# Minimizing the difference of convex and weakly convex functions via bundle method

Ksenia Syrtseva, Welington de Oliveira, Sophie Demassey, Wim van Ackooij

# Minimizing the difference of convex and weakly convex functions via bundle method

Ksenia Syrtseva
ksenia-k.syrtseva@edf.fr
Mines Paris, Université PSL, Centre de Mathématiques Appliquées (CMA), Sophia Antipolis, France
EDF R&D, Boulevard Gaspard Monge, Palaiseau, France

Welington de Oliveira, Sophie Demassey
Mines Paris PSL, CMA

Wim van Ackooij
EDF R&D

**Abstract:** We consider optimization problems with objective and constraint being the difference of convex and weakly convex functions. This framework covers a vast family of nonsmooth and nonconvex optimization problems, particularly those involving certain classes of composite and nonconvex value functions. We investigate several stationary conditions and extend the proximal bundle algorithm of [van Ackooij et al., Comput. Optim. Appl., 78 (2021), pp. 451–490 ] to compute critical points for problems in this class. Our modifications on that algorithm boil down to a different master program and an original rule to update the proximal parameter to ensure convergence in this more general setting. Thanks to this new rule, no pre-estimation of the underlying weakly-convex moduli is needed, opening the way to deal with optimization problems for which no practical and mathematically sound algorithms exist. Numerical experiments on some nonconvex stochastic problems illustrate the practical performance of the method.

**Keywords:** Non-smooth Optimization, DC Programming, Bundle Methods, weakly convex
**Mathematics Subject Classification:** 90C26, 65K05, 49J52, 49J53

# 1 Introduction.

This work presents a bundle method for nonsmooth and nonconvex optimization problems of the form

$$\min_{x\in X} f(x) \quad \text{s.t.} \quad c(x) \leq 0, \tag{1.1a}$$

where $X$ is a nonempty bounded polyhedron contained in an open convex set $\mathcal{O} \subset \mathbb{R}^n$, and functions $f : \mathcal{O} \to \mathbb{R}$ and $c : \mathcal{O} \to \mathbb{R}$ are decomposable as the difference of *convex* and (locally) *weakly convex* functions. More specifically, we assume that the following convex-weakly convex (CwC) decompositions are available:

$$f(x) = f_1(x) - f_2(x) \quad \text{and} \quad c(x) = c_1(x) - c_2(x), \tag{1.1b}$$

with $f_1, c_1 : \mathcal{O} \to \mathbb{R}$ convex and $f_2, c_2 : \mathcal{O} \to \mathbb{R}$ *weakly convex functions on some neighbourhood of each* $x \in \mathcal{O}$. We adopt the more general definition of weakly convex functions (see Definition 2.2 below) given in [50, Def. 4.2] so that we can exploit the equivalence between the families of locally weakly convex and Lower-$C^2$ functions [33, Thm. 1.3, Cor. 1.3] to highlight the breadth of our approach. In particular, we have in mind the following settings for $f_2$ (as well as for $c_2$):

i) $f_2(x) = \phi(x)$ is a (possibly nonsmooth) convex function;

ii) $f_2(x) = -h(x)$ with $h$ having Lipschitz continuous gradient;

iii) $f_2(x) = \phi(x) - h(x)$, with $\phi$ and $h$ as given above;

iv) $f_2(x)$ is the optimal value of $\max_{t\in T} F(t, x)$, with $T$ a (possibly nonconvex) compact set and $F$ of class $C^2$;

v) $f_2(x) = \phi(G(x))$, with $\phi : \mathbb{R}^m \to \mathbb{R}$ convex and Lipschitz and $G : \mathbb{R}^n \to \mathbb{R}^m$ a smooth mapping with Lipschitz Jacobian.

Analogous settings for $c_2$, and their combinations with the ones for $f_2$, are covered by our analysis (see Section 2 below for details).

Weakly convex functions enjoy favorable properties in so much as that they can be recast as *Difference-of-Convex* (DC) functions [10]. Hence, problem (1.1) can, in theory, be recast as a DC-constrained DC program, a setting that proves *practical if explicit DC decompositions are available*; see for instance [17, 16, 34, 24, 8, 49, 41, 35] and references therein. However, if no DC decomposition is known for $f$ or $c$, the DC machinery is unsuitable, and the methods proposed in these references are not applicable. This is already the case for the more straightforward items ii) and iii) above if the underlying Lipschitz constant is unknown and no upper bound is readily available[1]. The situation becomes even more complicated for items iv) and v): in general, there are no formulae, rules, or practical insights to obtain a DC decomposition for $f_2$ in these cases (see Example 1.1 below for a particular case of iv). A strategy to handle problem (1.1) via DC programming algorithms is to replace functions $f_i$ and $c_i$ $(i = 1, 2)$ with $f_i(x) + \frac{\mu}{2}\|x\|^2$ and $c_i(x) + \frac{\mu}{2}\|x\|^2$ for a large parameter $\mu > 0$ estimating upper bounds on the unknown weakly-convex moduli $\mu_f$ and $\mu_c$ of $f_2$ and $c_2$ (see Proposition 2.4), hoping that $f_2(x) + \frac{\mu}{2}\|x\|^2$ and $c_2(x) + \frac{\mu}{2}\|x\|^2$ are convex on $X$. As, in general, there is no reliable way to assert the convexity of these latter functions, DC programming algorithms applied in this context must be understood as heuristics. Remarkably, the work [39] exploits such a strategy by combining a dynamic rule to update $\mu$ with a nonconvexity test so that convergence is achieved, but only in a probability sense. Differently, for a class of nonconvex two-stage stochastic problems, the authors of [20] exploit an implicitly convex-concave structure of the objective function and propose an algorithm based on the so-called partial Moreau envelope that disregards DC decompositions at the price of nonnegligible computational costs.

In contrast to the above references, this work investigates a bundle method approach for tackling (1.1), which neither requires explicit DC decompositions of the involved functions (in particular, bounds on the weakly-convex moduli $\mu_f$ and $\mu_c$ need not be known), nor relies on (often costly) Moreau envelopes. For the method to work, it suffices to dispose of a difference of *convex and weakly convex* (CwC) decomposition of the involved functions, as in (1.1b). Compared to DC, the latter structure appears more naturally in applications (see [15, § 7.5]) and has yet to be exploited to design optimality conditions and numerical algorithms. This work aims to fill this gap.

Our approach broadens and enhances the method proposed in [49] for dealing with DC-constrained DC-problems in the following two ways. First, the availability of DC decompositions is no longer needed, which makes our approach applicable to a larger scope of problems in practice. Second, it is ensured to compute critical points for the original problem without any additional assumption on weakly-convex components ($f_2$ and $c_2$ need not be continuously differentiable as assumed in [49, Thm. 2]). In addition, it has a lower cost per iteration (the master subproblem has fewer constraints than the one of [49]). Similarly to [49], our approach builds upon a problem reformulation via improvement function, a well-known and successful strategy in the nonsmooth optimization literature [31, 1, 24]. However, due to the above modifications, the convergence analysis of our extension of the method proposed in [49] must be done anew. Furthermore, a new criticality definition for the reformulated problem links directly with (necessary) optimality conditions for the original problem (1.1), which makes it a major ingredient for these enhancements. Such a criticality concept is introduced and analyzed in Section 3 below, where we also extend the alternative characterization of Bouligand stationarity given in [27] to our CwC setting. Before that, we motivate this work with the following example that presents a class of problems (of great practical appeal) where the CwC decomposition arises upon applying a well-known interior-penalty strategy. An example of a real-life (chance-constrained optimal power flow) problem fitting our CwC structure without any approximation can be found in the Ph.D. thesis [15, § 7.5].

**Example 1.1** (Nonconvex two-stage programming)**.** Let $\Xi := \{\xi^1, \dots \xi^S\}$ be a set of scenarios and $\pi_s > 0$ the probability of occurrence of event $\xi^s$, $s = 1, \dots, S$. Consider the following nonconvex two-stage program

$$\begin{cases} \min\limits_{x \in X} & f_1(x) + \sum\limits_{s=1}^{S} \pi_s Q(x; \xi^s) \\ \text{s.t.} & c_1(x) - c_2(x) \leq 0 \end{cases} \quad \text{with} \quad Q(x;\xi) := \begin{cases} \min\limits_{y \in Y} & q(x, y; \xi) \\ \text{s.t.} & \psi_i(x, y; \xi) \leq 0, \ i = 1, \dots, m. \end{cases} \tag{1.2}$$

Assume that:

- $f_1, c_1, c_2 : \mathbb{R}^n \to \mathbb{R}$ are convex (possibly nonsmooth) functions;
- $X \subset \mathbb{R}^n$, $Y \subset \mathbb{R}^{n_2}$ are two (non-empty) convex and bounded polytopes;
- $q, \psi_i : \mathbb{R}^n \times \mathbb{R}^{n_2} \times \Xi \to \mathbb{R}$, $i = 1, \dots, m$, possess the following characteristics: $q(\cdot, \cdot, \xi)$ and $\psi_i(\cdot, \cdot, \xi)$ are twice-continuously differentiable for every $\xi \in \Xi$ fixed and, moreover, $q(x, \cdot, \xi)$ and $\psi_i(x, \cdot, \xi)$ are convex for every $x$ and $\xi$ fixed;

[1] It is worth noting that in many practical problems, mainly those from data science, an upper bound on such a constant can be computed.

- the constraints in the subproblem $Q(x;\xi)$ satisfy the Slater condition: for every $x \in X$ and $\xi \in \Xi$, there exists $y^\circ(x;\xi) \in Y$ such $\psi_i(x, y^\circ(x;\xi), \xi) < 0$, $i = 1, \ldots, m$.

As presented in [49], the DC constraint $c_1(x) - c_2(x) \leq 0$ above is particularly useful in this stochastic programming setting to model chance constraints.

Under the above assumptions, evaluating the recourse function $Q(x;\xi)$ amounts to solving a well-defined convex optimization problem on variable $y$. Although this essential property is present, the recourse function itself fails to be convex on variable $x$ (but $Q(\cdot;\xi)$ is continuous as a result of [2, Prop. 4.4]). Furthermore, without further assumptions, computing a (generalized) subgradient of $Q(\cdot;\xi)$ at $x$ as well as asserting additional properties about this function are challenging tasks. This could for instance be done if the constraints satisfy a further Aubin or Lipschitz like property upon exploiting [25, Chapter 4]. Still though, most likely, at best we would be dealing with subdifferentials inclusions - and concrete algorithms to handle such general "marginal functions" would be unavailable.

A possible manner to curtail these difficulties is to approximate the recourse function with a more tractable one. As explained in [3], with the help of the log-barrier penalty function and a penalization parameter $\varepsilon > 0$, we may approximate $Q(x;\xi)$ with

$$Q^\varepsilon(x;\xi) := \min_{y \in Y} \; q(x,y;\xi) - \frac{1}{\varepsilon} \sum_{i=1}^{m} \log(-\psi_i(x,y;\xi)). \tag{1.3}$$

Given the above assumptions, it is well known that $Q^\varepsilon(x;\xi) \downarrow Q(x;\xi)$ as $\varepsilon \downarrow 0$ (e.g., [3, § 2.2] and [2, p. 266]), and thus the model

$$\left\{ \begin{array}{ll} \min\limits_{x \in X} & f_1(x) - f_2(x) \\ \text{s.t.} & c_1(x) - c_2(x) \leq 0 \end{array} \right. \qquad \text{with} \quad f_2(x) := \sum_{s=1}^{S} \pi_s[-Q^\varepsilon(x;\xi^s)]$$

is an accurate approximation of (1.2) when $\varepsilon > 0$ is small enough. Furthermore, as $-Q^\varepsilon(x;\xi) = \max_{y \in Y} \; \frac{1}{\varepsilon} \sum_{i=1}^{m} \log(-\psi_i(x,y;\xi)) - q(x,y;\xi)$ is a weakly convex function (c.f. item iv) above), this model fits the structure (1.1). We highlight that $Q^\varepsilon(x;\xi)$ is generally a nonsmooth (nonconvex) function; hence, the above problem is challenging. To our knowledge, no practical and mathematically sound optimization algorithm could tackle this class of problems before this work. Indeed, [23] requires $f_2$ to be smooth, [20] assumes $q(\cdot,\cdot,\xi)$ and $\psi_i(\cdot,\cdot,\xi)$ to be concave-convex functions, and [3] requires another degree of approximation by adding a Tikhonov regularization to (1.3) to force $Q^\varepsilon(x;\xi)$ be smooth. In all these references, function $c_2$ is absent. Being nonsmooth, we mention in passing that a (generalized) subgradient of $f_2$ at $x$ can be computed and seen to be $\sum_{s=1}^{S} \pi_s g(y(x;\xi^s))$, where $g(\cdot) := \nabla_x[\frac{1}{\varepsilon} \sum_{i=1}^{m} \log(-\psi_i(x,\cdot;\xi^s)) - q(x,\cdot;\xi^s)]$ is the gradient w.r.t. $x$ of the objective function of (1.3) multiplied by $-1$, and $y(x;\xi)$ is an arbitrary optimal solution of (1.3) (see Proposition 2.1 and [30, Thm. 7.3]). □

Our approach is still applicable in more general case, where the probability vector $\pi$ in the above example is a function (of class $C^2$) of the first-stage variable $x$, i.e., $\pi_s(x)$, $i = 1, \ldots, S$. Hence, this work's class of optimization problems includes the challenging family of stochastic programming recourse models with decision-dependent uncertainty considered, for instance, in [12] and [20].

The remaining of this manuscript is organized as follows. Section 2 recalls essential definitions, key elements, and well-known concepts from variational analysis. Necessary optimality conditions for problem (1.1) are presented in Section 3 as well as the problem reformulation via an improvement function. Once the link between the reformulated and the original problem is established in the same section, Section 4 focuses on an improvement-function-based bundle method for problem (1.1). Section 5 presents the method's convergence analysis to critical points, and finally, Section 6 illustrates the practical performance of our approach on some nonconvex stochastic optimization problems.

**Notation** The following notation is employed throughout the text. For a real number $a$, we denote by $[a]_+$ the value $\max\{a, 0\}$. For any points $x, y \in \mathbb{R}^n$, $\langle x, y\rangle$ stands for the Euclidean inner product, and $\|\cdot\|$ for the associated norm, i.e., $\|x\| = \sqrt{\langle x, x\rangle}$. For a convex set $X$, $N_X(x)$ stands for its normal cone at the point $x$, i.e., the set $\{y : \langle y, z - x\rangle \leqslant 0 \;\text{ for all }\; z \in X\}$ if $x \in X$ and the empty set otherwise. The Bouligand tangent cone to a (possibly nonconvex) set $W \subset \mathbb{R}^n$ at a point $w \in W$ is the set $\mathcal{T}_W(w)$ of all tangent directions in the following sense: $d \in \mathcal{T}_W(w)$ if there exist a sequence of vectors $\{w^k\} \subset W$ converging to $w$ and a sequence of positive scalars $t_k \to 0$ such that $d = \lim_{k\to\infty}(w^k - w)/t_k$. The indicator function of $X \subset \mathbb{R}^n$ is defined as $i_X(x) = 0$ if $x \in X$ and $i_X(x) = +\infty$ otherwise. The convex hull of a set $X$ is $\texttt{conv}X$ and the relative interior is denoted by $\texttt{ri}\, X$. The domain of a function $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ is represented by $\mathcal{D}\text{om}(\varphi) = \{x \in \mathbb{R}^n : \varphi(x) < +\infty\}$. Notation $\mathcal{O}$ stands for an open convex set of the Euclidean space $\mathbb{R}^n$ and, given the definitions of $f$ and $c$, we have that $\mathcal{O} \subset \mathcal{D}\text{om}(f)$ and $\mathcal{O} \subset \mathcal{D}\text{om}(c)$. The component functions of $f$ and $c$ are $f_1, f_2$, and $c_1, c_2$ respectively: $f_1$ and $c_1$ are convex, whereas $f_2$ and $c_2$ are weakly convex on some neighbourhood of every $x \in \mathcal{O}$. Finally, $f^*$ stands for the Legendre-Fenchel transform of a function $f : \mathbb{R}^n \to (-\infty, +\infty]$.

## 2 Definition and prerequisites

This section starts by recalling the concept of (generalized) directional derivatives and subdifferentials. Basic subdifferential calculus is summarized in Proposition 2.1 below, followed by the definitions of weakly convex and lower-$C^2$ functions. The section closes with Proposition 2.4 asserting that the definition of (locally) weakly convex function can be globally extended to the whole convex and compact set $X$. Such a property is of crucial importance in this work.

A function $\mathsf{f} : \mathcal{O} \to \mathbb{R}$ is said to be *locally Lipschitz continuous* if for each $x' \in \mathcal{O}$ there is a neighbourhood $V_{x'} \subset \mathcal{O}$ of $x'$ such that, for some $L_{x'} \geq 0$,

$$|\mathsf{f}(x) - \mathsf{f}(y)| \leq L_{x'} \|x - y\| \quad \forall\, x, y \in V_{x'}.$$

The function $\mathsf{f}$ is said to be *Lipschitz continuous* on $\mathcal{O}$ if $L_{x'} = L$ can be taken independent of $x' \in \mathcal{O}$, and $V_{x'}$ in the above inequality is replaced with $\mathcal{O}$.

**Directional derivatives and subdifferentials.** Let $\phi : \mathcal{O} \to \mathbb{R}$ be a convex function. Then $\phi$ is locally Lipschitz continuous and, for each $x \in \mathcal{O}$, the directional derivative

$$\phi'(x; d) := \lim_{\tau \downarrow 0} \frac{\phi(x + \tau d) - \phi(x)}{\tau}$$

exists (and is finite) in every direction $d \in \mathbb{R}^n$ [26, Prop. 2.81 and Cor. 2.82]. Such a derivative can be represented by

$$\phi'(x; d) = \max_{s \in \partial\phi(x)} \langle s, d \rangle,$$

where $\partial\phi(x)$ is the *subdifferential* of $\phi$ at point $x$:

$$\partial\phi(x) := \{s \in \mathbb{R}^n : \phi(y) \geq \phi(x) + \langle s, y - x \rangle \quad \forall\, y \in \mathbb{R}^n\}. \tag{2.1}$$

The elements of $\partial\phi(x)$ are referred to as the *subgradients* of $\phi$ at $x$. The *approximate subdifferential* is defined, for $\epsilon \geq 0$, by

$$\partial_\epsilon\phi(x) := \{s \in \mathbb{R}^n : \phi(y) \geq \phi(x) + \langle s, y - x \rangle - \epsilon \quad \forall\, y \in \mathbb{R}^n\}.$$

Let $\mathsf{f} : \mathcal{O} \to \mathbb{R}$ be a locally Lipschitz continuous function. Then the generalized directional derivative defined by

$$\mathsf{f}^\circ(x; d) := \limsup_{x' \to x,\ \tau \downarrow 0} \frac{\mathsf{f}(x' + \tau d) - \mathsf{f}(x')}{\tau}$$

is finite for all $x \in \mathcal{O}$ in every direction $d \in \mathbb{R}^n$ [4, Prop. 2.1.1(a)]. Such a mathematical concept permits us to define the *Clarke subdifferential* of $\mathsf{f}$ at $x \in \mathcal{O}$,

$$\partial^{\mathsf{C}}\mathsf{f}(x) := \{g \in \mathbb{R}^n : \langle g, d \rangle \leq \mathsf{f}^\circ(x; d) \ \text{ for all } \ d \in \mathbb{R}^n\}, \tag{2.2}$$

which is a nonempty, convex, and compact subset of $\mathbb{R}^n$ [4, Prop. 2.1.2(a)] satisfying

$$\mathsf{f}^\circ(x; d) = \max_{g \in \partial^{\mathsf{C}}\mathsf{f}(x)} \langle g, d \rangle.$$

The elements of $\partial^{\mathsf{C}}\mathsf{f}(x)$ are referred to as *generalized (or Clarke) subgradients*, as they are the usual subgradients, i.e., $\partial^{\mathsf{C}}\mathsf{f} = \partial\mathsf{f}$, when $\mathsf{f}$ is convex [4, Prop. 2.2.7]. Furthermore, when $\mathsf{f}$ is continuously differentiable, $\partial^{\mathsf{C}}\mathsf{f}(x)$ reduces to the singleton $\{\nabla\mathsf{f}(x)\}$. An alternative representation, in finite dimensions, of $\partial^{\mathsf{C}}\mathsf{f}(x)$ is (see [4, Thm. 2.5.1])

$$\partial^{\mathsf{C}}\mathsf{f}(x) := \mathtt{conv}\left\{\lim_{\iota \to \infty} \nabla\mathsf{f}(x_\iota), x_\iota \to x, \mathsf{f} \text{ differentiable at } x_\iota\right\}.$$

A fundamental result, often evoked in this work, is the following one [4, Prop. 2.1.2]: the mapping $\partial^{\mathsf{C}}\mathsf{f}$ is locally bounded in the interior of $\mathcal{D}\mathrm{om}(\mathsf{f}) := \{x \in \mathbb{R}^n : \mathsf{f}(x) < \infty\}$. As a result, the image $\partial^{\mathsf{C}}\mathsf{f}(X)$ of every bounded set $X \subset \mathcal{O}$ ($\subset \mathcal{D}\mathrm{om}(\mathsf{f})$) is bounded in $\mathbb{R}^n$. Useful calculus rules of subdifferentials are listed in Proposition 2.1 below and rely on the concept of regularity.

