piecewise affine approximation of the cost function, the Level-Set method updates the current candidate to optimality by projecting it onto a level set of the piecewise affine approximation. The addition of this projection step attenuates the instabilities from which the basic cutting-plane methods suffer. Moreover, it enables the authors of [17] to perform a quantitative convergence analysis. They indeed prove that the Level-Set method exhibits a rate of convergence of $O(1/\sqrt{t})$. This rate of convergence is actually established for some quantity denoted $\Delta(t)$, which we interpret as a primal-dual gap. The fact that not only the optimality gap, but

also the primal-dual gap, converges (with a certain rate) is a crucial aspect, since it allows to show the convergence of FWA-type algorithm obtained through dualization. The convergence of the primal-dual gap in the Level-Set method is our main interest for it.

Our DLS method is not a direct dualization of the Level-Set method, but rather a dualization of an extension—in two directions—of the Level-Set method. In the original formulation of the Level-Set method, the full cost function is approximated with a piecewise affine cost. In our algorithm, the term f^* of the dual problem remains unchanged, which requires us to proceed to an extension of the convergence proof of the Level-Set method in which only some part of the cost function is approximated (through a piecewise affine function). The second extension of the Level-Set method that we need to perform concerns the projection step. In the original method, the projection step is done with respect to the Euclidean norm. The dualization of this step would require the knowledge of f^*, which we consider as too demanding. We propose to change the Euclidean norm by a specific Bregman distance, which ultimately yields a more tractable projection step. The DLS algorithm enjoys the same convergence rate as the original Level-Set method, in $O(1/\sqrt{t})$. Let us stress however that this convergence rate does not account for the possible increase of complexity of each iteration.

This article is organized as follows. In Section 2 we present our notations and some preliminary results. In Section 3 we give an insight on the primal and dual interpretation of the Frank–Wolfe Algorithm, and we then present our Extended Level-Set (ELS) method and its theoretical guarantees. In Section 4 we derive our DLS method, obtained by application of the ELS method to the dual problem, and we prove its convergence. The proofs of the technical results are postponed to Section 5. Finally, Section 6 is dedicated to numerical examples.

2 Preliminaries and notations

General notations In the following, $\bar{\mathbb{R}}$ denotes the ordered set $\mathbb{R} \cup \{+\infty, -\infty\}$ and $\bar{\mathbb{R}}^+$ the ordered set $\mathbb{R} \cup \{+\infty\}$. Let $\mathcal{E}$ be a Hilbert space. Its inner product is denoted as $\langle \cdot\,, \cdot \rangle_{\mathcal{E}}$ and the deriving norm $\|\cdot\|_{\mathcal{E}}$, defined for all $x \in \mathcal{E}$ as $\|x\|_{\mathcal{E}} = \sqrt{\langle x, x \rangle_{\mathcal{E}}}$. When we deem that no confusion is possible, we will drop the subscript for the inner product and the norm. Let $\mathcal{X}$ be a Banach space endowed with the norm $\|\cdot\|_{\mathcal{X}}$. We denote as $\mathcal{X}^*$ its topological dual, i.e., the space of continuous linear forms over $\mathcal{X}$, as $\|\cdot\|_{\mathcal{X}^*}$ the associated dual norm, and as $\langle \cdot\,, \cdot \rangle_{\mathcal{X}^*,\mathcal{X}}$ the natural pairing. Likewise, we will drop the subscripts when we deem that no confusion is possible. We recall that $\mathcal{X}^*$ endowed with $\|\cdot\|_{\mathcal{X}^*}$ is a Banach space and that a Hilbert space $\mathcal{E}$ is a Banach space. Also, we always identify $\mathcal{E}^*$ with $\mathcal{E}$. In this section, we make the convention that spaces denoted by the letter $\mathcal{E}$ (possibly with subscripts) are always assumed to be Hilbert spaces, while spaces denoted by the letter $\mathcal{X}$ (possibly with subscripts) are always assumed to be Banach spaces.

Unless specified otherwise when required, the definitions below are taken similarly over both $\mathcal{X}$ and $\mathcal{X}^*$. Let $f \colon \mathcal{X} \to \bar{\mathbb{R}}$. We denote as $\operatorname{epi}(f)$ the epigraph of f, defined as

$$\operatorname{epi}(f) = \left\{ (x, \lambda) \in \mathcal{X} \times \mathbb{R} \,\middle|\, \lambda \geq f(x) \right\}.$$

We recall that f is convex (respectively (weakly) lower semicontinuous) *iff* $\operatorname{epi}(f)$ is convex (respectively (weakly) closed). We denote as $\Gamma(\mathcal{X})$ the set of convex lower semicontinuous functions from $\mathcal{X}$ to $\bar{\mathbb{R}}$ and as $\Gamma_0(\mathcal{X})$ the subset of those which are proper, i.e., which never take the value $-\infty$ and are not constant equal to $+\infty$.

In what follows, we assume $f \in \Gamma_0(\mathcal{X})$. We denote as $\operatorname{dom}(f)$ the domain of f, which is the set

$$\operatorname{dom}(f) = \{f \in \mathbb{R}\} = \left\{ x \in \mathbb{R} \,\middle|\, f(x) \in \mathbb{R} \right\}.$$

For any $x \in \mathcal{X}$, we denote as $\partial f(x)$ the set of its subgradients at x, i.e., the set

$$\partial f(x) = \left\{ \mu \in \mathcal{X}^* \,\middle|\, \forall y \in \mathcal{X},\ f(y) \geq f(x) + \langle \mu, y - x \rangle \right\},$$

and as $\operatorname{dom}(\partial f)$ the set $\operatorname{dom}(\partial f) = \{\partial f \neq \varnothing\}$. Notice that $\operatorname{dom}(\partial f) \subset \operatorname{dom}(f)$. For a function $g \in \Gamma_0(\mathcal{X}^*)$ and $\mu \in \mathcal{X}^*$, we define $\partial g(\mu)$ as

$$\partial g(\mu) = \left\{ x \in \mathcal{X} \,\middle|\, \forall \lambda \in \mathcal{X}^*,\ g(\lambda) \geq g(\mu) + \langle \lambda - \mu, x \rangle \right\}.$$

Let $E \subset \mathcal{X}$, $\bar{x} \in \mathcal{X}$, and F be a subset of some topological space. We denote as

- $\operatorname{conv}(E)$ the convex hull of E, and $\overline{\operatorname{conv}}(E)$ the set $\overline{\operatorname{conv}(E)}$. We omit the parentheses when E is given as the description of its elements;

- $\iota_E \colon \mathcal{X} \to \{0, +\infty\}$ the characteristic function of E, defined for all $x \in \mathcal{X}$ as

$$\iota_E(x) = \begin{cases} 0 & \text{if } x \in E, \\ +\infty & \text{otherwise;} \end{cases}$$

- $\sigma_E \colon \mathcal{X}^* \to \bar{\mathbb{R}}$ the support function of E, defined for all $\mu \in \mathcal{X}^*$ as

$$\sigma_E(\mu) = \sup_{x \in E} \langle \mu, x \rangle.$$

  We recall that we have $\sigma_E = \sigma_{\overline{\mathrm{conv}}(E)}$.

We now assume that E is nonempty. We denote as

- $N_E(\bar{x})$, the normal cone of E at $\bar{x}$, defined as $\partial \iota_E(\bar{x})$, or in explicit terms as

$$N_E(\bar{x}) = \begin{cases} \varnothing & \text{if } \bar{x} \notin E, \\ \left\{ \mu \in \mathcal{X}^* \,\middle|\, \forall x \in E,\ \langle \mu, x - \bar{x} \rangle \leq 0 \right\} & \text{otherwise;} \end{cases}$$

- $d(\,\cdot\,, E) \colon \mathcal{X} \to \mathbb{R}$ the distance to E, defined for all $x \in \mathcal{X}$ as

$$d(x, E) = \inf_{x' \in E} \|x - x'\|;$$

- $\mathcal{C}(E; F)$ the set of continuous functions over E taking values in F, as $\mathcal{C}(E)$ the space $\mathcal{C}(E; \mathbb{R})$, and as $\mathcal{C}_b(E)$ the subset of $\mathcal{C}(E)$ of bounded functions over E;
- $\mathcal{M}(E)$ the space of signed Radon measures over E, $\mathcal{M}^+(E)$ the subset of $\mathcal{M}(E)$ of nonnegative measures, and $\mathcal{P}(E)$ the set of probability measures over E, i.e., the subset of $\mathcal{M}^+(E)$ of measures m of total mass $m(E) = 1$.

Duality We denote the Legendre–Fenchel transform, or conjugate, of f as $f^* \colon \mathcal{X}^* \to \bar{\mathbb{R}}$. It is defined for all $\mu \in \mathcal{X}^*$ as

$$f^*(\mu) = \sup_{x \in \mathcal{X}} \langle \mu, x \rangle - f(x).$$

Notice that f^* lies in $\Gamma(\mathcal{X}^*)$, as a supremum of convex (lower semi-)continuous functions of μ. For $g \colon \mathcal{X}^* \to \bar{\mathbb{R}}$, its conjugate is the function $g^* \colon \mathcal{X} \to \bar{\mathbb{R}}$ defined for all $x \in \mathcal{X}$ as

$$g^*(x) = \sup_{\mu \in \mathcal{X}^*} \langle \mu, x \rangle - g(\mu),$$

and we denote as $f^{**} = (f^*)^*$ the biconjugate of f. Notice that, with this definition, for any $E \subset \mathcal{X}$, we have $\sigma_E = {\iota_E}^*$. It follows from the definition of the Legendre–Fenchel transform that, for any $f \colon \mathcal{X} \to \bar{\mathbb{R}}$, $g \colon \mathcal{X}^* \to \bar{\mathbb{R}}$, and $(x, \mu) \in \mathcal{X} \times \mathcal{X}^*$,

$$f(x) + f^*(\mu) \geq \langle \mu, x \rangle \quad \text{and} \quad g(\mu) + g^*(x) \geq \langle \mu, x \rangle, \tag{3}$$

these inequalities being known as the Fenchel–Young inequality. We also recall the Fenchel–Moreau Theorem [23, Theorem 2.3.3].

Theorem 2.1 (Fenchel–Moreau)**.** *Assume that $f \colon \mathcal{X} \to \bar{\mathbb{R}}^+$ and $f \not\equiv +\infty$. Then $f \in \Gamma_0(\mathcal{X})$ iff $f^{**} = f$, and in that case $f^* \in \Gamma_0(\mathcal{X}^*)$.*

The following result is an easy consequence of Theorem 2.1 and the definitions of f^*, ∂f, and ∂f^*.

Lemma 2.2. *Let $f \in \Gamma_0(\mathcal{X})$ and $(x, \mu) \in \mathcal{X} \times \mathcal{X}^*$. Then*

$$\mu \in \partial f(x) \Leftrightarrow f(x) + f^*(\mu) = \langle \mu, x \rangle \Leftrightarrow x \in \partial f^*(\mu).$$

Remark 2.3. Let $E \subset \mathcal{X}$ be a nonempty closed convex set and $\mu \in \mathcal{X}^*$. From Lemma 2.2, we have

$$\partial \sigma_E(\mu) = \operatorname*{argmax}_{v \in E} \langle \mu, v \rangle = \left\{ v \in E \,\middle|\, \sigma_E(\mu) = \langle \mu, v \rangle \right\}.$$

Using Remark 2.3 and the fact that $\inf_{v\in K}\langle \mu, v\rangle = -\sup_{v\in K}\langle \mu, -v\rangle$, we obtain at once the following result.

Corollary 2.4. *Let $K \subset \mathcal{X}$ be a nonempty closed convex set and $\mu \in \mathcal{X}^*$. Then*

$$\operatorname*{argmin}_{v\in K} \langle \mu, v\rangle = -\partial\sigma_{-K}(\mu)$$

We next recall the Fenchel–Rockafellar duality theorem (see [23, Theorem 2.8.3 and Corollary 2.8.5]), which plays a major role in this work.

Theorem 2.5 (Fenchel–Rockafellar)**.** *Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be two Banach spaces, $f \in \Gamma_0(\mathcal{X}_1)$, $g \in \Gamma_0(\mathcal{X}_2)$, and $A\colon \mathcal{X}_1 \to \mathcal{X}_2$ be a bounded linear operator. Assume that*

$$0 \in \operatorname{int}(A\operatorname{dom}(f) - \operatorname{dom}(g)). \tag{4}$$

Then, the following two problems have opposite values:

$$\operatorname*{minimize}_{x\in\mathcal{X}_1} \ f(x) + g(Ax), \tag{5}$$

$$\operatorname*{minimize}_{\mu\in\mathcal{X}_2^*} \ f^*(A^*\mu) + g^*(-\mu), \tag{6}$$

where $A^*\colon \mathcal{X}_2^* \to \mathcal{X}_1^*$ is the adjoint operator of A. Moreover, if $V < +\infty$, where V denotes the value of Problem (5), then Problem (6) has a solution.

Problem (6) will be called dual problem to (5). When a problem and its dual have opposite values, we say that they are in strong duality. We will call primal-dual gap of Problems (5) and (6) the quantity $\Delta(x,\mu)$ defined for $(x,\mu) \in \mathcal{X}_1 \times \mathcal{X}_2^*$ by

$$\Delta(x,\mu) = \big(f(x) + g(Ax)\big) + \big(f^*(A^*\mu) + g^*(-\mu)\big). \tag{7}$$

We next collect some elementary properties of the primal-dual gap.

Corollary 2.6. *Let $(x,\mu) \in \mathcal{X}_1 \times \mathcal{X}_2^*$. Then $\Delta(x,\mu) \geq 0$. Moreover, if (4) holds true, then the following statements are equivalent:*

(i) $\Delta(x,\mu) = 0$,

(ii) *x is a solution to Problem (5) and μ is a solution to Problem (6),*

(iii) *$A^*\mu \in \partial f(x)$ and $-\mu \in \partial g(Ax)$,*

(iv) *$x \in \partial f^*(A^*\mu)$ and $Ax \in \partial g^*(-\mu)$.*

Proof. Observe that, by definition of the adjoint operator,

$$\Delta(x,\mu) = \big(f(x) + f^*(A^*\mu) - \langle A^*\mu, x\rangle_{\mathcal{X}_1^*,\mathcal{X}_1}\big) + \big(g(Ax) + g^*(-\mu) + \langle \mu, Ax\rangle_{\mathcal{X}_2^*,\mathcal{X}_2}\big).$$

The nonnegativity of the primal-dual gap then follows from the Fenchel–Young inequality (3). The equivalence between *i* and *ii* is a consequence of Theorem 2.5. Also, $\Delta(x,\mu)$ is null *iff* both terms in the above decomposition are null. This is equivalent to *iii* and to *iv*, by Lemma 2.2, which concludes the proof. □

Bregman distances Let $\Xi \in \Gamma_0(\mathcal{E})$ be β-strongly convex, i.e., such that $\Xi - \frac{\beta}{2}\|\cdot\|_{\mathcal{E}}^2$ is convex. We denote as $B_\Xi\colon \mathcal{E} \times \operatorname{dom}(\Xi) \times \mathcal{E} \to \bar{\mathbb{R}}$ the Bregman distance associated with Ξ, which we define for all $\mu \in \mathcal{E}, \mu' \in \operatorname{dom}(\Xi)$, and $w \in \mathcal{E}$ as

$$B_\Xi(\mu,(\mu',w)) = \begin{cases} +\infty & \text{if } w \notin \partial\Xi(\mu'), \\ \Xi(\mu) - \Xi(\mu') - \langle \mu-\mu', w\rangle & \text{otherwise.} \end{cases} \tag{8}$$

Let us note that this definition of the Bregman distance is not quite the standard one, in which one usually requires Ξ to be differentiable. In this case, one can eliminate w from the above definition and replace it by $\nabla\Xi(\mu')$, which yields the standard definition. In particular, when $\Xi = \frac{1}{2}\|\cdot\|^2$, we have $B_\Xi(\mu,(\mu',w)) = \frac{1}{2}\|\mu'-\mu\|^2 + \iota_{\{0\}}(w-\mu')$.

Remark 2.7. Notice that, since Ξ is β-strongly convex, we have for all $\mu \in \mathcal{E}, \mu' \in \mathrm{dom}(\Xi)$, and $w \in \mathcal{E}$

$$B_\Xi(\mu, (\mu', w)) \geq \frac{\beta}{2}\|\mu - \mu'\|^2$$

and thus $B_\Xi(\mu, (\mu', w)) = 0$ *iff* $\mu = \mu'$ and $w \in \partial\Xi(\mu')$.

The following lemma follows from direct calculations.

Lemma 2.8. *For all $a, b \in \mathrm{dom}(\partial\Xi)$, $c \in \mathrm{dom}(\Xi)$, $w_a \in \partial\Xi(a)$, and $w_b \in \partial\Xi(b)$, we have the identity*

$$B_\Xi(c, (b, w_b)) + B_\Xi(b, (a, w_a)) - B_\Xi(c, (a, w_a)) = \langle c - b, w_a - w_b\rangle.$$

Coercive functions Let $f\colon \mathcal{X} \to \bar{\mathbb{R}}$ and $\alpha \in \mathbb{R}$. We call sublevel set of f at height α the set

$$\{f \leq \alpha\} = \left\{x \in \mathcal{X} \,\middle|\, f(x) \leq \alpha\right\}.$$

We say that f is coercive if

$$f(x) \underset{\|x\| \to +\infty}{\longrightarrow} +\infty.$$

Remark 2.9. Notice that a function f is coercive *iff* its sublevel sets are bounded. Also notice that the sublevel sets of a convex function are convex.

When $\mathcal{E}_1$ and $\mathcal{E}_2$ are two Hilbert spaces, with inner products $\langle\,\cdot\,,\cdot\,\rangle_{\mathcal{E}_1}$ and $\langle\,\cdot\,,\cdot\,\rangle_{\mathcal{E}_2}$ respectively, unless specified otherwise, we endow the product space $\mathcal{E} = \mathcal{E}_1 \times \mathcal{E}_2$ with the canonical inner product, defined for all $x_1, y_1 \in \mathcal{E}_1$ and all $x_2, y_2 \in \mathcal{E}_2$ as

$$\langle (x_1, x_2), (y_1, y_2)\rangle_{\mathcal{E}} = \langle x_1, y_1\rangle_{\mathcal{E}_1} + \langle x_2, y_2\rangle_{\mathcal{E}_2}.$$

Let $f_1\colon \mathcal{X}_1 \to \bar{\mathbb{R}}^+$ and $f_2\colon \mathcal{X}_2 \to \bar{\mathbb{R}}^+$. We denote their direct sum as $f_1 \oplus f_2\colon \mathcal{X}_1 \times \mathcal{X}_2 \to \bar{\mathbb{R}}^+$. It is defined, for all $x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2$ as

$$f_1 \oplus f_2(x_1, x_2) = f_1(x_1) + f_2(x_2).$$

The proof of the following lemma is straightforward.

Lemma 2.10. *Let $\mathcal{E}_1$, $\mathcal{E}_2$ be two Hilbert spaces. Let $f_1\colon \mathcal{E}_1 \to \bar{\mathbb{R}}^+$ and $f_2\colon \mathcal{E}_2 \to \bar{\mathbb{R}}^+$ be two coercive functions. Assume that both f_1 and f_2 are bounded from below. Then $f_1 \oplus f_2$ is coercive.*

We also state here the next lemma on the duality between functions with Lipschitz gradients and strongly convex functions, whose proof can be found in [3, Theorem 18.15 and Corollary 11.16].

