

# Computing B-Stationary Points of Nonsmooth DC Programs\*

Jong-Shi Pang<sup>†</sup>

Meisam Razaviyayn<sup>‡</sup>

Alberth Alvarado<sup>§</sup>

Original: October 2014; revised August 2015

## Abstract

Motivated by a class of applied problems arising from physical layer based security in a digital communication system, in particular, by a secrecy sum-rate maximization problem, this paper studies a nonsmooth, difference-of-convex (dc) minimization problem. The contributions of this paper are: (i) clarify several kinds of stationary solutions and their relations; (ii) develop and establish the convergence of a novel algorithm for computing a d-stationary solution of a problem with a convex feasible set that is arguably the sharpest kind among the various stationary solutions; (iii) extend the algorithm in several directions including: a randomized choice of the subproblems that could help the practical convergence of the algorithm, a distributed penalty approach for problems whose objective functions are sums of dc functions, and problems with a specially structured (nonconvex) dc constraint. For the latter class of problems, a pointwise Slater constraint qualification is introduced that facilitates the verification and computation of a B(ouligand)-stationary point.

## 1 Introduction

A general difference-of-convex (dc) optimization problem refers to the minimization/maximization of an objective function that is the difference of two convex functions subject to constraints defined by functions of the same kind. Such optimization problems form a large class of nonconvex programs and have been studied extensively for more than three decades in the mathematical programming literature [21, 11, 13, 24, 26, 25, 44, 32, 45]. In particular, Pham Think Tao and Le Thi Hoai An have made pioneering contributions to this important subfield of contemporary optimization and are responsible for much of the development of theory, algorithms, and applications of dc programming to date. See the cited references [21, 17, 18, 13, 22, 32] for a sample of their voluminous writings in this area. In particular, the DCA (Difference-of-Convex Algorithm) has been a principal algorithm for computing a *critical point* of the problem.

Our interest in this class of nonconvex optimization problems stemmed initially from a particular application pertaining to physical layer based security in a digital communication system [1, 2] and a related one of joint base-station assignment and power allocation [38]. A first glance at their formulations does not immediately reveal that the resulting nonsmooth maximization problem (see (3)) is of the dc type. Yet, a careful look at the objective function shows that it can be expressed as the difference of two concave functions, one of which is differentiable and the other one is not. Furthermore, via a “lifting” of the

---

\*This work was based on research partially supported by the U.S. National Science Foundation grants CMMI 1402052 and 0969600.

<sup>†</sup>Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, California 90089-0193, U.S.A. *Email:* jongship@usc.edu.

<sup>‡</sup>Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A. *Email:* razav002@umn.edu

<sup>§</sup>Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, U.S.A. *Present address:* Department of Applied Mathematics, Universidad Galileo, Guatemala, 01010, Guatemala. *Email:* alberth@galileo.edu

problem using some auxiliary continuous variables to express a discrete pointwise maximum function as a value function of an optimization problem over the unit simplex, this applied problem can be formulated as a smooth, bi-concave (thus nonconcave), linearly constrained maximization problem. This special problem raises several interesting questions that do not seem to have been adequately addressed in the existing literature of dc programming. As a linearly constrained dc program, one can speak about the concept of a d(irectional)-stationary point of the problem, i.e., a point at which the one-sided derivatives of the objective function along any feasible (equivalently, tangent) direction are nonnegative. Since the lifted formulation is smooth, one can speak about the standard concept of stationarity, which we call *lifted stationarity*, in terms of the gradient of the objective function in the lifted space. The following is a set of questions that have partially motivated our research: for a dc program with convex constraints,

- (a) How are the concepts of a critical point and d-stationarity related to each other?
- (b) How is lifted stationarity defined in general? How is it related to criticality and d-stationarity of the un-lifted problem?
- (c) Are there algorithms that can provably compute a d-stationary point?

Providing answers to the first two questions constitute a major part of our study. In so doing, we are led to the contention that d-stationarity is arguably the sharpest among these stationarity concepts and yet the computation of such a point by an existing provably convergent algorithm seems to have been elusive to date. This lack of a computational scheme for obtaining a d-stationary point of a convex-constrained dc program leads to the other part of our work, namely, to propose a novel iterative algorithm to fill this gap. The design of the algorithm is interesting in its own right, namely, it contains innovative ideas that do not seem to have been introduced in the dc literature; in particular, we present a randomized version of the algorithm to deal with a potential weakness of the algorithm in practical implementation. Convergence of the algorithms is established.

Also included in our algorithmic development is the extension of the basic algorithm to a multi-agent context where the objective is the sum of dc functions with each summand being a private objective function (with coupled variables) of an individual agent. [The applied problem mentioned at the beginning of this introduction is a problem of this kind.] In such a context, it is desirable to develop a distributed algorithm wherein the optimization of each agent can be carried out independently of the other agents. The design of such a distributed algorithm is another major contribution of our work. This is accomplished via a double iteration wherein the outer loop is a penalty-based scheme and each inner loop applies the newly developed algorithm for computing a d-stationary point of a penalized subproblem. The separability into individual agent-based optimization occurs naturally in the latter loop.

Our last contribution is the extension of the basic algorithm to allow for the presence of a non-differentiable dc constraint, leading what has been called a *general dc program* [18, 20, 32]. Such a constraint adds considerable complication to the theory and computation for a convex feasible set, due to the nonsmoothness and nonconvexity of the dc constraint. For such a dc constrained dc program, we formally define the concept of a B(ouligand)-stationary point and show how it can be characterized by a reasonable number of convex programs, thus making the verification of such a stationary point practically implementable. A provably convergent algorithm is then developed for computing such a point.

## 2 Motivating Applied Problems

In this section, we discuss two applied problems pertaining to power allocation in digital communication systems that had motivated our research. These problems lead us to a unified class of value functions of a continuum family of bivariate functions which we show are of the dc type. For more applications of dc

programs to communication systems and other domains, we refer the readers to [17, 21]. Subsequent to the completion of the original version of this paper, the authors recognized that many interesting classes of nonconvex optimization problems are actually of the dc kind and have not been treated as such; examples include those arising from deviation measures in risk analysis as well as in the regularization of loss functions in statistical learning. Due to space limitations, we cannot give details of these other problems. We are hopeful that our work herein opens renewed opportunities for the dc methodology, in particular the results and algorithms in this paper and those existed in the literature, to be applied to deal with these nonconvex problems more effectively.

The concept of secrecy capacity is of fundamental importance in information theory [15]. Based on this concept, a design problem in physical layer based security is to allocate power budget to the network spectra so that the transmissions between legitimate parties can be kept secure. The problem stated below pertains to the single-input-single-output (SISO) paradigm where users of the network (consisting of transmitter-receiver pairs) communicate over multiple non-orthogonal subchannels. There are also a number of “friendly” jammers and one eavesdropper. Each legitimate user’s transmitter wants to communicate (in a secure way) with its corresponding receiver over a set of parallel subchannels. The friendly jammers are entities willing to cooperate with the legitimate parties by introducing judicious interferences so as to impair the eavesdropper’s ability to decode the messages between intended nodes. With  $H_{rq}(k)$ ,  $H_{r0}(k)$  and  $\hat{H}_{j0}(k)$  denoting the channel gains and  $\sigma_q^2(k)$  the variances of channel noise, all being constants in the model, and  $p_q(k)$  and  $\hat{p}_j(k)$  denoting, respectively, the variable power of user  $q$  and jammer  $j$  allocated to channel  $k$ , this multi-jammer secrecy-sum-rate maximization problem is formulated as follows [2]:

$$\begin{aligned} & \underset{(\mathbf{p}, \hat{\mathbf{p}}) \geq \mathbf{0}}{\text{maximize}} && \sum_{q=1}^Q \sum_{k=1}^N [R_{qqk}(\mathbf{p}, \hat{\mathbf{p}}) - R_{q0k}(\mathbf{p}, \hat{\mathbf{p}})]^+ \\ & \text{subject to:} && \sum_{k=1}^N p_q(k) \leq P_q^{\max} \quad \forall q = 1, \dots, Q \quad (\text{agents' private constraints}) \quad (1) \\ & \text{and} && \sum_{k=1}^N \hat{p}_j(k) \leq \hat{P}_j^{\max} \quad \forall j = 1, \dots, J \quad (\text{coupling constraints}), \end{aligned}$$

where  $\mathbf{p} \triangleq ((p_q(k))_{k=1}^N)_{q=1}^Q$ ,  $\hat{\mathbf{p}} \triangleq ((\hat{p}_j(k))_{k=1}^N)_{j=1}^J$ ,  $[\bullet]^+ \triangleq \max(0, \bullet)$  is the plus-function, and  $R_{qqk}(\mathbf{p}, \hat{\mathbf{p}})$  and  $R_{q0k}(\mathbf{p}, \hat{\mathbf{p}})$  are the Shannon information rate functions given by:

$$\begin{aligned} R_{qqk}(\mathbf{p}, \hat{\mathbf{p}}) &\triangleq \log \left( 1 + \frac{H_{qq}(k) p_q(k)}{\sigma_q^2(k) + \sum_{q \neq r=1}^Q H_{rq}(k) p_r(k) + \sum_{j=1}^J \hat{H}_{jq}(k) \hat{p}_j(k)} \right) \\ R_{q0k}(\mathbf{p}, \hat{\mathbf{p}}) &\triangleq \log \left( 1 + \frac{H_{q0}(k) p_q(k)}{\sigma_q^2(k) + \sum_{q \neq r=1}^Q H_{r0}(k) p_r(k) + \sum_{j=1}^J \hat{H}_{j0}(k) \hat{p}_j(k)} \right); \end{aligned}$$

both  $R_{qqk}$  and  $R_{q0k}$  are clearly differentiable differences of concave functions. Since the pointwise maximum of a finite number of dc functions is a dc function [13, 32], and so is the sum of a finite number of dc functions, it follows that the objective function of (1) is a dc function.

A related problem is that of optimal joint base station assignment and power allocation in a communication network [38]. Admitting a similar formulation with binary variables subject to a knapsack constraint, thus with multiple (more than two) discrete choices, this problem is

$$\begin{aligned}
& \underset{x_q, y_q}{\text{maximize}} && \sum_{q=1}^Q \sum_{\ell=1}^L y_{q\ell} \left[ \sum_{k=1}^N \log \left( 1 + \frac{H_{q\ell}(k) x_{q\ell}(k)}{\sigma_\ell(k)^2 + \sum_{r \neq q} y_{r\ell} H_{r\ell}(k) x_{rk}(f)} \right) - \underbrace{c_{q\ell}}_{\text{set-up cost}} \right] \\
& \text{subject to} && \sum_{\ell=1}^L \sum_{k=1}^N x_{q\ell}(k) \leq B_q^{\max} \\
& \text{and} && \text{for all } q = 1, \dots, Q, \\
& && 0 \leq x_{q\ell}(k) \leq \text{CAP}_{q\ell}(k), \quad \forall \ell = 1, \dots, L \text{ and } k = 1, \dots, N \\
& && \sum_{\ell=1}^L y_{q\ell} = 1, \quad y_{q\ell} \in \{0, 1\}, \quad \forall \ell = 1, \dots, L,
\end{aligned} \tag{2}$$

where the additional index  $\ell = 1, \dots, L$  labels the base stations. The problem is equivalent to

$$\begin{aligned}
& \underset{x_q}{\text{maximize}} && \theta(x) \triangleq \underbrace{\sum_{q=1}^Q \underbrace{\text{maximum}_{y_q \in Y_q} \sum_{\ell=1}^L y_{q\ell} \left[ \sum_{k=1}^N \log \left( 1 + \frac{H_{q\ell}(k) x_{q\ell}(k)}{\sigma_\ell(k)^2 + \sum_{r \neq q} y_{r\ell} H_{r\ell}(k) x_{rk}(f)} \right) - c_{q\ell} \right]}_{\text{pointwise max of dc functions}}}_{\text{pointwise max of dc functions}} \\
& \text{subject to} && \text{for all } q = 1, \dots, Q, \\
& && \sum_{\ell=1}^L \sum_{k=1}^N x_{q\ell}(k) \leq B_q^{\max} \\
& \text{and} && 0 \leq x_{q\ell}(k) \leq \text{CAP}_{q\ell}(k), \quad \forall \ell = 1, \dots, L \text{ and } k = 1, \dots, N,
\end{aligned}$$

where  $Y_q \triangleq \left\{ y_q \in \{0, 1\}^L \mid \sum_{\ell=1}^L y_{q\ell} = 1 \right\}$  remains a discrete set that contains a sum constraint which in

some applications could be generalized to a cardinality constraint of the type:  $K \geq \sum_{\ell=1}^L y_{q\ell} \geq 1$ . Since each  $Y_q$  is a discrete set, it follows again that the above objective  $\theta(x)$  is a dc function.

## 2.1 A digression: Continuum family of dc functions

While it is known that the pointwise maximum of finitely many dc functions is a dc function, it is not known whether this dc property extends to the pointwise maximum of a continuum family of dc functions. It turns out that this extension does not hold in general as the example below shows.

**Example 1.** Let  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  be a (globally) Lipschitz continuous function that is not directionally differentiable everywhere. Such a function exists as provided by a component function of the Euclidean projector onto a specially constructed compact convex set (see [43] for  $n = 2$  and [16] and [9, Exercise 4.8.5] for  $n = 3$ ). Let  $L > 0$  be a Lipschitz constant of  $g$  in the  $\ell_1$ -norm; i.e.,

$$|g(x) - g(y)| \leq L \|x - y\|_1, \quad \forall x, y \in \mathbb{R}^n.$$

Define  $f(x, y) \triangleq g(y) - L\|x - y\|_1$ . Obviously,  $g(x) = \underset{y \in \mathbb{R}^n}{\text{maximum}} f(x, y)$ . It is clear that  $f(\bullet, y)$  is a concave, thus dc, function. Yet  $g$  cannot be a dc function as every dc function must be directionally differentiable but  $g$  is not.  $\square$

In what follows, we present a class of value functions of bivariate functions that preserves the dc property; it turns out the structure of the component functions is important; such a structure includes the case of finitely many dc functions, and in particular the two problems (1) and (2). Specifically, consider the following non-convex, non-differentiable multi-agent optimization problem:

$$\underset{x \in X}{\text{maximize}} \theta(x) \triangleq \sum_{i=1}^I \left( \underset{\lambda^i \in \Lambda^i}{\text{maximum}} \sum_{j=1}^J h_{i,j}(\lambda^i) f_{i,j}(x) \right), \quad (3)$$

where the set  $X \subseteq \mathbb{R}^n$  is closed and convex, and for each  $i = 1, \dots, I$  (denoting the agents' labels),  $\Lambda^i \subseteq \mathbb{R}^{m_i}$  is a compact set (not assumed to be convex; cf. e.g., the set  $Y_q$  in the previous subsection). For each  $i = 1, \dots, I$  and  $j = 1, \dots, J$ ,  $f_{i,j} : \Omega \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ , where  $\Omega$  is an open convex superset of  $X$ , is either convex or concave on  $X$ . Finally, for each  $i = 1, \dots, I$  and  $j = 1, \dots, J$ ,  $h_{i,j} : \Omega^i \subseteq \mathbb{R}^{m_i} \rightarrow \mathbb{R}$ , where  $\Omega^i$  is open convex set containing  $\Lambda^i$ , is such that each product  $h_{i,j}(\lambda^i) f_{i,j}(x)$  is concave in  $\lambda^i$  for fixed  $x$ . A particularly important special case of  $\Lambda^i$  is when it is a unit simplex and each function  $h_{i,j}$  is affine so that the continuous pointwise maximum becomes a discrete pointwise maximum and the overall problem (3) is as follows:

$$\underset{x \in X}{\text{maximize}} \sum_{i=1}^I \max_{1 \leq j \leq J} f_{i,j}(x).$$

Our proof showing that the function  $\theta$  in (3) is of the dc kind appears to be new. In order not to further deviate from the discussion of the main topics of this paper, we provide the proof in an appendix at the end of the paper. Notice that (3) is “equivalent” to the “bivariate” maximization

$$\underset{x \in X; (\lambda^i \in \Lambda^i)_{i=1}^I}{\text{maximize}} \sum_{i=1}^I \sum_{j=1}^J h_{i,j}(\lambda^i) f_{i,j}(x), \quad (4)$$

where the equivalence pertains to the globally optimal solutions of these two problems. Nevertheless, when it comes to stationary solutions, the situation is quite different; see the subsequent discussion in Subsection 3.3. In particular, (4) is a differentiable program if all function  $h_{i,j}$  and  $f_{i,j}$  are differentiable while (3) remains non-differentiable due to the max operator; even with this advantage of differentiability, the concept of “d-stationarity” in the former problem is not as sharp as the same concept in the latter that has the  $\lambda$ -variable “hidden”, i.e., in the  $x$ -alone formulation:

$$\underset{x \in X}{\text{maximize}} \theta(x) \triangleq \sum_{i=1}^I \theta_i(x); \quad \text{with each} \quad \theta_i(x) \triangleq \underset{\lambda^i \in \Lambda^i}{\text{maximum}} \sum_{j=1}^J h_{i,j}(\lambda^i) f_{i,j}(x).$$

The upshot of this discussion is that different formulations of a non-differentiable, non-convex optimization problem could lead to stationary solutions with unequal likelihood for being a locally optimal solution. The search for a superior formulation is not an easy task in general, however.

