Optimization
A Journal of Mathematical Programming and Operations Research

ISSN: 0233-1934 (Print) 1029-4945 (Online) Journal homepage: www.tandfonline.com/journals/gopt20

# Algorithms for nonsmooth optimization models under distance-to-set penalties

W. van Ackooij & W. de Oliveira

**To cite this article:** W. van Ackooij & W. de Oliveira (06 Oct 2025): Algorithms for nonsmooth optimization models under distance-to-set penalties, Optimization, DOI: 10.1080/02331934.2025.2567647

**To link to this article:** https://doi.org/10.1080/02331934.2025.2567647

Published online: 06 Oct 2025.

OPTIMIZATION
https://doi.org/10.1080/02331934.2025.2567647

Taylor & Francis
Taylor & Francis Group



# Algorithms for nonsmooth optimization models under distance-to-set penalties

W. van Ackooij [a] and W. de Oliveira [b]

[a]OSIRIS, EDF Lab Paris-Saclay, Palaiseau, France ; [b]CMA – Centre de Mathématiques Appliquées, MINES Paris – PSL, Sophia Antipolis, France

**ABSTRACT**

We introduce two optimization methods for Difference-of-Convex (DC) optimization problems involving hard and soft constraints. While hard-constraints are strict requirements that must be satisfied for a solution to be valid, soft constraints are preferences that are desirable but not mandatory to be met. In this work, the hard constraints are considered convex, while the soft constraints (potentially nonconvex) are incorporated into the objective function via squared distance-to-set penalty terms. Our first algorithm requires only a *difference of convex and weakly convex* (CwC) decomposition of the objective function – a milder assumption than the standard DC decomposition, while preserving implementability and broadening the method's applicability. The second algorithm is an original bundle method that leverages a novel self-stabilizing model and operates under a standard DC decomposition framework. The proposed implementable algorithms come with convergence guarantees to critical points of the penalized, nonsmooth, nonconvex, optimization model. The theoretical framework is grounded in variational analysis and nonsmooth optimization, and our approaches have potential applications in signal processing, machine learning, and operations research.

**ARTICLE HISTORY**
Received 28 May 2025
Accepted 22 September 2025

**KEYWORDS**
Nonsmooth optimization; nonconvex optimization; variational analysis

**2020 MATHEMATICS SUBJECT CLASSIFICATIONS**
49-02; 65K10

## 1. Introduction

In many optimization models, or at least when setting them up, there is often a discussion among modellers regarding the use of soft versus hard constraints. This is a typical situation in various energy models, where one also seeks to incorporate preferred operational practices into the model. Such practices may aim to prevent premature material degradation or to conserve limited resources for future use.

A relevant example is the operation of battery energy storage systems (BESS), which are typically optimized over short time horizons while considering their long-term impact on battery lifespan [1]. Quantities of interest often include the effect on a state vector commonly referred to as the *state-of-health*, or constraints on the total number of charge-discharge 'cycles' allowed per day. In this context, one may wonder whether penalizing

**CONTACT** W. de Oliveira welington.oliveira@minesparis.psl.eu
Dedicated to Professor Michel Théra on the occasion of his 80th birthday

deviations from a preferred operating zone could offer greater flexibility than imposing hard constraints. Indeed, in challenging situations, exceeding the typical limit of two daily cycles might provide valuable system services that can be compensated for at a later time. Other examples arise in regression and image processing problems, where sparsity is often desirable. It is well-established practice to include a term in the objective function that promotes or induces sparse solutions.

Motivated by these applications, we are interested in the problem of minimizing a given structured function $f$ over a convex set $X$, complicated by the presence of one or more 'soft' constraint sets $K_1, \ldots, K_m$. The corresponding 'hard'-constrained optimization problem would then be formulated as:

$$\min_{x \in X \cap K_1 \cap \cdots \cap K_m} f(x). \tag{1}$$

However, as argued above, we are interested in the case wherein the sets $K_i$, $i = 1, \ldots, m$, represent conditions or requirements that are preferred but not strictly necessary to be met. Alternatively, these sets may represent potentially competing factors that must be balanced, but whose strict enforcement would render problem (1) infeasible. Consequently, we will consider a penalized form of (1), where the soft constraints are deployed to allow some degree of flexibility. To this end, let $\rho_i > 0$ be given penalties for $i = 1, \ldots, m$ and $\emptyset \neq K_i \subset \mathbb{R}^n$ closed sets. The problem of interest is then formulated as follows:

$$\min_{x \in X} \varphi(x), \quad \text{with} \quad \varphi(x) := f(x) + \sum_{i=1}^{m} \rho_i \mathrm{d}^2_{K_i}(x), \tag{2}$$

where $\mathrm{d}^2_{K_i}$ denotes the squared (Euclidean) distance to $K_i$, i.e.

$$\mathrm{d}^2_{K_i}(x) := \min_{p \in K_i} \frac{1}{2} \left\| p - x \right\|^2 = \frac{1}{2} \left\| \mathrm{P}_{K_i}(x) - x \right\|^2.$$

Throughout this work, the set $X \subset \mathbb{R}^n$ is assumed closed and convex and the projection $\mathrm{P}_{K_i}$ onto each (potentially nonconvex) set $K_i$ is convenient to execute. We are interested in situations where the nonsmooth function $f : \mathbb{R}^n \to \mathbb{R}$ is potentially nonconvex, but has a structure that can be exploited in numerical optimization. This is the case of difference-of-convex programming [2, Section 4.6].

One can also come at problems of the form (2) by looking at the Sharp Lagrangian of problem (1). Indeed, the constraint $x \in K_i$ can be equivalently expressed as $\mathrm{d}^2_{K_i}(x) = 0$. The Sharp Lagrangian, using the $\ell_1$-norm, fits naturally into the framework of (2); see [3, Chp. 11, Section K$*$]. In such a setting, the primal-dual bundle method proposed in [4] computes an (approximate) solution to problem (1) by solving, inexactly, a sequence of subproblems of the form (2).

In this work, we focus on optimization methods for computing critical points of problem (2), that is, points satisfying certain necessary optimality conditions; see (3) below. With the same goal, the authors of [5] look at a very similar setting but with $X = \mathbb{R}^n$ and wherein $f$ is assumed twice continuously differentiable. The authors moreover make the assumption that the algorithm never visits points where $\mathrm{P}_{K_i}$ is multivalued. A Newton descent approach is then constructed, grounded in the principle of majorizing the squared distance function.

The majorization principle is also present in [6]: a quadratic – or other simple – upper model is set up for $\varphi$, so that the proximal step, with $\mathrm{d}^2_{K_i}$ replaced by the distance to the last projected point, is easy to compute. In fact this is so because it is an unconstrained quadratic program. Once again the assumption is made that the projection is single-valued at visited points.

There is also a link between the above works and those involving so-called error bounds. Indeed, a key ingredient of [5, 6] is the capacity of being able to produce an upper estimate of some kind for the distance function $\mathrm{d}^2_{K_i}(x)$. If the constraint sets are given by explicit inequalities, for instance $K_1 = \{x \in \mathbb{R}^n : g_1(x) \leq 0\}$, then the existence of an estimate of the type:

$$\mathrm{d}_{K_1}(x) \leq c[\max\{g_1(x), 0\} + \max\{g_1(x), 0\}^\gamma],$$

for some $c > 0$ and exponent $\gamma$ is called an error bound. The existence of such a bound, would thus allow for a 'majorization model' in the parlance of [6]. We refer to the works by Michel Théra on this topic: [7–10]. In our situation however, we will not require error bounds as we assume projection onto the sets $K_i$ doable. Furthermore, we do not rely on the principle of majorization, as we are interested in situations where $f$, although structured, is a challenging function that is iteratively approximated by a model which need not be a majorizing one. We focus on the setting where the objective function $f$ is DC, i.e. it is expressible as the difference of two convex functions $f_1$ and $f_2$. As the squared distance function to $K_i$ is also DC (it is indeed convex if $K_i$ is convex) [2, Proposition 4.13], problem (2) is a DC programming problem for which several approaches exist. We refer the interested reader to the review paper [11] and tutorial [12] on DC programming. Recent papers on the subject focus on boosted and inertial algorithms [13–15], splitting methods [16–18], composition involving DC functions [19–21], and bundle methods [22–24].

In the DC setting, at least theoretically, the function $\varphi$ in (2) can be majorized, and the standard DC algorithm of [25] could be applied. However, such an algorithm, and many others in the DC literature, typically rely on two key assumptions: first, that a DC decomposition $f = f_1 - f_2$ is available (and therefore also a DC decomposition of $\varphi$; see [2, Proposition 4.13]); and second, that the resulting convex subproblem, derived via the majorization principle, is simple to solve. These assumptions together are often unrealistic in applications. First, a DC decomposition $f = f_1 - f_2$ is not always readily available. Second, even when it is, the convex subproblem that consists in minimizing the sum of $f_1$, a linearization of $f_2$, and a quadratic regularization term over the set $X$ can still be challenging. In some cases, relaxing the convexity requirement on $f_2$ allows for greater flexibility in selecting a simpler $f_1$, potentially leading to a more tractable convex subproblem.

For this reason, instead of assuming that a DC decomposition is available, our first algorithm (presented in Section 2) operates in a more general setting where a decomposition into a difference of *convex and weakly convex* (CwC) functions is assumed. In this case, $f_2$ need not be convex, but only weakly convex. As noted in [26], CwC decompositions arise more naturally in many applications. We stress that, in the CwC setting, the principle of majorization is not directly applicable.

In contrast, in Section 3, we assume that a DC decomposition is available, but our starting point is the recognition that the resulting convex subproblem may be challenging to solve. To address this, we draw on ideas from [22–24] to propose a new bundle algorithm

for solving problem (2). Our original algorithm builds upon a novel self-stabilizing model and dispenses with the proximal term used in these earlier works.

