JOTA manuscript No.
(will be inserted by the editor)

On the Linear Convergence to Weak/Standard D-stationary Points of DCA-based Algorithms for Structured Nonsmooth DC Programming

Hongbo Dong · Min Tao

Received: date / Accepted: date

Abstract We consider a class of structured nonsmooth difference-of-convex minimization. We allow nonsmoothness in both the convex and concave components in the objective function, with a finite max structure in the concave part. Our focus is on algorithms that compute a (weak or standard) d(irectional)-stationary point as advocated in a recent work of Pang et al. in 2017. Our linear convergence results are based on direct generalizations of the assumptions of error bounds and separation of isocost surfaces proposed in the seminal work of Luo et al. in 1993, as well as one additional assumption of locally linear regularity regarding the intersection of certain stationary sets and dominance regions. An interesting by-product is to present a sharper characterization of the limit set of the basic algorithm proposed by Pang et. al., which fits between d-stationarity and global optimality. We also discuss sufficient conditions under which these assumptions hold. Finally, we provide several realistic and nontrivial statistical learning models where all assumptions hold.

Keywords Difference-of-convex programming · Nonsmooth · Difference-of-convex algorithm · Linear convergence · Error bound

Mathematics Subject Classification (2000) 90C30 · 90C26

H. Dong

Facebook, Palo Alto, CA. E-mail: h.dong.acc@gmail.com

M. Tao (Corresponding author)

Department of Mathematics, National Key Laboratory for Novel Software Technology, Nanjing University, China.

E-mail: taom@nju.edu.cn

1 Introduction

Optimization with (structured) nonconvex and nonsmooth functions has become one of the primary focuses of contemporary research in optimization, partially motivated by interesting applications in statistics [1,2], machine learning and compressive sensing [3]. While analysis of general nonconvex nonsmooth functions is much more complicated than the convex case, many existing techniques in convex analysis can be exploited in the study of difference-of-convex (DC) functions [4,5]. It has been observed that DC functions is in fact a very rich class of nonconvex functions in optimization; e.g, see [6,7] for recent examples and discussions. DC programming and the DC algorithm (DCA) have been successfully employed in many applications [1–3,5,8,9]. See [10] for a recent survey.

In this paper, we focus on a class of structured nonsmooth DC programming where both of the convex and concave parts of the objective function are allowed to be nonsmooth and with a finite max structure in the concave part [1–3]. The first question immediately raised is about "stationary" which usually plays the role of computational solution. There are two commonly used stationarity notions in DC programming: *critical point* and *strongly critical point* [4,5,8,10,11]. Much of the literature of DC programming uses the notion of *critical point*, and the classical DCA converges to such point. Criticality (critical point) can be regarded as a relaxation of strong criticality (strongly critical point). Recently, Pang and coauthors in [12] advocated using the concept of *d(irectional)-stationary* point which is arguably the sharpest kind among the various stationary solutions for the structured nonsmooth DC programming. On the other hand, it has been verified that d stationarity is equivalent to strong criticality under some technique conditions [11, Theorem3.1]. Obviously, strong criticality is the strongest necessary condition for local optimality. Furthermore, under some extra assumptions, strong criticality is also sufficient for optimality [8]. From theoretically aspect, strong criticality can be achievable via using *the complete DCA*, see e.g. [4,8]. The complete DCA remains nonconvex programs and thus they are still of a difficult task [8]. Recently, some DCA-based novel algorithms were developed for converging to d-stationary points [9,12]. These algorithms inherited the advantage of the classical DCA, and they only required to solve a series of convex subproblem. In [9,12], its subsequential convergence were established. The authors did not provide sequential convergence analysis and convergence rate for such algorithms. At the same time, the authors of [12] also introduced another stationary notion, called *weak d-stationary* (also called *lifted stationary*). This notion is related to the concept of *quasi-Nash equilibrium* (QNE) [14] which is popular in characterizing the stationarity of Nash equilibrium problem. However, there does not present any algorithms to compute a weak d-stationary point in [12].

The main purpose of the present paper is to propose two types of algorithms for a class of structured nonsmooth DC programming: one is a specific version of DCA converging to a *weak* d-stationary point, while the other is a variant of DCA, called ε-DCA (which is equivalent to Algorithm 1 in [12]) for computing a d-stationary point. According to the existing literatures on the algorithms for computing d-stationarity [9,12], it seems to be that it is hard to establish the whole sequential convergence for such algorithms, let alone its convergence rate analysis. Therefore, we aim to prove its linear convergence rate of these two algorithms under some conditions, thus its sequential convergence results are also established as a corollary. More specifically, our assumptions include a locally linear regularity condition, the generalized versions of the classical error bound and proper separation of isocost surfaces conditions proposed in [13]. We also discuss sufficient conditions for such assumptions. To the best of our knowledge, our work is the first [1] to identify conditions under which the linear convergence to *a (weak or standard) d(irectional)-stationary point* for a large class of nonsmooth DC programs studied in [12] can be established.

As a by-product, we show that the basic algorithm in [12] in fact converges to a closed subset of all d-stationary points; hence providing a sharper characterization of the limit points of this algorithm and a stronger computable optimality condition than d-stationary. By using the notion of approximate subdifferentials, this characterization naturally lies between the d-stationary set and the global optimum as characterized in [8,15]. Although when preparing this manuscript, we became aware of a recent work [16] which proposed an equivalent concept, our interpretation by

[1] The first version of this paper with complete results appeared in 2018 on Optimization Online `http://www.optimization-online.org/DB_HTML/2018/08/6766.html`

using approximate subdifferentials is new and connects to the global optimality condition in a more stylish manner.

The rest of this paper proceeds as follows. In Section 2, we present technical setting and nomenclature and review some related works. In Section 3, some preliminary discussions, including the equivalence of a prox-linear mapping (used in the literature) and a DC step with a difference decomposition by two strongly convex functions and a key technical lemma are presented. In section 4, we consider a specific version of DCA (sDCA) where the subgradient is chosen to be an active gradient. We show this algorithm converges to a weak d-stationary point. Three key assumptions including the error bound and proper separation of isocost surfaces, as well as one additional assumption on the locally linear regularity of related sets are introduced in this section. We prove linear convergence (and as a corollary, sequential convergence) of sDCA under these assumptions. In section 5, we consider the basic algorithm proposed in [12] for computing a d-stationary point. Firstly, as a detour, we provide a sharper characterization of the set of points that this algorithm may converge to. It further motivates us to propose two notions: A-stationarity and A_ε-stationarity which naturally lying between d-stationarity and global optimum. We then prove the linear convergence (and sequential convergence) of the considered algorithm. The proof in this section is somewhat in parallel to that in Section 5, although with important differences. We conduct further discussions on checking these key assumptions in Section 6. Especially, we show that several statistical estimation models satisfy all these key assumptions. Finally, we present some concluding remarks in Section 7.

2 Technical Setting, Nomenclature and Related Works

In this paper, we consider the following structured nonsmooth DC program with convex constraints:

$$\min_{x\in X}\ F(x) := H(x) - \max_{1\le i\le m}\ g_i(x), \tag{1}$$

where $X \subseteq \mathbb{R}^n$ and $G(x) := \max_{1\le i\le m}\ g_i(x)$. We make the following blanket assumption throughout this paper:

Assumption 1 *The triplet $(H, X, \{g_i\}_{i=1}^m)$ satisfy the following properties with positive parameters $(\sigma, \{L_i\}_{i=1}^m)$:*

(a) X is a closed and convex set in $\mathbb{R}^n$ (or $\mathbb{R}^n$ itself);

(b) $H(\cdot)$ and $g_i(\cdot)$ $(i = 1, ..., m)$ are proper closed convex functions whose domains contain an open superset of X; $g_i(\cdot)$ is continuously differentiable at every point in X, and $\nabla g_i(\cdot)$ is Lipschitz on X with modulus $L_i > 0$;

(c) (Strong convexity) $H(\cdot)$ and $g_i(\cdot)$ $(i = 1, ..., m)$ are all strongly convex with parameter $\sigma > 0$;

(d) (Level-boundedness) For any $\alpha \in \mathbb{R}$, $\{x \in X \mid F(x) \le \alpha\}$ is bounded; $F(x)$ is bounded below on X, i.e., $\inf_{x\in X} F(x) > -\infty$.

It is easy to see that the strong convexity assumption incurs no loss of generality as a quadratic term $\sigma\|x\|^2/2$ can be simultaneously added to $H(\cdot)$ and all $g_i(\cdot)$ $(i = 1, ..., m)$. More specifically, by setting $\tilde{H}(x) := H(x) + \frac{\sigma}{2}\|x\|^2$ and $\tilde{g}_i(x) := g_i(x) + \frac{\sigma}{2}\|x\|^2$, it yields that

$$F(x) = \tilde{H}(x) - \tilde{G}(x), \;\; \tilde{G}(x) := \max_{1\le i\le m} \tilde{g}_i(x).$$

For the special case $m = 1$, we will often use lower case $g(\cdot)$ instead of $G(\cdot)$ in (1), and without confusion we say a triplet (H, X, g) satisfies Assumption 1 with parameter (σ, L). For the general case, we define $\hat{L} := \max_{1\le i\le m} L_i$.

Although, for general DC program, the symmetric relation between the primal DC program and the dual DC program can provide us with many inspiring ideas to work with [4,5,8]. However, the dual DC program of (1) can not be provided in an explicit way. Therefore, we only focus on the primal DC program (1) instead of considering both of them. Obviously, the difference-of-convex algorithm (DCA) (also called *simplified DCA*, see [5,8]) can be directly applied to (1). As we know, there exists an infinite number of DC decomposition of (1) and each of them results in a different DCA scheme. For succinctness, we only focus on the DC decomposition of $F := H - G$. With properly chosen x^0, in $(k + 1)$-th iteration, the following convex subproblem is solved to compute x^{k+1}:

$$x^{k+1} \leftarrow \operatorname*{argmin}_{x\in X} \left\{H(x) - \eta(x^k)^\top(x - x^k)\right\}, \tag{2}$$

where $\eta(x^k) \in \partial G(x^k)$, and $\partial G(\cdot)$ represents the subdifferentials of $G(\cdot)$ at x^k. Under Assumption 1, we have $\partial G(x) = \textbf{conv}\left\{\nabla g_i(x) \;:\; i = 1, ..., m\right\}, \; \forall x \in X$, where **conv** denotes the convex hull. The standard convergence theory of DCA (see e.g. [5]), guarantees that any limit point of the sequence generated by (2) is a *critical point*. A vector $\bar{x} \in X$ is a critical point to (1) if

$$\partial G(\bar{x}) \cap \big(\partial H(\bar{x}) + \mathcal{N}_X(\bar{x})\big) \neq \emptyset, \tag{3}$$

where $\mathcal{N}_X(\bar{x})$ is the normal cone of X at $\bar{x}$. (We will use ι_X to denote the indicator function of X which takes value 0 on X and $+\infty$ elsewhere. Under Assumption 1, it is easy to see that $\partial(H + \iota_X)(x) = \partial H(x) + \mathcal{N}_X(x)$ for all $x \in X$.) Another criticality notion for DC program is *strong criticality* [11]. A vector $\bar{x}$ is called a strongly critical point of (1) if

$$\partial G(\bar{x}) \subseteq \big(\partial H(\bar{x}) + \mathcal{N}_X(\bar{x})\big), \tag{4}$$

which is stronger than critical point. As already mentioned in some early works (e.g., [15, Remark 3.6]), the notion of critical point depends on the DC decomposition, which is not unique. Also, one can find univariate *convex* functions with non-minimizing critical points [17, Example 2].

Pang and coauthors in [12] advocated using the concept of *d(irectional)-stationary* points instead. A point $\bar{x} \in X$ is called a d-stationary point to (1) if $F'(\bar{x}; y - \bar{x}) \geq 0,\ \forall y \in X$, where $F'(\bar{x}; y - \bar{x})$ is the directional derivative of $F(\cdot)$ at $\bar{x}$ in the feasible direction $y - \bar{x}$. Obviously, this definition depends entirely on the geometry of F (and X) instead of the DC decomposition used. Since $F'(\bar{x}; y - \bar{x}) = H'(\bar{x}; y - \bar{x}) - G'(\bar{x}; y - \bar{x})$ and $G'(\bar{x}; y - \bar{x}) = \max_{v \in \partial G(\bar{x})} v^\top (y - \bar{x})$ under Assumption 1, it follows that $\bar{x}$ is a d-stationary of (1) if and only if

$$H'(\bar{x}; y - \bar{x}) \geq v^\top (y - \bar{x}),\ \forall v \in \partial G(\bar{x}),\ \forall y \in X,$$

which is equivalent to $\bar{x} \in \arg\min_{x \in X} (H(x) - v^\top x)$, for any $v \in \partial G(\bar{x})$. Consequently, it is easy to see that $\bar{x} \in X$ is d-stationary implies $\partial G(\bar{x}) \subseteq \partial H(\bar{x}) + \mathcal{N}_X(\bar{x})$ which is precisely the definition of strong criticality (see (4)) [11].