### 3 Stationarity: Convex Constraints

As a non-convex optimization problems, globally optimal solutions of a dc program are in general not possible to be computed. Thus, one has to settle for computing a “stationary” solution in practice. Even

so, one has to be cautious about the notion of stationarity, especially in the case where the constraints contain dc functions. The situation is simpler when the constraint set is convex; in this section, we consider this case first. Specifically, we deal with the following convex constrained dc minimization program:

$$\underset{x \in X}{\text{minimize}} \zeta(x) \triangleq f(x) - g(x), \quad (5)$$

where  $f$  and  $g$  are convex functions defined on an open convex set  $\Omega$  containing the closed convex set  $X \subseteq \mathbb{R}^n$ . [Note the change from maximization in the previous section to minimization in the problem (5).] Since  $\zeta$  is not differentiable, stationarity concepts of (5) are defined in terms of directional derivatives of the objective function, which we briefly review in the subsection below. Before doing so, we mention the references [4, 11] where a host of properties of dc functions are summarized.

### 3.1 Directional derivatives

The directional derivative of  $\zeta$  at a point  $x \in \Omega$  along a direction  $d \in \mathbb{R}^n$  is given by:

$$\zeta'(x; d) \triangleq \lim_{\tau \downarrow 0} \frac{\zeta(x + \tau d) - \zeta(x)}{\tau}.$$

It is well known that convex functions are directionally differentiable; so with  $\zeta = f - g$  being a dc function,  $\zeta'(x; d)$  is well defined for all  $x \in \Omega$  and  $d \in \mathbb{R}^n$ ; moreover

$$\zeta'(x; d) = f'(x; d) - g'(x; d).$$

Since dc functions are locally Lipschitz continuous (and are thus B(ouligand) differentiable [9, Definition 3.1.2]), the C(larke) directional derivative is also well defined:

$$\zeta^0(x; d) \triangleq \limsup_{\substack{y \rightarrow x \\ \tau \downarrow 0}} \frac{\zeta(y + \tau d) - \zeta(y)}{\tau}.$$

In general,  $\zeta^0(x; d) \geq \zeta'(x; d)$ . These two directional derivatives are equal if the function  $\zeta$  is C-regular [8]. However, dc functions are in general not C-regular. We recall that a function  $\zeta$  is *strictly differentiable* at a point  $x$  if the following limit holds:

$$\lim_{\substack{(y,z) \rightarrow (x,x) \\ y \neq z}} \frac{\zeta(y) - \zeta(z) - \nabla \zeta(x)^T (y - z)}{\|y - z\|} = 0,$$

where  $\nabla \zeta(x)$  denotes the gradient vector of  $\zeta$  at  $x$ . If  $g$  is strictly differentiable at  $x$ , then the dc function  $\zeta$  is C-regular at  $x$ . This class of dc functions deserves a name. Specifically, we say that  $\zeta$  is a *good* dc function on  $\Omega$  (with respect to a minimization problem) if there exists a strictly differentiable convex function  $v$  on  $\Omega$  such that  $\zeta + v$  is convex on  $\Omega$ ; in other words,  $\zeta$  is a good dc function if convex functions  $u$  and  $v$  exist such that  $\zeta = u - v$  and  $v$  is strictly differentiable. The class of good dc functions appears extensively in the machine learning area; see e.g. [44] and the references therein. These dc functions are particularly relevant in the context of computing stationary solutions and play an important role in the convergence of several families of iterative algorithms for solving dc programs, such as: the DCA [21, 13, 44, 32] that has been a fundamental algorithm with many applications, the S(uccessive)C(onvex)A(approximation) method [2, 7, 12, 36, 41, 42] that has attracted significant interest in recent years for solving non-convex non-differentiable optimization problems, and an alternating/successive minimization method [35, 36] for solving the joint minimization formulation of the dc program (to be introduced subsequently). In particular, the class of good dc functions will play an important role in two algorithms that we introduce later; see Propositions 16 and 17. Incidentally, since every quadratic function is a differentiable dc function, it follows that every convex constrained optimization problem with a quadratic objective is a “good dc program” while remaining possibly nonconvex.

### 3.2 Concepts of stationarity

As a non-convex, non-differentiable optimization program, there are many kinds of stationary solutions for a dc program. Ideally, we want to be able to identify a stationary solution of the sharpest kind. Arguably, for the convex constrained dc program (5), a d(irectional)-stationary solution defined in terms of the directional derivatives of the objective function would qualify for this purpose. In what follows, we clarify the relations of several major kinds of stationary solutions of (3) by starting with the definition of d-stationarity.

Specifically, we say that a vector  $x \in X$  is a (constrained) d(irectional)-stationary point of  $\zeta$  on  $X$  if

$$\zeta'(x; x' - x) \geq 0, \quad \forall x' \in X, \quad (6)$$

or equivalently,  $f'(x; x' - x) \geq g'(x; x' - x)$  for all  $x \in X$ . Since  $g'(x; d) = \max_{v \in \partial g(x)} v^T d$ , where  $\partial g(x)$  is the subdifferential of the convex function  $g$  at  $x$ , it follows that  $x$  is a (constrained) d-stationary point of the dc function  $\zeta$  on  $X$  if for all  $v \in \partial g(x)$ ,

$$f'(x; x' - x) \geq v^T (x' - x), \quad \forall x' \in X; \quad (7)$$

or equivalently, if

$$x \in \operatorname{argmin}_{x' \in X} f(x') - v^T x', \quad \forall v \in \partial g(x).$$

Letting  $\hat{f} \triangleq f + \delta_X$ , where  $\delta_X$  is the indicator function of the set  $X$ , i.e.,  $\delta_X(x) = \begin{cases} 0 & \text{if } x \in X \\ \infty & \text{otherwise} \end{cases}$ ,

we deduce that  $x \in X$  is d-stationary point of  $\zeta$  on  $X$  if and only if  $v \in \partial \hat{f}(x)$  for all  $v \in \partial g(x)$ ; i.e., if and only if  $\partial g(x) \subseteq \partial \hat{f}(x) = \partial f(x) + \mathcal{N}(x; X)$ , where  $\mathcal{N}(x; X)$  is the normal cone of the convex set  $X$  at  $x \in X$  [37]. This characterization of a d-stationary point is precisely the notion of a *generalized KKT point* employed in the dc literature [18, 20, 32] that is convex analysis based. We prefer to follow a directional derivative based definition with the constraint set  $X$  exposed in the condition (6) to facilitate the practical solution of the dc program (5). A weaker notion of stationarity, called *criticality* in the dc literature, is defined by the condition:  $\partial g(x) \cap (\partial f(x) + \mathcal{N}(x; X)) \neq \emptyset$ . In terms of directional derivatives, this condition says that  $x \in X$  is a *critical point* of  $\zeta$  of  $X$  if *there exists* (as opposed to *for all*)  $v \in \partial g(x)$  such that (7) holds.

Using the C-directional derivative, we say that a vector  $x \in X$  is *C(larke)-stationary* if  $\zeta^0(x; x' - x) \geq 0$  for all  $x' \in X$ . For a good dc function  $\zeta$ , d-stationarity and C-stationarity are equivalent. We will momentarily provide an example to show that if  $\zeta$  is not good, then the converse implication of C-stationarity implying d-stationarity is not always valid. This example uses the following fact which is by itself of independent interest.

**Proposition 2.** Let  $\zeta$  be a dc function defined on an open convex set  $\Omega \subseteq \mathbb{R}^n$ . The following two statements are equivalent:

- (a) both  $\zeta$  and its negative are good on  $\Omega$ ;
- (b)  $\zeta$  is strictly differentiable on  $\Omega$  and there exists a strictly differentiable function  $v$  on  $\Omega$  such that  $\zeta + v$  is convex.

**Proof.** (a)  $\Rightarrow$  (b). It suffices to show that  $\zeta$  is strictly differentiable on  $\Omega$  if (a) holds. Since both  $\zeta$  and  $-\zeta$  are C-regular, we have

$$\begin{aligned} -\liminf_{\substack{y \rightarrow x \\ \tau \downarrow 0}} \frac{\zeta(y + \tau d) - \zeta(y)}{\tau} &= \limsup_{\substack{y \rightarrow x \\ \tau \downarrow 0}} \frac{-\zeta(y + \tau d) + \zeta(y)}{\tau} = (-\zeta)^0(x; d) \\ &= (-\zeta)'(x; d) = -\lim_{\substack{y \rightarrow x \\ \tau \downarrow 0}} \frac{\zeta(x + \tau d) - \zeta(x)}{\tau} = -\limsup_{\substack{y \rightarrow x \\ \tau \downarrow 0}} \frac{\zeta(y + \tau d) - \zeta(y)}{\tau}. \end{aligned}$$

Hence,

$$\liminf_{\substack{y \rightarrow x \\ \tau \downarrow 0}} \frac{\zeta(y + \tau d) - \zeta(y)}{\tau} = \zeta'(x; d) = \limsup_{\substack{y \rightarrow x \\ \tau \downarrow 0}} \frac{\zeta(y + \tau d) - \zeta(y)}{\tau}.$$

Consequently,

$$\lim_{\substack{y \rightarrow x \\ \tau \downarrow 0}} \frac{\zeta(y + \tau d) - \zeta(y)}{\tau} = \zeta'(x; d), \quad \forall x \in \Omega \text{ and } \forall d \in \mathbb{R}^n.$$

Using this limit, we show that  $\zeta'(x; \bullet)$  is linear on  $\mathbb{R}^n$  for fixed  $x$ . Indeed, we have, for any  $d$  and  $d'$  in  $\mathbb{R}^n$ ,

$$\begin{aligned} \zeta'(x; d + d') - \zeta'(x; d) &= \lim_{\substack{y \rightarrow x \\ \tau \downarrow 0}} \left[ \frac{\zeta(y + \tau d + \tau d') - \zeta(y)}{\tau} - \frac{\zeta(y + \tau d) - \zeta(y)}{\tau} \right] \\ &= \lim_{\substack{y \rightarrow x \\ \tau \downarrow 0}} \frac{\zeta(y + \tau d + \tau d') - \zeta(y + \tau d)}{\tau} \\ &= \lim_{\substack{y' \rightarrow x \\ \tau \downarrow 0}} \frac{\zeta(y' + \tau d') - \zeta(y')}{\tau} = \zeta'(x; d'). \end{aligned}$$

The strict differentiability of  $\zeta$  follows readily from a direct verification of this property.

(b)  $\Rightarrow$  (a). This follows easily from the trivial equality  $\zeta = (\zeta + v) - v$ .  $\square$

The example below shows that for a dc function whose negative is good, a C-stationary point is not necessarily d-stationary.

**Example.** Consider the univariate dc function  $\zeta(x) \triangleq 1 + x^2 - 2|x|$  in the scalar variable  $x$ . Since  $-\zeta$  is a good dc function and  $\zeta$  is not differentiable at  $x = 0$ ,  $\zeta$  cannot be good. Clearly,  $\partial_C \zeta(0) = [-2, 2]$  contains the origin; thus  $x = 0$  is a C-stationary point. Yet,  $\zeta'(0; \pm 1) = -2$ ; thus  $x = 0$  is not d-stationary.  $\square$

### 3.3 Lifted stationarity $\Leftrightarrow$ weak d-stationary

A certain class of nonsmooth dc programs can be “lifted” to become a smooth, albeit still nonconvex, program to which standard stationarity conditions can be applied. Specifically, consider a dc function of the following kind:

$$\zeta(x) \triangleq \phi(x) - \max_{\mu \in \mathcal{M}} \psi(x, \mu), \quad (8)$$

where  $\phi$  is a convex function,  $\psi$  is convex-concave, i.e.,  $\psi(\bullet, \mu)$  is convex and  $\psi(x, \bullet)$  is concave, and  $\mathcal{M}$  is a compact set in  $\mathbb{R}^\ell$ . By not requiring  $\mathcal{M}$  to be convex allows us to include the case where  $\mathcal{M}$  is a discrete set such as  $P \cap \{0, 1\}^\ell$ , where  $P$  is a polyhedron in  $\mathbb{R}^\ell$ , so that the  $\mu$ -maximization problem corresponds to a binary optimization problem. In the event that  $\psi(x, \bullet)$  is linear for fixed  $x$ , the maximization of  $\psi(x, \mu)$  for  $\mu$  in a discrete set is equivalent to the maximization of  $\psi(x, \mu)$  for  $\mu$  in the convex hull of the set. If  $x$  is a d-stationary point of  $\zeta$  on  $X$ , then by the renowned Danskin’s Theorem,

$$\phi'(x; x' - x) - \max_{\mu \in \mathcal{M}(x)} \psi(\bullet, \mu)'(x; x' - x) \geq 0, \quad \forall x' \in X,$$

where  $\mathcal{M}(x) \triangleq \operatorname{argmax}_{\mu \in \mathcal{M}} \psi(x, \mu)$ . Equivalently,

$$\phi'(x; x' - x) \geq \psi(\bullet, \mu)'(x; x' - x), \quad \forall x' \in X \text{ and } \forall \mu \in \mathcal{M}(x). \quad (9)$$

We say that a vector  $x \in X$  is a *weak d-stationary point* of  $\zeta$  given by (8) on  $X$  if there exists  $\mu \in \mathcal{M}(x)$  such that

$$\phi'(x; x' - x) \geq \psi(\bullet, \mu)'(x; x' - x), \quad \forall x' \in X.$$

In contrast, since  $\partial \max_{\mu \in \mathcal{M}} \psi(x, \mu) = \text{convex hull of } \partial_x \psi(x, \mu)$  for  $\mu \in \mathcal{M}(x)$ , where  $\partial_x \psi(x, \mu)$  denotes the subdifferential of the function  $\psi(\bullet, \mu)$ , it follows that  $x$  is a critical point if there exist finitely many  $\mu^i \in \mathcal{M}(x)$  for  $i = 1, \dots, I$ , finitely many nonnegative scalars  $(\lambda_i)_{i=1}^I$  summing to unity, and subgradients  $v^i \in \partial_x \psi(x, \mu^i)$  such that

$$\sum_{i=1}^I \lambda_i v^i \in \partial \phi(x) + \mathcal{N}(x; X),$$

which is equivalent to

$$\phi'(x; x' - x) = \max_{u \in \partial \phi(x)} u^T (x' - x) \geq \left[ \sum_{i=1}^I \lambda_i v^i \right]^T (x' - x) \geq 0, \quad \forall x' \in X. \quad (10)$$

The latter inequality confirms that d-stationarity  $\Rightarrow$  weak d-stationary  $\Rightarrow$  criticality; the reason for these one-sided implications is twofold: (i) the multiplicity of the argmax  $\mathcal{M}(x)$ , and (ii) the multiplicity of the set  $\partial_x \psi(x, \bar{\mu})$  even if  $\mathcal{M}(x)$  is the singleton  $\{\bar{\mu}\}$ . When both  $\mathcal{M}(x)$  and  $\partial_x \psi(x, \bar{\mu})$  are singletons, then d-stationarity  $\Leftrightarrow$  weak d-stationary  $\Leftrightarrow$  criticality. This happens when  $\zeta$  is a good dc function.

Corresponding to the minimization problem (5) which takes the form

$$\text{minimize}_{x \in X} \left[ \phi(x) - \max_{\mu \in \mathcal{M}} \psi(x, \mu) \right], \quad (11)$$

is the lifted reformulation in the pair of variables  $(x, \mu)$ :

$$\text{minimize}_{(x, \mu) \in X \times \mathcal{M}} [\phi(x) - \psi(x, \mu)]. \quad (12)$$

In the case where both  $\phi$  and  $\psi$  are differentiable, the latter minimization has the advantage over the former in that it is a differentiable program in the variables  $(x, \mu)$  jointly, whereas with  $\mu$  hidden in the function  $\zeta$ , the minimization of  $\zeta$  over the  $x$ -variable alone is not a differentiable program unless  $\mathcal{M}(x)$  is a singleton for all  $x$  of interest.

In general, if  $\mathcal{M}$  is also convex, and  $\psi$  is directionally differentiable in both variables jointly (e.g.,  $\psi$  is continuously differentiable in  $(x, \mu)$ ) such that the total directional derivative is the sum of the partial directional derivatives with respect to the two arguments, i.e., suppose that

$$\psi'((x, \mu); (x' - x, \mu' - \mu)) = \psi(\bullet, \mu)'(x; x' - x) + \psi(x, \bullet)'(\mu; \mu' - \mu),$$

then it is not difficult to show that  $(x, \bar{\mu})$  is a stationary point of the function  $\phi(x) - \psi(x, \mu)$  on  $X \times \mathcal{M}$  if and only if  $\bar{\mu} \in \mathcal{M}(x)$  and

$$\phi'(x; x' - x) \geq \psi(\bullet, \bar{\mu})'(x; x' - x), \quad \forall x' \in X. \quad (13)$$

Thus,  $x$  is a weak d-stationary point of  $\zeta$  (given by (8)) on  $X$  if and only if there exists  $\bar{\mu} \in \mathcal{M}(x)$  such that  $(x, \bar{\mu})$  is a stationary point of the bivariate function  $\phi(x) - \psi(x, \mu)$  on  $X \times \mathcal{M}$ . In this sense, we can say that  $x$  is a *lifted stationary point* of  $\zeta$  on  $X$  after we have exposed the  $\mu$ -variable that is part of the bivariate function  $\psi(x, \mu)$ .