A locally Lipschitz continuous function $\mathsf{f} : \mathcal{O} \to \mathbb{R}$ is *subdifferentially regular* (or simply regular) at $x \in \mathcal{O}$ if for every $d \in \mathbb{R}^n$ the ordinary directional derivative at $x$ exists and coincides with the generalized one:

$$\mathsf{f}'(x; d) = \mathsf{f}^\circ(x; d) \quad \forall\, d \in \mathbb{R}^n.$$

It holds that smooth functions, as well as convex ones, are regular at every point in the interior of their domains. Moreover, a finite linear combination (by nonnegative scalars) of regular functions at $x$ is regular [4, Prop. 2.3.6].

**Proposition 2.1.** *Let $\mathsf{f}_t : \mathcal{O} \to \mathbb{R}$, $t = 1, 2, \dots, m$, be locally Lipschitz functions and $x \in \mathcal{O}$ an arbitrary point. Then*

*i) $\partial^{\mathsf{C}}[\sum_{t=1}^m a_t \mathsf{f}_t](x) \subset \sum_{t=1}^m a_t \partial^{\mathsf{C}} \mathsf{f}_t(x)$ for all $a \in \mathbb{R}^m$, and equality holds if*

- *all but one of $\mathsf{f}_t$ are smooth [4, Prop. 2.3.3 and Cor. 2];*
- *or if every $\mathsf{f}_t$ is regular at $x$ and $a \in \mathbb{R}^m_+$ [4, Cor. 3];*

*ii) $\partial^{\mathsf{C}} \mathsf{f}(x) \subset \mathtt{conv}\{\partial^{\mathsf{C}} \mathsf{f}_t(x) : t \in I(x)\}$, for $\mathsf{f}(x) = \max_{t=1,\dots,m} \mathsf{f}_t(x)$ and $I(x) := \arg\max_{t=1,\dots,m} \mathsf{f}_t(x)$, and equality holds and $f$ is regular if every $\mathsf{f}_t$ is regular at $x$. [4, Prop. 2.3.12].*

The last item can be strengthened when more structure is present, such as in the case of weakly convex functions (see Eq. (2.3) below).

## Weakly convex functions: definition and main properties

**Definition 2.2** (Def. 4.2 [50])**.** A function $\mathsf{f} : \mathcal{O} \to \mathbb{R}$ is said to be (locally) *weakly convex* on $\mathcal{O}$ if, on some neighbourhood $V_{x'} \subset \mathcal{O}$ of each $x' \in \mathcal{O}$, there exists $\mu_{x'} \geq 0$ such that, for all $\mu \geq \mu_{x'}$

$$\phi(x) := \mathsf{f}(x) + \frac{\mu}{2}\|x\|^2 \quad \text{is finite and convex on } V_{x'}.$$

Furthermore, $\mathsf{f}$ is said to be *weakly convex in the global sense* on $\mathcal{O}$ if the above property holds for $V_{x'} = \mathcal{O}$ and $\mu_{x'} = \bar{\mu} \geq 0$ regardless of $x' \in \mathcal{O}$. □

Clearly, a convex function on $\mathcal{O}$ is weakly convex in the global sense on $\mathcal{O}$: it suffices to take $\mu_{x'} = 0$ and $V_{x'} = \mathcal{O}$ for all $x' \in \mathcal{O}$. When $\mathsf{f}$ is a smooth function with a Lipschitz continuous gradient, then $\mathsf{f}$ is weakly convex in the global sense with $\mu = L$ the Lipschitz constant of $\nabla \mathsf{f}$ [6, Prop. 1]. Moreover, it follows from [7, Lemma 4.2] that the family of composite functions given in item v) of the Introduction is also weakly convex in the global sense.

Definition 2.2 implies that weakly convex functions are *locally DC*: the decomposition $\mathsf{f}(x) = \phi(x) - \frac{\mu}{2}\|x\|^2$ holds on some neighbourhood of every $x' \in \mathcal{O}$. As a result, [29, Thm. 10.33] ensures that the class of weakly convex functions coincide with that of Lower-$C^2$ functions; see also [33, Thm. 1.3, Cor. 1.3].

**Definition 2.3** (Def. 10.29 [29])**.** ($\mathsf{LC}^2$ functions). A function $\mathsf{f} : \mathcal{O} \to \mathbb{R}$ is said to be *Lower-$C^2$* or $\mathsf{LC}^2$ on $\mathcal{O}$ if, on some neighbourhood $V_{x'} \subset \mathcal{O}$ of each $x' \in \mathcal{O}$, there is a representation

$$\mathsf{f}(x) = \max_{t \in T} \mathsf{f}_t(x).$$

in which the functions $f_t$ are of differentiability class $C^2$ on $V_{x'}$ and the index set $T$ is a compact space such that $\mathsf{f}_t(x)$, $\nabla \mathsf{f}_t(x)$, and $\nabla^2 \mathsf{f}_t(x)$ depend continuously not just on $x \in V_{x'}$ but jointly on $(t, x) \in T \times V_{x'}$. □

In particular, if $\mathsf{f}$ is given by $\mathsf{f}(x) = \max\{\mathsf{f}_1(x), \dots, \mathsf{f}_m(x)\}$ and all functions $\mathsf{f}_1, ..., \mathsf{f}_m$ are of class $C^2$, then $\mathsf{f}$ is Lower-$C^2$/weakly convex. Furthermore, the functions of item iv) are also weakly convex, since they are Lower-$C^2$ by definition.

An important property of $\mathsf{LC}^2$/weakly convex functions is regularity [28, Thm. 1]: for every $x \in \mathcal{O}$, the equality $\mathsf{f}'(x; d) = \mathsf{f}^\circ(x; d)$ holds in every direction $d \in \mathbb{R}^n$. Furthermore, Theorem 7.3 in [30] gives the following characterization of the Clarke subdifferential of $\mathsf{f}$ at $x \in \mathcal{O}$: for $I(x) = \arg\max_{t \in T} \mathsf{f}_t(x)$,

$$\partial^{\mathsf{C}} \mathsf{f}(x) = \mathtt{conv}\,\{\nabla_x \mathsf{f}_t(x) \,:\, t \in I(x)\} \quad \text{for all } x \in \mathcal{O}. \tag{2.3}$$

When constrained to a compact convex set $X \neq \emptyset$, we can say more about weakly convex functions. Indeed, the local property in Definition 2.2 globally extends to the whole $X$, and we have the following key result (whose proof can be found in the Appendix A).

**Proposition 2.4.** *Let $\mathsf{f} : \mathcal{O} \to \mathbb{R}^n$ be a weakly convex function, and $X \subset \mathcal{O}$ a compact and convex set. Then there exist a real number $\mu_{\mathsf{f}} \geq 0$ and an open convex set $\mathcal{O}'$ satisfying $X \subset \mathcal{O}' \subset \mathcal{O}$ such that, for all $\mu \geq \mu_{\mathsf{f}}$:*

*i) the function $\phi(x) := \mathsf{f}(x) + \frac{\mu}{2}\|x\|^2$ is convex on $\mathcal{O}'$ and $\partial\phi(x) = \partial^{\mathsf{C}} \mathsf{f}(x) + \mu x$ for all $x \in \mathcal{O}'$;*

*ii) for all $s_f \in \partial^{\mathsf{C}} f(x)$ with $x \in \mathcal{O}'$, the following inequality holds*

$$\mathsf{f}(y) \geq \mathsf{f}(x) + \langle s_f, y - x\rangle - \frac{\mu}{2}\|y - x\|^2 \quad \forall y \in X. \tag{2.4}$$

Concerning the setting of this work where $X$ is compact, the appealing DC decomposition $\mathsf{f}(x) = \phi(x) - \frac{\mu_{\mathsf{f}}}{2}\|x\|^2$ is, unfortunately, unavailable: the threshold $\mu_{\mathsf{f}}$ in Proposition 2.4 is in general unknown. This fact precludes the application of DC techniques to optimization problems featuring general Lower-$C^2$/weakly convex functions. Interested readers are referred to [40] for a strategy that uses approximated DC decompositions based on item i) of Proposition 2.4.

# 3 Necessary optimality conditions and problem reformulation

Let $f, c : \mathcal{O} \to \mathbb{R}$ be given by (1.1b). We highlight that some properties of their components can be transferred to these functions. (To ease the presentation, let us focus only on $f(x) = f_1(x) - f_2(x)$, as the same conclusions hold for $c$.) For instance, $f$ is locally Lipschitz continuous because $f_1$ and $f_2$ are so. Furthermore, as $f_1$ is convex and $f_2$ is (locally) weakly convex, they are both directional differentiable and these properties extend to $f$ as well: for every $x \in \mathcal{O}$, the directionally derivative of $f$ is finite in every direction $d \in \mathbb{R}^n$ as result of the following relation:

$$\begin{aligned} f_1'(x;d) - f_2'(x;d) &= \lim_{\tau \downarrow 0} \frac{f_1(x+\tau d) - f_1(x)}{\tau} - \lim_{\tau \downarrow 0} \frac{f_2(x+\tau d) - f_2(x)}{\tau} \\ &= \lim_{\tau \downarrow 0} \left[ \frac{f_1(x+\tau d) - f_1(x)}{\tau} - \frac{f_2(x+\tau d) - f_2(x)}{\tau} \right] \\ &= \lim_{\tau \downarrow 0} \frac{f(x+\tau d) - f(x)}{\tau} = f'(x;d). \end{aligned}$$

However, the important regularity condition of both $f_1$ and $f_2$ does not extend to $f$ as a mere fact that the latter is not a linear combination with nonnegative coefficients of the two former functions (see Proposition 2.1.i)). Hence, we cannot expect to have equality in the following inclusion

$$\partial^{\mathsf{C}} f(x) \subset \partial f_1(x) - \partial^{\mathsf{C}} f_2(x)$$

unless one of the component functions is smooth at $x$. (Here we have used the fact that $\partial^{\mathsf{C}} f_1 = \partial f_1$ due to convexity of $f_1$.) Such an inclusion impacts stationary concepts as we will now discuss. Let us first consider the convexly-constrained problem

$$\min_{x \in X} \; f(x), \quad \text{with} \quad f(x) = f_1(x) - f_2(x). \tag{3.1}$$

A point $\bar{x} \in X$ is said to be directional ($d$)-stationary for this problem if $f'(\bar{x};d) \geq 0$ for all $d \in \mathcal{T}_X(\bar{x})$. Here, $\mathcal{T}_X(\bar{x})$ is the Bouligand tangent cone to $X$ at a point $\bar{x} \in X$. The following result generalizes [27, Prop. 5], where a specific case of problem (3.1) with $f_2(x) = \max\{\psi_1(x), \ldots, \psi_m(x)\}$ and convex $\psi_1, \ldots, \psi_m$ is considered.

**Proposition 3.1.** *A point $\bar{x} \in X$ is d-stationary of problem (3.1) if, and only if,*

$$\bar{x} \in \arg\min_{x \in X} \; f_1(x) - [f_2(\bar{x}) + \langle s_{f_2}, x - \bar{x} \rangle] \quad \forall\, s_{f_2} \in \partial^{\mathsf{C}} f_2(\bar{x}).$$

*Proof.* Observe that $\mathcal{T}_X(\bar{x}) = \mathtt{cl}\,\{d \in \mathbb{R}^n : d = t(x - \bar{x}),\ x \in X,\ t \in \mathbb{R}_+\}$ due to convexity of $X$. Therefore, the definition of $d$-stationarity can be equivalently written as $f'(\bar{x}; x - \bar{x}) \geq 0$ for all $x \in X$. Recall that $f_2$ is (locally) weakly convex and thus regular: $f_2'(\bar{x}; x - \bar{x}) = f_2^{\circ}(\bar{x}; x - \bar{x})$, which implies that $f_2'(\bar{x}; x - \bar{x}) = \max_{s_{f_2} \in \partial^{\mathsf{C}} f_2(\bar{x})} \langle s_{f_2}, x - \bar{x} \rangle$. Hence,

$$\begin{aligned} f'(\bar{x}; x - \bar{x}) = f_1'(\bar{x}; x - \bar{x}) - f_2'(\bar{x}; x - \bar{x}) \geq 0 \qquad & \forall\, x \in X \\ \Leftrightarrow\ f_1'(\bar{x}; x - \bar{x}) \geq f_2'(\bar{x}; x - \bar{x}) \qquad & \forall\, x \in X \\ \Leftrightarrow\ f_1'(\bar{x}; x - \bar{x}) \geq \langle s_{f_2}, x - \bar{x} \rangle \qquad & \forall\, s_{f_2} \in \partial^{\mathsf{C}} f_2(\bar{x}),\ \forall\, x \in X \\ \Leftrightarrow\ f_1'(\bar{x}; x - \bar{x}) - \langle s_{f_2}, x - \bar{x} \rangle \geq 0 \qquad & \forall\, s_{f_2} \in \partial^{\mathsf{C}} f_2(\bar{x}),\ \forall\, x \in X \\ \Leftrightarrow\ \bar{x} \in \arg\min_{x \in X} f_1(x) - \langle s_{f_2}, x - \bar{x} \rangle \qquad & \forall\, s_{f_2} \in \partial^{\mathsf{C}} f_2(\bar{x}). \end{aligned}$$

□

A point $\bar{x} \in X$ is said to be Clarke-stationary of problem (3.1) if

$$0 \in \partial^{\mathsf{C}} f(\bar{x}) + N_X(\bar{x}). \tag{3.2}$$

Furthermore, by following the lead of DC programming (see for instance [6, §3.1]), $\bar{x} \in X$ is said to be a critical point if

$$0 \in \partial f_1(\bar{x}) - \partial^{\mathsf{C}} f_2(\bar{x}) + N_X(\bar{x}). \tag{3.3}$$

It is not difficult to see that this inclusion means that

$$\bar{x} \in \arg\min_{x \in X} \; f_1(x) - [f_2(\bar{x}) + \langle s_{f_2}, x - \bar{x} \rangle] \quad \text{for some} \quad s_{f_2} \in \partial^{\mathsf{C}} f_2(\bar{x}).$$

Note that the concept of criticality is weaker than that of Clarke-stationarity, which in turn is weaker than $d$-stationarity (because $f'(\cdot;d) \leq f^\circ(\cdot;d)$ for all $d \in \mathbb{R}^n$). However, criticality and Clarke-stationarity coincide when at least one component function is smooth (in which case $f$ is regular). Furthermore, we can see from Proposition 3.1 and the above alternative characterization of criticality that the three concepts coincide when $f_2$ is continuously differentiable at $\bar{x}$.

For the more general problem (1.1), $\bar{x} \in X$ is said to be a Bouligand ($B$)-stationary point of (1.1) if $f'(\bar{x};d) \geq 0$ for all $d \in \mathcal{T}_{X^c}(\bar{x})$, with $X^c$ the feasible set of (1.1). $B$-stationary boils down to $d$-stationarity if the nonconvex constraint $c(x) \leq 0$ is absent or if the considered point satisfies strictly the nonlinear constraint (in the latter case, $\mathcal{T}_{X^c}(\bar{x}) = \mathcal{T}_X(\bar{x})$). Necessary and sufficient conditions for $B$-stationarity are given in [27, Prop. 4] for the case of DC-constrained DC problems. The next result deals with a more general case: we assume that only $c_2$ is convex, while $f_2$ remains a weakly-convex function.

**Proposition 3.2.** *In addition to our assumptions on problem (1.1), let $c_2 : \mathcal{O} \to \mathbb{R}$ be a convex function and $\bar{x} \in X^c := \{x \in X : c(x) \leq 0\}$ such that $c(\bar{x}) = 0$. Moreover, assume that the following constraint qualification (CQ) holds*

$$\texttt{cl}\left\{d \in \mathcal{T}_X(\bar{x}) : c'(\bar{x};d) < 0\right\} = \left\{d \in \mathcal{T}_X(\bar{x}) : c'(\bar{x};d) \leq 0\right\}. \tag{3.4}$$

*Then, $\bar{x}$ is a $B$-stationary point of problem (1.1) if and only if $\bar{x}$ solves the convex problems*

$$\left\{ \begin{array}{ll} \min\limits_{x \in X} & f_1(x) - [f_2(\bar{x}) + \langle s_{f_2}, x - \bar{x}\rangle] \\ s.t. & c_1(x) - [c_2(\bar{x}) + \langle s_{c_2}, x - \bar{x}\rangle] \leq 0 \end{array} \right\} \quad \forall\, s_{f_2} \in \partial^{\mathtt{C}} f_2(\bar{x}),\ \forall\, s_{c_2} \in \partial c_2(\bar{x}). \tag{3.5}$$

*Proof.* Denote $\bar{Y}(\bar{x}) = \{x \in X : c_1(x) \leq c_2(\bar{x}) + c_2'(\bar{x}; x - \bar{x})\}$. As the CQ (3.4) holds, Proposition 2.1 of [41] ensures that $\mathcal{T}_{X^c}(\bar{x}) = \mathcal{T}_{\bar{Y}(\bar{x})}(\bar{x}) = \texttt{cl}\left\{d \in \mathbb{R}^n : d = t(x - \bar{x}),\, x \in \bar{Y}(\bar{x}),\, t \in \mathbb{R}_+\right\}$. Thus, the $B$-stationary definition is equivalent to

$$\begin{aligned} f'(\bar{x}; x - \bar{x}) = f_1'(\bar{x}; x - \bar{x}) - f_2'(\bar{x}; x - \bar{x}) \geq 0 \qquad & \forall\, x \in \bar{Y}(\bar{x}) \\ \Leftrightarrow f_1'(\bar{x}; x - \bar{x}) \geq \langle s_{f_2}, x - \bar{x}\rangle \qquad & \forall\, s_{f_2} \in \partial^{\mathtt{C}} f_2(\bar{x}),\ \forall\, x \in \bar{Y}(\bar{x}). \end{aligned} \tag{3.6}$$

The stated result follows upon establishing the equivalence between (3.6) and (3.5).

[(3.6) $\Rightarrow$ (3.5)]. Suppose (3.6) holds and let $s_{c_2} \in \partial c_2(\bar{x})$ be arbitrary. As

$$Y(s_{c_2}) := \{x \in X :\, c_1(x) \leq c_2(\bar{x}) + \langle s_{c_2}, x - \bar{x}\rangle\} \subset \bar{Y}(\bar{x})$$

due to convexity of $c_2$, we conclude that $f_1'(\bar{x}; x - \bar{x}) \geq \langle s_{f_2}, x - \bar{x}\rangle$ for all $s_{f_2} \in \partial^{\mathtt{C}} f_2(\bar{x})$ and all $x \in Y(s_{c_2})$. Convexity of the latter set implies that $\bar{x}$ minimizes $f_1(x) - \langle s_{f_2}, x - \bar{x}\rangle$ over $Y(s_{c_2})$ for all $s_{f_2} \in \partial^{\mathtt{C}} f_2(\bar{x})$. Thus, condition (3.5) holds because $s_{c_2} \in \partial c_2(\bar{x})$ was taken arbitrarily.

[(3.5) $\Rightarrow$ (3.6)]. To show the reverse implication, we proceed with a proof by contrapositive. Suppose that there exist $s_{f_2}' \in \partial^{\mathtt{C}} f_2(\bar{x})$ and $x' \in \bar{Y}(\bar{x})$ such that $f_1'(\bar{x}; x' - \bar{x}) < \langle s_{f_2}', x' - \bar{x}\rangle$ (and hence $x' \neq \bar{x}$), i.e., (3.6) does not hold. Let $s_{c_2}' \in \partial c_2(\bar{x})$ be such that $c_2'(\bar{x}; x' - \bar{x}) = \langle s_{c_2}', x' - \bar{x}\rangle$. Therefore, $x'$ is feasible for the convex problem

$$\begin{aligned} &\min_{x \in X}\ f_1(x) - [f_2(\bar{x}) + \langle s_{f_2}', x - \bar{x}\rangle] \\ &\text{s.t. } c_1(x) - [c_2(\bar{x}) + \langle s_{c_2}', x - \bar{x}\rangle] \leq 0. \end{aligned}$$

Together with our assumption $f_1'(\bar{x}; x' - \bar{x}) < \langle s_{f_2}', x' - \bar{x}\rangle$, we have that $d = x' - \bar{x}$ is a feasible descent direction for the above problem, and thus $\bar{x}$ can not be one of its solution. Hence, $\bar{x}$ does not satisfy (3.5). The proof is thus complete. □

Note that convexity of $c_2$ plays an important role in the above proposition. Indeed, if $c_2$ is not convex, then the set $\{x \in X : c_1(x) \leq c_2(\bar{x}) + \langle s_{c_2}, x - \bar{x}\rangle\}$ is not necessarily a subset of $X^c$ and when solving the linearized subproblem (3.5) we may get a point that is infeasible for the original problem (1.1).

