manuscript No.
(will be inserted by the editor)

Error Bound and Isocost Imply Linear Convergence of DCA-based Algorithms to D-stationarity

Min Tao and Jiang-Ning Li

Received: date / Accepted: date

Abstract We consider a class of structured nonsmooth difference-of-convex minimization, which can be written as the difference of two convex functions possibly nonsmooth with the second one in the format of the maximum of a finite convex smooth functions. We propose two extrapolation proximal difference-of-convex based algorithms for potential acceleration to converge to a weak/standard d-stationary point of the structured nonsmooth problem, and prove its linear convergence of these algorithms under the assumptions of *piecewise* error bound and *piecewise* isocost condition. As a product, we refine the linear convergence analysis of sDCA and ε-DCA in a recent work of Dong and Tao (J. Optim. Theory Appl. 189, 190-220, 2021) by removing the assumption of locally linear regularity regarding the intersection of certain stationary sets and dominance regions. We also discuss sufficient conditions to guarantee these assumptions and illustrate that several sparse learning models satisfy all these assumptions. Finally, we conduct some elementary numerical simulations on sparse recovery to verify the theoretical results empirically.

Keywords Difference-of-convex programming · Difference-of-convex algorithm · linear convergence · error bound

Mathematics Subject Classification (2000) 90C30 · 90C26

1 Introduction

Minimization of structured nonsmooth difference-of-convex (DC) function [23] over a closed convex set has arisen from various applications such as statistical learning [1, 7], compressive sensing [8]. 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 the concave part is the minus of a finite max of convex functions. The first question raised is about *stationary* which plays the role of computational solution. There are several stationary notions for the structured nonsmooth DC programming: strongly critical point [24], A_ε-stationary [4] (equivalent to (α, η)-D-stationary [15]), d-stationary point [21], weak d-stationary point [21] and critical point [24]. Strong criticality is the strongest necessary condition for local/global optimality [4]. A_ε-stationary lies between the global optimum and the d-stationary set [4]. D-stationary point is arguably the sharpest kind among d-stationary point, weak d-stationary point and critical point [21]. It has been shown that d-stationarity is equivalent to strong criticality under some technique conditions [24].

The author was partially supported by National Key Research and Development Program of China (2018AAA0101100), the Natural Science Foundation of China (No. 11971228)

M. Tao
Department of Mathematics, National Key Laboratory for Novel Software Technology, Nanjing University, China.
E-mail: taom@nju.edu.cn

J.-N. Li
Department of Mathematics

We consider the following structured nonsmooth DC programming:

$$\min_x F(x) = \overbrace{H_s(x) + H_n(x)}^{H(x)} - \overbrace{\max_{1\le i\le m} g_i(x)}^{G(x)}. \tag{1}$$

Denote $F_i(x) = H(x) - g_i(x)$ for each i. By exploiting the potential structure, we separate $H(\cdot)$ into the smooth part $H_s(\cdot)$ and the nonsmooth part $H_n(\cdot)$ and suppose the proximal operator associated with $H_n(\cdot)$ can be evaluated. Throughout this paper, we make the following assumptions for problem (1).

Assumption 1 *(a) H_n is a proper closed convex function; H_s and g_i $(i = 1,\dots,m)$ are proper closed convex and continuously differentiable functions whose domains contain an open superset of $dom(H_n)$ and its Lipschitz continuous constant are L_s, L_i $(i = 1,\dots,m)$, respectively.*
(b) $H_n(\cdot)$ and $g_i(\cdot)$ $(i = 1,...,m)$ are all strongly convex with parameter $\sigma > 0$;
(c) For any $\alpha \in \mathbb{R}$, $\{x \in dom(F) \mid F(x) \le \alpha\}$ is bounded; $F(x)$ is bounded below on $dom(F)$, i.e., $\inf_x F(x) > -\infty$.

Assumption 1(b) is to facilitate the coming analysis which can be satisfied by adding simultaneously a quadratic term $\sigma\|x\|^2/2$ both to $H_n(\cdot)$ and all $g_i(\cdot)$ $(i = 1,...,m)$. Thus, $L_i \ge \sigma$ for all i. We call the triplet $(H_n, H_s, \{g_i\}_{i=1}^m)$ satisfying Assumption 1 with scalars $(\sigma,\ L_s,\ \{L_i\}_{i=1}^m)$ and denote $\hat{L} = \max_{i=1}^m L_i$. Given any $i \in [n]$ ($[n] = \{1, 2, \cdots, n\}$), we introduce the function

$$\ell_i(x, y, z) = H_s(y) + \langle \nabla H_s(y), x - y\rangle + H_n(x) - g_i(z) - \langle \nabla g_i(z), x - z\rangle. \tag{2}$$

A proximal projection step by employing the structure of H in (1) for each F_i generates the following mapping:

$$\mathcal{T}^{(i)}(x) = \operatorname*{argmin}_y \Big\{\ell_i(y, x, x) + \frac{L_s}{2}\|x - y\|^2\Big\}.$$

To prove linear convergence, we need extra assumptions: *piecewise* error bound and *piecewise* isocost. For each F_i, we define its stationary set and and its dominance region

$$\Omega^{(i)} = \{x : \nabla g_i(x) \in \partial H_n(x) + \nabla H_s(x)\},\ \mathcal{D}^{(i)} = \{x : g_i(x) \ge G(x)\}. \tag{3}$$

Assumption 2 *Let $\Omega^{(i)}$ be defined in (3) and $\mathcal{I} = \{i \ :\ \Omega^{(i)} \ne \emptyset\}$. For each $i \in \mathcal{I}$, the following two conditions hold:*

(a) For any $\zeta \ge \inf_x F(x)$, there exist $\tau, \varepsilon > 0$ such that

$$dist(x, \Omega^{(i)}) \le \tau\|x - \mathcal{T}^{(i)}(x)\| \quad whenever\ \ F(x) \le \zeta,\ \ \|x - \mathcal{T}^{(i)}(x)\| \le \varepsilon.$$

(b) There exists a positive constant $\mu > 0$ such that for any $x_1, x_2 \in \Omega^{(i)}$, it holds that $\|x_1 - x_2\| \ge \mu$ whenever $F_i(x_1) \ne F_i(x_2)$.

In fact, Assumption 2(a), 2(b) are equivalent to [27, Assumption 3.1] applied to the problem of $\min_x F_i(x)$.