Lemma 2.11. *Let $f \in \Gamma_0(\mathcal{E})$ and $\beta > 0$. Then f is Fréchet differentiable and its gradient ∇f is β-Lipschitz continuous if and only if the function f^* is $1/\beta$-strongly convex. In this case, f^* is also coercive.*

Perspective functions Let $f \in \Gamma_0(\mathcal{E})$. We denote its perspective function as $\tilde{f}\colon \mathcal{E} \times \mathbb{R} \to \bar{\mathbb{R}}^+$ and we define it, following [7], as

$$\tilde{f}(x, s) = \begin{cases} sf\left(\dfrac{x}{s}\right) & \text{if } s > 0, \\ \sup\limits_{y \in \mathrm{dom}(f)} f(y + x) - f(y) & \text{if } s = 0, \\ +\infty & \text{otherwise.} \end{cases}$$

We shall need in the sequel the following property of perspective functions, which can be found in [7, Proposition 2.3].

Lemma 2.12. *Let $f \in \Gamma_0(\mathcal{E})$. Then $\tilde{f} \in \Gamma_0(\mathcal{E} \times \mathbb{R})$. Moreover,*

$$(\tilde{f})^*(\mu, z) = \iota_{\mathrm{epi}(f^*)}(\mu, -z).$$

If f is Fréchet-differentiable, then for all $s > 0$ and $x \in \mathcal{E}$, $\tilde{f}$ is Fréchet-differentiable at (x, s) and

$$\nabla\tilde{f}(x, s) = \left(\nabla f\left(\frac{x}{s}\right), f\left(\frac{x}{s}\right) - \left\langle \nabla f\left(\frac{x}{s}\right), \frac{x}{s}\right\rangle\right). \tag{9}$$

Finally, if ∇f is continuous at $\frac{x}{s}$, then $\nabla\tilde{f}$ is continuous at (x, s).

3 The Extended Level-Set method

We introduce in Section 3.1 a prototype of the Frank–Wolfe algorithm, which contains as particular cases both the fully-corrective FWA and our DLS method. We give a dual interpretation of this method as a general prototype for a cutting-plane algorithm. In Section 3.2, we introduce our extension of the Level-Set method of [17], which we call Extended Level-Set (ELS) method. This method is a general cutting-plane-type algorithm, which will later yield the desired extension of the FWA by dualization. We give a convergence result for the ELS method in Section 3.3.

3.1 A dual point of view on the fully-corrective FWA

The FWA aims at solving problems of the form

$$\underset{x\in\mathcal{X}}{\text{minimize}}\ f(x)+\iota_K(x), \tag{p}$$

where $\mathcal{X}$ is a Banach space, K is a nonempty closed convex subset of $\mathcal{X}$, and $f\in\Gamma_0(\mathcal{X})$. The dual problem of Problem (p) is

$$\underset{\mu\in\mathcal{X}^*}{\text{minimize}}\ f^*(\mu)+\sigma_{-K}(\mu). \tag{d}$$

The primal dual-gap of Problems (p) and (d) is then given, for $(x,\mu)\in\mathcal{X}\times\mathcal{X}^*$, by

$$\Delta(x,\mu)=f(x)+\iota_K(x)+f^*(\mu)+\sigma_{-K}(\mu),$$

following (7). By Corollary 2.6, we have $\Delta(x,\mu)\geq 0$. Let us assume that $0\in\operatorname{int}(\operatorname{dom}(f)-K)$. Then Theorem 2.5 ensures that Problems (p) and (d) are in strong duality. Moreover, by Corollary 2.6, for any pair $(x,\mu)\in\mathcal{X}\times\mathcal{X}^*$, x and μ are respectively solutions to Problems (p) and (d) *iff* $\Delta(x,\mu)=0$.

Algorithm 1 below is a general form of the Frank–Wolfe algorithm. It relies on the linear minimization oracle $\mathrm{LMO}\colon\mathcal{X}^*\to K$, which is such that

$$\forall\mu\in\mathcal{X}^*,\quad \mathrm{LMO}(\mu)\in\underset{x\in K}{\operatorname{argmin}}\ \langle\mu,x\rangle. \tag{LMO}$$

Algorithm 1 covers the classical fully-corrective FWA, in the case where f is continuously differentiable with a Lipschitz-continuous gradient: it suffices to fix a point $x^0\in K$, to define $\mu^0=\nabla f(x^0)$, $K^{-1}=\{x^0\}$, and finally, at each iteration, to define μ^{t+1} as $\nabla f(\hat{x}^{t+1})$ in the last step. The **Dual update** step is facultative and can be omitted in general. Notice that in the case of the fully-corrective FWA, this step is implementable without explicit knowledge of f^* and σ_{-K} since, for $\mu^t=\nabla f(\hat{x}^t)$, we have

$$f^*(\mu^t)=\langle\mu^t,\hat{x}^t\rangle-f(\hat{x}^t)\quad\text{and}\quad\sigma_{-K}(\mu^t)=-\langle\mu^t,v^t\rangle,$$

by Lemma 2.2 and Corollary 2.4.

Algorithm 2 is equivalent to Algorithm 1. First we note that the two **Oracle** steps are equivalent, as a direct consequence of Corollary 2.4. We next notice that Problem (d^t) is the dual of Problem (p^t). Also, for the same reason as with Problems (p) and (d), Problems (p^t) and (d^t) are in strong duality, which ensures that the definitions of $\underline{h}^t$ in the algorithms are equivalent. We note that, in the case of the fully-corrective FWA, defining μ^{t+1} as $\nabla f(\hat{x}^{t+1})$ is equivalent to directly define μ^{t+1} as a solution to Problem (d^t).

Algorithm 2 can be seen as a general cutting-plane method for the dual problem (d). By construction, $-v^t\in\partial\sigma_{-K}(\mu^t)$, so the map $\mu\mapsto\langle\mu,-v^t\rangle$ is a linear lower bound of σ_{-K}, exact at $\mu=\mu^t$. This implies that σ_{-K^t} is a lower approximation of σ_{-K}, which is exact at the points $\mu^0,\dots,\mu^t$. If moreover K^{-1} is the convex hull of a finite number of points, then σ_{-K^t} is piecewise affine. We will call the map $\mu\mapsto\langle\mu,-v^t\rangle$ a cut, and by extension, we will simply call cut any element of K^t.

Let us comment on the role of the quantities $\bar{h}^t$ and $\underline{h}^t$. Denote by V_d the value of Problem (d). From the definition of $\bar{h}^t$, we directly see that it is an upper bound of V_d. Since $\sigma_{-\tilde{K}^t}\leq\sigma_{-K}$ and since $\underline{h}^t$ is the value of Problem (d^t), we deduce that $\underline{h}^t$ is a lower bound of V_d. This implies in particular that the candidate $\hat{\mu}^t$ obtained at iteration t of the algorithm is $(\bar{h}^t-\underline{h}_t)$-optimal. We can retrieve this property by noticing that

$$\bar{h}^t-\underline{h}^t=\Delta(\hat{x}^t,\hat{\mu}^t).$$

The interpretation of the quantity $(\bar{h}^t-\underline{h}^t)$ as a primal-dual gap is of a key importance for the design of the desired extension of the FWA.

Algorithm 1: FWA for Problem (p)

Require: $\mu^0 \in \mathrm{dom}(f^*)$;
Find $K^{-1} \subset K$ such that
$0 \in \mathrm{int}\big(\mathrm{dom}(f) - K^{-1}\big)$;
for $t = 0, \ldots$ **do**
Available at iteration t:
$\mu^t \in \mathcal{E}$, $K^{t-1} \subset K$;
Oracle:
Set $v^t = \mathrm{LMO}(\mu^t)$;
Dual update:
Optional. Take $\hat{\mu}^t$ a solution to

$$\underset{\mu \in \{\mu^0, \ldots, \mu^t\}}{\text{minimize}}\ f^*(\mu) + \sigma_{-K}(\mu);$$

Set $\bar{h}^t = f^*(\hat{\mu}^t) + \sigma_{-K}(\hat{\mu}^t)$;
Primal update:
Set $K^t = K^{t-1} \cup \{v^t\}$;
Find a solution $\hat{x}^t$ of

$$\underset{x \in \overline{\mathrm{conv}}(K^t)}{\text{minimize}}\ f(x); \qquad (p^t)$$

Set $\underline{h}^t = -f\left(\hat{x}^{t+1}\right)$;
Dual candidate:
Generate a new candidate μ^{t+1}.
end

Algorithm 2: Dual FWA for Problem (d)

Require: $\mu^0 \in \mathrm{dom}(f^*)$;
Find $K^{-1} \subset K$ such that
$0 \in \mathrm{int}\big(\mathrm{dom}(f) - K^{-1}\big)$;
for $t = 0, \ldots$ **do**
Available at iteration t:
$\mu^t \in \mathcal{E}$, $K^{t-1} \subset K$;
Oracle:
Find $v^t \in -\partial\sigma_{-K}(\mu^t)$;
Dual update:
Take a solution $\hat{\mu}^t$ to

$$\underset{\mu \in \{\mu^0, \ldots, \mu^t\}}{\text{minimize}}\ f^*(\mu) + \sigma_{-K}(\mu);$$

Set $\bar{h}^t = f^*(\hat{\mu}^t) + \sigma_{-K}(\hat{\mu}^t)$;
Primal update:
Set $K^t = K^{t-1} \cup \{v^t\}$;
Find a solution ν^{t+1} of

$$\underset{\mu \in \mathcal{E}}{\text{minimize}}\ f^*(\mu) + \sigma_{-K^t}(\mu); \qquad (d^t)$$

Set $\underline{h}^t = f^*\big(\nu^{t+1}\big) + \sigma_{-K^t}\big(\nu^{t+1}\big)$;
Dual candidate:
Generate a new candidate μ^{t+1}.
end

Let us recall our general objective: generalizing the FWA to the case of problems with nonsmooth costs f, utilizing the duality with cutting-plane-type algorithms. At a dual level, this means that we do not want to assume f^* to be strongly convex, which does not seem restrictive at the first sight, since for basic cutting-plane methods, f^* is simply the characteristic function of some given closed and bounded feasible set. By basic, we have in mind the methods for which one simply defines the next dual candidate μ^{t+1} as ν^{t+1}. Though these methods are known to converge, the convergence is in practice slow (see [5, Section 9.3.2]); moreover, to ensure the convergence of the corresponding FWA (obtained by "back"-dualization) to a minimizer, we need to ensure that the primal-dual gap $\bar{h}^t - \underline{h}^t$ converges to 0. In view of our objectives, our attention has focused on the cutting-plane method calld Level-Set method proposed and analyzed in [17, Section 2.2.1], for which the convergence of the primal-dual gap is known. As we already pointed out in the introduction, we need to utilize a double extension of this method, since f^* is restricted to be a characteristic function in [17, Section 2.2.1] and since a certain projection step realized for the generation of a novel dual candidate (in the last step of the algorithm) must also be generalized. The next section is dedicated to the generalization of the Level-Set method.

3.2 Statement of the ELS method

The aim of this section is to present an algorithm, which we call Extended Level-Set (ELS) method, to solve problems of the form

$$\underset{\mu \in \mathcal{E}}{\text{minimize}}\ \psi(\mu) + \sigma_E(\mu), \qquad (D)$$

where $\mathcal{E}$ is a Hilbert space, $\psi \in \Gamma_0(\mathcal{E})$, and $E \subset \mathcal{E}$ is a nonempty closed convex set. As its name suggests, the ELS method is an extension of the Level-Set method proposed in [17, Section 2.2.1], which is itself an extension of the Cutting-Plane Algorithm.

We denote by V_D the value of Problem (D). The problem dual to Problem (D) is

$$\underset{x \in -E}{\text{minimize}}\ \psi^*(x) \qquad (P)$$

and we denote by V_P its value. Given x and μ in $\mathcal{E}$, according to (7), the primal-dual gap between

Problems (D) and (P) is

$$\Delta(x,\mu) = \psi(\mu) + \sigma_E(\mu) + \psi^*(x) + \iota_{-E}(x).$$

A direct modification of the proof of Corollary 2.6 shows that $\Delta(x,\mu) = 0$ if and only if μ is a solution to Problem (D), x is a solution to Problem (P), and $V_P + V_D = 0$.

We fix a β-strongly convex function $\Xi \in \Gamma_0(\mathcal{E})$. Recall that B_Ξ denotes the associated Bregman distance, in the sense of the definition (8). The ELS method is described in Algorithm 3. Note that the **Primal update** involves a function called pruning, whose output is a subset of E. For the moment, we simply take $E^t = E^{t-1} \cup \{v^t\}$. We first investigate the convergence of the algorithm in this setting; we will later propose some pruning rules which preserve the convergence speed of our algorithm (see the last paragraph of Section 3.3).

Algorithm 3: Extended Level-Set method for Problem (D)

Require: $\mu^0 \in \mathrm{dom}(\sigma_E) \cap \mathrm{dom}(\partial\Xi) \cap \mathrm{dom}(\psi)$, $w^0 \in \partial\Xi(\mu^0)$, $E^{-1} \subset E$, $\lambda \in (0,1)$;
Set $\bar{h}^{-1} = +\infty$;
for $t = 0, \dots$ **do**
  Available at iteration t: $\mu^t \in \mathcal{E}$, $w^t \in \mathcal{E}$, $E^{t-1} \subset E$, $\bar{h}^{t-1} \in \bar{\mathbb{R}}^+$;
  Oracle:
    Find $v^t \in \partial\sigma_E(\mu^t)$;
  Dual update:
    Set $\bar{h}^t = \min\left\{\bar{h}^{t-1}, \psi(\mu^t) + \sigma_E(\mu^t)\right\}$;
  Primal update:
    Set $\underline{h}^t = \inf_{\mu\in\mathcal{E}} \psi(\mu) + \sigma_{E^{t-1}\cup\{v^t\}}(\mu)$;
    Set $E^t = \mathrm{pruning}\left(E^{t-1} \cup \{v^t\}\right) \subset \overline{\mathrm{conv}}\left(E^{t-1} \cup \{v^t\}\right)$;
  Dual candidate:
    Set $\Delta^t = \bar{h}^t - \underline{h}^t$;
    Set $\ell^t = \lambda\bar{h}^t + (1-\lambda)\underline{h}^t$;
    Set $Q^t = \{\psi + \sigma_{E^t} \le \ell^t\}$;
    Set a new candidate μ^{t+1} as the solution to:

$$\underset{\mu\in Q^t}{\text{minimize}}\ B_\Xi\left(\mu, \left(\mu^t, w^t\right)\right); \tag{10}$$

    Take $w^{t+1} \in \partial\Xi\left(\mu^{t+1}\right)$ such that $w^t - w^{t+1} \in N_{Q^t}(\mu^{t+1})$;
end

Remark 3.1. We want to highlight Algorithm 3 as being a specific instance of Algorithm 2, where ψ plays the role of f^*, E and E^t that of $-K$ and $-K^t$ respectively, and v^t is replaced by its opposite vector.

We now present the elements which support our claim that this algorithm is an extension of the Level-Set method.

- In our algorithm, we keep the function ψ as is and take subgradients of σ_E, whereas in [17, Section 2.2.1], subgradients of the whole cost function are taken.
- The step described in Problem (10) is a projection step of μ^t onto the set Q^t, following the Bregman distance associated with Ξ. Note that, when Ξ is differentiable, using a first-order optimality condition for μ^{t+1} in (10), we have that $w^t - \nabla\Xi(\mu^{t+1}) \in N_{Q^t}(\mu^{t+1})$, which ensures the existence of w^{t+1} in the very last step.

We now present a list of hypotheses under which the algorithm is well-defined and converges. Let us stress that some of these assumptions are not explicit, for example Assumption (H6) below. Though it would be easy to transform these assumptions into explicit ones by slightly weakening them, we recall that our main interest does not lie in the ELS method as such but rather in its dual counterpart, our DLS method, for which explicit assumptions will be made later on.

(H1) We have $\mathrm{dom}(\sigma_E) \cap \mathrm{dom}(\partial\Xi) \cap \mathrm{dom}(\psi) \neq \varnothing$.

(H2) For all $t \in \mathbb{N}$, $\partial\sigma_E(\mu^t) \neq \varnothing$. There exists a constant C_{oracle} such that, for all $t \in \mathbb{N}$, the vector v^t verifies $\|v^t\| \le C_{\text{oracle}}$.

(H3) The set E^{-1} required at the beginning of the algorithm is bounded and such that $\psi + \sigma_{E^{-1}}$ is coercive.

We fix a constant $C_{E^{-1}}$ such that $E^{-1} \subset B(0, C_{E^{-1}})$. We denote as Q^{-1} the set $\left\{\psi + \sigma_{E^{-1}} \leq \bar{h}^0\right\}$. This set is convex, nonempty as a consequence of Assumption (H1), and bounded as a consequence of Assumption (H3). Let then $C_{Q^{-1}} > 0$ be such that $Q^{-1} \subset B(0, C_{Q^{-1}})$.

(H4) We have $Q^{-1} \subset \mathrm{dom}(\Xi)$ and the function Ξ is bounded over the set Q^{-1} by a constant $C_{\Xi,Q^{-1}} > 0$.

(H5) For all $t \in \mathbb{N}$, there exists $w^{t+1} \in \partial\Xi\big(\mu^{t+1}\big)$ such that $w^t - w^{t+1} \in N_{Q^t}(\mu^{t+1})$. There exists a constant $C_{\partial\Xi} > 0$ such that for all $t \in \mathbb{N}$, $\|w^t\| \leq C_{\partial\Xi}$.

(H6) There exists a constant $C_\psi > 0$ such that, for all $t \in \mathbb{N}$,

$$\left|\psi\left(\mu^{t+1}\right) - \psi\left(\mu^t\right)\right| \leq C_\psi \left\|\mu^{t+1} - \mu^t\right\|.$$

In the rest of this section, we assume that all these assumptions are verified and we recall that, except for the last paragraph of this section, we assume that we take $E^t = E^{t-1} \cup \{v^t\}$ in the **Primal update** step of Algorithm 3. We next provide some first properties of Algorithm 3.

Proposition 3.2. *The Extended Level-Set method from Algorithm 3 is well-posed and, for any $t \in \mathbb{N}$, we have*

$$-\infty < \underline{h}^t \leq \underline{h}^{t+1} \leq -V_P \leq V_D \leq \bar{h}^{t+1} \leq \bar{h}^t < +\infty. \tag{11}$$

Moreover, for any $t \in \mathbb{N}$, there exist two elements $\hat{x}^t$ and $\hat{\mu}^t$ such that

$$\hat{x}^t \in \operatorname*{argmin}_{x\in\mathcal{E}} \psi^*(x) + \iota_{-\overline{\mathrm{conv}}(E^{t-1}\cup\{v^t\})}(x) \quad \textit{and} \quad \hat{\mu}^t \in \operatorname*{argmin}_{\mu\in\{\mu^0,\ldots,\mu^t\}} \psi(\mu) + \sigma_E(\mu). \tag{12}$$

We also have

$$-\underline{h}^t = \psi^*\left(\hat{x}^t\right) + \iota_{-E}\left(\hat{x}^t\right), \quad \bar{h}^t = \psi\left(\hat{\mu}^t\right) + \sigma_E\left(\hat{\mu}^t\right), \quad \textit{and} \quad \Delta^t = \Delta\left(\hat{x}^t, \hat{\mu}^t\right). \tag{13}$$

Finally, the sequence $(\Delta^t)_{t\in\mathbb{N}}$ is nonincreasing.