**Example 3.3.** Let $f_1 = x$, $f_2 = 0$, $c_1 = 0$, $c_2 = \frac{x^3}{3}$ and $X = [-2, 2]$. We are not in the framework of Proposition 3.2, since $c_2$ is not convex on $[-2, 2]$, but weakly convex (with mudulus $\mu = 4$). At $\bar{x} = 0$, which globally solves (1.1), the convex problem (3.5) becomes $\min_{x \in [-2,2]} x$ (because we have dropped the trivial constraint $0 \leq 0$), and thus does not provide a feasible point for the original problem. □

However, if the mudulus $\mu$ is known for the weakly convex function $c_2$, adding a quadratic term in the constraint of the convex problem (3.5) makes the corresponding set feasible for the original problem. Moreover, Proposition 3.2 is generalized in case of weakly convex $c_2$.

**Corollary 3.4.** *Let $c_2 : \mathcal{O} \to \mathbb{R}$ be a weakly convex function and $\mu_c \geq 0$ be a real number from Proposition 2.4 corresponding to $c_2$. Moreover, assume that the CQ (3.4) holds. Then, $\bar{x}$ is a B-stationary point of problem (1.1) if and only if, for any given $\mu \geq \mu_c$, $\bar{x}$ solves the convex problems*

$$\left\{ \begin{array}{ll} \min\limits_{x \in X} & f_1(x) - [f_2(\bar{x}) + \langle s_{f_2}, x - \bar{x}\rangle] \\ s.t. & c_1(x) - [c_2(\bar{x}) + \langle s_{c_2}, x - \bar{x}\rangle] + \frac{\mu}{2}\|x - \bar{x}\|^2 \leq 0 \end{array} \right\} \quad \forall\, s_{f_2} \in \partial^{\mathsf{C}} f_2(\bar{x}),\ \forall\, s_{c_2} \in \partial^{\mathsf{C}} c_2(\bar{x}). \tag{3.7}$$

*Proof.* Consider the convex functions $\tilde{c}_1(x) = c_1(x) + \frac{\mu}{2}\|x\|^2$ and $\tilde{c}_2(x) = c_2(x) + \frac{\mu}{2}\|x\|^2$ with $\mu \geq \mu_c$. The result follows from Proposition 3.2 by using instead the DC decomposition $c(x) = \tilde{c}_1(x) - \tilde{c}_2(x)$ and by noting that, for an arbitrary $\tilde{s}_{c_2} \in \partial \tilde{c}_2(\bar{x})$, we obtain

$$\tilde{c}_1(x) - [\tilde{c}_2(\bar{x}) + \langle \tilde{s}_{c_2}, x - \bar{x}\rangle] = c_1(x) - [c_2(\bar{x}) + \langle s_{c_2}, x - \bar{x}\rangle] + \frac{\mu}{2}\|x - \bar{x}\|^2$$

with $s_{c_2} = \tilde{s}_{c_2} - \mu\bar{x}$, $s_{c_2} \in \partial^{\mathsf{C}} c_2(\bar{x})$. □

Except for some particular cases, checking $B$-stationarity numerically is out of reach. Therefore, weaker stationarity concepts need to come into play: $\bar{x} \in X$ is said to be Clarke-stationary for (1.1) if there exists a Lagrange multiplier $\bar{\lambda}$ such that

$$\left\{ \begin{array}{l} 0 \in \partial^{\mathsf{C}} f(\bar{x}) + \bar{\lambda}\, \partial^{\mathsf{C}} c(\bar{x}) + N_X(\bar{x}) \\ c(\bar{x}) \leq 0,\ \bar{\lambda}\, c(\bar{x}) = 0,\ \bar{\lambda} \geq 0,\ \bar{x} \in X. \end{array} \right. \tag{3.8}$$

Analogously, $\bar{x}$ is a critical point of (1.1) if there exists a Lagrange multiplier $\bar{\lambda}$ such that

$$\left\{ \begin{array}{l} 0 \in \partial f_1(\bar{x}) - \partial^{\mathsf{C}} f_2(\bar{x}) + \bar{\lambda}\, [\partial c_1(\bar{x}) - \partial^{\mathsf{C}} c_2(\bar{x})] + N_X(\bar{x}) \\ c(\bar{x}) \leq 0,\ \bar{\lambda}\, c(\bar{x}) = 0,\ \bar{\lambda} \geq 0,\ \bar{x} \in X. \end{array} \right. \tag{3.9}$$

Observe that if $f_1$ or $f_2$ and $c_1$ or $c_2$ are smooth, then criticality boils down to Clarke stationarity. Next, we revisit the proximal bundle method of [49] and extend it to the more general setting of problem (1.1). To this end, the method must be modified, and its convergence analysis must be done anew.

## 3.1 Problem reformulation via improvement function

Nonsmooth and noncovex constraints in optimization problems are in general dealt with numerically via exact penalization [18, 21, 34, 14], linearization of certain components [27, 39], and improvement functions [31, 1, 49]. The latter has a recognized good practical performance, does not require the additional assumptions normally assumed in exact penalization methods, and employs parameters that are simple to set. For these reasons, we handle problem (1.1) via the *improvement function* $H : \mathcal{O} \times \mathcal{O} \to \mathbb{R}$ given by

$$H(x; y) = \max\left\{ f(x) - \tau_f(y),\ c(x) - \tau_c(y) \right\}, \tag{3.10a}$$

$$\text{with} \quad \tau_f(y) = f(y) + \rho[c(y)]_+ \quad \text{and} \quad \tau_c(y) = \sigma[c(y)]_+, \quad \text{for } \rho \geq 0 \text{ and } \sigma \in [0, 1). \tag{3.10b}$$

Observe that if $\bar{x}$ is a global solution of (1.1), then $H(x; \bar{x}) \geq 0$ for all $x \in X$ and $H(\bar{x}; \bar{x}) = 0$.

Improvement functions (also known as progress functions) have been considered within bundle methods in [31, 47, 38] for convex problems, in [1] for a class of (nonconvex) optimal control problems, and in [24, 49] for DC-constrained DC programs. In what follows we exploit some relevant mathematical properties of (3.10) and its link to the original problem (1.1). To this end, we need to consider necessary conditions for a point $\bar{x}$ to be a local solution of the reformulated problem

$$\min_{x \in X}\ H(x; \bar{x}). \tag{3.11}$$

As the second argument of $H$ is fixed, it follows from (3.2) that $\bar{x} \in X$ is a Clarke-stationary point of (3.11) if

$$0 \in \partial_1^{\mathsf{C}} H(\bar{x}; \bar{x}) + N_X(\bar{x}), \tag{3.12}$$

where $\partial_1^{\mathsf{C}} H$ stands for the generalized subdifferential of $H$ with respect to the first argument. Proposition 2.1 ii) yields

$$\partial_1^{\mathsf{C}} H(\bar{x}; \bar{x}) \subset \left\{ \begin{array}{ll} \partial^{\mathsf{C}} c(\bar{x}) & \text{if } f(\bar{x}) - \tau_f(\bar{x}) < c(\bar{x}) - \tau_c(\bar{x}) \\ \mathtt{conv}\left\{\partial^{\mathsf{C}} f(\bar{x}),\, \partial^{\mathsf{C}} c(\bar{x})\right\} & \text{if } f(\bar{x}) - \tau_f(\bar{x}) = c(\bar{x}) - \tau_c(\bar{x}) \\ \partial^{\mathsf{C}} f(\bar{x}) & \text{if } f(\bar{x}) - \tau_f(\bar{x}) > c(\bar{x}) - \tau_c(\bar{x}). \end{array} \right.$$

Since we do not work with generalized subgradients of either $f$ or $c$, but only with subgradients of the functions yielding their CwC decompositions (1.1b), we must consider a weaker stationary definition: we say that $\bar{x} \in X$ is a critical point of the composite problem (3.11) with CwC decompositions (1.1b) if

$$0 \in N_X(\bar{x}) + \left\{ \begin{array}{ll} \partial c_1(\bar{x}) - \partial^{\mathsf{C}} c_2(\bar{x}) & \text{if } f(\bar{x}) - \tau_f(\bar{x}) < c(\bar{x}) - \tau_c(\bar{x}) \\ \mathtt{conv}\left\{\partial f_1(\bar{x}) - \partial^{\mathsf{C}} f_2(\bar{x}),\, \partial c_1(\bar{x}) - \partial^{\mathsf{C}} c_2(\bar{x})\right\} & \text{if } f(\bar{x}) - \tau_f(\bar{x}) = c(\bar{x}) - \tau_c(\bar{x}) \\ \partial f_1(\bar{x}) - \partial^{\mathsf{C}} f_2(\bar{x}) & \text{if } f(\bar{x}) - \tau_f(\bar{x}) > c(\bar{x}) - \tau_c(\bar{x}). \end{array} \right. \tag{3.13}$$

Note that if both $f$ and $c$ are regular, then the above condition coincides with that of Clarke stationarity: recall (3.2), Proposition 2.1 i), and observe that the set defined by the expressions in the curly brackets above is nothing but $\partial_1^{\mathsf{C}} H(\bar{x}; \bar{x})$. The following result, inspired by both [1, Lemma 5.1] that deals with the (stronger) Clarke stationarity and [49, Thm. 2] that works with the (weaker) criticality definition from DC programming, links condition (3.13) with criticality of the original problem.

**Theorem 3.5.** *Let $\bar{x} \in X$ be a point satisfying condition (3.13). Then, the following hold:*

*i) If $c(\bar{x}) > 0$, then $\bar{x}$ is a critical point (in the sense of (3.3)) of the optimization problem*

$$\min_{x \in X} \; c_1(x) - c_2(x). \tag{3.14}$$

*ii) If $c(\bar{x}) = 0$ and $\bar{x}$ is not a critical point of (3.14), then $\bar{x}$ satisfies (3.9) for some $\bar{\lambda} > 0$.*

*iii) If $c(\bar{x}) < 0$, then $\bar{x}$ satisfies (3.9) with $\bar{\lambda} = 0$.*

*Proof.* With the $\tau$ function defined in (3.10b), note that

$$f(\bar{x}) - \tau_f(\bar{x}) - [c(\bar{x}) - \tau_c(\bar{x})] = \begin{cases} -[\rho + (1-\sigma)]c(\bar{x}) < 0 & \text{if } c(\bar{x}) > 0, \\ 0 & \text{if } c(\bar{x}) = 0, \\ -c(\bar{x}) > 0 & \text{if } c(\bar{x}) < 0. \end{cases}$$

Hence,

$$\begin{aligned} c(\bar{x}) > 0 \quad &\Leftrightarrow \quad f(\bar{x}) - \tau_f(\bar{x}) < c(\bar{x}) - \tau_c(\bar{x}), \\ c(\bar{x}) = 0 \quad &\Leftrightarrow \quad f(\bar{x}) - \tau_f(\bar{x}) = c(\bar{x}) - \tau_c(\bar{x}), \\ c(\bar{x}) < 0 \quad &\Leftrightarrow \quad f(\bar{x}) - \tau_f(\bar{x}) > c(\bar{x}) - \tau_c(\bar{x}), \end{aligned}$$

and items i) and iii) follow directly from (3.13). To show item ii), recall that $c(\bar{x}) = 0$ and condition (3.13) ensures the existence of $\lambda \in [0, 1]$ such that

$$0 \in \lambda[\partial f_1(\bar{x}) - \partial^{\mathsf{C}} f_2(\bar{x})] + (1-\lambda)[\partial c_1(\bar{x}) - \partial^{\mathsf{C}} c_2(\bar{x})] + N_X(\bar{x}).$$

By assumption, $\bar{x}$ is not a critical point of (3.14). Then $0 \notin \partial c_1(\bar{x}) - \partial^{\mathsf{C}} c_2(\bar{x}) + N_X(\bar{x})$, implying that $\lambda$ above must be strictly positive. Dividing the displayed inclusion by $\lambda > 0$ we obtain the criticality condition (3.9) with $\bar{\lambda} = (1-\lambda)/\lambda > 0$. □

At item ii) above, the assumption that $\bar{x}$ is not a critical point of (3.14) can be seen as a constraint qualification, which turns out to be more restrictive than (3.4). Indeed, the latter excludes $d$-stationary points of (3.14), but not necessarily critical ones. The following example gives a critical point $\bar{x}$ of (3.14) that satisfies (3.4) but not the criticality condition (3.9) for the nonlinearly-constrained problem (1.1).

**Example 3.6.** Take $c_1(x) = \max\{x, 2x\}$, $c_2(x) = \max\{2x, 4x\}$, $X = [-2, 2]$ and $\bar{x} = 0$. Then $\mathcal{T}_X(\bar{x}) = \mathbb{R}$, $N_X(\bar{x}) = \{0\}$, and $\bar{x}$ is a critical point of (3.14) because $0 \in \partial c_1(\bar{x}) - \partial c_2(\bar{x}) = [1, 2] - [2, 4] = [-3, 0]$. Furthermore, note that

$$c_1'(\bar{x}; d) = \max_{s \in [1,\,2]} sd = \begin{cases} 2d & \text{if } d \geq 0 \\ d & \text{if } d \leq 0 \end{cases} \quad \text{and} \quad c_2'(\bar{x}; d) = \max_{s \in [2,\,4]} sd = \begin{cases} 4d & \text{if } d \geq 0 \\ 2d & \text{if } d \leq 0 \end{cases},$$

thus $c'(\bar{x}; d) = \min\{-d, -2d\}$. We conclude that $\{d \in \mathcal{T}_X(\bar{x}) : c'(\bar{x}; d) < 0\} = \mathbb{R}_+$, whereas $\{d \in \mathcal{T}_X(\bar{x}) : c'(\bar{x}; d) \leq 0\} = \mathbb{R}_+ \cup \{0\}$, showing that $\bar{x} = 0$ satisfies the CQ (3.4). However, if we take $f_1(x) = 0$ and $f_2(x) = -\frac{1}{2}x^2 + x$, the following system does not have a solution:

$$\begin{cases} 0 & \in \partial f_1(0) - \partial^{\mathsf{C}} f_2(0) + \bar{\lambda}[\partial c_1(0) - \partial c_2(0)] \\ \bar{\lambda} & \geq 0 \end{cases} \equiv \begin{cases} 0 & \in -1 + \bar{\lambda}[-3, 0] \\ \bar{\lambda} & \geq 0 \end{cases} \equiv \begin{cases} 0 & \in [-3\bar{\lambda} - 1, -1] \\ \bar{\lambda} & \geq 0, \end{cases}$$

i.e., $\bar{x}$ does not satisfy (3.9). Figure 1 illustrates the objective and constraint function in this example: it is clear that $\bar{x}$ is indeed a global maximizer of $f(x)$ under the constraints $x \in X$ and $c(x) \leq 0$.

□

This example shows that, at item ii) of Theorem 3.5, we cannot replace the assumption that $\bar{x}$ is not a critical point of (3.14) with the CQ (3.4).

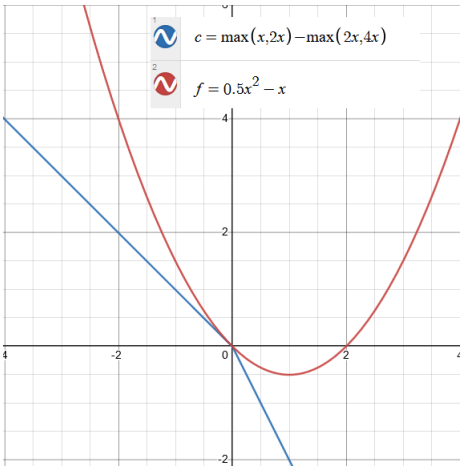


Figure 1: Function $f(x) = \frac{1}{2}x^2 - x$ in red and $c(x) = \max\{x, 2x\} - \max\{2x, 4x\}$ in blue.

## 3.2 The DC setting.

In the DC setting, functions $f_2$ and $c_2$ are convex and the improvement function (3.10) is DC. Indeed, for $\bar{x}$ fixed, we can write

$$H(x;\bar{x}) = F(x;\bar{x}) - G(x), \quad \text{with} \quad \begin{cases} F(x;\bar{x}) & = \max\left\{f_1(x) + c_2(x) - \tau_f(\bar{x}),\ f_2(x) + c_1(x) - \tau_c(\bar{x})\right\}, \\ G(x) & = f_2(x) + c_2(x). \end{cases} \tag{3.15}$$

Since $F$ and $G$ are convex functions, the criticality condition (3.3) for (3.11) (under this DC decomposition) reads as

$$0 \in \partial_1 F(\bar{x};\bar{x}) - \partial G(\bar{x}) + N_X(\bar{x}), \tag{3.16}$$

where $\partial_1 F$ stands for the subdifferential of $F$ with respect to the first argument. It turns out that our new condition (3.13) is stronger than (3.16), used in [49].

**Lemma 3.7.** *In addition to our basic assumptions on problem* (1.1)*, suppose that $f_2$ and $c_2$ are convex. Then the necessary optimality condition* (3.13) *implies* (3.16).

*Proof.* Let $\bar{x} \in X$ be a point satisfying (3.13). Let us first observe that since $f_2, c_2$ are convex and thus regular, we have $\partial G(\bar{x}) = \partial f_2(\bar{x}) + \partial c_2(\bar{x})$. A similar observation can be made concerning the computation for $F$. Our analysis splits into three possible cases.

a) $f(\bar{x}) - \tau_f(\bar{x}) < c(\bar{x}) - \tau_c(\bar{x})$. It follows from (3.13) that

$$0 \in N_X(\bar{x}) + \partial c_1(\bar{x}) - \partial c_2(\bar{x}) \subset N_X(\bar{x}) + \partial f_2(\bar{x}) + \partial c_1(\bar{x}) - [\partial f_2(\bar{x}) + \partial c_2(\bar{x})].$$

We claim that this inclusion implies (3.16). To see that, observe that the above inequality implies $f_1(\bar{x}) + c_2(\bar{x}) - \tau_f(\bar{x}) < f_2(\bar{x}) + c_1(\bar{x}) - \tau_c(\bar{x})$, which in turn gives $\partial_1 F(\bar{x};\bar{x}) = \partial f_2(\bar{x}) + \partial c_1(\bar{x})$. Therefore, the right-hand-side of the above inclusion is (3.16).

b) $f(\bar{x}) - \tau_f(\bar{x}) = c(\bar{x}) - \tau_c(\bar{x})$. It follows from (3.13) that there exists $\lambda \in [0,1]$ such that

$$\begin{aligned} 0 &\in N_X(\bar{x}) + \lambda[\partial f_1(\bar{x}) - \partial f_2(\bar{x})] + (1-\lambda)[\partial c_1(\bar{x}) - \partial c_2(\bar{x})] \\ &= N_X(\bar{x}) + \lambda[\partial f_1(\bar{x}) + \partial c_2(\bar{x})] - \lambda\partial f_2(\bar{x}) + (1-\lambda)\partial c_1(\bar{x}) - \partial c_2(\bar{x}) \\ &\subset N_X(\bar{x}) + \lambda[\partial f_1(\bar{x}) + \partial c_2(\bar{x})] + (1-\lambda)[\partial f_2(\bar{x}) + \partial c_1(\bar{x})] - [\partial f_2(\bar{x}) + \partial c_2(\bar{x})] \\ &\subset N_X(\bar{x}) + \partial_1 F(\bar{x};\bar{x}) - \partial G(\bar{x}). \end{aligned}$$

c) $f(\bar{x}) - \tau_f(\bar{x}) > c(\bar{x}) - \tau_c(\bar{x})$. Again, (3.13) gives

$$0 \in N_X(\bar{x}) + \partial f_1(\bar{x}) - \partial f_2(\bar{x}) \subset N_X(\bar{x}) + \partial f_1(\bar{x}) + \partial c_2(\bar{x}) - [\partial f_2(\bar{x}) + \partial c_2(\bar{x})].$$

The proof is complete because in this case $\partial_1 F(\bar{x};\bar{x}) = \partial f_1(\bar{x}) + \partial c_2(\bar{x})$ due to the fact that $f_1(\bar{x}) + c_2(\bar{x}) - \tau_f(\bar{x}) > f_2(\bar{x}) + c_1(\bar{x}) - \tau_c(\bar{x})$.

□

**Remark 3.8.** In the DC setting, the three concepts of criticality (3.12), (3.13), and (3.16) are equivalent when $f_2$ and $c_2$ are continuously differentiable at $\bar{x}$. Indeed, in this case $f$ and $c$ are regular at $\bar{x}$ and (3.12) coincides with (3.13) (regardless of convexity of $f_2$ and $c_2$). Theorem 2 in [49] ensures that, under these assumptions, (3.16) is equivalent to (3.12). □

The following example shows that (3.16) does not necessarily imply (3.13) in the nondifferentiable DC case.