The standard DCA [22] applied to problem (1) generates the new iterate x^{k+1} by

$$x^{k+1} \leftarrow \operatorname*{argmin}_x \{H(x) - \langle \xi^k, x\rangle\}, \tag{4}$$

where $\xi^k \in \partial G(x^k)$. The standard DCA converges to a critical point [22]. The sDCA [4] aims to compute a weak d-stationary point and updates the new iterate by choosing the subgradient $\xi^k \in \partial G(x^k)$ from the "active piece". Pang, Razaviyayn and Alvarado [21] proposed a basic algorithm [21, Algorithm 1] (PRA) and generates the iterate as

$$\begin{cases} x^{k,i} \leftarrow \operatorname{argmin}_x \{H(x) - \langle \nabla g_i(x^k), x\rangle + \frac{1}{2}\|x - x^k\|^2\} \ \text{for each } i \in \mathcal{M}_\varepsilon(x^k), \\ \hat{i} \in \operatorname{argmin}_{i\in\mathcal{M}_\varepsilon(x^k)}\{F(x^{k,i}) + \frac{1}{2}\|x^{k,i} - x^k\|^2\}, \\ x^{k+1} = x^{k,\hat{i}}, \end{cases}$$

where $\mathcal{M}_\varepsilon(x) = \{i \ : \ g_i(x) \geq G(x) - \varepsilon\}$ for some $\varepsilon > 0$. The PRA was shown subsequently to a d-stationary point [21]. Dong and Tao proposed ε-DCA [4] and it generates the new iterate by

$$\begin{cases} \hat{x}^{k,i} \leftarrow \underset{x}{\operatorname{argmin}} \left\{H(x) - \langle \nabla g_i(x^k), x\rangle\right\}, \ \ i \in \mathcal{M}_\varepsilon(x^k), & \text{(5a)} \\ \hat{i} \in \underset{i\in\mathcal{M}_\varepsilon(x^k)}{\operatorname{argmin}} \{F(\hat{x}^{k,i}) + \frac{\sigma}{2}\|\hat{x}^{k,i} - x^k\|^2\}, & \text{(5b)} \\ x^{k+1} = \hat{x}^{k,\hat{i}}. & \text{(5c)} \end{cases}$$

Any accumulation point of the sequence generated by ε-DCA is an $A_{\varepsilon'}$-stationary point ($0 < \varepsilon' < \varepsilon$) which is stronger than d-stationary [4, Section 5.1].

The proximal DCA (PDCA), its nonmonotone line search version and the proximal DCA with extrapolation (PDCA$_e$) were developed in [8,28]. In particular, PDCA$_e$ [28] computes the new iterate by

$$\begin{cases} z^k = x^k + \beta_k(x^k - x^{k-1}), \\ x^{k+1} = \operatorname{argmin}_x\{H_n(x) + \langle x, \nabla H_s(z^k) - \xi^k\rangle + \frac{L_s}{2}\|x - z^k\|^2\}, \end{cases}$$

where $\xi^k \in \partial G(x^k)$ and $\beta_k \subseteq [0, 1)$ with $\sup_k \beta_k < 1$. The extra extrapolation step is for the potential accelerating. An enhanced PDCA with extrapolation (EPDCA$_e$) for computing a d-stationary point of (1) was proposed in [15], i.e.,

$$\begin{cases} z^k = x^k + \beta_k(x^k - x^{k-1}), \ \beta_k \in [0, \sqrt{c/L_s}), & \text{(6a)} \\ \quad \text{for each } i \in \mathcal{M}_\varepsilon(x^k) & \text{(6b)} \\ x^{k,i} = \underset{x}{\operatorname{argmin}}\{H_n(x) + \langle x, \nabla H_s(z^k) - \nabla g_i(x^k)\rangle + \frac{L_s}{2}\|x - z^k\|^2 + \frac{c}{2}\|x - x^k\|^2\}, & \text{(6c)} \\ \hat{i} = \underset{i\in\mathcal{M}_\varepsilon}{\operatorname{argmin}}\{F(x^{k,i}) + \frac{c}{2}\|x^{k,i} - x^k\|^2\}; \ \ x^{k+1} = x^{k,\hat{i}}, & \text{(6d)} \end{cases}$$

and $0 < c < L_s$. Furthermore, a nonmonotone enhanced PDCA with extrapolation (nEPDCA$_e$) was proposed in [16] by replacing (6c) with

$$x^{k,i} = \underset{x}{\operatorname{argmin}}\{H_n(x) + \langle x, \nabla H_s(z^k) - \nabla g_i(x^k)\rangle + \frac{L_s}{2}\|x - z^k\|^2\}.$$

The advantage of nEPDCA$_e$ over EPDCA$_e$ is to permit a larger step size in solving the subproblem of $x^{k,i}$. Any accumulation point of the sequence generated by the PDCA (PDCA$_e$) is a critical point [28] while any accumulation point of EPDCA or nEPDCA$_e$ is a d-stationary point [15,16]

To establish sequential convergence analysis, one common way is to exploit the Kurdyka-Łojasiewicz (KL) property [2]. Le Thi, Huynh and Pham Dinh established the sequential convergence result of DCA under the assumptions that one of ∇H and ∇G in (1) is Lipschitz continuous and the objective function with the KL property [24]. Wen, Chen and Pong established the sequential convergence of PDCA$_e$ under the Lipschitz continuousness of $\nabla G(\cdot)$ and an auxiliary function with the KL property [28]. Liu, Pong and Takeda proved the sequential convergence of PDCA$_e$ by removing the Lipschitz continuousness of $\nabla G(\cdot)$ [12]. All these convergence results [3, 24,28,12] are shown to a critical point instead of a weak/standard d-stationary solution.

For converging to a d-stationary solution, Lu, Zhou and Sun [15] established the sequential convergence to a d-stationary point by assuming one of these conditions holds: (1) One of the elements of the accumulation set is isolated; (2) A certain merit function is with the KL property and for each accumulation point $\bar{x}$, $|\mathcal{M}(\bar{x})| = 1$ where $|\cdot|$ denotes the number of the indices and the parameter ε in (6b) satisfies $0 < \varepsilon < \bar{\varepsilon}$ where $\bar{\varepsilon} = \frac{1}{2}(G(\bar{x}) - \max_{i\in\mathcal{M}^o(\bar{x})} \nabla g_i(\bar{x}))$ and $(\cdot)^o$ denotes its complementary set. The concepts of critical point, weak d-stationary and d-stationary are identical when $|\mathcal{M}(\bar{x})| = 1$ [4, Lemma 3.2]. The sequential convergence to a d-stationary point still remains open when $|\mathcal{M}(\bar{x})| \neq 1$.

Another approach to prove sequential (linear) convergence is by using the error bound condition; see some seminal work [17,18,19,25,27] and references therein. The linear convergence of sDCA and ε-DCA are established in [4] under the assumptions of the generalized piecewise error bound and piecewise isocost, as well as locally linear regularity condition (LLR) regarding the intersection of certain stationary sets and dominance regions. More specifically, the piecewise error bound condition is indeed "Luo-Tseng" error bound condition which is commonly used together with isocost condition (concerning separation of stationary values) for establishing local linear convergence of various first-order algorithms [17,18,19]. The sufficient conditions [4, Proposition 6.6] to guarantee the LLR condition regarding the intersection of certain sets are not easy to be satisfied since the stationary set of F_i might not be convex. Therefore, *a very natural question is whether we can still establish the linear convergence of sDCA/ε-DCA only under piecewise error bound and piecewise isocost condition without the LLR condition.* In this paper, we answer this question affirmatively.

Note that either sDCA or ε-DCA has a favorite convergence guarantee than DCA in terms of solution quality. However, its subproblem might not be easy to solve and ε-DCA usually converges slowly due to solving multiple subproblems. We propose two proximal DCA with extrapolation step for potential acceleration: one is a specific version of proximal DCA with extrapolation (sPDCA$_e$) to compute a weak d-stationary point, while the other is a proximal ε-DCA with extrapolation (PEDCA$_e$) for computing a d-stationarity point. The latter is equivalent to nEPDCA$_e$ [16, Algorithm 3]. Although the subsequential convergence of nEPDCA$_e$ to a d-stationary point was established in [16], the sequential convergence is absent. In this paper, we show that every accumulation point of the sequence generated by PEDCA$_e$ is an $A_{\varepsilon'}$-stationary point ($0 < \varepsilon' < \varepsilon$, ε defined in (6b)) which is stronger than d-stationary. Based on this observation, we establish the linear convergence rate both for sPDCA$_e$ and PEDCA$_e$ under piecewise error bound and piecewise isocost. As a corollary, we obtain the linear convergence of sDCA and ε-DCA also under the same conditions without the LLR condition.