The lifted problem (12) in the joint variables  $(x, \mu)$  can be interpreted as a 2-person Nash equilibrium problem. Indeed, consider two optimization problems: one is a minimization problem in the  $x$ -variable parameterized by  $\mu$  and the other is a maximization in the  $\mu$ -variable parameterized by  $x$ :

$$\text{minimize}_{x \in X} \phi(x) - \psi(x, \mu) \quad \text{and} \quad \text{maximize}_{\mu \in \mathcal{M}} \psi(x, \mu). \quad (14)$$

A Nash equilibrium of (14) is a pair  $(x^*, \mu^*)$  such that

$$x^* \in \operatorname{argmin}_{x \in X} \phi(x) - \psi(x, \mu^*) \quad \text{and} \quad \mu^* \in \operatorname{argmax}_{\mu \in \mathcal{M}} \psi(x^*, \mu).$$

Since,  $\phi - \psi(\bullet, \mu)$  is not necessarily a convex function, we say that  $(x^*, \mu^*)$  is a *quasi-Nash equilibrium* (QNE) [30, 31] if  $x^*$  is a stationary point of the differentiable program

$$\operatorname{minimize}_{x \in X} \phi(x) - \psi(x, \mu^*)$$

and  $\mu^* \in \operatorname{argmax}_{\mu \in \mathcal{M}} \psi(x^*, \mu)$ . It is then easy to see that if a pair  $(x^*, \mu^*)$  is a QNE of the pair of programs (14), then  $x^*$  is a lifted stationary point of  $\zeta$  (given by (8)) on  $X$ . Conversely, if  $x^*$  is such a stationary solution, then  $(x^*, \mu^*)$  is a QNE of the pair of programs (14) for some  $\mu^*$ . The upshot of this observation is that a dc program is intimately related to games through its equivalent lifted program formulation.

For the dc minimization problem (11) and its lifted formulation (12) with  $\mathcal{M}$  convex, we have the following string of implications that relates different concepts of stationarity.

local minimizer of (11)	=====>	d-stationary	=====>	lifted stationary
C-stat. $\Leftarrow$ d-stationary	$\mathcal{M}(x)$ singleton <===== =>	weak d-stationary	$\psi(x; \bullet)$ linear <===== => $\psi(\bullet; \mu)$ diff.	critical
QNE	<===== =>	lifted stationary	<===== => $\phi$ diff $\psi(\bullet; \mu)$ diff	C-stationary

Two of the above implications are not accounted for in the above discussion; namely, criticality implies lifted stationarity if  $\psi(x; \bullet)$  is linear and  $\psi(\bullet; \mu)$  is differentiable on  $\Omega$  for all  $\mu \in \mathcal{M}$ , and lifted stationarity implies C-stationarity if  $\phi$  and  $\psi(\bullet, \mu)$  are both differentiable. To prove the former, let  $\{\mu^i, \lambda_i, v^i\}_{i=1}^I$  be as given in the derivation of (10). Since  $\psi(x, \bullet)$  is linear, it follows that  $\sum_{i=1}^I \lambda_i v^i \in \partial_x \psi \left( x, \sum_{i=1}^I \lambda_i u^i \right)$ .

Thus,  $\sum_{i=1}^I \lambda_i v^i = \nabla_x \psi \left( x, \sum_{i=1}^I \lambda_i u^i \right)$ . Since  $\bar{\mu} \triangleq \sum_{i=1}^I \lambda_i u^i \in \mathcal{M}(x)$ , weak d-stationarity follows. To prove the remaining implication, we recall that C-stationarity of a vector  $\hat{x} \in X$  means that  $\zeta^0(\hat{x}; x - \hat{x}) \geq 0$  for all  $x \in X$ . By the definition of the C-generalized gradient, we have, for any vector  $d$ ,

$$\begin{aligned} \zeta^0(\hat{x}; d) &= \limsup_{\substack{y \rightarrow \hat{x} \\ \tau \downarrow 0}} \frac{\phi(y + \tau d) - \phi(y) - (\varphi(y + \tau d) - \varphi(y))}{\tau} \\ &\geq \limsup_{\tau \downarrow 0} \frac{\phi(\hat{x}) - \phi(\hat{x} - \tau d) - (\varphi(\hat{x}) - \varphi(\hat{x} - \tau d))}{\tau}. \end{aligned}$$

Let  $\mu \in \mathcal{M}(\hat{x})$  be such that  $\phi'(\hat{x}; x' - \hat{x}) \geq \psi(\bullet, \mu)'(\hat{x}; x' - \hat{x})$  for all  $x' \in X$ . We then have, provided that  $\phi$  and  $\psi(\bullet, \mu)$  are both differentiable at  $\hat{x}$ ,

$$\begin{aligned} \zeta^0(\hat{x}; x - \hat{x}) &\geq \limsup_{\tau \downarrow 0} \frac{\phi(\hat{x}) - \phi(\hat{x} - \tau(x - \hat{x})) + \psi(\hat{x} - \tau(x - \hat{x}), \mu) - \psi(\hat{x}, \mu)}{\tau} \\ &\geq \nabla \phi(\hat{x})^T (x - \hat{x}) - \nabla_x \psi(\hat{x})^T (x - \hat{x}). \end{aligned}$$

If the value function  $\varphi(x)$  is strictly differentiable (thus  $\zeta = \phi - \varphi$  is a good dc function), then all the stationarity concepts discussed so far are equivalent. When  $\mathcal{M}$  is a finite set,  $\varphi(x)$  is a piecewise smooth function; its strict differentiability has been characterized in terms of the gradients  $\nabla_x \psi(x, \mu)$  at the maximizing  $\mu$ 's; see [33]. This class of dc programs, which can be good or not, will be the focus of our subsequent algorithmic development.

**Counterexamples.** We make two remarks with regard to the above string of implications:

(1) In general, a critical point of (11) is not necessarily weakly d-stationary; a counterexample is provided by the univariate function:  $\zeta(x) \triangleq -|x|$  obtained by letting  $\phi(x) \triangleq 0$ ,  $\psi(x, \pm 1) \triangleq \pm x$ ,  $\mathcal{M} \triangleq \{\pm 1\}$ , and  $X \triangleq [-1, 1]$  for simplicity. Since  $\partial_C \zeta(0) = [-1, 1]$  and  $\mathcal{N}_X(0) = \{0\}$ , it follows that  $0 \in \partial_C \zeta(0) + \mathcal{N}_X(0)$ . Yet  $\frac{\partial \psi(0, \pm 1)}{\partial x} = \pm 1 \neq 0$ .

(2) If  $\phi$  is not differentiable, then a weak d-stationary point is not necessarily C-stationary. Take  $\phi(x) \triangleq x + |x|$  and the same  $\psi$ ,  $\mathcal{M}$ , and  $X$  as above, resulting in  $\varphi(x) = |x|$ ; thus  $\zeta(x) = x$ . Clearly, 0 is not a C-stationary point. Yet, with  $\mu = 1$ , we have  $\phi'(0; d) - \frac{\partial \psi(0, 1)}{\partial x} d = d + |d| - d = |d| \geq 0$  for all  $d \in \mathbb{R}$ . Hence 0 is a weak d-stationary point; yet this point has no minimizing property whatsoever with regard to the problem of minimizing  $\zeta(x)$  on  $X$ .  $\square$

Derived from the above discussion, particularly from the counterexamples, the following conclusions refine our understanding of dc programs and add insights to the existing literature of this class of non-convex optimization problems.

- The class of good dc programs, i.e., convex constrained programs whose objectives are good dc functions, is a favorable class of nonsmooth dc problems for which many advanced concepts of stationarity are equivalent to the basic d-stationarity that is easily described and understood in terms of the elementary directional derivatives.
- Given a dc function (even a differentiable one), a “bad” representation as the difference of two convex functions can yield a weak d-stationary point that is not C-stationary.
- For general nonsmooth minimization problems, the search for a sharp notion of stationarity has always been a challenge. Ideally, one wants to be able to design an algorithm that will compute a stationary point that has the best chance to be a local minimum. For the class of dc minimization problems exemplified by (11), the above examples show that the critical points, C-stationary points, and even weak d-stationary points are not ideal because it is less likely for them to correspond to local minima.

## 4 dc Constrained dc Programs

In this section, we study the B-stationarity concept (to be defined momentarily) associated with a general dc program, i.e., a dc program subject to dc constraints:

$$\begin{aligned} & \underset{x \in X}{\text{minimize}} && \zeta(x) \triangleq \phi(x) - \varphi(x) \\ & \text{subject to} && \phi_{c,j}(x) - \varphi_{c,j}(x) \leq 0, \quad j = 1, \dots, J, \end{aligned} \tag{15}$$

where  $\phi$ ,  $\varphi$ ,  $\phi_{c,j}$ , and  $\varphi_{c,j}$  for all  $j$  are all convex functions defined on the open convex set  $\Omega$  containing the closed convex set  $X$ . This study is not only interesting for its own sake but the results are needed subsequently in the convergence analysis of an iterative scheme for solving the problem. Before proceeding, we mention a variation of the problem (1) that leads to dc constraints; see [1]. Namely, in this version of the problem, we impose some Quality-of-Service (QoS) constraints defined by a prescribed level of

minimum secrecy rate profile  $\mathbf{s}^* \triangleq (s_q^*)_{q=1}^Q$  that need to be satisfied in the power allocation. Specifically, the problem is

$$\begin{aligned}
& \underset{(\mathbf{p}, \widehat{\mathbf{p}}) \geq \mathbf{0}}{\text{maximize}} && \sum_{q=1}^Q \sum_{k=1}^N [R_{qqk}(\mathbf{p}, \widehat{\mathbf{p}}) - R_{q0k}(\mathbf{p}, \widehat{\mathbf{p}})]^+ \\
& \text{subject to:} && \sum_{k=1}^N p_q(k) \leq P_q^{\max} \quad \forall q = 1, \dots, Q \quad (\text{private constraints}) \\
& && \sum_{k=1}^N \widehat{p}_j(k) \leq \widehat{P}_j^{\max} \quad \forall j = 1, \dots, J \quad (\text{coupling constraints}) \\
& \text{and} && \sum_{k=1}^N [R_{qqk}(\mathbf{p}, \widehat{\mathbf{p}}) - R_{q0k}(\mathbf{p}, \widehat{\mathbf{p}})]^+ \geq s_q^*, \quad q = 1, \dots, Q \quad (\text{QoS constraints}).
\end{aligned}$$

Since each term  $[R_{qqk}(\mathbf{p}, \widehat{\mathbf{p}}) - R_{q0k}(\mathbf{p}, \widehat{\mathbf{p}})]^+$  is a dc function of the power variables, the QoS constraints are of the dc type. Another class of problems that leads to a dc constrained dc program is the class of quadratic programs with (linear) complementarity constraints (QPCC) [5, 6, 14]. Specifically, consider

$$\begin{aligned}
& \underset{(x,y) \in Z}{\text{minimize}} && q(x, y) \\
& \text{subject to} && 0 \leq y \perp r + Nx + My \geq 0,
\end{aligned} \tag{16}$$

where  $q(x, y)$  is a (possibly nonconvex) quadratic function,  $Z$  is a polyhedron in  $\mathbb{R}^{n+m}$ ,  $r$  is an  $m$ -dimensional vector,  $N$  is an  $m \times n$  matrix,  $M$  is an  $m \times m$  matrix (not necessarily positive semidefinite), and the  $\perp$  notation denotes the complementarity between the variables  $y$  and  $w \triangleq r + Nx + My$ . Since the (linear) complementarity constraint is clearly equivalent to 3 conditions, 2 linear and 1 quadratic:

$$y \geq 0, \quad r + Nx + My \geq 0, \quad \text{and} \quad y^T(r + Nx + My) \leq 0,$$

the QPCC is a linearly constrained dc program with one additional dc constraint.

In general, multiple dc constraints can be combined into a single nondifferentiable dc constraint using the max-function. Indeed, the  $J$  dc constraints in (15) are equivalent to the single dc constraint:

$$\max_{1 \leq j \leq J} (\phi_{c,j}(x) - \varphi_{c,j}(x)) \leq 0.$$

Note that

$$\max_{1 \leq j \leq J} (\phi_{c,j}(x) - \varphi_{c,j}(x)) = \underbrace{\max_{1 \leq j \leq J} \left( \phi_{c,j}(x) + \sum_{j \neq \ell=1}^J \varphi_{c,\ell}(x) \right)}_{\phi_c(x)} - \underbrace{\sum_{\ell=1}^J \varphi_{c,\ell}(x)}_{\varphi_c(x)},$$

where  $\phi_c(x)$  and  $\varphi_c(x)$  are both convex functions with the latter being (strictly) differentiable if each  $\varphi_{c,\ell}$  is so. Thus  $\zeta_c(x) \triangleq \phi_c(x) - \varphi_c(x)$  is a good dc function if each  $\varphi_{c,\ell}$  is strictly differentiable. Thus, we restrict the discussion below to a singly dc constrained dc program:

$$\begin{aligned}
& \underset{x \in X}{\text{minimize}} && \zeta(x) \triangleq \phi(x) - \varphi(x) \\
& \text{subject to} && \zeta_c(x) \triangleq \phi_c(x) - \varphi_c(x) \leq 0.
\end{aligned} \tag{17}$$

Due to the last nonconvex constraint:  $\phi_c(x) \leq \varphi_c(x)$ , the above problem is considerably more complicated than the convex constrained problem (5). For one thing, constraint qualifications (CQs) are needed to yield a constructive description of the stationarity condition of the problem (17); this is not a trivial task as the dc constraint is both nondifferentiable and nonconvex. We focus on the well-known concept of stationarity based on the Bouligand tangent cone of a constraint set at a feasible point. Applied to (17), this concept, called B-stationarity [28], pertains to a feasible vector  $x^* \in \widehat{X}$  satisfying

$$\zeta'(x^*; d) \geq 0, \quad \forall d \in \mathcal{T}_{\widehat{X}}(x^*), \quad (18)$$

where  $\widehat{X} \triangleq \{x \in X \mid \phi_c(x) \leq \varphi_c(x)\}$  is the (nonconvex) feasible set of (17) and  $\mathcal{T}_{\widehat{X}}(x^*)$  is the Bouligand tangent cone of  $\widehat{X}$  at  $x^* \in \widehat{X}$ , i.e.,  $d \in \mathcal{T}_{\widehat{X}}(x^*)$  if there exist a sequence of vectors  $\{x^k\} \subset \widehat{X}$  converging to  $x^*$  and a sequence of positive scalars  $\{\tau_k\}$  converging to 0 such that  $d = \lim_{k \rightarrow \infty} \frac{x^k - x^*}{\tau_k}$ . For a nonconvex set such as  $\widehat{X}$  that involves the nondifferentiable function  $\zeta_c$ , it is difficult to derive a constructive description of  $\mathcal{T}_{\widehat{X}}(x^*)$ . Thus the B-stationarity condition (18) is hard to verify in general and no existing computational scheme can compute a B-stationary point for the problem (17) according to this definition. Incidentally, B-stationarity reduces to d-stationarity without the dc constraint; we use the former terminology to highlight the nonconvexity and nondifferentiability of the dc constraint.

#### 4.1 A subclass of dc constraints

Our goal here is to introduce a constraint qualification for the special case of (17) where

$$\varphi_c(x) \triangleq \max_{1 \leq k \leq L} \psi_{c,k}(x) \quad (19)$$

is the pointwise maximum of finitely many differentiable convex functions but there is no structural assumption on  $\phi_c(x)$  except its convexity. Under the stipulation, the feasible set  $\widehat{X}$  is the union of finitely many convex sets consisting of the “smooth” pieces of  $\widehat{X}$ . Specifically, we have

$$\widehat{X} = \bigcup_{j=1}^L \widehat{X}^j, \quad \text{with} \quad \widehat{X}^j \triangleq \{x \in X \mid \phi_c(x) \leq \psi_{c,j}(x)\}.$$

The approach below is reminiscent of the study of stationarity for the class of mathematical programs with complementarity constraints [27, 29, 39, 40], and more generally, problems with piecewise smooth constraints. In particular, the stationarity theory in [39] is in principle applicable to the above representation of the feasible  $\widehat{X}$ . Yet, by focusing on each individual set  $\widehat{X}^j$ , we are able to derive a full characterization of the tangent cone of this set at a feasible point  $\bar{x}$  under a pointwise CQ of the Slater type. For the discussion below to be meaningful, we assume that  $\bar{x} \in \widehat{X}$  is such that  $\phi_c(\bar{x}) = \varphi_c(\bar{x})$ . Indeed, if  $\phi_c(\bar{x}) < \varphi_c(\bar{x})$ , then  $\mathcal{T}_{\widehat{X}}(\bar{x}) = \mathcal{T}_X(\bar{x})$  and there is no need to analyze  $\mathcal{T}_{\widehat{X}}(\bar{x})$  further because we assume that  $\mathcal{T}_X(\bar{x})$  is well behaved.

We introduce a convex subset of  $\widehat{X}^j$  by linearizing the function  $\psi_{c,j}(x)$  at the given point  $\bar{x} \in \widehat{X}^j$ , obtaining a convex subset of  $\widehat{X}^j$ :

$$\widehat{Y}^j(\bar{x}) \triangleq \{x \in X \mid \phi_c(x) \leq \psi_{c,j}(\bar{x}) + \nabla \psi_{c,j}(\bar{x})^T (x - \bar{x})\}.$$

Notice that we cannot linearize the function  $\phi_c(x)$  because we do not assume that it is differentiable; moreover, the set  $\widehat{Y}^j(\bar{x})$  depends on the given vector  $\bar{x}$  whereas  $\widehat{X}^j$  does not. Clearly,  $\mathcal{T}_{\widehat{Y}^j(\bar{x})}(\bar{x}) \subseteq \mathcal{T}_{\widehat{X}^j}(\bar{x})$ . The example below shows that this inclusion is proper in general.