**Example 3.9.** Let $X = [-1, 1]$, $f_1 = 2x$, $f_2 = |x|$, $c_1 = 4x$ and $c_2 = 2|x|$. At $\bar{x} = 0$, $f(\bar{x}) = c(\bar{x}) = 0$ and thus, $\tau_f(\bar{x}) = \tau_c(\bar{x}) = 0$ due to (3.10b). Furthermore, we have that

$$\partial f_1(0) = \{2\}, \quad \partial f_2(0) = [-1, 1], \quad \partial c_1(0) = \{4\}, \quad \text{and} \quad \partial c_2(0) = [-2, 2].$$

As a result, $\partial f_1(0) - \partial f_2(0) = [1, 3]$, $\partial c_1(0) - \partial c_2(0) = [2, 6]$, $\partial f_1(0) + \partial c_2(0) = [0, 4]$, and $\partial f_2(0) + \partial c_1(0) = [3, 5]$. As in $N_X(0) = \{0\}$, we conclude that

$$0 \notin [1, 6] = \texttt{conv}\{\partial f_1(0) - \partial f_2(0), \partial c_1(0) - \partial c_2(0)\} + N_X(0),$$

whereas

$$\texttt{conv}\{\partial f_1(0) + \partial c_2(0), \partial c_1(0) + \partial f_2(0)\} + N_X(0) = [0, 5]$$

and $\partial f_2(0) + \partial c_2(0) = [-3, 3]$, showing that (3.16) is satisfied but not (3.13). □

The paper [49] proposes a bundle method for DC-constrained DC programs employing the DC decomposition $H = F - G$ above. Once a critical point satisfying (3.16) is computed, the link with criticality of the original problem is adequate if $f_2$ and $c_2$ are continuously differentiable at $\bar{x}$. In the next section we modify that method to compute a point satisfying the stronger criticality condition (3.13). As a result, the link with criticality of the original problem is nicely established by Theorem 3.5 without any additional assumption. In fact, $f_2$ and $c_2$ need not even be convex, but weakly convex on some neighbourhood of each $x \in \mathcal{O}$. We, therefore, strengthen the analysis provided in [49] even though significantly fewer assumptions are required: [49] works in the DC configuration, whereas here, we deal with the more general CwC structure. These improvements, together with the optimality conditions presented above, feature the main contributions of this work.

# 4 Proximal bundle method with improvement function

This section extends the proximal bundle method of [49] for computing a critical point of problem (1.1). The main tool in our analysis is the improvement function $H$ given in (3.10). In the DC setting, the algorithm of [49] works with the explicit DC decomposition (3.15) of $H$ and computes a point $\bar{x} \in X$ satisfying the classic criticality condition in DC programming (3.16). In this section we do not decompose $H$ and consider the milder assumption that $f_2$ and $c_2$ are weakly convex and target the stronger criticality condition (3.13).

## 4.1 The method's main ingredients: model, subproblem, and descent test

The algorithm requires four oracles (black-boxes) providing, for every given $x \in X$, $i \in \{1, 2\}$, the function values $f_i(x)$, $c_i(x)$, arbitrary subgradients $s_{f_1} \in \partial f_1(x)$, $s_{c_1} \in \partial c_1(x)$ (c.f., (2.1)) and arbitrary generalized subgradients $s_{f_2} \in \partial^{\mathcal{C}} f_2(x)$, $s_{c_2} \in \partial^{\mathcal{C}} c_2(x)$ (c.f. (2.2)). We do not impose any assumption on these (generalized) subgradients, as they are assumed to be computed by (external) oracles that do not accept any intervention from the algorithm. (This is particularly useful in industrial applications where companies do not want or cannot share information on the underlying functions with optimizers.)

At iteration $k \in \mathbb{N}$, given a trial point $x^k \in X$, we construct a linearization of every component (here $s^k_{f_i}$, $s^k_{c_i}$, $i \in \{1, 2\}$, denote the respective - generalized - subgradients at $x^k$):

$$\bar{f}^k_1(x) := f_1(x^k) + \langle s^k_{f_1}, x - x^k \rangle \tag{4.1a}$$

$$\bar{f}^k_2(x) := f_2(x^k) + \langle s^k_{f_2}, x - x^k \rangle \tag{4.1b}$$

$$\bar{c}^k_1(x) := c_1(x^k) + \langle s^k_{c_1}, x - x^k \rangle \tag{4.1c}$$

$$\bar{c}^k_2(x) := c_2(x^k) + \langle s^k_{c_2}, x - x^k \rangle. \tag{4.1d}$$

Due to convexity of $f_1$ and $c_1$, we have the following inequalities

$$\bar{f}_1^k(x) \le f_1(x) \quad \text{and} \quad \bar{c}_1^k(x) \le c_1(x) \quad \text{for all } x \in \mathbb{R}^n. \tag{4.2}$$

Since $X$ is compact and components $f_2$ and $c_2$ are assumed to be only weakly convex on some neighbourhood of each $x \in \mathcal{O}$, we have weaker inequalities for these functions. Let $\mathcal{O}' \subset O$ be an open convex set and $\mu_f, \mu_c$ real numbers ensured by Proposition 2.4. As $x^k \in X \subset O'$, the following inequalities are due to Proposition 2.4, item ii)

$$\bar{f}_2^k(x) \le f_2(x) + \tfrac{\bar{\mu}}{2}\|x - x^k\|^2 \quad \text{and} \quad \bar{c}_2^k(x) \le c_2(x) + \tfrac{\bar{\mu}}{2}\|x - x^k\|^2 \quad \text{for all } x \in X, \tag{4.3}$$

where $\bar{\mu} := \max\{\mu_f, \mu_c\} > 0$. Observe that the threshold $\bar{\mu}$ is in general unknown, and the inequalities in (4.3) are only supposed to hold for $x$ in $X$, in contrast with the (subgradient) inequalities in (4.2).

Let $\mathcal{B}_f^k$ and $\mathcal{B}_c^k$ be two index sets gathering the *bundle* of information (function values and subgradients) given by the oracles. In general, $\mathcal{B}_f^k, \mathcal{B}_c^k \subset \{0, \ldots, k\}$ but other possibilities exist making it possible to design a limited-memory method (see Remark 5.5 below). These index sets are useful to define the following individual cutting-plane models for the convex functions $f_1$ and $c_1$:

$$\check{f}_1^k(x) := \max_{j \in \mathcal{B}_f^k} \bar{f}_1^j(x) \quad \le \quad f_1(x) \quad \text{for all } x \in \mathbb{R}^n$$

$$\check{c}_1^k(x) := \max_{j \in \mathcal{B}_c^k} \bar{c}_1^j(x) \quad \le \quad c_1(x) \quad \text{for all } x \in \mathbb{R}^n.$$

Furthermore, let $\ell_k \in \{0, \ldots, k\}$ be the iteration index of the best candidate solution (*stability center*, in the parlance of bundle methods) among the trial points $\{x^0, \ldots, x^k\}$: whenever a better candidate solution $x^{k+1}$ is computed by the algorithm, at a so-called *serious step*, such a point becomes the new stability center and the counter $\ell$ is increased by one: for $\kappa \in (0, \frac{1}{2})$, we declare a *serious step* and let $\ell_{k+1} := k+1$ if $x^{k+1} \neq x^{\ell_k}$ and the inequality

$$H(x^{k+1}; x^{\ell_k}) \le H(x^{\ell_k}; x^{\ell_k}) - \frac{\kappa}{2}\|x^{k+1} - x^{\ell_k}\|^2 \tag{4.5}$$

holds, and declare a *null step* and let $\ell_{k+1} := \ell_k$ otherwise. Since the descent test is independent of the model, the following result from [49] also holds in our framework.

**Lemma 4.1** (Lemma 1 in [49]). *Let $x^{\ell_k} \in X$ be the stability center at iteration $k$. Then $H(x^{\ell_k}; x^{\ell_k}) \ge 0$ and if inequality (4.5) holds, we have that either*

*i) $f(x^{k+1}) \le f(x^{\ell_k}) - \frac{\kappa}{2}\|x^{k+1} - x^{\ell_k}\|^2$ and $c(x^{k+1}) \le 0$ when $c(x^{\ell_k}) \le 0$; or*

*ii) $c(x^{k+1}) \le c(x^{\ell_k}) - \frac{\kappa}{2}\|x^{k+1} - x^{\ell_k}\|^2$ when $c(x^{\ell_k}) > 0$.*

The rationale of serious iterates is to ensure sufficient decrease on one component function of $H(\cdot; x^{\ell_k})$ while maintaining feasibility for (1.1) once reached.

Having all these ingredients at our disposal, we can now define our convex model for the improvement function (3.10) at iteration $k$ :

$$\check{\mathsf{H}}^k(x; x^{\ell_k}) = \max\left\{\check{f}_1^k(x) - \bar{f}_2^{\ell_k}(x) - \tau_f(x^{\ell_k}),\ \check{c}_1^k(x) - \bar{c}_2^{\ell_k}(x) - \tau_c(x^{\ell_k})\right\}. \tag{4.6}$$

(Even in the particular setting where $f_2$ and $c_2$ are convex functions, this model differs from the one employed in [49] and is crucial to obtain convergence results stronger than the ones in that paper.) Given a prox-parameter $\mu^k > 0$ estimating the threshold $\bar{\mu}$ in (4.3), the next iterate is the solution of the following strict convex subproblem

$$x^{k+1} = \operatorname*{arg\,min}_{x \in X}\ \check{\mathsf{H}}^k(x; x^{\ell_k}) + \frac{\mu^k}{2}\|x - x^{\ell_k}\|^2, \tag{4.7}$$

which can be transformed into a QP (provided $X$ is a polyhedron) by adding an extra variable $r \in \mathbb{R}$

$$\left\{\begin{array}{lll} \min\limits_{x,\,r} & r + \frac{\mu^k}{2}\|x - x^{\ell_k}\|^2 & \\ \text{s.t.} & \bar{f}_1^j(x) - \bar{f}_2^{\ell_k}(x) - r \le \tau_f(x^{\ell_k}) & \forall\, j \in \mathcal{B}_f^k \\ & \bar{c}_1^j(x) - \bar{c}_2^{\ell_k}(x) - r \le \tau_c(x^{\ell_k}) & \forall\, j \in \mathcal{B}_c^k \\ & x \in X, \quad r \in \mathbb{R}. & \end{array}\right. \tag{4.8}$$

The optimality condition for (4.7) gives

$$x^{k+1} = x^{\ell_k} - \frac{1}{\mu^k}[p^{k+1} + s_X^{k+1}], \quad \text{with} \quad \left\{\begin{array}{l} p^{k+1} \in \partial_1 \check{\mathsf{H}}^k(x^{k+1}; x^{\ell_k}) \\ s_X^{k+1} \in N_X(x^{k+1}). \end{array}\right. \tag{4.9}$$

As usual in bundle methods, we may remove from the model the inactive linearizations to keep (4.8) small. To this end, we denote by $\bar{\mathcal{B}}_f^k \subset \mathcal{B}_f^k$ and $\bar{\mathcal{B}}_c^k \subset \mathcal{B}_c^k$ the index set of active linearizations in the QP subproblem (4.8), i.e.,

$$\bar{\mathcal{B}}_f^k := \left\{ j \in \mathcal{B}_f^k : \alpha_f^j > 0 \right\} \quad \text{and} \quad \bar{\mathcal{B}}_c^k := \left\{ j \in \mathcal{B}_c^k : \alpha_c^j > 0 \right\} \tag{4.10}$$

where $\alpha_f^j \geq 0$, $j \in \mathcal{B}_f^k$, denote the Lagrange multipliers associated with the first set of constraints and $\alpha_c^j \geq 0$, $j \in \mathcal{B}_c^k$, the ones associated with the second family of constraints. We mention in passing that the index sets $\mathcal{B}_f^k$ and $\mathcal{B}_c^k$ can be kept bounded at the price of including artificial (aggregate) linearizations. We postpone this discussion to Remark 5.5, right after the analysis of null steps (the only place in the convergence analysis where bundle management plays a role.)

We can now present the following proximal bundle method algorithm for `CwC`-constrained `CwC` programs (1.1), which modifies [49, Alg. 1] in two ways. First, the convex model (4.6) for the improvement function is distinct. On the one hand, it is a key element to obtain the stronger criticality condition (3.13), and on the other hand, it leads to a simpler/smaller strongly convex QP (4.8) (more details can be found in Subsection 4.2 below). Second, Algorithm 1 employs an ad-hoc rule to update the proximal parameter $\mu^k$ so that no pre-estimation of the underlying weakly-convex moduli is needed. The proposed rule employs the following value

$$\nu^k := 2 \max \left\{ \frac{\bar{f}_2^{\ell_k}(x^{k+1}) - f_2(x^{k+1})}{\|x^{k+1} - x^{\ell_k}\|^2}, \; \frac{\bar{c}_2^{\ell_k}(x^{k+1}) - c_2(x^{k+1})}{\|x^{k+1} - x^{\ell_k}\|^2}, \; 0 \right\}. \tag{4.11}$$

**Algorithm 1** Proximal Bundle Method for `CwC`-constrained `CwC` programs - `CwC-PBM`

**Step 0 (Initialization)** Let $x^0 \in X$, $\kappa \in (0, \frac{1}{2})$, $\kappa \leq \mu^0$, $\rho \geq 0$, $\delta \in [0, 1)$, and $\texttt{Tol} \geq 0$ be given.
Call the oracles to compute $f_i(x^0)$, $c_i(x^0)$, and (generalized) subgradients $s_{f_i}^0$, $s_{c_i}^0$, $i = 1, 2$.
Define $k := \ell_k = 0$ and $\mathcal{B}_f^0 = \mathcal{B}_c^0 := \{0\}$.

**Step 1 (Trial point)** Compute $x^{k+1}$ by solving the QP (4.8).

**Step 2 (Stopping test)** If $\|x^{k+1} - x^{\ell_k}\| \leq \texttt{Tol}$, then stop and return $x^{\ell_k}$.

**Step 3 (Oracles call)** Compute $f_i(x^{k+1})$, $c_i(x^{k+1})$, and subgradients $s_{f_i}^{k+1}$, $s_{c_i}^{k+1}$, $i = 1, 2$.

**Step 4 (Descent test)**

(a) If (4.5) holds, then declare a *serious step*: define $\ell_{k+1} := k + 1$, choose $\mathcal{B}_f^{k+1}, \mathcal{B}_c^{k+1} \subset \{0, \ldots, k+1\}$ with $\{k+1\} \in \mathcal{B}_f^{k+1} \cap \mathcal{B}_c^{k+1}$ and arbitrarily select $\mu^{k+1} \in (0, \mu^k]$.

(b) Else, declare a *null step*: define $\ell_{k+1} := \ell_k$ and choose $\mathcal{B}_f^{k+1}, \mathcal{B}_c^{k+1} \subset \{0, \ldots, k+1\}$ with $\bar{\mathcal{B}}_f^k \cup \{k+1, \ell_k\} \subset \mathcal{B}_f^{k+1}$ and $\bar{\mathcal{B}}_c^k \cup \{k+1, \ell_k\} \subset \mathcal{B}_c^{k+1}$ ($\bar{\mathcal{B}}_f^k$ and $\bar{\mathcal{B}}_c^k$ as in (4.10)).
Compute $\nu^k$ by (4.11). If $\nu^k \geq \mu^k - 2\kappa$, set $\mu^{k+1} = \nu^k + 1$; otherwise $\mu^{k+1} = \mu^k$.

**Step 5 (Loop)** Set $k := k + 1$ and go back to Step 1.

A drawback of the rule for updating the prox-parameter is that $\mu^k$ only increases after a null step when the inequality $\nu^k \geq \mu^k - 2\kappa$ is verified. As a result, $\mu^k$ may never increase: this is, for instance, the case when $f_2$ and $c_2$ are convex (thus $\nu^k = 0$ for all $k$). The motivation for this rule is to eventually keep the prox-parameter fixed if the algorithm performs an infinite sequence of null steps after a last serious step (see Lemma 5.1). This is a condition necessary to prove Proposition 5.4 below. We care to mention that increasing $\mu^k$ after a null step is a simple strategy that pays off in practice: it helps the algorithm to either stop or produce a new serious step, and thus accelerate the numerical performance.

## 4.2 The DC setting: a comparison with the earlier bundle method for DC probgrams

In the DC setting, both functions $f_2$ and $c_2$ are convex and the improvement function (3.10) is DC. The DC decomposition given in (3.15), with $\bar{x}$ replaced with $x^{\ell_k}$, was exploited in the bundle method of [49] through the following model for the improvement function $H(\cdot; x^{\ell_k})$ (see Eq. (18) therein):

$$\max\left\{\check{f}_1^k(x) + \check{c}_2^k(x) - \tau_f(x^{\ell_k}),\ \check{f}_2^k(x) + \check{c}_1^k(x) - \tau_c(x^{\ell_k})\right\} - [\bar{f}_2^{\ell_k}(x) + \bar{c}_2^{\ell_k}(x)]. \tag{4.12}$$

Differently from our model (4.6), the above gathers also cutting-planes for $f_2$ and $c_2$ and, although gathering more information, only the weaker criticality condition (3.16) is ensured by the method of [49]. Hence, the proposed model (4.6) is more advantageous than (4.12) from both practical and theoretical point of view:

- the quadratic program (QP) issued by our model has only half of the linearizations, and is thus simpler to solve;
- convexity of $f_2$ and $c_2$ are required in (4.12), but not in (4.6);
- both models (4.6) and (4.12) are iteratively updated to ensure that every cluster point $\bar{x} \in X$ of the sequence of stability centers satisfies a criticality condition. To show that such a point is also critical for (the DC counterpart of) (1.1), [49, Thm. 2] requires both $f_2$ and $c_2$ to be continuously differentiable at $\bar{x}$. As we will show in Theorem 5.8 below, neither convexity nor differentiability of $f_2$ and $c_2$ are required to establish that $\bar{x}$ issued by Algorithm 1 is also critical for (1.1) in the sense of (3.9). Thus, Algorithm 1 strengthens the results of [49] even though significantly fewer assumptions are required.

Although the apparently small changes concerning [49, Alg. 1], the convergence analysis in that paper cannot be reused here. The reason is that the analysis in [49] strongly depends on the DC decomposition of the employed model for the improvement function. That reasoning is no longer valid for our new model, even if $f_2$ and $c_2$ were convex. Furthermore, our more general setting requires extra steps to cope with the weakly convex functions.

# 5 Convergence Analysis

The goal of this section is to show that every cluster point $\bar{x}$ of the sequence $\{x^{\ell_k}\}_k \subset X$ generated by Algorithm 1 satisfies the necessary optimality condition (3.13). To this end, we first observe that the sequence of prox-parameters issued by Algorithm 1 is bounded.