Another stationarity tailed for the nonsmooth DC program (1) is called *weak d-stationarity*. As indicated in [12], $x \in X$ is called a *weak d-stationary* point for (1) if and only if there exists $i \in \mathcal{M}(x)$ such that $H'(x; x' - x) \geq g_i'(x; x' - x)\ \left(= \nabla g_i(x)^\top (x' - x)\right),\ \forall x' \in X$, where $\mathcal{M}(x)$ is the active set and defined as:

$$\mathcal{M}(x) \triangleq \underset{1 \leq j \leq m}{\operatorname{argmax}}\, g_j(x) = \{ j : g_j(x) = G(x) \}. \tag{5}$$

Local convergence rates for DCA and its variants have been studied in the literature [8] for some special cases of (1). One popular approach for convergence rate analysis is to exploit Kurdyka-Łojasiewicz (KL) property established in [18], which holds for proper closed subanalytic functions. In [11], Le Thi, Huynh and Pham Dinh analyzed the convergence rate of DCA under the assumptions that one of the gradients of $(H + \iota_X)$ and G of (1) is Lipschitz continuous and that KL property holds for the objective function. For the special case that $G(\cdot)$ in (1) is continuously differentiable (in other words $m = 1$), Wen, Chen and Pong [19] studied a variant of DCA (with extrapolation steps), proved the sequential convergence under the assumption of an auxiliary function with KL property. Very recently, Liu, Pong and Takeda [20] further studied the sequential convergence of the variant of DCA proposed in [19], and removed the assumption of $\nabla G(\cdot)$ continuously differentiable while still requiring the KL property of another merit function [20]. The convergence rate results in [8, 11, 19, 20] depend on the (usually unknown) exponent in the KL inequality, and locally linear convergence can only be achieved when this exponent is no more than $\frac{1}{2}$. Moreover, all these convergence results are aiming at converging to a critical point defined in (3) instead of a (weak or standard) d-stationary point. In a very recent work [16], Lu, Zhou and Sun proved the sequential convergence of proximal DC algorithm with extrapolation step converging to a d-stationary by assuming that one of the two conditions holds: (a) one of the elements of Γ is isolated (Γ denotes the accumulation set of the generated sequence); (b) A certain merit function is a KL function and for each $\bar{x} \in \Gamma$, $\mathcal{M}(\bar{x})$ defined in (5) is a singleton and the parameter ε in the proximal DCA satisfies $0 < \varepsilon < (G(\bar{x}) - \max_{i \in \mathcal{M}^c(\bar{x})} \nabla g_i(\bar{x}))/2$ where $\mathcal{M}^c(\cdot)$ denotes its complementary set of $\mathcal{M}(\cdot)$. However, when the concerned stationary point $\bar{x}$ is satisfying that the index number of $\mathcal{M}(\bar{x})$ is a singleton, these three concepts of critical point, weak d-stationary and d-stationary are equivalent for the problem (1) which will be shown in Lemma 3.2.

An alternative approach to derive linear convergence rate is based on error bound assumption. Some early references include [13, 21–23]. In some recent papers (e.g., [24, 25]), an error bound has been defined for the case of $m = 1$ and $G(\cdot)$ not necessarily convex. It is a direct generalization of the error bound assumption defined in [13] for convex-constrained smooth minimization. We adapt related notions to our problem setting (1) with Assumptions 1 and $m = 1$. For clarity, we use $g(\cdot)$ to replace $G(\cdot)$ when $m = 1$. Firstly, the *subdifferential set* of function $(H + \iota_X - g)$ is given by:

$$\partial(H + \iota_X - g)(x) = \partial H(x) + \mathcal{N}_X(x) - \nabla g(x), \quad \forall x \in X. \tag{6}$$

This definition coincides with Frechet, limiting and Clarke subdifferentials of $(H + \iota_X - g)$ in variational analysis. Let Ω be the set of all stationary points defined as,

$$\Omega := \left\{\bar{x} \in X \ : \ \nabla g(\bar{x}) \in \partial H(\bar{x}) + \mathcal{N}_X(\bar{x})\right\}. \tag{7}$$

Note that Ω is exactly the set of d-stationary points in this case. In [26], an error bound holds around $\bar{x} \in X$ if there exists an open neighborhood of $\bar{x}$, denoted by $\mathcal{N}$, such that

$$\operatorname{dist}(x, \Omega) \leq \tau \cdot \left\|x - \operatorname{Prox}_{\rho^{-1}(H+\iota_X)}(x + \rho^{-1}\nabla g(x))\right\|, \quad \forall x \in \mathcal{N}, \tag{8}$$

where $\rho > 0$ is some (fixed) parameter and $\operatorname{Prox}_f(v) := \arg\min_x \left\{f(x) + \frac{1}{2}\|x - v\|^2\right\}$ is the usual proximal mapping defined for a convex function f. ($\|\cdot\|$ is the Euclidean norm throughout this paper.) An important observation made in [26] (which generalizes a similar observation in [27] for the convex setting) is that (8) is the same as the *subregularity* of the prox-gradient mapping

$$x \to \sigma\left(x - \operatorname{Prox}_{\sigma^{-1}(H+\iota_X)}(x + \sigma^{-1}\nabla g(x))\right),$$

where " $\to$ " represents a mapping. It is then equivalent to (up to a constant factor) the subregularity of the set-valued mapping

$$x \to \partial(H + \iota_X - g)(x)$$

as defined in (6). In [25, Assumption 3.1] and [24, Assumption 2], the inequality (8) with a uniform τ is assumed to hold for all x such that $F(x) \leq \zeta$ and $\|x - \operatorname{Prox}_{H+\iota_X}(x + \nabla G(x))\| \leq \varepsilon$ where ζ and ε are positive scalars such that $\inf_{x\in X} F(x) \leq \zeta$. Convergence rate analysis, including conditions under which linear convergence holds, are aiming at converging to a critical point provided in these papers [13, 21–25]. Therefore, none of these results apply to our setting due to the existence of nonsmoothness in the "concave part" $G(\cdot)$, where the characterization of d-stationarity is more complicated than (7).

3 Preliminary and Some Technical Lemmas

In this section, we present some preliminary discussions and technical lemmas to be used in later section. Throughout this paper, let $\mathbb{R}^n$ denote the n-dimensional Euclidean space, $\langle \cdot, \cdot \rangle$ denote the standard inner product, and $\|\cdot\|$ denote the Euclidean norm. For any $\delta > 0$, $\mathbf{B}_\delta(x)$ is the open neighborhood $\{y : \|y - x\| < \delta\}$. Given an index set $\mathcal{M}$, $|\mathcal{M}|$ denotes the index number in the set $\mathcal{M}$. We use $\mathcal{O}^c$ to denote the complement set of $\mathcal{O}$. The domain of a function $f : \mathbb{R}^n \to [-\infty, \infty]$ is denoted by $\mathbf{dom}\, f = \{x \in \mathbb{R}^n : f(x) < +\infty\}$. A proper closed function f is said to be level bounded if for any $\alpha \in \mathbb{R}$, the set $\{x \in \mathbb{R}^n : f(x) \le \alpha\}$ is bounded. Given a convex function f defined on $\mathbb{R}^n$ and a positive scalar ε, a vector $y \in \mathbb{R}^n$ is an ε-subgradient of f at $\bar{x}$ if

$$f(x) \ge f(\bar{x}) + y^\top (x - \bar{x}) - \varepsilon, \qquad \forall x \in \mathbb{R}^n\,. \tag{9}$$

The set of all ε-subgradients of f at $\bar{x}$, i.e., the ε-subdifferential, is denoted as $\partial_\varepsilon f(\bar{x})$. By rearranging terms in (9) and taking supremum over x, it is straightforward to see that $y \in \partial_\varepsilon f(\bar{x})$ if and only if $f^*(y) + f(\bar{x}) - \bar{x}^\top y \le \varepsilon$. When $\varepsilon = 0$, $\partial_\varepsilon f$ is the usual convex subdifferential. It is easy to verify that for any $0 \le \varepsilon \le \varepsilon'$, $\partial_\varepsilon f \subseteq \partial_{\varepsilon'} f$. More results on approximate subdifferentials can be found in [28]. For a convex set $X \subseteq \mathbb{R}^n$ and $x \in X$, the normal cone of X at x is denoted by $\mathcal{N}_X(x)$ [29].

The following lemma includes a few basic facts of weak d-stationary points.

Lemma 3.1 *(i) A point $\bar{x} \in X$ is a weak d-stationary point for* (1) *if and only if there exists $i \in \mathcal{M}(\bar{x})$ such that*

$$\nabla g_i(\bar{x}) \in \partial(H + \iota_X)(\bar{x}) = \partial H(\bar{x}) + \mathcal{N}_X(\bar{x}). \tag{10}$$

(ii) Let $\{x^k\}_{k=1}^\infty$ be a sequence in X converging to x^∞. Suppose that there exists a sequence $\{z^k\}$ such that $z^k \in \partial(H + \iota_X)(x^k)$ and $\lim_{k \mapsto +\infty} z^k = z^\infty$, then $z^\infty \in \partial(H + \iota_X)(x^\infty)$.

Proof The proof is elementary, thus we omit here. □

Similarly, as clarified in [12], $x \in X$ is called a *d-stationary* point for (1) if and only if for *any* $i \in \mathcal{M}(x)$ such that $\nabla g_i(\bar{x}) \in \partial H(\bar{x}) + \mathcal{N}_X(\bar{x})$.

The following lemma present a sufficient condition to show when these three stationary concepts are equivalent.

Lemma 3.2 *Considering the problem* (1)*, suppose that $\bar{x} \in X$, and $|\mathcal{M}(\bar{x})| = 1$. Then, the following three assertions are equivalent:*

(i) $\bar{x}$ is a critical point;

(ii) $\bar{x}$ is a weak d-stationary point;

(iii) $\bar{x}$ is a d-stationary point.

Proof For (i)⇒(ii), suppose that $\bar{x}$ is a critical point, i.e., $\partial G(\bar{x}) \cap (\partial H(\bar{x}) + \mathcal{N}_X(\bar{x})) \neq \emptyset$. Note that $|\mathcal{M}(\bar{x})| = 1$, and let $i := \mathcal{M}(\bar{x})$, one has $\partial G(\bar{x}) = \nabla g_i(\bar{x})$. Consequently, we gets

$$\nabla g_i(\bar{x}) \in \partial(H + \iota_X)(\bar{x}).$$

Hence, the assertion (ii) holds. Note the converse direction is also true according to the above proof. For (ii)⇔(iii), it follows directly from $|\mathcal{M}(\bar{x})| = 1$. □

Next, we establish explicitly the equivalence of a prox-linear step used in algorithms in [12, 19, 25, 26] and a DC step with a perturbed DC decomposition. Consider the triplet (H, X, g) satisfying Assumption 1 with parameters (σ, L) and the associated DC program

$$\min_{x \in X} \; f(x) := H(x) - g(x), \tag{11}$$

a prox-linear mapping $\mathcal{G} : X \to X$ for (11) with parameter $\rho > 0$ is

$$\mathcal{G}_\rho^{H+\iota_X, g}(x) := \mathrm{Prox}_{\rho^{-1}(H+\iota_X)}(x + \rho^{-1}\nabla g(x)) = \operatorname*{argmin}_{y \in X}\{H(y) - \nabla g(x)^\top y + \frac{\rho}{2}\|y - x\|^2\}.$$

A typical DC step employs the following mapping which amounts to the prox-linear mapping with $\rho = 0$:

$$T^{H+\iota_X, g}(x) := \operatorname*{argmin}_{y \in X} \Big\{ H(y) - \nabla g(x)^\top y \Big\}.$$

Note that the right hand side is a singleton as H is assumed to be strongly convex in Assumption 1. Obviously $\mathcal{G}_\rho^{H+\iota_X, g} \neq T^{H+\iota_X, g}$ in general. However if we define $\hat{H}$ and $\hat{g}$ by

$$\hat{H}(x) := H(x) - \frac{\sigma}{2}\|x\|^2, \; \hat{g}(x) := g(x) - \frac{\sigma}{2}\|x\|^2.$$

Note that $H - g = \hat{H} - \hat{g}$. Then, for any $x \in X$,

$$T^{H+\iota_X, g}(x) = \operatorname*{argmin}_{y \in X} \Big\{ \hat{H}(y) + \frac{\sigma}{2}\|y\|^2 - (\nabla \hat{g}(x) + \sigma x)^\top y \Big\} = \mathcal{G}_\sigma^{\hat{H}+\iota_X, \hat{g}}(x).$$

Note that the convexity of $\hat{H}$ and $\hat{g}$ is guaranteed by the σ-strong convexity of H and g. This equivalence suggests to use the DC-step mapping in our later discussion, which enjoys a slightly simpler form with one less parameter to carry. Another immediate implication is the Lipschitz continuity of the DC-step mapping, i.e., for any $x, y \in X$,

$$\begin{aligned}
&\left\|T^{H+\iota_X,g}(x) - T^{H+\iota_X,g}(y)\right\| = \left\|\mathcal{G}_\sigma^{\hat{H}+\iota_X,\hat{g}}(x) - \mathcal{G}_\sigma^{\hat{H}+\iota_X,\hat{g}}(y)\right\| \\
&= \left\|\mathrm{Prox}_{\sigma^{-1}(\hat{H}+\iota_X)}(x+\sigma^{-1}\nabla\hat{g}(x)) - \mathrm{Prox}_{\sigma^{-1}(\hat{H}+\iota_X)}(y+\sigma^{-1}\nabla\hat{g}(y))\right\| \\
&\le \|x - y + \sigma^{-1}[(\nabla g(x) - \sigma x) - (\nabla g(y) - \sigma y)]\| \le \sigma^{-1}L\|x-y\|. \qquad (12)
\end{aligned}$$

The first inequality is due to the nonexpansivity of proximal operators [30], and the last inequality is due to the Lipschitz continuity of ∇g. Without confusion, in the rest of this paper, we will use T to denote $T^{H+\iota_X,g}$ when we discuss the triplet (H, X, g), and $T^{(i)} := T^{H+\iota_X,g_i}$ when we discuss the general case of $\left(H, X, \{g_i\}_{i=1}^m\right)$.

Next we introduce the concept of *locally linearly regularity* (LLR) for the intersection of two closed sets, and this concept stated in [31] as follows.

Definition 3.1 Given two closed sets C and D in $\mathbb{R}^n$, and $\bar{x} \in C \cap D$. We say $C \cap D$ is *locally linearly regular* at $\bar{x}$ with parameter (δ, η) if there exists an open neighborhood $\mathbf{B}_\delta(\bar{x})$ and a constant $\eta \ge 1$ such that

$$\mathrm{dist}(x, C\cap D) \le \eta \max(\mathrm{dist}(x, C), \mathrm{dist}(x, D)), \quad \forall x \in \mathbf{B}_\delta(\bar{x}). \qquad (13)$$

This definition is symmetric over C and D. Since in our context, the two sets involved have different meanings, we find it convenient to work with the following equivalent, but asymmetric, definition.