The rest of this paper proceeds as follows. In Section 2, we describe the notations and definitions. In section 3, we consider sPDCA$_e$ and show it converges to a weak d-stationary point. We prove linear convergence (and sequential convergence) under piecewise error bound and piecewise isocost. In section 4, we consider PEDCA$_e$ for computing a d-stationary point and prove it converges to a $A_{\varepsilon'}$-stationarity ($0 < \varepsilon' < \varepsilon$). Then, we prove the linear convergence (and sequential convergence) under piecewise error bound and piecewise isocost. We further discuss on sufficient conditions for ensuring these two key assumptions in Section 5. Especially, we show that several statistical estimation and sparse recovery models satisfy all these key assumptions in Section 6. Section 7 is devoted to some elementary numerical experiments to showcase the superior performance of PEDCA$_e$ in sparse recovery problems. In Section 8, we draw some conclusions.

2 Preliminary

Let $\mathbb{R}^n$ denote the n-dimensional Euclidean space, $\langle\cdot,\cdot\rangle$ denote the standard inner product, and $\|\cdot\|$ denote the Euclidean 2-norm. We define $[n] = \{1, 2, \cdots, n\}$ as an index set up to n. Given a matrix $A \in \mathbb{R}^{n\times n}$, we use $\lambda_{\max}(A)$ to denote its maximum eigenvalue. An extended-real-valued function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is said to be proper if its domain $\mathrm{dom} f = \{x \ : \ f(x) < \infty\}$ is nonempty. For any $\delta > 0$, $\mathcal{B}(x, \delta)$ is the open neighborhood $\{y : \|y - x\| < \delta\}$. A proper function f is said to be closed if it is lower semi-continuous. For a proper closed function f and $\hat{x} \in \mathrm{dom} f$, the regular subdifferential $\hat{\partial} f(\hat{x})$ and the limiting subdifferential $\partial f(x)$ [31] are defined as

$$\hat{\partial} f(\hat{x}) = \left\{ v : \lim_{x\neq\hat{x}, x\to\hat{x}} \frac{f(x) - f(\hat{x}) - \langle v, x - \hat{x}\rangle}{\|x - \hat{x}\|} \geq 0 \right\},$$

$$\partial f(\hat{x}) = \left\{ v : \ \exists\, x^k \to \hat{x},\ f(x^k) \to f(\hat{x}),\ v^k \in \hat{\partial} f(x^k) \text{ with } v^k \to v \right\},$$

respectively. The directional derivative of f at a point $x \in \mathrm{dom}(f)$ along a direction $d \in \mathbb{R}^n$ is defined as $f'(x; d) = \lim_{\tau\downarrow 0} \frac{f(x+\tau d) - f(x)}{\tau}$. 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 $f(x) \geq f(\bar{x}) + y^\top(x - \bar{x}) - \varepsilon, \quad \forall x \in \mathbb{R}^n$.

The set of all ε-subgradients of f at $\bar{x}$ is denoted as $\partial_\varepsilon f(\bar{x})$. When $\varepsilon = 0$, $\partial_\varepsilon f$ is the usual subdifferential. For a smooth function g, we use ∇g to represent its gradient mapping. Suppose f be a lower semicontinuous function, we define the proximal mapping [31, Definition 1.22]:

$$\mathrm{Prox}_f(v) = \arg\min_x \Big\{ f(x) + \frac{1}{2}\|x - v\|^2 \Big\}.$$

Given a scalar x, we define the sign function: $\mathrm{sign}(x) = 1$ if $x \geq 0$ and otherwise -1. In what follows, we recall several different types of stationarity.

Definition 2.1 A point x is said to be a critical point of (1) if $\partial H(x) \cap \partial G(x) \neq \phi$.

Definition 2.2 A point x is said to be a strong criticality of (1) if and only if $\emptyset \neq \partial G(x) \subseteq \partial H(x)$.

Definition 2.3 A point x is called d-stationary to (1) if $F'(x; y - x) \geq 0,\ \forall y \in \mathrm{dom}(F)$.

As shown in [21,4], another equivalent definition of d-stationary is presented below.

Definition 2.4 x is called a d-stationary point of (1) if and only if for any $i \in \mathcal{M}(x)$ such that $\nabla g_i(x) \in \partial H_n(x) + \nabla H_s(x)$.

It is easy to see that x is a d-stationary if and only if x is a strong criticality when $x \in \mathrm{ri}(\mathrm{dom} H_n)$ [24, Theorem 3.1].

Definition 2.5 [21] x is called a weak d-stationary point of (1) if and only if there exists $i \in \mathcal{M}(x)$ such that $\nabla g_i(x) \in \partial H_n(x) + \nabla H_s(x)$.

D-stationary point is definitely weak d-stationary while the converse is not true. A much stronger optimality condition than d-stationarity, i.e., A_ε-stationary (for $\varepsilon > 0$) proposed in [4] and shown to be equivalent to the notion of (α, η)-D-stationary [15] (see Remark 5.1 of [4]).

Definition 2.6 x is called A_ε-stationary of (1) if and only if for any $i \in \mathcal{M}_\varepsilon(x)$, $\nabla g_i(x) \in \partial_{(G(x) - g_i(x))}(H_n + H_s)(x)$.

An A_ε-stationary point must be a d-stationarity, but the converse generally does not hold (see [4, *Counterexample* (ii)]). We refer to [4, Fig. 1] for the relationship between these different types of stationary, including global optimality condition.

Next, we recall the concept of *subregularity* [31] and two types of error bounds.

Definition 2.7 A set-valued mapping (or a multi-function) $G : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is said to be *subregular* at $(\bar{x}, \bar{y})$ with parameter $\ell > 0$ if $\bar{y} \in G(\bar{x})$ and there exists a neighborhood $\mathcal{N}$ of $\bar{x}$, such that

$$\mathrm{dist}(x, G^{-1}(\bar{y})) \leq \ell\, \mathrm{dist}(\bar{y}, G(x)), \qquad \forall x \in \mathcal{N}\,.$$

Next, we consider the DC program:

$$\min_x \ f(x) = H_n(x) + H_s(x) - g(x). \tag{7}$$

The set of stationary points of f is defined by

$$\Omega = \{x \ : \ \nabla g(x) \in \partial H_n(x) + \nabla H_s(x)\}\,. \tag{8}$$

Suppose $\Omega \neq \emptyset$. We say that the "Luo-Tseng" error bound holds if for any $\zeta \geq \inf f$, there exist $c,\ \varepsilon > 0$ such that $\mathrm{dist}(x, \Omega) \leq c \| \mathrm{Prox}_{H_n}(x - (\nabla H_s(x) - \nabla g(x))) - x\|$ whenever $\zeta \geq f(x)$ and $\| \mathrm{Prox}_{H_n}(x - (\nabla H_s(x) - \nabla g(x))) - x\| \leq \varepsilon$; see, e.g., [17,18,19,14,27,25,30]. In [5], an error bound holds around $\bar{x}$ if there exists an open neighborhood $\mathcal{N}$ of $\bar{x}$, such that

$$\mathrm{dist}(x, \Omega) \leq c \left\| x - \mathrm{Prox}_{\rho^{-1} H_n}(x - \rho^{-1}(\nabla H_s(x) - \nabla g(x))) \right\|, \quad \forall x \in \mathcal{N},$$

where $\rho > 0$ is a fixed parameter. In Section 5, we will show these two types of error bound are equivalent under certain conditions and both of them are equivalent to the subregularity of the set-valued mapping $\partial(H_n + H_s - g)(x)$.