**Lemma 5.1.** *The value $\mu_{\max} := \sup_{k\in\mathbb{N}} \mu^k$ is finite. Furthermore, if the algorithm produces an infinite sequence of null steps after a last serious step, then the prox-parameter becomes eventually constant.*

*Proof.* Let $\bar{\mu} := \max\{\mu_{f_2}, \mu_{c_2}, \mu^0\} > 0$ be given, where $\mu_{f_2}$ and $\mu_{c_2}$ are as in Proposition 2.4 for the weakly convex functions $f_2$ and $c_2$, and $\mu^0$ is the parameter given to the algorithm at initialization. Then, by taking $y := x^{k+1}$ and $x := x^{\ell_k}$ in (2.4) it follows that

$$2\frac{\bar{f}_2^{\ell_k}(x^{k+1}) - f_2(x^{k+1})}{\|x^{k+1} - x^{\ell_k}\|^2} \leq \bar{\mu}, \quad \text{and} \quad 2\frac{\bar{c}_2^{\ell_k}(x^{k+1}) - c_2(x^{k+1})}{\|x^{k+1} - x^{\ell_k}\|^2} \leq \bar{\mu} \quad \text{for all } k \text{ with } x^{k+1} \neq x^{\ell_k}.$$

As a result, $\nu^k \leq \bar{\mu}$ for all $k$. Note that the prox-parameter is only increased after a null step such that $\nu^k \geq \mu^k - 2\kappa$. In this case, the rule employed in Step 4 of the algorithm sets $\mu^{k+1} = \nu^k + 1$, which gives $\mu^{k+1} = \nu^k + 1 \leq \bar{\mu} + 1$. Since the algorithm does not increase the prox-parameter after a serious step or null step such that $\nu^k < \mu^k - 2\kappa$, we conclude that $\mu_{\max} := \sup_{k\in\mathbb{N}} \mu^k \leq \bar{\mu}+1$ is finite. Finally, note that the prox-parameter is sharply increased after a null step such that $\nu^k \geq \mu^k - 2\kappa$: $\mu^{k+1} = \nu^k + 1 \geq \mu^k - 2\kappa + 1 > \mu^k + \delta$ because $\kappa \in (0, \frac{1}{2})$, with $\delta = \frac{1}{2} - \kappa > 0$. As a result, if the algorithm produces an infinite sequence of null steps after a last serious step, then the inequality $\nu^k < \mu^k - 2\kappa$ will be satisfied for all $k$ large enough and the prox-parameter will become constant (otherwise $\mu^k$ would increase indefinitely, which contradicts that $\mu_{\max}$ is finite). □

We now define the following function $\bar{\mathsf{H}} : \mathcal{O} \times \mathcal{O} \to \mathbb{R}$, which is of key importance in our analysis:

$$\bar{\mathsf{H}}(x; y) := \max \left\{ f_1(x) - [f_2(y) + \langle s_{f_2}, x - y \rangle] - \tau_f(y),\ c_1(x) - [c_2(y) + \langle s_{c_2}, x - y \rangle] - \tau_c(y) \right\}, \tag{5.1}$$

with $s_{f_2} \in \partial^{\mathsf{C}} f_2(y)$ and $s_{c_2} \in \partial^{\mathsf{C}} c_2(y)$. As these subgradients are not specified, the above definition is ambiguous. However, when $y$ is a point previously computed by the algorithm, say $y = x^j$ for $j \leq k$, then $s^j_{f_2} \in \partial^{\mathsf{C}} f_2(x^j)$ and $s^j_{c_2} \in \partial^{\mathsf{C}} c_2(x^j)$ are the subgradients provided by the oracles and ambiguity disappears:

$$\bar{\mathsf{H}}(x; x^j) := \max \left\{ f_1(x) - \bar{f}^j_2(x) - \tau_f(x^j),\ c_1(x) - \bar{c}^j_2(x) - \tau_c(x^j) \right\}.$$

It follows from convexity of $f_1$ and $c_1$ that, for every $y \in \mathcal{O}$ fixed, the function $\bar{\mathsf{H}}(\cdot; y)$ is convex and satisfies $\bar{\mathsf{H}}(\cdot; x^{\ell_k}) \geq \check{\mathsf{H}}^k(\cdot; x^{\ell_k})$ for all $k$. Furthermore, as $\ell_k \in \mathcal{B}^k_f \cap \mathcal{B}^k_c$ for all $k$, we have that $\check{f}^k_i(x^{\ell_k}) = f_i(x^{\ell_k})$, $\check{c}^k_i(x^{\ell_k}) = c_i(x^{\ell_k})$, $i = 1, 2$, and thus

$$\bar{\mathsf{H}}(x^{\ell_k}; x^{\ell_k}) = \check{\mathsf{H}}^k(x^{\ell_k}; x^{\ell_k}) = H(x^{\ell_k}; x^{\ell_k}). \tag{5.2}$$

The following lemma is of particular interest in the remainder of this work.

**Lemma 5.2.** *Suppose that $\bar{x}$ minimizes $\bar{\mathsf{H}}(\cdot; \bar{x})$ over $X$. Then, $\bar{x}$ satisfies the necessary optimality condition (3.13).*

*Proof.* Convexity of $\bar{\mathsf{H}}(\cdot; \bar{x})$ in the first argument and assumption on $\bar{x} \in X$ imply that $0 \in \partial_1 \bar{\mathsf{H}}(\bar{x}; \bar{x}) + N_X(\bar{x})$. The result follows by noting that, for some pair of generalized subgradients $\bar{s}_{f_2} \in \partial^{\mathsf{C}} f_2(\bar{x})$ and $\bar{s}_{c_2} \in \partial^{\mathsf{C}} c_2(\bar{x})$, the following set

$$\partial_1 \bar{\mathsf{H}}(\bar{x}; \bar{x}) = \begin{cases} \partial c_1(\bar{x}) - \bar{s}_{c_2} & \text{if } f(\bar{x}) - \tau_f(\bar{x}) < c(\bar{x}) - \tau_c(\bar{x}) \\ \mathtt{conv} \left\{ \partial f_1(\bar{x}) - \bar{s}_{f_2},\ \partial c_1(\bar{x}) - \bar{s}_{c_2} \right\} & \text{if } f(\bar{x}) - \tau_f(\bar{x}) = c(\bar{x}) - \tau_c(\bar{x}) \\ \partial f_1(\bar{x}) - \bar{s}_{f_2} & \text{if } f(\bar{x}) - \tau_f(\bar{x}) > c(\bar{x}) - \tau_c(\bar{x}) \end{cases}$$

is contained in the one defined by the curly brackets in (3.13). □

We begin the convergence analysis for the case $\mathtt{Tol} = 0$ with the remark that the sequence of stability centers $\{x^{\ell_k}\}_k$ has at least one cluster point, since it is contained in the compact set $X$. We split the analysis into three cases: the algorithm performs only finitely many steps; the algorithm performs infinitely many steps and the sequence $\{x^{\ell_k}\}_k$ is either finite or infinite.

**Proposition 5.3** (Finitely many iterations)**.** *Assume that Algorithm 1 stops at iteration $k$ with $\mathtt{Tol} = 0$. Then, the last stability center $\bar{x} := x^{\ell_k} = x^{k+1}$ satisfies condition (3.13).*

*Proof.* It follows from the model's definition (4.6) and (5.1) that $\bar{\mathsf{H}}(x; x^{\ell_k}) \geq \check{\mathsf{H}}^k(x; x^{\ell_k})$ for all $x \in \mathcal{O}$. Hence, as $x^{\ell_k} \in X$ we have that

$$\begin{aligned} \bar{\mathsf{H}}(x^{\ell_k}; x^{\ell_k}) &\geq \min_{x \in X} \ \bar{\mathsf{H}}(x; x^{\ell_k}) + \frac{\mu_k}{2} \|x - x^{\ell_k}\|^2 \\ &\geq \min_{x \in X} \ \check{\mathsf{H}}^k(x; x^{\ell_k}) + \frac{\mu_k}{2} \|x - x^{\ell_k}\|^2 \\ &= \check{\mathsf{H}}^k(x^{k+1}; x^{\ell_k}) + \frac{\mu_k}{2} \|x^{k+1} - x^{\ell_k}\|^2 \\ &= \check{\mathsf{H}}^k(x^{\ell_k}; x^{\ell_k}) = \bar{\mathsf{H}}(x^{\ell_k}; x^{\ell_k}), \end{aligned}$$

where the first equality is due to (4.7), the second one follows by the fact that $x^{k+1} = x^{\ell_k}$ since the algorithm stops at iteration $k$ with $\mathtt{Tol} = 0$, and the last one is due to (5.2). Hence, $x^{\ell_k}$ minimizes $\bar{\mathsf{H}}(\cdot; x^{\ell_k}) + \frac{\mu_k}{2} \|\cdot - x^{\ell_k}\|^2$ over $X$ and the quadratic term vanishes in the corresponding optimality condition: $0 \in \partial_1 \bar{\mathsf{H}}(\bar{x}; x^{\ell_k}) + N_X(\bar{x})$ and the stated result follows from Lemma 5.2. □

If the algorithm performs finitely many serious steps and infinite number of null steps, the following result shows that the last stability center satisfies (3.13).

**Proposition 5.4** (Finitely many serious steps)**.** *Suppose that Algorithm 1 with $\mathtt{Tol} = 0$ does not stop but produces only finitely many serious steps. Then the last stability center $\bar{x}$ satisfies the condition (3.13).*

*Proof.* Let $\ell \in \mathbb{N}$ denote the last serious iteration, then $\bar{x} = x^\ell$ and note that, for all subsequent (null) iterations $k > \ell$, $\ell_k = \ell$ and the linearizations $\bar{f}_2^\ell$ and $\bar{c}_2^\ell$ are fixed in the model $\check{\mathsf{H}}^k(\cdot;\bar{x})$, which is in this case a cutting-plane model for the convex function $\bar{\mathsf{H}}(\cdot;\bar{x})$. Here we take $\tau_f^\ell = \tau_f(x^\ell)$, $\tau_c^\ell = \tau_c^\ell(x^\ell)$, and function $\bar{\mathsf{H}}(\cdot;\bar{x})$ defined with the fixed linearizations $\bar{f}_2^\ell$ and $\bar{c}_2^\ell$, i.e.,

$$\bar{\mathsf{H}}(\cdot;x^\ell) := \max\left\{ f_1(\cdot) - \bar{f}_2^\ell(\cdot) - \tau_f^\ell,\ c_1(\cdot) - \bar{c}_2^\ell(\cdot) - \tau_c^\ell \right\}.$$

We highlight that the updating rule for $\mu^k$ in Algorithm 1 ensures that the sequence $\{\mu^k\}_{k>\ell}$ is non-decreasing and becomes constant at a certain value $\mu' \in (0, \mu_{\max}]$ after finitely many steps $k' > \ell$, as a consequence of Lemma 5.1. More precisely, the updating rule at Step 4(b) of Algorithm 1 ensures that

$$\mu^k = \mu' \quad \text{and} \quad \nu^k + 2\kappa < \mu' \quad \text{for all } k > k'. \tag{5.3}$$

Hence, from iteration $k'$ on, Algorithm 1 becomes a cutting-plane procedure to compute the unique solution $\tilde{x}$ of

$$\min_{x\in X}\ \bar{\mathsf{H}}(x;\bar{x}) + \frac{\mu'}{2}\|x-\bar{x}\|^2. \tag{5.4}$$

As the algorithm keeps all the active linearizations in the bundles (Step 4(b)), standard arguments from the convex bundle methods' theory (see [5, Prop. 4.3]) ensure that

$$\lim_{k\to\infty} x^k = \tilde{x} \quad \text{and} \quad \lim_{k\to\infty}[\check{\mathsf{H}}^k(x^{k+1};\bar{x}) - \bar{\mathsf{H}}(x^{k+1};\bar{x})] = 0.$$

(The last inequality implies that the convex model asymptotically coincides with the function at the limit point.) We claim that $\tilde{x} = \bar{x}$. To show that, let us assume the opposite, i.e., $\tilde{x} \neq \bar{x}$, and arrive to a contradiction. In this case, for some $\delta > 0$, we may find and index $k_1$ such that $\|x^{k+1} - \bar{x}\|^2 > \delta$ for all $k \geq k_1$. We may furthermore find an index $k_2$ such that $\check{\mathsf{H}}^k(x^{k+1};\bar{x}) - \bar{\mathsf{H}}(x^{k+1};\bar{x}) \geq -\frac{\kappa}{2}\delta$ for all $k \geq k_2$ as the left-hand side vanishes. Therefore, for $k'' \geq \max\{k_1, k_2, k'\}$, we have

$$\check{\mathsf{H}}^k(x^{k+1};\bar{x}) - \bar{\mathsf{H}}(x^{k+1};\bar{x}) \geq -\frac{\kappa}{2}\|x^{k+1}-\bar{x}\|^2 \neq 0 \quad \text{for all } k > k''.$$

The following chain of inequalities holds at every iteration $k > k''$:

$$\begin{aligned}
\check{\mathsf{H}}^k(\bar{x};\bar{x}) &\geq \check{\mathsf{H}}^k(x^{k+1};\bar{x}) + \tfrac{\mu'}{2}\|x^{k+1}-\bar{x}\|^2 && \text{(by (4.7) and (5.3))}\\
&= [\check{\mathsf{H}}^k(x^{k+1};\bar{x}) - \bar{\mathsf{H}}(x^{k+1};\bar{x})] + \bar{\mathsf{H}}(x^{k+1};\bar{x}) + \tfrac{\mu'}{2}\|x^{k+1}-\bar{x}\|^2\\
&\geq -\tfrac{\kappa}{2}\|x^{k+1}-\bar{x}\|^2 + \bar{\mathsf{H}}(x^{k+1};\bar{x}) + \tfrac{\mu'}{2}\|x^{k+1}-\bar{x}\|^2\\
&\geq -\tfrac{\kappa}{2}\|x^{k+1}-\bar{x}\|^2 + \max\left\{ \begin{array}{l} f_1(x^{k+1}) - \bar{f}_2^\ell(x^{k+1}) - \tau_f^\ell + \frac{\mu'}{2}\|x^{k+1}-\bar{x}\|^2 \\ c_1(x^{k+1}) - \bar{c}_2^\ell(x^{k+1}) - \tau_c^\ell + \frac{\mu'}{2}\|x^{k+1}-\bar{x}\|^2 \end{array} \right\} && \text{(by (5.1))}\\
&> -\tfrac{\kappa}{2}\|x^{k+1}-\bar{x}\|^2 + \max\left\{ \begin{array}{l} f_1(x^{k+1}) - \bar{f}_2^\ell(x^{k+1}) - \tau_f^\ell + \frac{\nu^k+2\kappa}{2}\|x^{k+1}-\bar{x}\|^2 \\ c_1(x^{k+1}) - \bar{c}_2^\ell(x^{k+1}) - \tau_c^\ell + \frac{\nu^k+2\kappa}{2}\|x^{k+1}-\bar{x}\|^2 \end{array} \right\} && \text{(by (5.3))}\\
&\geq -\tfrac{\kappa}{2}\|x^{k+1}-\bar{x}\|^2 + \max\left\{ \begin{array}{l} f_1(x^{k+1}) - f_2(x^{k+1}) - \tau_f^\ell + \frac{2\kappa}{2}\|x^{k+1}-\bar{x}\|^2 \\ c_1(x^{k+1}) - c_2(x^{k+1}) - \tau_c^\ell + \frac{2\kappa}{2}\|x^{k+1}-\bar{x}\|^2 \end{array} \right\} && \text{(by (4.11))}\\
&= H(x^{k+1};\bar{x}) + \tfrac{\kappa}{2}\|x^{k+1}-\bar{x}\|^2. && \text{(by (3.10))}
\end{aligned}$$

As $\bar{x} = x^\ell$ and $\check{\mathsf{H}}^k(x^\ell;x^\ell) = H(x^\ell;x^\ell)$ due to (5.2), we have shown that the descent test (4.5) is satisfied at $x^{k+1} \neq x^\ell$:

$$H(x^{k+1};x^\ell) \leq H(x^\ell;x^\ell) - \frac{\kappa}{2}\|x^{k+1}-x^\ell\|^2,$$

contradicting thus the assumption that only null steps are performed for $k > \ell$. Hence, $\tilde{x} = \bar{x}$ and the last stability center solves (5.4). This allows us to conclude (thanks to convexity of $\bar{\mathsf{H}}(\cdot;\bar{x})$) that $\bar{x} = x^\ell$ solves $\min_{x\in X}\bar{\mathsf{H}}(x;x^\ell)$. Lemma 5.2 then concludes the proof. □

**Remark 5.5** (Bundle compression)**.** It is worth mentioning that the index sets $\mathcal{B}_f^k$ and $\mathcal{B}_c^k$ gathering the information bundle can be kept bounded; each one having at most $M_{\max}$ indices, for a chosen integer $M_{\max} \geq 3$. Indeed, it suffices to keep in the bundles the linearizations issued by the stability center $x^{\ell_k}$, the new trial point $x^{k+1}$ and the so-called aggregate linearization as in [5, Eq. 4.5]. When transcribed to our setting, the aggregate linearizations for $f_1$ and $c_1$ read as

$$\bar{f}_1^{a_f^k}(x) := \check{f}_1^k(x^{k+1}) + \langle p_f^k, x - x^{k+1}\rangle \leq f_1(x) \quad \forall x \in \mathbb{R}^n$$

$$\bar{c}_1^{a_c^k}(x) := \check{c}_1^k(x^{k+1}) + \langle p_c^k, x - x^{k+1}\rangle \leq c_1(x) \quad \forall x \in \mathbb{R}^n,$$

with $p_f^k := \sum_{j\in\mathcal{B}_f^k} \alpha_f^j s_{f_1}^j$, $p_c^k := \sum_{j\in\mathcal{B}_c^k} \alpha_c^j s_{c_1}^j$ and multipliers $\alpha_f, \alpha_c$ as in (4.10). We claim that the following economical rule for managing $\mathcal{B}_f^k$ and $\mathcal{B}_c^k$ (in Step 4 of Algorithm 1) is enough to ensure convergence:

**Serious step**: set $\mathcal{B}_f^{k+1} = \{k+1\}$ and $\mathcal{B}_c^{k+1} = \{k+1\}$;

**Null step**: set $\mathcal{B}_f^{k+1} = \{k+1, \ell_k, a_f^k\}$ and $\mathcal{B}_c^{k+1} = \{k+1, \ell_k, a_c^k\}$.

Indeed, Proposition 5.4 is still valid if the algorithm employs the above economical rule for updating the bundles: the key Proposition 4.3 from [5] still applies and thus the displayed equations right after (5.4) hold. As it can be noted in the sequel, no bundle management restriction (besides the requirement that $k+1 \in \mathcal{B}_f^{k+1} \cap \mathcal{B}_c^{k+1}$) is required after a serious steps. □

We consider now the case of infinitely many serious steps. To this end, we need the following auxiliary result.

**Lemma 5.6.** *There exist constants $L, M > 0$ such that, for all $k \in \mathbb{N}$, the three following conditions hold for $p^{k+1} \in \partial_1 \check{\mathsf{H}}^k(x^{k+1}; x^{\ell_k})$, $s_X^{k+1} \in N_X(x^{k+1})$, and $e^{k+1} = L\|x^{k+1} - x^{\ell_k}\|$:*

$$\|p^{k+1} + s_X^{k+1}\| \le \mu_{\max} \|x^{k+1} - x^{\ell_k}\| \le M, \tag{5.5a}$$

$$p^{k+1} + s_X^{k+1} \in \partial_{e^{k+1}} \big[\bar{\mathsf{H}}(x^{\ell_k}; x^{\ell_k}) + i_X(x^{\ell_k})\big], \tag{5.5b}$$

$$p^{k+1} \in \partial_{e^{k+1}} \bar{\mathsf{H}}(x^{\ell_k}; x^{\ell_k}). \tag{5.5c}$$

*Proof.* As $\mu^k \in (0, \mu_{\max}]$ (c.f. Lemma 5.1), expression (4.9) yields the first inequality in (5.5a). Recall that the iterates $x^{k+1}$ and $x^{\ell_k}$ are contained in the bounded set $X$ for all $k$. The second inequality in (5.5a) then follows. Convexity of the function $\check{\mathsf{H}}^k + i_X$ and (4.9) gives that, for all $x \in \mathbb{R}^n$,

$$\begin{aligned}
\check{\mathsf{H}}^k(x; x^{\ell_k}) + i_X(x) &\ge \check{\mathsf{H}}^k(x^{k+1}; x^{\ell_k}) + \langle p^{k+1} + s_X^{k+1}, x - x^{k+1}\rangle \\
&\ge \check{\mathsf{H}}^k(x^{k+1}; x^{\ell_k}) + \langle p^{k+1} + s_X^{k+1}, x - x^{\ell_k}\rangle + \langle p^{k+1} + s_X^{k+1}, x^{\ell_k} - x^{k+1}\rangle \\
&\ge \check{\mathsf{H}}^k(x^{k+1}; x^{\ell_k}) + \langle p^{k+1} + s_X^{k+1}, x - x^{\ell_k}\rangle - M\|x^{\ell_k} - x^{k+1}\|,
\end{aligned} \tag{5.6}$$

where the last inequality is due to (5.5a) and Cauchy-Schwarz inequality. Definition (4.6) of $\check{\mathsf{H}}^k(\cdot; x^{\ell_k})$ as well as the fact that $\ell_k \in \mathcal{B}_f^k \cap \mathcal{B}_f^k$ give the following chain of inequalities:

$$\begin{aligned}
\check{\mathsf{H}}^k(x^{k+1}; x^{\ell_k}) &\ge \max\Big\{\bar{f}_1^{\ell_k}(x^{k+1}) - \bar{f}_2^{\ell_k}(x^{k+1}) - \tau_f(x^{\ell_k}), \bar{c}_1^{\ell_k}(x^{k+1}) - \bar{c}_2^{\ell_k}(x^{k+1}) - \tau_c(x^{\ell_k})\Big\} \\
&\ge \max\Big\{f_1(x^{\ell_k}) - f_2(x^{\ell_k}) - \tau_f(x^{\ell_k}),\, c_1(x^{\ell_k}) - c_2(x^{\ell_k}) - \tau_c(x^{\ell_k})\Big\} \\
&\quad + \min\Big\{\langle s_{f_1}^{\ell_k} - s_{f_2}^{\ell_k}, x^{k+1} - x^{\ell_k}\rangle,\, \langle s_{c_1}^{\ell_k} - s_{c_2}^{\ell_k}, x^{k+1} - x^{\ell_k}\rangle\Big\}.
\end{aligned}$$