**Example 3.** Consider the convex univariate functions  $\phi_c(x) = x^4$  and  $\psi_c(x) = x^2$  so that the set  $\widehat{X} \triangleq \{x \in \mathbb{R} \mid x^4 - x^2 \leq 0\} = [-1, 1]$  is a simple interval. Let  $\bar{x} = 0$ . It follows that  $\widehat{Y}(0) = \{0\} = \mathcal{T}_{\widehat{Y}(0)}(0)$ ; yet  $\mathcal{T}_{\widehat{X}}(0) = \mathbb{R}$ . For this example, note that  $x = 1/2$  is an ‘‘algebraic Slater’’ point of  $\widehat{X}$ , i.e., the inequality  $x^4 \leq x^2$  holds strictly at this point.  $\square$

We next introduce a convex cone that is a candidate for the tangent cones  $\mathcal{T}_{\widehat{Y}^j(\bar{x})}(\bar{x})$  and  $\mathcal{T}_{\widehat{X}^j}(\bar{x})$ :

$$\widehat{C}^j(\bar{x}) \triangleq \{d \in \mathcal{T}_X(\bar{x}) \mid \phi'_c(\bar{x}; d) \leq \nabla\psi_{c,j}(\bar{x})^T d\},$$

for  $j \in \mathcal{M}_c(\bar{x}) \triangleq \{k \mid \varphi_c(\bar{x}) = \psi_{c,k}(\bar{x})\}$ . The result below shows that if the above cone has an element that satisfies the inequality therein strictly, then the two tangent cones  $\mathcal{T}_{\widehat{Y}^j(\bar{x})}(\bar{x})$  and  $\mathcal{T}_{\widehat{X}^j}(\bar{x})$  are both equal to  $\widehat{C}^j(\bar{x})$ . This result is the key for us to show the convergence of the iterative algorithm to be presented later for computing a B-stationary point of the dc constrained dc program (17).

**Proposition 4.** Let  $\bar{x} \in \widehat{X}$  be such that  $\phi_c(\bar{x}) = \varphi_c(\bar{x})$  and let  $j \in \mathcal{M}_c(\bar{x})$ . Suppose an element  $\bar{d} \in \mathcal{T}_X(\bar{x})$  exists such that  $\phi'_c(\bar{x}; \bar{d}) < \nabla\psi_{c,j}(\bar{x})^T \bar{d}$ . Then

$$\mathcal{T}_{\widehat{Y}^j(\bar{x})}(\bar{x}) = \mathcal{T}_{\widehat{X}^j}(\bar{x}) = \widehat{C}^j(\bar{x}). \quad (20)$$

Thus  $\mathcal{T}_{\widehat{X}^j}(\bar{x})$  is a closed convex cone.

**Proof.** We have the following inclusions:

$$\mathcal{T}_{\widehat{Y}^j(\bar{x})}(\bar{x}) \subseteq \mathcal{T}_{\widehat{X}^j}(\bar{x}) \subseteq \widehat{C}^j(\bar{x}),$$

where the second inclusion can easily be proved as follows. Let  $d \in \mathcal{T}_{\widehat{X}^j}(\bar{x})$  with unit norm be given. Clearly  $d \in \mathcal{T}_X(\bar{x})$ . Let  $\{x^k\} \subset \widehat{X}^j(\bar{x}) \setminus \{\bar{x}\}$  be a sequence converging to  $\bar{x}$  such that  $d = \lim_{k \rightarrow \infty} \frac{x^k - \bar{x}}{\|x^k - \bar{x}\|}$ . Since  $\phi_c(x^k) \leq \psi_{c,j}(x^k)$  for all  $k$  and  $\phi_c(\bar{x}) = \psi_{c,j}(\bar{x})$ , it follows readily that  $\phi'_c(\bar{x}; d) \leq \nabla\psi_{c,j}(\bar{x})^T d$ . It remains to show that  $\widehat{C}^j(\bar{x}) \subseteq \mathcal{T}_{\widehat{Y}^j(\bar{x})}(\bar{x})$ . We first show that any  $\bar{d} \in \mathcal{T}_X(\bar{x})$  satisfying  $\phi'_c(\bar{x}; \bar{d}) < \nabla\psi_{c,j}(\bar{x})^T \bar{d}$  must belong to  $\mathcal{T}_{\widehat{Y}^j(\bar{x})}(\bar{x})$ . Indeed, for any such  $\bar{d}$ , let  $\{\bar{x}^k\} \subset X \setminus \{\bar{x}\}$  be a sequence converging to  $\bar{x}$  such that  $\bar{d} = \lim_{k \rightarrow \infty} \frac{\bar{x}^k - \bar{x}}{\|\bar{x}^k - \bar{x}\|}$ . Since  $\phi'_c(\bar{x}; d) = \lim_{k \rightarrow \infty} \frac{\phi_c(\bar{x}^k) - \phi_c(\bar{x})}{\|\bar{x}^k - \bar{x}\|}$ , it follows that for all  $k$  sufficiently large,  $\phi_c(\bar{x}^k) < \psi_{c,j}(\bar{x}) + \nabla\psi_{c,j}(\bar{x})^T(\bar{x}^k - \bar{x})$ . Thus  $\bar{x}^k \in \widehat{Y}^j(\bar{x})$  for all  $k$  sufficiently large. Hence  $\bar{d} \in \mathcal{T}_{\widehat{Y}^j(\bar{x})}(\bar{x})$ . For any  $d \in \widehat{C}^j(\bar{x})$ ,  $d + \tau\bar{d}$  remains in  $\mathcal{T}_X(\bar{x})$  and satisfies:  $\phi'_c(\bar{x}; d + \tau\bar{d}) < \nabla\psi_{c,j}(\bar{x})^T(d + \tau\bar{d})$  for all  $\tau > 0$ , by the subadditivity and positive homogeneity of the directional derivative  $\phi'_c(\bar{x}; \bullet)$ . Therefore,  $d + \tau\bar{d} \in \mathcal{T}_{\widehat{Y}^j(\bar{x})}(\bar{x})$  for all  $\tau > 0$ . Since the tangent cone is a closed set, it follows that  $d \in \mathcal{T}_{\widehat{Y}^j(\bar{x})}(\bar{x})$ , establishing the equalities in (20). The last assertion of the proposition follows readily from the closedness and convexity of  $\widehat{C}^j(\bar{x})$ .  $\square$

The following remarks are worth noting.

- The existence of a vector  $\bar{d} \in \mathcal{T}_X(\bar{x})$  such that  $\phi'_c(\bar{x}; \bar{d}) < \nabla\psi_{c,j}(\bar{x})^T \bar{d}$  is equivalent to the existence of a vector  $\widehat{x} \in X$  such that  $\phi_c(\widehat{x}) < \psi_{c,j}(\bar{x}) + \nabla\psi_{c,j}(\bar{x})^T(\widehat{x} - \bar{x})$ . Such a vector satisfies  $\phi_c(\widehat{x}) < \psi_{c,j}(\widehat{x})$ .
- For an index  $j \in \mathcal{M}_c(\bar{x})$  for which  $\psi_{c,j}$  is an affine function, we have  $\widehat{X}^j = \widehat{Y}^j(\bar{x})$  for any  $\bar{x} \in \widehat{X}^j$ . Hence  $\mathcal{T}_{\widehat{Y}^j(\bar{x})}(\bar{x}) = \mathcal{T}_{\widehat{X}^j}(\bar{x})$  always holds; nevertheless for the second equality in (20) to hold, we still need the existence of the vector  $\bar{d}$  as in Proposition 4.  $\square$

We define a Slater concept for a vector  $\bar{x}$  satisfying the assumptions of Proposition 4.

**Definition 5.** The *pointwise Slater CQ* is said to hold for the set  $\widehat{X}$  at a vector  $\bar{x} \in \widehat{X}$  satisfying  $\phi_c(\bar{x}) = \varphi_c(\bar{x})$  if for every index  $j \in \mathcal{M}_c(\bar{x})$ , there exists  $\bar{d}^j \in \mathcal{T}_X(\bar{x})$  satisfying  $\phi'_c(\bar{x}; \bar{d}^j) < \nabla \psi_{c,j}(\bar{x})^T \bar{d}^j$ .  $\square$

Since  $\phi'_c(\bar{x}; \bullet)$  is convex, it follows that the pointwise Slater CQ holds at a vector  $\bar{x} \in \widehat{X}$  satisfying  $\phi_c(\bar{x}) = \varphi_c(\bar{x})$  if and only if there exists a single vector  $\bar{d} \in \mathcal{T}_X(\bar{x})$  such that  $\phi'_c(\bar{x}; \bar{d}) < \nabla \psi_{c,j}(\bar{x})^T \bar{d}$  for all  $j \in \mathcal{M}_c(\bar{x})$ . At such a point  $\bar{x}$ , we have the following string of implications:

$$\begin{aligned} \text{pointwise Slater at } \bar{x} &\quad \Rightarrow \quad \text{set algebraic Slater} \quad \Rightarrow \quad \text{set topological Slater; i.e.,} \\ \phi_c(\widehat{x}) < \psi_{c,j}(\bar{x}) + \nabla \psi_{c,j}(\bar{x})^T (\widehat{x} - \bar{x}) &\quad \Rightarrow \quad \phi_c(\widehat{x}) < \psi_{c,j}(\widehat{x}) \quad \Rightarrow \quad \widehat{x} \text{ interior pt. of } \Phi_{c,j} \\ \Downarrow & \\ \phi'_c(\bar{x}; \bar{d}) < \nabla \psi_{c,j}(\bar{x})^T \bar{d}, & \end{aligned}$$

where  $\Phi_{c,j} \triangleq \{x \mid \phi_c(x) \leq \psi_{c,j}(x)\}$ . Nevertheless, the reverse of each of the two implications is in general not true, the main reason being the nonconvexity of the set  $\Phi_{c,j}$ . The corollary below is an immediate consequence of Proposition 4.

**Corollary 6.** If  $\bar{x} \in \widehat{X}$  satisfies the pointwise Slater CQ, then

$$\mathcal{T}_{\widehat{X}}(\bar{x}) = \bigcup_{j \in \mathcal{M}_c(\bar{x})} \widehat{C}^j(\bar{x}). \quad (21)$$

Hence,  $\mathcal{T}_{\widehat{X}}(\bar{x})$  is the union of finitely many closed convex cone.  $\square$

The next result identifies another situation in which the equalities in (21) will hold. We recall that a function  $\theta$  is *piecewise affine* on a domain  $\mathcal{D}$  [9, Definition 4.1.3] if it is continuous and there exist finitely many affine functions  $\{\theta_i\}_{i=1}^K$  such that  $\theta(x) \in \{\theta_i(x)\}_{i=1}^K$  for all  $x \in \mathcal{D}$ .

**Proposition 7.** Let  $\bar{x} \in \widehat{X}$  be such that  $\phi_c(\bar{x}) = \varphi_c(\bar{x})$ . If  $X$  is a polyhedron and the (convex) function  $\phi_c$  is piecewise affine on  $X$ , then (21) holds.

**Proof.** It suffices to show the inclusion:  $\widehat{C}^j(\bar{x}) \subseteq \mathcal{T}_{\widehat{Y}^j(\bar{x})}(\bar{x})$  for all  $j \in \mathcal{M}_c(\bar{x})$ . Let  $d \in \widehat{C}^j(\bar{x})$ . Since  $X$  is polyhedral, it follows that  $\bar{x} + \tau d \in X$  for all  $\tau > 0$  sufficiently small. Moreover, for all such  $\tau$ , we have

$$\phi_c(\bar{x} + \tau d) = \phi_c(\bar{x}) + \tau \phi'_c(\bar{x}; d)$$

by the piecewise affine property of  $\phi_c$ ; see Exercise 4.8.10 in [9]. From this equality, we easily deduce that  $d \in \mathcal{T}_{\widehat{Y}^j(\bar{x})}(\bar{x})$ .  $\square$

Based on the last two results, we can derive the following necessary and sufficient conditions for a B-stationary point of the program (17).

**Proposition 8.** Let  $\bar{x} \in \widehat{X}$ . Provided that either  $\bar{x}$  satisfies the pointwise Slater CQ or the assumptions of Proposition 7 hold, the following statements are equivalent:

- (a)  $\bar{x}$  is a B-stationary point of (17);
- (b) for every  $j \in \mathcal{M}_c(\bar{x})$ ,  $\zeta'(\bar{x}; d) \geq 0$  for all  $d$  in  $\widehat{C}^j(\bar{x})$ ;
- (c) for every  $j \in \mathcal{M}_c(\bar{x})$ ,  $\bar{x}$  is a d-stationary point of  $\zeta(x)$  on the convex subset  $\widehat{Y}^j(\bar{x})$  of  $\widehat{X}^j$ ; i.e.,  $\zeta'(\bar{x}; x - \bar{x}) \geq 0$  for all  $x \in \widehat{Y}^j(\bar{x})$ .  $\square$

Thus, under the assumptions of Proposition 8, checking if  $\bar{x}$  is a B-stationary point of (17) can be determined by showing that  $\bar{x}$  is a d-stationary of  $|\mathcal{M}_c(\bar{x})|$  convex-constrained dc programs (part (c)). An algorithm for accomplishing the latter task is presented in Section 6. Nevertheless, the identification of  $\bar{x}$  requires an extension of this algorithm to deal with the dc constraint.

## 5 Computing d-Stationary Points

We are now ready to discuss the next main topic of this paper, namely the computation of a d-stationary point of a convex constrained dc program and its extension to a B-stationary point when there is a dc constraint present in the problem. The discussion is divided into 3 parts: the first part (Subsection 5.1) deals with the convex-constrained dc program (5), the second part extends the discussion to a problem with a dc constraint, and the third and last part discusses a parallel implementation when the problem objective function has a sum structure.

### 5.1 The basic algorithm

The DCA is a well-known algorithm for solving a dc program. In its abstract form [21, 13, 32, 45], the algorithm works with the subgradients of the two convex functions  $\phi(x)$  and  $\varphi(x)$ , taken to be extended valued, whose difference is the objective function of the problem; the constraints are all embedded in  $\phi$  and  $\varphi$ . Subsequential convergence to a critical point is proved, among other properties of the algorithm. Assuming that  $\varphi(x)$  is differentiable, the paper [44] revisited the DCA and extended it, called the convex-concave procedure (CCCP), to a problem with dc constraints defined by good dc functions. Thus the setting in the latter reference pertains to good dc programs. To motivate the discussion below, we give an example of a dc function that is not good and show that the limit point obtained by the DCA is not a d-stationary solution.

**Example 9.** Consider the univariate, unconstrained minimization of the dc function  $\frac{1}{2}x^2 - \max(-x, 0)$  whose unique d-stationary point is  $x = -1$ . Choose a positive  $x^0$  as the initial iterate. Without regularization, the DCA computes  $x^1$  by minimizing  $\frac{1}{2}x^2$ , yielding  $x^1 = 0$ . At this point, if the subgradient of the plus-function is incorrectly picked, the algorithm could stay at the origin forever. A better illustration is to consider a regularization of the DCA wherein at each iteration  $\nu$ , the algorithm minimizes the regularized function  $\frac{1}{2}x^2 - (\partial \max(-x, 0)|_{x=x^\nu})(x - x^\nu) + \frac{1}{2}(x - x^\nu)^2$ . It is not hard to see that starting at the same positive  $x^0$ , the regularized DCA generates a sequence of iterates satisfying the recursive equation  $x^{\nu+1} = \frac{1}{2}x^\nu$  for  $\nu = 0, 1, \dots$ , which converges to the non-d-stationary point  $x^\infty = 0$ . For this example, it is easy to modify the DCA so that the unique d-stationary point can be computed; one such modification is that at each iteration  $\nu$ , we consider 2 subproblems: (i) minimizing  $\frac{1}{2}x^2 + (x - x^\nu) + \frac{1}{2}(x - x^\nu)^2$ , and (ii) minimizing  $\frac{1}{2}x^2 + \frac{1}{2}(x - x^\nu)^2$ , and choose the next iterate to be the minimizer of these two subproblems that gives a lower value of the original objective function. We leave it to the reader to verify that this modified procedure will converge to the d-stationary solution of  $-1$ .  $\square$

Consider the dc program:

$$\underset{x \in X}{\text{minimize}} \zeta(x) \triangleq \phi(x) - \varphi(x), \quad \varphi(x) \triangleq \max_{1 \leq i \leq \ell} \psi_i(x) \quad (22)$$

where  $\phi$  and each  $\psi_i$  are convex functions defined on an open convex set  $\Omega$  containing the feasible set  $X$ , which is a closed convex set in  $\mathbb{R}^n$ . Moreover, we assume that each  $\psi_i$  is continuously differentiable ( $C^1$ ) on  $\Omega$ . Being a pointwise maximum of finitely many  $C^1$  convex functions,  $\varphi$  is a convex piecewise smooth function with directional derivative at a point  $x$  along a direction  $d \in \mathbb{R}^n$  given by

$$\varphi'(x; d) = \max_{i \in \mathcal{M}(x)} \nabla \psi_i(x)^T d,$$

where  $\mathcal{M}(x) \triangleq \underset{1 \leq i \leq \ell}{\text{argmax}} \psi_i(x)$ . For a given scalar  $\varepsilon > 0$ , let  $\mathcal{M}_\varepsilon(x) \triangleq \{i \mid \psi_i(x) \geq \varphi(x) - \varepsilon\}$  which is a superset of  $\mathcal{M}(x)$ . The following result gives necessary and sufficient condition for a d-stationary point of (22) that is useful for its computation.

**Proposition 10.** A vector  $\bar{x} \in X$  is a d-stationary solution of (22) if and only if for every  $i \in \mathcal{M}(\bar{x})$ ,  $\bar{x} \in \operatorname{argmin}_{x \in X} [\phi(x) - \nabla\psi_i(\bar{x})^T(x - \bar{x})]$ , or equivalently,  $\bar{x} = \operatorname{argmin}_{x \in X} [\phi(x) - \nabla\psi_i(\bar{x})^T(x - \bar{x}) + \frac{1}{2}\|x - \bar{x}\|^2]$ .