Proof. We do a proof by induction. We claim that, for any $t \in \mathbb{N}$, the algorithm can be run until the beginning of iteration t and that the following is satisfied:

$$\mu^t \in \mathrm{dom}(\psi), \quad w^t \in \partial\Xi(\mu^t), \quad E^{-1} \subset \overline{\mathrm{conv}}\big(E^{t-1}\big), \quad \text{and} \quad t \geq 1 \Rightarrow \bar{h}^{t-1} \leq \bar{h}^0.$$

We also claim that E^{t-1} is a bounded subset of E. For $t = 0$, the claim follows from Assumption (H1). Let us assume that it is satisfied for some $t \in \mathbb{N}$ and let us consider the execution of the iteration t of the method.

The **Oracle** step is well-defined, by Assumption (H2), which also implies that $\mu^t \in \mathrm{dom}(\sigma_E)$. Concerning the **Dual update** step, since $\mu^t \in \mathrm{dom}(\sigma_E) \cap \mathrm{dom}(\psi)$, we have $\psi(\mu^t) + \sigma_E(\mu^t) < +\infty$ and thus $\bar{h}^t < \infty$. If $t = 0$, then $\bar{h}^t = \bar{h}^0$. If $t \geq 1$, then $\bar{h}^t \leq \bar{h}^{t-1} \leq \bar{h}^0$.

Let us move to the **Primal update** step. Since $E^{-1} \subset \overline{\mathrm{conv}}\big(E^{t-1} \cup \{v^t\}\big)$, we have that $\psi + \sigma_{E^{-1}} \leq \psi + \sigma_{E^{t-1}\cup\{v^t\}}$, which implies that $\psi + \sigma_{E^{t-1}\cup\{v^t\}}$ is coercive, by Assumption (H3). Since this function is also convex and lower semicontinuous, it has a minimizer ν^t and $\underline{h}^t$ is finite, by [3, Proposition 11.14]. Since $E^{t-1} \cup \{v^t\}$ is bounded, then so is E^t, and we have $\mathrm{dom}\big(\sigma_{E^{t-1}\cup\{v^t\}}\big) = \mathcal{E}$. Applying Theorem 2.5, we deduce that

$$-\underline{h}^t = \inf_{x\in\mathcal{E}} \psi^*(x) + \iota_{-\overline{\mathrm{conv}}(E^{t-1}\cup\{v^t\})}(x)$$

and that the above problem has a solution $\hat{x}^t$. This implies that $\underline{h}^t \leq -V_P \leq V_D$. By construction, $\hat{x}^t \in -\overline{\mathrm{conv}}\big(E^{t-1} \cup \{v^t\}\big) \subset -E$, which implies that $-\underline{h}^t = \psi^*(\hat{x}^t) + \iota_{-E}(\hat{x}^t)$. From the definition of E^t, we have that $E^{-1} \subset \overline{\mathrm{conv}}(E^t)$.

Finally, we discuss the **Dual candidate** step. We have $\Delta^t \geq 0$, because $\bar{h}^t \geq V_D \geq -V_P$ and $\underline{h}^t \leq -V^P$. By definition of ℓ^t, we have $\underline{h}^t \leq \ell^t \leq \bar{h}^t$. Since moreover $E^t \subset \overline{\mathrm{conv}}\big(E^{t-1} \cup \{v^t\}\big)$, we have

$$\psi(\nu^t) + \sigma_{E^t}(\nu^t) \leq \psi(\nu^t) + \sigma_{E^{t-1}\cup\{v^t\}}(\nu^t) = \underline{h}^t \leq \ell^t,$$

which implies that $\nu^t \in Q^t$ and thus Q^t is nonempty. Since $\bar{h}^t \leq \bar{h}^0$ and $\sigma_{E^t} \geq \sigma_{E^{-1}}$, we deduce that $Q^t \subset Q^{-1} \subset \mathrm{dom}(\Xi) = \mathrm{dom}(B_\Xi(\cdot, (\mu^t, w^t)))$, by Assumption (H4). This implies that Problem (10) has

a solution μ^{t+1}. Note that $\mu^{t+1} \in Q^t \subset \mathrm{dom}(\psi)$. Finally, the existence of $w^{t+1} \in \partial\Xi(\mu^{t+1}) \cap \big(w^t - N_{Q^t}(\mu^{t+1})\big)$ is ensured by Assumption (H5). Therefore, the claim is verified for $t+1$, which proves the well-posedness of the algorithm.

Concerning the proof of (11), it remains to prove that $\bar{h}^t$ is nonincreasing and that $\underline{h}^t$ is nondecreasing. It is obvious that $\bar{h}^t$ is nonincreasing. The fact that $\underline{h}^t \leq \underline{h}^{t+1}$ is a consequence of the inclusion $E^{t-1} \cup \{v^t\} \subset \overline{\mathrm{conv}}\big(E^t \cup \{v^{t+1}\}\big)$, which ensures that $\sigma_{E^{t-1}\cup\{v^t\}} \leq \sigma_{E^t\cup\{v^{t+1}\}}$.

Next, concerning (12) and (13), we have already proved the existence of $\hat{x}^t$ and we have already justified that $-\underline{h}^t = \psi^*(\hat{x}^t) + \iota_{-E}(\hat{x}^t)$. The existence of $\hat{\mu}^t$ and the fact that $\bar{h}^t = \psi(\hat{\mu}^t) + \sigma_E(\hat{\mu}^t)$ is straightforward. The equality $\Delta^t = \Delta(\hat{x}^t, \hat{\mu}^t)$ is then a simple consequence of the definition of the primal-dual gap. Finally, $(\Delta^t)_{t\in\mathbb{N}}$ in nonincreasing because $(\bar{h}^t)_{t\in\mathbb{N}}$ and $(\underline{h}^t)_{t\in\mathbb{N}}$ are respectively nonincreasing and nondecreasing. This concludes the proof of the proposition. □

Remark 3.3. Problem (10) has a unique solution since Ξ is strongly convex and Q^t is convex.

3.3 Convergence analysis

The aim of this subsection is to prove the convergence of Algorithm 3 under the previous assumptions and provide its convergence speed. We start by noticing that, exactly as in the proof of the lemma inside [17, Theorem 2.2.1], we have the following property.

Lemma 3.4. *Let $t_1 \leq t_2$ be such that $\Delta^{t_2} \geq (1-\lambda)\Delta^{t_1}$. Then $\underline{h}^{t_2} \leq \ell^{t_1}$.*

The next theorem deals with the convergence of Algorithm 3. Its proof follows closely the analysis in [17, Theorem 2.2.1], but additional care is needed in order to take into account our generalizations, in particular the use of the Bregman distance B_Ξ for the projection step. We hence provide a detalied proof below, which makes use of the constants C_{oracle}, $C_{E^{-1}}$, $C_{Q^{-1}}$, $C_{\Xi,Q^{-1}}$, $C_{\partial\Xi}$, and C_ψ, introduced with Assumptions (H2) to (H6).

Theorem 3.5. *Consider the Extended Level-Set method from Algorithm 3 under Assumptions (H1) to (H6) and with $E^t = E^{t-1} \cup \{v^t\}$ in the* **Primal update** *step. The primal-dual gap Δ^t converges to 0 with a speed of order $1/\sqrt{t}$, i.e., there exists $C > 0$ such that, for all $t \in \mathbb{N}^*$,*

$$\Delta^t \leq \frac{C}{\sqrt{t}}.$$

Proof. **Step 1.** Let $T \in \mathbb{N}$ and set $\varepsilon = \Delta^T$ and $I = \{0, \ldots, T\}$. We recall that, using the monotonicity of $(\Delta^t)_{t\in\mathbb{N}}$, we have $\varepsilon = \inf_{t\in I} \Delta^t$. We split I in a partition $I_1, \ldots, I_m$ as follows:

- We set $p = 0$ and $i_0 = -1$.
- While $i_p < T$, we set

$$i_{p+1} = \max\big\{t \in \{0, \ldots, T\} \,\big|\, \Delta^t \geq (1-\lambda)\Delta^{i_p+1}\big\} \quad \text{and} \quad I_{p+1} = \{i_p+1, \ldots, i_{p+1}\}$$

  and we increment p by 1.

Following [17], for all $p \in \{0, \ldots, m-1\}$, the iteration $i_p + 1$ is called *critical*. Notice that, using the monotonicity of the sequence $(\Delta^t)_{t\in\mathbb{N}}$, we have, for all $p \in \{1, \ldots, m\}$ and $t \in I_p$,

$$\Delta^t \geq (1-\lambda)\Delta^{i_{p-1}+1}. \tag{14}$$

Now, let $p \in \{1, \ldots, m\}$ and χ^p be a minimizer of $\psi + \sigma_{E^{i_p}}$. Notice that such a minimizer exists, since $\psi + \sigma_{E^{i_p}} \in \Gamma_0(\mathcal{E})$ and is coercive using Assumption (H3). Then, Lemma 3.4 applied with $t_1 = t \in I_p$ and $t_2 = i_p$ shows that

$$\psi(\chi^p) + \sigma_{E^{i_p}}(\chi^p) = \underline{h}^{i_p} \leq \ell^t.$$

Since for all $t \in I_p$, we have $\sigma_{E^t} \leq \sigma_{E^{i_p}}$, this yields $\chi^p \in \bigcap_{t\in I_p} Q^t$. This construction holds for any $p \in \{1, \ldots, m\}$.

Step 2. Let $p \in \{1, \ldots, m\}$. For all $t \in \mathbb{N}$, we set $\tau_p^t = B_\Xi(\chi^p, (\mu^t, w^t)) \geq 0$. Now, let $t \in I_p$. We use the identity from Lemma 2.8 with $a = \mu^t$, $b = \mu^{t+1}$, $c = \chi^p$, $w_a = w^t$, and $w_b = w^{t+1}$, which yields

$$\tau_p^{t+1} + B_\Xi\big(\mu^{t+1}, (\mu^t, w^t)\big) - \tau_p^t = \big\langle w^t - w^{t+1}, \chi^p - \mu^{t+1}\big\rangle.$$

Reordering and using Assumption (H5) yields

$$0 \leq \tau_p^{t+1} \leq \tau_p^t - B_\Xi\big(\mu^{t+1}, (\mu^t, w^t)\big). \tag{15}$$

In turn, using Remark 2.7, we have

$$0 \leq \tau_p^{t+1} \leq \tau_p^t - \frac{\beta}{2}\left\|\mu^t - \mu^{t+1}\right\|^2 \leq \tau_p^t - \frac{\beta}{2C^2}\left|\psi\left(\mu^{t+1}\right) + \sigma_{E^t}\left(\mu^{t+1}\right) - \psi\left(\mu^t\right) - \sigma_{E^t}\left(\mu^t\right)\right|^2, \tag{16}$$

where $C = C_\psi + \max(C_{E^{-1}}, C_{\text{oracle}})$. Indeed,

$$\begin{aligned}\left|\psi\left(\mu^{t+1}\right) + \sigma_{E^t}\left(\mu^{t+1}\right) - \psi\left(\mu^t\right) - \sigma_{E^t}\left(\mu^t\right)\right| &\leq \left|\psi\left(\mu^{t+1}\right) - \psi\left(\mu^t\right)\right| + \left|\sigma_{E^t}\left(\mu^{t+1}\right) - \sigma_{E^t}\left(\mu^t\right)\right| \\ &\leq C\left\|\mu^{t+1} - \mu^t\right\|,\end{aligned}$$

using Assumption (H6) and the fact that $E^t \subset B(0, \max(C_{E^{-1}}, C_{\text{oracle}}))$, which is itself a consequence of the definition of E^t and of Assumptions (H2) and (H3). Moreover, we know that

$$\left|\psi\left(\mu^{t+1}\right) + \sigma_{E^t}\left(\mu^{t+1}\right) - \psi\left(\mu^t\right) - \sigma_{E^t}\left(\mu^t\right)\right| \geq (1-\lambda)\Delta^t, \tag{17}$$

since

$$\begin{aligned}\psi\left(\mu^t\right) + \sigma_{E^t}\left(\mu^t\right) - \psi\left(\mu^{t+1}\right) - \sigma_{E^t}\left(\mu^{t+1}\right) &\geq \psi\left(\mu^t\right) + \sigma_{E^t}\left(\mu^t\right) - \ell^t \\ &\geq \bar{h}^t - \ell^t \\ &= (1-\lambda)\Delta^t,\end{aligned}$$

where we use, in order, the fact that $\psi\left(\mu^{t+1}\right) + \sigma_{E^t}\left(\mu^{t+1}\right) \leq \ell^t$ and the definitions of $\bar{h}^t$ and of Δ^t. Combining eqs. (16) and (17) yields

$$0 \leq \tau_p^{t+1} \leq \tau_p^t - \frac{\beta}{2C^2}\left((1-\lambda)\Delta^t\right)^2,$$

which implies, using the nonnegativeness and nonincreasingness of the sequence $(\Delta^t)_{t\in\mathbb{N}}$,

$$0 \leq \tau_p^{t+1} \leq \tau_p^t - \frac{\beta}{2C^2}\left((1-\lambda)\Delta^{i_p}\right)^2. \tag{18}$$

Taking $t = i_p$ and iterating $i_p - i_{p-1} - 1$ times eq. (18) yields

$$0 \leq \tau_p^{i_p+1} \leq \tau_p^{i_{p-1}+1} - (i_p - i_{p-1})\frac{\beta}{2C^2}\left((1-\lambda)\Delta^{i_p}\right)^2. \tag{19}$$

Notice also that

$$\tau_p^{i_{p-1}+1} \leq 2\left(C_{\Xi,Q^{-1}} + C_{\partial\Xi}C_{Q^{-1}}\right), \tag{20}$$

since, for all $t \in \mathbb{N}$,

$$\tau_p^t \leq \left|\Xi\left(\chi^p\right) - \Xi\left(\mu^t\right)\right| + \left|\left\langle w^t, \chi^p - \mu^t\right\rangle\right|.$$

Combining eqs. (19) and (20) yields

$$0 \leq 2\left(C_{\Xi,Q^{-1}} + C_{\partial\Xi}C_{Q^{-1}}\right) - (i_p - i_{p-1})\frac{\beta}{2C^2}\left((1-\lambda)\Delta^{i_p}\right)^2,$$

and thus

$$|I_p| = i_p - i_{p-1} \leq \frac{4C^2\left(C_{\Xi,Q^{-1}} + C_{\partial\Xi}C_{Q^{-1}}\right)}{\beta\left((1-\lambda)\Delta^{i_p}\right)^2} = \frac{\bar{C}}{\left((1-\lambda)\Delta^{i_p}\right)^2}. \tag{21}$$

Step 3. Since we have, by definition of the indices i_p and nonincreasingness of $(\Delta^t)_{t\in\mathbb{N}}$,

$$\Delta^{i_{m-1}+1} \geq \Delta^{i_m} = \Delta^T = \varepsilon \quad \text{and} \quad \Delta^{i_{p+1}} \geq (1-\lambda)\Delta^{i_p+1} > \Delta^{i_{p+1}+1} \text{ for all } p \in \{1, \ldots, m-2\},$$

then for all $p \in \{1, \ldots, m-1\}$, we have

$$\Delta^{i_p} \geq \frac{\varepsilon}{(1-\lambda)^{m-1-p}}. \tag{22}$$

Summing eq. (21) then yields

$$T+1=\sum_{p=1}^{m}|I_p| \leq \frac{\bar{C}}{(1-\lambda)^2}\sum_{p=1}^{m}\left(\frac{1}{\Delta^{i_p}}\right)^2 \leq \frac{\bar{C}}{(1-\lambda)^2}\left(\frac{1}{\varepsilon^2}+\sum_{p=1}^{m-1}\frac{(1-\lambda)^{2(m-1-p)}}{\varepsilon^2}\right)$$
$$\leq \frac{\bar{C}}{(1-\lambda)^2\varepsilon^2}\left(1+\sum_{p\in\mathbb{N}}(1-\lambda)^{2p}\right)=\frac{\bar{C}}{(1-\lambda)^2\varepsilon^2}\left(1+\frac{1}{\lambda(2-\lambda)}\right), \tag{23}$$

which shows the expected result. □

Remark 3.6. We can minimize eq. (23) with respect to λ, which gives $\bar{\lambda}=1-\sqrt{2-\sqrt{2}}\approx 0.23$.

The pruning step As is, the main issue with the Extended Level-Set method from Algorithm 3 is that we need to keep track of all the subgradients v^t. This may become costly in terms of memory and of computation time of $\underline{h}^t$ and of μ^{t+1}, as the set E^t appears in the definition of Q^t. This is why we propose to apply pruning steps in our algorithm. By pruning, we mean that we want to define E^t as a smaller set than $E^{t-1}\cup\{v^t\}$, in the sense that $E^t\subset\overline{\text{conv}}\big(E^{t-1}\cup\{v^t\}\big)$. Note that imposing $E^t\subset\overline{\text{conv}}\big(E^{t-1}\cup\{v^t\}\big)$ at each iteration implies that $E^t\subset\overline{\text{conv}}\big(E^{-1}\cup\{v^0,\ldots,v^t\}\big)$ and gives thus the boundedness of E^t.

Pruning offers the possibility to chose a set E^t with a small cardinality, so as to simplify the implementation of the **Primal update** and **Dual candidate** steps. We establish in this paragraph some sufficient properties on the choice of E^t which ensure that Algorithm 3 remains well-posed and that its convergence properties are preserved. The proof of Proposition 3.2 reveals that the algorithm indeed remains well-posed if we require that $E^{-1}\subset\overline{\text{conv}}(E^t)$. The convergence proof of Theorem 3.5 still holds if the following holds.

(i) The sequences $\big(\bar{h}^t\big)_{t\in\mathbb{N}}$ and $\big(\underline{h}^t\big)_{t\in\mathbb{N}}$ keep the same monotonicity as shown in (11).

(ii) The function $\psi+\sigma_{E^t}$ remains coercive.

(iii) We are able to find $\chi^p\in\bigcap_{t\in I_p}Q^t$, where I_p describes a subinterval defined in the proof of Theorem 3.5.

For property (iii) to hold, we decide to only make pruning steps at the critical iterations. Notice that detecting a critical iteration is easy, since it only requires to keep track of Δ^j, where j denotes the last critical iteration. For property (ii) to hold, we only need to keep $E^{-1}\subset\overline{\text{conv}}(E^t)$ after the pruning step, as we already required for the well-posedness of the algorithm. Lastly, for property (i), notice that the monotonicity of the sequence $\big(\bar{h}^t\big)_{t\in\mathbb{N}}$ is preserved, and for the monotonicity of $\big(\underline{h}^t\big)_{t\in\mathbb{N}}$, it suffices that $\underline{h}^t$ be preserved by pruning steps, i.e., that, for all $t\in\mathbb{N}$,

$$\underline{h}^t=\tilde{h}^t:=\inf_{\mu\in\mathcal{E}}\psi(\mu)+\sigma_{E^t}(\mu). \tag{24}$$

By definition, this equality holds at noncritical iterations. Let then t be a critical iteration. Since $E^t\subset\overline{\text{conv}}\big(E^{t-1}\cup\{v^t\}\big)$, we have $\bar{h}^t\geq\tilde{h}^t$. Since E^t is bounded, we have

$$\tilde{h}^t=-\min_{x\in\mathcal{E}}\psi^*(x)+\iota_{-\overline{\text{conv}}(E^t)}(x). \tag{25}$$

We recall that we proved in Proposition 3.2 that $\underline{h}_t=-\psi^*(\hat{x}^t)$ and that $\hat{x}^t\in\overline{\text{conv}}\big(E^{t-1}\cup\{v^t\}\big)$. Therefore, to ensure that $\tilde{h}^t\geq\bar{h}^t$ (and thus for property (i) to hold) it suffices to require that $\hat{x}^t\in\overline{\text{conv}}(E^t)$.