Definition 3.2 (Locally Linearly Regularity) Given an ordered pair of closed sets (C, D) in $\mathbb{R}^n$ and $\bar{x} \in C \cap D$, we say *locally linearly regularity* (LLR) holds at $\bar{x}$ with parameter (δ, η) if there exists an open neighborhood $\mathbf{B}_\delta(\bar{x})$ and a constant $\eta \ge 1$ such that

$$\mathrm{dist}(x, C\cap D) \le \eta\, \mathrm{dist}(x, C), \quad \forall x \in D \cap \mathbf{B}_\delta(\bar{x}). \qquad (14)$$

The following proposition proves the equivalence of these two definitions:

Proposition 3.1 *Let C and D be two closed sets in $\mathbb{R}^n$, and let $\bar{x} \in C \cap D$. If* (13) *holds with parameters (δ, η), then* (14) *holds with the same parameters (δ, η). If* (14) *holds with parameters (δ, η), then* (13) *holds with parameters $(\delta/2, 2\eta + 1)$.*

Proof The proof for (13) ⇒ (14) is straightforward. Suppose that $x \in D \cap \mathbf{B}_\delta(\bar{x})$, then we have $\mathrm{dist}(x, D) = 0$. Hence, $\max(\mathrm{dist}(x, C), \mathrm{dist}(x, D)) = \mathrm{dist}(x, C)$. Now, suppose that (14) holds with parameters (δ, η). Consider any arbitrary $y \in \mathbf{B}_{\delta/2}(\bar{x})$, we let $\hat{x}$ be a point in D such that

$\|y-\hat{x}\| = \operatorname{dist}(y, D) \leq \|y-\bar{x}\| < \delta/2$. By triangular inequality $\|\hat{x}-\bar{x}\| < \delta$, i.e., $\hat{x} \in D \cap \mathbf{B}_\delta(\bar{x})$. By (14), $\operatorname{dist}(\hat{x}, C \cap D) \leq \eta \operatorname{dist}(\hat{x}, C)$. Therefore

$$\begin{aligned}
&\operatorname{dist}(y, C \cap D) \leq \|y-\hat{x}\| + \operatorname{dist}(\hat{x}, C \cap D) \leq \operatorname{dist}(y, D) + \eta \operatorname{dist}(\hat{x}, C) \\
&\leq \operatorname{dist}(y, D) + \eta(\|y-\hat{x}\| + \operatorname{dist}(y, C)) = (\eta+1) \operatorname{dist}(y, D) + \eta \operatorname{dist}(y, C) \\
&\leq (2\eta+1) \max(\operatorname{dist}(y, C), \operatorname{dist}(y, D)). \qquad \square
\end{aligned}$$

The following technical lemma establish an inequality that will be used multiple times in our later analysis.

Lemma 3.3 *Let (H, X, g) be a triplet satisfying Assumption 1 with parameters (σ, L), and set f with $H-g$. Let $T := T^{H+\iota_X, g}$, then for any $x,\ y \in X$,*

$$f(y) \geq f(T(x)) - \frac{L}{2}\|x-y\|^2 + \frac{\sigma}{2}\|T(x)-y\|^2 + \frac{\sigma}{2}\|T(x)-x\|^2.$$

Proof By the definition of $T(x)$,

$$\nabla g(x)^\top (T(x)-y) \geq H(T(x)) - H(y) + \frac{\sigma}{2}\|T(x)-y\|^2,\ \forall\, y \in X. \tag{15}$$

By strong convexity of $g(\cdot)$, $g(T(x)) \geq g(x) + \nabla g(x)^\top (T(x)-x) + \frac{\sigma}{2}\|T(x)-x\|^2$. By Lipschitz continuity of $\nabla g(\cdot)$, we get $g(x) - g(y) + \frac{L}{2}\|x-y\|^2 \geq \nabla g(x)^\top (x-y)$. Adding these three inequalities together, we have

$$H(y) - g(y) \geq H(T(x)) - g(T(x)) - \frac{L}{2}\|x-y\|^2 + \frac{\sigma}{2}\|T(x)-y\|^2 + \frac{\sigma}{2}\|T(x)-x\|^2.$$

Then, by invoking the definition of f, the desired inequality follows directly. □

The following is a simple fact regarding the accumulation points of a bounded sequence.

Lemma 3.4 *Suppose that $\{x^k\}_{k=1}^\infty$ is a bounded sequence with the set of all accumulation points $\mathcal{L}$. Then $\mathcal{L}$ is compact and $\lim_{k \mapsto +\infty}$ dist$(x^k, \mathcal{L}) = 0$.*

Proof The proof is similar to [32, Proposition 1], thus we omit here. □

4 Linear Convergence of sDCA Converging to a Weak d-Stationary Point

In this section, we consider a specific version of DCA (sDCA) where the subgradient $\eta(x^k)$ in the subproblem (2) is chosen to be one of the "active gradients". See Algorithm 1 for details. As a

Algorithm 1 A specific version of DCA for computing a weak d-stationary point

```
Initialization: choose $x^0 \in X$.
for $k = 0, 1, 2, ...$ do
    Choose an index $i_k \in \mathcal{M}(x^k)$, then compute $x^{k+1}$ by
        $x^{k+1} \longleftarrow T^{(i_k)}(x^k)$.        (16)
end for
```

preparation of our convergence rate analysis, we first show that this algorithm converges to a weak d-stationary point. Our main result in this section is the proof of (locally) linear convergence of this algorithm under three assumptions. The same set of assumptions will be used in Section 5 for analyzing an algorithm computing a d-stationary point of (1). Along with natural generalizations of the error bound and separation of isocost surfaces assumptions proposed in [13], we propose an additional regularity condition regarding the intersection of pairs of related sets. Each pair includes one set characterizing stationarity of $(H - g_i)$ while another set being the region where g_i is "active". Our linear convergence results include sequential convergence as a corollary.

We let $\varOmega^{(i)}$ to denote the set of all points where (10) holds for i, and $\mathcal{D}^{(i)}$ be the region where $g_i(x)$ is active, i.e.,

$$\varOmega^{(i)} := \{x \in X : \nabla g_i(x) \in \partial(H + \iota_X)(x)\}\,, \quad \mathcal{D}^{(i)} := \{x \in X : g_i(x) \geq G(x)\}\,. \tag{17}$$

Then, the set of all weak d-stationary points can be written as

$$\bigcup_{i=1}^{m} \left(\varOmega^{(i)} \cap \mathcal{D}^{(i)}\right).$$

For Algorithm 1, it is easy to show that the optimality condition of the subproblem (16) is characterized with

$$\nabla g_{i_k}(x^k) \in \partial(H + \iota_X)(x^{k+1}), \tag{18}$$

which will be repeatedly used in our later analysis. The following is a simple lemma regarding the active sets of the points in an open neighborhood of $\bar{x} \in X$.

Lemma 4.1 *For any $\bar{x} \in X$, there exists an open neighborhood $\mathbf{B}_\delta(\bar{x})$ ($\delta > 0$) such that for any $x \in \mathbf{B}_\delta(\bar{x})$, $\mathcal{M}(x) \subseteq \mathcal{M}(\bar{x})$.*

Proof By definition for any $i \notin \mathcal{M}(\bar{x})$, $g_i(\bar{x}) < G(\bar{x})$. By continuity, there exists a neighborhood $\mathbf{B}_\delta(\bar{x})$, such that for any $x \in \mathbf{B}_\delta(\bar{x})$ and any $i \notin \mathcal{M}(\bar{x})$, $g_i(x) < G(x)$, which implies that $\mathcal{M}(x) \subseteq \mathcal{M}(\bar{x})$. □

The following lemma is known in the literature; e.g., [8]. For completeness, we present here as a simple corollary of Lemma 3.3. Consequently, subsequential convergence of Algorithm 1 follows directly from this lemma.

Lemma 4.2 *Let $\{x^k\}$ be the sequence generated by Algorithm 1. For each k,*

$$F(x^{k+1}) \leq F(x^k) - \frac{\sigma}{2}\|x^{k+1} - x^k\|^2. \tag{19}$$

Proof Recall that i_k is the active index chosen in step k of Algorithm 1. By applying Lemma 3.3 with $f = F_{i_k} := H - g_{i_k}$, $x = y = x^k$ and $T = T^{H+\iota_X, g_{i_k}}$, we have

$$F_{i_k}(x^k) \geq F_{i_k}(x^{k+1}) + \sigma \|x^{k+1} - x^k\|^2.$$

Then, (19) follows directly by noting the facts that $F(x^k) = F_{i_k}(x^k)$ and $F_{i_k}(x^{k+1}) \geq F(x^{k+1})$. □

Next, we prove the convergence of Algorithm 1. Note that the assertions (i)-(iii) of Theorem 4.1 are standard in the literature [11], and thus we omit the proof here.

Theorem 4.1 *Let the sequence $\{x^k\}$ be generated by Algorithm 1. Then, the following properties hold:*

(i) The sequence $\{F(x^k)\}$ is convergent, i.e.,

$$F^* := \lim_{k \to +\infty} F(x^k); \tag{20}$$

(ii) The sequence $\{x^k\}$ is bounded;

(iii) $\sum_{k=1}^{\infty} \|x^k - x^{k+1}\|^2 < +\infty$ (which implies that $\lim_{k \to +\infty} \|x^k - x^{k+1}\| = 0$);

(iv) Let $\{x^{k_j}\}_j$ be a subsequence of $\{x^k\}_k$ such that $x^{k_j} \to x^\infty$, and $\bar{i} \in \{1, ..., m\}$ is an index appearing infinitely many times in $\{i_{k_j}\}_j$, then $x^\infty \in \Omega^{(\bar{i})}$; Consequently, any accumulation point of $\{x^k\}$ is a weak d-stationary point;

(v) All accumulation points of $\{x^k\}$ have the same objective values;

(vi) Suppose that one of the elements in the accumulation set of $\{x^k\}$ is isolated. Then, the whole sequence $\{x^k\}$ converges to a weak d-stationary point.

Proof (iv) Let x^∞ be any accumulation point of $\{x^k\}$, and $\{x^{k_j}\}_{j=1}^\infty$ be a subsequence converging to x^∞. Since $\{i_{k_j}\}_{j=1}^\infty$ consists of finitely many distinct values, there are some indices appearing infinitely many times in this sequence. Without loss of generality (by restricting to a subsequence if necessary), we assume that $i_{k_j} \equiv \bar{i} \in \{1, ..., m\}$ for all j. By (iii), one has that $\lim_{k \to +\infty} \|x^k - x^{k+1}\| = 0$, which implies that $\{x^{k_j+1}\}_{j=1}^\infty$ also converges to x^∞. By the optimality condition (18), we have $\nabla g_{\bar{i}}(x^{k_j}) \in \partial H(x^{k_j+1}) + \mathcal{N}_X(x^{k_j+1})$. Since $\nabla g_{\bar{i}}(x^{k_j}) \to \nabla g_{\bar{i}}(x^\infty)$ as $j \to +\infty$, by Lemma 3.1, we have $\nabla g_{\bar{i}}(x^\infty) \in \partial H(x^\infty) + \mathcal{N}_X(x^\infty)$. Thus, we have $x^\infty \in \Omega^{(\bar{i})}$. Since $g_{\bar{i}}(\cdot)$ is active at infinitely many points $\{x_{k_j}\}_{j=1}^\infty$, it implies that $g_{\bar{i}}(x^{k_j}) \geq G(x^{k_j})$. Invoking $\{x_{k_j}\}_{j=1}^\infty$ converging to x^∞, it yields that $g_{\bar{i}}(x^\infty) \geq G(x^\infty)$. Thus, one gets $x^\infty \in \mathcal{D}^{(\bar{i})}$. Consequently, $\bar{i} \in \mathcal{M}(x^\infty)$. Therefore, $x^\infty \in \Omega^{(\bar{i})} \cap \mathcal{D}^{(\bar{i})}$, and it is a weak d-stationary point to (1). Finally, we prove (v). Let x^∞ be a accumulation point of $\{x^k\}$ and $\{x^{k_j}\}_{j=1}^\infty$ be a subsequence converging to x^∞. Then, by the continuity of F (over X) and the assertion (1),

$$F(x^\infty) = \lim_{j \to +\infty} F(x^{k_j}) = \lim_{k \to +\infty} F(x^k) = F^*.$$

For the assertion (vi), it follows by combining the assertion (iii) and Proposition 8.3.10 in [33]. □

To prove linear convergence of Algorithm 1, we will need the following construction and assumptions. For each k, we construct an auxiliary point $\bar{x}^k$ as follows

$$\bar{x}^k \in \arg\min_x \left\{ \|x - x^k\| : x \in \Omega^{(i_k)} \cap \mathcal{D}^{(i_k)} \right\}, \tag{21}$$

provided that $\Omega^{(i_k)} \cap \mathcal{D}^{(i_k)}$ is nonempty. Obviously, $\Omega^{(i_k)} \cap \mathcal{D}^{(i_k)}$ is a closed set due to Lemma 3.1 and the continuity of functions $\{g_i\}_{i=1}^m$. The following lemma shows that the projection (21) is well-defined for all sufficiently large k.