We summarize our contributions as follows:

- In contrast to [5, 6], we allow for the presence of a closed convex constraint set $X$, forego the explicit use of the majorization principle in the subproblems to enhance computational efficiency and simplicity, and explicitly address the challenges arising from the potential multivaluedness of the projection operator in the nonconvex setting. In particular, for any given $x \in \mathbb{R}^n$, our algorithms only require arbitrary points in $\mathrm{P}_{K_i}(x)$, $i = 1, \ldots, m$.
- We consider the case where the mapping $f$ is nonconvex, specifically expressed as the difference of a convex and a weakly convex function, without requiring any particular smoothness assumptions. Moreover, in this setting, we propose an algorithm that operates without requiring prior knowledge of the modulus of weak convexity. Our approach extends Algorithm 23 in [2] to the setting under distance-to-set penalties.
- Under the assumption that a DC decomposition of $f$ is available, we propose a bundle method with a self-stabilizing model. The method's subproblem can be as simple as a strongly convex quadratic problem provided $X$ is polyhedral (or simply absent).
- We present convergence analyses for our two algorithms, establishing guarantees of convergence to critical points of the nonsmooth, nonconvex optimization problem (2).
- We present preliminary numerical experiments using our second algorithm applied to the sparse Wasserstein barycenter problem [27], which finds applications in image processing.

We will use mostly standard notation. In our work, $\mathrm{N}_X(x)$ will denote the (convex) normal cone to the set $X$ at the point $x$. Moreover, for a locally Lipschitzian function $f$, $\partial^{\mathrm{C}} f$ will denote Clarke's subdifferential, which in finite dimensions can be expressed as:

$$\partial^{\mathrm{C}} f(x) = \mathrm{Co}\left\{x^* : \nabla f(x_k) \to x^*, x_k \xrightarrow{\mathrm{d}} x\right\},$$

where $\xrightarrow{\mathrm{d}}$ means that only sequences are to be considered along which $f$ is continuously differentiable, e.g. [28, Theorem 2.5.1], see also [2, Proposition 3.10]. Furthermore, for a set $C \subseteq \mathbb{R}^n$, Co $C$ denotes the convex hull. Throughout our work, for a convex function $f$, $\partial f$ will denote the convex subdifferential – at points where such $f$ is also locally Lipschitz, this subdifferential coincides with Clarke's one. For a DC function $f = f_1 - f_2$, it is well known that it is locally Lipschitzian in the interior of its domain and that $\partial^{\mathrm{C}} f \subseteq \partial f_1 - \partial f_2$. This inclusion can be strict.

The paper is organized into three main sections: two are dedicated to the presentation and analysis of each algorithm, and the third reports numerical experiments that illustrate the performance of our bundle method algorithm.

## 2. CwC algorithm for optimization under distance-to-set penalties

This section presents an implementable algorithm for computing critical points of problem (2). Throughout, we assume that the following conditions hold.

**Assumption 2.1 (Basic requirements):** (a) (*Hard constraints*) The set $X \subset \mathbb{R}^n$ is nonempty, compact, and convex;

(b) (*Soft constraints*) The sets $\emptyset \neq K_i \subset \mathbb{R}^n$, for $i = 1, \ldots, m$, are closed, and their projections $\mathtt{P}_{K_i}$ are assumed to be efficiently computable;

(c) (*Basic objective function*) $f : \mathbb{R}^n \to \mathbb{R}$ has a CwC decomposition, that is, $f = f_1 - f_2$, with $f_1 : \mathbb{R}^n \to \mathbb{R}$ convex and $f_2 : \mathbb{R}^n \to \mathbb{R}$ a weakly convex function on an open set $\mathcal{O}$ containing $X$, that is, there exists a threshold (modulus) $\bar{\mu} \geq 0$ such that $f_2(x) + \frac{\mu}{2}\|x\|^2$ is convex on $\mathcal{O}$ for all $\mu \geq \bar{\mu}$.

It is important to note that CwC problems fall under the broader category of DC problems. The motivation for developing specialized algorithms to address CwC problems arises from the fact that while DC algorithms require a DC decomposition of the involved function, CwC problems require a CwC decomposition, which appears more naturally in applications [26]. Numerical algorithms for CwC problems have been recently proposed in [18, 26] and [2, Chapter 14]. In this section, we leverage the distance-to-set penalty in (2) to extend Algorithm 23 in [2] to the current setting.

Observe that in the framework of problem (2), a necessary optimality condition, called criticality, reads as

$$0 \in \partial f_1(\bar{x}) - \partial^{\mathrm{C}} f_2(\bar{x}) + \sum_{i=1}^{m} \rho_i \partial^{\mathrm{C}} \mathrm{d}^2_{K_i}(\bar{x}) + \mathrm{N}_X(\bar{x}). \tag{3}$$

We will check (approximate) criticality at iteration $k$ of our algorithm by examining a simpler inequality of the form $\nu_k \leq \mathtt{Tol}$, where $\nu_k \geq 0$ is a criticality measure motivated by the following observation: for $g_2^k \in \partial^{\mathrm{C}} f_2(x^k)$ and arbitrary $p_i^k \in \mathtt{P}_{K_i}(x^k)$, if the inclusion

$$0 \in \partial f_1(x^k) - g_2^k + \sum_{i=1}^{m} \rho_i (x^k - p_i^k) + \mathrm{N}_X(x^k),$$

holds, then $x^k$ is a critical point (recall that $x - \mathrm{Co}\,\mathtt{P}_{K_i}(x) = \partial^{\mathrm{C}} \mathrm{d}^2_{K_i}(x)$ [6, p. 9]; see also 29, Lemma 2.9]; or from the chain rule and [3, Example 8.53]). This inclusion is equivalent to saying that $x^k$ solves the strongly convex subproblem

$$\min_{x \in X} f_1(x) - \left[ f_2(x^k) + \left\langle g_2^k, x - x^k \right\rangle \right] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x - p_i^k \right\|^2. \tag{4}$$

In this case, the optimal value of this problem is $f_1(x^k) - f_2(x^k) + \sum_{i=1}^m \frac{\rho_i}{2} \left\| x^k - p_i^k \right\|^2$, which coincides with $\varphi(x^k)$ because $\frac{1}{2} \left\| x^k - p_i^k \right\|^2 = \mathrm{d}^2_{K_i}(x^k)$. Otherwise, if $x^k$ does not solve this subproblem, then its solution $\tilde{x}^k$ yields an optimal value strictly less than $\varphi(x^k)$. As a result, we propose the following measure of criticality for problem (2):

$$\nu_k = \varphi(x^k) - \left[ f_1(\tilde{x}^k) - \left[ f_2(x^k) + \left\langle g_2^k, \tilde{x}^k - x^k \right\rangle \right] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| \tilde{x}^k - p_i^k \right\|^2 \right] \geq 0.$$

One can readily compute this $\nu_k$ after having solved problem (4). With all the ingredients in place, we now present our first approach in Algorithm 1, which extends the (CwC)

**Algorithm 1** CwC algorithm for optimization under distance-to-set penalties

Let $x^0 \in X$, scalars $\mu_0 > 0$, $0 < \eta_1 < 1 < \eta_2$, and a tolerance $\texttt{Tol} \geq 0$ be given
**for** $k = 0, 1, \ldots$ **do**
  Compute $g_2^k \in \partial^{\mathrm{C}} f_2(x^k)$ and $p_i^k \in \mathrm{P}_{K_i}(x^k)$, $i = 1, \ldots, m$, arbitrary
  Let $\tilde{x}^k$ be a solution of subproblem (4)
  Set $v_k \leftarrow \varphi(x^k) - \left[ f_1(\tilde{x}^k) - [f_2(x^k) + \langle g_2^k, \tilde{x}^k - x^k \rangle] + \sum_{i=1}^m \frac{\rho_i}{2} \left\| \tilde{x}^k - p_i^k \right\|^2 \right]$
  **if** $v_k \leq \texttt{Tol}$ **then**
    Stop and return $x^k$
  **end if**
  Set $\mu \leftarrow \eta_1 \mu_{k-1}$, $d^k \leftarrow \tilde{x}^k - x^k$ and $t \leftarrow \min\left\{1, \frac{v_k}{\mu \|d^k\|^2}\right\}$
  **while** $f(x^k + td^k) + \sum_{i=1}^m \frac{\rho_i}{2} \left\| x^k + td^k - p_i^k \right\|^2 > \varphi(x^k) - tv_k + \frac{\mu t^2}{2} \left\| d^k \right\|^2$ **do**
    Set $\mu \leftarrow \eta_2 \mu$ and $t \leftarrow \min\left\{1, \frac{v_k}{\mu \|d^k\|^2}\right\}$
  **end while**
  Define $t_k \leftarrow t$, $\mu_k \leftarrow \mu$, and $x^{k+1} \leftarrow x^k + t_k d^k$
**end for**

algorithm introduced in 2, Section 14.1] to the distance-to-set penalization framework of problem (2).

Note that subproblem (4) consists of minimizing a strongly convex function over a closed convex set. Therefore, $\tilde{x}^k$ is well defined for all $k$. Computing $\tilde{x}^k$ is presumably an easy task, but if it is not case, then (standard) convex bundle methods are potential choices for the job. Observe also that $x^{k+1}$ is a convex combination of $\tilde{x}^k$ and $x^k$. Thus, a simple argument by induction shows that $\{x^k\}_k \subset X$ ($\subset \mathcal{O}$).

When $v_k > 0$, the step size that minimizes $\varphi(x^k) - tv_k + \frac{\mu t^2}{2} \left\| d^k \right\|^2$ with respect to $t \in [0, 1]$ is given by $t_\mu = \min\{1, \frac{v_k}{\mu \|d^k\|^2}\}$ (see details in the proof of Lemma 2.2.) Note that the larger the value of $\mu > 0$, the smaller $t_\mu \geq 0$ becomes. Lemma 2.2 shows that if $\mu \geq \bar{\mu}$ (the modulus of weak convexity of $f_2$), then the line search procedure terminates. Accordingly, using $\mu \geq \bar{\mu}$ can limit the algorithm's effectiveness: first, estimating this constant may be computationally expensive; second, $\bar{\mu}$ is a global upper bound on the curvature of $f_2$, which can lead to suboptimal step sizes in regions where the local curvature is significantly smaller. To avoid slow convergence, a strategy explored in the literature (see, for instance, 2, Section 14.1]) is to allow $\mu > 0$ to be smaller than the constant $\bar{\mu}$ in Assumption 2.1(c). However, $\mu$ must not be too small, as this could hinder convergence. A practical compromise is to begin the line search with a parameter $\mu$ smaller than that used in the previous iteration (i.e. $\mu = \eta_1 \mu_{k-1}$, with $\eta_1 \in (0, 1)$), and to increase it iteratively (by setting $\mu = \mu \eta_2$, with $\eta_2 > 1$) until

$$\varphi(x^k) - t_\mu v_k + \frac{\mu t_\mu^2}{2} \left\| d^k \right\|^2$$
$$\geq f(x^k + t_\mu d^k) + \sum_{i=1}^m \frac{\rho_i}{2} \left\| x^k + t_\mu d^k - p_i^k \right\|^2 \quad \left( \geq \varphi(x^k + t_\mu d^k) \right).$$

The pair $(\mu, t_\mu)$ satisfying this inequality is then set to $(\mu_k, t_k)$. With such a line search, there is no need to know $\bar{\mu}$.