3 Linear Convergence of sPDCA$_e$ to a Weak D-Stationary

By further exploring the particular structure of H in (1), we propose a specific version of the proximal DCA with extrapolation (sPDCA$_e$) for the potential acceleration. The subgradient ξ^k in the proximal DCA-step is chosen to be one of the "active gradients". The advantage of sPDCA$_e$ over sDCA [4, Algorithm 1] lies in two aspects: (1) Introducing one extrapolation step ($\beta_k \in \left[0, \sqrt{\frac{\sigma}{L_s}}\right)$) for potential acceleration; (2) The new iterate is updated via minimizing a tight quadratic upper bound function F_{i_k}, i.e.,

$$\begin{aligned}\widehat{F}_{i_k}(x, z^k, x^k) = H_n(x) + H_s(z^k) + \langle x - z^k, \nabla H_s(z^k)\rangle - g_{i_k}(x^k) \\ - \langle x - x^k, \nabla g_{i_k}(x^k)\rangle + \frac{L_s}{2}\|x - z^k\|^2. \end{aligned} \tag{9}$$

Such upper-bound functions often leads to much simpler subproblems than solving the original function; see, e.g., [32,16]. Note that the function $\widehat{F}_{i_k}(x, x^k, x^k)$ (setting $z^k = x^k$ in the above upper bound) satisfies

(i) $\widehat{F}_{i_k}(x, x^k, x^k)\,|_{x=x^k} = F_{i_k}(x^k)$;
(ii) $\partial_x \widehat{F}_{i_k}(x, x^k, x^k)\,|_{x=x^k} = \partial F_{i_k}(x^k)$;
(iii) $\widehat{F}_{i_k}(x, x^k, x^k) \geq F_{i_k}(x)$ for any x.

Recalling the definition of $\ell_i(x, y, z)$ in (2), $\widehat{F}_{i_k}(x, z^k, x^k) = \ell_{i_k}(x, z^k, x^k) + \frac{L_s}{2}\|x - z^k\|^2$. We summarize the details of sPDCA$_e$ in Algorithm 1.

Algorithm 1 sPDCA$_e$

Initialization: choose $x^0 \in \mathrm{dom}(F)$, $\varepsilon > 0$ and $\{\beta_k\}_k \subseteq [0, \sqrt{\frac{\sigma}{L_s}})$ with $\bar{\beta} = \sup_k \beta_k < \sqrt{\frac{\sigma}{L_s}}$. Set $x^{-1} = x^0$.

for $k = 0, 1, 2, \ldots$ **do**

Set

$$z^k = x^k + \beta_k(x^k - x^{k-1}). \tag{10}$$

Choose an index $i_k \in \mathcal{M}(x^k)$, then compute x^{k+1} by

$$x^{k+1} \longleftarrow \operatorname*{argmin}_x \left[\ell_{i_k}(x, z^k, x^k) + \frac{L_s}{2}\|x - z^k\|^2\right]. \tag{11}$$

end for

As a preparation for our convergence rate analysis, we first show that this algorithm converges to a weak d-stationary point. Next, we prove our main result in this section, i.e., (locally) linear convergence of this algorithm under Assumption 2. Finally, since sDCA is a special case of sPDCA$_e$, the linear convergence result of this algorithm can be established under Assumption 2.

Next, we introduce an auxiliary function:

$$\Delta(x, y, \beta) = F(x) + \frac{\beta^2 L_s}{2}\|y - x\|^2. \tag{12}$$

The following lemma shows the sufficient descent property of the merit function $\Delta(x^{k+1}, x^k, \beta_{k+1})$.

Lemma 3.1 *Suppose Assumption 1 holds. Let $\{x^k\}$ be the sequence generated by sPDCA$_e$. For each k,*

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

Proof Recall that i_k is the active index chosen in step k of sPDCA$_e$, then

$$\begin{aligned}
&H_s(x^k) + H_n(x^k) + \frac{L_s}{2}\|x^k - z^k\|^2 \\
&\geq H_s(z^k) + \langle x^k - z^k, \nabla H_s(z^k)\rangle + \frac{L_s}{2}\|x^k - z^k\|^2 + H_n(x^k) \\
&\geq H_s(z^k) + \langle x^{k+1} - z^k, \nabla H_s(z^k)\rangle + \frac{L_s}{2}\|x^{k+1} - z^k\|^2 + H_n(x^{k+1}) \\
&\quad -\langle x^{k+1} - x^k, \nabla g_{i_k}(x^k)\rangle + \frac{L_s}{2}\|x^{k+1} - x^k\|^2 \\
&\geq H_s(x^{k+1}) + H_n(x^{k+1}) - g_{i_k}(x^{k+1}) + g_{i_k}(x^k) + \frac{L_s + \sigma}{2}\|x^{k+1} - x^k\|^2.
\end{aligned}$$

The first inequality is due to the convexity of H_s, the second is due to (11) and the strong convexity of the objective function of (11) with the modulus L_s, the last is due to L_s-smoothness of H_s and the σ-strong convexity of g_{i_k}. Consequently, subtracting both sides of the above inequality by $g_{i_k}(x^k)$ and using (10), it yields that

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

Then, the assertion follows from $F(x^k) = F_{i_k}(x^k)$ and $F_{i_k}(x^{k+1}) \geq F(x^{k+1})$. □

From Lemma 3.1, we have

$$\Delta(x^{k+1}, x^k, \beta_{k+1}) \leq \Delta(x^k, x^{k-1}, \beta_k) - \frac{1}{2}(\sigma + L_s - \beta_{k+1}^2 L_s)\|x^{k+1} - x^k\|^2. \tag{13}$$

It implies that the sequence $\{\Delta(x^k, x^{k-1}, \beta_k)\}$ is monotonic decreasing whenever $\bar{\beta} = \sup_k \beta_k < \sqrt{\frac{\sigma}{L_s}}$. Moreover, it follows directly from (13) that

$$\|x^k - x^{k+1}\|^2 \leq \frac{2}{\sigma + L_s - (\bar{\beta})^2 L_s}\left(\Delta(x^k, x^{k-1}, \beta_k) - \Delta(x^{k+1}, x^k, \beta_{k+1})\right).$$

Equipped with this lemma, subsequential convergence of sPDCA$_e$ follows directly. The proof of the assertion (i) of Theorem 3.1 are standard in the literature [24], the assertion (ii), (iv) are similar to [4, Theorem 4.1(iv),(vi)] and thus we omit them. We only present the proof for (iii) here.