**Proof.** This follows readily because both functions  $\phi(x) - \nabla\psi_i(\bar{x})^T(x - \bar{x})$  and  $\phi(x) - \nabla\psi_i(\bar{x})^T(x - \bar{x}) + \frac{1}{2}\|x - \bar{x}\|^2$  are convex in  $x$ .  $\square$

**Algorithm I.** Let  $\varepsilon > 0$  be given. For a given  $x^\nu \in X$  at iteration  $\nu$  and for each index  $i \in \mathcal{M}_\varepsilon(x^\nu)$ , let

$$\hat{x}^{\nu,i} = \operatorname{argmin}_{x \in X} \phi(x) - \psi_i(x^\nu) - \nabla\psi_i(x^\nu)^T(x - x^\nu) + \frac{1}{2}\|x - x^\nu\|^2. \quad (23)$$

Let  $\hat{i} \in \operatorname{argmin}_{i \in \mathcal{M}_\varepsilon(x^\nu)} \zeta(\hat{x}^{\nu,i}) + \frac{1}{2}\|\hat{x}^{\nu,i} - x^\nu\|^2$ ; set  $x^{\nu+1} \triangleq \hat{x}^{\nu,\hat{i}}$ . If  $x^{\nu+1} = x^\nu$ , terminate; otherwise replace  $\nu$  by  $\nu + 1$  and repeat the iteration.

Before proving the convergence of the above algorithm, we offer a few comments. First of all, with  $\varepsilon = 0$  and noting that  $\nabla\psi_i(x^\nu) \in \partial\varphi(x^\nu)$ , the algorithm is the “complete primal DCA” described in Section 3 of the unpublished report [19]. Nevertheless, as shown by the example above, convergence to a d-stationary point of the algorithm with  $\varepsilon = 0$  cannot be proved. Thus, the introduction of the scalar  $\varepsilon$  is essential. Second, the proximal regularization is perhaps not needed as one can always strongly convexify the functions  $\phi$  and  $\psi_i$  without changing the difference function  $\zeta$ ; indeed we always have

$$\zeta(x) = [\phi(x) + c(x)] - \max_{1 \leq i \leq \ell} [\psi_i(x) + c(x)],$$

for any strongly convex function  $c(x)$  such as  $\frac{1}{2}\|x\|^2$ . We adopt the term  $\frac{1}{2}\|x - x^\nu\|^2$  as this is a common regularization in many nonlinear programming algorithms. Third, we have left open the practical solution of the subproblems (23) which may require yet an iterative process. With the many advances in convex programming in recent years, this is a safe omission as we are adopting this technology as the workhorse in the above algorithm. In this regard, one could incorporate a variable step size in the quadratic term instead of a unit step size to increase flexibility in the practical implementation of the algorithm. We have omitted all these refinements as we want to present the basic version of the algorithm and establish its (subsequential) convergence to a d-stationary point of the program (22) which we present in the result below.

**Proposition 11.** Suppose that the dc function  $\zeta$  is bounded below on the closed convex set  $X$ . Starting at any  $x^0 \in X$  for which the level set  $L(x^0) \triangleq \{x \in X \mid \zeta(x) \leq \zeta(x^0)\}$  is bounded, Algorithm I generates a well-defined bounded sequence  $\{x^\nu\}$  such that every accumulation point, at least one of which must exist, is a d-stationary solution of (22). Moreover, if the algorithm does not terminate in a finite number of iterations, any such point cannot be a local maximizer of  $\zeta$  on  $X$ .

**Proof.** By the update rule of the algorithm, we have

$$\begin{aligned}
\zeta(x^\nu) &= \phi(x^\nu) - \max_{1 \leq i \leq \ell} \{\psi_i(x^\nu)\} \\
&= \phi(x^\nu) - \psi_i(x^\nu), \quad \forall i \in \mathcal{M}(x^\nu) \\
&\geq \phi(\widehat{x}^{\nu,i}) - \psi_i(x^\nu) - \nabla \psi_i(x^\nu)^T (\widehat{x}^{\nu,i} - x^\nu) + \frac{1}{2} \|\widehat{x}^{\nu,i} - x^\nu\|^2, \quad \forall i \in \mathcal{M}(x^\nu) \\
&\quad \text{by the definition of } \widehat{x}^{\nu,i} \\
&\geq \phi(\widehat{x}^{\nu,i}) - \psi_i(\widehat{x}^{\nu,i}) + \frac{1}{2} \|\widehat{x}^{\nu,i} - x^\nu\|^2, \quad \forall i \in \mathcal{M}(x^\nu) \\
&\quad \text{by the convexity of } \psi_i \\
&\geq \phi(\widehat{x}^{\nu,i}) - \max_{1 \leq j \leq \ell} \psi_j(\widehat{x}^{\nu,i}) + \frac{1}{2} \|\widehat{x}^{\nu,i} - x^\nu\|^2, \quad \forall i \in \mathcal{M}(x^\nu) \\
&= \zeta(\widehat{x}^{\nu,i}) + \frac{1}{2} \|\widehat{x}^{\nu,i} - x^\nu\|^2, \quad \forall i \in \mathcal{M}(x^\nu) \\
&\geq \zeta(x^{\nu+1}) + \frac{1}{2} \|x^{\nu+1} - x^\nu\|^2 \quad \text{by the definition of } x^{\nu+1}.
\end{aligned}$$

Hence, the sequence of objective values  $\{\zeta(x^\nu)\}$  is non-increasing, and strictly decreasing if  $x^{\nu+1} \neq x^\nu$  for all  $\nu$ . Since  $\zeta$  is bounded below on  $X$ , it follows that  $\lim_{\nu \rightarrow \infty} \zeta(x^\nu)$  exists and

$$\lim_{\nu \rightarrow \infty} [\zeta(x^\nu) - \zeta(x^{\nu+1})] = \lim_{\nu \rightarrow \infty} \|x^{\nu+1} - x^\nu\| = 0. \quad (24)$$

Since the sequence  $\{x^\nu\}$  is contained in the bounded set  $L(x^0)$ , it has at least one accumulation point. Let  $\{x^\nu\}_{\nu \in \kappa}$  be a subsequence converging to a limit  $x^\infty$ , which must necessarily belong to  $X$ . By restricting the subsequence on hand, a simple limiting argument shows that  $\mathcal{M}(x^\nu) \subseteq \mathcal{M}(x^\infty) \subseteq \mathcal{M}_\varepsilon(x^\nu)$  for all  $\nu \in \kappa$  sufficiently large. Therefore, using the update rule of the algorithm, for all  $i \in \mathcal{M}(x^\infty)$ , we have

$$\begin{aligned}
\zeta(x^{\nu+1}) + \frac{1}{2} \|x^{\nu+1} - x^\nu\|^2 &\leq \zeta(\widehat{x}^{\nu,i}) + \frac{1}{2} \|\widehat{x}^{\nu,i} - x^\nu\|^2 \\
&\leq \phi(x) - (\psi_i(x^\nu) + \nabla \psi_i(x^\nu)^T (x - x^\nu)) + \frac{1}{2} \|x - x^\nu\|^2 \quad \forall x \in X.
\end{aligned}$$

Taking the limit  $\nu \in \kappa \rightarrow \infty$  yields

$$\zeta(x^\infty) \leq \phi(x) - (\psi_i(x^\infty) + \nabla \psi_i(x^\infty)^T (x - x^\infty)) + \frac{1}{2} \|x - x^\infty\|^2, \quad \forall x \in X, \forall i \in \mathcal{M}(x^\infty),$$

or equivalently

$$\phi(x^\infty) \leq \phi(x) - \nabla \psi_i(x^\infty)^T (x - x^\infty) + \frac{1}{2} \|x - x^\infty\|^2, \quad \forall x \in X, \forall i \in \mathcal{M}(x^\infty).$$

The d-stationarity of  $x^\infty$  for the minimization problem (22) follows from Proposition 10. To prove the last statement of the proposition, we note that the sequence  $\{\zeta(x^\nu)\}$  must be strictly decreasing (since the algorithm does not terminate in finite number of iterations); moreover, if  $\widehat{x}$  is any accumulation point of the sequence  $\{x^\nu\}$ , then the sequence  $\{\zeta(x^\nu)\}$  converges to  $\zeta(\widehat{x})$ . Let  $\{x^\nu\}_{\nu \in \kappa'}$  be a subsequence converging to  $\widehat{x}$ . We must have  $\zeta(x^{\nu-1}) > \zeta(x^\nu) \geq \zeta(\widehat{x})$  for all  $\nu \in \kappa'$ . Since  $\{x^{\nu-1}\}_{\nu \in \kappa'}$  also converges to  $\widehat{x}$ , it follows that  $\widehat{x}$  cannot be a local maximizer of  $\zeta$  on  $X$ .  $\square$

We make several additional remarks regarding the algorithm and its convergence proof.

- The choice of  $\varepsilon > 0$  is important as Example 9 shows the failure of the algorithm with  $\varepsilon = 0$ .
- A major departure of the algorithm from the DCA is that instead of choosing a subgradient from  $\partial\varphi(x^\nu)$  at iteration  $\nu$ , we choose the family of gradients  $\{\nabla\psi_k(x^\nu)\}_{k \in \mathcal{M}_\varepsilon(x^\nu)}$ , which is a finite subset of

$\partial\varphi(x^\nu)$ , at the expense of solving multiple convex subprograms. This extra effort per iteration leads to the (subsequential) convergence to a d-stationary point of (22).

- A referee asked the question of whether the non-increasing property of the sequence of objective values  $\{\zeta(x^\nu)\}$  can be derived from the well-known properties of the proximal map. This does not appear to be the case; however, the proof given above is fairly elementary.
- If the function  $\phi = \phi_{\text{nd}} + \phi_{\text{d}}$  is the sum of two convex functions with  $\phi_{\text{nd}}$  being nondifferentiable and  $\phi_{\text{d}}$  being differentiable, then we keep  $\phi_{\text{nd}}$  as it is but approximate  $\phi_{\text{d}}$  by its first-order Taylor expression at  $x^\nu$ . Specifically, we may define  $\hat{x}^{\nu,i}$  to be the minimizer of

$$\underset{x \in X}{\text{minimize}} \phi_{\text{nd}}(x) + \phi_{\text{d}}(x^\nu) + \nabla\phi_{\text{d}}(x^\nu)^T(x - x^\nu) + \frac{1}{2}\|x - x^\nu\|^2 - \psi_i(x^\nu) - \nabla\psi_i(x^\nu)^T(x - x^\nu),$$

and the same convergence result can be proved.

- At this time, we are not able to extend the algorithm to treat the case where  $\varphi(x)$  is the value function of a continuum family of convex functions, i.e., when  $\varphi(x) = \max_{y \in Y} \psi(x, y)$  where  $\psi(x, \bullet)$  is a concave function and  $Y$  is a compact convex set in  $\mathbb{R}^m$  for some positive integer  $m$ . It remains an open challenge to develop a practically implementable and provably convergent algorithm to compute a d-stationary solution of (22) in this case.

Proposition 11 yields the subsequential convergence of Algorithm I. There are various additional conditions under which sequential convergence can be established. One such condition is the existence of an *isolated* accumulated point of the sequence; such a point has the property that it is the unique accumulation point of the sequence within a certain neighborhood of the point. We formally state this result in the corollary below; its proof follows immediately from [9, Proposition 8.3.10] and is omitted.

**Corollary 12.** Under the assumptions of Proposition 11, if one of the accumulation points of the sequence  $\{x^\nu\}$  is isolated, then the sequence converges to it.  $\square$

A referee pointed out that a recent paper [3] has established the convergence of the whole sequence (as opposed to subsequences) produced by various classes of algorithms to a “critical point” for a broad class of nonconvex semi-algebraic problems. It would be interesting to investigate whether such a sequential convergence result could be established for Algorithm I applied to the dc program (22) with semi-algebraic functions.

## 5.2 A randomized version

When the set  $\mathcal{M}_\varepsilon(x^\nu)$  contains a large number of elements, then many subproblems (23) have to be solved. Although each of them is convex and presumably easy to solve, it would be desirable not to solve too many of them in practical implementation. Randomization could help in this regard; i.e., we randomize the choice of an appropriate subproblem to be solved at each iteration. We present this randomized algorithm below and show that it will produce a d-stationary point of the problem (22) almost surely.

**The Randomized Version.** Let a scalar  $p_{\min} \in (0, 1)$  be given and let  $\varepsilon > 0$  be arbitrary. For a given  $x^\nu \in X$  at iteration  $\nu$ , choose an index  $i \in \mathcal{M}_\varepsilon(x^\nu)$  randomly so that

$$p_i^\nu \triangleq \Pr(\text{index } i \text{ is chosen} \mid x^1, \dots, x^\nu) \geq p_{\min} > 0.$$

Let  $x^{\nu+1} = \underset{x \in X}{\text{argmin}} \phi(x) + \frac{1}{2}\|x - x^\nu\|^2 - (\psi_i(x^\nu) + \nabla\psi_i(x^\nu)^T(x - x^\nu))$ .

In what follows, we establish the almost sure convergence of the above randomized algorithm. For each index  $j \in \mathcal{M}_\varepsilon(x^\nu)$ , let  $\widehat{x}^{\nu,j} \triangleq \operatorname{argmin}_{x \in X} \widehat{\zeta}_j(x; x^\nu) \triangleq \phi(x) + \frac{1}{2} \|x - x^\nu\|^2 - (\psi_j(x^\nu) + \nabla\psi_j(x^\nu)^T(x - x^\nu))$ . We have,

$$\zeta(x^\nu) = \widehat{\zeta}_j(x^\nu; x^\nu) \geq \widehat{\zeta}_j(\widehat{x}^{\nu,j}; x^\nu) \geq \zeta(\widehat{x}^{\nu,j}) + \frac{1}{2} \|\widehat{x}^{\nu,j} - x^\nu\|^2.$$

Moreover,  $x^{\nu+1} = \widehat{x}^{\nu,j}$  with probability  $p_j^\nu$ . Taking conditional expectations, the above inequality implies

$$\mathbb{E}[\zeta(x^{\nu+1}) \mid x^\nu] = \sum_{i \in \mathcal{M}_\varepsilon(x^\nu)} p_i^\nu \zeta(\widehat{x}^{\nu,i}) \leq \zeta(x^\nu) - \frac{1}{2} \sum_{i \in \mathcal{M}_\varepsilon(x^\nu)} p_i^\nu \|\widehat{x}^{\nu,i} - x^\nu\|^2.$$

Consequently, the random sequence  $\{\zeta(x^\nu)\}$  is a supermartingale and assuming that  $\zeta$  is bounded from below on  $X$ , we may conclude that  $\{\zeta(x^\nu)\}$  converges almost surely and, letting  $p_i^\nu = 0$  for all  $i \notin \mathcal{M}_\varepsilon(x^\nu)$ ,

$$\lim_{\nu \rightarrow \infty} p_i^\nu \|\widehat{x}^{\nu,i} - x^\nu\| = 0, \quad \forall i = 1, \dots, \ell \quad (25)$$

with probability one. In the rest of the proof, we restrict ourselves to the set of probability one in which the above limit holds. Consider a point  $x^\infty$  that is the limit of the subsequence  $\{x^\nu\}_{\nu \in \kappa}$ . By further restricting the subsequence on hand, we can assume that  $i_\nu = \bar{i}$  for all  $\nu \in \kappa$  with  $\bar{i} \in \mathcal{M}(x^\infty)$ . Let  $i \in \mathcal{M}(x^\infty)$  be given. It then follows that  $i \in \mathcal{M}_\varepsilon(x^\nu)$  for all  $\nu \in \kappa$  sufficiently large. Since  $p_i^\nu \geq p_{\min}$ , it follows from (25) that  $\lim_{\nu(\in \kappa) \rightarrow \infty} \widehat{x}^{\nu,i} = \lim_{\nu(\in \kappa) \rightarrow \infty} x^\nu = x^\infty$ . Therefore, by the definition of  $\widehat{x}^{\nu,i}$ , we have, for every  $x \in X$ ,

$$\phi(x) + \frac{1}{2} \|x - x^\nu\|^2 - \psi_i(x^\nu) - \nabla\psi_i(x^\nu)^T(x - x^\nu) \geq \phi(\widehat{x}^{\nu,i}) - \psi_i(x^\nu) - \nabla\psi_i(x^\nu)^T(\widehat{x}^{\nu,i} - x^\nu).$$

Letting  $\nu(\in \kappa) \rightarrow \infty$  in the above inequality, we deduce

$$\phi(x) + \frac{1}{2} \|x - x^\infty\|^2 - \nabla\psi_i(x^\infty)^T(x - x^\infty) \geq \phi(x^\infty),$$

from which we can deduce that  $\phi(x) - \nabla\psi_i(x^\infty)^T(x - x^\infty) \geq \phi(x^\infty)$  for all  $x \in X$ . By Proposition 10, it follows that  $x^\infty$  is a d-stationary solution of (22) almost surely. This completes the proof of the following convergence result.

**Proposition 13.** Suppose that the dc function  $\zeta$  is bounded below on the closed convex set  $X$ . Every limit point of the iterates generated by the randomized algorithm is a d-stationary point of the dc program (22) with probability one.  $\square$

## 6 Algorithmic Extension: I

In this and the next section, we present two extensions of the deterministic Algorithm I and omit their randomized versions. When providing the convergence of the extended algorithms, we focus on their subsequential convergence and rely on Corollary 12 and the recent reference [3] for the issue of sequential convergence. The first extension of Algorithm I is to the dc constrained dc program (17). We start by presenting an immediate consequence of Propositions 8 and 10.