### 2.1. Convergence analysis

Throughout this subsection we consider Algorithm 1 with $\texttt{Tol} = 0$. Should the algorithm stop at iteration $k$ with $\texttt{Tol} = 0$ then $\bar{x} = x^k$ is a critical point for problem (2), as $\hat{x}^k$ solves the master program (4), the optimality conditions of which are precisely criticality. In what follows we may thus assume that the algorithm loops forever.

Our first result provides a technical estimate, necessary to ensure that the line search procedure terminates finitely.

**Lemma 2.1:** *Suppose Assumption 2.1 holds. Let $\tilde{x}^k$ be a solution to (4) and $d^k = \tilde{x}^k - x^k$. If $\mu \geq \bar{\mu}$ (the modulus of weak convexity of $f_2$), then the following inequalities holds for all $t \in [0, 1]$:*

$$\varphi(x^k + td^k) \leq f(x^k + td^k) + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x^k + td^k - p_i^k \right\|^2 \tag{5a}$$

$$\leq \varphi(x^k) - tv_k + \frac{\mu t^2}{2} \left\| d^k \right\|^2. \tag{5b}$$

***Proof:*** The first inequality follows from the definitons of $\varphi$ and the squared distance function: $\mathrm{d}^2_{K_i}(x^k + td^k) \leq \frac{1}{2} \left\| x^k + td^k - p_i^k \right\|^2$. To show the second inequality, we first recall that $\left\{\tilde{x}^k\right\}, \left\{x^k\right\} \subset X \subset \mathcal{O}$. Thus, by employing the subgradient inequality to the convex function $f_2(x) + \frac{\mu}{2} \|x\|^2$ on $\mathcal{O}$ we get that, for $g_2^k \in \partial^{\mathrm{C}} f_2(x^k)$,

$$f_2(x^k) + \left\langle g_2^k, x - x^k \right\rangle - \frac{\mu}{2} \left\| x - x^k \right\|^2 \leq f_2(x) \quad \forall\, \tilde{x} \in \mathcal{O}.$$

In particular, for $x = x^k + td^k$, we have that $x \in X \subset \mathcal{O}$ and thus $f_2(x^k) + t\left\langle g_2^k, d^k \right\rangle - \frac{\mu t^2}{2} \left\| d^k \right\|^2 \leq f_2(x^k + td^k)$. This gives

$$\begin{aligned}
&f(x^k + td^k) + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x^k + td^k - p_i^k \right\|^2 \\
&\quad = f_1(x^k + td^k) - f_2(x^k + td^k) + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x^k + td^k - p_i^k \right\|^2 \\
&\quad \leq f_1(x^k + td^k) - \left[ f_2(x^k) + t\left\langle g_2^k, d^k \right\rangle \right] + \frac{\mu t^2}{2} \left\| d^k \right\|^2 + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x^k + t^k d^k - p_i^k \right\|^2 \\
&\quad = f_1(t\tilde{x}^k + (1-t)x^k) + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| t\tilde{x}^k + (1-t)x^k - p_i^k \right\|^2 \\
&\qquad - \left[ f_2(x^k) + t\left\langle g_2^k, d^k \right\rangle \right] + \frac{\mu t^2}{2} \left\| d^k \right\|^2.
\end{aligned}$$

By convexity of $f_1(\cdot) + \sum_{i=1}^m \frac{\rho_i}{2} \left\| \cdot - p_i^k \right\|^2$, we have that the latter right-hand side is less than or equal to

$$\begin{aligned}
&t\left[f_1(\tilde{x}^k) + \sum_{i=1}^m \frac{\rho_i}{2} \left\| \tilde{x}^k - p_i^k \right\|^2\right] + (1-t)\left[f_1(x^k) + \sum_{i=1}^m \frac{\rho_i}{2} \left\| x^k - p_i^k \right\|^2\right] \\
&\quad - \left[f_2(x^k) + t\left\langle g_2^k, d^k \right\rangle\right] + \frac{\mu t^2}{2} \left\| d^k \right\|^2 \\
&= \varphi(x^k) + t\left[f_1(\tilde{x}^k) + \sum_{i=1}^m \frac{\rho_i}{2} \left\| \tilde{x}^k - p_i^k \right\|^2\right] \\
&\quad - t\left[f_1(x^k) + \sum_{i=1}^m \rho_i \mathrm{d}_{K_i}^2(x^k) + \left\langle g_2^k, d^k \right\rangle\right] + \frac{\mu t^2}{2} \left\| d^k \right\|^2.
\end{aligned}$$

By adding and subtracting $tf_2(x^k)$, the right-hand side of the above equality is equal to

$$\begin{aligned}
&\varphi(x^k) + t\left[f_1(\tilde{x}^k) - \left(f_2(x^k) + \left\langle g_2^k, d^k \right\rangle\right) + \sum_{i=1}^m \frac{\rho_i}{2} \left\| \tilde{x}^k - p_i^k \right\|^2\right] \\
&\quad - t\left[f_1(x^k) - f_2(x^k) + \sum_{i=1}^m \rho_i \mathrm{d}_{K_i}^2(x^k)\right] + \frac{\mu t^2}{2} \left\| d^k \right\|^2 \\
&= \varphi(x^k) - t\left[-f_1(\tilde{x}^k) + \left(f_2(x^k) + \left\langle g_2^k, d^k \right\rangle\right) - \sum_{i=1}^m \frac{\rho_i}{2} \left\| \tilde{x}^k - p_i^k \right\|^2\right] \\
&\quad - t\varphi(x^k) + \frac{\mu t^2}{2} \left\| d^k \right\|^2 \\
&= \varphi(x^k) - tv_k + \frac{\mu t^2}{2} \left\| d^k \right\|^2,
\end{aligned}$$

which gives (5b). ■

With the help of the previous result we can now show that the line search procedure terminates and does not hinder progression of the algorithm.

**Lemma 2.2:** *Let Assumption 2.1 hold true. At every iteration k, the line search stops after finitely many steps with $t_k$ and $\mu_k$ satisfying*

$$\varphi(x^k + t_k d^k) \leq \varphi(x^k) - \frac{v_k}{2} \min\left\{\frac{v_k}{\mu_k \left\| d^k \right\|^2}, 1\right\}. \tag{6}$$

***Proof:*** Let $k$ be fixed. Clearly, as $\mu$ increases at every line search iteration, inequality (5b) ensures that after finitely many steps the line search stops with $\mu_k \leq \eta_2 \bar{\mu}$ satisfying

$$f(x^k + t_k d^k) + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x^k + t_k d^k - p_i^k \right\|^2 \leq \varphi(x^k) - t_k v_k + \frac{\mu_k t_k^2}{2} \left\| d^k \right\|^2.$$

Inequality (5a) thus gives $\varphi(x^k + t_k d^k) \leq \varphi(x^k) - t_k v_k + \frac{\mu_k t_k^2}{2} \left\| d^k \right\|^2$. It remains to show that, for the specific $t_k$ set by the algorithm, this last inequality is the same as (6). To this end, let us define $q(t) := \varphi(x^k) - t v_k + \frac{\mu_k t^2}{2} \left\| d^k \right\|^2$. Since $v_k > 0$ in this case (otherwise the algorithm would have terminated), the minimizer of $q(t)$ over $[0, +\infty)$ is strictly positive, and thus satisfies the optimality condition $-v_k + \mu_k \bar{t} \left\| d^k \right\|^2 = 0$, i.e. $\bar{t} = v_k / (\mu_k \left\| d^k \right\|^2)$. Hence, the minimizer of $q(t)$ over $[0, 1]$ is $t_k = \min\left\{1, \bar{t}\right\}$. We now split our analysis into two cases:

(a) $t_k = 1$. This happens when $\bar{t} \geq 1$, i.e. $v_k \geq \mu_k \left\| d^k \right\|^2$. It then follows that $q(t_k) = q(1) = \varphi(x^k) - v_k + \frac{\mu_k}{2} \left\| d^k \right\|^2 \leq \varphi(x^k) - v_k/2$.
(b) $t_k < 1$. In this case, $q(t_k) = \varphi(x^k) - \frac{v_k}{2} \frac{v_k}{\mu_k \| d^k \|^2}$.

Inequality (6) follows from (a) and (b) with the observation that $\varphi(x^k + t_k d^k) \leq q(t_k)$. ■

We can now show convergence of Algorithm 1.

**Theorem 2.3:** *Consider Algorithm 1 with* `Tol` $= 0$ *applied to problem (2) satisfying Assumption 2.1. Then* $\lim_{k\to\infty} v_k = 0$ *and all cluster points of* $\{x^k\}_k$ *are critical.*

***Proof:*** Inequality (5b) ensures that after finitely many steps the line search stops with $\mu_k \leq \eta_2 \bar{\mu}$ $(> \bar{\mu})$, which together with inequality (6) gives $0 \leq \frac{v_k}{2} \min\left\{ \frac{v_k}{\eta_2 \bar{\mu} (\mathtt{diam}(X))^2}, 1 \right\} \leq \varphi(x^k) - \varphi(x^{k+1})$. Since $\{\varphi(x^k)\}$ is nonincreasing and bounded from below (recall that $X$ is compact and $\varphi$ is continuous), it follows that it is convergent, and in particular, $\lim_{k\to\infty} [\varphi(x^k) - \varphi(x^{k+1})] = 0$. Therefore, $\lim_{k\to\infty} v_k = 0$. Let $\bar{x} \in X$ be a cluster point of $\{x^k\}$. Then there exists an infinite index set $J$ such that $\lim_{J \ni k\to\infty} x^k = \bar{x}$. As $\{g_2^k\}_J$ is bounded and the projection onto a nonempty closed set is an outer-semicontinuous and locally bounded operator [3, Example 5.23], there exist $J' \subset J$, $\bar{g}_2 \in \partial^{\mathrm{C}} f_2(\bar{x})$, and $\bar{p}_i \in \mathrm{P}_{K_i}(\bar{x})$ such that

$$\lim_{J' \ni k\to\infty} x^k = \bar{x}, \quad \lim_{J' \ni k\to\infty} g_2^k = \bar{g}_2 \quad \text{and} \quad \lim_{J' \ni k\to\infty} p_i^k = \bar{p}_i, \quad i = 1, \ldots, m.$$