We summarize the previous discussion by presenting, in Algorithm 4, the extension of Algorithm 3 with a pruning step satisfying the above requirements, and we deduce at once the following convergence result.

Theorem 3.7. *Consider the Extended Level-Set method with pruning from Algorithm 4 under Assumptions* (H1) *to* (H6)*. The primal-dual gap Δ^t converges to* 0 *with a speed of order* $1/\sqrt{t}$.

Algorithm 4: Extended Level-Set method for Problem (D) with pruning

Require: $\mu^0 \in \operatorname{dom}(\sigma_E) \cap \operatorname{dom}(\partial\Xi) \cap \operatorname{dom}(\psi)$, $w^0 \in \partial\Xi(\mu^0)$, $E^{-1} \subset E$, $\lambda \in (0,1)$;
Set $\bar{h}^{-1} = +\infty$ and $\bar{\Delta} = +\infty$;
for $t = 0, \dots$ **do**
  Available at iteration t: μ^t, w^t, E^{t-1}, $\bar{h}^{t-1}$ as in Algorithm 3, and $\bar{\Delta} \in \mathbb{R}_+ \cup \{+\infty\}$;
  Oracle:
    Find $v^t \in \partial\sigma_E(\mu^t)$;
    Set $\tilde{E}^t = \overline{\operatorname{conv}}(E^{t-1} \cup v^t)$;
  Dual update:
    Set $\bar{h}^t$, $\underline{h}^t$, Δ^t, ℓ^t as in Algorithm 3, and take $\hat{x}^t$ as a solution to

$$\underset{x\in\mathcal{E}}{\text{minimize}}\ \psi^*(x) + \iota_{-\tilde{E}^t}(x)\ ;$$

  Pruning:
    if $\Delta^t < (1-\lambda)\bar{\Delta}$ **then**
      Take $E^t \subset \overline{\operatorname{conv}}(E^{t-1} \cup \{v^t\})$ such that $\{-\hat{x}^t\} \cup E^{-1} \subset \overline{\operatorname{conv}}(E^t)$;
      Set $\bar{\Delta} = \Delta^t$;
    else
      Set $E^t = E^{t-1} \cup \{v^t\}$;
    end
  Dual candidate:
    Take Q^t, μ^{t+1}, and w^{t+1} as in Algorithm 3;
end

4 The Dualized Level-Set method

We present in this section the general nonsmooth problem that we aim at solving with our DLS algorithm, Problem (**P**). Our approach consists in applying the ELS method from Algorithm 4 to its dual, Problem (**D**).

4.1 Framework and mathematical assumptions

Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be Hilbert spaces and $\mathcal{E}$ their product space. We aim at solving problems of the form

$$\underset{x_1\in\mathcal{E}_1}{\text{minimize}}\ f(x_1) + \sigma_Q(Ax_1 - b) + \iota_K(x_1), \tag{P}$$

where $A\colon \mathcal{E}_1 \to \mathcal{E}_2$ is a linear operator, $b \in \mathcal{E}_2$, $Q \subset \mathcal{E}_2$, and $K \subset \mathcal{E}_1$. We denote by V the value of Problem (**P**).

Structural assumptions We make the following assumptions.

(A1) The function $f\colon \mathcal{E}_1 \to \mathbb{R}$ is convex and has β-Lipschitz gradient, for some $\beta > 0$.

(A2) The set $Q \subset \mathcal{E}_2$ is nonempty and can be decomposed in $Q = Q_1 + Q_2$, where Q_1 is a closed convex bounded set and Q_2 is a closed convex cone. Let then C_{Q_1} be such that $Q_1 \subset B(0, C_{Q_1})$.

(A3) The set $K \subset \mathcal{E}_1$ is nonempty, closed, and convex.

(A4) The operator $A\colon \mathcal{E}_1 \to \mathcal{E}_2$ is linear and bounded.

(A5) There exists $\mu^0 \in \operatorname{dom}(\partial f^*) \times Q$ such that $\mu_1^0 + A^*\mu_2^0 \in \operatorname{dom}(\sigma_{-K})$.

Notice that, in this context, we have

$$\operatorname{dom}(\sigma_Q(\cdot - b)) = b + Q_2^{\ominus}, \quad \text{where } Q_2^{\ominus} = \left\{x_2 \in \mathcal{E}_2 \,\middle|\, \forall \mu_2 \in Q_2, \langle \mu_2, x_2\rangle_{\mathcal{E}_2} \le 0\right\}.$$

Moreover, Problem (**P**) is equivalent to

$$\underset{x\in\mathcal{E}}{\text{minimize}}\ f(x_1) + \sigma_Q(x_2 - b) + \iota_{\mathbf{K}}(x), \tag{26}$$

where

$$\mathbf{K} = \left\{ x \in \mathcal{E} \,\middle|\, x_1 \in K,\, x_2 = Ax_1 \right\},$$

in the sense that the value of Problem (26) is V and x is solution to Problem (26) *iff* $x_2 = Ax_1$ and x_1 is solution to Problem (**P**).

Qualification condition We require the following qualification condition.

(A6) There exists a compact set $K^{-1} \subset K$ such that

$$-b \in \operatorname{int}\big(Q_2^{\ominus} - A\,\overline{\operatorname{conv}}(K^{-1})\big). \tag{27}$$

Remark 4.1. Note that, if Q is bounded, then $Q_2 = \{0\}$, so that $Q_2^{\ominus} = \mathcal{E}_2$ and the qualification condition (27) is satisfied.

The dual problem to Problem (26) writes:

$$\underset{\mu \in \mathcal{E}}{\text{minimize}}\ f^*(\mu_1) + \iota_Q(\mu_2) + \langle \mu_2, b \rangle_{\mathcal{E}_2} + \sigma_E(\mu), \quad \text{where } E = -\mathbf{K}. \tag{**D**}$$

Lemma 4.2. *Under Assumptions* (A1) *to* (A6)*, the value of Problem (**D**) is equal to* $-V$*. Moreover, if* $V < +\infty$*, then Problem (**D**) has a solution.*

Proof. By Theorem 2.5, it suffices to verify that $0_{\mathcal{E}} \in \operatorname{int}(\operatorname{dom}(f) \times \operatorname{dom}(\sigma_Q(\cdot - b)) - \mathbf{K})$. The inclusion (27) indeed implies that

$$\begin{aligned} 0_{\mathcal{E}} \in \mathcal{E}_1 \times \operatorname{int}\big(b + Q_2^{\ominus} - A\,\overline{\operatorname{conv}}(K^{-1})\big) &\subset \mathcal{E}_1 \times \operatorname{int}\big(b + Q_2^{\ominus} - AK\big) \\ &= \operatorname{int}\big(\mathcal{E}_1 \times \big(b + Q_2^{\ominus} - AK\big)\big) \\ &= \operatorname{int}\big(\big(\mathcal{E}_1 \times \big(b + Q_2^{\ominus}\big)\big) - \mathbf{K}\big), \end{aligned}$$

as was to be proved. □

We define $\psi \colon \mathcal{E} \to \bar{\mathbb{R}}$ as, for all $\mu \in \mathcal{E}$

$$\psi(\mu) = f^*(\mu_1) + \iota_Q(\mu_2) + \langle \mu_2, b \rangle_{\mathcal{E}_2} \tag{28}$$

so that Problem (**D**) has the form of Problem (D). For future reference, we let $C_{K^{-1}} > 0$ be such that $K^{-1} \subset B(0, C_{K^{-1}})$ and we set

$$E^{-1} = \left\{ x \in \mathcal{E} \,\middle|\, x_1 \in -K^{-1}, x_2 = Ax_1 \right\}.$$

4.2 Numerical assumptions and statement of the algorithm

Before providing the statement of the DLS algorithm, we list the required numerical assumptions for its implementation.

(N1) For a given $\mu_1 \in \mathcal{E}_1$, we are able to find efficiently a solution to the problem

$$\underset{v_1 \in K}{\text{minimize}}\ \langle \mu_1, v_1 \rangle_{\mathcal{E}_1}. \tag{29}$$

In other words, we have a LMO on K at μ_1.

(N2) For a given $\mu_2 \in \mathcal{E}_2$, we can find efficiently a vector $\bar{\mu}_2 \in \underset{\nu_2 \in Q}{\operatorname{argmin}}\ \|\mu_2 - \nu_2\|_{\mathcal{E}_2}$. In other words, we can project efficiently onto Q. Notice that this also gives an easy access to $d(\mu_2, Q)$.

(N3) For any simple set K', we can solve efficiently the problem

$$\underset{x_1 \in \mathcal{E}_1}{\text{minimize}}\ f(x_1) + \sigma_Q(Ax_1 - b) + \iota_{\overline{\operatorname{conv}}(K')}(x_1), \tag{30}$$

where we say that K' is simple if it is of the form $K^{-1} \cup S$ for some finite set S.

Before we state our last numerical assumption, we specify the Bregman distance B_Ξ that will be used and state a few basic properties that it satisfies. We define $\Xi\colon \mathcal{E} \to \bar{\mathbb{R}}$ as, for all $\mu \in \mathcal{E}$,

$$\Xi(\mu) = f^*(\mu_1) + \frac{1}{2}\|\mu_2\|^2_{\mathcal{E}_2}. \tag{31}$$

Remark 4.3. Notice that, thanks to Assumption (A1), Ξ is strongly convex, and that we have, for all $\mu \in \mathcal{E}$,

$$\partial\Xi(\mu) = \partial f^*(\mu_1) \times \{\mu_2\}.$$

Thus, for all $\hat{\mu} \in \mathcal{E}$, $\mu_1 \in \mathrm{dom}(\partial f^*)$, $w_1 \in \partial f^*(\mu_1)$, and $\mu_2 \in \mathcal{E}_2$, we have

$$B_\Xi(\hat{\mu}, (\mu, (w_1, \mu_2))) = B_{f^*}(\hat{\mu}_1, (\mu_1, w_1)) + \frac{1}{2}\|\hat{\mu}_2 - \mu_2\|^2_{\mathcal{E}_2}. \tag{32}$$

Remark 4.4. The idea of using a Bregman distance derived from f^* is reminiscent from the article [2], in which a dual interpretation of the agnostic FWA is given as a mirror descent algorithm. The Bregman distance involved in the mirror descent is the one associated with the Fenchel conjugate of the cost function.

The following lemma is required for the statement of our last numerical assumption. Its proof is deferred to Section 5.

Lemma 4.5. *Let $S \subset \mathcal{E}_1$ be compact. There exists a bounded linear operator $\mathfrak{E}\colon \mathcal{M}(S) \to \mathcal{E}_1$ such that, for all $m \in \mathcal{M}(S)$, $\mathfrak{E}m$ is the unique vector verifying*

$$\forall \mu \in \mathcal{E}_1, \quad \int_S \langle \mu, v\rangle_{\mathcal{E}_1} dm(v) = \langle \mu, \mathfrak{E}m\rangle_{\mathcal{E}_1}.$$

Moreover, if $m \in \mathcal{P}(S)$, then $\mathfrak{E}m \in \overline{\mathrm{conv}}(S)$.

Remark 4.6. The vector $\mathfrak{E}m$ can be interpreted as the m-average over S. We also point out that simple sets are compact.

Our last numerical assumption is the following.

(N4) Let $S \subset \mathcal{E}_1$ be simple. We can compute efficiently a solution to the problem

$$\underset{m \in -\mathcal{M}^+(S)}{\text{minimize}}\ \tilde{f}\left(-\mathfrak{E}m + w_1^t, 1 - m(S)\right) + \frac{1}{2}\left(\|\tilde{\mu}_2\|^2_{\mathcal{E}_2} - d(\tilde{\mu}_2, Q)^2\right) - \ell^t m(S), \tag{33}$$

where $\tilde{\mu}_2 = -A\,\mathfrak{E}m + m(S)b + \mu_2^t$.

Remark 4.7. If S is a finite set $\{s_j, j \in \{1, \dots, J\}\}$, then finding $m \in \mathcal{M}(S)$ consists in finding a vector $(m_j)_{j\in\{1,\dots,J\}} \in \mathbb{R}^J$, and in this case $\mathfrak{E}m = \sum_{j=1}^J m_j s_j$.

Remark 4.8. The availability of efficient methods for the resolution of Problems (30) and (33), involved in the numerical assumptions (N3) and (N4), heavily depends on the nature of f and Q. Let us insist on the fact that they do not need the knowledge of f^*. Concerning the resolution of Problem (30), we note that, if Q is bounded and K^{-1} is finite, then the problem amounts to minimizing a Lipschitz-continuous cost function over a convex hull, which can be done with the mirror descent algorithm [4]. Problem (33) involves a smooth cost function to be minimized with simple constraints. Thus it can therefore easily be handled with the projected gradient method. Let us also note that the complexity of the two problems possibly increases along the iterations as we have no a priori upper bounds on the cardinality of K'. We will further comment on this issue in the conclusion of the article.

We are finally in position to state our DLS algorithm, its statement is provided in Algorithm 5. The main idea to obtain it is to apply Algorithm 4 to problem (**D**), which gives Algorithm 6 (where we use (28) as well as Remark 4.3 to rewrite expressions depending on ψ and Ξ in terms of f^* and Q). Suitable manipulations, which we detail below in Subsection 4.3, allow one to prove that our DLS algorithm, Algorithm 5, is an instance of Algorithm 6.

4.3 Duality between the ELS and the DLS methods

We justify in this subsection the fact that our DLS method, Algorithm 5, is an instance of the ELS method applied to the dual problem Problem (**D**), Algorithm 6.

We first note that the requirements of Algorithms 5 and 6 coincide, since $\mathrm{dom}(\sigma_E) = \{\mu \in \mathcal{E} \mid \mu_1 + A^*\mu_2 \in \mathrm{dom}(\sigma_{-K})\}$. We next study the relations between corresponding steps of Algorithms 5 and 6.

Algorithm 5: Dualized Level-Set method for Problem (26)

Require: $\mu^0 \in \mathrm{dom}(\partial f^*) \times Q$ such that $\mu_1^0 + A^*\mu_2^0 \in \mathrm{dom}(\sigma_{-K})$, $w_1^0 \in \partial f^*\big(\mu_1^0\big)$, $\lambda \in (0,1)$;
Set $\bar{h}^{-1} = +\infty$, $\hat{\mu}^{-1} = \mu^0$, and $\bar{\Delta} = +\infty$;
for $t = 0, \dots$ **do**

Available at iteration t: $\mu^t \in \mathcal{E}$, $\hat{\mu}^{t-1} \in \mathcal{E}$, $\hat{x}_1^t \in \mathcal{E}_1$, $w_1^t \in \mathcal{E}_1$, $\bar{h}^{t-1} \in \bar{\mathbb{R}}^+$, $\bar{\Delta} \in \mathbb{R}_+ \cup \{+\infty\}$, $K^{t-1} \subset \mathcal{E}_1$;

Oracle:

Find $v_1^t \in -\underset{v_1 \in K}{\mathrm{argmin}} \ \langle \mu_1^t + A^*\mu_2^t, v_1\rangle_{\mathcal{E}_1}$ and set $\tilde{K}^t = \overline{\mathrm{conv}}\big(K^{t-1} \cup \{-v_1^t\}\big)$;

Dual update:

Set $\bar{h}^t = \min\big\{\bar{h}^{t-1}, \langle \mu_2^t, b\rangle_{\mathcal{E}_2} + \langle \mu_1^t, w_1^t\rangle_{\mathcal{E}_1} + \langle \mu_1^t + A^*\mu_2^t, v_1^t\rangle_{\mathcal{E}_1} - f\left(w_1^t\right)\big\}$;

if $\bar{h}^t < \bar{h}^{t-1}$ **then**

Set $\hat{\mu}^t = \mu^t$

else

Set $\hat{\mu}^t = \hat{\mu}^{t-1}$

end

Primal update:

Take a solution $\hat{x}_1^t$ to: $\underset{x_1 \in \tilde{K}^t}{\mathrm{minimize}} \ f(x_1) + \sigma_Q(Ax_1 - b)$;

Set $\underline{h}^t = -f\left(\hat{x}_1^t\right) - \sigma_Q(A\hat{x}_1^t - b)$, $\Delta^t = \bar{h}^t - \underline{h}^t$, and $\ell^t = \lambda\bar{h}^t + (1-\lambda)\underline{h}^t$;

Pruning:

if $\Delta^t < (1-\lambda)\bar{\Delta}$ **then**

Take $K^t \subset \overline{\mathrm{conv}}\big(K^{t-1} \cup \{-v^t\}\big)$ simple and such that $\hat{x}_1^t \in \overline{\mathrm{conv}}(K^t)$;

Set $\bar{\Delta} = \Delta^t$;

else

Set $K^t = K^{t-1} \cup \{-v_1^t\}$;

end

Dual candidate:

Find $m^t \in -\mathcal{M}^+(K^t)$ solution to Problem (33) with $S = K^t$;

Set $w_1^{t+1} = \frac{w_1^t - \mathfrak{E}m^t}{1 - m^t(K^t)}$, $\mu_1^{t+1} = \nabla f\big(w_1^{t+1}\big)$, and $\mu_2^{t+1} = \mathrm{proj}_Q(\mu_2^t + m^t(K^t)b - A\,\mathfrak{E}m^t)$;

end

Algorithm 6: Extended Level-Set method for Problem (**D**)

Require: $\mu^0 \in \mathrm{dom}(\sigma_E) \cap (\mathrm{dom}(\partial f^*) \times Q)$, $w_1^0 \in \partial f^*\big(\mu_1^0\big)$, $\lambda \in (0,1)$;
Set $\bar{h}^{-1} = +\infty$ and $\bar{\Delta} = +\infty$;
for $t = 0, \dots$ **do**

Available at iteration t: $\mu^t \in \mathcal{E}$, $w^t \in \mathcal{E}$, $\bar{h}^{t-1} \in \bar{\mathbb{R}}^+$, $\bar{\Delta} \in \mathbb{R}_+ \cup \{+\infty\}$, $E^{t-1} \subset \mathcal{E}$;

Oracle:

Choose $v^t \in \partial\sigma_E(\mu^t)$ and set $\tilde{E}^t = \overline{\mathrm{conv}}\big(E^{t-1} \cup \{v^t\}\big)$;

Dual update:

Set $\bar{h}^t = \min\big\{\bar{h}^{t-1}, f^*(\mu_1^t) + \langle \mu_2^t, b\rangle_{\mathcal{E}_2} + \sigma_E(\mu^t)\big\}$;

Primal update:

Take a solution $\hat{x}^t$ to: $\underset{(x_1,x_2) \in -\tilde{E}^t}{\mathrm{minimize}} \ f(x_1) + \sigma_Q(x_2 - b)$;

Set $\underline{h}^t = \inf\limits_{\mu \in \mathcal{E}_1 \times Q} f^*(\mu_1) + \langle \mu_2, b\rangle_{\mathcal{E}_2} + \sigma_{\tilde{E}^t}(\mu)$, $\Delta^t = \bar{h}^t - \underline{h}^t$, and $\ell^t = \lambda\bar{h}^t + (1-\lambda)\underline{h}^t$;

Pruning:

As in Algorithm 4.