Theorem 3.1 *Suppose Assumption 1 holds. Let the sequence $\{x^k\}$ be generated by sPDCA$_e$. Then, the following statements holds:*

(i) $\sum_{k=1}^{\infty}\|x^k - x^{k+1}\|^2 < +\infty$ *(which implies* $\lim_{k\to+\infty}\|x^k - x^{k+1}\| = 0$*); The sequence $\{F(x^k)\}$ is convergent, and denote $F^* = \lim_{k\to+\infty} F(x^k)$; The sequence $\{x^k\}$ is bounded;*

(ii) Any accumulation point of $\{x^k\}$ is a weak d-stationary point;

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

(iv) Suppose that one of the elements in the accumulation set of $\{x^k\}$ is isolated. Then, the whole sequence $\{x^k\}$ converges.

Proof (iii) Let x^∞ be an accumulation point of $\{x^k\}$. There is a subsequence $\{x^{k_j}\}_{j=1}^{\infty}$ converging to x^∞. Since $\{i_{k_j}\}_{j=1}^{\infty}$ have finite choices from $\{1, \cdots, m\}$, there is subsequence of $\{i_{k_j}\}_{j\in\kappa} \subseteq \{i_{k_j}\}_{j=1}^{\infty}$ such that $i_{k_j} \equiv \bar{i}$ when $j \in \kappa$. Without loss of generality, we can assume that $\{x^{k_j}\}_{j=1}^{\infty}$ be a subsequence converging to x^∞ and $i_{k_j} \equiv \bar{i}$. Invoking (11), we have

$$\begin{aligned}
&H_s(x^{k_j+1}) + H_n(x^{k_j+1}) - g_{\bar{i}}(x^{k_j+1}) \leq H_s(z^{k_j}) + \langle \nabla H_s(z^{k_j}), x^{k_j+1} - z^{k_j}\rangle \\
&+ H_n(x^{k_j+1}) - g_{\bar{i}}(x^{k_j}) - \langle \nabla g_{\bar{i}}(x^{k_j}), x^{k_j+1} - x^{k_j}\rangle + \frac{L_s}{2}\|x^{k_j+1} - z^{k_j}\|^2 \\
&\leq H_s(z^{k_j}) + \langle \nabla H_s(z^{k_j}), x^\infty - z^{k_j}\rangle + H_n(x^\infty) - g_{\bar{i}}(x^{k_j}) \\
&- \langle \nabla g_{\bar{i}}(x^{k_j}), x^\infty - x^{k_j}\rangle + \frac{L_s}{2}\|x^\infty - z^{k_j}\|^2.
\end{aligned}$$

By taking limit on both sides of the above inequality, we have that $F^* \le F_{\bar{i}}(x^\infty) = F(x^\infty)$ due to $\bar{i} \in \mathcal{M}(x^\infty)$. Since F is lower semicontinuous, we have

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

Thus, $F(x^\infty) = F^*$. Since x^∞ is arbitrary, the assertion (iii) follows immediately. □

For convenience of analysis, we introduce the following auxiliary vector $\bar{x}^k$:

$$\bar{x}^k \in \arg\min_x \left\{ \|x - x^k\| \ : \ x \in \Omega^{(i_k)} \right\}, \tag{14}$$

provided that $\Omega^{(i_k)}$ is nonempty and i_k is defined in (11). Obviously, $\Omega^{(i_k)}$ is a closed set and nonempty (see Lemma 3.2). Therefore, the vector of $\bar{x}^k$ in (14) is well-defined when k sufficiently large. The proof of the following lemma is similar to [4, Lemma 4.3] and thus omitted here.

Lemma 3.2 *Suppose Assumption 1 holds. Let $\{x^k\}$ be the sequence generated by sPDCAe and denote $\mathcal{C}$ the accumulation set of $\{x^k\}$. Then there exists K such that for all $k \ge K$, $\mathcal{C} \cap \Omega^{(i_k)} \cap \mathcal{D}^{(i_k)}$ is nonempty.*

Remark 3.1 From Lemma 3.2, we see that $\Omega^{(i_k)}$ is nonempty when k is sufficiently large.

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

Lemma 3.3 *Suppose Assumption 1 holds. Let $\{x^k\}$ be the sequence generated by sPDCA$_e$, and let i_k be the index chosen in step k and $\bar{x}^k$ defined in (14). If Assumption 2(b) holds, then for all k sufficiently large, $F_{i_k}(\bar{x}^k) = F^*$.*

Proof First, we use the contradiction to show that $\lim_{k\to+\infty} \mathrm{dist}(x^k, \mathcal{C} \cap \Omega^{(i_k)} \cap \mathcal{D}^{(i_k)}) = 0$. Suppose not. There exists a subsequence $x^{k_j} \to x^\infty$ as $j \to \infty$ and $\varepsilon > 0$ such that $\mathrm{dist}(x^{k_j}, \mathcal{C} \cap \Omega^{(i_{k_j})} \cap \mathcal{D}^{(i_{k_j})}) \ge \varepsilon$ for all j. Without loss of generality, $i_{k_j} \equiv \bar{i}$ for $j \in \kappa$ where κ denoted a subsequence of k_j. Then, we have $x^\infty \in \mathcal{C} \cap \Omega^{(\bar{i})} \cap \mathcal{D}^{(\bar{i})}$. This contradicts $\mathrm{dist}(x^{k_j}, \mathcal{C} \cap \Omega^{(i_{k_j})} \cap \mathcal{D}^{(i_{k_j})}) \ge \varepsilon$. Thus, we have that $\lim_{k\to+\infty} \mathrm{dist}(x^k, \mathcal{C} \cap \Omega^{(i_k)} \cap \mathcal{D}^{(i_k)}) = 0$. Furthermore, when k is sufficiently large, we have $\mathrm{dist}(x^k, \mathcal{C} \cap \Omega^{(i_k)} \cap \mathcal{D}^{(i_k)}) < \mu/2$ for the μ given in Assumption 2(b). By the definition of $\bar{x}^k$ (see (14)), we have $\|x^k - \bar{x}^k\| = \mathrm{dist}(x^k, \Omega^{(i_k)}) \le \mathrm{dist}(x^k, \mathcal{C} \cap \Omega^{(i_k)} \cap \mathcal{D}^{(i_k)}) < \mu/2$. Then, let $\hat{x}^k \in \arg\min_x \left\{ \|x - x^k\| \ : \ x \in \mathcal{C} \cap \Omega^{(i_k)} \cap \mathcal{D}^{(i_k)} \right\}$. Hence, it yields that $\|\hat{x}^k - \bar{x}^k\| \le \|x^k - \hat{x}^k\| + \|x^k - \bar{x}^k\| < \mu$. Since $\hat{x}^k \in \mathcal{C} \cap \Omega^{(i_k)} \cap \mathcal{D}^{(i_k)}$, then $F_{i_k}(\hat{x}^k) = F(\hat{x}^k)$ due to $\hat{x}^k \in \mathcal{D}^{(i_k)}$. Next, $F(\hat{x}^k) = F^*$ since $\hat{x}^k \in \mathcal{C}$. Finally, Assumption 2(b) implies that $F_{i_k}(\bar{x}^k) = F_{i_k}(\hat{x}^k) = F^*$. □