Now, by using the definition of $v_k$ and passing to the limit with $k \in J'$ tending to infinity we obtain

$$\begin{aligned} 0 &= \lim_{J' \ni k\to\infty} v_k \\ &= \lim_{J' \ni k\to\infty} \left\{ \varphi(x^k) - \left[ f_1(\tilde{x}^k) - \left[ f_2(x^k) + \left\langle g_2^k, \tilde{x}^k - x^k \right\rangle \right] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| \tilde{x}^k - p_i^k \right\|^2 \right] \right\} \end{aligned}$$

$$= \varphi(\bar{x}) - \lim_{J' \ni k \to \infty} \left[ f_1(\tilde{x}^k) - \left[ f_2(x^k) + \left\langle g_2^k, \tilde{x}^k - x^k \right\rangle \right] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| \tilde{x}^k - p_i^k \right\|^2 \right],$$

showing that the last limit above exists and equals $\varphi(\bar{x})$. The definition of $\tilde{x}^k$ implies that, for all $x \in X$,

$$\begin{aligned} & f_1(x) - \left[ f_2(x^k) + \left\langle g_2^k, x - x^k \right\rangle \right] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x - p_i^k \right\|^2 \\ & \geq f_1(\tilde{x}^k) - \left[ f_2(x^k) + \left\langle g_2^k, \tilde{x}^k - x^k \right\rangle \right] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| \tilde{x}^k - p_i^k \right\|^2. \end{aligned}$$

By passing to the limit with $k \in J'$ tending to infinity we get, for all $x \in X$,

$$\begin{aligned} & f_1(x) - [f_2(\bar{x}) + \langle \bar{g}_2, x - \bar{x} \rangle] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x - \bar{p}_i \right\|^2 \\ & \geq \lim_{J' \ni k \to \infty} \left[ f_1(\tilde{x}^k) - \left[ f_2(x^k) + \left\langle g_2^k, \tilde{x}^k - x^k \right\rangle \right] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| \tilde{x}^k - p_i^k \right\|^2 \right] \\ & = \varphi(\bar{x}), \end{aligned}$$

i.e. $\min_{x \in X} \left\{ f_1(x) - [f_2(\bar{x}) + \langle \bar{g}_2, x - \bar{x} \rangle] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x - \bar{p}_i \right\|^2 \right\} \geq \varphi(\bar{x})$. As $\bar{x}$ is feasible to this problem, its optimal value is bounded from above by $\varphi(\bar{x})$. Hence, $\bar{x}$ solves this problem, which entails criticality of $\bar{x}$ for problem (2) (under Assumption 2.1). ■

**Remark 2.1:** The assumption that $X$ is compact can be relaxed when $f_2$ is convex and the level set $\left\{ x \in X : \varphi(x) \leq \varphi(x^0) \right\}$ is compact. In this case, it can be shown that the sequence $\left\{ \tilde{x}^k \right\}$ remains within this level set. Since $\bar{\mu} = 0$, there is no need for the line search in Algorithm 2. Consequently, $x^k = \tilde{x}^k$ and $d^k = 0$. The proof of Theorem 2.3 can be adapted to show that the algorithm asymptotically computes a critical point even when $X$ is unbounded. We omit this proof, as in this case Algorithm 1 reduces to the well-known DC algorithm [30]; see also 2, Section 14.2].

Finally, we close this section by mentioning that Algorithm 1 ensures that $\lim_{k \to \infty} (\min_{j \leq k} v_j) = 0$ at convergence rate of order $1/\sqrt{k}$. This is the well-known convergence rate of Frank-Wolfe algorithms. Once (5a) holds, a formal proof of this rate of convergence can be found, for instance, in [31, Theorem 7].

## 3. DC bundle algorithm for optimization under distance-to-set penalties

We now consider the situation wherein the function $f$ is explicitly given as a difference of convex functions. Formally, the set of assumptions is follows.

**Assumption 3.1 (Basic requirements):** (a) (*Hard constraints*) The set $X \subset \mathbb{R}^n$ is nonempty, closed, and convex;

(b) (*Soft constraints*) The sets $\emptyset \neq K_i \subset \mathbb{R}^n$, $i = 1, \ldots, m$, are closed, and their projections $\mathrm{P}_{K_i}$ are assumed to be efficiently computable;
(c) (*Basic objective function*) The function $f : \mathbb{R}^n \to \mathbb{R}$ has a readily available DC decomposition, that is, $f = f_1 - f_2$, with $f_1$ and $f_2$ convex functions.

In contrast to Assumption 1, item (a) does not require $X$ to be compact, and item (c) imposes additional structure on $f_2$, which is now assumed to be convex. On the other hand, in the following analysis, $f_1$ can be only accessible via a first-order oracle.

Under these basic assumptions, we address problem (2) using a bundle method. The main idea is to replace $f_1$ in the convex subproblem (4) with a model $f_k^{\mathrm{M}}$. This substitution results in a subproblem that is significantly easier to solve. Such a model is assumed to be convex and to approximate $f_1$ from below, that is, $f_k^{\mathrm{M}}(x) \leq f_1(x)$ for all $x \in X$. A simple classic way to set up such a model is to construct a cutting-plane model for $f_1$: at iteration $k$, having collected first-order information of $f_1$, that is, $(f_1(x^j), g_1^j \in \partial f_1(x^j))$, $j = 1, \ldots, k$, we may set $f_k^{\mathrm{M}} = \check{f}_1^k$, with

$$\check{f}_1^k(x) := \max_{j=1,\ldots,k} \left\{ f_1(x^j) + \left\langle g_1^j, x - x^j \right\rangle \right\} \leq f_1(x).$$

The bundle of information indexed by $\{1, \ldots, k\}$ does not need to be full, that is, this information can be managed following 'standard' bundle management assumptions, essentially allowing us to throw away all elements except for the last one and provided the aggregate (artificial) linearization is added; see [32] or [2, Section 11.4.2]. In short, any rule to update $f_k^{\mathrm{M}}$ to $f_{k+1}^{\mathrm{M}}$ satisfying the following minimal requirements (taken from [32, Eqs. (4.7)–(4.8)]) is enough to ensure convergence of proximal bundle methods. (Observe that $f_{k+1}^{\mathrm{M}}$ need not be piecewise linear as the cutting-plane model above.)

**Assumption 3.2 (Model Management):** At iteration $k$, having a trial point $x^{k+1}$, let $a_k(x)$ be the *aggregate* linearization $a_k(x) = f_k^{\mathrm{M}}(x^{k+1}) + \langle s^k, x - x^{k+1} \rangle$, with $s^k \in \partial f_k^{\mathrm{M}}(x^{k+1})$. The convex model for $f_1$ must satisfy the following conditions:

$$\begin{aligned} f_k^{\mathrm{M}}(x) &\leq f_1(x) \quad \text{for } k = 0, 1, \ldots, \text{ and all } x \in X, \\ a_k(x) &\leq f_{k+1}^{\mathrm{M}}(x) \\ f_1(x^{k+1}) + \left\langle g_1^{k+1}, x - x^{k+1} \right\rangle &\leq f_{k+1}^{\mathrm{M}}(x). \end{aligned}$$

The cutting-plane model above automatically satisfies these conditions. This is carefully discussed in the literature, e.g. [33], [34, pattern P, p. 246], [35], 2, Section 11.4.2]. Presenting the approach in this way also allows for more sophisticated models. For instance when $f_1(x) = \sum_{\ell=1}^{L} f_1^\ell(x)$, with $f_1^\ell : \mathbb{R}^n \to \mathbb{R}$ convex functions, then one can think of the disaggregate cutting plane model $f_k^{\mathrm{M}} = \sum_{\ell=1}^{L} \check{f}_k^\ell$, which is more precise than the cutting plane model for the sum of functions. In the litterature on Benders decomposition, this is referred to as 'multi-cut'. There is an obvious link between the size of $L$, the difficulty of the master-program, the cost of the oracles and the interest of using the disaggregate model in practice. This can be sidestepped by partitioning $L$ in a smart way and use some intermediate model – also fitting the class. Yet further possibilities are partially exact models as for instance discussed in [36]. This also goes under the name of 'easy components', e.g. [37].

Standard bundle methods use an element $\hat{x}^k \in \{x^0, \ldots, x^k\} \subset X$ as a stability centre for the model, i.e. the best known candidate to solve (2). Here, instead of using $\hat{x}^k$ as an stability centre, we employ its projections $\hat{p}_i^k \in \mathrm{P}_{K_i}(\hat{x}^k)$, $i = 1, \ldots, m$, and define the following master program

$$\min_{x \in X} f_k^{\mathrm{M}}(x) - \left[ f_2(\hat{x}^k) + \left\langle \hat{g}_2^k, x - \hat{x}^k \right\rangle \right] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x - \hat{p}_i^k \right\|^2, \tag{7}$$

which is a 'lower' approximation of (4) (note that $\hat{g}_2^k \in \partial f_2(\hat{x}^k)$). Observe that this is a convex optimization problem, whose solution $x^{k+1}$ we assume is convenient to compute. This is the case when $X$ is a polyhedron and $f_k^{\mathrm{M}}$ is a piecewise linear model: the master program becomes a quadratic programming (QP) problem, for which several commercial and open-source solvers exist. We mention in passing that if $X = X_1 \times \cdots \times X_L$ and $f_1(x) = \sum_{\ell=1}^{L} f_1^\ell(x_\ell)$, where each $f_1^\ell \colon \mathbb{R}^{n_\ell} \to \mathbb{R}$ is convex, then the master problem (7) with the disaggregate cutting-plane model decomposes into $L$ smaller and independent subproblems. These can be solved in parallel to compute the trial point $x^{k+1}$.