Dual candidate:

Set $Q^t = \{\psi + \sigma_{E^t} \le \ell^t\}$;

Take μ^{t+1} as the solution to: $\underset{\mu \in Q^t}{\mathrm{minimize}} \ B_{f^*}(\mu_1, (\mu_1^t, w_1^t)) + \frac{1}{2}\|\mu_2 - \mu_2^t\|_{\mathcal{E}_2}^2$;

Take $w_1^{t+1} \in \partial f^*\big(\mu_1^{t+1}\big)$ such that $(w_1^t - w_1^{t+1}, \mu_2^t - \mu_2^{t+1}) \in N_{Q^t}(\mu^{t+1})$;

end

Oracle From Corollary 2.4 and the definition of the adjoint operator, we get that, for all $\mu' \in \mathcal{E}$, if we set $\mu_1 = \mu_1' + A^*\mu_2'$ and let v_1 be given by the LMO on K at μ_1, then $-(v_1, Av_1) \in \partial\sigma_E(\mu')$. Thus, the LMO on K allows us to find an element of $\partial\sigma_E(\mu^t)$. Also, for all $t \in \mathbb{N}$, since A is linear and bounded, we have $\tilde{E}^t = \left\{x \in \mathcal{E} \,|\, x_1 \in -\tilde{K}^t, x_2 = Ax_1\right\}$.

Dual update The update of $\bar{h}^t$ in Algorithm 5 is justified using the **Oracle**, Lemma 2.2, and Remark 2.3.

Primal update First notice that, for all $\mu \in \mathcal{E}$, we have

$$(f \oplus \sigma_Q(\cdot - b))^*(\mu) = f^*(\mu_1) + \langle \mu_2, b\rangle + \iota_Q(\mu_2).$$

Let now $t \in \mathbb{N}$. Since $\tilde{E}^t$ is bounded, we have

$$\begin{aligned}\inf_{\mu\in\mathcal{E}_1\times Q} f^*(\mu_1) + \langle \mu_2, b\rangle_{\mathcal{E}_2} + \sigma_{\tilde{E}^t}(\mu) &= -\inf_{x\in\mathcal{E}} f(x_1) + \sigma_Q(x_2 - b) + \iota_{-\tilde{E}_t}(x)\\ &= -\inf_{x_1\in\mathcal{E}_1} f(x_1) + \sigma_Q(Ax_1 - b) + \iota_{\tilde{K}^t}(x_1).\end{aligned}$$

This justifies the updates of $\underline{h}^t$ and $\hat{x}^t$.

Pruning Notice that, for all $t \in \mathbb{N}$, the set K^t is simple, and thus compact, and that the set E^t can always be taken as

$$E^t = \left\{x \in \mathcal{E} \,|\, x_1 \in -K^t,\, x_2 = Ax_1\right\}. \tag{34}$$

Dual candidate We focus on the projection problem, that is,

$$\underset{\mu\in Q^t}{\text{minimize}}\; B_{f^*}\big(\mu_1, (\mu_1^t, w_1^t)\big) + \frac{1}{2}\|\mu_2 - \mu_2^t\|_{\mathcal{E}_2}^2. \tag{35}$$

The next proposition, whose proof is provided in Section 5, collects the properties of this problem that allow one to justify that the **Dual candidate** step of Algorithm 5 is an instance of the corresponding step of Algorithm 6.

Proposition 4.9. *Let $t \in \mathbb{N}$ and m^t be a solution to Problem (33) with $S = K^t$. The following hold:*

i) Up to a shift in its value, Problem (33) is the dual of Problem (35).

ii) Set w_1^{t+1} and μ^{t+1} as in Algorithm 5, i.e., as

$$\begin{aligned} w_1^{t+1} &= \frac{w_1^t - \mathfrak{E}m^t}{1 - m^t(K^t)}\\ \mu_1^{t+1} &= \nabla f\big(w_1^{t+1}\big)\\ \mu_2^{t+1} &= \mathrm{proj}_Q\big(\mu_2^t + m^t(K^t)b - A\,\mathfrak{E}m^t\big).\end{aligned}$$

Then:

a) The solution to Problem (35) is μ^{t+1}.

b) We have $w_1^{t+1} \in \partial f^*\big(\mu_1^{t+1}\big)$ and $(w_1^t - w_1^{t+1}, \mu_2^t - \mu_2^{t+1}) \in N_{Q^t}(\mu^{t+1})$. Moreover, if Assumption (H2) is satisfied, then we can bound w_1^t uniformly in t.

4.4 Convergence analysis

We now want to prove that, under our standing assumptions, Algorithm 5 converges. At the light of Subsection 4.3, it suffices to show that Algorithm 6 converges and, for that purpose, one is left to verify that the assumptions of the convergence theorem for Algorithm 4, Theorem 3.7, are satisfied in our setting. We assume in the sequel that Assumptions (A1) to (A6) are verified.

We start by remarking that, thanks to (28), Remark 4.3, and the definition of E in Problem (**D**), we immediately obtain the following result from Assumption (A5) and a straightforward computation.

Lemma 4.10. *Assumption (H1) is satisfied.*

We next turn to the verification of Assumption (H3).

Lemma 4.11. *Assumption* (H3) *is satisfied.*

Proof. We recall that we defined the set E^{-1} as

$$E^{-1} = \left\{ x \in \mathcal{E} \,\middle|\, x_1 \in -K^{-1},\, x_2 = Ax_1 \right\},$$

and that we want to show that E^{-1} is bounded and

$$\psi(\mu) + \sigma_{E^{-1}}(\mu) \underset{\|\mu\|_{\mathcal{E}} \to +\infty}{\longrightarrow} +\infty.$$

Boundedness of E^{-1} is immediate since K^{-1} is bounded and A is a bounded linear operator.

The qualification condition (27) implies that there exists $\varepsilon > 0$ such that for all $\mu_2 \in \mathcal{E}_2 \setminus \{0\}$ we have

$$\varepsilon \frac{\mu_2}{\|\mu_2\|_{\mathcal{E}_2}} - b \in Q_2^{\ominus} - A\,\overline{\mathrm{conv}}\big(K^{-1}\big).$$

In other words, there exists $\varepsilon > 0$ such that for all $\mu_2 \in \mathcal{E}_2 \setminus \{0\}$, there exists $x_{\mu_2} \in \overline{\mathrm{conv}}\big(K^{-1}\big)$ and $z_{\mu_2} \in Q_2^{\ominus}$ such that

$$\varepsilon \frac{\mu_2}{\|\mu_2\|_{\mathcal{E}_2}} - b = z_{\mu_2} - Ax_{\mu_2}. \tag{36}$$

Now, we fix such an $\varepsilon > 0$. Equation (36) implies that, for all $\mu_2 \in \mathcal{E}$, there exists $x_{\mu_2} \in \overline{\mathrm{conv}}\big(K^{-1}\big)$ and $z_{\mu_2} \in Q_2^{\ominus}$ such that

$$\varepsilon \|\mu_2\|_{\mathcal{E}_2} - \langle \mu_2, b \rangle_{\mathcal{E}_2} = \langle \mu_2, z_{\mu_2} \rangle_{\mathcal{E}_2} - \langle \mu_2, Ax_{\mu_2} \rangle_{\mathcal{E}_2}. \tag{37}$$

Notice that eq. (37) also holds for $\mu_2 = 0$ (with arbitrary choices of $x_{\mu_2} \in \overline{\mathrm{conv}}\big(K^{-1}\big)$ and $z_{\mu_2} \in Q_2^{\ominus}$). Let then $\mu_1 \in \mathcal{E}_1$ and $\mu_2 \in Q$. We know that

- Using the decomposition $Q = Q_1 + Q_2$, there exist $\mu_2^a \in Q_1$ and $\mu_2^b \in Q_2$ such that

$$\mu_2 = \mu_2^a + \mu_2^b. \tag{38}$$

- There exist $x_{\mu_2} \in \overline{\mathrm{conv}}\big(K^{-1}\big)$ and $z_{\mu_2} \in Q_2^{\ominus}$ satisfying eq. (37).

We have

$$\begin{aligned}
\psi(\mu) + \sigma_{E^{-1}}(\mu) &= f^*(\mu_1) + \langle \mu_2, b \rangle_{\mathcal{E}_2} + \iota_Q(\mu_2) + \sigma_{E^{-1}}(\mu) \\
&\geq f^*(\mu_1) + \langle \mu_2, b \rangle_{\mathcal{E}_2} - \langle \mu_1, x_{\mu_2} \rangle_{\mathcal{E}_1} - \langle \mu_2, Ax_{\mu_2} \rangle_{\mathcal{E}_2} && (39)\\
&\geq f^*(\mu_1) - C_{K^{-1}} \|\mu_1\|_{\mathcal{E}_1} + \varepsilon \|\mu_2\|_{\mathcal{E}_2} - \langle \mu_2, z_{\mu_2} \rangle_{\mathcal{E}_2} && (40)\\
&\geq f^*(\mu_1) - C_{K^{-1}} \|\mu_1\|_{\mathcal{E}_1} + \varepsilon \|\mu_2\|_{\mathcal{E}_2} - \langle \mu_2^a, z_{\mu_2} \rangle_{\mathcal{E}_2} && (41)\\
&\geq f^*(\mu_1) - C_{K^{-1}} \|\mu_1\|_{\mathcal{E}_1} + \varepsilon \|\mu_2\|_{\mathcal{E}_2} - C_{Q_1} \left(\varepsilon + C_{K^{-1}} \|A\| + \|b\|_{\mathcal{E}_2} \right), && (42)
\end{aligned}$$

where

- Equation (39) derives from the facts that $\mu_2 \in Q$ and $(-x_{\mu_2}, -Ax_{\mu_2}) \in \overline{\mathrm{conv}}\big(E^{-1}\big)$.
- Equation (40) is a consequence of the Cauchy–Schwarz inequality, of the fact that $\|x_{\mu_2}\|_{\mathcal{E}_2} \leq C_{K^{-1}}$, and of eq. (37).
- Equation (41) derives from the fact that $z_{\mu_2} \in Q_2^{\ominus}$, and thus $\left\langle \mu_2^b, z_{\mu_2} \right\rangle_{\mathcal{E}_2} \leq 0$ since $\mu_2^b \in Q_2$.
- Equation (42) is a consequence of the Cauchy–Schwarz inequality, of the fact that eq. (36) yields $\|z_{\mu_2}\|_{\mathcal{E}_2} \leq \varepsilon + C_{K^{-1}} \|A\| + \|b\|_{\mathcal{E}_2}$, and of the fact that $\|\mu_2^a\|_{\mathcal{E}_2} \leq C_{Q_1}$.

Finally, notice that the function $\mu_2 \in \mathcal{E}_2 \mapsto \varepsilon \|\mu_2\|_{\mathcal{E}_2}$ is coercive and lower bounded, and so is the function $\mu_1 \in \mathcal{E}_1 \mapsto f^*(\mu_1) - C\|\mu_1\|_{\mathcal{E}_1} \in \mathbb{R}$ as a consequence of Lemma 2.11. The expected result then derives from Lemma 2.10. □

We next use the inequality (42) from the proof of Lemma 4.11 in order to verify Assumption (H4).

Lemma 4.12. *Assumption* (H4) *is satisfied.*

Proof. Recall that $Q^{-1} = \{\psi + \sigma_{E^{-1}} \leq \bar{h}^0\}$. Thus, by (42), we obtain that $f^*(\mu_1)$ and $\|\mu_2\|_{\mathcal{E}_2}$ are finite for every $\mu \in Q^{-1}$, yielding that $Q^{-1} \subset \mathrm{dom}(\Xi)$. In addition, it also follows from (42) that $\mu \mapsto f^*(\mu_1) - C_{K^{-1}}\|\mu_1\|_{\mathcal{E}_1} + \varepsilon\|\mu_2\|_{\mathcal{E}_2}$ is bounded over Q^{-1}, and we conclude thanks to the definition of Ξ and the strong convexity of f^*. □

Lemma 4.13. *If Assumption (H2) is satisfied, then Assumption (H5) is satisfied.*

Proof. Recall that, for all $\mu \in \mathcal{E}$, $\partial\Xi(\mu) = \partial f^*(\mu_1) \times \{\mu_2\}$. Thus, we have $w^t = (w_1^t, \mu_2^t) \in \partial\Xi(\mu^t)$, for all $t \in \mathbb{N}$. The only thing left to show is that there exists a constant $C_{\partial\Xi} > 0$ such that, for all $t \in \mathbb{N}$, we have $\|w^t\| \leq C_{\partial\Xi}$. This derives from the fact that w_1^t is uniformly bounded in t, as stated in Proposition 4.9.*ii.b*, and that, for all $t \in \mathbb{N}$, we have $\mu^{t+1} \in Q^t \subset Q^{-1}$ and Q^{-1} is bounded. □

Lemma 4.14. *If Assumption (H2) is satisfied, then Assumption (H6) is satisfied.*

Proof. We want to show there exists $C_\psi > 0$ such that for all $t \in \mathbb{N}$ we have

$$\left|\psi\left(\mu^{t+1}\right) - \psi\left(\mu^t\right)\right| \leq C\left\|\mu^{t+1} - \mu^t\right\|_{\mathcal{E}}.$$

Let $t \in \mathbb{N}$, we have

$$\left|\psi(\mu^{t+1}) - \psi(\mu^t)\right| = \left|f^*\left(\mu_1^{t+1}\right) - f^*\left(\mu_1^t\right) + \left\langle \mu_2^{t+1} - \mu_2^t, b\right\rangle_{\mathcal{E}_2}\right| \leq \left|f^*\left(\mu_1^{t+1}\right) - f^*\left(\mu_1^t\right)\right| + \left|\left\langle \mu_2^{t+1} - \mu_2^t, b\right\rangle_{\mathcal{E}_2}\right|.$$

Recall that $w_1^{t+1} \in \partial f^*\left(\mu_1^{t+1}\right)$, and thus

$$f^*\left(\mu_1^{t+1}\right) - f^*\left(\mu_1^t\right) \leq \left\langle \mu_1^{t+1} - \mu_1^t, w_1^{t+1}\right\rangle_{\mathcal{E}} \leq \left\|w_1^{t+1}\right\|_{\mathcal{E}_1}\left\|\mu_1^{t+1} - \mu_1^t\right\|_{\mathcal{E}_1}.$$

Likewise, we have

$$f^*\left(\mu_1^t\right) - f^*\left(\mu_1^{t+1}\right) \leq \left\|w_1^t\right\|_{\mathcal{E}_1}\left\|\mu_1^t - \mu_1^{t+1}\right\|_{\mathcal{E}_1}.$$

This yields

$$\left|\psi\left(\mu^{t+1}\right) - \psi\left(\mu^t\right)\right| \leq \left(\max\left\{\left\|w_1^{t+1}\right\|_{\mathcal{E}_1}, \left\|w_1^t\right\|_{\mathcal{E}_1}\right\} + \|b\|_{\mathcal{E}_2}\right)\left\|\mu^{t+1} - \mu^t\right\|_{\mathcal{E}},$$

and the conclusion follows since w_1^t in uniformly bounded in t, as stated in Proposition 4.9.*ii.b*. □

Gathering Lemmas 4.10 to 4.14 and combining with the discussion of Subsection 4.3, we obtain at once the following result.

Theorem 4.15. *Consider the Dualized Level-Set method from Algorithm 5 under Assumption (H2) and Assumptions (A1) to (A6). Then the primal-dual gap Δ^t converges to 0 with a speed of order $1/\sqrt{t}$. Also, we have, for all $t \in \mathbb{N}$, $\hat{x}_2^t = A\hat{x}_1^t$, and we have the same convergence speed for $f\left(\hat{x}_1^t\right) + \sigma_Q(A\hat{x}_1^t - b) - V$ and for $f^*(\hat{\mu}_1) + \iota_Q(\hat{\mu}_2) + \langle\hat{\mu}_2, b\rangle_{\mathcal{E}_2} + \sigma_E(\hat{\mu}) + V$.*

Remark 4.16.
- We have kept Assumption (H2) in its non-explicit formulation on purpose. It is of course satisfied if K is bounded, since then K is weakly compact and thus Problem (29) has a solution for any μ_1.
- In the general context of the ELS algorithm, it is actually sufficient to require the existence of a constant $C_{\text{oracle}} > 0$ such that for any $\mu \in Q^{-1}$, the linear minimization oracle has a solution in $B(0, C_{\text{oracle}})$. In the more specific context of the DLS algorithm, we have $Q^{-1} \subset \mathrm{dom}(f^*) \times Q$. Moreover, Q^{-1} is a bounded set. Therefore, Assumptions (H2) and (N1) can be replaced by the following one:

  (A0) For any $R > 0$, there exists C_{oracle} such that for any $(\mu_1', \mu_2') \in (\mathrm{dom}(f^*) \times Q) \cap B(0, R)$, Problem (29) has a solution $v \in B(0, C_{\text{oracle}})$, when called with $\mu_1 = \mu_1' + A^*\mu_2'$.

4.5 Extension of the Generalized Conditional Gradient

A now classical extension of the Frank–Wolfe algorithm, called generalized Frank–Wolfe (or generalized conditional gradient method, see [14]), consists in linearizing only partially the cost function. The contribution of the cost function which is not linearized remains then in the oracle and replaces the characteristic function of the feasible set.

In this subsection, we show that our algorithm can handle a situation of this kind, thanks to a natural augmentation of the problem through a slack variable. Let $\mathcal{E}_{1a}$ and $\mathcal{E}_2$ be Hilbert spaces. We aim at solving the following generalization of Problem (**P**):

$$\underset{x_{1a} \in \mathcal{E}_{1a}}{\text{minimize}}\ f_a\left(x_{1a}\right) + \sigma_Q(A_a x_{1a} - b) + h\left(x_{1a}\right). \tag{43}$$

Structural assumptions We make the same assumptions on f_a, Q, and A_a as we made on f, Q, and A in Section 4.1, and we assume that $h \in \Gamma_0(\mathcal{E}_{1a})$ and is supercoercive, i.e., $\lim_{\|x\|\to+\infty} h(x)/\|x\| = +\infty$.

Qualification condition We assume that we can find $K_a^{-1} \subset \mathrm{dom}(h)$ compact such that h is bounded over $\overline{\mathrm{conv}}\big(K_a^{-1}\big)$ and $-b \in \mathrm{int}\big(Q_2^{\ominus} - A_a\,\overline{\mathrm{conv}}\big(K_a^{-1}\big)\big)$.

Oracle We assume that we have the following oracle: for all $\mu_{1a} \in \mathcal{E}_{1a}$, we can find a solution to the problem

$$\underset{v_{1a}\in\mathcal{E}_{1a}}{\text{minimize}}\ \langle \mu_{1a}, v_{1a}\rangle_{\mathcal{E}_{1a}} + h\left(v_{1a}\right).$$

Remark 4.17. This is indeed a more general framework, since we can take $h = \iota_{K_a}$, in which case we make exactly the same **Structural assumptions** and make the same **Qualification condition** as in Section 4.1, and the **Oracle** is a LMO on K_a.