Lemma 3.4 *Let $\{x^k\}$ be a sequence generated by sPDCA$_e$. Suppose that Assumptions 1 and 2(b) hold. Then for all k sufficiently large,*

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

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

Proof Recall that i_k is the active index chosen in step k of sPDCA$_e$, then

$$\begin{aligned}
&H_s(x) + \frac{L_s}{2}\|x - z^k\|^2 + H_n(x) - g_{i_k}(x) \\
&\ge H_s(z^k) + \langle x - z^k, \nabla H_s(z^k)\rangle + \frac{L_s}{2}\|x - z^k\|^2 + H_n(x) - \langle x - x^k, \nabla g_{i_k}(x^k)\rangle - \frac{L_{i_k}}{2}\|x - x^k\|^2 - g_{i_k}(x^k) \\
&\ge H_s(z^k) + \langle x^{k+1} - z^k, \nabla H_s(z^k)\rangle + \frac{L_s}{2}\|x^{k+1} - z^k\|^2 \\
&+ H_n(x^{k+1}) - \langle x^{k+1} - x^k, \nabla g_{i_k}(x^k)\rangle - \frac{L_{i_k}}{2}\|x - x^k\|^2 - g_{i_k}(x^k).
\end{aligned} \tag{15}$$

Furthermore, invoking that g_{i_k} is σ-strongly convex, then

$$-\langle x^{k+1}-x^k, \nabla g_{i_k}(x^k)\rangle - g_{i_k}(x^k) \geq -g_{i_k}(x^{k+1}) + \frac{\sigma}{2}\|x^{k+1}-x^k\|^2.$$

Using Lipschitz continuity of ∇H_s, one has

$$H_s(z^k) + \langle x^{k+1}-z^k, \nabla H_s(z^k)\rangle + \frac{L_s}{2}\|x^{k+1}-z^k\|^2 \geq H_s(x^{k+1}).$$

Substituting above two inequalities into (15), one further has

$$\begin{aligned} &H_s(x) + \frac{L_s}{2}\|x-z^k\|^2 + H_n(x) - g_{i_k}(x) \\ &\geq H_s(x^{k+1}) + H_n(x^{k+1}) - g_{i_k}(x^{k+1}) - \frac{L_{i_k}}{2}\|x-x^k\|^2 + \frac{\sigma}{2}\|x^{k+1}-x^k\|^2. \end{aligned}$$

By setting $x=\bar{x}^k$ in the above inequality and using Lemma 3.3, we obtain that

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

□

Now, we are ready to prove linear convergence of sPDCA$_e$.

Theorem 3.2 *Suppose Assumptions 1 and 2 hold. Let the sequence $\{x^k\}$ be generated by sPDCA$_e$. Then,*

(i) the sequence of $\{\Delta^k\}$ converges R-linearly to zero where $\Delta^{k+1} = \Delta(x^{k+1}, x^k, \beta_{k+1}) - F^*$;
(ii) the sequence $\{x^k\}$ converges R-linearly to a weak d-stationary point x^∞;
(iii) the sequence of objective function values $\{F(x^k)\}$ converges to F^* R-linearly.

Proof In the following, we divide into two cases to verify.

Case (I): $\beta_k \not\equiv 0$.
(i) First, invoking (13), one has

$$\Delta(x^{k+1}, x^k, \beta_{k+1}) \leq \Delta(x^k, x^{k-1}, \beta_k) - \frac{(L_s + \sigma - \beta_{k+1}^2 L_s)}{2}\|x^{k+1}-x^k\|^2.$$

Second, invoking Lemma 3.4 and recalling $\beta_{k+1}^2 < \frac{\sigma}{L_s}$, one obtain that

$$\Delta(x^{k+1}, x^k, \beta_{k+1}) \leq F^* + \frac{L_s}{2}\|\bar{x}^k - z^k\|^2 + \frac{L_{i_k}}{2}\|\bar{x}^k - x^k\|^2. \tag{16}$$

Next, consider the second term in the right hand side of the above inequality, one gets

$$\frac{L_s}{2}\|\bar{x}^k - z^k\|^2 \leq L_s(\|\bar{x}^k - x^k\|^2 + \|x^k - z^k\|^2).$$

Furthermore, using (10), it leads to

$$\frac{L_s}{2}\|\bar{x}^k - z^k\|^2 \leq L_s(\|\bar{x}^k - x^k\|^2 + \beta_k^2\|x^k - x^{k-1}\|^2). \tag{17}$$

Then, combining (16) with the above inequality, one gets

$$\Delta(x^{k+1}, x^k, \beta_{k+1}) \leq F^* + \left(L_s + \frac{\hat{L}}{2}\right)\|\bar{x}^k - x^k\|^2 + L_s\beta_k^2\|x^k - x^{k-1}\|^2, \tag{18}$$

where $\hat{L} = \max_{i=1}^m L_i$. Moreover, by the assertion (i) of Theorem 3.1, we have $\lim_{k\to+\infty}\|x^k - x^{k+1}\| = 0$. By considering the second term in the above inequality, we have

$$\|x^k - \bar{x}^k\| = \mathrm{dist}(x^k, \Omega^{(i_k)}) \leq \tau\|x^k - \mathcal{T}^{(i_k)}(x^k)\|, \tag{19}$$

where the inequality follows from the (i_k)th error bound in Assumption 2(a). Then, by using

$$\mathcal{T}^{(i_k)}(x^k) = \mathrm{Prox}_{(H_s/L_s)}\left[x^k - (L_s)^{-1}(\nabla H_s(x^k) - \nabla g_{i_k}(x^k))\right],$$

and

$$x^{k+1} = \mathrm{Prox}_{(H_s/L_s)}\left[z^k - (L_s)^{-1}(\nabla H_s(z^k) - \nabla g_{i_k}(x^k))\right],$$

we have

$$\|x^{k+1} - \mathcal{T}^{(i_k)}(x^k)\| \le 2\|x^k - z^k\|,$$

which is due to the nonexpansivity of proximal operators and Lipschitz continuity of ∇H_s. Combining (19) with the above inequality, we have

$$\|x^k - \bar{x}^k\|^2 \le \tau^2\|x^k - \mathcal{T}^{(i_k)}(x^k)\|^2 \le 2\tau^2(\|x^k - x^{k+1}\|^2 + 4\|x^k - z^k\|^2). \tag{20}$$

Substituting (10) into the above inequality, it leads to

$$\|x^k - \bar{x}^k\|^2 \le 2\tau^2(\|x^k - x^{k+1}\|^2 + 4\beta_k^2\|x^k - x^{k-1}\|^2).$$

Therefore, substituting the above inequality into (18), it yields

$$\Delta(x^{k+1}, x^k, \beta_{k+1}) - F^* \le c_1\|x^k - x^{k+1}\|^2 + c_2\|x^k - x^{k-1}\|^2, \tag{21}$$

where $c_1 = 2\tau^2(L_s + \hat{L}/2)$, $c_2 = (8\tau^2 L_s + 4\tau^2\hat{L} + L_s)\bar{\beta}^2$. By substituting (13) and (13) with $k = k-1$ into (21), we obtain

$$\Delta^{k+1} \le \frac{2c_1}{\alpha}(\Delta^k - \Delta^{k+1}) + \frac{2c_2}{\alpha}(\Delta^{k-1} - \Delta^k).$$

where $\alpha = \sigma + L_s - \bar{\beta}^2 L_s > 0$. In addition, $\Delta^{k+1} \le \Delta^k$ for any k. It further leads to

$$(1 + \frac{2c_1}{\alpha})\Delta^{k+1} \le (\frac{2c_1}{\alpha} - \frac{2c_2}{\alpha})\Delta^k + \frac{2c_2}{\alpha}\Delta^{k-1}. \tag{22}$$