Note that as $\hat{x}^k$ is feasible for (7), we get that

$$\begin{aligned}
&f_k^{\mathrm{M}}(x^{k+1}) - \left[ f_2(\hat{x}^k) + \left\langle \hat{g}_2^k, x^{k+1} - \hat{x}^k \right\rangle \right] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x^{k+1} - \hat{p}_i^k \right\|^2 \\
&\quad \leq f_k^{\mathrm{M}}(\hat{x}^k) - f_2(\hat{x}^k) + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| \hat{x}^k - \hat{p}_i^k \right\|^2 \\
&\quad \leq f_1(\hat{x}^k) - f_2(\hat{x}^k) + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| \hat{x}^k - \hat{p}_i^k \right\|^2 \\
&\quad = f(\hat{x}^k) + \sum_{i=1}^{m} \rho_i \mathrm{d}_{K_i}^2(\hat{x}^k) = \varphi(\hat{x}^k).
\end{aligned}$$

As a result, the predicted decrease $v_k^{\mathrm{M}}$ is nonnegative:

$$v_k^{\mathrm{M}} := \varphi(\hat{x}^k) - \left[ f_k^{\mathrm{M}}(x^{k+1}) - \left[ f_2(\hat{x}^k) + \left\langle \hat{g}_2^k, x^{k+1} - \hat{x}^k \right\rangle \right] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x^{k+1} - \hat{p}_i^k \right\|^2 \right] \geq 0.$$

The descent test

$$f(x^{k+1}) + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x^{k+1} - \hat{p}_i^k \right\|^2 \leq \varphi(\hat{x}^k) - \kappa v_k^{\mathrm{M}}, \quad \text{with } \kappa \in (0, 1), \tag{8}$$

decides whether the stability centres can be updated to elements in $\mathrm{P}_{K_i}(x^{k+1})$, $i = 1, \ldots, m$, or must be kept fixed for the next iteration (which will be performed with an updated $f_{k+1}^{\mathrm{M}}$). Note that to perform this test, there is no need to compute the projection of $x^{k+1}$ onto each $K_i$. Furthermore, when this test is satisfied then the objective function of (2) decreases by

at least $\kappa v_k^{\mathrm{M}} \geq 0$. This is because the definition of the distance function implies

$$\varphi(x) = f(x) + \sum_{i=1}^{m} \rho_i \mathrm{d}_{K_i}^2(x) \leq f(x) + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x - \hat{p}_i^k \right\|^2 \quad \forall\, x.$$

Given these ingredients, we can now present our specialized bundle method in Algorithm 2.

**Algorithm 2** DC bundle method for optimization under distance-to-set penalties

Let $x^0 \in X$, $\kappa \in (0,1)$, and $\texttt{Tol} \geq 0$ be given
Compute $f_i(x^0)$, $g_i^0 \in \partial f_i(x^0)$, $i = 1, 2$, and $\hat{p}_i^0 \in \mathrm{P}_{K_i}(x^0)$
Choose an initial model $f_0^M \leq f_1$, set $\hat{x}^0 \leftarrow x^0$, $\hat{g}_2^0 \leftarrow g_2^0$, $\mathcal{B}_k \leftarrow \{0\}$ and $k \leftarrow 0$
**for** $k = 0, 1, 2, \ldots$ **do**
  Let $x^{k+1}$ be the solution to the master program (7) ▷ Master program
  Set $v_k^{\mathrm{M}} := \varphi(\hat{x}^k) - \left[ f_k^{\mathrm{M}}(x^{k+1}) - [f_2(\hat{x}^k) + \langle \hat{g}_2^k, x^{k+1} - \hat{x}^k \rangle] + \sum_{i=1}^m \frac{\rho_i}{2} \left\| x^{k+1} - \hat{p}_i^k \right\|^2 \right]$
  **if** $v_k^{\mathrm{M}} \leq \texttt{Tol}$ **then**
    Stop: and return $\hat{x}^k$ ▷ Stopping test
  **end if**
  Compute $f_i(x^{k+1})$, $g_i^{k+1} \in \partial f_i(x^{k+1})$, $i = 1, 2$ ▷ Oracle call
  **if** $f(x^{k+1}) + \sum_{i=1}^m \frac{\rho_i}{2} \left\| x^{k+1} - \hat{p}_i^k \right\|^2 \leq \varphi(\hat{x}^k) - \kappa v_k^{\mathrm{M}}$ **then**
    Set $\hat{x}^{k+1} \leftarrow x^{k+1}$, $\hat{g}_2^{k+1} \leftarrow g_2^{k+1}$, and select arbitrarily $\hat{p}_i^{k+1} \in \mathrm{P}_{K_i}(\hat{x}^{k+1})$, $i = 1, \ldots, m$ ▷ Serious Step
  **else**
    Set $\hat{x}^{k+1} \leftarrow \hat{x}^k$, $\hat{g}_2^{k+1} \leftarrow \hat{g}_2^k$, and $\hat{p}_i^{k+1} \leftarrow \hat{p}_i^k$, $i = 1, \ldots, m$ ▷ Null Step
  **end if**
  Update $f_k^{\mathrm{M}}$ to a convex model $f_{k+1}^{\mathrm{M}}$ satisfying Assumption 3.2 ▷ Bundle management
**end for**

In contrast with other DC bundle methods in the literature (see for instance, [22–24] and [2, Algorithm 25]) Algorithm 2 dismisses the stabilizing (proximal) term $\frac{1}{2t_k} \left\| x - \hat{x}^k \right\|^2$, with $t_k > 0$ a prox-parameter. More specifically, for the DC decomposition

$$f_1(x) - f_2(x) + \sum_{i=1}^{m} \rho_i \mathrm{d}_{K_i}^2(x) = \tilde{f}_1(x) - \tilde{f}_2(x) \quad \text{with}$$

$$\tilde{f}_1(x) = f_1(x) + \left( \sum_{i=1}^{m} \frac{\rho_i}{2} \right) \|x\|^2 \quad \text{and} \quad \tilde{f}_2(x) = f_2(x) + \sum_{i=1}^{m} \rho_i \left\{ \max_{p_i \in K_i} \langle x, p_i \rangle - \frac{1}{2} \left\| p_i \right\|^2 \right\},$$

the master program of DC bundle methods is typically formulated as

$$\min_{x \in X} \tilde{f}_k^{\mathrm{M}}(x) - \left[ \tilde{f}_2(\hat{x}^k) + \left\langle \hat{s}_2^k, x - \hat{x}^k \right\rangle \right] + \frac{1}{2t_k} \left\| x - \hat{x}^k \right\|^2,$$

with $\hat{s}_2^k = \hat{g}_2^k + \sum_{i=1}^m \rho_i \hat{p}_i^k \in \partial \tilde{f}_2(\hat{x}^k)$ and $\tilde{f}_k^{\mathrm{M}}(x)$ a model for $\tilde{f}_1$. This is significantly different from our master program (7), which considers a model for $f_1$ rather than for $f_1(x) + (\sum_{i=1}^m \frac{\rho_i}{2}) \|x\|^2$, and omits $\frac{1}{2t_k} \left\| x - \hat{x}^k \right\|^2$. This omission is justified because the objective function in our master program (7), which does not employ a DC decomposition of the squared distance functions, is inherently self-stabilizing in the parlance of [35].

We mention in passing that if the convexity assumption on $f_2$ is relaxed to weak convexity, then the iterates must be defined using the subproblem described above, rather than

the one in Equation (7). Furthermore, an ad-hoc rule for selecting $t_k$ must be employed to ensure convergence. The reader is referred to [26] or [2, Algorithm 25] for further details on the more general setting of nonlinearly constrained CwC problems.

### 3.1. Convergence analysis

With the algorithm stated, we can now proceed with the analysis of its convergence. Here as usual in bundle methods, we need to check what happens if the algorithm produces an infinite sequence of serious steps (which are steps updating the stability centre $\hat{x}^k$) or only finitely many. First let us investigate what happens if $v_k^{\mathrm{M}}$ is found to be zero and the algorithm stops finitely.

**Lemma 3.1:** *Consider Algorithm 2 applied to problem (2) under Assumption 3.1. If at a certain iteration k it holds that $v_k^M = 0$, then $\bar{x} = \hat{x}^k$ is a critical point.*

***Proof:*** Note that

$$\begin{aligned} & f_k^{\mathrm{M}}(x^{k+1}) - \left[f_2(\hat{x}^k) + \left\langle \hat{g}_2^k, x^{k+1} - \hat{x}^k \right\rangle\right] + \sum_{i=1}^m \frac{\rho_i}{2} \left\| x^{k+1} - \hat{p}_i^k \right\|^2 \\ &= \min_{x \in X} f_k^{\mathrm{M}}(x) - \left[f_2(\hat{x}^k) + \left\langle \hat{g}_2^k, x - \hat{x}^k \right\rangle\right] + \sum_{i=1}^m \frac{\rho_i}{2} \left\| x - \hat{p}_i^k \right\|^2 \\ &\leq \min_{x \in X} f_1(x) - \left[f_2(\hat{x}^k) + \left\langle \hat{g}_2^k, x - \hat{x}^k \right\rangle\right] + \sum_{i=1}^m \frac{\rho_i}{2} \left\| x - \hat{p}_i^k \right\|^2 \leq \varphi(\hat{x}^k). \end{aligned}$$

Thus, $v_k^{\mathrm{M}} = 0$ implies $\varphi(\hat{x}^k) = f_k^{\mathrm{M}}(x^{k+1}) - [f_2(\hat{x}^k) + \langle \hat{g}_2^k, x^{k+1} - \hat{x}^k \rangle] + \sum_{i=1}^m \frac{\rho_i}{2} \left\| x^{k+1} - \hat{p}_i^k \right\|^2$, i.e. $\hat{x}^k$ is also optimal for $\min_{x \in X} f_1(x) - [f_2(\hat{x}^k) + \langle \hat{g}_2^k, x - \hat{x}^k \rangle] + \sum_{i=1}^m \frac{\rho_i}{2} \left\| x - \hat{p}_i^k \right\|^2$. With a strictly convex objective function, the solution is unique, hence this shows that $x^{k+1}$ equals $\hat{x}^k$ and this point solves the latter problem. Its optimality condition gives (3) with $\bar{x} = \hat{x}^k$, i.e. $\bar{x}$ is a critical point to problem (2). ■

**Proposition 3.2 (Null steps):** *Consider Algorithm 2 with* `Tol` $= 0$ *applied to problem (2) under Assumption 3.1. Assume furthermore that the algorithm produces only finitely many serious steps and that the model $f_k^M$ is managed following Assumption 3.2. Then $v_k^M \to 0$ and the last serious iterate is critical for problem (2).*