Lemma 4.3 *Let $\{x^k\}$ be the sequence generated by Algorithm 1, and i_k be the index chosen in step k. Let $\mathcal{L}$ to denote the set of all accumulation points of $\{x^k\}$. Then there exists K such that for all $k \geq K$, $\mathcal{L} \cap \mathcal{D}^{(i_k)} \cap \Omega^{(i_k)}$ is nonempty, where i_k is the index chosen in* (16)*.*

Proof It is easy to see that $\mathcal{L}$ is a closed set. Since $\{x^k\}$ is bounded, so is $\mathcal{L}$. By Lemma 3.4, we have $\lim_{k\to\infty} \operatorname{dist}(x^k, \mathcal{L}) = 0$. We show that $\mathcal{L} \cap \mathcal{D}^{(i_k)} \cap \Omega^{(i_k)}$ is non-empty for all k sufficiently large. It suffices to show the non-emptiness of $\mathcal{L} \cap \mathcal{D}^{(\bar{i})} \cap \Omega^{(\bar{i})}$ for all $\bar{i}$ appearing infinitely many times in the sequence $\{i_k\}_k$. Take such a subsequence $\{x^{k_j}\}_j$ with $i_{k_j} \equiv \bar{i}$. Any of its accumulation points is obviously in $\mathcal{L}$, and also in $\mathcal{D}^{(\bar{i})}$ due to the continuity of functions $\{g_j\}_{j=1}^m$. Also, such a limit point is in $\Omega^{(\bar{i})}$ due to the assertion (iv) of Theorem 4.1. Hence, $\mathcal{L} \cap \mathcal{D}^{(\bar{i})} \cap \Omega^{(\bar{i})} \neq \emptyset$ for all $\bar{i}$ appearing infinitely many times. It further implies that $\mathcal{L} \cap \mathcal{D}^{(i_k)} \cap \Omega^{(i_k)} \neq \emptyset$ for all sufficiently large k. □

Our analysis of linear convergence depends on three assumptions. The first two are natural generalization of the error bound and the proper separation of isocost surfaces assumptions as discussed in [13] and the references therein. The third assumption is a regularity condition on the intersection of $\Omega^{(i)}$ and $\mathcal{D}^{(i)}$ at a limit point. It is reasonable to expect the necessity of such a regularity condition, as the step $x^k \to T^{(i_k)}(x^k)$ moves x^k "towards" $\Omega^{(i_k)}$, while "knowing nothing" about $\mathcal{D}^{(i_k)}$.

We are now ready to state the main assumptions for proving linear convergence.

Assumption 2 *Let $\Omega^{(i)}$, $\mathcal{D}^{(i)}$ be sets defined in* (17)*. Define the index set*

$$\mathcal{I} = \left\{ j : \Omega^{(j)} \cap \mathcal{D}^{(j)} \neq \emptyset \right\}. \tag{22}$$

The following three conditions hold:

(A) For any $\zeta \geq \inf_{x\in X} F(x)$, there exists τ, $\varepsilon > 0$ such that for any $i \in \mathcal{I}$,

$$\mathit{dist}(x, \Omega^{(i)}) \leq \tau \|x - T^{(i)}(x)\|, \quad \forall x \in \mathcal{D}^{(i)},\ F(x) \leq \zeta,\ \|x - T^{(i)}(x)\| \leq \varepsilon.$$

(B) There exists a positive constant μ such that for any $i \in \mathcal{I}$ and $x_1, x_2 \in \Omega^{(i)} \cap \mathcal{D}^{(i)}$, it holds that $\|x_1 - x_2\| \geq \mu$ whenever $F(x_1) \neq F(x_2)$.

(C) For each $i \in \mathcal{I}$, the intersection $\Omega^{(i)} \cap \mathcal{D}^{(i)}$ satisfies the LLR condition at every point of $\Omega^{(i)} \cap \mathcal{D}^{(i)}$.

Remark 4.1 The readers may compare Assumption 2(A), 2(B) with [25, Assumption 3.1] applied to the problem of minimizing $(H - g_i)$ over X. Note that our assumption 2(A) is slightly weaker than [25, Assumption 3.1(i)] in the sense that we only need the error bound to hold for all x

in $\mathcal{D}^{(i)}$. Obviously, if Assumption 3.1(i) in [25] holds for $F_i := H + \iota_X - g_i$ with $i \in \mathcal{I}$, then Assumption 2(A) holds. One sufficient (but not necessary) condition for our Assumption 2(B) is that [25, Assumption 3.1(ii)] holds for all $(H - g_i)$ $(i = 1, ..., m)$. If for all $i \in \mathcal{I}$ and $x_1, x_2 \in \Omega^{(i)}$, $H(x_1) - g_i(x_1) \neq H(x_2) - g_i(x_2)$ implies that $\|x_1 - x_2\| \geq \mu$, then Assumption 2(B) holds. This is due to the simple fact that if $x_1, x_2 \in \mathcal{D}^{(i)}$, then $F(x_j) = H(x_j) - g_i(x_j),\ j = 1, 2$.

Now, consider Assumption 2(C). In fact, it will become evident that we only require the LLR condition to hold at all accumulation points of the sequence generated by Algorithm 1. Since the set of all accumulation points is compact under our blanket Assumption 1, the LLR (see Definition 3.2) parameters (δ, η) defined with respect to the concerned point x can be made uniform for all the points lying in the set $\mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)}$ which is nonempty.

Lemma 4.4 (Existence of uniform LLR parameters) *Suppose Assumption 2(C) holds and $\mathcal{L}$ is a compact subset of $\mathbb{R}^n$, then for each $i \in \{1, ..., m\}$ such that $\mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)} \neq \emptyset$, there exists uniform parameters $\bar{\delta}_i > 0$ and $\bar{\eta}_i \geq 1$ such that*

$$\mathit{dist}(x, \Omega^{(i)} \cap \mathcal{D}^{(i)}) \leq \bar{\eta}_i \cdot \mathit{dist}(x, \Omega^{(i)}), \qquad \forall x \in \mathbf{B}_{\bar{\delta}_i}(\mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)}) \cap \mathcal{D}^{(i)}. \tag{23}$$

Proof Let i be any index such that $\mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)} \neq \emptyset$. For any $\bar{x} \in \mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)}$, by Assumption 2(C), there exists $\delta_i(\bar{x}) > 0$ and $\eta_i(\bar{x}) \geq 1$ such that,

$$\mathrm{dist}(y, \Omega^{(i)} \cap \mathcal{D}^{(i)}) \leq \eta_i(\bar{x}) \cdot \mathrm{dist}(y, \Omega^{(i)}), \qquad \forall y \in \mathbf{B}_{\delta_i(\bar{x})}(\bar{x}) \cap \mathcal{D}^{(i)}. \tag{24}$$

Since $\mathcal{V} := \left\{\mathbf{B}_{\delta_i(\bar{x})/2}(\bar{x}) : \bar{x} \in \mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)}\right\}$ is an open cover of the compact set $\mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)}$, there exists a finite subcover. Let S to denote the set of centers of open balls in this finite subcover. We claim that (23) holds with $\bar{\eta}_i := \max_{\bar{x} \in S} \eta_i(\bar{x})$ and $\bar{\delta}_i := \min_{\bar{x} \in S} \delta_i(\bar{x})/2$. Indeed, note that for any $y \in \mathbf{B}_{\bar{\delta}_i}(\mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)}) \cap \mathcal{D}^{(i)}$, there exists $\hat{x} \in S$ such that $\|y - \hat{x}\| < \delta_i(\hat{x})/2 + \bar{\delta}_i \leq \delta_i(\hat{x})$. Thus, (23) follows easily from (24). □

The following lemma implies that for all k sufficiently large, $F(\bar{x}^k) = F^*$ where F^* is defined in (20).

Lemma 4.5 *Let $\{x^k\}$ be the sequence generated by Algorithm 1, and let i_k be the index chosen in step k, we have*

$$\lim_{k \to +\infty} \mathit{dist}(x^k, \mathcal{L} \cap \Omega^{(i_k)} \cap \mathcal{D}^{(i_k)}) = 0. \tag{25}$$

Moreover, if the Assumption 2(B) holds, then for all k sufficiently large, $F(\bar{x}^k) = F^*$ with $\bar{x}^k$ defined in (21)*.*

Proof Assume otherwise, i.e., $\lim_{k \to +\infty} \mathrm{dist}(x^k, \mathcal{L} \cap \Omega^{(i_k)} \cap \mathcal{D}^{(i_k)}) \neq 0$, then there exists a subsequence $\{x^{k_j}\}_j$ and $\varepsilon > 0$ such that $\mathrm{dist}(x^{k_j}, \mathcal{L} \cap \Omega^{(i_{k_j})} \cap \mathcal{D}^{(i_{k_j})}) \geq \varepsilon$ for all j. Without loss of generality, we assume the subsequence converges to $x^\infty \in \mathcal{L}$ such that $i_{k_j} \equiv \bar{i}$. Since $\bar{i}$ is active at all points in this subsequence, and combining its continuity of $\{g_i\}_{i=1}^m$, one has $\bar{i} \in \mathcal{M}(x^\infty)$, i.e.,

$x^\infty \in \mathcal{D}^{(\bar{i})}$. Therefore, by the assertion (iv) of Theorem 4.1, we have $x^\infty \in \mathcal{L} \cap \Omega^{(\bar{i})} \cap \mathcal{D}^{(\bar{i})}$. This contradicts with our assumption that $\operatorname{dist}(x^{k_j}, \mathcal{L} \cap \Omega^{(\bar{i})} \cap \mathcal{D}^{(\bar{i})}) \geq \varepsilon$. Consequently, we must have that $\lim_{k\to+\infty} \operatorname{dist}(x^k, \mathcal{L} \cap \Omega^{(i_k)} \cap \mathcal{D}^{(i_k)}) = 0$. Next, we further assume Assumption 2(B) holds. For all k sufficiently large, we have $\operatorname{dist}(x^k, \mathcal{L} \cap \Omega^{(i_k)} \cap \mathcal{D}^{(i_k)}) < \mu/2$ due to (25), where μ is defined in Assumption 2(B). By the definition of $\bar{x}^k$ in (21), we have

$$\|x^k - \bar{x}^k\| = \operatorname{dist}(x^k, \Omega^{i_k} \cap \mathcal{D}^{(i_k)}) \leq \operatorname{dist}(x^k, \mathcal{L} \cap \Omega^{i_k} \cap \mathcal{D}^{(i_k)}) < \mu/2.$$

Then, using triangle inequality, there exists $\hat{x}^k \in \mathcal{L} \cap \Omega^{i_k} \cap \mathcal{D}^{(i_k)}$ such that

$$\|\hat{x}^k - \bar{x}^k\| \leq \|x^k - \hat{x}^k\| + \|x^k - \bar{x}^k\| < \mu/2 + \mu/2 = \mu.$$

Finally, Assumption 2(B) implies that $F(\bar{x}^k) = F(\hat{x}^k) = F^*$. □

Lemma 4.6 *Let $\{x^k\}$ be a sequence generated by Algorithm 1. Suppose that Assumption 2(B) holds. Then for all k sufficiently large,*

$$F(x^{k+1}) \leq F^* + \frac{L_{i_k}}{2}\|x^k - \bar{x}^k\|^2,$$

where L_{i_k} is the Lipschitz constant for $\nabla g_{i_k}(\cdot)$.

Proof Applying Lemma 3.3 with $f = F_{i_k} := H - g_{i_k}$, $x = x^k$, $y = \bar{x}^k$ and $T = T^{H+\iota_X, g_{i_k}}$, we have

$$F_{i_k}(\bar{x}^k) \geq F_{i_k}(x^{k+1}) - \frac{L_{i_k}}{2}\|x^k - \bar{x}^k\|^2.$$

By definition, $F_{i_k}(x^{k+1}) \geq F(x^{k+1})$. Invoking (21) and Lemma 4.5, $F_{i_k}(\bar{x}^k) = F(\bar{x}^k) = F^*$. Therefore, we have the desired inequality. □

Now, we are ready to prove linear convergence of Algorithm 1 by piecing together all previous lemmas and constructions.

Theorem 4.2 *Let the sequence $\{x^k\}$ be generated by Algorithm 1, and let i_k be the index chosen in step k. Suppose that all conditions in Assumption 2 hold. Then, the following properties hold:*

(i) The sequence of objective function values $\{F(x^k) - F^*\}$ converges to zero Q-linearly;

(ii) The sequence $\{x^k\}$ converges R-linearly to a weak d-stationary point x^∞; furthermore, if $\hat{i}$ appears infinitely many times in $\{i_k\}_k$, then $x^\infty \in \Omega^{(\hat{i})} \cap \mathcal{D}^{(\hat{i})}$.

Proof (i) Invoking Lemma 4.6, we get $F(x^{k+1}) \leq F^* + \frac{\hat{L}}{2}\|x^k - \bar{x}^k\|^2$. Let $\mathcal{L}$ be the set of all accumulation points of $\{x^k\}$, and $(\bar{\delta}_i, \bar{\eta}_i)$ be the uniform LLR parameters as in Lemma 4.4. Furthermore, define

$$\delta := \min_i\{\bar{\delta}_i : \mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)} \text{ nonempty }\}, \quad \eta := \max_i\{\bar{\eta}_i : \mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)} \text{ nonempty }\}.$$

For all k sufficiently large such that $x^k \in \mathbf{B}_\delta(\mathcal{L} \cap \Omega^{(i_k)} \cap \mathcal{D}^{(i_k)})$, we have

$$\|x^k - \bar{x}^k\| = \operatorname{dist}(x^k, \Omega^{(i_k)} \cap \mathcal{D}^{(i_k)}) \leq \eta \operatorname{dist}(x^k, \Omega^{(i_k)}) \leq \eta\tau\|x^k - x^{k+1}\|,$$

where the last inequality follows from Assumption 2(A). Therefore,

$$F(x^{k+1}) \leq F^* + \frac{\hat{L}\eta^2\tau^2}{2}\|x^k - x^{k+1}\|^2.$$

Combining Lemma 4.2 and the above inequality, we obtain

$$F(x^{k+1}) \leq F^* + \frac{\hat{L}\eta^2\tau^2}{\sigma}(F(x^k) - F(x^{k+1})).$$

By rearranging terms, we have the desired inequality that for all k sufficiently large,

$$F(x^{k+1}) - F^* \leq \frac{M}{1+M}(F(x^k) - F^*),$$

where $M := \hat{L}\eta^2\tau^2/\sigma$.
(ii) Recalling Lemma 4.2, we obtain that

$$\|x^k - x^{k+1}\| \leq \sqrt{2/\sigma}\sqrt{F(x^k) - F(x^{k+1})} \leq \sqrt{2/\sigma}\sqrt{F(x^k) - F^*}.$$

Consequently,

$$\begin{aligned}&\sum_{i=0}^{+\infty}\|x^{k+i+1} - x^{k+i}\| \leq \sqrt{2/\sigma}\sum_{i=0}^{+\infty}\sqrt{F(x^{k+i}) - F^*} \\ &\leq \sqrt{2/\sigma}\left(1 - \sqrt{M/(M+1)}\right)^{-1}\sqrt{F(x^k) - F^*} < +\infty,\end{aligned}$$

where the last inequality is due to $\sqrt{F(x^k) - F^*}$ converges to zero Q-linearly. Since $\{x^k\}$ is bounded, it implies that its limit point of $\{x^k\}$ is unique, and we denote it by x^*. Then, using triangle inequality yields that

$$\|x^k - x^*\| \leq \sqrt{2/\sigma}\left(1 - \sqrt{M/(M+1)}\right)^{-1}\sqrt{F(x^k) - F^*}.$$

Because $\sqrt{F(x^k) - F^*}$ converges to zero Q-linearly, $\{x^k\}$ converges to x^* R-linearly. The last assertion follows easily from the assertion (iv) of Theorem 4.1 and the continuity of functions $\{g_i\}_{i=1}^m$. □

5 Linear Convergence of ε-DCA

We now consider the basic algorithm proposed in [12, Section 5.1] for computing a d-stationary solution. This algorithm is described in Algorithm 2. The ε-active set at any $x \in X$ is defined as:

$$\mathcal{M}_\varepsilon(x) := \{i : g_i(x) \geq G(x) - \varepsilon\}.$$

Note that the authors of [12] defined $x^{k,i}$ by a prox-linear step, which is equivalent to our mapping $T^{(i)}$ as argued by our discussion in Section 3. It has been shown in [12] that any accumulation point of the sequence generated by Algorithm 2 is a d-stationary point.