Since $X \subset \mathcal{O}$ is compact, we have that $\partial f_1(X)$, $\partial c_1(X)$, $\partial^{\mathsf{C}} f_2(X)$, and $\partial^{\mathsf{C}} c_2(X)$ are bounded sets (see Section 2). Hence, there exist $K_f > 0$ and $K_c > 0$ such that, $\|s_{f_1}^{\ell_k} - s_{f_2}^{\ell_k}\| \le K_f$ and $\|s_{c_1}^{\ell_k} - s_{c_2}^{\ell_k}\| \le K_c$ for all $k$. Applying the Cauchy-Schwarz inequality to the inequalities above and recalling that

$$\max\Big\{f_1(x^{\ell_k}) - f_2(x^{\ell_k}) - \tau_f(x^{\ell_k}),\, c_1(x^{\ell_k}) - c_2(x^{\ell_k}) - \tau_c(x^{\ell_k})\Big\} = H(x^{\ell_k}; x^{\ell_k}) = \bar{\mathsf{H}}(x^{\ell_k}; x^{\ell_k})$$

by definition, we get

$$\check{\mathsf{H}}^k(x^{k+1}; x^{\ell_k}) \ge \bar{\mathsf{H}}(x^{\ell_k}; x^{\ell_k}) - L_0\|x^{k+1} - x^{\ell_k}\|, \quad \text{with } L_0 = \max\{K_f, K_c\}. \tag{5.7}$$

Recall that $\bar{\mathsf{H}}(x; x^{\ell_k}) \ge \check{\mathsf{H}}^k(x; x^{\ell_k})$ for all $x \in \mathbb{R}^n$ and combine (5.6) with (5.7) to obtain

$$\begin{aligned}
\bar{\mathsf{H}}(x; x^{\ell_k}) + i_X(x) &\ge \check{\mathsf{H}}^k(x; x^{\ell_k}) + i_X(x) \\
&\ge \check{\mathsf{H}}^k(x^{k+1}; x^{\ell_k}) + \langle p^{k+1} + s_X^{k+1}, x - x^{\ell_k}\rangle - M\|x^{\ell_k} - x^{k+1}\| \\
&\ge \bar{\mathsf{H}}(x^{\ell_k}; x^{\ell_k}) - (L_0 + M)\|x^{k+1} - x^{\ell_k}\| + \langle p^{k+1} + s_X^{k+1}, x - x^{\ell_k}\rangle.
\end{aligned}$$

We have thus shown (5.5b) with $L = M + L_0$. To prove the last inclusion (5.5c), observe that this chain of inequalities remains true if the term $i_X(x)$ is excluded together with corresponding subdifferential $s_X^{k+1}$: for all $x \in \mathbb{R}^n$,

$$\begin{aligned}
\bar{\mathsf{H}}(x; x^{\ell_k}) &\ge \check{\mathsf{H}}^k(x; x^{\ell_k}) \ge \check{\mathsf{H}}^k(x^{k+1}; x^{\ell_k}) + \langle p^{k+1}, x - x^{\ell_k}\rangle - M\|x^{\ell_k} - x^{k+1}\| \\
&\ge \bar{\mathsf{H}}(x^{\ell_k}; x^{\ell_k}) + \langle p^{k+1}, x - x^{\ell_k}\rangle - (L_0 + M)\|x^{k+1} - x^{\ell_k}\|.
\end{aligned}$$

□

Observe that $\check{\mathsf{H}}^k(\cdot; x^{\ell_k})$ given in (4.6) is the pointwise maximum of finitely many affine functions. Hence, its subdifferential is the convex hull of the "active" linearization slopes, i.e., Proposition 2.1 ii) asserts that

$$\partial_1 \check{\mathsf{H}}^k(x^{k+1}; x^{\ell_k}) := \mathtt{conv}\left\{ \left\{ s_{f_1}^j - s_{f_2}^{\ell_k} \right\}_{j \in \bar{\mathcal{B}}_f^k}, \left\{ s_{c_1}^j - s_{c_2}^{\ell_k} \right\}_{j \in \bar{\mathcal{B}}_c^k} \right\}, \tag{5.8}$$

with $\bar{\mathcal{B}}_f^k$ and $\bar{\mathcal{B}}_c^k$ given in (4.10). Since $X \subset \mathcal{O}$ is compact, we have that $\partial f_1(X)$, $\partial c_1(X)$, $\partial^{\mathsf{C}} f_2(X)$, and $\partial^{\mathsf{C}} c_2(X)$ are bounded sets (see Section 2). Thus, (4.9) and (5.8) certificate that the sequence of model's subgradients $\{p^k\}$ is bounded. This property is used in the proof of the following proposition.

**Proposition 5.7** (Infinitely many serious steps)**.** *Assume that the algorithm performs infinitely many serious steps. Then, any cluster point $\bar{x} \in X$ of the sequence $\{x^{\ell_k}\}_k$ satisfies the necessary optimality condition* (3.13).

*Proof.* We first show that

$$\lim_{k \to \infty} \|x^{\ell_{k+1}} - x^{\ell_k}\| = 0\,. \tag{5.9}$$

To this end, we must analyze the two cases of Lemma 4.1. In case i), Algorithm 1 produces a feasible point for (1.1) after finitely many serious steps and all subsequent points are feasible. Let $x^{\ell_{k_1}}$ be the first feasible serious iterate. Then, Lemma 4.1 i) yields

$$f(x^{\ell_{k+1}}) \leq f(x^{\ell_k}) - \tfrac{\kappa}{2}\|x^{\ell_{k+1}} - x^{\ell_k}\|^2 \text{ and } c(x^{\ell_{k+1}}) \leq 0 \text{ for all } k \geq \ell_{k_1}\,.$$

The telescopic sum of the first inequality above yields

$$\begin{aligned} \sum_{k=k_1}^{\infty} \|x^{\ell_{k+1}} - x^{\ell_k}\|^2 &\leq \frac{2}{\kappa} \sum_{k=k_1}^{\infty} \Big( f\big(x^{\ell_k}\big) - f\big(x^{\ell_{k+1}}\big) \Big) \\ &\leq \frac{2}{\kappa} \Big( f\big(x^{\ell_{k_1}}\big) - \lim_{k \to \infty} f\big(x^{\ell_{k+1}}\big) \Big)\,. \end{aligned}$$

Since $f$ is finite-valued and continuous over the bounded set $X$, the right-hand side of the above inequality is finite. Hence, (5.9) holds. Assume now that the sequence $\{x^{\ell_k}\}$ is infeasible for (1.1). Lemma 4.1 ii) yields

$$0 < c(x^{\ell_{k+1}}) \leq c(x^{\ell_k}) - \tfrac{\kappa}{2}\|x^{\ell_{k+1}} - x^{\ell_k}\|^2 \text{ for all } \ell.$$

Once again, by using the telescopic sum we get (5.9).

As $X \subset \mathcal{O}$ is compact, with $\mathcal{O}$ an open set contained in the domains of component functions, and the generalized subdifferential is locally compact, we conclude that for $s_{f_2}^{\ell_k} \in \partial^{\mathsf{C}} f_2(x^{\ell_k})$ and $s_{c_2}^{\ell_k} \in \partial^{\mathsf{C}} c_2(x^{\ell_k})$

$$\left\{x^{\ell_k}\right\}, \quad \left\{s_{f_2}^{\ell_k}\right\} \quad \text{and} \quad \left\{s_{c_2}^{\ell_k}\right\} \quad \text{are bounded sequences.}$$

By taking subsequences, we can define an index set $\mathcal{K} \subset \{0, 1, 2, \ldots\}$ such that

$$\lim_{\mathcal{K} \ni k \to \infty} x^{\ell_k} = \bar{x} \in X, \quad \lim_{\mathcal{K} \ni k \to \infty} s_{f_2}^{\ell_k} = \bar{s}_{f_2} \in \partial^{\mathsf{C}} f_2(\bar{x}) \quad \text{and} \quad \lim_{\mathcal{K} \ni k \to \infty} s_{c_2}^{\ell_k} = \bar{s}_{c_2} \in \partial^{\mathsf{C}} c_2(\bar{x}),$$

where the two last limits are due to the fact that the generalized subdifferential is outer-semicontinuous [4, Prop. 2.1.5(b)]. Let us now define $\varphi_k(\cdot) = \bar{\mathsf{H}}(\cdot; x^{\ell_k})$ and $\varphi(\cdot) = \bar{\mathsf{H}}(\cdot; \bar{x})$, with the latter defined in (5.1) with $y = \bar{x}$ and the pair of generalized subgradients $(\bar{s}_{f_2}, \bar{s}_{c_2})$ above. For every $x \in \mathbb{R}^n$ fixed, the above limits imply

$$\lim_{\mathcal{K} \ni k \to \infty} \bar{\mathsf{H}}(x; x^{\ell_k}) = \bar{\mathsf{H}}(x; \bar{x}),$$

i.e., $\{\varphi_k\}_{k \in \mathcal{K}}$ converges pointwise to $\varphi$. With $k+1 = \ell_{k+1}$, recall that $p^{\ell_{k+1}} \in \partial_{e^{\ell_{k+1}}} \varphi_\ell(x^{\ell_k})$ due to Lemma 5.6, Eq. (5.5c). Furthermore, as the sequence $\{p^{\ell_{k+1}}\}$ is bounded (see the paragraph right before this proposition), we may take another subsequence indexed by $\mathcal{K}' \subset \mathcal{K}$ so that $\lim_{\mathcal{K}' \ni \ell \to \infty} p^{\ell_{k+1}} = \bar{p} \in \mathbb{R}^n$ and $\lim_{\mathcal{K}' \ni \ell \to \infty} e^{\ell_{k+1}} = 0$ in view of (5.9) and definition of $e^{k+1}$ given in Lemma 5.6. With these conditions at hand, Lemma A.1 (in Appendix A) ensures that $\bar{p} \in \partial\varphi(\bar{x})$, i.e., $\bar{p} \in \partial_1 \bar{\mathsf{H}}(\bar{x}; \bar{x})$. Next, observe that $\{s_X^{\ell_{k+1}}\}$ is a bounded sequence as the inequality

$$||p^{\ell_{k+1}} + s_X^{\ell_{k+1}}|| \leq \mu_{\max} ||x^{\ell_{k+1}} - x^{\ell_k}|| \tag{5.10}$$

holds due to Lemma 5.6 (with $\mu_{\max}$ finite due to Lemma 5.1). By definition of the convex normal cone, it follows that there exists a suitable subsequence of $\{s_X^{\ell_{k+1}}\}_{k \in \mathcal{K}''}$, with $\mathcal{K}'' \subset \mathcal{K}'$ converging to a cluster point $\bar{s} \in N_X(\bar{x}) = \partial i_X(\bar{x})$. Hence, since $X$ is polyhedral and $\mathtt{ri}(\mathcal{D}\mathrm{om}(\bar{\mathsf{H}}(\cdot; \bar{x}))) = \mathcal{O} \neq \emptyset$,

$$\bar{p} + \bar{s} \in \partial_1 \bar{\mathsf{H}}(\bar{x}; \bar{x}) + \partial i_X(\bar{x}) = \partial_1 \big[\bar{\mathsf{H}}(\bar{x}; \bar{x}) + i_X(\bar{x})\big]. \tag{5.11}$$

Finally, inequality (5.10) combined with (5.9) yield $\bar{p} + \bar{s} = 0$. Hence, $0 \in \partial_1 \big[\bar{\mathsf{H}}(\bar{x}; \bar{x}) + i_X(\bar{x})\big]$, showing that $\bar{x}$ minimizes $\bar{\mathsf{H}}(\cdot; \bar{x})$ over $X$. Lemma 5.2 thus concludes the proof. □

The following theorem sums up the algorithm's convergence analysis.

**Theorem 5.8** (Convergence analysis)**.** *Let $X \neq \emptyset$ be a bounded polyhedron contained in the open set $\mathcal{O}$, $f_1, c_1 : \mathcal{O} \to \mathbb{R}$ convex, and $f_2, c_2 : \mathcal{O} \to \mathbb{R}$ weakly convex functions on some neighbourhood of each $x \in \mathcal{O}$. If in Algorithm 1 the stopping test tolerance* `Tol` *is set to zero, then any cluster point $\bar{x}$ of the sequence of stability centers $\{x^{\ell_k}\}$ satisfies the necessary optimality condition (3.13).*

*Furthermore, concerning the original problem (1.1), the following holds:*

*i) If $c(\bar{x}) > 0$, then $\bar{x}$ is a critical point of $\min_{x \in X} c_1(x) - c_2(x)$.*

*ii) If $c(\bar{x}) = 0$ and $\bar{x}$ is not a critical point of $\min_{x \in X} c_1(x) - c_2(x)$, then $\bar{x}$ satisfies the criticality condition (3.9) with $\bar{\lambda} > 0$.*

*iii) If $c(\bar{x}) < 0$, then $\bar{x}$ satisfies the criticality condition (3.9) with $\bar{\lambda} = 0$.*

*If* `Tol` $> 0$*, then the algorithm stops after finitely many steps $k \in \mathbb{N}$ with an approximate critical point $x^{\ell_k}$ of (3.13).*

*Proof.* For the case `Tol` $= 0$, condition (3.13) follows directly from Proposition 5.3 if $\{x^{\ell_k}\}$ is finite, from Proposition 5.4 if the the algorithm produces only finitely many serious steps followed by an infinite sequence of null steps, and from Proposition 5.7 if infinitely many serious steps are produced.

Furthermore, the connection with the necessary optimality condition for the original problem (1.1) is established by Theorem 3.5.

Proposition 5.4 ensures that $\lim_{k \to \infty} ||x^{k+1} - x^{\ell}|| = 0$ if $x^{\ell}$ is the last stability center. Otherwise, $\lim_{k \to \infty} ||x^{\ell_{k+1}} - x^{\ell_k}|| = 0$, as shown in the proof of Proposition 5.7. Thus, Algorithm 1 stops after finitely many steps provided `Tol` $> 0$. □

# 6 Illustrative numerical examples

A deep analysis of the numerical performance of Algorithm 1 is out of the scope of this paper. Instead, this section aims to illustrate our approach to solving some challenging test problems. Here we have two goals: show that it provides good-quality critical points (examples of Subsection 6.1, Subsection 6.2 and Subsection 6.4) and is able to solve problems that, to the best of our knowledge, could not be resolved with other solvers (example of Subsection 6.3). We consider three nonconvex stochastic optimization problems and one coming from signal processing. Notice that the stochastic problems do not have explicit DC decompositions, and thus the algorithm programming algorithms are not directly applicable.

We have coded Algorithm 1 in `MATLAB` using `Gurobi` for solving the master QP subproblem (4.8). Our implementation allows for simpler problems without non-linear constraints (as in the case of the problem in Subsection 6.1). We invite the interested reader to check Appendix B for a brief presentation of how the approach can be simplified in this case. As for the algorithm's parameters, we have chosen `Tol` $= 10^{-4}$, $\rho = \sigma = \frac{1}{2}$, and $\kappa = 0.3$. Parameter $\mu^0 > 0$ has been selected differently (in the range $[10^{-2}, 10^2]$) depending on the problem.

## 6.1 Investment like problems

We will see how the structure of Example 1.1 can appear in practice. Here we will follow the general discussion in [45]. We are interested in the situation wherein we dispose of a set of different technologies $i = 1, ..., m$ capable of generating electricity. Each technology comes with a specific and detailed set of constraints $P_i$, cost function $c_i$ attributing to $p_i \in \mathbb{R}^T$ the cost of generation. Altogether, the various technologies are meant to ensure the satisfaction of a given customer load $d \in \mathbb{R}^T$. We are interested in finding the optimal mix. Thus for $i = 1, ..., m$, we are given $\theta_i \in \mathbb{N}$, the number of "units" of a given type we would like to invest in. The vector $\theta$ comes with an investment cost $F(\theta)$. In a deterministic setting this would amount to solving

$$\begin{aligned} \min_{\theta \in \Theta, p_i^j \in P_i} \quad & F(\theta) + \sum_{i=1}^{m} \sum_{j=1}^{\theta_i} c_i(p_i^j) \\ \text{s.t.} \quad & \sum_{i=1}^{m} \sum_{j=1}^{\theta_i} p_i^j \geq d. \end{aligned}$$

Now should for each $i$, the mappings $c_i$ as well as the feasible sets $P_i$ be convex, then it must be so that the averaged solution: $p_i^* = \frac{1}{\theta_i} \sum_{j=1}^{\theta_i} (p_i^j)^*$, in which each power plant of technology $i$ produces exactly this

amount is also optimal. This follows from using convexity of $P_i$ showing feasibility of $p_i^*$ and through using Jensen's inequality for $c_i$. This is also exactly what would happen if we would solve the subproblems of the Lagrangian dual (w.r.t. the load constraint) for a given fixed investment vector. The convexifying effect of the Lagrangian is well known, e.g., [19, 43] and thus for this dual setting convexity of $c_i$ or $P_i$ would not be essential. Either way, as a result we may thus assume that each power plant of the same technology produces the same amount. This would thus lead to the simpler problem (less variables):

$$\begin{aligned} \min_{\theta\in\Theta, p_i\in P_i} \quad & F(\theta) + \sum_{i=1}^{m} \theta_i c_i(p_i) \\ \text{s.t.} \quad & \sum_{i=1}^{m} \theta_i p_i \geq d. \end{aligned}$$

We will investigate a two-stage stochastic version of the last problem, wherein $d$ is uncertain. We thus define:

$$Q(\theta, \xi) = \min_{p_i\in P_i} \sum_{i=1}^{m} \theta_i c_i(p_i) \text{ s.t. } \sum_{i=1}^{m} \theta_i p_i \geq d. \tag{6.1}$$

and consider the optimization problem

$$\min_{\theta\in\Theta} F(\theta) + \mathbb{E}[Q(\theta, \xi)], \tag{6.2}$$

where for the sake of simplicity we will assume $\theta$ to be allowed to take continuous values ($\Theta$ is a polytope). We will also assume that the feasible set $P_i$ is convex, although one could investigate problem (6.2) without this assumption - for instance by arguing through Lagrangian duality.

Let us now look at a concrete instance. We will consider a time horizon of $t = 1, ..., T$ time steps where each time step is considered to be $\Delta t$ hours long. The problem disposes of $m$ types of technology, having the following characteristics. Each technology type has a maximum power output level $p_i^{\texttt{mx}}$, proportional cost $c_i$ and gradient condition $g_i$. Additionally, each unit is assumed to dispose of a carbon emission rate $e_i$, and the system subject to a carbon cost $f$. Concretely this means that the proportional cost gets updated through the formula $c_i \leftarrow c_i + fe_i$.

The system is moreover endowed with a given customer load that we will assume to be multivariate Gaussian with a given mean and positive definite Covariance matrix. We refer to [36, § 5] for the description of $P_i$ (polyhedral). Furthermore for the various technologies we will assume that $F(\theta) = F^\top\theta$. The purpose of our experiment is to showcase how concretely the new algorithm can process specifically structured problems such as these.

Following the description of Example 1.1, we need to compute $Q^\varepsilon(\theta, \xi)$ at each given $\theta$. The latter involves the solution of a convex optimization problem, wherein $\psi_t$ is given by the $t$-th component of $d - \sum_{i=1}^{m} \theta_i p_i$. As a result of the logarithm, the objective function defining $Q^\varepsilon$ is convex non-linear in $y$. We will therefore use a cutting plane approach to internally compute $Q^\varepsilon$, as well as it's gradient. The inner optimization is initialized from the optimal solution $y_0$ of the inner optimization problem of $Q(x, \xi)$. The latter can be computed by solving a linear program. This will give us the oracle for $f_2$ (in the notation of Subsection 4.1).

Table 1 provides the concrete data.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $p^{\texttt{mx}}$ (MW) | 900 | 900 | 900 | 300 | 300 | 200 | 200 | 200 | 100 | 10000 |
| $g$ (MW/h) | 100 | 100 | 100 | 30 | 30 | 20 | 20 | 70 | 70 | 5000 |
| $c$ (€/MWh) | 30 | 35 | 37 | 45 | 47 | 60 | 100 | 110 | 150 | 10000 |
| $F^{\texttt{inv}}$ (€) | 493151 | 493151 | 493151 | 41096 | 41096 | 32877 | 32877 | 32877 | 21918 | 0 |
| e (t/MW) | 0 | 0 | 0 | 1 | 1 | 0.5 | 0.5 | 0.5 | 1.1 | 0 |

Table 1: Data for the stylized investment problem

We can observe that the last unit described in the previous table is an imbalance unit - a computational trick to ensure that one can always meet the load, in this case even despite a potentially completely unbalanced set of invested assets. In terms of constraints on investment, we do not allow investment in this last unit, the capacity will remain at 1. The cost of investment was set up taking inspiration from https://www.e-education.psu.edu/eme801/node/530, upon rescaling to match $T$ and while accounting for life span of the various technologies. The data of the case is stylized and the general purpose of the study

is more a demonstration of the capabilities of the algorithm rather than an attempt to provide practical insights into investment regarding the electrical system.