***Proof:*** An important observation is that since no serious steps are made, $\hat{x}^k$, $\hat{g}_2^k$ and $\hat{p}_i^k \in \mathtt{P}_{K_i}(\hat{x}^k)$ are never again updated (by assumption) when $k$ is large enough. To avoid confusion, we call them $\hat{x}$, $\hat{g}_2$ and $\hat{p}_i$. Observe furthermore that the master program objective function is identical – up to a constant – to

$$\min_{x \in X} f_k^{\mathrm{M}}(x) - \left[f_2(\hat{x}) + \langle \hat{g}_2, x - \hat{x} \rangle\right] + \frac{\mu}{2} \left\| x - \hat{p} \right\|^2,$$

with $\hat{p} = \sum_{i=1}^m \frac{\rho_i}{\mu} \hat{p}_i$ and $\mu = \sum_{i=1}^m \rho_i > 0$. This follows by expanding squares. As a result, the infinite sequence of null-steps is also defined by solving the above convex program. We

can now apply classic arguments from (convex) bundle methods, e.g. [32, Proposition 4.3] to entail that: $f_1(x^{k+1}) - f_k^{\mathrm{M}}(x^{k+1}) \to 0$ and $x^{k+1} \to \bar{x}$, with

$$\bar{x} = \operatorname*{argmin}_{x \in X} f_1(x) - [f_2(\hat{x}) + \langle \hat{g}_2, x - \hat{x} \rangle] + \frac{\mu}{2} \left\| x - \hat{p} \right\|^2.$$

Of course, any solution is characterized by

$$0 \in \partial f_1(\bar{x}) - \hat{g}_2 + \mu(\bar{x} - \hat{p}) + \mathrm{N}_X(\bar{x})$$

and this solution is unique because $f_1$ is convex – thus making the overall objective function strictly convex. Thus, to show that $\hat{x}$ satisfies the criticality condition (3) we only need to show that $\bar{x} = \hat{x}$.

By adding appropriate constants, it follows also that $\bar{x}$ is the unique solution of

$$\min_{x \in X} f_1(x) - [f_2(\hat{x}) + \langle \hat{g}_2, x - \hat{x} \rangle] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x - \hat{p}_i \right\|^2. \tag{9}$$

Furthermore, as a result of the subgradient inequality $f_2(\hat{x}) + \langle \hat{g}_2, \bar{x} - \hat{x} \rangle \leq f_2(\bar{x})$ and feasibility of $\hat{x}$ we get

$$\begin{aligned} f(\bar{x}) + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| \bar{x} - \hat{p}_i \right\|^2 &\leq f_1(\bar{x}) - [f_2(\hat{x}) + \langle \hat{g}_2, \bar{x} - \hat{x} \rangle] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| \bar{x} - \hat{p}_i \right\|^2 \\ &= \min_{x \in X} f_1(x) - [f_2(\hat{x}) + \langle \hat{g}_2, x - \hat{x} \rangle] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x - \hat{p}_i \right\|^2 \\ &\leq \varphi(\hat{x}). \end{aligned} \tag{10}$$

Now, upon using the definition of $v_k^{\mathrm{M}}$ and by moving to the limit, using the above, we get:

$$\begin{aligned} 0 \leq \lim_{k \to \infty} v_k^{\mathrm{M}} &= \varphi(\hat{x}) - \left[ f_1(\bar{x}) - [f_2(\hat{x}) + \langle \hat{g}_2, \bar{x} - \hat{x} \rangle] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| \bar{x} - \hat{p}_i \right\|^2 \right] \\ &\leq \varphi(\hat{x}) - \left[ f(\bar{x}) + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| \bar{x} - \hat{p}_i \right\|^2 \right]. \end{aligned}$$

Failure of the descent test implies:

$$\kappa v_k^{\mathrm{M}} > \varphi(\hat{x}) - \left[ f(x^{k+1}) + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x^{k+1} - \hat{p}_i \right\|^2 \right].$$

By passing to the limit with $k$ going to infinity we obtain

$$\begin{aligned} &\kappa \left( \varphi(\hat{x}) - \left[ f(\bar{x}) + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| \bar{x} - \hat{p}_i \right\|^2 \right] \right) \\ &\geq \lim_{k \to \infty} \kappa v_k^{\mathrm{M}} \geq \varphi(\hat{x}) - \left[ f(\bar{x}) + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| \bar{x} - \hat{p}_i \right\|^2 \right]. \end{aligned}$$

As $\kappa \in (0,1)$, this inequality implies that $\varphi(\hat{x}) \leq f(\bar{x}) + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\|\bar{x} - \hat{p}_i\right\|^2$. Combined with (10), we have that this last inequality holds as equality. Now $\hat{x}$ is feasible for (9) and has optimal objective function value (it realizes the lower bound). It must thus follow that $\bar{x} = \hat{x}$, since problem (9) has a unique solution (strictly convex objective function). Moreover, as noted above, $\bar{x} = \hat{x}$ satisfies the criticality condition (3). ■

We can now turn our attention to the case of infinitely many serious steps.

**Proposition 3.3 (Serious steps):** *Consider Algorithm 2 with* `Tol` $= 0$ *applied to problem (2) under Assumption 3.1. Assume furthermore that the set* $\left\{x : \varphi(x) \leq \varphi(x^0)\right\}$ *is compact and that the algorithm produces infinitely many serious steps. Then the sequence of serious steps admits a cluster point that is critical for problem (2).*

***Proof:*** With $\hat{x}^{k+1}$ the next serious step, we have

$$\varphi(\hat{x}^{k+1}) = f(\hat{x}^{k+1}) + \sum_{i=1}^{m} \rho_i \mathrm{d}^2_{K_i}(\hat{x}^{k+1}) \leq f(\hat{x}^{k+1}) + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\|\hat{x}^{k+1} - \hat{p}_i^k\right\| \leq \varphi(\hat{x}^k) - \kappa v_k^{\mathrm{M}}.$$

Since we assume that $\left\{x : \varphi(x) \leq \varphi(x^0)\right\}$ is compact, a telescoping-sum argument implies that $v_k^{\mathrm{M}}$ tends to zero along the subsequence of serious iterates. Furthermore, the sequence $\left\{\hat{x}^k\right\}$ belongs to this compact set and must thus admit a converging subsequence. Let $\hat{x} \in X$ be a cluster point of $\{\hat{x}^k\}$. Then there exists an infinite index set $J$ such that $\lim_{J \ni k \to \infty} \hat{x}^k = \hat{x}$. As $\{\hat{g}_2^k\}_J$ is bounded and the projection onto a nonempty closed set is an outer-semicontinuous and locally bounded operator [3, Example 5.23], there exist $J' \subset J$, $\hat{g}_2 \in \partial f_2(\hat{x})$, and $\hat{p}_i \in \mathrm{P}_{K_i}(\hat{x})$ such that

$$\lim_{J' \ni k \to \infty} \hat{x}^k = \hat{x}, \quad \lim_{J' \ni k \to \infty} \hat{g}_2^k = \hat{g}_2 \quad \text{and} \quad \lim_{J' \ni k \to \infty} \hat{p}_i^k = \hat{p}_i, \quad i = 1, \ldots, m.$$

Now, by using the definition of $v_k^{\mathrm{M}}$ and passing to the limit with $k \in J'$ tending to infinity we obtain

$$\begin{aligned}
0 &= \lim_{J' \ni k \to \infty} v_k^{\mathrm{M}} \\
&= \lim_{J' \ni k \to \infty} \left\{ \varphi(\hat{x}^k) - \left[ f_k^{\mathrm{M}}(x^{k+1}) - \left[ f_2(\hat{x}^k) + \left\langle \hat{g}_2^k, x^{k+1} - \hat{x}^k \right\rangle \right] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x^{k+1} - \hat{p}_i^k \right\|^2 \right] \right\} \\
&= \varphi(\hat{x}) - \lim_{J' \ni k \to \infty} \left[ f_k^{\mathrm{M}}(x^{k+1}) - \left[ f_2(\hat{x}^k) + \left\langle \hat{g}_2^k, x^{k+1} - \hat{x}^k \right\rangle \right] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x^{k+1} - \hat{p}_i^k \right\|^2 \right],
\end{aligned}$$

showing that the last limit above exists and equals $\varphi(\hat{x})$. As $f_k^{\mathrm{M}}$ approximates $f_1$ from below, the definition of $x^{k+1}$ implies that, for all $x \in X$,

$$f_1(x) - \left[ f_2(\hat{x}^k) + \left\langle \hat{g}_2^k, x - \hat{x}^k \right\rangle \right] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x - \hat{p}_i^k \right\|^2$$

$$\geq f_k^{\mathrm{M}}(x) - \left[f_2(\hat{x}^k) + \left\langle \hat{g}_2^k, x - \hat{x}^k \right\rangle\right] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x - \hat{p}_i^k \right\|^2$$

$$\geq f_k^{\mathrm{M}}(x^{k+1}) - \left[f_2(\hat{x}^k) + \left\langle \hat{g}_2^k, x^{k+1} - \hat{x}^k \right\rangle\right] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x^{k+1} - \hat{p}_i^k \right\|^2.$$

By passing to the limit with $k \in J'$ tending to infinity we get, for all $x \in X$,

$$f_1(x) - [f_2(\hat{x}) + \langle \hat{g}_2, x - \hat{x} \rangle] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x - \hat{p}_i \right\|^2$$

$$\geq \lim_{J' \ni k \to \infty} \left[ f_k^{\mathrm{M}}(x^{k+1}) - \left[f_2(\hat{x}^k) + \left\langle \hat{g}_2^k, x^{k+1} - \hat{x}^k \right\rangle\right] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x^{k+1} - \hat{p}_i^k \right\|^2 \right]$$

$$= \varphi(\hat{x}),$$

i.e. $\min_{x \in X} \left\{ f_1(x) - [f_2(\hat{x}) + \langle \hat{g}_2, x - \hat{x} \rangle] + \sum_{i=1}^{m} \frac{\rho_i}{2} \left\| x - \hat{p}_i \right\|^2 \right\} \geq \varphi(\hat{x})$. As $\hat{x}$ is feasible for this problem, its optimal value is bounded from above by $\varphi(\hat{x})$. Hence, $\hat{x}$ solves this problem, which entails (from (3)) criticality of $\hat{x}$ to problem (2) (under Assumption 3.1). ■

We can now summarize these results by combining both propositions.