In the following, we further divide into two cases to verify: (a) $c_1 \ge c_2$; (b) $c_1 < c_2$.
Case (a): $c_1 \ge c_2$.

$$(1 + \frac{2c_1}{\alpha})\Delta^{k+1} \le \frac{2c_1}{\alpha}\Delta^{k-1},$$

due to $\Delta^k \le \Delta^{k-1}$. Thus, $\Delta^{k+1} \le \frac{2c_1}{\alpha+2c_1}\Delta^{k-1}$. By defining $x^{-1} = x^0$,

$$\Delta^k \le \begin{cases} \Delta^0 q_1^{2t}, & k = 2t; \\ \Delta^1 q_1^{2t}, & k = 2t+1, \end{cases} \tag{23}$$

where $q_1 = \sqrt{\frac{2c_1}{\alpha+2c_1}} \in (0,1)$.
Case (b): $c_1 < c_2$. It follows from (22), we have that

$$(1 + \frac{2c_1}{\alpha})\Delta^{k+1} + \frac{2c_2 - 2c_1}{\alpha}\Delta^k \le \frac{2c_2}{\alpha}\Delta^{k-1}.$$

By using $\Delta^{k+1} \le \Delta^k$, we get $\Delta^{k+1} \le \frac{2c_2}{\alpha+2c_2}\Delta^{k-1}$. Similarly to Case (a), the recursive relation of (23) holds with replacing q_1 with $q_2 = \sqrt{\frac{2c_2}{\alpha+2c_2}} \in (0,1)$. In both of Cases (a) and (b), we have that $\Delta^k \le \max(\Delta^0, \frac{\Delta^1}{q})q^k$ with $q = \max(q_1, q_2) \in (0,1)$. In summary, the sequence of $\{\Delta^k\}$ converges to zero R-linearly.
(ii) First, combining the assertion (i) of Theorem 3.1 and (13), one has

$$\lim_{k\to\infty} \Delta(x^{k+1}, x^k, \beta_{k+1}) = F^*.$$

Furthermore, invoking (13) and using the fact of $\Delta^{k+1} \ge 0$, we obtain that

$$\|x^k - x^{k+1}\| \le \sqrt{\frac{2}{\alpha}(\Delta^k - \Delta^{k+1})} \le \sqrt{\frac{2}{\alpha}\Delta^k}.$$

Note,

$$\sum_{i=0}^{+\infty}\|x^{k+i+1}-x^{k+i}\| \le \sqrt{\frac{2}{\alpha}}\sum_{i=0}^{+\infty}\sqrt{\Delta^{k+i}} \le \frac{1}{\sqrt{q}-q}\sqrt{\frac{2}{\alpha}\Delta^k},$$

where the last inequality is due to

$$\sqrt{\Delta^{k+i}} \le \max(\sqrt{\Delta^k}, \sqrt{\Delta^{k+1}/q})\sqrt{q}^i \le \max(\sqrt{\Delta^k}, \sqrt{\Delta^k/q})\sqrt{q}^i = \sqrt{\Delta^k/q}\sqrt{q}^i = \sqrt{\Delta^k}(\sqrt{q})^{i-1}.$$

Since $\sqrt{\Delta^k} = \sqrt{\Delta(x^k, x^{k-1}, \beta^k) - F^*}$ converges to zero R-linearly, thus $\sum_{i=0}^{+\infty}\|x^{k+i+1} - x^{k+i}\| < +\infty$. Thus, $\{x^k\}$ is a Cauchy sequence. Suppose that $\{x^k\}$ converges to x^∞ which is a weak d-stationary. Therefore,

$$\|x^k - x^\infty\| \le \sum_{i=0}^{+\infty}\|x^{k+i+1}-x^{k+i}\| \le \frac{1}{\sqrt{q}-q}\sqrt{\frac{2}{\alpha}\Delta^k}.$$

It follows that $\{x^k\}$ converges to x^∞ R-linearly.
(iii) Define $F^k = F(x^k) - F^*$. Since

$$\left|\left|F^{k+1}\right| - \frac{\beta_{k+1}^2 L_s}{2}\|x^{k+1}-x^k\|^2\right| \le \left|F^{k+1} + \frac{\beta_{k+1}^2 L_s}{2}\|x^{k+1}-x^k\|^2\right|$$

and $\Delta^{k+1} \ge 0$, we have

$$|\, F^{k+1}\,| \le \Delta^{k+1} + \frac{\beta_{k+1}^2 L_s}{2}\|x^{k+1}-x^k\|^2 \le \Delta^{k+1} + \frac{\bar{\beta}^2 L_s}{\alpha}\Delta^k \le (1 + \frac{\bar{\beta}^2 L_s}{\alpha})\Delta^k.$$

Thus, the sequence of objective function values $\{F(x^k)\}$ converges to F^* R-linearly.

Case (II): $\beta_k \equiv 0$.
(i) First, analogous to the proof for Lemma 3.1, we have

$$F(x^{k+1}) + \frac{\sigma + L_s}{2}\|x^{k+1}-x^k\|^2 \le F(x^k). \tag{24}$$

By invoking Lemma 3.4, we obtain that

$$F(x^{k+1}) + \frac{\sigma}{2}\|x^{k+1}-x^k\|^2 \le F^* + \frac{L_s + \hat{L}}{2}\|\bar{x}^k - x^k\|^2.$$

By invoking the assertion (i) of Theorem 3.1, we have $\lim_{k\to+\infty}\|x^k - x^{k+1}\| = 0$. Using (i_k)th error bound in Assumption 2(a) and $x^{k+1} = \mathcal{T}^{(i_k)}(x^k)$, it leads to

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

Combining the above inequality with (24), we have that

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

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

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

where $M = \frac{L_s+\hat{L}}{\sigma+L_s}\tau^2$. Since $\Delta^k = F(x^k) - F^*$, the sequence of $\{\Delta^k\}$ converges to zero Q-linearly, and thus R-linearly. The proof for (ii) and (iii) are similar to Case (I), thus omitted. □

By setting $L_s = 0$, $H_n(x) = H(x) + \iota_X$, $H_s(x) = 0$ and $\beta_k \equiv 0$ in Algorithm 1, it reduces to sDCA [4, Algorithm 1]. The following theorem shows that the linear convergence of sDCA can be established under Assumption 2 without the LLR condition.

Theorem 3.3 *Suppose Assumptions 1 and 2 hold. Let the sequence $\{x^k\}$ be generated by sDCA. Then,*

(i) the sequence of objective function values $\{F(x^k)\}$ converges Q-linearly to F^*;
(ii) the sequence $\{x^k\}$ converges R-linearly to a weak d-stationary point x^∞.