Algorithm 2 ε-DCA for computing a d-stationary point

```
Require: A triplet $(H, X, \{g_i\}_{i=1}^m)$ satisfying Assumption 1 with strong convexity modulus $\sigma$;
  Initialization: Choose $x^0 \in X$ and $\varepsilon > 0$;
  for $k = 0, 1, 2, ...$ do
      for $i \in \mathcal{M}_\varepsilon(x^k)$ do
          $x^{k,i} \leftarrow T^{(i)}(x^k)$          (26)
      end for
      Let
          $i_k \leftarrow \operatorname*{argmin}_{i \in \mathcal{M}_\varepsilon(x^k)} \left\{ F(x^{k,i}) + \frac{\sigma}{2}\|x^{k,i} - x^k\|^2 \right\};$          (27)
      Set $x^{k+1} \leftarrow x^{k,i_k}$.
  end for
```

5.1 A Sharper Characterization of the Limit Set of Algorithm 2

Before we get into the convergence rate analysis of Algorithm 2, we digress to derive a sharper characterization of the point that Algorithm 2 converges to. By examining the algorithm, it is intuitive that for any fixed $\varepsilon > 0$, there are some d-stationary points where Algorithm 2 cannot converge to. For each x of those d-stationary points, one could find an index $i \in \mathcal{M}_\varepsilon(x) \setminus \mathcal{M}(x)$ such that $T_i(x)$ would significantly reduce the objective function value in (1). We illustrate this by the following simple one-dimensional example. We apply Algorithm 2 to solve it with $\sigma = 1$:

$$F(x) = -\max(0, x) = \underbrace{x^2/2}_{:=H(x)} - \underbrace{\max(g_1(x), g_2(x))}_{:=G(x)}, \quad \text{where } g_1(x) = x^2/2,\ g_2(x) = x + x^2/2.$$

Apparently, $F(x)$ is unbounded below. However, every point in $]-\infty, 0[$ is a local optimal solution (hence a d-stationary point). Suppose that $x^0 \in]\max(-0.5, -\varepsilon), 0[$ (let $0 < \varepsilon < 0.5$). With some elementary calculations, it can be shown that Algorithm 2 cannot converge to the point in the range of $]\max(-0.5, -\varepsilon), 0[$ while every point in this range is a d-stationary point.

Next, we will provide a stronger optimality condition, which we call it $A_{\varepsilon'}$-stationarity, that holds at all accumulation points of Algorithm 2 with $0 < \varepsilon' < \varepsilon$. This optimality condition has some "global" flavor. In the literature, d-stationarity is sometimes considered as the sharpest kind of stationary point for nonsmooth DC programs [12], while being "sharpest" is under the implicit stipulation that only local first-order information is considered. However, note that every step of Algorithm 2 exploits the "global" information of the convex part $H(\cdot)$. It is meaningful to consider optimality conditions much stronger than d-stationarity. Furthermore, by using the notion of *approximate subdifferentials*, in a stylish manner, we show that our proposed optimality condition connects to the sufficient and necessary condition for *global optimality* for DC programs. The *global optimality* was proposed by J. B. Hiriart-Urruty in 1989 [15], and further studied in [8]. Therefore, our approach sheds some new lights in understanding the gap between the computed solutions by Algorithm 2 and the global optimal solutions of (1).

During the last stage of preparation of this paper, we were brought some attention to a recent manuscript [16]. In this paper, a concept called "(α,η)-D-stationary" was proposed. We will show that A_ε-stationarity is essentially equivalent to this concept. However, our interpretation by using approximate subdifferentials is novelty.

Recall that at every point $\bar{x} \in X$ and for any $i \in \{1,\dots,m\}$, the convex surrogate function

$$H(\cdot) - \nabla g_i(\bar{x})^\top(\cdot - \bar{x}) - g_i(\bar{x})$$

is a global over-estimator of $F(\cdot)$ at $\bar{x}$. If $\bar{x}$ was a global optimal solution of (1), minimizing this convex function over X would not yield a value better than $F(\bar{x})$. In fact, this is similar to the motivation for defining "(α,η)-D-stationarity" in [16]. The following proposition provides a concise characterization of this condition with approximate subdifferentials.

Proposition 5.1 *Let $\bar{x} \in X$, then for a fixed $i \in \{1,...,m\}$,*

$$F(\bar{x}) \le \min_{x\in X} \left\{H(x) - \nabla g_i(\bar{x})^\top(x-\bar{x}) - g_i(\bar{x})\right\} \tag{28}$$

if and only if $\nabla g_i(\bar{x}) \in \partial_{G(\bar{x})-g_i(\bar{x})}(H+\iota_X)(\bar{x})$.

Proof By the definition of ε-subgradients and $\bar{x} \in X$, $\nabla g_i(\bar{x}) \in \partial_{G(\bar{x})-g_i(\bar{x})}(H+\iota_X)(\bar{x})$ is equivalent to

$$\begin{aligned} & (H+\iota_X)^*(\nabla g_i(\bar{x})) + H(\bar{x}) - \bar{x}^\top\nabla g_i(\bar{x}) \le G(\bar{x}) - g_i(\bar{x}) \\ \Longleftrightarrow \quad & H(\bar{x}) - G(\bar{x}) \le \inf_{x\in X}\left\{H(x) - x^\top\nabla g_i(\bar{x})\right\} + \bar{x}^\top\nabla g_i(\bar{x}) - g_i(\bar{x}). \end{aligned}$$

which is the same as (28). □

Therefore, a necessary condition for $\bar{x}$ to be optimal to (1) is:

$$\bar{x} \in X \qquad \text{and} \qquad \nabla g_i(\bar{x}) \in \partial_{G(\bar{x})-g_i(\bar{x})}(H+\iota_X)(\bar{x}), \quad \forall i = 1,...,m. \tag{29}$$

It is easy to see that $\bar{x}$ is a d-stationary point to (1) if and only if the following weaker condition holds:

$$\bar{x} \in X \qquad \text{and} \qquad \nabla g_i(\bar{x}) \in \partial(H+\iota_X)(\bar{x}), \quad \forall i \in \mathcal{M}(\bar{x}). \tag{30}$$

For $\varepsilon > 0$, an optimality condition between (29) and d-stationarity is:

$$\bar{x} \in X \qquad \text{and} \qquad \nabla g_i(\bar{x}) \in \partial_{G(\bar{x})-g_i(\bar{x})}(H + \iota_X)(\bar{x}), \quad \forall i \in \mathcal{M}_\varepsilon(\bar{x}). \tag{31}$$

We call $\bar{x}$ to be *A-stationary* if it satisfies (29), and *A_ε-stationary* if it satisfies (31). It is interesting to compare these optimality conditions with the *necessary and sufficient* condition for global optimality in 1989 [15].

Theorem 5.1 (Theorem 4.4, [15]) *Let $f(x) = h(x) - g(x)$ be a difference-of-convex function for $x \in \mathbb{R}^n$, where h and g are proper closed convex functions. $\bar{x} \in \mathbb{R}^n$ is a global minimizer of $f(\cdot)$ if and only if $\partial_\alpha g(\bar{x}) \subseteq \partial_\alpha h(\bar{x})$, $\forall \alpha \geq 0$.*

Applying this theorem to (1), $\bar{x}$ is a global optimal solution to (1) if and only if

$$\bar{x} \in X, \qquad \partial_\alpha G(\bar{x}) \subseteq \partial_\alpha (H + \iota_X)(\bar{x}), \qquad \forall \alpha \geq 0. \tag{32}$$

A full characterization of $\partial_\alpha G(\bar{x})$ is rather complicated, we refer the readers to [28, Theorem 3.5.1] for a calculus rule for approximate subdifferentials of the max of finitely many convex functions, while only pointing out the following fact that [28, Page 124] for any $x \in X$ and $\alpha \geq 0$,

$$\big\{\nabla g_i(x) : i \in \mathcal{M}_\alpha(x)\big\} \subseteq \partial_\alpha G(x).$$

It is easy to verify this inclusion from definitions (although equality does not hold in general). Therefore, a condition weaker than (32) is

$$\bar{x} \in X \qquad \text{and} \qquad \nabla g_i(\bar{x}) \in \partial_\alpha (H + \iota_X)(\bar{x}), \quad \forall i \in \mathcal{M}_\alpha(\bar{x}), \ \forall \alpha \geq 0. \tag{33}$$

Note that $i \in \mathcal{M}_\alpha(\bar{x})$ is the same as $\alpha \geq G(\bar{x}) - g_i(\bar{x})$. By monotonicity of approximate subdifferentials it suffices to enforce $\nabla g_i(\bar{x}) \in \partial_{G(\bar{x})-g_i(\bar{x})}(H + \iota_X)(\bar{x})$ for all i. In other words, (33) is equivalent to A-stationarity defined in (29). For structured nonsmooth DC program (1), the relations between these conditions, strong criticality and criticality are summarized in the diagram in Fig. 1. By virtue of [11, Theorem 3.1], d-stationarity implies strong criticality under Assumption 1.

$$\begin{tikzcd}
\begin{array}{c}\text{Global optim.}\\ (32)\end{array} \arrow[r, Rightarrow] & \begin{array}{c}A\text{-stat.}\\ (33) \Leftrightarrow (29)\end{array} \arrow[r, Rightarrow] & \begin{array}{c}A_\varepsilon\text{-stat.}\\ (31)\end{array} \arrow[r] & \begin{array}{c}\text{d-stat.}\\ (30)\end{array} \arrow[r, Rightarrow, "\text{(Assumption 1)}"] & \begin{array}{c}\text{strong crit.}\\ (4)\end{array} \arrow[r, Rightarrow] & \begin{array}{c}\text{crit.}\\ (3)\end{array}
\end{tikzcd}$$

Fig. 1: Relations among different types of optimality conditions

Counterexample. We make several remarks with regard to the above string of implications:

(i) In general, a A-stationary point of (1) is not necessarily a global minimizer. A counterexample is provided by the univariate example:

$$\min_x F_1(x) = \underbrace{2 + \iota_{[-\sqrt{2},\sqrt{2}]}(x)}_{:=H(x)} - \underbrace{\max(g_1(x), g_2(x))}_{:=G(x)}, \text{ where } g_1(x) = 1,\ g_2(x) = x^2.$$

Obviously, $x = 0$ is a A-stationary point, but it is not a global minimizer.

(ii) Second, a d-stationary point is not necessarily A$_\varepsilon$-stationary. We consider the univariate example:

$$\min_x F_2(x) = \underbrace{x^2/2}_{:=H(x)} - \underbrace{\max(g_1(x), g_2(x))}_{:=G(x)}, \text{ where } g_1(x) = x^2/2,\ g_2(x) = x + x^2/2.$$

Let $0 < \nu < 0.5$, then every point in the range of $(-\nu, 0)$ is a local minimizer, and thus is a d-stationary point. However, every point in the range of $(-\nu, 0)$ is not A$_\varepsilon$-stationary with $\varepsilon \geq \nu$.

(iii) Third, an A$_\varepsilon$-stationary point is not necessarily A-stationary. We consider the univariate example:

$$\min_x F_3(x) = \underbrace{x^2 + x/2}_{:=H(x)} - \underbrace{\max(g_1(x), g_2(x))}_{:=G(x)}, \text{ where } g_1(x) = x/2,\ g_2(x) = 12x - 1.$$

Obviously, $x = 0$ is A$_\varepsilon$-stationarity with $0 < \varepsilon < 1$ and it is not a A-stationary point.

Remark 5.1 (Equivalence to (α,η)-D-stationarity in [16]*)* Let $(\hat{H}, X, \{\hat{g}_i\}_{i=1}^m)$ be a triplet satisfying Assumption 1 except for strong convexity. Let $F := \hat{H} - \max_{1\leq i\leq m} \hat{g}_i$. Then, by [16], $\bar{x} \in X$ is called (α,η)-D-stationary with $(\alpha,\eta) = (\sigma^{-1}, \varepsilon)$ if

$$F(\bar{x}) \leq \min_{x\in X} \left\{ \hat{H}(x) - \nabla \hat{g}_i(\bar{x})^\top (x - \bar{x}) - \hat{g}_i(\bar{x}) + \frac{\sigma}{2}\|x - \bar{x}\|^2 \right\}, \quad \forall i \in \mathcal{M}_\varepsilon(\bar{x}). \tag{34}$$

Note that if we define $H := \hat{H} + \sigma\|\cdot\|^2/2$ and $g_i := \hat{g}_i + \sigma\|\cdot\|^2/2$ for all i, then

$$\hat{H}(x) - \nabla \hat{g}_i(\bar{x})^\top (x - \bar{x}) - \hat{g}_i(\bar{x}) + \frac{\sigma}{2}\|x - \bar{x}\|^2 = H(x) - \nabla g_i(\bar{x})^\top (x - \bar{x}) - g_i(\bar{x}).$$

Then, by Proposition 5.1, (34) is equivalent to the A_ε-stationarity of $\bar{x}$ for minimizing $F(x)$ over X.