**Theorem 3.4 (Convergence):** *Under the assumptions of Propositions 3.2 and 3.3, Algorithm 2 with* `Tol` $= 0$ *produces a sequence* $\{\hat{x}^k\}_k$*, which if infinite has a cluster point satisfying condition (3) for problem (2) and if finite, has a last element satisfying this condition. Furthermore, if* `Tol` $> 0$ *the algorithm stops after finitely many iterations with an approximate critical point.*

***Proof:*** The result follows by combining Lemma 3.1 with Propositions 3.2 and 3.3. Since these propositions ensure that $\liminf_{k \to \infty} v_k^{\mathrm{M}} = 0$, the algorithm terminates after finitely many iterations, provided that `Tol` $> 0$. The last $\hat{x}^k$ is thus an approximate critical point, as can be seen from the optimality conditions of the subproblem (7). ■

## 4. Preliminary numerical experiments

This section experiments Algorithm 2 on the problem

$$\min_{x \in X \cap K} f(x), \quad \text{with } f(x) := \frac{1}{M} \sum_{m=1}^{M} \mathtt{OT}_m(x), \tag{11a}$$

where $\mathtt{OT}_m(x)$ is the optimal value of the *optimal transportation problem*

$$\mathtt{OT}_m(x) := \min_{\pi \in \mathbb{R}_+^{n \times S^m}} \sum_{i=1}^{n} \sum_{s=1}^{S^m} D_{is}^m \pi_{is} \quad \text{s.t.} \quad (\pi)^\top \mathbf{1}_n = q^m \quad \text{and} \quad \pi\, \mathbf{1}_{S^m} = x. \tag{11b}$$

In this notation, $D^m \in \mathbb{R}_+^{n \times S^m}$ is a given (distance) matrix and $q^m \in \mathbb{R}^{S^m}$ is a probability vector for all $m = 1, \ldots, M$. Furthermore, $X = \{x \in \mathbb{R}_+^n : \sum_{i=1}^n x_i = 1\}$ is the standard

simplex and $K = \{x \in \mathbb{R}^n : \|x\|_0 \leq \texttt{nnz}\}$ is the set of vectors with at most $\texttt{nnz}$ ($< n$) non-zero components. Note that for every $x \in X$, the linear programming problem (11b) admits a solution, which can be numerically computed using standard linear programming solvers or, in high-dimensional settings, via specialized methods such as the interior point algorithm proposed in [38]. We emphasize that $f$ is a nonsmooth convex function that is real-valued over $X$. The subdifferential of $f$ at a point $x \in X$ consists of vectors of the form $\frac{1}{M}\sum_{m=1}^{M} u_i^m$, where each $u_i^m$ is a Lagrange multiplier associated with the constraint $\pi\ \mathbf{1}_{S^m} = x$ in (11b). Since the set of Lagranange multipliers is unbounded, we normalize each computed $u_i^m$ to sum zero, as detailed in [39, Section 4].

When the constraint set $K$ is omitted, problem (11a) is referred to in the literature as the Wasserstein barycenter (WB) problem. It summarizes the given $M$ empirical measures, represented by the probability vectors $q^m$, $m = 1, \ldots, M$. This class of problems has been extensively studied in the machine learning and probability communities, and has found applications in areas such as clustering, applied probability, and image processing. We refer the interested reader to [39, 40] and the references therein.

The constrained Wasserstein barycenter problem arises when the additional constraint set $K$ is introduced. Such a set of constraints is particularly relevant in applications where the barycenter must adhere to specific structural or operational requirements, or align with prior knowledge about the desired properties of the barycentric measure [27]. In some applications (as in the image processing exercise below) the constraints represented by $K$ can be considered as soft. Indeed, the goal of having $K$ in the WB problem is to influence – not to impose - the geometry of the barycenter, its statistical properties, its support, or its physical suitability.

In our application, the specific choice of $K$ is to induce sparsity. We emphasize that the commonly employed strategy of promoting sparsity via the (convex) $\ell_1$-norm is not effective in our setting, as $\|x\|_1 = 1$ for all $x \in X$. Therefore, the (nonconvex) $\ell_0$-norm becomes crucial for computing a sparse WB, as discussed in [27].

In our numerical experiments, we are interested in summarizing $M = 100$ images $28 \times 28$ of handwritten digits from the MNIST database. As standard in the literature, we normalize and reshape the images into $M$ probability vectors $q^m \in \mathbb{R}^{784}$ (pixel values), and consider the matrix $D^m$ as the distance between the pixel positions as described in [39]. The number of decision variables is $n = 784$. Observe that the projection of $x$ onto $K$ is convenient to execute: it amounts to assigning zero to $n - \texttt{nnz}$ smallest components of $(|x|_1, \ldots, |x_n|)$. (Note that this projection is not unique; however, this is not an issue for our algorithms, which only require an arbitrary point in $\texttt{P}_K(x)$.) In our experiments, the number of non-zero components $\texttt{nnz}$ was set as follows: for every digit $\{0, 1, \ldots, 9\}$, we computed the number $a^m$ of non-zero pixels of each one of the $M = 100$ considered images, and set $\texttt{nnz} = \lfloor 0.45 \min_{m=1,\ldots,M} a^m \rfloor$.

We applied Algorithm 2 to the distance-to-set penalized formulation of problem (11a). Since the objective function in this problem is convex, we have $f = f_1$ and $f_2 \equiv 0$ in the algorithm's description. To compute a sparse barycenter of good quality, the algorithm was executed twice for each problem instance, with an increasing penalty parameter. In the first run, we set $\rho = 10$, initialized the algorithm with $x^0 = \mathbf{1}_n/n$, and used a tolerance parameter of $\texttt{Tol} = 10^{-2}$. The solution obtained from this initial run was then used as the starting point for a second run, where we set $\rho = 100$ and $\texttt{Tol} = 10^{-4}$. The

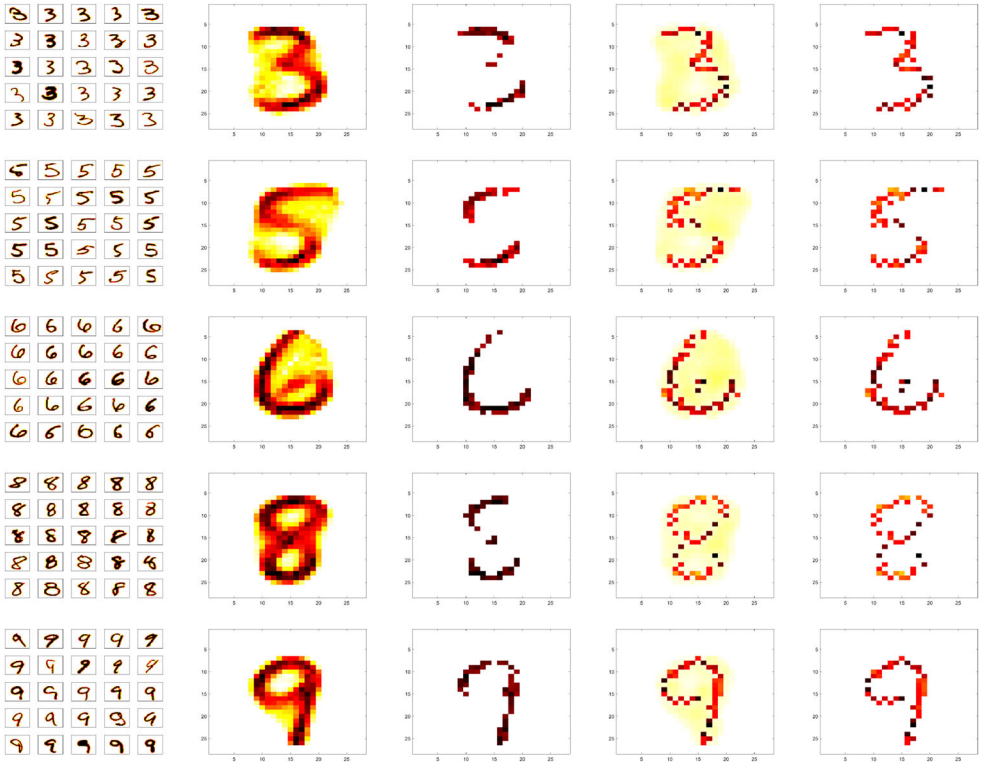

**Figure 1.** Comparison between the computed unconstrained and soft-constrained WBs. First column: 25 out of the 100 input images. Second column: unconstrained WBs. Third column: projection of the unconstrained WBs onto $X$. Fourth column: soft-constrained WBs (Algorithm 2). Fifth column: projection of the soft-constrained WBs onto $X$.

results presented below correspond to the second run. However, the reported number of iterations, serious steps, and CPU time reflect the cumulative values from both runs.

For comparison, we also computed an unconstrained Wasserstein barycenter for the same problem instance. To this end, we removed the nonconvex constraint set $K$ from formulation (11a), and applied the proximal bundle algorithm described in [2, p. 347], using the same stopping criterion as in Algorithm 2, with $\texttt{Tol} = 10^{-4}$. The algorithm was initialized with $x^0 = \mathbf{1}_n / n$ (as in the first run of Algorithm 2).

The numerical experiments were conducted in `MATLAB` (version 2022b) on a personal computer: 12th Gen Intel(R) Core(TM) i9 clocked at 2.5 GHz (64 GB RAM). We used `Gurobi` (version 10.0) to solve the optimal transport problems (11b). For the latter task, the subproblems were solved in parallel using 12 workers. Our `MATLAB` codes, dataset, and results are freely available for download at the following link: www.oliveira.mat.br/solvers.

Figure 1 presents our preliminary results. The first column displays 25 out of the 100 input images to be summarized. The second column shows the unconstrained Wasserstein barycenters (WBs), while their projections onto the simplex $X$ are shown in the third column. The fourth and fifth columns display the WBs computed using Algorithm 2 and their respective projections onto $X$.We do not report results for digits 0, 1, 2, 4, and 7, as the naive approach of computing an unconstrained Wasserstein barycenter and subsequently

**Table 1.** Results of Algorithm 2 applied to the WB problem (11a), with $K$ treated as a soft constraint (cumulative values for the two runs).

| Problem instance | Unconstrained WB | | Soft-constrained WB | | | |
|---|---|---|---|---|---|---|
| Digit | $f(\bar{x})$ | CPU(s) | $f(\bar{x})$ | #Iter | #SS | CPU(s) |
| 0 | 0.35467 | 27 | 0.37958 | 90 | 34 | 60 |
| 1 | 0.27455 | 9 | 0.32160 | 38 | 23 | 10 |
| 2 | 0.62541 | 49 | 0.65147 | 97 | 32 | 88 |
| 3 | 0.47859 | 38 | 0.50527 | 92 | 31 | 77 |
| 4 | 0.51716 | 30 | 0.54592 | 83 | 34 | 59 |
| 5 | 0.60131 | 33 | 0.62276 | 98 | 32 | 90 |
| 6 | 0.52126 | 36 | 0.54467 | 77 | 24 | 61 |
| 7 | 0.58986 | 23 | 0.62498 | 71 | 32 | 41 |
| 8 | 0.40300 | 42 | 0.43551 | 91 | 38 | 84 |
| 9 | 0.35376 | 23 | 0.38255 | 62 | 22 | 42 |

Note: For comparison, results for the unconstrained WB problem were obtained using a bundle method applied to the convex reformulation of problem (11a), where the constraint set $K$ is omitted.

projecting it onto $X$ already yielded satisfactory outcomes. This was not case of the digits reported in Figure 1.