**Proposition 14.** Let  $\varphi(x) \triangleq \max_{1 \leq j \leq \ell} \psi_j(x)$  and  $\varphi_c(x) \triangleq \max_{1 \leq j \leq L} \psi_{c,j}(x)$ , where  $\psi_i$  and  $\psi_{c,j}$  are convex differentiable functions on  $\Omega$ . Let  $\bar{x} \in \widehat{X}$  satisfy the pointwise Slater CQ. It holds that  $\bar{x}$  is a B-stationary solution of (17) if and only if for every  $i \in \mathcal{M}(\bar{x})$  and every  $j \in \mathcal{M}_c(\bar{x})$ ,  $\bar{x} \in \operatorname{argmin}_{x \in \widehat{Y}^j(\bar{x})} [\phi(x) - \nabla\psi_i(\bar{x})^T(x - \bar{x})]$ ,

or equivalently,  $\bar{x} = \operatorname{argmin}_{x \in \widehat{Y}^j(\bar{x})} [\phi(x) - \nabla\psi_i(\bar{x})^T(x - \bar{x}) + \frac{1}{2} \|x - \bar{x}\|^2]$ .  $\square$

In the rest of this section, we assume that the two functions  $\varphi$  and  $\varphi_c$  are as given in Proposition 14. Till now, the issue of feasibility of the problem (17) has not been addressed. Indeed, this is a very difficult issue and we will not directly deal with it. In what follows, we propose two approaches to compute a B-stationary point of (17). The first approach assumes that a feasible solution of the problem is available which we will use to initiate the algorithm. The second approach does not assume that such a (feasible) solution is readily available, perhaps because the problem is actually not feasible. We propose a double-loop scheme in which the outer loop solves a sequence of convex-constrained subproblems by penalizing the dc constraint and the inner loop applies the basic Algorithm I (or its randomized version) to compute a d-stationary point of the penalized subproblems. Convergence of both algorithms will be analyzed.

## 6.1 Feasibility assumed

In this subsection, we assume that a vector  $x^0 \in \widehat{X}$  is available. Similar to the index set  $\mathcal{M}_\varepsilon(\bar{x})$  pertaining to the max-function  $\varphi(x)$  in the objective, we define, for each  $\varepsilon > 0$  and each  $\bar{x} \in X$  the set

$$\mathcal{M}_{c,\varepsilon}(\bar{x}) \triangleq \{k \mid \psi_{c,k}(\bar{x}) \geq \varphi_c(\bar{x}) - \varepsilon\}$$

pertaining to the max-function  $\varphi_c(x)$  in the constraint. We also recall the set

$$\widehat{Y}^j(\bar{x}) \triangleq \{x \in X \mid \phi_c(x) \leq \psi_{c,j}(\bar{x}) + \nabla \psi_{c,j}(\bar{x})^T(x - \bar{x})\},$$

which we have previously defined for a vector  $\bar{x} \in \widehat{X}^j$  is now extended to an arbitrary vector  $\bar{x} \in X$ . Note: if  $\bar{x} \notin \widehat{X}^j$ , the nonemptiness of  $\widehat{Y}^j(\bar{x})$  is not guaranteed. Nevertheless  $\widehat{Y}^j(\bar{x})$  must be nonempty if  $\bar{x} \in \widehat{X}$  and  $j \in \mathcal{M}_c(\bar{x})$ .

**Algorithm II.** Let  $\varepsilon > 0$  and  $x^0 \in \widehat{X}$  be given. At iteration  $\nu$ , given  $x^\nu \in \widehat{X}$ , we let, for every pair of indices  $i \in \mathcal{M}_\varepsilon(x^\nu)$  and  $j \in \mathcal{M}_{c,\varepsilon}(x^\nu)$ ,  $\widehat{x}^{\nu,i,j}$  be the (unique) optimal solution of the strongly convex program:

$$\operatorname{argmin}_{x \in \widehat{Y}^j(x^\nu)} \phi(x) - \psi_i(x^\nu) - \nabla \psi_i(x^\nu)^T(x - x^\nu) + \frac{1}{2} \|x - x^\nu\|^2 \quad (26)$$

if  $\widehat{Y}^j(x^\nu) \neq \emptyset$ ; otherwise we let  $\widehat{x}^{\nu,i,j} = x^\nu$ . Let  $(\widehat{i}, \widehat{j}) \in \operatorname{argmin}_{(i,j) \in \mathcal{M}_\varepsilon(x^\nu) \times \mathcal{M}_{c,\varepsilon}(x^\nu)} \zeta(\widehat{x}^{\nu,i,j}) + \frac{1}{2} \|\widehat{x}^{\nu,i,j} - x^\nu\|^2$ ;  
set  $x^{\nu+1} \triangleq \widehat{x}^{\nu,\widehat{i},\widehat{j}}$ .

We have the following (subsequential) convergence result of the above algorithm.

**Proposition 15.** Suppose that the dc function  $\zeta$  is bounded below on the feasible set  $\widehat{X}$ . Starting at any  $x^0 \in \widehat{X}$  for which the level set  $\widehat{L}(x^0) \triangleq \{x \in \widehat{X} \mid \zeta(x) \leq \zeta(x^0)\}$  is bounded, Algorithm II generates a well-defined bounded sequence  $\{x^\nu\} \subset \widehat{X}$  such that every accumulation point  $x^\infty$ , at least one of which must exist, is feasible to (17); moreover, if  $x^\infty$  satisfies the pointwise Slater CQ, then  $x^\infty$  is a B-stationary point of (17).

**Proof.** Since  $x^\nu \in \widehat{X}$ , it follows that the subproblem (26) is feasible for all  $j \in \mathcal{M}_c(x^\nu)$  and thus has a unique optimal solution. Moreover,  $x^{\nu+1} \in \widehat{X}$  by the gradient inequality applied to the function  $\psi_{c,\widehat{j}}$ .

We now follow the proof of Proposition 11 to deduce the following string of (in)equalities:

$$\begin{aligned}
\zeta(x^\nu) &= \phi(x^\nu) - \max_{1 \leq i \leq \ell} \{\psi_i(x^\nu)\} \\
&= \phi(x^\nu) - \psi_i(x^\nu), \quad \forall i \in \mathcal{M}(x^\nu) \\
&\geq \phi(\widehat{x}^{\nu,i,j}) - \psi_i(x^\nu) - \nabla \psi_i(x^\nu)^T (\widehat{x}^{\nu,i,j} - x^\nu) + \frac{1}{2} \|\widehat{x}^{\nu,i,j} - x^\nu\|^2, \\
&\quad \forall (i,j) \in \mathcal{M}(x^\nu) \times \mathcal{M}_c(x^\nu); \text{ by the definition of } \widehat{x}^{\nu,i,j} \\
&\geq \phi(\widehat{x}^{\nu,i,j}) - \psi_i(\widehat{x}^{\nu,i,j}) + \frac{1}{2} \|\widehat{x}^{\nu,i,j} - x^\nu\|^2, \quad \forall (i,j) \in \mathcal{M}(x^\nu) \times \mathcal{M}_c(x^\nu) \\
&\quad \text{by the convexity of } \psi_i \\
&\geq \phi(\widehat{x}^{\nu,i}) - \max_{1 \leq k \leq \ell} \psi_k(\widehat{x}^{\nu,i,j}) + \frac{1}{2} \|\widehat{x}^{\nu,i,j} - x^\nu\|^2, \quad \forall (i,j) \in \mathcal{M}(x^\nu) \times \mathcal{M}_c(x^\nu) \\
&= \zeta(\widehat{x}^{\nu,i,j}) + \frac{1}{2} \|\widehat{x}^{\nu,i,j} - x^\nu\|^2, \quad \forall (i,j) \in \mathcal{M}(x^\nu) \times \mathcal{M}_c(x^\nu) \\
&\geq \zeta(x^{\nu+1}) + \frac{1}{2} \|x^{\nu+1} - x^\nu\|^2 \quad \text{by the definition of } x^{\nu+1}.
\end{aligned}$$

As before, it follows that (24) holds. Let  $\{x^\nu\}_{\nu \in \kappa}$  be a subsequence converging to a limit  $x^\infty$ , which can easily be seen to belong to  $\widehat{X}$ . Suppose that  $x^\infty$  satisfies the pointwise Slater CQ. According to Proposition 14, it suffices to show that  $\bar{x} = \operatorname{argmin}_{x \in \widehat{Y}^j(\bar{x})} [\phi(x) - \nabla \psi_i(\bar{x})^T (x - \bar{x}) + \frac{1}{2} \|x - \bar{x}\|^2]$  for all pairs

of indices  $(i,j) \in \mathcal{M}(x^\infty) \times \mathcal{M}_c(x^\infty)$ . Let  $j$  be such an index and  $x \in \widehat{Y}^j(x^\infty)$  be arbitrary. Then  $j \in \mathcal{M}_{c,\varepsilon}(x^\nu)$  for all  $\nu \in \kappa$  sufficiently large. Let  $\bar{x}^j$  satisfy:  $\phi_c(\bar{x}^j) < \psi_{c,j}(x^\infty) + \nabla \psi_{c,j}(x^\infty)^T (\bar{x}^j - x^\infty)$ . For all scalars  $\tau \in [0, 1)$ , with  $x^\tau \triangleq \bar{x}^j + \tau(x - \bar{x}^j) \in X$ , we have

$$\phi_c(x^\tau) < \psi_{c,j}(x^\infty) + \nabla \psi_{c,j}(x^\infty)^T (x^\tau - x^\infty).$$

For each fixed  $\tau \in [0, 1)$ , it follows that for all  $\nu \in \kappa$  sufficiently large, we have

$$\phi_c(x^\tau) < \psi_{c,j}(x^\nu) + \nabla \psi_{c,j}(x^\nu)^T (x^\tau - x^\nu)$$

for all  $\nu \in \kappa$  sufficiently large. Thus,  $x^\tau$  is feasible to (26). Similar to the proof of Proposition 11, we have for all  $i \in \mathcal{M}(x^\infty)$ , which is a subset of  $\mathcal{M}_\varepsilon(x^\nu)$  for all  $\nu$  sufficiently large

$$\begin{aligned}
\zeta(x^{\nu+1}) + \frac{1}{2} \|x^{\nu+1} - x^\nu\|^2 &\leq \zeta(\widehat{x}^{\nu,i,j}) + \frac{1}{2} \|\widehat{x}^{\nu,i,j} - x^\nu\|^2 \\
&\leq \phi(x^\tau) - (\psi_i(x^\nu) + \nabla \psi_i(x^\nu)^T (x^\tau - x^\nu)) + \frac{1}{2} \|x^\tau - x^\nu\|^2.
\end{aligned}$$

Passing to the limit  $\nu \in \kappa \rightarrow \infty$ , we deduce, for all  $\tau \in [0, 1)$ ,

$$\zeta(x^\infty) \leq \phi(x^\tau) - \psi_i(x^\infty) - \nabla \psi_i(x^\infty)^T (x^\tau - x^\infty) + \frac{1}{2} \|x^\tau - x^\infty\|^2.$$

Passing to the limit  $\tau \uparrow 1$ , we deduce

$$\phi(x^\infty) \leq \phi(x) - \nabla \psi_i(x^\infty)^T (x - x^\infty) + \frac{1}{2} \|x - x^\infty\|^2$$

for all  $x \in \widehat{Y}^j(x^\infty)$  and all  $(i,j) \in \mathcal{M}(x^\infty) \times \mathcal{M}_c(x^\infty)$  as desired.  $\square$

## 6.2 Feasibility not assumed

Without assuming the feasibility of the problem (17), we propose a penalization of the dc constraint and establish a limiting result when the penalization tends to infinity. Penalization techniques in dc

programming and DCA have been investigated for solving dc constrained dc programs [32, 22, 23, 18]. The departure of our discussion from these references is that we aim to compute a B-stationary solution of such a program. This is accomplished by considering the following penalized convex-constrained dc program: for  $\rho > 0$ ,

$$\underset{x \in X}{\text{minimize}} \quad \zeta(x) + \rho \max(0, \zeta_c(x))$$

and letting  $x^\rho$  be a d-stationary point of this problem. Suppose that for a sequence of penalty parameters  $\{\rho_\nu\} \uparrow \infty$ , the corresponding sequence of d-stationary solutions  $\{x^{\rho_\nu}\}$  converges to a limit  $x^\infty$ . What can we say about  $x^\infty$  with regard to the stationarity properties of (17)? Incidentally, since

$$\phi(x) - \varphi(x) + \rho \max(0, \phi_c(x) - \varphi_c(x)) = \left[ \underbrace{\phi(x) + \rho \max(\phi_c(x), \varphi_c(x))}_{\text{convex}} \right] - \left( \underbrace{\varphi(x) + \rho \varphi_c(x)}_{\text{convex}} \right),$$

the computation of each  $x^\rho$  can be accomplished by Algorithm I or its randomized version. It should be noted that during the computation of the sequence  $\{x^\nu\}$ , where  $x^\nu \triangleq x^{\rho_\nu}$ , if at any time, an iterate satisfies the dc constraint and is thus a feasible solution of the problem (17), we have the option of abandoning this penalization approach and return to the previous direct approach wherein feasibility is maintained throughout the algorithm. Here, we do not concern ourselves with these algorithmic details and focus on an understanding of the asymptotic property of the penalization approach employing an unbounded sequence of penalty parameters. For practical implementation, one should introduce a penalty update rule that circumvents the unboundedness of such a sequence. Details like this should best be left for future studies.

**Proposition 16.** In the above setting, the following three statements hold:

- (a) If  $\zeta_c(x^\nu) > 0$  for infinitely many  $\nu$ 's, and if  $\varphi_c$  is strictly differentiable at  $x^\infty$ , then provided that  $X$  is bounded and  $\zeta$  is globally Lipschitz continuous on  $X$ ,  $x^\infty$  is a d-stationary solution of

$$\underset{x \in X}{\text{minimize}} \quad \zeta_c(x) \triangleq \phi_c(x) - \varphi_c(x);$$

- (b) If  $\zeta_c(x^\nu) < 0$  for infinitely many  $\nu$ 's, then  $x^\infty$  is a d-stationary point of  $\zeta$  on  $X$ , thus a B-stationary point on  $\widehat{X}$ ;
- (c) Suppose  $\zeta_c(x^\nu) = 0$  for all but finitely many  $\nu$ 's. If  $x^\infty$  satisfies the pointwise Slater CQ and  $\mathcal{M}_c(x^\infty)$  is a singleton, then  $x^\infty$  is a B-stationary point of (17).

**Proof.** (a) If  $\zeta_c(x^\nu) > 0$ , then the stationarity condition of  $x^\nu$  is

$$\zeta'(x^\nu; x - x^\nu) + \rho_\nu \zeta'_c(x^\nu; x - x^\nu) \geq 0, \quad \forall x \in X. \quad (27)$$

In the following, we restrict  $\nu$  so that  $\zeta_c(x^\nu) > 0$ . On one hand, we have

$$\zeta'_c(x^\nu; x - x^\nu) = \phi'_c(x^\nu; x - x^\nu) - (g^\nu)^T(x - x^\nu),$$

where  $g^\nu \in \partial\varphi_c(x^\nu)$  that depends on the vector  $x$ . With  $x$  fixed, the sequence of subgradients  $\{g^\nu\}$  has an accumulation point  $g^\infty$  which belongs to  $\partial\varphi_c(x^\infty)$ . Without loss of generality, we may assume that  $g^\infty$  is the limit of the sequence  $\{g^\nu\}$ . Consequently, we deduce that

$$\limsup_{\nu \rightarrow \infty} \zeta'_c(x^\nu; x - x^\nu) \leq \phi'_c(x^\infty; x - x^\infty) - (g^\infty)^T(x - x^\infty).$$

On the other hand, by the boundedness of  $X$  and the global Lipschitz continuity of  $\zeta$ , it follows that  $\zeta'(x^\nu; x - x^\nu)$  is bounded. Hence diving by  $\rho_\nu$  in (27), we deduce

$$\phi'_c(x^\infty; x - x^\infty) - (g^\infty)^T(x - x^\infty) \geq 0,$$

where the limit  $g^\infty$  depends on  $x$ . Thus, we have proved that

$$\phi'_c(x^\infty; x - x^\infty) \geq \min_{g \in \partial \varphi_c(x^\infty)} g^T(x - x^\infty), \quad \forall x \in X.$$

Hence, if  $\varphi_c$  is strictly differentiable at  $x^\infty$ , the above inequality yields the d-stationarity of the dc constraint function  $\zeta_c$  on  $X$ .

(b) If  $\zeta_c(x^\nu) < 0$ , then the stationarity condition of  $x^\nu$  is

$$\zeta'(x^\nu; x - x^\nu) \geq 0, \quad \forall x \in X.$$

Passing to the limit  $\nu \rightarrow \infty$  for these  $\nu$ 's easily yields  $\zeta'(x^\infty; x - x^\infty) \geq 0$  for all  $x \in X$ , as desired.