Note that (29) and (30) can be understood as the limit cases of (31) by letting $\varepsilon \uparrow +\infty$ and $\varepsilon \downarrow 0^+$, respectively. We will show that for the fixed $\varepsilon > 0$ chosen in Algorithm 2, any limit point of the sequence generated by Algorithm 2 is $A_{\varepsilon'}$-stationary [2] for any $\varepsilon' \in]0, \varepsilon[$. Although this has essentially been proved in [16] by the equivalence shown in Remark 5.1, we find it easier for presentation and completeness purposes to provide a concise proof using our notation and construction. The following results are also in parallel to Lemma 4.2 and Theorem 4.1.

Proposition 5.2 *Let the sequence $\{x^k\}$ be generated by Algorithm 2. Then, for all k,*

$$F(x^{k+1}) + \frac{\sigma}{2}\|x^{k+1} - x^k\|^2 \leq F(x^k).$$

Proof By choosing $i \in \mathcal{M}(x^k) \subseteq \mathcal{M}_\varepsilon(x^k)$, we apply Lemma 3.3 with setting (f, x, y, T) as $\left(H - g_i, x^k, x^k, T^{H+\iota_X, g_i}\right)$. The remaining proof is similar to Lemma 4.2, thus we omit here. □

Theorem 5.2 *Let the sequence $\{x^k\}$ be generated by Algorithm 2. Then, the following properties hold:*

(i) The sequence $\{F(x^k)\}$ is convergent;
(ii) The sequence of $\{x^k\}$ is bounded;
(iii) $\sum_{k=1}^{\infty} \|x^k - x^{k+1}\|^2 < +\infty$;
(iv) Any accumulation point of $\{x^k\}$ has the same objective value, i.e.,

$$\bar{F}^* := \lim_{k\to+\infty} F(x_k);$$

(v) Any accumulation point of $\{x^k\}$ is an $A_{\varepsilon'}$-stationary point, for any $\varepsilon' \in]0, \varepsilon[$;
(vi) Suppose that one of the elements in the accumulation set of $\{x^k\}$ is isolated. Then, the whole sequence $\{x^k\}$ converges to an $A_{\varepsilon'}$-stationary point for any $\varepsilon' \in]0, \varepsilon[$.

Proof The proof for the first four properties are identical to that of Theorem 4.1 by using Proposition 5.2. Now, we prove (v). Let x^∞ be an accumulation point of $\{x^k\}$, and $\{x^{k_j}\}_j$ is a subsequence converging to it. By the assertion (iii), $\{x^{k_j+1}\}_j$ also converges to x^∞. For any $i \in \mathcal{M}_{\varepsilon'}(x^\infty)$ where $\varepsilon' \in]0, \varepsilon[$, we have $i \in \mathcal{M}_\varepsilon(x^{k_j})$ for all sufficiently large j. Therefore, for all $i \in \mathcal{M}_{\varepsilon'}(x^\infty)$,

$$\begin{aligned} &F(x^{k_j+1}) \leq F(x^{k_j,i}) + \frac{\sigma}{2}\|x^{k_j,i} - x^{k_j}\|^2 \leq H(x^{k_j,i}) - g_i(x^{k_j,i}) + \frac{\sigma}{2}\|x^{k_j,i} - x^{k_j}\|^2 \\ &\leq H(x^{k_j,i}) - g_i(x^{k_j}) - \nabla g_i(x^{k_j})^\top (x^{k_j,i} - x^{k_j}) \leq H(y) - g_i(x^{k_j}) - \nabla g_i(x^{k_j})^\top (y - x^{k_j}),\ \forall y \in X. \end{aligned}$$

The first inequality is due to (27), the last is due to the update rule (26), respectively. Now, taking $j \to +\infty$ on both sides, we obtain

$$F(x^\infty) \leq \min_{y\in X}\left\{H(x) - g_i(x^\infty) - \nabla g_i(x^\infty)^\top (x - x^\infty)\right\}, \quad \forall i \in \mathcal{M}_{\varepsilon'}(x^\infty).$$

By Proposition 5.1, we have $\nabla g_i(x^\infty) \in \partial_{G(x^\infty) - g_i(x^\infty)}(H + \iota_X)(x^\infty)$ for any $i \in \mathcal{M}_{\varepsilon'}(x^\infty)$, i.e., x^∞ is $A_{\varepsilon'}$-stationary for any $\varepsilon' \in]0, \varepsilon[$. For the assertion (vi), it follows directly from the assertion (iii) and Proposition 8.3.10 in [33]. □

[2] Note the discrepancy between ε and ε'. This is consistent with the observation in [12] that if we take $\varepsilon = 0$ in Algorithm 2, a limit point is not necessarily d-stationary.

5.2 Convergence Rate Analysis

Finally, we prove the linear convergence of Algorithm 2 under Assumption 2. Our proof is somewhat parallel to the proofs in Section 4. However, there are important differences between them. The main difference is the way we define the projected auxiliary points. In fact, we introduce $\{y^{k,i}\}_{k,i}$ for all $i \in \mathcal{M}(x^k)$ while we only define one projection vector with the index i_k chosen at k-step of Algorithm 1 (see the definition (21)). More especially, we define

$$y^{k,i} \in \arg\min_y \left\{ \|y - x^k\| : y \in \Omega^{(i)} \cap \mathcal{D}^{(i)} \right\}. \tag{35}$$

We will show in the following lemma that for all k sufficiently large and $i \in \mathcal{M}(x^k)$, $y^{k,i}$ is properly defined (i.e., $\Omega^{(i)} \cap \mathcal{D}^{(i)}$ is nonempty). For a sequence $\{x^k\}$ generated by Algorithm 2, let us define the following index set

$$\mathcal{I}^\infty := \Big\{ i \in \{1, ..., m\} : i \in \mathcal{M}(x^k) \text{ for infinitely many } k \Big\}. \tag{36}$$

Lemma 5.1 *Let $\{x^k\}$ be the sequence generated by Algorithm 2, and $\mathcal{I}^\infty$ defined in (36). Then, (i) there exists K sufficiently large such that for all $k \geq K$, $\mathcal{M}(x^k) \subseteq \mathcal{I}^\infty$. (ii) $\mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)} \neq \emptyset$ for all $i \in \mathcal{I}^\infty$.*

Proof (i) We use contradiction to prove it. If it does not hold, it means that there exists a infinite index sequence $k_j \to +\infty$ such that $\exists\ i_{k_j} \in \mathcal{M}(x^{k_j})$ and $i_{k_j} \notin \mathcal{I}^\infty$. Note that $\{1, \ldots, m\}$ is a finite set. Invoking the Pigeonhole Principle, and by restricting the subsequence on hand, we have $i_{k_j} \equiv \bar{i}$ when $j \in \kappa$ where κ represents the subsequence. Then, we have $\bar{i} \notin \mathcal{I}^\infty$ due to $i_{k_j} \notin \mathcal{I}^\infty$. However, it contradicts to $\bar{i} \in \mathcal{I}^\infty$ due to $\bar{i} \in \mathcal{M}(x^{k_j})$ for infinitely many k_j. In other words, $\mathcal{M}(x^k) \subseteq \mathcal{I}^\infty$ for all $k \geq K$. (ii) We show that $\mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)} \neq \emptyset$ for all $i \in \mathcal{I}^\infty$. Take any $\bar{i} \in \mathcal{I}^\infty$. Suppose $\{x^{k_j}\}_j$ is a subsequence of $\{x^k\}$ such that $\bar{i} \in \mathcal{M}(x^{k_j})$ for all j and $x^{k_j} \to x^\infty$ as $j \to +\infty$. We show that $x^\infty \in \mathcal{L} \cap \mathcal{D}^{(\bar{i})} \cap \Omega^{(\bar{i})}$. Apparently, we have $x^\infty \in \mathcal{L}$ and $x^\infty \in \mathcal{D}^{(\bar{i})}$ (by continuity of $\{g_i\}_{i=1}^m$) and $x^\infty \in \Omega^{(\bar{i})}$ (as x^∞ is d-stationary by Theorem 5.2). □

Next, we will prove an analogous version of Lemma 4.5.

Lemma 5.2 *Let $\{x^k\}$ be the sequence generated by Algorithm 2. Then*

$$\lim_{k \to +\infty} \max_{i \in \mathcal{M}(x^k)} \left\{ dist(x^k, \mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)}) \right\} = 0. \tag{37}$$

Moreover, if Assumption 2(B) holds, then for all k sufficiently large and $i \in \mathcal{M}(x^k)$, $F(y^{k,i}) = \bar{F}^*$.

Proof Suppose the first argument (37) is not true. Then, there exists a $\varepsilon > 0$ and a subsequence $\{x^{k_j}\}$ such that $i_{k_j} \in \mathcal{M}(x^{k_j})$ and $\operatorname{dist}(x^{k_j}, \mathcal{L} \cap \Omega^{(i_{k_j})} \cap \mathcal{D}^{(i_{k_j})}) \geq \varepsilon$ for all j. Since the number of the index set $\{1, \ldots, m\}$ is finite, there is one index appears infinite numbers in the sequence $\{i_{k_j}\}$. Thus, there is a subsequence $\{x^{k_j}\}_{j \in \kappa}$ converges to $x^\infty \in \mathcal{L}$ and $i_{k_j} \equiv \bar{i}$ for $j \in \kappa$. Furthermore, by Theorem 5.2 and invoking the fact that x^∞ is d-stationary, one has $x^\infty \in \mathcal{L} \cap \Omega^{(\bar{i})} \cap \mathcal{D}^{(\bar{i})}$. This contradicts with our assumption that $\text{dist}(x^{k_j}, \mathcal{L} \cap \Omega^{(\bar{i})} \cap \mathcal{D}^{(\bar{i})}) \geq \varepsilon$. Therefore, we must have (37) holds.

Next, we further assume Assumption 2(B). For all k sufficiently large, we have

$$\text{dist}(x^k, \mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)}) < \mu/2, \quad \forall\, i \in \mathcal{M}(x^k),$$

where μ is defined in Assumption 2(B). By definition of $y^{k,i}$ defined in (35), we have for all $i \in \mathcal{M}(x^k)$, $\|x^k - y^{k,i}\| = \text{dist}(x^k, \Omega^{(i)} \cap \mathcal{D}^{(i)}) \leq \text{dist}(x^k, \mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)}) < \mu/2$. Therefore, by triangle inequality, there exists $\hat{x}^k \in \mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)}$ such that

$$\|\hat{x}^k - y^{k,i}\| \leq \|x^k - \hat{x}^k\| + \|x^k - y^{k,i}\| < \mu/2 + \mu/2 = \mu.$$

Then, Assumption 2(B) implies that $F(y^{k,i}) = \bar{F}^*$ for all $i \in \mathcal{M}(x^k)$ and k sufficiently large. □

Now, we are ready to prove the linear convergence of Algorithm 2 under Assumption 2. In fact, since all accumulation points of the sequence generated by Algorithm 2 are $A_{\varepsilon'}$-stationary for all $\varepsilon' \in]0, \varepsilon[$, Assumption 2(C) can be replaced by the following weaker version:

Assumption 3 *Suppose that ε is used in Algorithm 2. For each $i \in \mathcal{I}$, the intersection $\Omega^{(i)} \cap \mathcal{D}^{(i)}$ satisfies the LLR condition at every $A_{\varepsilon'}$-stationary point of (1) where $0 < \varepsilon' < \varepsilon$.*

Our proof is again different from that of Theorem 4.2. Here, we need to consider the quantity $Q(x^k, x^{k-1}) := F(x^k) + \frac{\sigma}{2}\|x^k - x^{k-1}\|^2$, and first show that $\{Q(x^k, x^{k-1})\}$ is linearly convergent. Note that by Theorem 5.2, $\lim_{k\to+\infty} Q(x^k, x^{k-1}) = \bar{F}^*$.

Theorem 5.3 *Suppose Assumption 2(A), 2(B) and Assumption 3 hold, and let $\{x^k\}$ be the sequence generated by Algorithm 2. Then, the following properties hold:*

(i) the sequence $\{Q(x^k, x^{k-1}) - \bar{F}^*\}$ converges to zero Q-linearly;
(ii) $\{F(x^k) - \bar{F}^*\}$ converges to zero R-linearly;
(iii) $\{x^k\}$ converges to an $A_{\varepsilon'}$-stationary point R-linearly, where ε' is an arbitrary value in $]0, \varepsilon[$.

Proof (i) We apply Lemma 3.3 with $(f, x, y, T) = (H - g_i, x^k, x^k, T^{H+\iota_X, g_i})$ for any $i \in \mathcal{M}_\varepsilon(x^k)$:

$$\begin{aligned} F_i(x^k) &\geq F_i(x^{k,i}) + \frac{\sigma}{2}\|x^{k,i} - x^k\|^2 + \frac{\sigma}{2}\|x^{k,i} - x^k\|^2 \\ &\geq Q(x^{k+1}, x^k) + \frac{\sigma}{2}\|x^{k,i} - x^k\|^2, \qquad \forall i \in \mathcal{M}_\varepsilon(x^k). \end{aligned}$$

The second inequality is due to the definition of $Q(x^{k+1}, x^k)$ and (27). Take $i \in \mathcal{M}(x^k)$. Then $F_i(x^k) = F(x^k) = Q(x^k, x^{k-1}) - \frac{\sigma}{2}\|x^k - x^{k-1}\|^2$. Then, it yields that

$$Q(x^{k+1}, x^k) - Q(x^k, x^{k-1}) \leq -\frac{\sigma}{2}\|x^{k-1} - x^k\|^2 - \frac{\sigma}{2}\|x^{k,i} - x^k\|^2 \leq -\frac{\sigma}{2}\|x^{k,i} - x^k\|^2.$$

Recalling Lemma 5.1, $y^{k,i}$ is properly defined for all k sufficiently large and all $i \in \mathcal{M}(x^k)$. We have the following inequality for such k and $i \in \mathcal{M}(x^k)$:

$$Q(x^{k+1}, x^k) \leq F_i(x^{k,i}) + \frac{\sigma}{2}\|x^{k,i} - x^k\|^2 \leq F_i(y^{k,i}) + \frac{\hat{L}}{2}\|x^k - y^{k,i}\|^2 = \bar{F}^* + \frac{\hat{L}}{2}\|x^k - y^{k,i}\|^2.$$

The first inequality is due to $Q(x^{k+1}, x^k)$ and the update rule (27). The second inequality is obtained by applying Lemma 3.3 by setting $(f, x, y, T) = (H - g_i, x^k, y^{k,i}, T^{H+\iota_X, g_i})$. The last equality is due to Lemma 5.2. The remaining proof is similar to the assertion (i) of Theorem 4.2, thus is omiited.