The quality of the barycenters produced by our approach is clearly superior to that of the naive strategy (third column). This evidences the interest of employing soft (nonconvex) constraints to the classic (convex) WB formulation.

Table 1 reports some results obtained from applying Algorithm 2 to the WB problem (11a), with $K$ treated as a soft constraint. We recall that results for the unconstrained WB problem were obtained using a bundle method applied to the convex reformulation of problem (11a), where the constraint set $K$ is omitted. The table presents the function values and CPU time (in seconds) required to solve the unconstrained and sof-constrained constrained variants of the problem. We also report the number of iterations (#Iter) and number of serious steps (#SS) performed by Algorithm 2 to satisfy its stopping test (cumulative values for the two runs). We do not report this information for the bundle method applied to the unconstrained variant of the problem, as it is not relevant to our comparison. It is important to keep in mind that we are presenting results from two different solvers applied to two distinct optimization models, and this distinction should be considered when interpreting Table 1.

## 5. Concluding remarks

In this work, we proposed two new algorithms for solving problems in which the squared distance to a collection of sets is penalized. These algorithms are designed to accommodate a broad range of nominal problem data, offering considerable generality. We illustrated the numerical performance of one of our algorithms for computing sparse Wasserstein barycenters using a set of images from the MNIST database. As future work, we plan to implement and test the proposed methods on problems arising from the field of stochastic optimization, and benchmark their performance against existing approaches.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## ORCID

*W. van Ackooij* http://orcid.org/0000-0002-9943-3572

## References

[1] Bousnina D, de Oliveira W, Pflaum P. A stochastic optimization model for frequency control and energy management in a microgrid. In: Nicosia G, Ojha V, La Malfa E, et al., editors. Cham: Springer International Publishing; 2020. p. 177–189. (Machine Learning, Optimization, and Data Science).
[2] van Ackooij WS, de Oliveira WL. Methods of nonsmooth optimization in stochastic programming: from conceptual algorithms to real-world applications. 1st ed. Cham: Springer; 2025. (International Series in Operations Research & Management Science).
[3] Rockafellar RT, Wets RJB. Variational analysis. 3rd ed. Berlin: Springer Verlag; 2009. (Grundlehren der mathematischen Wissenschaften; Vol. 317).
[4] Cordova M, de Oliveira W, Sagastizábal C. Revisiting augmented Lagrangian duals. Math Program. 2022 Nov;196(1):235–277.
[5] Xu J, Chi EC, Lange K. Generalized linear model regression under distance-to-set penalties. In: NIPS 2017; Long Beach, CA, USA: Curran Associates, Inc.; 2017.
[6] Keys KL, Zhou H, Lange K. Proximal distance algorithms: theory and practice. J Mach Learn Res. 2019;20(66):1–38.
[7] Van Ngai H, Théra M. Error bounds in metric spaces and application to the perturbation stability of metric regularity. SIAM J Optim. 2008;19:1–20. doi: 10.1137/060675721
[8] Van Ngai H, Théra M. Error bounds for systems of lower semicontinuous functions in Asplund spaces. Math Program. 2009;116:397–427. doi: 10.1007/s10107-007-0121-9
[9] Kruger A, Van Ngai H, Théra M. Stability of error bounds for convex constraint systems in Banach spaces. SIAM J Optim. 2010;20(6):3280–3296. doi: 10.1137/100782206
[10] Van Ngai H, Kruger A, Théra M. Stability of error bounds for semi-infinite convex constraint systems. SIAM J Optim. 2010;20(4):2080–2096. doi: 10.1137/090767819
[11] Le Thi HA, Tao PD. DC programming and DCA: thirty years of developments. Math Program. 2018 May;169(1):5–68. doi: 10.1007/s10107-018-1235-y
[12] de Oliveira W. The ABC of DC programming. Set-Valued Var Anal. 2020 Dec;28(4):679–706. doi: 10.1007/s11228-020-00566-w
[13] Aragon Artacho F, Fleming R, Vuong P. Accelerating the DC algorithm for smooth functions. Math Program. 2018;169:95–118. doi: 10.1007/s10107-017-1180-1
[14] Ferreira OP, Santos EM, Souza JCO. A boosted DC algorithm for non-differentiable dc components with non-monotone line search. Comput Optim Appl. 2024 Jul;88(3):783–818. doi: 10.1007/s10589-024-00578-4
[15] de Oliveira W, Tcheou MP. An inertial algorithm for DC programming. Set-Valued Var Anal. 2019 Dec;27(4):895–919. doi: 10.1007/s11228-018-0497-0
[16] Pham TN, Dao MN, Amjady N, et al. A proximal splitting algorithm for generalized DC programming with applications in signal recovery. Eur J Oper Res. 2025;326(1):42–53.
[17] He S, Dong QL, Rassias MT. Contractive difference-of-convex algorithms. J Optim Theory Appl. 2025 May;206(1):12. doi: 10.1007/s10957-025-02689-2
[18] de Oliveira W, Souza JCdO. A progressive decoupling algorithm for minimizing the difference of convex and weakly convex functions. J Optim Theory Appl. 2025 Jan;204(3):36. doi: 10.1007/s10957-024-02574-4
[19] Cui Y, Pang JS, Sen B. Composite difference-max programs for modern statistical estimation problems. SIAM J Optim. 2018;28(4):3344–3374. doi: 10.1137/18M117337X
[20] Cui Y, Pang JS. Modern nonconvex nondifferentiable optimization. Philadelphia, PA: SIAM; 2022.
[21] Sempere GM, de Oliveira W, Royset JO. A proximal-type method for nonsmooth and nonconvex constrained minimization problems. J Optim Theory Appl. 2025 Feb;204(3):54. doi: 10.1007/s10957-024-02597-x

[22] Joki K, Bagirov AM, Karmitsa N, et al. A proximal bundle method for nonsmooth dc optimization utilizing nonconvex cutting planes. J Global Optim. 2017;68(3):501–535. doi: 10.1007/s10898-016-0488-3
[23] Gaudioso M, Giallombardo G, Miglionico G, et al. Minimizing nonsmooth DC functions via successive DC piecewise-affine approximations. J Global Optim. 2018;71:37–55. doi: 10.1007/s10898-017-0568-z
[24] de Oliveira W. Proximal bundle methods for nonsmooth DC programming. J Global Optim. 2019;75(2):523–563. doi: 10.1007/s10898-019-00755-4
[25] Tao PD, An LTH. Convex analysis approach to DC programming: theory, algorithms and applications. Acta Math Vietnam. 1997;22(1):289–355.
[26] Syrtseva K, de Oliveira W, Demassey S, et al. Minimizing the difference of convex and weakly convex functions via bundle method. Pac J Optim. 2024;20(4):699–741.
[27] Mimouni D, de Oliveira W, Sempere GM On the computation of constrained Wasserstein barycenters. Pac J Optim. 2025: 1–16. doi: 10.61208/pjo-2025-040
[28] Clarke F. Optimisation and nonsmooth analysis. Society for Industrial and Applied Mathematics; 1987. (Classics in Applied Mathematics).
[29] van Ackooij W, Pérez-Aros P, Soto C. Differentiability of probability functions involving star-shaped valued set-valued maps. SIAM J Optim. 2025;35(2):1216–1245. doi: 10.1137/24M1665465
[30] An LTH, Tao PD. The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. Ann Oper Res. 2005;133(1):23–46. doi: 10.1007/s10479-004-5022-1
[31] de Oliveira W. Short paper – a note on the frank–Wolfe algorithm for a class of nonconvex and nonsmooth optimization problems. Open J Math Optim. 2023;4:2. doi: 10.5802/ojmo.21
[32] Correa R, Lemaréchal C. Convergence of some algorithms for convex minimization. Math Program. 1993;62(1–3):261–275. doi: 10.1007/BF01585170
[33] van Ackooij W, Frangioni A. Incremental bundle methods using upper models. SIAM J Optim. 2018;28(1):379–410. doi: 10.1137/16M1089897
[34] de Oliveira W, Sagastizábal C, Lemaréchal C. Convex proximal bundle methods in depth: a unified analysis for inexact oracles. Math Prog Series B. 2014;148:241–277. doi: 10.1007/s10107-014-0809-6
[35] Frangioni A. Standard bundle methods: Untrusted models and duality. Cham: Springer International Publishing; 2020. Chapter 3; p. 61–116.
[36] van Ackooij W, Berge V, de Oliveira W, et al. Probabilistic optimization via approximate p-efficient points and bundle methods. Comput Oper Res. 2017;77:177–193. doi: 10.1016/j.cor.2016.08.002
[37] Frangioni A, Gorgone E. Generalized bundle methods for sum-functions with "easy" components: applications to multicommodity network design. Math Program. 2014;145(1):133–161. doi: 10.1007/s10107-013-0642-3
[38] Zanetti F, Gondzio J. An interior point–inspired algorithm for linear programs arising in discrete optimal transport. INFORMS J Comput. 2023;35(5):1061–1078. doi: 10.1287/ijoc.2022.0184
[39] Cuturi M, Doucet A. Fast computation of Wasserstein barycenters. In: Xing EP, Jebara T, editors. Proceedings of the 31st International Conference on Machine Learning; (Proceedings of Machine Learning Research; Vol. 32); 22–24 Jun; Bejing, China. PMLR; 2014. p. 685–693.
[40] Mimouni DW, Malisani P, Zhu J, et al. Computing Wasserstein barycenters via operator splitting: the method of averaged marginals. SIAM J Math Data Sci. 2024;6(4):1000–1026. doi: 10.1137/23M1584228