The nominal set of assets consists of one asset of each type. This is the initial vector chosen to start the algorithm. The first run considers a situation wherein $f = 0$, i.e. with zero carbon cost. The algorithm in a total of 13 iterations performs a single serious step and essentially keeps the nominal investment vector fixed - the total installed capacity is 4000 MW. In a second run we have set $f = 100$ to see the potential impact of such a penalized setting for emitting technologies. In this case the algorithm updates the set of installed capacities quite significantly, by shifting essentially all generation from carbon emitting technologies to technologies 1 - 3 not emitting CO2 at all. The algorithm performs a total of 37 iterations and 28 serious steps. The resulting total amount of installed capacity is 3500, slightly less that the nominal vector. This can be explained since the original setting was slightly overcapacitated - and as a result of the introduction of the "fictive" imbalance unit - infeasibility is no longer an issue. The maximum load over the considered scenarios being roughly 3500 as well. The overall objective function value is of the same magnitude as the earlier run. This numerical experiment thus clearly shows that the algorithm has potential to provide meaningful solutions for this type of problem.

## 6.2 Decision dependent probability constraints in two stage problems

In this section we consider the following stochastic problem having different random vectors:

$$\begin{cases} \min\limits_{x\in X} & f_1(x) + \sum\limits_{s=1}^{S} \pi_s Q(x;\xi^s) \\ \text{s.t.} & \mathbb{P}[A_1 x + b_1 \geq \omega_1] \geq p_1 \end{cases} \quad \text{with} \quad Q(x;\xi) := \begin{cases} \min\limits_{y\in Y} & q(x,y;\xi) \\ \text{s.t.} & \mathbb{P}_x[A_2(\xi)y + b_2(\xi) \geq \omega_2(x)] \geq p_2. \end{cases}$$

In this problem, $\xi \in \Xi := \{\xi^1, \ldots, \xi^S\}$, $\omega_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$, and $\omega_2(x) \sim \mathcal{N}(\mu_2(x), \Sigma_2(x))$. The latter random vector depends on the first-stage decisions. We assume that the covariance matrices $\Sigma_1$ and $\Sigma_2(x)$ are positive definite for all $x \in X$. As a result, the probability functions are twice-differentiable [13, 44]. Furthermore, as the multivariate Gaussian distribution is log-concave, we get that $c_1(x) = \log(p_1) - \log(\mathbb{P}[A_1 x + b_1 \geq \omega_1])$ is a convex function and so is the objective of the penalized subproblem

$$Q^\varepsilon(x;\xi) = \min_{y\in Y} \quad q(x,y;\xi) - \tfrac{1}{\varepsilon}\log\Big(\mathbb{P}_x[A_2(\xi)y + b_2(\xi) \geq \omega_2(x)] - p_2\Big).$$

We are thus in the setting of Example 1.1 with $f_2(x) = \sum_{s=1}^{S} \pi_s[-Q^\varepsilon(x;\xi^s)]$. We can observe that the just given optimization problem is a version of two-stage stochastic program having unhedgeable, or post-decision random realizations.

Now in order to compute the gradient of both of the involved probability functions, we can rely on two different formulæ for the gradients. The mapping $c_1$ is continuously differentiable and its gradient can be evaluated by employing the results shown in [48]. The second stage probability function is also differentiable under a mild regularity condition, its gradient can be evaluated using the formulæ from [42, Thm. 5.1]. Indeed with $L_2(x)$ the matrix resulting from the Cholesky decomposition $\Sigma_2(x) = L_2(x)L_2(x)^\top$, we may write

$$c_2^i(x,y) = \mathbb{P}[-A_2(\xi^i)y - b_2(\xi^i) + \mu_2(x) + L_2(x)\omega_2 \leq 0],$$

where $\omega_2 \in \mathbb{R}^m$, $\omega_2 \sim \mathcal{N}(0, I)$. Hence we can observe that:

$$\nabla c_2^i(x,y) = \int\limits_{v\in\mathbb{S}^{m-1}: J^*(v)\neq\emptyset, |J^{**}(v)|=1} -\frac{\chi(\hat{\rho}(v))}{(L_2(x))_{j(v)}v}\Big(\hat{\rho}(v)\nabla(L_2(x))_{j(v),.}v) + \nabla\mu_2(x), -(A_2)^\top_{j(v),:}\Big)\, d\mu_\zeta(v) \tag{6.3}$$

with

$$\begin{aligned} J^*(v) &= \{j = 1, ..., r \ : (L_2(x))_j v > 0\} \\ \hat{\rho}(v) &= \min_{j\in J^*(v)} \frac{A_2 y + b_2 - \mu_2(x)}{(L_2(x))_j v} \\ J^{**}(v) &= \left\{ j \in J^*(v) \ : \hat{\rho}(v) = \frac{A_2 y + b_2 - \mu_2(x)}{(L_2(x))_j v} \right\}. \end{aligned}$$

and $j(v)$ being the unique element of $J^{**}(v)$. In this case since $L_2$ has linearly independent rows - which is the case since $\Sigma_2$ is positive definite - the aforementioned regularity condition (R2CQ) holds true. In fact (R2CQ) holds true locally and as a consequence it is indeed so that both $c_1$ and $c_2$ are twice continuously differentiable. This was already clear for $c_1$ upon using well known classic arguments.

Let us now consider the following concrete example of a problem of this kind. We are interested in a situation considering a manufacturer capable of producing two different products. The first-stage decision variables of the problem consist of setting prices for the products and an advertisement levels. The price will be assumed to be in relation to the average second-stage demand for the given product. We will use the following rule $\mu_2(x) = (\bar{\mu}_1/x_1, \bar{\mu}_2/x_2)$, with $x_1, x_2$ being the price levels for product 1 and 2 respectively. Advertisement is assumed to have a beneficial effect on the variance of the demand, but simultaneous advertisement for both products will be counterproductive. In other words:

$$\Sigma_2(x) := \begin{bmatrix} (0.1\bar{\mu}_1/x_1x_3)^2 & -0.4(0.01\bar{\mu}_1/x_1x_3\bar{\mu}_2/x_2x_4) \\ -0.4(0.01\bar{\mu}_1/x_1x_3\bar{\mu}_2/x_2x_4) & (0.1\bar{\mu}_2/x_2x_4)^2 \end{bmatrix}.$$

The second stage decision $y$ involves the production of the goods. The production process of the goods is subject to some possible unreliability as the amount of actually produced goods are concerned. The matrix $A_2$ is thus a diagonal matrix, where the first entry is a uniform random variable over the interval $[0.9, 1]$ - on average only 95% of the commissioned products actually get manufactured. The second diagonal entry is uniform over the interval $[0.8, 1]$ - the process of production here is more unreliable. However producing with the more unreliable process is slightly cheaper. Any products that are manufactured but not sold, will incur a penalty. The second stage cost function is thus given by

$$q(x, y, \xi) = (2 - x_1)y_1 + (1 - x_2)y_2 + 12\mathbb{E}[\max\{y_1 - (\omega_2)_1, 0\}] + 12\mathbb{E}[\max\{y_2 - (\omega_2)_2, 0\}].$$

The last two terms correspond to the penalization of produced, but not sold goods. It turns out that the latter expectations can be computed "analytically" as they are related to the computation of an expectation of a truncated Gaussian random variable. Therefore, we can observe that the following identity holds true:

$$\begin{aligned}\mathbb{E}[\max\{y_1 - (\omega_2)_1, 0\}] = {} & \Phi((y_1 - (\mu_2(x))_1)x_1/(0.1\bar{\mu}_1x_3))(y_1 - (\mu_2(x))_1) \\ & - \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}((y_1-(\mu_2(x))_1)x_1/(0.1\bar{\mu}_1x_3))^2} 0.1\bar{\mu}_1/x_1x_3.\end{aligned}$$

The second formula is of course immediately deduced as it is analogous. Both products require a different setup of the factory, so $y_1 + y_2 \leq 10$. Furthermore, the first-stage cost function is related to the cost of advertisement $f_1(x) = q_1x_3^2 + q_2x_4^2$. Furthermore all first-stage variables are bounded.

The implementation of this example requires first the implementation of the formulae for the gradient of the probability function. Here we can exploit the earlier given formula immediately. It can be observed (see the more extensive discussion in [37]) that the probability value itself can be computed with exactly the same cost. Subsequently the algorithm scheme is very similar to the one of the investment problem. In particular, combining the computations for probability function value and subgradient (6.3) with the reasoning of the previous example, we will obtain the oracle for $f_2$ component (in the notation of Subsection 4.1), while the oracle for $f_1$ is straightforward from the formula of the advertisement cost.

Therefore, we have also run this case through the algorithm and found an approximate critical solution after a total of 12 iterations (1600 seconds on personal laptop). The found solution is $x = (3.37, 3.21, 0.096, 0.784)$, showing that there is interest in balancing the prices, i.e., not taking maximal prices, while also investing in advertisement. We have done the same test with IPOPT solver. The computation was aborted after 50 000 seconds with the resulting infeasible point $x = (9.997, 9.983, 0.0004, 0.167)$ slowly approaching the bound $(10, 10, 0, 0)$.

This example thus shows that the new algorithm allows us to consider settings beyond classic convexity, even when dealing with probability functions - in this case with decision dependent random vectors.

## 6.3 Highly nonconvex chance-constrained problem

In this section we investigate the following optimization problem (having weakly-convex constraint):

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & c^\top x \\ \text{s.t.} \quad & \mathbb{P}\left[\frac{1}{2}\xi^\top Q_j(x)\xi + q_j(x)^\top\xi + d_j(x) \leq 0,\ j = 1, ..., k\right] \geq p \\ & \underline{x} \leq x \leq \bar{x}. \end{aligned} \tag{6.5}$$

We first note that as a result of [46] and the upfollowing concrete data, that the probability function is continuously differentiable. Furthermore the underlying feasible set is compact and so we are in case ii) of the introduction: $c_1(x) = p$ and $-c_2$ indicating the probability function. The underlying data is not convex in the parameter replaced by the random vector. As a result, the underlying feasible set is not expected to be convex. Concretely we will consider the following data, for $k = 2$:

$$Q_1(x) = \begin{bmatrix} 3(x_1 - 1) & -x_2 \\ -x_2 & 3(x_1 - 1) \end{bmatrix}, \quad Q_2(x) = \begin{bmatrix} -2x_2 & x_1 - 1 \\ x_1 - 1 & -2x_2 \end{bmatrix}$$

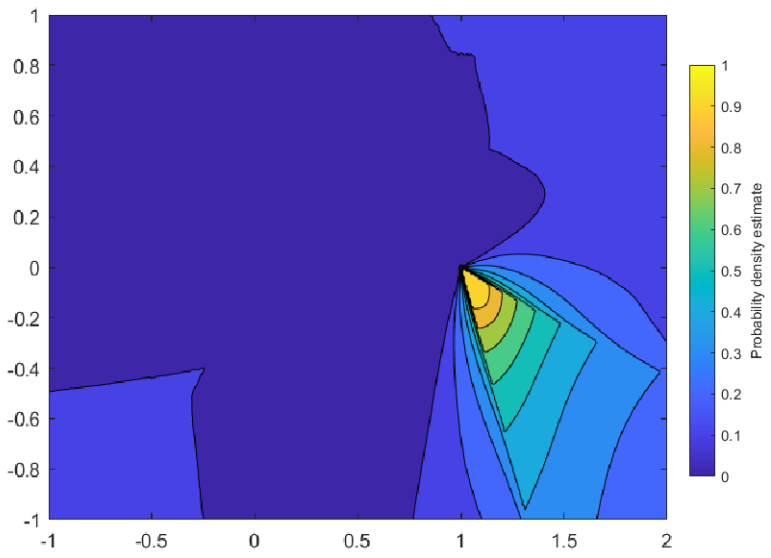


Figure 2: Probability distribution associated with the chance constraint in (6.5)

as well as

$$q_1(x) = \begin{bmatrix} 3 \\ 1 \end{bmatrix} (x_1 - 1), \quad q_2(x) = \begin{bmatrix} 1 \\ 4 \end{bmatrix} x_2$$

$d_1(x) = -2, d_2(x) = -2$. We have also picked $p = 0.7$ together with $c = (-1, -1)$, $\underline{x} = (-2, -2)$, $\overline{x} = (2, 2)$. The random vector is taken to be multivariate Gaussian with mean vector 0 and covariance matrix

$$\Sigma = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}.$$

This optimization problem is quite challenging. First it can be observed that an alternative sample average approximation along the lines of [22] would be a MILP. It is thus tempting to first try to solve the resulting optimization problem with such a formulation. We have done so for the following sample sizes $\{100, 1000, 5000, 10000, 50000\}$, using CPLEX 12.10. The resulting computation times are $0.5, 4, 22, 100, >$ 8000 seconds respectively. The last computation was aborted still showcasing a 25.9% gap after more than 2 hours of computation. Unfortunately, none of the obtained solutions turn out to be feasible, quite to the contrary: the typically obtained final probability value is roughly 0.04 being far from the required 0.7. We have also performed a run of a sampled problem with 10000 random realizations, but with a significantly higher probability level of 0.9. In this case, after roughly one hour of computation, the resulting solution being at a MIPgap of 3.3 %, is still highly infeasible having only a probability value of 0.02. Furthermore, we have tested IPOPT solver for the problem resolution: tests have been performed for the six initial points listed in Table 6.3. After at most 26 seconds of computation, IPOPT halted with highly infeasible points with a probability constraint value equal to 0. The difficulty of generating feasible points might come from the form of probability distribution as the probability level sharply raises from zero (blue, Figure 2) to nearly 1 (yellow, Figure 2) when approaching the feasible area from most directions, which causes the loss of gradient information in a large zone of zero probability. But of course this gradient information is not exploited by the MILP formulation at all.

In contrast, as Table 6.3 shows, our algorithm manages to improve the probability level if the starting point is infeasible, and to improve the objective function value while satisfying probability constraint for a feasible initial case. Moreover, for one of the tested starting points we have managed to generate the near (globally) optimal solution $(1.2400; -0.1126)$. Since the problem is indeed very difficult, a precise internal sampling scheme for the probability function is required. This amounts to the number of samples used to compute a formula of the type (6.3), which subsequently leads to design of the oracle for $c_2$ component. We can prematurely end further sampling, very much like the implementation of Genz' code [9], by checking if sampling variance - in fact the confidence interval bounds - are sufficiently small. Unlike Genz' code a

| Initial point | Time (s) | Iterations | Initial probability level | Final probability level | Objective value |
|---|---|---|---|---|---|
| $(1.2, -0.1)^\top$ | 47.6 | 17 | 0.80 | 0.70 | -1.1097 |
| $(1.5, -0.4)^\top$ | 78.4 | 28 | 0.43 | 0.70 | -1.1346 |
| $(1.6, -0.3)^\top$ | 36.9 | 14 | 0.42 | 0.70 | -1.0815 |
| $(1.7, -0.4)^\top$ | 59.8 | 20 | 0.36 | 0.70 | -1.1443 |
| $(1.1, -0.4)^\top$ | 21.3 | 10 | 0.43 | 0.43 | -0.7001 |
| $(1, -0.8)^\top$ | 10.8 | 9 | 0.18 | 0.18 | -0.2001 |

Table 2: Results obtained with Algorithm 1 for problem (6.5) depending on the initial point

crude antithetic Monte-Carlo scheme has been used for sampling, thus leaving much room for significant improvement in terms of speed and precision, by using for instance QMC as, for instance, in [11].

### 6.4 Compressed sensing problem

In this section we focus on the problem of compressed sensing considered in [51]:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \|x\|_1 - \|x\| \\ \text{s.t.} \quad & \|Ax - b\|_{LL_2,\gamma} \leq \delta, \end{aligned} \tag{6.6}$$

where $A \in \mathbb{R}^{q\times n}$ is a full row rank matrix and $b \in \mathbb{R}^q$. For given $\gamma > 0$, Lorentzian norm $\|\cdot\|_{LL_2,\gamma}$ of a vector $y \in \mathbb{R}^q$ is defined as

$$\|y\|^2_{LL_2,\gamma} = \sum_{i=1}^{q} \log\left(1 + \frac{y_i^2}{\gamma^2}\right).$$

As discussed in [51], the problem (6.6) is DC with twice continuously differentiable constraint, whose modulus of Lipschitz gradient is known. This allows us to construct the oracle for the constraint component. To compute subgradient of the component $f_1(x) = \|x\|_1$, we have chosen the sign function.

As in [51], we have taken $q = 3\,600$, $n = 12\,800$ and generated $A \in \mathbb{R}^{q\times n}$ with normally distributed random entries normalizing it so that each column has a unit norm. To set the original point, we have chosen a subset of size $s_0 = [\frac{q}{9}]$ among basis vectors and generated a $s_0$-sparse vector $x_{orig}$ with i.i.d. normally distributed random entries. We have taken $b = Ax_{orig} + 0.01\eta$, each $\eta_i$ having a standard Cauchy distribution, and $\delta = 1.1\|0.01\eta\|_{LL_2,\gamma}$ with $\gamma = 0.02$.

The CwC-bundle algorithm manages to obtain a feasible solution after 60 iterations and 52 serious steps, as well as to recover a global solution within tolerance $9.5 \times 10^{-4}$ after 2400 iterations and 1613 serious steps with execution time of 8800 seconds (on personal laptop). This example shows the ability of our algorithm to solve problems with an explicit DC structure, even if its performance is not as good as for the method in [51] developed for more specific framework.

## 7 Conclusion

In this manuscript, we have considered nonsmooth and nonconvex optimization problems where the objective function and nonlinear constraint are represented as the difference of convex and weakly convex functions (CwC). Our work studies various stationary conditions and a bundle method approach enabling to compute critical (generalized KKT) points. The latter broadens and enhances the algorithm developed in [49] for

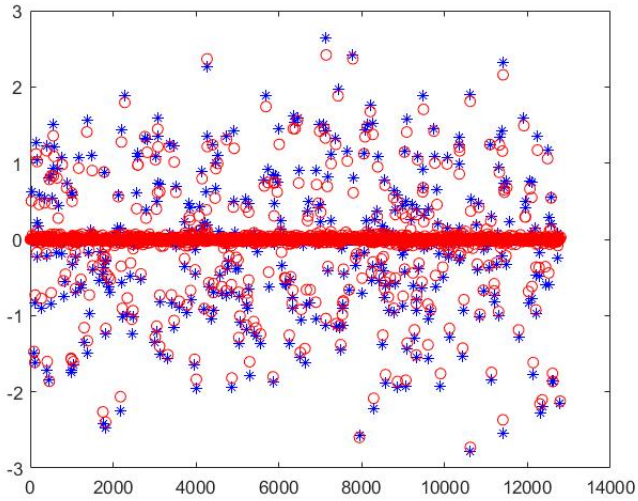


Figure 3: Computed solution (marked by circle) of (6.6) and $x_{orig}$ (marked by asterisk)

the case of Difference-of-Convex (DC)-constrained DC-problems, and likewise relies on problem reformulation via an improvement function. To the best of our knowledge, proposed method is the first one that directly exploits the CwC-structure of the involved functions and does not require additional assumptions or transformations as, for instance, explicit Difference-of-Convex decompositions or Moreau envelopes. We have illustrated the method performance with a few stochastic problems, including two-stage and chance-constrained problems and a compressed sensing problem with nonlinear constraint. Preliminary results are meaningful and show that the algorithm enables tackling settings beyond the classic Difference-of-Convex setting.

# Appendices

## A Some mathematical results

**Proof of Proposition 2.4.** Since $\mathsf{f} : \mathcal{O} \to \mathbb{R}^n$ is weakly convex, it follows by definition that, relative to some neighbourhood $V_{x'}$ of each point $x' \in \mathcal{O}$, there exist $\mu_{x'} > 0$ such that for all $\mu \geq \mu_{x'}$ the function $\phi(x) = \mathsf{f}(x) + \frac{\mu}{2}\|x\|^2$ is finite and convex on $V_{x'}$. In such a representation, there is no loss of generality in assuming that $V_{x'} \subset \mathcal{O}$ (if necessary we can define a new/smaller neighbourhood as $V_{x'} \cap \mathcal{O}$ for which the above conclusion obviously stands). By considering all the points in $X$, let $V := \{V_{x'} : x' \in X\}$ be the collection of all such neighbourhoods. Then, by construction, $V$ is an open cover of the compact set $X$ and, by definition of compactness, it has a finite open subcover, i.e., there exists finitely many points $\{x'_1, \ldots, x'_m\} \subset X$ such that $\mathcal{O}' := \cup_{i=1}^m V_{x'_i} \supset X$, and by construction $\mathcal{O}'$ is an open subset of $\mathcal{O}$. The first part of item i) thus follows by taking $\mu_{\mathsf{f}} := \max_{i=1,\ldots,m} \mu_{x'_i} < \infty$. By writing $\mathsf{f}(x) = \phi(x) - \frac{\mu}{2}\|x\|^2$ and recalling Proposition 2.1 i) we get $\partial^{\mathsf{C}}\mathsf{f}(x) = \partial\phi(x) - \mu x$ for all $x \in \mathcal{O}'$. This concludes item i).