For the assertions (ii) and (iii), it follows from the fact that $Q(x^k, x^{k-1}) - \bar{F}^* = F(x^k) - \bar{F}^* + \frac{\sigma}{2}\|x^k - x^{k-1}\|^2$. □

6 Discussions on the Key Assumptions

In this final section, we focus on the key assumptions used to prove linear convergence of two algorithms in Sections 4 and 5. We discuss conditions under which these assumptions hold.

Firstly, consider the error bound condition, i.e., Assumption 2(A). We show a general result that this condition is equivalent to *subregularity*, an important concept in variational analysis, of a related subdifferential mapping. This result is inspired by a result in [26, Section 8], where a slightly different definition of error bound is used. This equivalence implies that Assumption 2(A) in fact does not depend on specific DC decompositions.

A set-valued mapping (or a multi-function) $M : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is said to be *subregular* at $(\bar{x}, \bar{y})$ with parameter $\ell > 0$ if $\bar{y} \in M(\bar{x})$ and there exists a neighborhood of $\bar{x}$, denoted by $\mathcal{O}$, such that

$$\operatorname{dist}(x, M^{-1}(\bar{y})) \le \ell \cdot \operatorname{dist}(\bar{y}, M(x)), \qquad \forall x \in \mathcal{O}\,.$$

For the sake of completeness, we first prove the following lemma which is an adapted version of a result in [26, Section 8] to our setting.

Proposition 6.1 *Let the triplet (H, X, g) satisfy Assumption 1 with parameters (σ, L). The subdifferential $\partial(H + \iota_X - g)(x)$ and the stationary set Ω are defined as in* (7)*. Suppose that $\Omega \neq \emptyset$ and $\bar{x} \in \Omega$, consider the following two conditions:*

(i) The multi-function $\partial(H + \iota_X - g)(\cdot)$ is subregular at $(\bar{x}, 0)$;

(ii) dist$(x, \Omega) \le \hat{\ell}\,\|x - T(x)\|$ for all $x \in X \cap \mathcal{O}$, where $\mathcal{O}$ is an open neighborhood of $\bar{x}$.

If condition (i) *holds with constant ℓ, then condition* (ii) *holds with constant $\hat{\ell} = 1 + \ell L$; if condition* (ii) *holds with $\hat{\ell}$, then condition* (i) *holds with $\ell = 2\hat{\ell}/\sigma$.*

Proof For (ii)⇒(i), suppose that condition (ii) holds with constant $\hat{\ell}$ and neighborhood $\mathcal{O}$ of $\bar{x}$. We show that in the open neighborhood $\mathcal{O}$, $\operatorname{dist}(x, \Omega) \le 2\hat{\ell}\sigma^{-1} \operatorname{dist}(0, \partial(H + \iota_X - g)(x))$. For any $x \notin X$, $\partial(H + \iota_X - g)(x) = \emptyset$ [29, Proposition 23.4] and this inequality holds trivially. Now consider $x \in X \cap \mathcal{O}$. Let ν be a vector in $\partial(H + \iota_X - g)(x)$ and $\nu = \omega + h - \nabla g(x)$ for $\omega \in \partial H(x)$ and $h \in \mathcal{N}_X(x)$. By convexity of $(H + \iota_X)$ and the definition of normal cone, for any $x \in X \cap \mathcal{O}$,

$$\nu^\top(x - T(x)) \ge H(x) - \big(H(T(x) - \nabla g(x)^\top (T(x) - x)\big) \ge \frac{\sigma}{2}\|x - T(x)\|^2,$$

where the second inequality is due to inequality (15) with $y = x$. Thus,

$$\|\nu\| \cdot \|T(x) - x\| \geq \frac{\sigma}{2}\|x - T(x)\|^2 \Rightarrow \|x - T(x)\| \leq 2\sigma^{-1}\|\nu\|.$$

Therefore, $\mathrm{dist}(x, \Omega) \leq \hat{\ell}\,\|x - T(x)\| \leq 2\hat{\ell}\sigma^{-1}\|\nu\|,$ for all $x \in \mathcal{O}$. Since $\nu \in \partial(H + \iota_X - g)(x)$, condition (i) holds with factor $\ell = 2\hat{\ell}\sigma^{-1}$.

To prove the converse, suppose condition (i) holds with constant ℓ and neighborhood $\mathcal{O}$, we show that condition (ii) holds $\hat{\ell} = 1 + \ell L$. Firstly, note that $\bar{x} \in \Omega \Rightarrow T(\bar{x}) = \bar{x}$, by the Lipschitz continuity of T, i.e., inequality (12), there exists a neighborhood $\tilde{\mathcal{O}}$ of $\bar{x}$ such that for all $x \in \tilde{\mathcal{O}}$, $T(x) \in \mathcal{O}$. Then, we claim that for all $x \in \tilde{\mathcal{O}}$,

$$\begin{aligned}\mathrm{dist}(x, \Omega) &\leq \mathrm{dist}(T(x), \Omega) + \|x - T(x)\| \leq \ell\,\mathrm{dist}(0, \partial(H + \iota_X - g)(T(x))) + \|x - T(x)\| \\ &\leq [1 + \ell L]\,\|x - T(x)\|.\end{aligned}$$

The first inequality is the triangle inequality. The second is by applying condition (i) to $T(x)$, and the third inequality is because of the optimality condition of $T(x)$:

$$\nabla g(x) \in \partial H(T(x)) + \mathcal{N}_X(T(x)) \Longleftrightarrow \nabla g(x) - \nabla g(T(x)) \in \partial H(T(x)) + \mathcal{N}_X(T(x)) - \nabla g(T(x)),$$

which implies that $\mathrm{dist}(0, \partial(H + \iota_X - g)(T(x))) \leq \| - \nabla g(T(x)) + \nabla g(x)\| \leq L\|x - T(x)\|$. This concludes our proof. □

Note that the second condition in the previous proposition is not exactly the error bound we use. The following proposition fills this gap and shows that Assumption 2(A) is equivalent to subregularity of subdifferential mappings.

Proposition 6.2 *Let $(H, X, \{g_i\}_{i=1}^m)$ be a triplet satisfying Assumption 1 with scalars $\big(\sigma, \{L_i\}_{i=1}^m\big)$. Suppose that for any $i \in \mathcal{I}$ defined in (22) and $\bar{x} \in \Omega^{(i)} \cap \mathcal{D}^{(i)}$, the multifunction $\partial(H + \iota_X - g_i)$ is subregular at $(\bar{x}, 0)$ if and only if Assumption 2(A) holds.*

Proof We first show that if for all $i \in \mathcal{I}$ defined in (22) and $\partial(H + \iota_X - g_i)$ is subregular at $(\bar{x}, 0)$ for all $\bar{x} \in \Omega^{(i)} \cap \mathcal{D}^{(i)}$, then the following holds:

– For any $\zeta \geq \inf_{x \in X} F(x)$, there exists $\rho, \tau > 0$ such that for any $i \in \mathcal{I}$,

$$\text{(EBN)} \quad \mathrm{dist}(x, \Omega^{(i)}) \leq \tau\|x - T^{(i)}(x)\|, \text{ for all } \ x \in X,\ F(x) \leq \zeta,\ \mathrm{dist}(x, \Omega^{(i)} \cap \mathcal{D}^{(i)}) \leq \rho. \tag{38}$$

Since $\{x : F(x) \leq \zeta\}$ is compact, so is $\Omega^{(i)} \cap \mathcal{D}^{(i)} \cap \{x : F(x) \leq \zeta\}$. By Proposition 6.1, for any $\bar{x} \in \Omega^{(i)} \cap \mathcal{D}^{(i)}$ there exists $\hat{\tau} := \hat{\tau}(\bar{x})$ such that $\mathrm{dist}(x, \Omega^{(i)}) \leq \hat{\tau}\|x - T^{(i)}(x)\|$ in an open neighborhood of $\bar{x}$ in X. Such open neighborhoods form an open cover of

$$\Omega^{(i)} \cap \mathcal{D}^{(i)} \cap \{x : F(x) \leq \zeta\},$$

hence there exists a finite sub-cover. Let τ be the largest $\hat{\tau}$ among all $\bar{x}$ associated to this finite sub-cover, and $\mathcal{O}$ be the union of such finite open neighborhoods, then (38) holds with the quantifier "$\mathrm{dist}(x, \Omega^{(i)} \cap \mathcal{D}^{(i)}) \leq \rho$" replaced by "$x \in \mathcal{O}$". The existence of ρ can then be proved by applying the Weierstrass Theorem to the continuous function $\mathrm{dist}(\cdot, \mathcal{O}^c)$ defined on the compact set $\Omega^{(i)} \cap \mathcal{D}^{(i)} \cap \{x : F(x) \leq \zeta\}$.

Now, we show that (38) implies Assumption 2(A). It suffices to show that for any $\zeta \geq \inf_{x\in X} F(x)$, and the associated ρ in (38), there exists $\varepsilon > 0$ such that

$$x \in \mathcal{D}^{(i)},\ \ F(x) \leq \zeta,\ \|x - T^{(i)}(x)\| \leq \varepsilon \Rightarrow x \in X,\ F(x) \leq \zeta,\ \operatorname{dist}(x, \Omega^{(i)} \cap \mathcal{D}^{(i)}) \leq \rho.$$

Since $x \in X$ and $F(x) \leq \zeta$ are trivial consequences of conditions on the left, we focus on the last condition $\operatorname{dist}(x, \Omega^{(i)} \cap \mathcal{D}^{(i)}) \leq \rho$. Suppose otherwise, then there exists $\bar{\zeta} \geq \inf_{x\in X} F(x)$ and a sequence $\epsilon_k \downarrow 0$ and a sequence $\{x^{(k)}\}_k \subseteq \mathcal{D}^{(i)}$ such that for all k that

$$F(x^{(k)}) \leq \bar{\zeta},\ \|x^{(k)} - T^{(i)}(x^{(k)})\| \leq \epsilon_k \text{ and } \operatorname{dist}(x^{(k)}, \Omega^{(i)} \cap \mathcal{D}^{(i)}) > \rho.$$

Since F is level-bounded, let x^{∞} be a limit point of $\{x^{(k)}\}$. By the Lipschitz condition of $T^{(i)}$, we have $x^{\infty} = T^{(i)}(x^{\infty})$. In other words, $x^{\infty} \in \Omega^{(i)} \cap \mathcal{D}^{(i)}$. This is contradictory to the presumption $\operatorname{dist}(x^{(k)}, \Omega^{(i)} \cap \mathcal{D}^{(i)}) > \rho$ for all k.

Now, suppose Assumption 2(A) holds, we prove the subregularity of $\partial(H + \iota_X - g_i)$ at any $\bar{x} \in \Omega^{(i)} \cap \mathcal{D}^{(i)}$. Firstly, by choosing ζ such that $F(\bar{x}) < \zeta$ and using continuity, we have $F(x) < \zeta$ and $x \in \mathcal{D}^{(i)}$ for all $x \in X$ in an open neighborhood of $\bar{x}$. Consequently, there exists an open neighborhood $\mathcal{O}$ of $\bar{x}$ such that

$$\operatorname{dist}(x, \Omega^{(i)}) \leq \tau \|x - T^{(i)}(x)\|,\ \text{for all } x \in X \cap \mathcal{O}\,.$$

Then, Proposition 6.1 implies the multifunction $\partial(H + \iota_X - g_i)$ is subregular at $(\bar{x}, 0)$ for any $\bar{x} \in \Omega^{(i)} \cap \mathcal{D}^{(i)}$. □

Furthermore, existing results on types of functions satisfying error bound conditions (e.g., see [13, 21–24, 27, 34, 35]) can be used to check Assumption 2(A) in a piecewise manner. For example, by exploiting the following theorem in [24].

Proposition 6.3 [24] *We consider the following minimization problem:*

$$\min_{x\in\mathbb{R}^n}\ f(x) := H(x) + g(x), \tag{39}$$

where H is a proper closed convex function and g is a function (not necessarily concave nor convex) that has a Lipschitz continuous gradient. Suppose that the objective function f of (39) *is level-bounded and the stationary point set of* (39)*, denoted by $\bar{\Omega}$, is nonempty. Further, assume that H and g satisfied with one of the following conditions:*

(i) $g(x) = q(Ax)$ for all $x \in \mathbb{R}^n$, where $A \in \mathbb{R}^{m\times n}$ and q is a strongly convex differentiable function with ∇q Lipschitz continuous on any compact convex set and H is polyhedral;
(ii) g is a quadratic function (possibly nonconvex), and H is a polyhedral function.

Then, for any $\zeta \geq \inf_{x\in\mathbb{R}^n} f(x)$, there exist $\tau, \varepsilon > 0$ such that

$$\mathit{dist}(x, \Omega) \leq \tau \|x - T(x)\|, \quad \textit{whenever}\ f(x) \leq \zeta,\ \ \|x - T(x)\| \leq \varepsilon.$$

Proposition 6.4 *Let $\left(H, \mathbb{R}^n, \{g_i\}_{i=1}^m\right)$ be a triplet satisfying Assumption 1 with parameters $(\sigma, \{L_i\}_{i=1}^m)$. Let $F_i = H - g_i$ for all i. If F_i satisfies the assumptions in Proposition 6.3 for all $i \in \mathcal{I}$, then Assumption 2(A) holds.*

Proof Since $F = H - \max_i g_i$, then $F = \min_i F_i$. For any $\zeta \geq \inf_x F(x)$, we denote the index set

$$\mathcal{I}_\xi^* := \Big\{ i \ : \ \{x \ : \ \zeta \geq \inf_x F_i(x)\} \neq \phi \Big\} \cap \mathcal{I}.$$

By assumption, $\operatorname{dist}(x, \Omega^{(i)}) \leq \tau_i \|x - T^{(i)}(x)\|$, for all $F_i(x) \leq \zeta$, $\|x - T^{(i)}(x)\| \leq \varepsilon_i$. By setting $\tau := \max_{i \in \mathcal{I}_\xi^*} \tau_i$ and $\varepsilon := \min_{i \in \mathcal{I}_\xi^*} \varepsilon_i$ in Assumption 2(A), it holds directly. □

Next, we show that with the level boundedness assumption in Assumption 1, Assumption 2(B) holds whenever F_i takes only finitely many different values on $\Omega^{(i)} \cap \mathcal{D}^{(i)}$.