Under these assumptions, we can rewrite our problem in the framework of Section 4.1. For this, we take $\mathcal{E}_1 = \mathcal{E}_{1a} \times \mathbb{R}$, the function f defined as, for all $x_1 = (x_{1a}, x_{1b}) \in \mathcal{E}_1$

$$f(x_1) = f_a(x_{1a}) + x_{1b},$$

which is indeed convex with gradient β-Lipschitz continuous, the bounded linear operator A defined as, for all $x_1 \in \mathcal{E}_1$

$$A\left(x_{1a}, x_{1b}\right) = A_a x_{1a},$$

and the closed convex set $K = \mathrm{epi}(h)$. In this context, f^* is given by, for all $\mu_1 = (\mu_{1a}, \mu_{1b}) \in \mathcal{E}_1$

$$f^*(\mu_{1a}, \mu_{1b}) = {f_a}^*(\mu_{1a}) + \iota_{\{1\}}(\mu_{1b})$$

and the adjoint operator of A is the operator A^* given, for all $\mu_2 \in \mathcal{E}_2$, by $A^*\mu_2 = (A_a^*\,\mu_2, 0)$.

Also, this new problem verifies the **Qualification condition** introduced in Section 4.1, with $K^{-1} = K_a^{-1} \times \{M\}$, where M denotes an upper bound of h over K_a^{-1}. Clearly $K^{-1} \subset K$ is compact. Since $A\,\overline{\mathrm{conv}}\big(K^{-1}\big) = A_a\,\overline{\mathrm{conv}}\big(K_a^{-1}\big)$, we have $-b \in \mathrm{int}\big(Q_2^{\ominus} - A\,\overline{\mathrm{conv}}\big(K^{-1}\big)\big)$ and the problem is indeed qualified.

We now need to verify that we indeed have an **Oracle** for the rewritten problem. As was explained in Remark 4.16, it is sufficient to verify Assumption (A0). We fix an arbitrary constant $R > 0$ and take $(\mu_1', \mu_2') \in (\mathrm{dom}(f^*) \times Q) \cap B(0, R)$. Therefore $\mu_1' = (\mu_{1a}', 1)$, with $\mu_{1a} \in \mathrm{dom}({f_a}^*)$. Let $\mu_{1a} = \mu_{1a}' + A_a^*\mu_2'$. We have

$$\begin{aligned}\bar{v}_1 \in \underset{v_1\in K}{\mathrm{argmin}}\ \langle \mu_1' + A^*\mu_2', v_1\rangle_{\mathcal{E}_1} &\Leftrightarrow \bar{v}_1 \in \underset{(v_{1a},v_{1b})\in\mathrm{epi}(h)}{\mathrm{argmin}}\ \langle \mu_{1a}' + A_a^*\,\mu_2', v_{1a}\rangle_{\mathcal{E}_{1a}} + v_{1b}\\ &\Leftrightarrow \bar{v}_{1b} = h\left(\bar{v}_{1a}\right) \text{ and } \bar{v}_{1a} \in \underset{v_{1a}\in\mathcal{E}_{1a}}{\mathrm{argmin}}\ \langle \mu_{1a}, v_{1a}\rangle_{\mathcal{E}_{1a}} + h\left(v_{1a}\right).\end{aligned}$$

Then $\bar{v}_{1a}$ is given by our **Oracle**. As a direct consequence of Lemma 2.2 the above statements are equivalent to: $\bar{v}_{1a} \in \partial h^*(-\mu_{1a})$ and $\bar{v}_{1b} = \langle -\mu_{1a}, \bar{v}_{1a}\rangle_{\mathcal{E}_{1a}} - h^*(-\mu_{1a})$. Since we have a bound on (μ_1', μ_2'), we also have one on $-\mu_{1a}$. Applying [3, Propositions 14.15(ii) and 16.17(iii)], we deduce that $|h^*(-\mu_{1a})|$ and $\|\bar{v}_{1a}\|_{\mathcal{E}_{1a}}$ are bounded by some constant independent of (μ_1', μ_2'), depending only on R. Then $|\bar{v}_{1b}|$ is bounded (in the same sense). This concludes the verification of Assumption (A0).

5 Technical proofs

In this section, we provide the proofs of Lemma 4.5 and Proposition 4.9. Both these results concern the dualization of the projection problem with respect to the Bregman distance associated with the function Ξ from (31), Problem (35). Recall that, for $t \in \mathbb{N}$, Problem (35) is

$$\underset{\mu\in Q^t}{\text{minimize}}\ B_{f^*}\left(\mu_1, \left(\mu_1^t, w_1^t\right)\right) + \frac{1}{2}\left\|\mu_2 - \mu_2^t\right\|_{\mathcal{E}_2}^2, \tag{35}$$

and we denote here its value by V^t.

We want to write a problem equivalent to Problem (35) which fits in the framework of the Fenchel–Rockafellar duality, i.e., which is of the form

$$\underset{X\in\mathcal{X}_1}{\text{minimize}}\ F(X) + G(LX) \tag{44}$$

where $L\colon \mathcal{X}_1 \to \mathcal{X}_2$ is a bounded linear operator, $\mathcal{X}_1$ and $\mathcal{X}_2$ are Banach spaces, $F \in \Gamma_0(\mathcal{X}_1)$, and $G \in \Gamma_0(\mathcal{X}_2)$.

We recall that B_{f^*} is defined, for all $\mu_1 \in \mathcal{E}_1$, $\mu_1' \in \operatorname{dom}(\partial f^*)$, and $w_1 \in \partial f^*(\mu_1')$ as

$$B_{f^*}(\mu_1, (\mu_1', w_1)) = f^*(\mu_1) - f^*(\mu_1') - \langle \mu_1 - \mu_1', w_1 \rangle_{\mathcal{E}_1}$$

and that, for all $\mu \in \mathcal{E}$, we have

$$\mu \in Q^t \Leftrightarrow \begin{cases} \mu_2 \in Q, \\ \forall v \in E^t, f^*(\mu_1) + \langle \mu_2, b \rangle_{\mathcal{E}_2} + \langle \mu, v \rangle_{\mathcal{E}} - \ell^t \leq 0. \end{cases}$$

Thus, Problem (35) is equivalent to

$$\begin{aligned} &\underset{\mu \in \mathcal{E}_1 \times Q}{\text{minimize}}\ f^*(\mu_1) - \langle \mu_1, w_1^t \rangle_{\mathcal{E}_1} + \frac{1}{2}\|\mu_2\|_{\mathcal{E}_2}^2 - \langle \mu_2, \mu_2^t \rangle_{\mathcal{E}_2} \\ &\text{s.t. } \forall v \in E^t, f^*(\mu_1) + \langle \mu_2, b \rangle_{\mathcal{E}_2} + \langle \mu, v \rangle_{\mathcal{E}} - \ell^t \leq 0, \end{aligned} \tag{45}$$

in the sense that the value of Problem (45) is $V^t - f^*(\mu^t) + \langle \mu_1^t, w_1^t \rangle_{\mathcal{E}_1} - \frac{1}{2}\|\mu_2^t\|_{\mathcal{E}_2}^2$ and that Problems (35) and (45) have the same solution. Since $v \in E^t$ *iff* $v_1 \in -K^t$ and $v_2 = Av_1$, this last problem is itself equivalent to

$$\begin{aligned} &\underset{(\mu, z) \in (\mathcal{E}_1 \times Q) \times \mathbb{R}}{\text{minimize}}\ z - \langle \mu_1, w_1^t \rangle_{\mathcal{E}_1} + \frac{1}{2}\|\mu_2\|_{\mathcal{E}_2}^2 - \langle \mu_2, \mu_2^t \rangle_{\mathcal{E}_2} + \iota_{\operatorname{epi}(f^*)}(\mu_1, z) \\ &\text{s.t. } \forall v_1 \in -K^t, z + \langle \mu_2, b \rangle_{\mathcal{E}_2} + \langle \mu_1 + A^*\mu_2, v_1 \rangle_{\mathcal{E}_1} - \ell^t \leq 0, \end{aligned} \tag{46}$$

in the sense that they have same value, and that (μ, z) is the solution to Problem (46) *iff* $z = f^*(\mu_1)$ and μ is the solution to Problem (45).

Let us now show that Problem (46) has the expected shape. Let $L^t\colon \mathcal{E} \times \mathbb{R} \to \mathcal{C}(K^t)$ be the bounded linear operator defined, for $(\mu, z) \in \mathcal{E} \times \mathbb{R}$, by

$$L^t(\mu, z) = z + \langle \mu_2, b \rangle_{\mathcal{E}_2} - \langle \mu_1 + A^*\mu_2, \cdot\, \rangle_{\mathcal{E}_1}.$$

For any (μ, z), $L^t(\mu, z)$ is indeed a continuous and bounded function, since it is affine and defined on a bounded set. Clearly L^t is a linear operator, it is easy to verify that it is bounded. Next, given $(\mu, z) \in (\mathcal{E}_1 \times Q) \times \mathbb{R}$, we see that the constraint in Problem (46) is satisfied *iff* $L^t(\mu, z) - \ell^t \in \mathcal{C}(K^t; \mathbb{R}_-)$. From this last point, we define $G^t\colon \mathcal{C}(K^t) \to \bar{\mathbb{R}}_+$ as, for all $\phi \in \mathcal{C}(K^t)$,

$$G^t(\phi) = \iota_{\mathcal{C}(K^t; \mathbb{R}_-)}\big(\phi - \ell^t\big)$$

and $F^t\colon \mathcal{E} \times \mathbb{R} \to \bar{\mathbb{R}}$ as, for all $\mu \in \mathcal{E}$ and $z \in \mathbb{R}$,

$$F^t(\mu, z) = z - \langle \mu_1, w_1^t \rangle_{\mathcal{E}_1} + \iota_{\operatorname{epi}(f^*)}(\mu_1, z) + \frac{1}{2}\|\mu_2\|_{\mathcal{E}_2}^2 - \langle \mu_2, \mu_2^t \rangle_{\mathcal{E}_2} + \iota_Q(\mu_2), \tag{47}$$

and we notice that $F^t \in \Gamma_0(\mathcal{E} \times \mathbb{R})$ and $G^t \in \Gamma_0(\mathcal{C}(K^t))$. Finally, Problem (46) reads

$$\underset{(\mu, z) \in \mathcal{E} \times \mathbb{R}}{\text{minimize}}\ F^t(\mu, z) + G^t(L^t(\mu, z)), \tag{48}$$

which is under the form (44), as required. The dual problem to Problem (48) is then

$$\underset{m \in \mathcal{C}(K^t)^*}{\text{minimize}}\ F^{t^*}\Big(L^{t^*}m\Big) + G^{t^*}(-m) \tag{49}$$

where $L^{t^*}\colon \mathcal{C}(K^t)^* \to \mathcal{E} \times \mathbb{R}$ is the dual operator of L^t.

The aim of what follows is to provide a more explicit expression of Problem (49). We start by identifying, in the following lemma, the set $\mathcal{C}(K^t)^*$ over which we minimize as a set of measures. Such an identification is immediate since K^t is compact.

Lemma 5.1. *We have $\mathcal{C}(K^t) = \mathcal{C}_b(K^t)$ and $\mathcal{C}(K^t)^* = \mathcal{M}(K^t)$. Moreover, $\mathcal{M}(K^t)$ is endowed with the total variation norm $\|m\| = |m|(K^t)$.*

Let us now prove Lemma 4.5.

Proof of Lemma 4.5. Let S be a compact subset of $\mathcal{E}$. Let us first show that there exists a bounded linear operator $\mathfrak{E}\colon \mathcal{M}(S) \to \mathcal{E}$ such that, for all $m \in \mathcal{M}(S)$, $\mathfrak{E}m$ is the unique vector verifying

$$\forall \mu \in \mathcal{E},\ \langle \mu, \mathfrak{E}m\rangle_{\mathcal{E}} = \int_S \langle \mu, v\rangle_{\mathcal{E}} dm(v).$$

First, we show that for all $m \in \mathcal{M}(S)$, there exists a unique vector v_m such that, for all $\mu \in \mathcal{E}$, we have

$$\langle \mu, v_m\rangle_{\mathcal{E}} = \int_S \langle \mu, v\rangle_{\mathcal{E}}\, dm(v).$$

Since S is compact, there exists $C_S \in \mathbb{R}$ such that $S \subset B(0, C_S)$. Let $m \in \mathcal{M}(S)$. We define $I_m\colon \mathcal{E} \to \bar{\mathbb{R}}$ as, for all $\mu \in \mathcal{E}$,

$$I_m(\mu) = \int_S \langle \mu, v\rangle_{\mathcal{E}}\, dm(v).$$

Since for all $\mu \in \mathcal{E}$, $|I_m(\mu)| \leq C_S\|m\|\|\mu\|_{\mathcal{E}}$ and for all $v \in \mathcal{E}$, $\mu \mapsto \langle \mu, v\rangle_{\mathcal{E}}$ is linear, then I_m is linear and continuous. Thus, using Riesz's representation theorem, there exists a unique $v_m \in \mathcal{E}$ such that for all $\mu \in \mathcal{E}$, $I_m(\mu) = \langle \mu, v_m\rangle_{\mathcal{E}}$. Furthermore, $m \in \mathcal{M}(S) \mapsto v_m$ is also linear. Let then $\mathfrak{E}\colon \mathcal{M}(S) \to \mathcal{E}$ be defined as, for all $m \in \mathcal{M}(S)$, $\mathfrak{E}m = v_m$. Then $\|\mathfrak{E}m\|_{\mathcal{E}} \leq C_S\|m\|$ and thus $\mathfrak{E}$ is continuous.

It remains to show that, for a given $m \in \mathcal{P}(S)$, we have $\mathfrak{E}m \in \overline{\mathrm{conv}}(S)$. It suffices for this to show that $\iota_{\overline{\mathrm{conv}}(S)}(v_m) \leq 0$. We have

$$\iota_{\overline{\mathrm{conv}}(S)}(v_m) = {\sigma_S}^*(v_m) = \sup_{\mu\in\mathcal{E}} \langle \mu, v_m\rangle_{\mathcal{E}} - \sigma_S(\mu).$$

For $\mu \in \mathcal{E}$, we have

$$\langle \mu, v_m\rangle_{\mathcal{E}} - \sigma_S(\mu) = \int_S \left(\langle \mu, v\rangle_{\mathcal{E}} - \sigma_S(\mu)\right) dm(v)$$

by definition of v_m and using the linearity of the integral and the fact that $m(S) = 1$. Moreover, for all $v \in S$, we have

$$\langle \mu, v\rangle_{\mathcal{E}} - \sigma_S(\mu) \leq \iota_S(v) = 0$$

using inequality (3). Thus, by positivity of the integral, we have $\langle \mu, v_m\rangle_{\mathcal{E}} - \sigma_S(\mu) \leq 0$. Since this holds for any $\mu \in \mathcal{E}$, we have $\iota_{\overline{\mathrm{conv}}(S)}(v_m) \leq 0$, which is the required result. □

We next turn to the question of providing explicit expressions for the functions F^{t^*} and G^{t^*} and the linear operator L^{t^*} appearing in Problem (49).

Lemma 5.2. *The functions F^{t^*} and G^{t^*} and the linear operator L^{t^*} are as follows:*

i) *The function $F^{t^*}\colon \mathcal{E} \times \mathbb{R} \to \bar{\mathbb{R}}^+$ is given, for all $(x, s) \in \mathcal{E} \times \mathbb{R}$, by*

$$F^{t^*}(x, s) = \tilde{f}(x_1 + w_1^t, 1 - s) + \frac{1}{2}\left(\left\|x_2 + \mu_2^t\right\|_{\mathcal{E}_2}^2 - d\big(x_2 + \mu_2^t, Q\big)^2\right).$$

ii) *The function $G^{t^*}\colon \mathcal{M}(K^t) \to \bar{\mathbb{R}}^+$ is given, for all $m \in \mathcal{M}(K^t)$, by*

$$G^{t^*}(m) = \ell^t m(K^t) + \iota_{\mathcal{M}^+(K^t)}(m).$$

iii) *The operator L^{t^*} is given, for all $m \in \mathcal{M}(K^t)$, by*

$$L^{t^*}m = \big(\big(-\mathfrak{E}m, -A\mathfrak{E}m + m(K^t)b\big), m(K^t)\big)$$

Proof. *i)* Let $x \in \mathcal{E}$ and $s \in \mathbb{R}$. We have

$$\begin{aligned} F^{t^*}(x, s) &= \sup_{(\mu,z)\in\mathcal{E}\times\mathbb{R}} \langle(\mu, z), (x, s)\rangle - F^t(\mu, s) \\ &= \sup_{(\mu_1,z)\in\mathrm{epi}(f^*)} \left\langle \mu_1, x_1 + w_1^t\right\rangle_{\mathcal{E}_1} + z(s-1) + \sup_{\mu_2\in Q} \left\langle \mu_2, x_2 + \mu_2^t\right\rangle_{\mathcal{E}_2} - \frac{1}{2}\|\mu_2\|_{\mathcal{E}_2}^2. \end{aligned}$$

By Lemma 2.12, it holds that

$$\sup_{(\mu_1,z)\in\mathrm{epi}(f^*)} \left\langle \mu_1, x_1 + w_1^t\right\rangle_{\mathcal{E}_1} + z(s-1) = \tilde{f}\left(x_1 + w_1^t, 1 - s\right).$$

Then, using [3, Example 13.5], we have

$$\sup_{\mu_2\in Q} \left\langle \mu_2, x_2 + \mu_2^t\right\rangle_{\mathcal{E}_2} - \frac{1}{2}\|\mu_2\|_{\mathcal{E}_2}^2 = \frac{1}{2}\left(\left\|x_2 + \mu_2^t\right\|_{\mathcal{E}_2}^2 - d\big(x_2 + \mu_2^t, Q\big)^2\right).$$

ii) Let $m \in \mathcal{M}(K^t)$. We have

$$G^{t^*}(m) = \sup_{\phi \in \mathcal{C}(K^t)} \int_{K^t} \phi \, dm - \iota_{\mathcal{C}(K^t;\mathbb{R}_-)}\big(\phi - \ell^t\big) = \sup_{\phi \in \mathcal{C}(K^t;\mathbb{R}_-)} \int_{K^t} (\phi + \ell^t) \, dm$$
$$= \ell^t m(K^t) + \sup_{\phi \in \mathcal{C}(K^t;\mathbb{R}_-)} \int_{K^t} \phi \, dm \;=\; \ell^t m(K^t) + \iota_{\mathcal{M}^+(K^t)}(m).$$

iii) The operator L^{t^*} is characterized by the relation

$$\forall \mu \in \mathcal{E}, \forall z \in \mathbb{R}, \forall m \in \mathcal{M}\big(K^t\big), \int_{K^t} L^t(\mu, z)(v) \, dm(v) = \Big\langle (\mu, z), L^{t^*} m \Big\rangle.$$

Using Lemma 4.5, we have, for all $\mu \in \mathcal{E}$, $z \in \mathbb{R}$, and $m \in \mathcal{M}(K^t)$

$$\int_{K^t} L^t(\mu, z)(v) \, dm(v) = z \, m(K^t) + \big\langle \mu_2, m(K^t) b \big\rangle_{\mathcal{E}_2} - \int_{K^t} \langle \mu_1 + A^* \mu_2, v \rangle_{\mathcal{E}_1} \, dm(v)$$
$$= \big\langle (\mu_1, \mu_2, z), (-\mathfrak{E} m, -A\mathfrak{E} m + m(K^t) b, m(K^t)) \big\rangle.$$

This concludes the proof. □

Now that we have computed F^{t^*} and G^{t^*}, we compute in the next lemma their subgradients.