(c) Suppose  $\zeta_c(x^\nu) = 0$  for all but finitely many  $\nu$ 's. The stationarity condition of  $x^\nu$  is

$$\zeta'(x^\nu; x - x^\nu) + \rho_\nu \max(0, \zeta'_c(x^\nu; x - x^\nu)) \geq 0, \quad \forall x \in X, \quad (28)$$

from which we want to show:  $\zeta'(x^\infty; x - x^\infty) \geq 0$  for all  $x \in \widehat{Y}^j(x^\infty)$ , where

$$\widehat{Y}^j(x^\infty) \triangleq \{x \in X \mid \phi_c(x) \leq \psi_{c,j}(x^\infty) + \nabla \psi_{c,j}(x^\infty)^T(x - x^\infty)\},$$

with  $j$  being the single element of  $\mathcal{M}_c(x^\infty)$ . Thus  $\mathcal{M}_c(x^\nu) = \{j\}$  also, for all  $\nu$  sufficiently large. Hence,

$$\zeta'_c(x^\nu; x - x^\nu) = \phi'_c(x^\nu; x - x^\nu) - \nabla \psi_c(x^\nu)^T(x - x^\nu).$$

Consider a vector  $x \in \widehat{Y}^j(x^\infty)$  that satisfies the inequality therein strictly. Since  $\psi_{c,j}(x^\infty) = \varphi_c(x^\infty) = \phi_c(x^\infty)$ , we deduce

$$\begin{aligned} \limsup_{\nu \rightarrow \infty} \phi'_c(x^\nu; x - x^\nu) &\leq \phi'_c(x^\infty; x - x^\infty) \\ &\leq \phi_c(x) - \phi_c(x^\infty) < \nabla \psi_{c,j}(x^\infty)^T(x - x^\infty) = \lim_{\nu \rightarrow \infty} \nabla \psi_{c,j}(x^\nu)^T(x - x^\nu). \end{aligned}$$

It follows that for all  $\nu$  sufficiently large,

$$\phi'_c(x^\nu; x - x^\nu) < \nabla \psi_{c,j}(x^\nu)^T(x - x^\nu)$$

Hence

$$0 \leq \zeta'(x^\nu; x - x^\nu) + \rho_\nu \max(0, \zeta'_c(x^\nu; x - x^\nu)) = \zeta'(x^\nu; x - x^\nu).$$

If  $x \in \widehat{Y}^j(x^\infty)$  is such that  $\phi_c(x) = \psi_{c,j}(x^\infty) + \nabla \psi_{c,j}(x^\infty)^T(x - x^\infty)$ , then the vector  $x^\tau \triangleq \bar{x}^j + \tau(x - \bar{x}^j)$ , where  $\bar{x}^j$  is the Slater point under the CQ, i.e.,  $\bar{x}^j \in X$  satisfies the dc constraint strictly, remains a Slater point of  $\widehat{Y}^j$  for all  $\tau \in [0, 1)$ . By the above proof, we have  $\zeta'(x^\nu; x^\tau - x^\nu) \geq 0$ . Letting  $\tau \uparrow 1$  completes the proof.  $\square$

**Remarks.** The assumption that  $\mathcal{M}_c(x^\infty)$  is a singleton, or equivalently that  $\varphi_c$  is strictly differentiable at  $x^\infty$ , is a pointwise goodness of the dc function  $\zeta_c(x) = \phi_c(x) - \varphi_c(x)$  at  $x^\infty$ . In spite of the differentiability of the function  $\varphi_c$  at  $x^\infty$ , the difference function  $\zeta_c$  remains not necessarily so. This is another instance where the class of good dc functions offers an advantage over the class of not-good dc functions.

Another noteworthy remark is that while the assumption  $\zeta_c(x^\nu) = 0$  renders  $x^\nu$  feasible to the problem (17), this vector is obtained as a limit point of a presumably infinite process when the penalized subproblem is solved, and is thus generally not readily available in practical computation. From this perspective, Proposition 16 should be considered a conceptual result in that it offers insights into the asymptotic property of the penalization approach for solving a dc constrained dc program without assuming feasibility. How this result can be turned into a constructive approach for use in practice in solving such a problem requires further investigation.  $\square$

## 7 Algorithmic Extension: II

In this section, we discuss how we can develop a distributed algorithm for solving the following extended dc program:

$$\underset{x \in X}{\text{minimize}} \zeta(x) \triangleq \phi(x) - \sum_{i=1}^I \varphi_i(x), \quad \text{with each } \varphi_i(x) \triangleq \max_{1 \leq k \leq \ell_i} \psi_{i,k}(x) \quad (29)$$

where  $\phi$  and each  $\psi_{i,k}$  are convex functions defined on  $\Omega$  with each  $\psi_{i,k}$  being  $C^1$  on  $\Omega$ . The goal is to exploit the sum structure in the objective function so that each summand can be treated separately from the others. One motivation of this consideration arises from a multi-agent optimization context wherein each agent  $i$  has a private performance function  $\varphi_i(x)$  and it is desirable to be able to implement an algorithm requiring minimal communication among the agents. Two challenges of this goal is the dc and nondifferentiable features of the overall objective function and the coupling of variables in each summand.

Before presenting the distributed algorithm, we mention that while it is possible to apply the basic Algorithm I to the problem (29), the sum structure makes a straightforward application of this centralized algorithm rather laborious when there are many summands. In this case, it may be necessary to solve many subproblems at each iteration that are derived by selecting the functions  $\psi_{i,j}$  for all  $j \in \{k \mid \psi_{i,k}(x^\nu) \geq \varphi_i(x^\nu) - \varepsilon\}$  and all  $i = 1, \dots, I$ . To give an example, consider the case  $I = n$  and each  $\ell_i = 2$ . In this case, the number of subproblems to be solved in each iteration could be exponential in  $n$ . Randomization could help in this regard by not exhausting such a selection per iteration; yet the resulting algorithm remains a centralized scheme that does not take advantage of the sum structure for possible parallel processing. To see how the probabilistic approach can be applied, we define the tuple

$t \triangleq (k_i)_{i=1}^I$  and let  $\mathcal{K} \triangleq \prod_{i=1}^I \{1, \dots, \ell_i\}$ . For each such tuple  $t$ , let  $\psi_t(x) \triangleq \sum_{i=1}^I \psi_{i,k_i}(x)$ . It is easy to see that

$$\zeta(x) = \phi(x) - \varphi(x), \quad \text{where } \varphi(x) \triangleq \max_{t \in \mathcal{K}} \psi_t(x).$$

The total number of elements in  $\mathcal{K}$  is  $\prod_{i=1}^I \ell_i$ , which could be very large. [For instance, if each  $\ell_i = 2$  and  $I = n$ , then the product of these  $\ell_i$ 's is equal to  $2^n$ , which is exponential in the dimension of the variable  $x$ .] In this case, the randomized version of the algorithm becomes useful. For  $\varepsilon > 0$  and  $i = 1, \dots, I$ , let

$$\mathcal{M}_i(x) \triangleq \{k \mid \psi_{i,k}(x) = \varphi_i(x)\} \quad \text{and} \quad \mathcal{M}_{i,\varepsilon}(x) \triangleq \{k \mid \psi_{i,k}(x) \geq \varphi_i(x) - \varepsilon\}.$$

At each iteration  $\nu$ , given an iterate  $x^\nu$ , we select a random tuple  $t \triangleq (k_i)_{i=1}^I \in \mathcal{K}$  such that  $k_i \in \mathcal{M}_{i,\varepsilon}(x^\nu)$  for every  $i = 1, \dots, I$ . We then solve the following strongly convex subproblem,

$$\underset{x \in X}{\text{argmin}} \left[ \phi(x) + \frac{1}{2} \|x - x^\nu\|^2 - \sum_{i=1}^I (\psi_{k_i}(x^\nu) + \nabla \psi_{k_i}(x^\nu)^T (x - x^\nu)) \right]. \quad (30)$$

Although the randomized selection of the tuple  $t$  avoids the enumeration of a possibly large number of elements of the set  $\mathcal{K}$  and significantly reduces the number of subproblems to be solved at each iteration, the global resolution of (30) remains a centralized task. While it may be possible to simplify this task under some structural assumptions on the set  $X$  and differentiability properties of the functions  $\phi(x)$  and  $\varphi(x)$  (see [1, 2]), we present below a distributed penalty approach that is by itself a novel idea for computing d-stationary points of dc programs of the kind (29) and requires no such additional structures. Variations of this approach can be applied to other separable forms of a dc program (e.g., when  $\phi(x)$  is also a sum of agents' functions or a sum of a differentiable and a non-differentiable function). In what follows, we restrict our discussion to (29) where a sum structure is present only in the concave term

of the objective. This distributed approach recognizes the sum structure  $\sum_{i=1}^I \max_{1 \leq k \leq \ell_i} \psi_{i,k}(x)$  and solves subproblems that naturally decomposes according to the latter structure; each decomposed subproblem can be solved in parallel per individual summand.

## 7.1 A penalty approach

The penalty approach for computing a d-stationary solution to the problem (29) consists of two main iterative steps implemented by a sequence of outer iterations each in turn composed of a sequence of inner iterations. Each outer iteration is based on the simple observation that the problem (29) is clearly equivalent to the following one where the single variable  $x$  is duplicated  $I$  times with the addition of the constraints:  $z^i = x$ . This results in a reformulated problem with  $I + 1$  variables:

$$\begin{aligned} & \underset{x, z^i \in X}{\text{minimize}} && \phi(x) - \sum_{i=1}^I \varphi_i(z^i) \\ & \text{subject to} && z^i = x, \quad i = 1, \dots, I. \end{aligned} \quad (31)$$

We next penalize the duplication constraints by replacing them with a sum-of-squares term in the objective using a penalty scalar  $\rho > 0$ :

$$\underset{x, z^i \in X}{\text{minimize}} \theta_\rho(x, z) \triangleq \phi(x) - \sum_{i=1}^I \varphi_i(z^i) + \frac{\rho}{2} \sum_{i=1}^I \|z^i - x\|^2; \text{ where } z \triangleq (z^i)_{i=1}^I. \quad (32)$$

The outer iterations consist of solving the problem (32) for an increasing sequence of positive scalars  $\{\rho_\nu\}$  tending to  $\infty$ . This is accomplished by applying the basic Algorithm I or its randomized version to the following problem:

$$\underset{x, z^i \in X}{\text{minimize}} \theta_{\rho_\nu}(x, z) \triangleq \phi(x) - \sum_{i=1}^I \varphi_i(z^i) + \frac{\rho_\nu}{2} \sum_{i=1}^I \|z^i - x\|^2 \quad (33)$$

for each  $\nu$ , yielding a sequence of d-stationary points  $\{x^\nu, (z^{\nu,i})_{i=1}^I\}_{\nu=1}^\infty$ . Thus, for all  $x$  and  $z^i$  in  $X$ ,

$$\begin{aligned} & \phi'(x^\nu; x - x^\nu) - \sum_{i=1}^I \max_{k \in \mathcal{M}_i(z^{\nu,i})} \nabla \psi_{i,k}(z^{\nu,i})^T (z^i - z^{\nu,i}) + \\ & \rho_\nu \sum_{i=1}^I [(z^{\nu,i} - x^\nu)^T (z^i - z^{\nu,i}) + (x^\nu - z^{\nu,i})^T (x - x^\nu)] \geq 0. \end{aligned} \quad (34)$$

Before describing the inner iterations to generate such stationary solutions of (33), we first establish the desired limiting property of such solutions; namely, every accumulation point  $(x^\infty, (z^{\infty,i})_{i=1}^I)$  of  $\{x^\nu, (z^{\nu,i})_{i=1}^I\}_{\nu=1}^\infty$  must satisfy  $z^{\infty,i} = x^\infty$  for all  $i = 1, \dots, I$ ; thus we recover the feasibility condition of (31).

**Convergence of penalization.** Throughout the following analysis, we assume that each  $\|\nabla \psi_{i,k}\|$  is bounded on  $X$ . By letting  $\eta \triangleq \max_{1 \leq i \leq I} \max_{1 \leq k \leq \ell_i} \max_{x \in X} \|\nabla \psi_{i,k}(x)\|$  and  $z^i = x = x^\nu$  for all  $i$ , we deduce from (34),

$$0 \leq \eta \sum_{i=1}^I \|x^\nu - z^{\nu,i}\| - \frac{\rho_\nu}{2} \sum_{i=1}^I \|z^{\nu,i} - x^\nu\|^2,$$

which easily implies that  $\lim_{\nu \rightarrow \infty} \|z^{\nu,i} - x^\nu\| = 0$  for all  $i$ . Hence, if  $x^\infty$  is the limit of a convergent subsequence  $\{x^\nu\}_{\nu \in \mathcal{N}}$ , which must exist by the boundedness of  $X$ , then  $\lim_{\nu(\in \mathcal{N}) \rightarrow \infty} z^{\nu,i} = x^\infty$  for all  $i$ . With  $z^i = x$  for every  $i$ , (34) also implies

$$\phi'(x^\nu; x - x^\nu) \geq \sum_{i=1}^I \max_{k \in \mathcal{M}_i(z^{\nu,i})} \nabla \psi_{i,k}(z^{\nu,i})^T (x - z^{\nu,i}).$$

Since  $\mathcal{M}_i(z^{\nu,i}) \subseteq \mathcal{M}_i(x^\infty)$  for all  $\nu \in \mathcal{N}$  sufficiently large, we deduce that for some nonnegative scalars  $\{\lambda_{i,k}^\nu\}_{k \in \mathcal{M}_i(x^\infty)}$ , satisfying  $\sum_{k \in \mathcal{M}_i(x^\infty)} \lambda_{i,k}^\nu = 1$  and possibly dependent on  $x$ ,

$$\phi'(x^\nu; x - x^\nu) \geq \sum_{i=1}^I \sum_{k \in \mathcal{M}_i(x^\infty)} \lambda_{i,k}^\nu \nabla \psi_{i,k}(z^{\nu,i})^T (x - z^{\nu,i}).$$

For  $x$  fixed, we may assume, without loss of generality, that for each pair  $(i, k)$ , the sequence of scalars  $\{\lambda_{i,k}^\nu\}_{\nu \in \mathcal{N}}$  converges to  $\lambda_{i,k}^\infty$ , which must be nonnegative and satisfies:  $\sum_{k \in \mathcal{M}_i(x^\infty)} \lambda_{i,k}^\infty = 1$ . By a known limiting property of the directional derivatives of convex functions [37, Theorem 24.5], we have

$$\phi'(x^\infty; x - x^\infty) \geq \limsup_{\nu(\in \mathcal{N}) \rightarrow \infty} \phi'(x^\nu; x - x^\nu).$$

Hence,

$$\phi'(x^\infty; x - x^\infty) \geq \sum_{i=1}^I \sum_{k \in \mathcal{M}_i(x^\infty)} \lambda_{i,k}^\infty \nabla \psi_{i,k}(x^\infty)^T (x - x^\infty).$$

Since  $\sum_{k \in \mathcal{M}_i(x^\infty)} \lambda_{i,k}^\infty \nabla \psi_{i,k}(x^\infty) \in \partial \varphi_i(x^\infty)$ , we deduce that

$$\phi'(x^\infty; x - x^\infty) \geq \sum_{i=1}^I \min_{g^i \in \partial \varphi_i(x^\infty)} (g^i)^T (x - x^\infty), \quad \forall x \in X.$$

Hence, if each  $\partial \varphi_i(x^\infty)$  is a singleton, it follows that  $x^\infty$  is a d-stationarity solution of (29). We have therefore proved the next result.

**Proposition 17.** Suppose that each  $\|\nabla \psi_{i,k}\|$  is bounded on  $X$ .

- (a) (Recovering feasibility) Every accumulation point  $(x^\infty, (z^{\infty,i})_{i=1}^I)$  of the sequence  $\{x^\nu, (z^{\nu,i})_{i=1}^I\}_{\nu=1}^\infty$  of penalized d-stationary points corresponding to a sequence of penalty parameters  $\{\rho_\nu\} \uparrow \infty$  must satisfy  $z^{\infty,i} = x^\infty$  for all  $i = 1, \dots, I$ .
- (b) (Achieving stationarity) Moreover, if each  $\varphi_i$  is strictly differentiable at  $x^\infty$ , then  $x^\infty$  is a d-stationary solution of (29).  $\square$

**Remark.** Once again, the goodness of the objective function of (29) is needed to complete the last step of the proof of the above proposition.  $\square$

**A distributed algorithm for (33).** Based on Algorithm I, we present in this section a distributed algorithm for computing a d-stationary solution of each penalized problem (33). To do this, we need to

take care of one detail of this problem having to do with the non-separability of the term  $\|z^i - x\|^2$  in the objective function. Namely, we linearize this term at a base tuple  $(x^\nu, (z^{\nu,i})_{i=1}^I)$  as follows:

$$\|z^i - x\|^2 \approx \|z^{\nu,i} - x^\nu\|^2 + 2 \left[ (x - x^\nu)^T (x^\nu - z^{\nu,i}) + (z^i - z^{\nu,i})^T (z^{\nu,i} - x^\nu) \right] \quad (35)$$

and use this linearization in each step of the algorithm.