Proposition 6.5 *Let $\Big(H, X, \{g_i\}_{i=1}^m\Big)$ be a triplet satisfying Assumption 1. Assumption 2(B) holds if for all $i \in \mathcal{I}$ defined in* (22)*, $F_i(:= H - g_i)$ takes only finitely many distinct values on $\Omega^{(i)} \cap \mathcal{D}^{(i)}$.*

Proof Assume otherwise, there exists two sequences $\{x_n\}_n$, $\{y_n\}_n$ both in $\Omega^{(i)} \cap \mathcal{D}^{(i)}$ such that $F_i(x_n) \neq F_i(y_n)$ for all n and $\|x_n - y_n\| \to 0$ as $n \to +\infty$. Note that as $F(x_n) = F_i(x_n)$ and $F(y_n) = F_i(y_n)$ for all n, both sequences belong to the compact set $\{x \in X \mid F(x) \leq M\}$ where M is the largest value of F_i on $\Omega^{(i)} \cap \mathcal{D}^{(i)}$. Let x^* be a accumulation point of $\{x_n\}_n$. Since $\|x_n - y_n\| \to 0$, x^* is also a accumulation point of $\{y_n\}_n$. Therefore $|F_i(x_n) - F_i(y_n)|$ can be arbitrarily small as $n \to +\infty$. This contradicts with the assumption that $F_i(x_n) \neq F_i(y_n)$ and F_i takes only finitely many distinct values on $\Omega^{(i)} \cap \mathcal{D}^{(i)}$. □

Apparently, this condition holds if each F_i is convex while $F = \min_i F_i$ is nonconvex. Another classical example is when X is polyhedral, and each F_i is the summation of a nonconvex quadratic function and a convex polyhedral function [34, Lemma 3.1].

Finally, we consider Assumption 2(C) and Assumption 3, i.e., the locally linear regularity (LLR) of the intersection $\Omega^{(i)}$ and $\mathcal{D}^{(i)}$ at relevant points. Note that LLR is an nonconvex counterpart of the (bounded) linear regularity for convex sets. The following proposition is a direct application of [36, Corollary 3].

Proposition 6.6 *LLR holds if $\Omega^{(i)}$ and $\mathcal{D}^{(i)}$ are convex, and $\mathbf{ri}(\Omega^{(i)}) \cap \mathbf{ri}(\mathcal{D}^{(i)}) \neq \emptyset$; if $\Omega^{(i)}$ is polyhedral, then $\mathbf{ri}(\Omega^{(i)})$ can be replaced by $\Omega^{(i)}$, and similarly for $\mathcal{D}^{(i)}$.*

This result implies if both $\Omega^{(i)}$ and $\mathcal{D}^{(i)}$ are polyhedral for all i such that $\Omega^{(i)} \cap \mathcal{D}^{(i)} \neq \emptyset$, then Assumption 2(C) and Assumption 3 hold. This is true when X is polyhedral, each F_i is convex or (nonconvex) quadratic, and $\{g_i\}_{i=1}^m$ only differ from each other with an affine function.

6.1 Discussions on Regularized Statistical Learning Models

Taking the following optimization problem for regularized linear regression:

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2 + P(x; \lambda), \tag{40}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^n$ and λ is a positive tuning parameter. The regularization function P usually takes a separable form:

$$P(x;\lambda) = \sum_{i=1}^{n} \rho_i(x_i;\lambda). \tag{41}$$

Example 6.1 For the least squares problems (40) with capped-ℓ_1 [37] regularization, each regularizer ρ_i takes the form with an additional tuning parameter $\theta > 0$:

$$\rho(t;\lambda) = \begin{cases} \lambda|t|/\theta, & \text{if } |t| \le \theta, \\ \lambda, & \text{otherwise.} \end{cases}$$

Note that $\rho(t;\lambda)$ can be decomposed as $\lambda|t|/\theta - \lambda\max(0, t/\theta - 1, -t/\theta - 1)$, the optimization (40) can be formulated as our model (1) with triplet $(H, X, \{g_i\}_i)$ being

$$H(x) = \frac{1}{2}\|Ax - b\|^2 + \lambda\frac{\|x\|_1}{\theta}, \quad X = \mathbb{R}^n,$$

and each g_i being an affine function. By Proposition 6.3, Assumption 2(A) holds. Assumption 2(B) holds because of Proposition 6.5 and that each $H - g_i$ is convex (hence taking a fixed value at stationarity). Moreover, Assumption 2(C) (and hence the weaker Assumption 3) holds because of Proposition 6.6 (noting that both $\Omega^{(i)}$ and $\mathcal{D}^{(i)}$ are convex polyhedral).

Example 6.2 For the least squares problem (40) with MCP [38] regularization, the corresponding function ρ in (41) is defined as:

$$\rho_{\mathrm{MCP}}(t;\lambda) = \begin{cases} \lambda|t| - \frac{t^2}{2\theta}, & \text{if } |t| \le \theta\lambda, \\ \frac{\theta\lambda^2}{2}, & \text{if } |t| > \theta\lambda, \end{cases} \quad \text{with } \lambda > 0,\ \theta > 0.$$

It is not difficult to see that problem (40) with MCP [38] regularization is a special case of our model (1) with $m = 1$, where

$$H(x) = \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1,\ X = \mathbb{R}^n,\ g(x) = \lambda\sum_{i=1}^{n}\int_0^{|x_i|} \min\{1, \frac{x}{\theta\lambda}\}dx.$$

In the case of $m = 1$, d-stationarity coincides the weaker condition of criticality, and the linear convergence of DCA-type algorithms have been solved by proving the Kurdyka-Łojasiewicz exponent of the MCP regularizer model is $\frac{1}{2}$ [39]. It is not necessary to invoke our theory which deals with the case of general m and d-stationarity.

7 Conclusions

In this paper, we consider a class of structured nonsmooth DC minimization. Both of convex functions in the DC decomposition of the objective function are not necessarily smooth, and the second is with a finite max structure. We are the first to establish the linear convergence of these algorithms that compute a (weak or standard) d-stationary point under some error bound based assumptions. Furthermore, we discuss some sufficient conditions to ensure these key assumptions.

Acknowledgements Both authors would like to thank Prof. Jong-Shi Pang for numerous inspiring discussions on related topics during their individual visits to the University of Southern California in 2016. Prof. Pang also proposed the main question studied in this paper. Part of this work was completed during the second author's visit of Prof. Kung-Fu Ng at the Chinese University of Hong Kong. Min Tao was supported partially by National Key Research and Development Program of China (No. 2018AAA0101100), the Chinese National Natural Science Foundation Grant (No. 11971228) and the JiangSu Provincial National Natural Science Foundation of China (No. BK20181257). Both of the authors are grateful to anonymous referees and the associate editor for their valuable comments and suggestions which have helped to improve the presentation of this paper.

References

1. Ahn, M., Pang, J. S., Xin, J.: Difference-of-convex learning: Directional stationarity, optimality, and sparsity. SIAM J. Optim. 27, 1637–1665 (2017)
2. Gong, P., Zhang, C., Lu, Z., Huang, J. Z., Ye, J.: A general iterative shinkage and thresholding algorithm for non-convex regularized optimization problems. Proc. Int. Conf. Mach. Learn. 28, 37–45 (2013)
3. Gotoh, J. Y., Takeda, A., Tono, K.: DC formulations and algorithms for sparse optimization problems. Math. Program. 169, 141–176 (2018)
4. Pham Dinh, T., Souad, E. B.: Duality in D.C. (difference of convex functions) optimization. Subgradient Methods. In: Hoffmann KH., Zowe J., Hiriart-Urruty JB., Lemarechal C. (eds) Trends in Mathematical Optimization. International Series of Numerical Mathematics, vol 84. Birkhäuser Basel (1988).
5. Le Thi, H. A., Pham Dinh, T.: The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. Ann. Oper. Res. 133, 23–46 (2005)
6. Bačák, M., Borwein, J.: On difference convexity of locally lipschitz functions. Optimization 60, 961–978 (2011)
7. Nouiehed, M., Pang, J. S., Razaviyayn, M.: On the pervasiveness of difference-convexity in optimization and statistics. Math. Program. 174, 195–222 (2019)
8. Pham Dinh, T., Le Thi, H. A.: Convex analysis approach to DC programming: theory, algorithms and applications. Acta Mathematica Vietnamica 22, 289–355 (1997)
9. Pang, J. S., Tao, M.: Decomposition methods for computing directional stationary solutions of a class of non-smooth non-convex optimization problems. SIAM J. Optim. 28, 1640–1669 (2018)
10. Le Thi, H. A., Pham Dinh, T.: DC programming and DCA: thirty years of developments. Math. Program. Ser. B 169, 5–68 (2018)
11. Le Thi, H. A., Huynh, V. N., Pham Dinh, T.: Convergence analysis of difference-of-convex algorithm with subanalytic data. J. Optim. Theory Appl. 179, 103–126 (2018)
12. Pang, J. S., Razaviyayn, M., Alvarado, A.: Computing B-stationary points of nonsmooth DC programs. Math. Oper. Res. 42, 95–118 (2017)
13. Luo, Z. Q., Tseng, P.: Error bounds and convergence analysis of feasible descent methods: a general approach. Ann. Oper. Res. 46–47, 157–178 (1993)
14. Pang, J. S., Scutari, G. : Nonconvex games with side constraints. SIAM J. Optim. 21, 1491-1522 (2010)
15. Hiriart-Urruty, J. B.: From convex optimization to nonconvex optimization. Necessary and sufficient conditions for global optimality, pp. 219–239. Springer US, Boston, MA (1989)
16. Lu, Z. S., Zhou, Z. R., Sun, Z.: Enhanced proximal DC algorithms with extrapolation for a class of structured nonsmooth DC minimization. Math. Program. 176, 369–401 (2018)
17. Cui, Y., Pang, J. S., Sen, B.: Composite difference-max programs for modern statistical estimation problems. SIAM J. Optim. 28, 3344–3374 (2018)
18. Bolte, J., Daniilidis, A., Lewis, A.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM J. Optim. 17, 1205–1223 (2007)
19. Wen, B., Chen, X. J., Pong, T. K.: A proximal difference-of-convex algorithm with extrapolation. Comput. Optim. Appl. 69, 297–324 (2018)
20. Liu, T. X., Pong, T. K., Takeda, A.: A refined convergence analysis of pDCA$_e$ with applications to simultaneous sparse recovery and outlier detection. Comput. Optim. Appl. 73, 69–100 (2019)
21. Luo, Z. Q., Pang, J. S.: Error bounds for analytic systems and their application. Math. Program. 67, 1–28 (1994)
22. Luo, Z. Q., Sturm, J. F.: Error bounds for quadratic systems. In: Frenk, H., Roos, K., Terlaky, T., Zhang, S. (eds.): High Performance Optimization. Applied Optimization, vol 33. Springer, Boston, MA (1985)

23. Pang, J. S.: Error bounds in mathematical programming. Math. Program. 79, 299–332 (1997)
24. Tseng, P., Yun, S.: A coordinate gradient descent method for nonsmooth separable minimization. Math. Program. Ser. B 117, 387–423 (2009)
25. Wen, B., Chen, X., Pong, T. K.: Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconve nonsmooth minimization problems. SIAM J. Optim. 27, 124–145 (2017)
26. Drusvyatskiy, D., Lewis, A. S.: Error bounds, quadratic growth, and linear convergence of proximal methods. Math. Opera. Res. 43, 919–948 (2018)
27. Zhou, Z., So, A. M. C.: A unified approach to error bounds for structured convex optimization problems. Math. Program., Ser. A 165, 689–728 (2017)
28. Hiriart-Urruty, J. B., Lemaréchal, C.: Convex analysis and minimization algorithms. II. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 306, Springer-Verlag, Berlin (1993)
29. Rockafellar, R. T.: Convex Analysis. Princeton University Press (1970)
30. Eckstein, J., Bertsekas, D. P.: On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. 55, 293–318 (1992)
31. Hesse, R., Luke, D. R.: Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems. SIAM J. Optim. 23, 2397–2419 (2013)
32. Attouch, H. and Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Program. 116, 5–16 (2009)
33. Harker, P. T., Pang, J. S.: Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms and applications. Math. Program. 48, 161–220 (1990)
34. Luo, Z. Q., Tseng, P.: Error bound and convergence analysis of matrix splitting algorithms for the affine variational inequality problem. SIAM J. Optim. 2, 43–54 (1992)
35. Luo, Z. Q., Tseng, P.: On linear convergence of descent methods for convex essentially smooth minimization. SIAM J. Control Optim. 30, 408–425 (1992)
36. Bauschke, H. H., Borwein, J. M., Li, W.: Strong conical hull intersection property, bounded linear regularity, Jameson's property (G), and error bounds in convex optimization. Math. Program., Ser. A 86, 135–160 (1999)
37. Zhang, T., Bach, F.: Analysis of multi-stage convex relaxation for sparse regularization. Journal of Machine Learning Research 11, 1081–1107 (2010)
38. Zhang, C. H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Statist. 38, 894–942 (2010)
39. Li, G. Y., Pong, T. K.: Calculus of the exponent of Kurdyka–Łojasiewicz inequality and its applications to linear convergence of first-order methods. Found. Comput. Math. 18, 1199–1232 (2018)