Lemma 5.3. *i) The function F^{t^*} is Fréchet-differentiable over $\mathcal{E} \times \mathbb{R}_-$ with continuous gradient, and we have, for all $(x, s) \in \mathcal{E} \times \mathbb{R}_-$,*

$$\nabla F^{t^*}(x, s) = \Big(\nabla f(y_1), \operatorname{proj}_Q\big(x_2 + \mu_2^t\big), -f(y_1) + \langle \nabla f(y_1), y_1 \rangle_{\mathcal{E}_1} \Big), \quad \text{where } y_1 = \frac{x_1 + w_1^t}{1 - s}.$$

ii) For all $m \in \mathcal{M}^+(K^t)$

$$\partial G^{t^*}(m) = \ell^t \mathbb{1}_{K^t} + \left\{ \phi \in \mathcal{C}\big(K^t; \mathbb{R}_-\big) \,\middle|\, \int_{K^t} \phi \, dm = 0 \right\},$$

where $\mathbb{1}_{K^t} \in \mathcal{C}(K^t)$ is the function constantly equal to 1 in K^t, and, for all $m \in \mathcal{M}(K^t) \setminus \mathcal{M}^+(K^t)$, $\partial G^{t^*}(m) = \varnothing$.

Proof. *i)* The conclusion follows from the following facts:

- The function $\tilde{f}$ is differentiable with continuous gradient over $\mathcal{E} \times \mathbb{R}_+^*$, with its formula given in (9).
- If $s \leq 0$, then $1 - s > 0$.
- Using [3, Corollary 12.30], we have $\nabla d(\,\cdot\,, Q)^2 = 2(\mathrm{Id} - \operatorname{proj}_Q)$, which is continuous.

ii) We denote by $\psi^t \colon \mathcal{M}(K^t) \to \mathbb{R}$ the function defined for all $m \in \mathcal{M}(K^t)$ as $\psi^t(m) = \ell^t m(K^t)$, so that $G^{t^*} = \psi^t + \iota_{\mathcal{M}^+(K^t)}$. Notice that ψ^t is linear, and thus convex, and continuous, and that $\iota_{\mathcal{M}^+(K^t)} \in \Gamma_0(\mathcal{M}(K^t))$. Thus, we have, for all $m \in \mathcal{M}(K^t)$,

$$\partial G^{t^*}(m) = \partial \psi^t(m) + \partial \iota_{\mathcal{M}^+(K^t)}(m).$$

Clearly, this implies that, for all $m \in \mathcal{M}(K^t) \setminus \mathcal{M}^+(K^t)$, we have $\partial G^{t^*}(m) = \varnothing$. Now, let $m \in \mathcal{M}^+(K^t)$. We have

$$\psi^t(m) = \int_{K^t} \ell^t \mathbb{1}_{K^t} \, dm$$

and thus, $\partial \psi^t(m) = \{\ell^t \mathbb{1}_{K^t}\}$.

We are thus left to compute $\partial \iota_{\mathcal{M}^+(K^t)}(m)$ for $m \in \mathcal{M}^+(K^t)$. By definition, we have

$$\partial \iota_{\mathcal{M}^+(K^t)}(m) = \left\{ \phi \in \mathcal{C}\big(K^t\big) \,\middle|\, \forall \bar{m} \in \mathcal{M}^+\big(K^t\big), \int_{K^t} \phi(d\bar{m} - dm) \leq 0 \right\}.$$

Let $\phi \in \mathcal{C}(K^t)$. We consider three cases.

a) There exists $\bar{v}_1 \in K^t$ such that $\phi(\bar{v}_1) > 0$. To show that $\phi \notin \partial\iota_{\mathcal{M}^+(K^t)}(m)$, we have to find $\bar{m} \in \mathcal{M}^+(K^t)$ such that $\int_{K^t} \phi(d\bar{m} - dm) > 0$. We set $\bar{m} = m + \delta_{\bar{v}_1}$. Then clearly $\bar{m} \in \mathcal{M}^+(K^t)$, and we have

$$\int_{K^t} \phi(d\bar{m} - dm) = \phi(\bar{v}) > 0.$$

And thus, $\phi \notin \partial\iota_{\mathcal{M}^+(K^t)}(m)$.

b) We have $\phi \in \mathcal{C}(K^t; \mathbb{R}_-)$ and $\int_{K^t} \phi\, dm < 0$. Set $\bar{m} = \frac{1}{2}m$. Then $\bar{m} \in \mathcal{M}^+(K^t)$ and

$$\int_{K^t} \phi(d\bar{m} - dm) = -\frac{1}{2}\int_{K^t} \phi\, dm > 0.$$

Thus, $\phi \notin \partial\iota_{\mathcal{M}^+(K^t)}(m)$.

c) We have $\phi \in \mathcal{C}(K^t; \mathbb{R}_-)$ and $\int_{K^t} \phi\, dm = 0$. Let $\bar{m} \in \mathcal{M}^+(K^t)$. Then

$$\int_{K^t} \phi(d\bar{m} - dm) = \int_{K^t} \phi\, d\bar{m} \leq 0.$$

Thus, $\phi \in \partial\iota_{\mathcal{M}^+(K^t)}(m)$.

Since those three cases make a partition of $\mathcal{C}(K^t)$, the result follows. □

Our next result proves that there is strong duality between between Problems (48) and (49) and provides optimality conditions for these problems.

Proposition 5.4. *There exists $(\bar{\mu}, \bar{z}) \in \mathcal{E} \times \mathbb{R}$ such that $F^t(\bar{\mu}, \bar{z}) < +\infty$ and G^t is continuous at $L^t(\bar{\mu}, \bar{z})$. Thus, Problems (48) and (49) have opposite values and, for all $(\mu, z, m) \in \mathcal{E} \times \mathbb{R} \times \mathcal{M}(K^t)$, the following are equivalent:*

i) *(μ, z) is solution to Problem (48) and m is solution to Problem (49).*

ii) *$L^{t*}m \in \partial F^t(\mu, z)$ and $-m \in \partial G^t(L^t(\mu, z))$.*

iii) *$(\mu, z) = \nabla F^{t*}\big(L^{t*}m\big)$ and $L^t(\mu, z) \in \partial G^{t*}(-m)$.*

Proof. Notice that $\psi + \sigma_{E^t} \in \Gamma_0(\mathcal{E})$ and it is coercive since it is lower bounded by $\psi + \sigma_{E^t}$, which is coercive thanks to Lemma 4.11. Let then $\bar{\mu} \in \mathcal{E}$ be a solution to

$$\underset{\mu \in \mathcal{E}}{\text{minimize}}\ \psi(\mu) + \sigma_{E^t}(\mu)$$

and set $\alpha = \psi(\bar{\mu}) + \sigma_{E^t}(\bar{\mu})$. Note that, since $E^t \subset \tilde{E}^t$ (where $\tilde{E}^t$ is the set defined in Algorithm 6), we deduce that $\alpha \leq \underline{h}^t$ (where we use once again the notations of Algorithm 6).

It follows from (28) that $f^*(\bar{\mu}_1) < +\infty$ and $\bar{\mu}_2 \in Q$, and we deduce from (47) that $F^t(\bar{\mu}, f^*(\bar{\mu}_1)) < +\infty$. Notice also that, for every $v \in K^t$, we have

$$L^t(\bar{\mu}, f^*(\bar{\mu}_1))(v) = \psi(\bar{\mu}) - \langle \bar{\mu}, (v, Av)\rangle_{\mathcal{E}_2} \leq \psi(\bar{\mu}) + \sigma_{E^t}(\bar{\mu}) = \alpha \leq \underline{h}^t < \ell^t,$$

and thus G^t is continuous at $L^t(\bar{\mu}, f^*(\bar{\mu}_1))$.

We thus obtain that $0 \in \operatorname{int}(L^t\operatorname{dom}(F^t) - \operatorname{dom}(G^t))$, and the other conclusions of the proof follow from Theorem 2.5, Corollary 2.6, and the fact that $1 - m(K^t) > 0$. □

The aim of our next results is to show that, if w^{t+1} and μ^{t+1} are defined as in Algorithm 5, then they necessarily satisfy the properties required for the corresponding elements in Algorithm 6. We start with the following property of w^{t+1}.

Lemma 5.5. *Let m^t be a solution to Problem (49) and set w^{t+1} as in Algorithm 5, i.e., as*

$$w^{t+1} = \frac{w^t - \mathfrak{E}m^t}{1 - m^t(K^t)}.$$

Then $w^{t+1} \in \overline{\operatorname{conv}}(\{w^t\} \cup K^t)$.

Proof. We recall that $m^t \in -\mathcal{M}^+(K^t)$. There are two cases. If $m^t(K^t) = 0$, then $w^{t+1} = w^t$. Otherwise, if $m^t(K^t) < 0$ we set $\tilde{m}^t = \frac{m^t}{m^t(K^t)} \in \mathcal{P}(K^t)$ so that $\mathfrak{E}\tilde{m}^t \in \overline{\mathrm{conv}}(K^t)$. Then, using the linearity of $\mathfrak{E}$, we have

$$w^{t+1} = \frac{1}{1 - m^t(K^t)} w^t - \frac{m^t(K^t)}{1 - m^t(K^t)} \mathfrak{E}\tilde{m}^t.$$

The result follows. □

Note that, if Assumption (H2) is satisfied, then the sets K^t are uniformly bounded in t. Hence, as an immediate consequence of Lemma 5.5, we obtain the following result.

Corollary 5.6. *Assume that Assumption (H2) is satisfied. Then we can bound w^t uniformly in t.*

We next verify that the choices of w_1^{t+1} and μ^{t+1} in Algorithm 5 satisfy the required conditions from Algorithm 6.

Lemma 5.7. *Let $t \in \mathbb{N}$ and consider the elements μ^t, μ^{t+1}, w_1^t, and w_1^{t+1} defined as in Algorithm 5. Define Q^t as in Algorithm 6. Then*

$$Q^t = \{\mu \in \mathcal{E}_1 \times Q \mid f^*(\mu_1) + \langle \mu_2, b\rangle_{\mathcal{E}_2} - \langle \mu_1 + A^*\mu_2, v_1\rangle_{\mathcal{E}_1} \leq \ell^t \text{ for all } v_1 \in \overline{\mathrm{conv}}(K^t)\}, \tag{50}$$

μ^{t+1} is the solution to Problem (35), and

$$(w_1^t - w_1^{t+1}, \mu_2^t - \mu_2^{t+1}) \in N_{Q^t}(\mu^{t+1}). \tag{51}$$

Proof. First, Equation (50) derives from the definition of Q^t in Algorithm 6 and from eq. (34).

To prove the second statement, note that, thanks to Assumption (A1), f^* is strongly convex, and thus Problem (35) admits a unique solution $\boldsymbol{\mu}$. Due to the discussion at the beginning of the section, $\boldsymbol{\mu}$ is solution to Problem (35) if and only if $(\boldsymbol{\mu}, f^*(\boldsymbol{\mu}_1))$ is solution to Problem (48).

On the other hand, the element m^t from Algorithm 5 is solution to Problem (33) with $S = K^t$, which, thanks to Lemmas 5.1 and 5.2, is equivalent to m^t being a solution to Problem (49). Thus, by Proposition 5.4, we necessarily have $(\boldsymbol{\mu}, f^*(\boldsymbol{\mu}_1)) = \nabla F^{t^*}(L^{t^*} m^t)$. Using the expression of ∇F^{t^*} from Lemma 5.3 and that of L^{t^*} from Lemma 5.2, we finally deduce that $\boldsymbol{\mu} = \mu^{t+1}$.

Let us now turn to the proof of (51). Note that, by definition of the normal cone, (51) is equivalent to having

$$\langle w_1^t - w_1^{t+1}, \mu_1 - \mu_1^{t+1}\rangle_{\mathcal{E}_1} + \langle \mu_2^t - \mu_2^{t+1}, \mu_2 - \mu_2^{t+1}\rangle_{\mathcal{E}_2} \leq 0$$

for every $\mu = (\mu_1, \mu_2) \in Q^t$.

For $\mu \in Q^t$, we set

$$\Lambda^t(\mu) = \langle w_1^t - w_1^{t+1}, \mu_1 - \mu_1^{t+1}\rangle_{\mathcal{E}_1} + \langle \mu_2^t - \mu_2^{t+1}, \mu_2 - \mu_2^{t+1}\rangle_{\mathcal{E}_2}.$$

First, we notice that

$$\begin{aligned}
&\langle \mu_2^t - \mu_2^{t+1}, \mu_2 - \mu_2^{t+1}\rangle_{\mathcal{E}_2} \\
&\quad = \langle \mu_2^t - A\,\mathfrak{E}m^t + m^t(K^t)b - \mu_2^{t+1}, \mu_2 - \mu_2^{t+1}\rangle_{\mathcal{E}_2} + \langle A\,\mathfrak{E}m^t - m^t(K^t)b, \mu_2 - \mu_2^{t+1}\rangle_{\mathcal{E}_2} \\
&\quad \leq \langle A\,\mathfrak{E}m^t - m^t(K^t)b, \mu_2 - \mu_2^{t+1}\rangle_{\mathcal{E}_2},
\end{aligned}$$

since $\mu_2^{t+1} = \mathrm{proj}_Q(\mu_2^t - A\,\mathfrak{E}m^t + m(K^t)b)$. Also notice that, by definition of w_1^{t+1}, we have

$$w_1^t = \mathfrak{E}m^t + \left(1 - m^t(K^t)\right) w_1^{t+1}$$

and thus

$$w_1^t - w_1^{t+1} = \mathfrak{E}m^t - m^t(K^t) w_1^{t+1}$$

and

$$\Lambda^t(\mu) \leq \langle \mu - \mu^t, \left(\mathfrak{E}m^t, A\,\mathfrak{E}m^t\right)\rangle_{\mathcal{E}} - m^t(K^t)\langle \mu - \mu^{t+1}, \left(w_1^{t+1}, b\right)\rangle_{\mathcal{E}}.$$

Clearly, the result holds if $m^t = 0_{\mathcal{M}(K^t)}$.

Now, assume that $m^t \neq 0_{\mathcal{M}(K^t)}$. We recall that, by definition of μ_1^{t+1} in Algorithm 5, we have $w_1^{t+1} \in \partial f^*\left(\mu_1^{t+1}\right)$, and thus

$$\langle \mu_1 - \mu_1^{t+1}, w_1^{t+1}\rangle_{\mathcal{E}_1} \leq f^*(\mu_1) - f^*\left(\mu_1^{t+1}\right)$$

which yields, using the fact that $-m^t \in \mathcal{M}^+(K^t)$,

$$\Lambda^t(\mu) \leq \left\langle \mu - \mu^{t+1}, \left(\mathfrak{E}m^t, A\,\mathfrak{E}m^t\right)\right\rangle_{\mathcal{E}} - m^t(K^t)\left(f^*(\mu_1) - f^*\left(\mu_1^{t+1}\right) + \left\langle \mu_2 - \mu_2^{t+1}, b\right\rangle_{\mathcal{E}_2}\right).$$

Recalling that m^t is solution to Problem (49) and $(\mu^{t+1}, f^*\left(\mu_1^{t+1}\right))$ is solution to Problem (48), it follows from Proposition 5.4 that $L^t\left(\mu^{t+1}, f^*\left(\mu_1^{t+1}\right)\right) \in \partial G^{t^*}(-m^t)$. Hence, from Lemma 5.3, we have

$$\int_{K^t}\left[f^*\left(\mu_1^{t+1}\right) + \left\langle \mu_2^{t+1}, b\right\rangle_{\mathcal{E}_2} - \ell^t - \left\langle \mu_1^{t+1} + A^*\mu_2^{t+1}, v_1\right\rangle_{\mathcal{E}}\right] dm^t(v_1)$$
$$= m^t(K^t)\left(f^*\left(\mu_1^{t+1}\right) + \left\langle \mu_2^{t+1}, b\right\rangle_{\mathcal{E}_2} - \ell^t\right) - \left\langle \mu_1^{t+1} + A^*\mu_2^{t+1}, \mathfrak{E}m^t\right\rangle_{\mathcal{E}} = 0,$$

and thus

$$\Lambda^t(\mu) \leq \left\langle \mu_1 + A^*\mu_2, \mathfrak{E}m^t\right\rangle_{\mathcal{E}} - m^t(K^t)\left(f^*(\mu_1) + \left\langle \mu_2, b\right\rangle_{\mathcal{E}_2} - \ell^t\right).$$

Since $\mu \in Q^t$, it follows from (50) that, for all $v_1 \in \overline{\mathrm{conv}}(K^t)$,

$$\Lambda^t(\mu) \leq \left\langle \mu_1 + A^*\mu_2, \mathfrak{E}m^t\right\rangle_{\mathcal{E}} - m^t(K^t)\langle \mu_1 + A^*\mu_2, v_1\rangle_{\mathcal{E}_1}.$$

Set $\bar{v}_1 = \dfrac{\mathfrak{E}m^t}{m^t(K^t)}$. We have $\bar{v}_1 \in \overline{\mathrm{conv}}(K^t)$ since $\dfrac{m^t}{m^t(K^t)} \in \mathcal{P}\left(K^t\right)$ and $\mathfrak{E}$ is linear, and taking $v_1 = \bar{v}_1$ in the above inequality yields $\Lambda^t(\mu) \leq 0$, which concludes the proof. □

We finally collect the results of this section in order to provide a proof for Proposition 4.9.

Proof of Proposition 4.9. *i)* Thanks to Lemmas 5.1 and 5.2, Problem (33) coincides with Problem (49), and the latter is the dual of Problem (48), which is equivalent (up to a change in its value and a transformation in its variables, as discussed in the beginning of this section) to Problem (35).

ii) Part *a* of the conclusion follows from Lemma 5.7, while part *b* follows by combining the definition of μ_1^{t+1} in Algorithm 5, Lemma 2.2, Corollary 5.6, and Lemma 5.7. □

6 Numerical examples

In this section, given $n \in \mathbb{N}^*$ and $p \in [1, +\infty]$, we denote by B_p^n the closed unit ball in $\mathbb{R}^n$ for the ℓ^p norm, which is denoted by $\|\cdot\|_p$. We denote by $\mathcal{M}_{n,m}(\mathbb{R})$ the set of real matrices of size $n \times m$. For square matrices, we write $\mathcal{M}_n(\mathbb{R})$ instead of $\mathcal{M}_{n,n}(\mathbb{R})$. The vector space of symmetric matrices of size n is denoted by $\mathcal{S}_n(\mathbb{R})$ and the subset of those which are positive semidefinite is denoted by $\mathcal{S}_n^+(\mathbb{R})$.

6.1 A projection problem

We first test our algorithm in a simple problem taken from [20]. Let $p \in \{1, 2\}$ and $A \in \mathcal{M}_{1,2}(\mathbb{R})\backslash\{0\}$. We aim at solving

$$\underset{x \in B_p^2}{\text{minimize}}\ \frac{1}{2}\|x - y\|^2 \text{ s.t. } Ax = 0. \tag{52}$$

Notice that the feasible set of the problem is a segment, obtained as the intersection of a convex set and a line, whose extremities can be computed analytically. The projection of a point onto a segment being easy to solve, the problem can be solved analytically without difficulty.

Structure This problem fits in our framework if we take $\mathcal{E}_1 = \mathbb{R}^2$, $f = \frac{1}{2}\|\cdot - y\|^2$, $\mathcal{E}_2 = \mathbb{R}$, $Q_1 = \{0\}$, $Q_2 = \mathbb{R}$, $b = 0$, and $K = B_p^2$. For this choice, we indeed have $\sigma_Q = \iota_{\{0\}}$. Notice that, since K is compact, Assumption (H2) is satisfied.