To show item ii), let us now define $\tilde{\phi}(x) = \mathsf{f}(x) + \frac{\mu}{2}\|x\|^2 + i_{\mathcal{O}'}(x)$, an extended real-valued convex function: $\tilde{\phi} : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. Note that for each $x \in \mathcal{O}'$, there exists a neighbourhood $V_x \subset \mathcal{O}'$ such that $\tilde{\phi}(x') = \phi(x')$ for all $x' \in V_x$. This fact permits us to conclude that $\partial\tilde{\phi}(x) = \partial\phi(x)$ for all $x \in \mathcal{O}'$. It thus follows from item i) that, for every $x \in \mathcal{O}'$ and every $s \in \partial\tilde{\phi}(x)$, there exists $s_f \in \partial^{\mathsf{C}} f(x)$ such that

$s = s_f + \mu x$ and the subgradient inequality reads as

$$\tilde{\phi}(y) \geq \tilde{\phi}(x) + \langle s_f + \mu x, y - x\rangle \quad \forall y \in \mathbb{R}^n,$$

i.e., $\mathtt{f}(y) + \frac{\mu}{2}\|y\|^2 + i_{\mathcal{O}'}(y) \geq \mathtt{f}(x) + \frac{\mu}{2}\|x\|^2 + i_{\mathcal{O}'}(x) + \langle s_f + \mu x, y - x\rangle$ for all $y \in \mathbb{R}^n$. The latter simplifies to

$$\mathtt{f}(y) + i_{\mathcal{O}'}(y) \geq \mathtt{f}(x) + \langle s_f, y - x\rangle - \frac{\mu}{2}\|y - x\|^2 \quad \forall y \in \mathbb{R}^n.$$

By restricting $y$ to the set $X$ and recalling that $s_f = s - \mu x \in \partial^{\mathsf{C}}\mathtt{f}(x)$ is an arbitrary subgradient (because no restriction was imposed to $s \in \partial\tilde{\phi}(x)$), the above inequality becomes (2.4). □

**Lemma A.1.** *Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be a convex function, and $\{\varphi_\ell\}_{\ell\in\mathbb{N}}$ a sequence of convex functions $\varphi_\ell : \mathbb{R}^n \to \mathbb{R}$ converging pointwise to $\varphi$, i.e., $\lim_{\ell\to\infty} \varphi_\ell(x) = \varphi_\ell(x)$ for every given point $x$. Furthermore, let $\{x^\ell\} \subset \mathbb{R}^n$ be such that $\lim_{\ell\to\infty} x^\ell = \bar{x}$ and $\{\epsilon^\ell\} \subset \mathbb{R}_+$ satisfy $\lim_{\ell\to\infty} \epsilon^\ell = 0$. If $g^\ell \in \partial_{\epsilon^\ell}\varphi_\ell(x^\ell)$ for all $\ell$ and $\lim_{\ell\to\infty} g^\ell = \bar{g}$, then $\bar{g} \in \partial\varphi(\bar{x})$.*

*Proof.* First, let us prove that $\liminf_\ell \varphi_\ell(x^\ell) \geq \varphi(\bar{x})$. Since $\mathtt{dom}(\varphi_\ell) = \mathtt{dom}(\varphi) = \mathbb{R}^n$, it follows from [32, Cor. 2C] that the pointwise convergence of $\{\varphi_\ell\}_{\ell\in\mathbb{N}}$ is equivalent to epi-convergence, which in turn is equivalent (see [32, Eq. (4.2)]) to epi-convergence of $\{\varphi_\ell^*\}_{\ell\in\mathbb{N}}$, the sequence of conjugate functions to $\varphi_\ell$. Hence, it follows that $\lim_\ell \varphi_\ell^*(x) = \varphi^*(x)$ for every given $x \in \mathbb{R}^n$. Now consider the following development:

$$\varphi_\ell(x^\ell) = (\varphi_\ell^*)^*(x^\ell) = \sup_{y\in\mathbb{R}^n} [\langle y, x^\ell\rangle - \varphi_\ell^*(y)] \geq \langle y, x^\ell\rangle - \varphi_\ell^*(y) \quad \forall y \in \mathbb{R}^n.$$

Accordingly, $\liminf_\ell \varphi_\ell(x^\ell) \geq \liminf_\ell [\langle y, x^\ell\rangle - \varphi_\ell^*(y)] = \langle y, \bar{x}\rangle - \varphi^*(y)$ for all $y \in \mathbb{R}^n$, showing that

$$\liminf_\ell \varphi_\ell(x^\ell) \geq \sup_y [\langle y, \bar{x}\rangle - \varphi^*(y)] = (\varphi^*)^*(\bar{x}) = \varphi(\bar{x}).$$

Recall that $g^\ell \in \partial_{\epsilon^\ell}\varphi_\ell(x^\ell)$. Then, $\varphi_\ell(x) \geq \varphi_\ell(x^\ell) + \langle g^\ell, x - x^\ell\rangle - \epsilon^\ell$ for all $x \in \mathbb{R}^n$. By taking the limit when $\ell$ goes to infinity we get

$$\begin{aligned}\varphi(x) = \lim_\ell \varphi_\ell(x) = \liminf_\ell \varphi_\ell(x) &\geq \liminf_\ell [\varphi_\ell(x^\ell) + \langle g^\ell, x - x^\ell\rangle - \epsilon^\ell] \\ &\geq \liminf_\ell \varphi_\ell(x^\ell) + \liminf_\ell \langle g^\ell, x - x^\ell\rangle - \limsup_\ell \epsilon^\ell \\ &\geq \varphi(\bar{x}) + \langle \bar{g}, x - \bar{x}\rangle,\end{aligned}$$

showing that $\bar{g} \in \partial\varphi(\bar{x})$. □

# B Simplified algorithm for the case without nonlinear constraints

This section describes how Algorithm 1 can be simplified to deal with the simpler convexly-constrained problem

$$\min_{x\in X} f(x), \quad \text{with} \quad f(x) = f_1(x) - f_2(x). \tag{B.1}$$

In this case, the problem's model (4.6) reduces to $\check{\mathsf{H}}^k(x; x^{\ell_k}) = \check{f}_1^k(x) - \bar{f}_2^{\ell_k}(x)$, and the descent test (4.5) becomes $f(x^{k+1}) \leq f(x^{\ell_k}) - \frac{\kappa}{2}\|x - x^{\ell_k}\|^2$. Hence, Algorithm 1 boils down to the following plainer scheme.

Convergence analysis for Algorithm 2 follows from that of Algorithm 1 upon several simplifications. Instead of doing this exercise, we simply state the following result.

**Theorem B.1.** *Consider problem (B.1) with $X \neq \emptyset$ a bounded polyhedron contained in the open set $\mathcal{O}$, $f_1 : \mathcal{O} \to \mathbb{R}$ convex, and $f_2 : \mathcal{O} \to \mathbb{R}$ weakly convex on some neighbourhood of each $x \in \mathcal{O}$. If in Algorithm 2 the stopping test tolerance `Tol` is set to zero, then any cluster point $\bar{x}$ of the sequence of stability centers $\{x^{\ell_k}\}$ satisfies the necessary optimality condition (3.3).*

*If `Tol` $> 0$, then the algorithm stops after finitely many steps $k \in \mathbb{N}$ with an approximate critical point $x^{\ell_k}$ of (3.3).*

To have an intuition of why the above theorem is valid, the reader may think of adding a dummy convex nonlinear convex function $c(x) = c_1(x) - 0$ to (B.1) and rely on the results from Sections 4 and 5. Indeed, by selecting a constant $M > 0$ large enough and function $c$ such that $c(x) \leq -M < 0$ for all $x \in X$, we can see that Algorithm 1 applied to (B.1) with the additional and superfluous constraint $c(x) \leq 0$ boils down to Algorithm 2. Furthermore, in this artificial setting, the above convergence result follows directly from Theorem 5.8, item iii).

**Algorithm 2** Proximal Bundle Method for Convexly-Constrained `CwC` programs

**Step 0 (Initialization)** Let $x^0 \in X$, $\kappa \in (0, \frac{1}{2})$, $\kappa \leq \mu^0$, and $\texttt{Tol} \geq 0$ be given. Call the oracles to compute $f_i(x^0)$ and (generalized) subgradients $s_{f_i}^0$, $i = 1, 2$. Define $k := \ell_k = 0$ and $\mathcal{B}_f^0 := \{0\}$.

**Step 1 (Trial point)** Compute $x^{k+1}$ the ($x$-part) solution of the QP

$$\left\{ \begin{array}{lll} \min\limits_{x,\, r} & r + \frac{\mu^k}{2}\|x - x^{\ell_k}\|^2 & \\ \text{s.t.} & \bar{f}_1^j(x) - \bar{f}_2^{\ell_k}(x) - r & \leq 0 \quad \forall\, j \in \mathcal{B}_f^k \\ & x \in X, \quad r \in \mathbb{R}. & \end{array} \right.$$

**Step 2 (Stopping test)** If $\|x^{k+1} - x^{\ell_k}\| \leq \texttt{Tol}$, then stop and return $x^{\ell_k}$.

**Step 3 (Oracles call)** Compute $f_i(x^{k+1})$, and subgradients $s_{f_i}^{k+1}$, $i = 1, 2$.

**Step 4 (Descent test)**

(a) If $f(x^{k+1}) \leq f(x^{\ell_k}) - \frac{\kappa}{2}\|x - x^{\ell_k}\|^2$ holds, then declare a *serious step*: define $\ell_{k+1} := k + 1$, choose $\mathcal{B}_f^{k+1} \subset \{0, \ldots, k+1\}$ with $\{k+1\} \in \mathcal{B}_f^{k+1}$ and arbitrarily select $\mu^{k+1} \in (0, \mu^k]$.

(b) Else, declare a *null step*: define $\ell_{k+1} := \ell_k$ and choose $\mathcal{B}_f^{k+1} \subset \{0, \ldots, k+1\}$ with $\bar{\mathcal{B}}_f^k \cup \{k+1, \ell_k\} \subset \mathcal{B}_f^{k+1}$ ($\bar{\mathcal{B}}_f^k$ as in (4.10)).
Compute $\nu^k := 2 \max\left\{ \frac{\bar{f}_2^{\ell_k}(x^{k+1}) - f_2(x^{k+1})}{\|x^{k+1} - x^{\ell_k}\|^2},\ 0 \right\}$. If $\nu^k \geq \mu^k - 2\kappa$, set $\mu^{k+1} = \nu^k + 1$; otherwise $\mu^{k+1} = \mu^k$.

**Step 5 (Loop)** Set $k := k + 1$ and go back to Step 1.

# Acknowledgments.

W. Oliveira acknowledges financial support from the Gaspard-Monge Program for Optimization and Operations Research (PGMO) project "Scalable Optimization for Learning and Energy Management."

# References

[1] P. Apkarian, D. Noll, and A. Rondepierre. Mixed $H_2/H_\infty$ Control via Nonsmooth Optimization. *SIAM J. Control Optim.*, 47(3):1516–1546, Jan 2008.

[2] J. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization problems*. Springer Ser. Oper. Res. Financ. Eng. . Springer - New York, 1st edition, 2000.

[3] P. Borges, C. Sagastizábal, and M. Solodov. A regularized smoothing method for fully parameterized convex problems with applications to convex and nonconvex two-stage stochastic programming. *Math. Program.*, 189(1):117–149, Sep 2021. doi: 10.1007/s10107-020-01582-2.

[4] F. Clarke. *Optimisation and Nonsmooth Analysis*. Classics Appl. Math. SIAM, 1990.

[5] R. Correa and C. Lemaréchal. Convergence of some algorithms for convex minimization. *Math. Program.*, 62(2):261–275, 1993.

[6] W. de Oliveira. The ABC of DC programming. *Set-Valued Var. Anal.*, 28(4):679–706, 2020.

[7] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Math. Program.*, 178(1):503–558, Nov 2019.

[8] M. Gaudioso, G. Giallombardo, G. Miglionico, and A. M. Bagirov. Minimizing nonsmooth DC functions via successive DC piecewise-affine approximations. *J. Global Optim.*, 71(1):37–55, May 2018. doi: 10.1007/s10898-017-0568-z.

[9] A. Genz. Numerical computation of multivariate normal probabilities. *J. Comput. Graph. Statist.*, 1: 141–149, 1992.

[10] P. Hartman. On functions representable as a difference of convex functions. *Pacific J. Math.*, 9(3): 167–198, 1959.

[11] H. Heitsch. On probabilistic capacity maximization in a stationary gas network. *Optimization*, 69(3): 575–604, 2020. doi: 10.1080/02331934.2019.1625353.

[12] L. Hellemo, P. I. Barton, and A. Tomasgard. Decision-dependent probabilities in stochastic programs with recourse. *Comput. Manag. Sci.*, 15(3):369–395, Oct 2018. doi: 10.1007/s10287-018-0330-0.

[13] R. Henrion and A. Möller. A gradient formula for linear chance constraints under Gaussian distribution. *Math. Oper. Res.*, 37:475–488, 2012. doi: 10.1287/moor.1120.0544.

[14] F. Jara-Moroni, J.-S. Pang, and A. Wächter. A study of the difference-of-convex approach for solving linear programs with complementarity constraints. *Math. Program.*, 169(1):221–254, May 2018.

[15] P. Javal. *Integrating uncertainties in short-term operational planning*. PhD thesis, Mines Paris PSL - Centre de Mathématiques Appliquées - CMA, December 2021. URL https://pastel.archives-ouvertes.fr/tel-03693993v2/document.

[16] W. Khalaf, A. Astorino, P. d'Alessandro, and M. Gaudioso. A DC optimization-based clustering technique for edge detection. *Optim. Lett.*, 11(3):627–640, 2017.

[17] H. A. Le Thi, T. Pham Dinh, and H. V. Ngai. Exact penalty and error bounds in DC programming. *J. Global Optim.*, 52(3):509–535, Mar 2012.

[18] H. A. Le Thi, H. V. Ngai, and P. D. Tao. DC programming and DCA for general DC programs. In T. Van Do, H. A. Le Thi, and N. T. Nguyen, editors, *Advanced Computational Methods for Knowledge Engineering: Proceedings of the 2nd International Conference on Computer Science, Applied Mathematics and Applications (ICCSAMA 2014)*, pages 15–35. Springer International Publishing, 2014.

[19] C. Lemaréchal and A. Renaud. A geometric study of duality gaps, with applications. *Math. Program.*, 90:399–427, 2001.

[20] H. Li and Y. Cui. A decomposition algorithm for two-stage stochastic programs with nonconvex recourse, 2022. URL https://arxiv.org/abs/2204.01269.

[21] T. Lipp and S. Boyd. Variations and extension of the convex–concave procedure. *Optim. Eng.*, 17(2): 263–287, 2016. doi: 10.1007/s11081-015-9294-x.

[22] J. Luedtke and S. Ahmed. A sample approximation approach for optimization with probabilistic constraints. *SIAM J. Optim.*, 19:674–699, 2008.

[23] H. Mine and M. Fukushima. A minimization method for the sum of a convex function and a continuously differentiable function. *J. Optim. Theory Appl.*, 33(1):9–23, Jan 1981. doi: 10.1007/BF00935173.

[24] O. Montonen and K. Joki. Bundle-based descent method for nonsmooth multiobjective DC optimization with inequality constraints. *J. Global Optim.*, 72(3):403–429, Nov 2018.

[25] B. S. Mordukhovich. *Variational Analysis and Applications*. Springer Monogr. Math. Springer, Cham, 2018. doi: https://doi.org/10.1007/978-3-319-92775-6.

[26] B. S. Mordukhovich and N. M. Nam. *An Easy Path to Convex Analysis and Applications*. Synth. Lect. Math. Stat. Springer International Publishing, 2014.

[27] J. S. Pang, M. Razaviyayn, and A. Alvarado. Computing B-stationary points of nonsmooth DC programs. *Math. Oper. Res.*, 42(1):95–118, 2017. doi: 10.1287/moor.2016.0795.

[28] R. Rockafelar. Favorable classes of lipschitz continuous functions in subgradient optimization. In *Progress in Nondifferentiable Optimization*, IIASA Collaborative Proceedings Series, International Institute of Applied Systems Analysis, Laxenburg, Austria, pages 125–144, 1982.

[29] R. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren Math. Wiss.* Springer Verlag Berlin, 3rd edition, 2009. doi: 10.1007/978-3-642-02431-3.

[30] R. T. Rockafellar. Generalized subgradients in mathematical programming. In *Mathematical Programming The State of the Art: Bonn 1982*, pages 368–390. Springer, 1982.

[31] C. Sagastizábal and M. Solodov. An infeasible bundle method for nonsmooth convex constrained optimization without a penalty function or a filter. *SIAM J. Optim.*, 16(1):146–169, 2005. doi: 10.1137/040603875.

[32] G. Salinetti and R. J. Wets. On the relations between two types of convergence for convex functions. *J. Math. Anal. Appl.*, 60(1):211–226, aug 1977.

[33] A. Shapiro and Y. Yomdin. On functions representable as a difference of two convex functions, and necessary conditions in a constrainted optimization. Technical report, Ben-Gurion University of the Negev, 1981. URL https://sites.gatech.edu/alexander-shapiro/publications.

[34] A. S. Strekalovsky and I. M. Minarchenko. A local search method for optimisation problem with d.c. inequality constraints. *Appl. Math. Model.* , 1(58):229–244, 2018.

[35] K. Sun and X. A. Sun. Algorithms for difference-of-convex programs based on difference-of-moreau-envelopes smoothing. *INFORMS J. Optim.*, 2022. doi: 10.1287/ijoo.2022.0087.

[36] W. van Ackooij. Decomposition approaches for block-structured chance-constrained programs with application to hydro-thermal unit commitment. *Math. Methods Oper. Res.*, 80(3):227–253, 2014.

[37] W. van Ackooij. A discussion of probability functions and constraints from a variational perspective. *Set-Valued Var. Anal.*, 28(4):585–609, 2020. doi: 10.1007/s11228-020-00552-2.

[38] W. van Ackooij and W. de Oliveira. Level bundle methods for constrained convex optimization with various oracles. *Comput. Optim. Appl.*, 57(3):555–597, 2014.

[39] W. van Ackooij and W. de Oliveira. Nonsmooth and nonconvex optimization via approximate difference-of-convex decompositions. *J. Optim. Theory Appl.*, 182(1):49–80, 2018. doi: 10.1007/s10957-019-01500-3.

[40] W. van Ackooij and W. de Oliveira. Nonsmooth and nonconvex optimization via approximate difference-of-convex decompositions. *J. Optim. Theory Appl.*, 182(1):49–80, mar 2019.

[41] W. van Ackooij and W. de Oliveira. Addendum to the paper 'nonsmooth DC-constrained optimization: constraint qualification and minimizing methodologies'. *Optim. Methods Softw.*, pages 1–10, May 2022.

[42] W. van Ackooij and R. Henrion. (Sub-) Gradient formulae for probability functions of random inequality systems under Gaussian distribution. *SIAM/ASA J. Uncertain. Quantif.* , 5(1):63–87, 2017. doi: 10.1137/16M1061308.

[43] W. van Ackooij and J. Malick. Decomposition algorithm for large-scale two-stage unit-commitment. *Ann. Oper. Res.*, 238(1):587–613, 2016. doi: 10.1007/s10479-015-2029-8.

[44] W. van Ackooij and J. Malick. Second-order differentiability of probability functions. *Optim. Lett.*, 11 (1):179–194, 2017. doi: 10.1007/s11590-016-1015-7.

[45] W. van Ackooij and N. Oudjane. On investment in power systems. *Preprint*, pages 1–12, 2022.

[46] W. van Ackooij and P. Pérez-Aros. Gradient formulae for probability functions depending on a heterogenous family of constraints. *Open J. Math. Optim.*, 2:1–29, 2021. doi: 10.5802/ojmo.9.

[47] W. van Ackooij and C. Sagastizábal. Constrained bundle methods for upper inexact oracles with application to joint chance constrained energy problems. *SIAM J. Optim.*, 24(2):733–765, 2014.

[48] W. van Ackooij, R. Henrion, A. Möller, and R. Zorgati. On probabilistic constraints induced by rectangular sets and multivariate normal distributions. *Math. Methods Oper. Res.*, 71(3):535–549, 2010.

[49] W. van Ackooij, S. Demassey, P. Javal, H. Morais, W. de Oliveira, and B. Swaminathan. A bundle method for nonsmooth DC programming with application to chance-constrained problems. *Comput. Optim. Appl.*, 78(2):451–490, 2021.

[50] J.-P. Vial. Strong and weak convexity of sets and functions. *Math. Oper. Res.*, 8(2):231–259, 1983.

[51] P. Yu, T. K. Pong, and Z. Lu. Convergence rate analysis of a sequential convex programming method with line search for a class of constrained difference-of-convex optimization problems. *SIAM J. Optim.*, 31(3):2024–2054, 2021. doi: 10.1137/20M1314057.