At the beginning of an outer iteration  $\nu$  (thus  $\rho_\nu$  is fixed), starting at a tuple  $(x^{\nu,0}, (z^{\nu,i,0})_{i=1}^I)$  of vectors in  $X$ , the algorithm generates a sequence of inner iterates

$$\{x^{\nu,\mu}, (z^{\nu,i,\mu})_{i=1}^I\}_{\mu=0}^\infty. \quad (36)$$

At each inner iteration  $\mu = 1, 2, \dots$ , for every tuple  $t_{\nu,\mu} \triangleq (k_{\nu,i,\mu})_{i=1}^I$  consisting of indices  $k_{\nu,i,\mu} \in \mathcal{M}_{i,\varepsilon}(z^{\nu,i,\mu})$  for  $i = 1, \dots, I$ , we solve the strongly convex subprogram:

$$\begin{aligned} \text{minimize}_{x, z^i \in X} & \left\{ \phi(x) + \frac{1}{2} \left( \underbrace{\|x - x^{\nu,\mu}\|^2 + \sum_{i=1}^I \|z^i - z^{\nu,i,\mu}\|^2}_{\text{regularization}} \right) + \right. \\ & \left. \rho_\nu \sum_{i=1}^I \left[ \underbrace{(x - x^{\nu,\mu})^T (x^{\nu,\mu} - z^{\nu,i,\mu}) + (z^i - z^{\nu,i,\mu})^T (z^{\nu,i,\mu} - x^{\nu,\mu})}_{\text{linear approximation of } \frac{1}{2}\|z^i - x\|^2} \right] - \right. \\ & \left. \sum_{i=1}^I (\psi_{i,k_{\nu,i,\mu}}(z^{\nu,i,\mu}) + \nabla \psi_{i,k_{\nu,i,\mu}}(z^{\nu,i,\mu})^T (z^i - z^{\nu,i,\mu})) \right\}, \end{aligned} \quad (37)$$

which naturally decomposes into  $I + 1$  subproblems:

- a strongly convex subproblem in the  $x$ -variable,

$$\text{minimize}_{x \in X} \left[ \phi(x) + \frac{1}{2} \|x - x^{\nu,\mu}\|^2 + \rho_\nu (x - x^{\nu,\mu})^T \sum_{i=1}^I (x^{\nu,\mu} - z^{\nu,i,\mu}) \right];$$

- a problem of the same kind in the  $z^i$ -variable, for  $i = 1, \dots, I$ ,

$$\begin{aligned} \text{minimize}_{z^i \in X} & \left\{ \frac{1}{2} \|z^i - z^{\nu,i,\mu}\|^2 + \rho_\nu (z^i - z^{\nu,i,\mu})^T (z^{\nu,i,\mu} - x^{\nu,\mu}) - \right. \\ & \left. [\psi_{i,k_{\nu,i,\mu}}(z^{\nu,i,\mu}) + \nabla \psi_{i,k_{\nu,i,\mu}}(z^{\nu,i,\mu})^T (z^i - z^{\nu,i,\mu})] \right\}. \end{aligned} \quad (38)$$

Among the optimal solutions to (37) for various choices of the index tuples  $t_{\nu,\mu}$ , one of them leads to the new iterates  $(x^{\nu,\mu+1}, (z^{\nu,i,\mu+1})_{i=1}^I)$ . [This is the deterministic version of the algorithm; we leave the probabilistic version for the reader to complete and remind the reader that the latter could be more efficient than the former in practice especially when synchronization among agents is costly.] In total, for each tuple  $t$ , a total of  $I + 1$  strongly convex subprograms are solved at each inner iteration, each of them can be solved separately from the others. The choice of the individual indices  $k_{\nu,i,\mu}$  can be carried out in parallel per agent  $i$ . Thus, the overall implementation of the algorithm is totally distributed. At the completion of the inner iterations according to a prescribed termination criterion, the penalty parameter  $\rho_\nu$  is updated and a new sequence of inner iterations is entered. The convergence of the inner iterations can be proved in a similar way as the basic Algorithm I and is not repeated.

The outer-inner scheme for solving (29) has its advantage of being implementable distributedly according to the individual summands. Ideally, it would be desirable to have a single-loop algorithm wherein the update of the penalty parameter  $\rho$  can be incorporated into the inner iterations. At this time, we are not able to develop a provably convergent single-loop algorithm that can be implemented distributedly.

**Acknowledgement.** The authors are grateful to two referees for offering many constructive comments that have improved the presentation of the paper. Moreover, the first author has benefitted from a fruitful visit to Lorraine University where he had a very productive discussion with Professors Pham Dinh Tao and Le Thi Hoa An on dc programming in general and this paper in particular.

## References

- [1] A. ALVARADO. *Centralized and distributed resource allocation with applications to signal processing in communications*. Ph.D. thesis. Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign (August 2014).
- [2] A. ALVARADO, G. SCUTARI, AND J.S. PANG. A new decomposition method for multiuser DC-programming and its applications. *IEEE Transactions on Signal Processing* 62 (2014) 2984–2998.
- [3] H. ATTOUCH, J. BOLTE, AND B. SVAITER. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Mathematical Programming* 137 (2013) 91–129.
- [4] M. BAČÁK AND J.M. BORWEIN. On difference convexity of locally Lipschitz functions. *Optimization: A Journal of Mathematical Programming and Operations Research* 60 (2011) 961–978.
- [5] L. BAI, J.E. MITCHELL, AND J.S. PANG. On convex quadratic programs with linear complementarity constraints with *Computational Optimization and Applications* 54 (2013) 517–554.
- [6] L. BAI, J.E. MITCHELL, AND J.S. PANG. On QPCCs, QCQPs, and completely positive programs. Manuscript under revision. (August 2014).
- [7] J. BOLTE, S. SHOHAM, AND M. TEBoulLE. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming* (2013) 1–36.
- [8] F.H. CLARKE. *Optimization and Nonsmooth Analysis*. Classic Series in Applied Mathematics 5. SIAM Publications (Philadelphia 1987). [Originally published by John Wiley, New York 1983].
- [9] F. FACCHINEI AND J.S. PANG. *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Volumes I and II, Springer-Verlag, New York (2003).
- [10] F. FACCHINEI, S. SAGRATELLA, AND G. SCUTARI. Flexible parallel algorithms for big data optimization. *arXiv preprint arXiv:1311.2444* (2013).
- [11] J.B. HIRIART-URRUTY. Generalized differentiability, duality and optimization for problems dealing with differences of convex functions. In J. Ponstein, editor. *Convexity and Duality in Optimization*. Proceedings of the Symposium on Convexity and Duality in Optimization Held at the University of Groningen, The Netherlands June 22, 1984. (1985) 37–70.
- [12] M. HONG, T.H. CHANG, X. WANG, M. RAZAVIYAYN, S. MA, AND Z.-Q. LUO. A block successive upper bound minimization method of multipliers for linearly constrained convex optimization. *arXiv preprint arXiv:1401.7079* (2013).

- [13] R. HORST AND N.V. THOAI. D.C. programming: Overview. *Journal of Optimization Theory and Applications* 103 (1999) 1–43.
- [14] H. JIANG AND D. RALPH. QPECgen, a MATLAB generator for mathematical programs with quadratic objectives and affine variational inequality constraints. *Computational Optimization and Applications* 13 (1999) 25–59.
- [15] E. JORSWIECK, A. WOLF, AND S. GERBRACHT. Secrecy on the physical layer in wireless networks. *INTECH* Chapter 20 (2010) pp. 413–435.
- [16] J. KRUSKAL. Two convex counterexamples: A discontinuous envelope function and a nondifferentiable nearest-point mapping. *Proceedings of the American Mathematical Society* 23 (1969) 697–703.
- [17] H.A. LE THI AND D.T. PHAM. DC programming in communication systems: challenging problems and methods. *Vietnam Journal of Computer Science* 1 (2014) 15–28.
- [18] H.A. LE THI AND D.T. PHAM. Recent advances in DC programming and DCA. *Transactions on Computational Collective Intelligence* 8342 (2014) 1–37.
- [19] H.A. LE THI AND D.T. PHAM. The state of the art in DC programming and DCA. Research Report (60 pages), Lorraine University (2013).
- [20] H.A. LE THI, V.N. HUYNH, AND D.T. PHAM. DC programming and DCA for general DC programs. *Advances in Intelligent Systems and Computing* ISBN 978-3-319-06568-7, pp. 15-35, Springer 2014.
- [21] H.A. LE THI AND D.T. PHAM. The DC programming and DCA revised with DC models of real world nonconvex optimization problems. *Annals of operations research* 133 (2005) 25–46.
- [22] H.A. LE THI, D.T. PHAM, AND V.N. HUYNH. Exact penalty and error bounds in DC Programming. *Journal of Global Optimization* 52 (2012) 509–535.
- [23] H.A. LE THI, D.T. PHAM, AND D.M. LE. Exact penalty in d.c. programming. *Vietnam Journal of Mathematics* 27 (1999) 169–178.
- [24] A. MOUDAFI. On the difference of two maximal monotone operators: Regularization and algorithmic approaches. *Applied Mathematics and Computation* 202 (2008) 446-452.
- [25] J.E. MAINGÉ AND A. MOUDAFI. On the convergence of an approximate proximal method for DC functions. *Journal of Computational Mathematics* 24 (2006) 475-480.
- [26] J.E. MAINGÉ AND A. MOUDAFI. Convergence of new inertial proximal methods for DC programming. *SIAM Journal on Optimization* 19 (2008) 397–413.
- [27] Z.Q. LUO, J.S. PANG, AND D. RALPH. *Mathematical Programs With Equilibrium Constraints*. Cambridge University Press (Cambridge, England 1996).
- [28] J.S. PANG. Partially B-regular optimization and equilibrium problems. *Mathematics of Operations Research* 32 (2007) 687–699.
- [29] J.S. PANG AND M. FUKUSHIMA. Complementarity constraint qualifications and simplified B-stationarity conditions for mathematical programs with equilibrium constraints. *Computational Optimization and Applications* 13 (1999) 111–136.

- [30] J.S. PANG AND M. RAZAVIYAYN. A unified distributed algorithm for non-cooperative games with non-convex and non-differentiable objectives. In S.G. Cui, A. Hero, Z.Q. Luo, and J.M.F. Moura, editors. *Big Data over Networks*. Cambridge University Press; in press.
- [31] J.S. PANG AND G. SCUTARI. Nonconvex games with side constraints. *SIAM Journal on Optimization* 21 (2011) 1491–1522.
- [32] D.T. PHAM AND H.A. LE THI. Convex analysis approach to DC programming: Theory, algorithm and applications. *Acta Mathematica Vietnamica* 22 (1997) 289–355.
- [33] L. QI AND P. TSENG. On piecewise smooth functions and almost smooth functions. *Nonlinear Analysis* 67 (2007) 773–794.
- [34] M. RAZAVIYAYN. *Successive Convex Approximation: Analysis and Applications*. Ph.D. thesis. Department of Electrical and Computer Engineering, University of Minnesota (May 2014).
- [35] M. RAZAVIYAYN, M. HONG, AND Z.-Q. LUO. A unified convergence analysis of block successive minimization methods for non-smooth optimization. *SIAM Journal on Optimization* 23 (2013) 1126–1153.
- [36] M. RAZAVIYAYN, M. HONG, Z.-Q. LUO, AND J.-S. PANG. Parallel successive convex approximation for nonsmooth nonconvex optimization. *arXiv preprint arXiv:1406.3665* (2014).
- [37] R.T. ROCKAFELLAR. *Convex Analysis*. Princeton University Press (New Jersey 1970).
- [38] M. SANJABI, M. RAZAVIYAYN, AND Z.Q. LUO. Optimal joint base station assignment and beamforming for heterogeneous networks. *IEEE Transactions on Signal Processing* 62 (2014) 1950–1961.
- [39] S. SCHOLTES. Nonconvex structures in nonlinear programming *Operations Research* 52 (2004) 368–383.
- [40] H. SCHEEL AND S. SCHOLTES. Mathematical programs with equilibrium constraints: Stationarity, optimality, and sensitivity. *Mathematics of Operations Research* 25 (2000) 1–25.
- [41] G. SCUTARI, F. FACCHINEI, L. LAMPARIELLO, AND P. SONG. Parallel and distributed methods for nonconvex optimization. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 14), May 4-9, 2014, Florence, Italy.
- [42] G. SCUTARI, F. FACCHINEI, P. SONG, D.P. PALOMAR, AND J.S. PANG. Decomposition by partial linearization: Parallel optimization of multiuser systems. *IEEE Transaction on Signal Processing* 63 (2014) 641–656.
- [43] A. SHAPIRO. Directionally nondifferentiable metric projection. *Journal of Optimization Theory and Applications* 81 (1994) 203–204.
- [44] B.K. SRIPERUMBUDUR AND G.R.G. LANCKRIET. A proof of convergence of the concave-convex procedure using Zangwill’s theory. *Neural Computation* 24 (2012) 1391–1407.
- [45] H. TUY. Global minimization of a difference of two convex functions. *Mathematical Programming Study* 30 (1987) 150–187.

**Appendix: Proof of the dc-property of  $\theta$  in (3).** We first introduce some notational definitions. For every  $i = 1, \dots, I$  and  $j = 1, \dots, J$  we let

$$\rho_{i,j}^{\max} \triangleq \underset{\lambda^i \in \Lambda^i}{\text{maximum}} h_{i,j}(\lambda^i) \quad \text{and} \quad \rho_{i,j}^{\min} \triangleq \underset{\lambda^i \in \Lambda^i}{\text{minimum}} h_{i,j}(\lambda^i), \quad (39)$$

which are finite scalars by the compactness of  $\Lambda^i$  and the continuity of  $h_{i,j}$ . Based on these two extremum values, we rewrite each product  $h_{i,j}(\lambda^i)f_{i,j}(x)$  as follows:

(i) if  $f_{i,j}(x)$  is convex (cvx):

$$h_{i,j}(\lambda^i) f_{i,j}(x) = \rho_{i,j}^{\min} f_{i,j}(x) + (h_{i,j}(\lambda^i) - \rho_{i,j}^{\min}) f_{i,j}(x),$$

(ii) if  $f_{i,j}(x)$  is concave (cve):

$$h_{i,j}(\lambda^i) f_{i,j}(x) = \rho_{i,j}^{\max} f_{i,j}(x) + (\rho_{i,j}^{\max} - h_{i,j}(\lambda^i)) (-f_{i,j}(x)).$$

Thus, we can rewrite each  $\theta_i(x)$  for  $i = 1, \dots, I$  as

$$\begin{aligned} \theta_i(x) &= \text{maximum}_{\lambda^i \in \Lambda^i} \sum_{j=1}^J h_{i,j}(\lambda^i) f_{i,j}(x) \\ &= \sum_{j: f_{i,j} \text{ cvx}} \rho_{i,j}^{\min} f_{i,j}(x) + \sum_{j: f_{i,j} \text{ cve}} \rho_{i,j}^{\max} f_{i,j}(x) \\ &\quad + \text{maximum}_{\lambda^i \in \Lambda^i} \left( \sum_{j: f_{i,j} \text{ cvx}} (h_{i,j}(\lambda^i) - \rho_{i,j}^{\min}) f_{i,j}(x) + \sum_{j: f_{i,j} \text{ cve}} (\rho_{i,j}^{\max} - h_{i,j}(\lambda^i)) (-f_{i,j}(x)) \right) \end{aligned} \quad (40)$$

It is important to highlight the following facts with regard to the above representation:

- If  $\rho_{i,j}^{\min} \leq 0$  for some  $j$  such that  $f_{i,j}$  is convex then  $\rho_{i,j}^{\min} f_{i,j}(x)$  is a *concave* function on  $X$ .
- If  $\rho_{i,j}^{\max} \geq 0$  for some  $j$  such that  $f_{i,j}$  is concave then  $\rho_{i,j}^{\max} f_{i,j}(x)$  is a *concave* function on  $X$ .
- Let

$$g_i(x) \triangleq \text{maximum}_{\lambda^i \in \Lambda^i} \left[ \underbrace{\sum_{j: f_{i,j} \text{ cvx}} (h_{i,j}(\lambda^i) - \rho_{i,j}^{\min}) f_{i,j}(x) + \sum_{j: f_{i,j} \text{ cve}} (\rho_{i,j}^{\max} - h_{i,j}(\lambda^i)) (-f_{i,j}(x))}_{\triangleq \varphi_i(x, \lambda^i)} \right]. \quad (41)$$

Since the maximand  $\varphi_i(x, \lambda^i)$  is a convex function in  $x$  for each  $\lambda^i$ , it follows readily that  $g_i(x)$  is a convex function on  $X$ .

• By combining the above three observations, it is clear that each  $\theta_i(x)$  is a dc-function. To be explicit, we define some index sets:

$$\begin{aligned} \mathcal{J}_i^{-, \text{cvx}} &\triangleq \left\{ j \mid \rho_{i,j}^{\min} \leq 0 \text{ and } f_{i,j} \text{ is cvx} \right\} \\ \mathcal{J}_i^{+, \text{cvx}} &\triangleq \left\{ j \mid \rho_{i,j}^{\min} \geq 0 \text{ and } f_{i,j} \text{ is cvx} \right\} \\ \mathcal{J}_i^{+, \text{cve}} &\triangleq \left\{ j \mid \rho_{i,j}^{\max} \geq 0 \text{ and } f_{i,j} \text{ is cve} \right\} \\ \mathcal{J}_i^{-, \text{cve}} &\triangleq \left\{ j \mid \rho_{i,j}^{\max} \leq 0 \text{ and } f_{i,j} \text{ is cve} \right\}. \end{aligned}$$

We can then write

$$\begin{aligned}
\theta_i(x) &= \sum_{j:f_{i,j} \text{ cvx}} \rho_{i,j}^{\min} f_{i,j}(x) + \sum_{j:f_{i,j} \text{ cve}} \rho_{i,j}^{\max} f_{i,j}(x) + g_i(x) \\
&= \left[ \underbrace{\sum_{j \in \mathcal{J}_i^-, \text{cvx}} \rho_{i,j}^{\min} f_{i,j}(x) + \sum_{j \in \mathcal{J}_i^+, \text{cve}} \rho_{i,j}^{\max} f_{i,j}(x)}_{\triangleq u_i(x)} \right] + \left[ \underbrace{\sum_{j \in \mathcal{J}_i^+, \text{cvx}} \rho_{i,j}^{\min} f_{i,j}(x) + \sum_{j \in \mathcal{J}_i^-, \text{cve}} \rho_{i,j}^{\max} f_{i,j}(x) + g_i(x)}_{\triangleq v_i(x)} \right]
\end{aligned}$$

where the function  $u_i(x)$  is concave and differentiable while  $v_i(x)$  is convex and non-differentiable. Hence

$\theta_i(x)$  is a dc function; thus so is  $\theta(x) = \sum_{i=1}^I \theta_i(x) = u(x) + v(x)$ , where  $u(x) \triangleq \sum_{i=1}^I u_i(x)$  is concave and

differentiable and  $v(x) \triangleq \sum_{i=1}^I v_i(x)$  is convex and nondifferentiable. Consequently, (3) is a nondifferentiable

dc program. The noteworthy point is that in the representation  $\theta(x) = u(x) + v(x)$ , the concave summand  $u(x)$  is differentiable whereas the convex summand  $v(x)$  is not; thus  $\theta$  is not a good dc function.