Qualification condition This problem is qualified. Indeed, if $\mathrm{Ker}(A) = \mathbb{R} \times \{0\}$, then we can take $K^{-1} = \{(0, -1), (0, 1)\}$, and otherwise we can take $K^{-1} = \{(-1, 0), (1, 0)\}$.

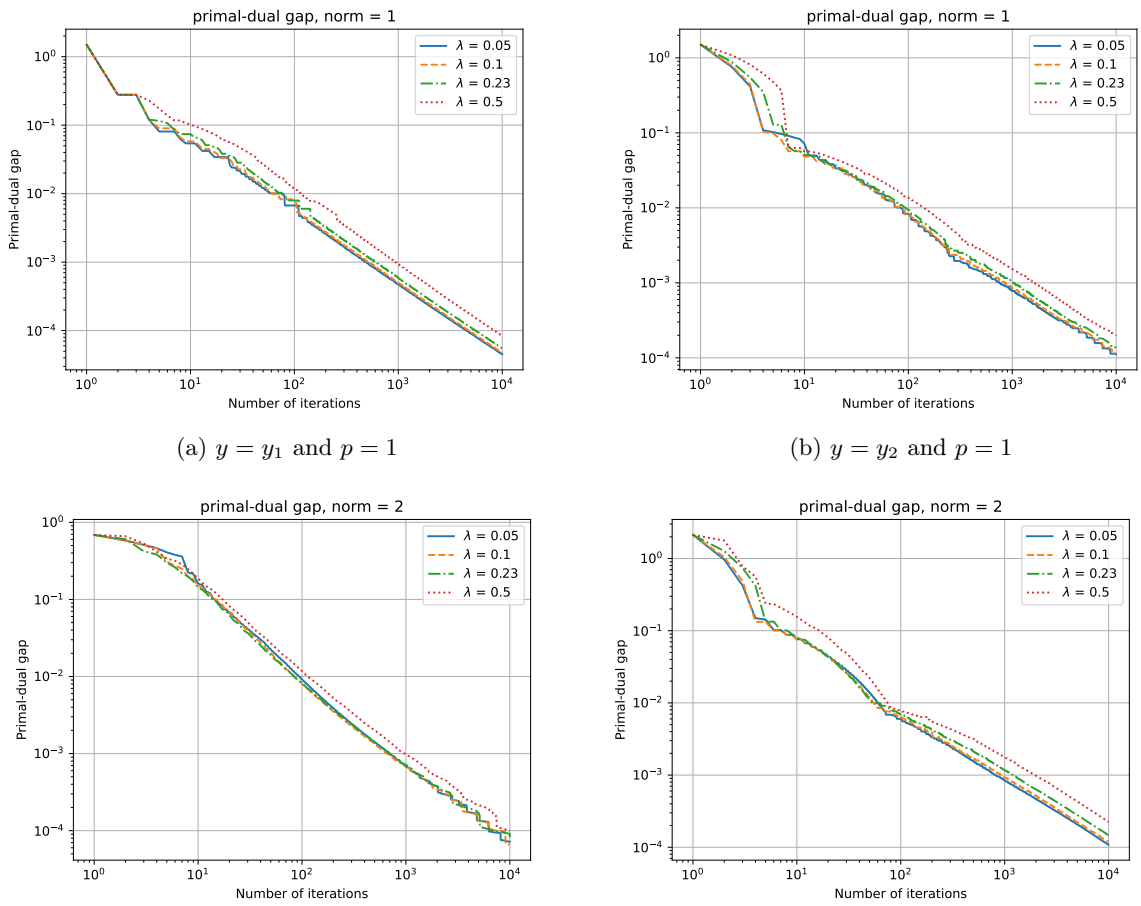
Oracle The LMO writes: $\underset{x \in B_p^2}{\text{minimize}}\ \langle \mu, x\rangle$, for a given $\mu \in \mathbb{R}^2$. For $p = 1$, the set $K = B_1^2$ is the convex hull of the four points $(0, 1)$, $(0, -1)$, $(1, 0)$, and $(-1, 0)$, therefore, for any $\mu \in \mathbb{R}^2$, one of them is solution to the LMO, making its resolution easy. For $p = 2$, two cases must be considered. If $\mu = 0$, any point in K is a solution. Otherwise, the unique solution is given by $-\mu/\|\mu\|_2$.

Numerical results We provide numerical results for a fixed value of A and for 10^4 iterations of the algorithm. For the two possible values of p, we consider two values of y, denoted y_1 and y_2 and chosen in such a way that, for $y = y_1$, the solution to Problem (52) lies on the boundary of K (i.e., on the unit sphere) and, for $y = y_2$, the solution lies in the interior of K.

Let us note that for $p = 1$, the (primal) solution is obtained after finitely many iterations, since K is the convex hull of finitely many points. We also obtain an optimal primal solution in a finite number of iterations if it lies in the interior of K (for $y = y_2$). We do not expect in general to find the dual solution in finitely many iterations.

In Figures 1a to 1d, we show the evolution of the primal-dual gap Δ^t at each iteration, in log-log scale, for $\lambda \in \{0.05, 0.1, 1-\sqrt{2-\sqrt{2}}, 0.5\}$, for the pairs (y, p) taken respectively as $(y_1, 1)$, $(y_2, 1)$, $(y_1, 2)$, and $(y_2, 2)$. For the pruning rule, we simply take $K^t = \{\hat{x}^t\} \cup K^{-1}$ at critical iterations.

We notice that, in all of these cases, the primal-dual gap show a numerical decrease toward 0 with a speed of order $1/t$, with the curves for $\lambda = 0.5$ decreasing slightly more slowly than those with the other choices of λ. We recall that our proof shows only a speed of order $1/\sqrt{t}$. In Figures 2a and 2b, we show the number of cuts at each iteration for the same values of λ and for the pairs (y, p) taken respectively as $(y_1, 2)$ and $(y_2, 2)$. We do not show the number of cuts for $p = 1$ since in that case it is bounded by 4 (the four extremal points of K, which contains K^{-1}). We notice that the number of cuts stays lower for smaller values of λ. This result was to be expected, since pruning steps should happen more often for lower values of λ.



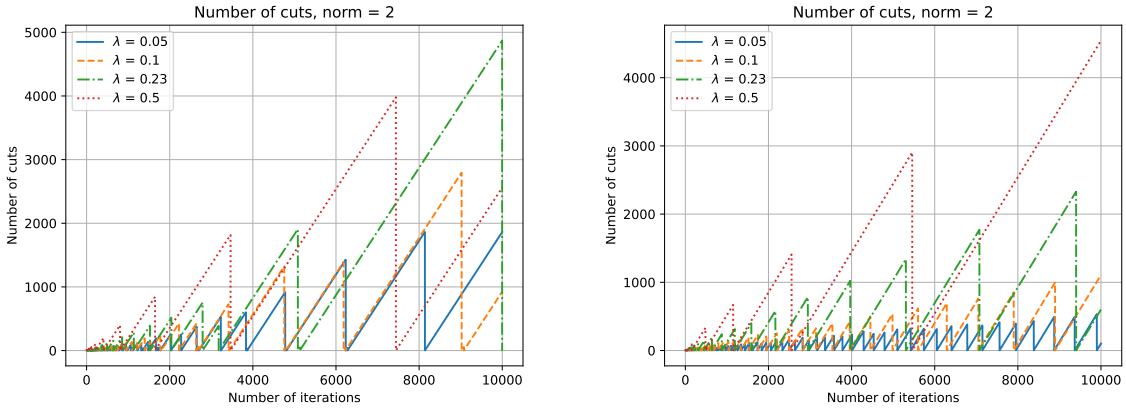
(a) $y = y_1$ and $p = 1$

(b) $y = y_2$ and $p = 1$

(c) $y = y_1$ and $p = 2$

(d) $y = y_2$ and $p = 2$

Figure 1: Primal-dual gap for 4 instances of Problem (52)



(a) $y = y_1$ and $p = 2$

(b) $y = y_2$ and $p = 2$

Figure 2: Evolution of the number of cuts for 2 instances of Problem (52)

6.2 A semidefinite program

We now test our algorithm with the following problem, extracted from [22]. Let $n \in \mathbb{N}^*$ and let $C \in \mathcal{S}_n(\mathbb{R})$. We aim at solving

$$\underset{X \in \mathcal{S}_n^+(\mathbb{R})}{\text{minimize}} \ \langle C, X \rangle_{\mathcal{M}_n(\mathbb{R})} \text{ s.t. } \operatorname{diag}(X) = 1_{\mathbb{R}^n} \tag{53}$$

where $\langle A, B \rangle_{\mathcal{M}_n(\mathbb{R})} = \operatorname{tr}(A^T B)$ is the canonical inner product.

Structure This problem fits in our framework if we take $\mathcal{E}_1 = \mathcal{S}_n(\mathbb{R})$, $f = \langle C, \cdot \rangle_{\mathcal{M}_n(\mathbb{R})}$, $\mathcal{E}_2 = \mathbb{R}^n$, A the linear operator which maps X to its diagonal, $Q_1 = \{0\}$, $Q_2 = \mathbb{R}^n$, $b = 1_{\mathbb{R}^n}$ and $K = \mathcal{S}_n^+(\mathbb{R}) \cap \{\operatorname{tr} \leq n+1\}$. Notice that, in this context, A^* is the linear operator which maps a vector Y to the diagonal matrix whose diagonal is Y. The definition of the set K is dictated by the qualification condition and by the necessity to have an oracle.

Qualification condition This problem verifies the qualification condition by taking K^{-1} as the convex hull of the null matrix and of the matrices $(n+1)E_{i,i}$, $i \in \{1, \ldots, n\}$, where $E_{i,j}$, $i, j \in \{1, \ldots, n\}$ are the elementary matrices.

Oracle A direct application of the spectral theorem reveals that $K = (n+1)\operatorname{conv}(K')$, where the set K' is defined by $K' = \{0\} \cup \left\{vv^T \,|\, \|v\|_2 = 1\right\}$. This allows to show the following lemma.

Lemma 6.1. *Let $M \in \mathcal{S}_n(\mathbb{R})$. Consider the problem*

$$\underset{V \in K}{\text{minimize}} \ \langle M, V \rangle_{\mathcal{M}_n(\mathbb{R})}. \tag{54}$$

Let s denote the small eigenvalue of M. If $s \geq 0$, then the null matrix is a solution to the problem. Otherwise, if $s < 0$, let $\bar{v}$ be an eigenvector associated with s, of norm equal to 1. Then $(n+1)\bar{v}\bar{v}^T$ is a solution to the problem.

As in the previous subsection, since K is bounded, Assumption (H2) is satisfied.

Numerical results We set $n = 10$ and take a random matrix $C \in \mathcal{S}_n(\mathbb{R})$. We then run the algorithm until the primal-dual gap Δ^t reaches 10^{-6}. We use the same pruning rule as in Section 6.1. In Figure 4, we show the number of cuts at each iteration, respectively for $\lambda = 0.05$ and $\lambda = 0.7$. Although we do not show it, the number of cuts for the other values of λ stays below 45. We notice that, as we expected, the maximal number of cuts is lower for smaller values of λ.

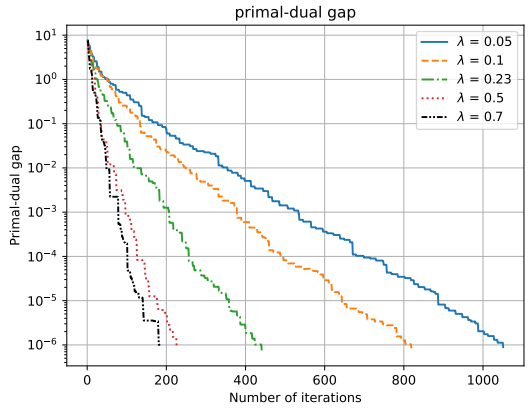
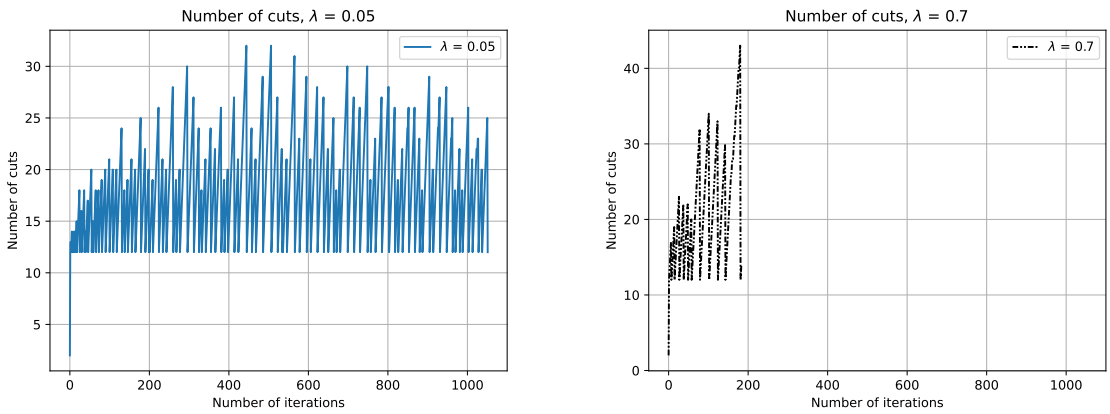


Figure 3: Primal-dual gap

In Figure 3, we show the primal-dual gap at each iteration for $\lambda \in \left\{0.05, 0.1, 1-\sqrt{2-\sqrt{2}}, 0.5, 0.7\right\}$, in y-log scale. We can see that, numerically, our algorithm has a linear convergence speed for this problem. Also notice that, unlike the previous problem, the algorithm seems to converge faster for larger values of λ. We do not know the reason for this behavior.



(a) $\lambda = 0.05$ (b) $\lambda = 0.7$

Figure 4: Number of cuts at each iteration

7 Conclusion and open problems

In this article, we proposed a Dualized Level-Set method for problems of the form

$$\underset{x_1 \in \mathcal{E}_1}{\text{minimize}} \; f(x_1) + \sigma_Q(Ax_1 - b) + \iota_K(x_1),$$

which we derived from an extension of the Level-Set method. We showed guarantees for the convergence of this algorithm. There are some improvement perspectives we can think of.

- Our experiments show better numerical convergence rates than we expected, which might be due to the specific form of these problems. This should not be surprising since some classical versions of the FWA are also known to have an improved convergence rate in some contexts, see for instance [6, Section 2.2].
- We assume that we are able to solve exactly our subproblems. An extension of our results could concern the situation where the problems are only solved up to a certain precision, as was done in [11] for the FWA.
- We chose to keep the parameter λ fixed along the iterations. Our numerical results, in particular those done for the semidefinite program, show that the value of λ may have a significant impact

on the efficiency of the algorithm. Future improvements may therefore concern variants of the algorithm in which λ is adapted along the algorithm. For instance, we think our proof could be adapted if we take $0 < \lambda_{\min} \leq \lambda \leq \lambda_{\max} < 1$, with changes of λ occurring only at critical iterations.

- The nonsmooth term in the cost function is restricted to be of the form $\sigma_Q(Ax - b)$, with Q of a specific shape. A possible extension may deal with the case of a general nonsmooth term g. From an algorithmic point of view, this would lead to a more general projection problem, whose dual (Problem (33)) would involve the Moreau envelope of g^*. Yet some difficulties would arise in the extension of Lemma 4.11, whose proof heavily relies on the structure of the nonsmooth term.
- Our numerical experience shows that pruning rules are unavoidable. Indeed, an implementation of the algorithm which retains all cuts computed at each iteration quickly becomes intractable. Yet the pruning rule that we proposed is not completely satisfactory in so far as we do not have a priori bounds on the number of cuts which need to be stored. As was seen on Figure 2, the critical iterations may occur at diminishing frequency, leading to an accumulation of many cuts. In [13], the author develops a variant of the Level-Set method with a different pruning rule, for which the number of cuts to be stored is bounded by the dimension of the space. We expect that his method can be extended in the same manner as we extended the Level-Set method of [17].

References

[1] Francis Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends® in Machine Learning*, 6(2-3):145–373, 2013.

[2] Francis Bach. Duality between subgradient and conditional gradient methods. *SIAM J. Optim.*, 25(1):115–129, 2015.

[3] Heinz H Bauschke and Patrick L Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

[4] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[5] Joseph-Frédéric Bonnans, Jean Charles Gilbert, Claude Lemaréchal, and Claudia A Sagastizábal. *Numerical optimization: theoretical and practical aspects*. Springer Science & Business Media, 2006.

[6] Gábor Braun, Alejandro Carderera, Cyrille W. Combettes, Hamed Hassani, Amin Karbasi, Aryan Mokhtari, and Sebastian Pokutta. Conditional gradient methods, 2023.

[7] Patrick L Combettes. Perspective functions: Properties, constructions, and examples. *Set-Valued and Variational Analysis*, 26:247–264, 2018.

[8] Rémi Flamary, Nicholas Courty, Davis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell*, 1(1-40):2, 2016.

[9] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

[10] Gauthier Gidel, Fabian Pedregosa, and Simon Lacoste-Julien. Frank-Wolfe splitting via augmented Lagrangian method. In *International Conference on Artificial Intelligence and Statistics*, pages 1456–1465. PMLR, 2018.

[11] Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International conference on machine learning*, pages 427–435. PMLR, 2013.

[12] Armand Joulin, Kevin Tang, and Li Fei-Fei. Efficient image and video co-localization with Frank-Wolfe algorithm. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 253–268. Springer, 2014.

[13] Krzysztof C. Kiwiel. Proximal level bundle methods for convex nondifferentiable optimization, saddle-point problems and variational inequalities. *Math. Programming*, 69(1):89–109, 1995.

[14] Karl Kunisch and Daniel Walter. On fast convergence rates for generalized conditional gradient methods with backtracking stepsize. *Numerical Algebra, Control and Optimization*, pages 0–0, 2022.

[15] Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *International Conference on Machine Learning*, pages 53–61. PMLR, 2013.

[16] Pierre Lavigne and Laurent Pfeiffer. Generalized conditional gradient and learning in potential mean field games. *Applied Mathematics & Optimization*, 88(3):89, 2023.

[17] Claude Lemaréchal, Arkadii Nemirovskii, and Yurii Nesterov. New variants of bundle methods. *Mathematical programming*, 69:111–147, 1995.

[18] Kang Liu and Laurent Pfeiffer. Mean field optimization problems: stability results and Lagrangian discretization. *arXiv preprint arXiv:2310.20037*, 2023.

[19] Sebastian Pokutta. The Frank-Wolfe algorithm: a short introduction. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, pages 1–33, 2023.

[20] Antonio Silveti-Falls, Cesare Molinari, and Jalal Fadili. Generalized conditional gradient with augmented Lagrangian for composite minimization. *SIAM Journal on Optimization*, 30(4):2687–2725, January 2020.

[21] Elias Wirth, Thomas Kerdreux, and Sebastian Pokutta. Acceleration of Frank-Wolfe algorithms with open-loop step-sizes. In *International Conference on Artificial Intelligence and Statistics*, pages 77–100. PMLR, 2023.

[22] Alp Yurtsever, Olivier Fercoq, and Volkan Cevher. A conditional-gradient-based augmented Lagrangian framework. In *International Conference on Machine Learning*, pages 7272–7281. PMLR, 2019.

[23] Constantin Zălinescu. *Convex analysis in general vector spaces.* World Scientific Publishing Co., Inc., River Edge, NJ, 2002.

[24] Song Zhou, Swati Gupta, and Madeleine Udell. Limited memory Kelley's method converges for composite convex and submodular objectives. *Advances in Neural Information Processing Systems*, 31, 2018.