4 Linear Convergence of PEDCA$_e$ to an $A_{\varepsilon'}$-stationary point

Since the basic algorithm (PRA) proposed in [21, Section 5.1] is a specific of ε-DCA [4, Algorithm 2], it has been shown to converge to an $A_{\varepsilon'}$-stationary point ($0 < \varepsilon' < \varepsilon$) [4]. To speed up ε-DCA and make full use of its structure in (1), we consider a proximal ε-DCA with an extrapolation step (PEDCA$_e$). This algorithm is described in Algorithm 2. In contrast to ε-DCA, PEDCA$_e$ introduces an extrapolation step for acceleration, and generates the vector $x^{k,i}$ different from ε-DCA while similar to sPDCA$_e$. Indeed, PEDCA$_e$ is equivalent to nEPDCA$_e$ [16, Algorithm 3]. It has been shown in [16] that any accumulation point of the sequence generated by nEPDCA$_e$ is a d-stationary point. We will sharpen the convergence of sPDCA$_e$ to an $A_{\varepsilon'}$-stationary point (for any $0 < \varepsilon' < \varepsilon$). Then, we establish (locally) linear convergence of this algorithm under Assumption 2. Finally, we prove the linear convergence of ε-DCA under Assumption 2 without the LLR condition.

Algorithm 2 PEDCA$_e$

Initialization: Choose $x^0 \in \mathrm{dom}(F)$, $\varepsilon > 0$ and $\{\beta_k\}_k \subseteq [0, \sqrt{\frac{\sigma}{L_s}})$ with $\bar{\beta} = \sup_k \beta_k < \sqrt{\frac{\sigma}{L_s}}$. Set $x^{-1} = x^0$.

for $k = 0, 1, 2, \ldots$ **do**

Set

$$z^k = x^k + \beta_k(x^k - x^{k-1}). \tag{25}$$

for $i \in \mathcal{M}_\varepsilon(x^k)$, then compute $x^{k,i}$ by **do**

$$x^{k,i} \leftarrow \operatorname*{argmin}_x \left[\ell_i(x, z^k, x^k) + \frac{L_s}{2}\|x - z^k\|^2\right]. \tag{26}$$

end for

Let

$$i_k \leftarrow \operatorname*{argmin}_{i \in \mathcal{M}_\varepsilon(x^k)} \left[F(x^{k,i}) + \frac{\sigma}{2}\|x^{k,i} - x^k\|^2\right]. \tag{27}$$

Set $x^{k+1} \leftarrow x^{k,i_k}$.

end for

The following proposition plays a key role to establish the subsequential convergence.

Proposition 4.1 *Suppose Assumption 1 holds. Let the sequence $\{x^k\}$ be generated by PEDCA$_e$. Then, for all k and $i \in \mathcal{M}(x^k)$,*

$$F(x^{k+1}) + \frac{\bar{\beta}^2 L_s}{2}\|x^{k+1} - x^k\|^2 + \frac{L_s}{2}\|x^{k,i} - x^k\|^2 \le F(x^k) + \frac{\bar{\beta}^2 L_s}{2}\|x^k - x^{k-1}\|^2 - \frac{\sigma - L_s\bar{\beta}^2}{2}\|x^{k+1} - x^k\|^2.$$

Proof Take any $i \in \mathcal{M}(x^k) \subseteq \mathcal{M}_\varepsilon(x^k)$, similar to the proof of Lemma 3.1, we obtain that

$$\begin{aligned}
&F_i(x^k) + \frac{L_s}{2}\|x^k - z^k\|^2 \\
&\ge H_s(z^k) + \langle x^k - z^k, \nabla H_s(z^k)\rangle + H_n(x^k) + \frac{L_s}{2}\|x^k - z^k\|^2 - g_i(x^k) \\
&\ge H_s(z^k) + \langle x^{k,i} - z^k, \nabla H_s(z^k)\rangle + H_n(x^{k,i}) - \langle x^{k,i} - x^k, \nabla g_i(x^k)\rangle - g_i(x^k) \\
&\quad + \frac{L_s}{2}\|x^{k,i} - z^k\|^2 + \frac{L_s}{2}\|x^{k,i} - x^k\|^2 \\
&\ge H_s(x^{k,i}) + H_n(x^{k,i}) - g_i(x^{k,i}) + \frac{\sigma}{2}\|x^{k,i} - x^k\|^2 + \frac{L_s}{2}\|x^{k,i} - x^k\|^2.
\end{aligned} \tag{28}$$

The first inequality is due to the convexity of H_s. The second inequality is by (26), while the last one is due to the strong convexity of g_i. Then, using $i \in \mathcal{M}(x^k)$, it follows from (25), (27) and

(28) that

$$F(x^k) + \frac{L_s\bar{\beta}^2}{2}\|x^{k-1} - x^k\|^2 \\ \geq F(x^{k+1}) + \frac{L_s\bar{\beta}^2}{2}\|x^{k+1} - x^k\|^2 + \frac{L_s}{2}\|x^{k,i} - x^k\|^2 + \frac{\sigma - L_s\bar{\beta}^2}{2}\|x^{k+1} - x^k\|^2.$$

□

Since PEDCA$_e$ is equivalent to nEPDCA$_e$, thus the subsequential convergence of PEDCA$_e$ to a d-stationary point can be obtained from [16, Theorem 4]. In what follows, we sharpen the convergence result to an $A_{\varepsilon'}$-stationary point ($0 < \varepsilon' < \varepsilon$).

Theorem 4.1 *Suppose Assumption 1 holds. Let the sequence $\{x^k\}$ be generated by PEDCA$_e$. Then,*

(i) The sequence $\{F(x^k)\}$ is convergent and the sequence of $\{x^k\}$ is bounded; $\sum_{k=1}^{\infty}\|x^k - x^{k+1}\|^2 < +\infty$;
(ii) Any accumulation point of $\{x^k\}$ has the same objective value, i.e., $\hat{F}^* = \lim_{k\to+\infty} F(x_k)$;
(iii) Any accumulation point of $\{x^k\}$ is an $A_{\varepsilon'}$-stationary point, for any $\varepsilon' \in (0, \varepsilon)$.
(iv) 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 assertions of (i) and (ii) can be found in [16, Theorem 4]. (iii) Let x^∞ be an accumulation point of $\{x^k\}$, and $\{x^{k_j}\}_j$ is a subsequence converging to it. By (i), we know that $\lim_{k\to\infty}\|x^k - x^{k+1}\| = 0$. Thus, $\{x^{k_j+1}\}_j$ also converges to x^∞. Since $\mathcal{M}_{\varepsilon'}(x^\infty) \subseteq \mathcal{M}_\varepsilon(x^{k_j})$ when j is sufficiently large. Thus, for any $i \in \mathcal{M}_{\varepsilon'}(x^\infty)$, one has

$$\begin{aligned}
&F(x^{k_j+1}) + \frac{\sigma}{2}\|x^{k_j+1} - x^{k_j}\|^2 \leq F(x^{k_j,i}) + \frac{\sigma}{2}\|x^{k_j,i} - x^{k_j}\|^2 \\
&\leq H_s(x^{k_j,i}) + H_n(x^{k_j,i}) - g_i(x^{k_j,i}) + \frac{\sigma}{2}\|x^{k_j,i} - x^{k_j}\|^2 \\
&\leq H_s(z^{k_j}) + \langle x^{k_j,i} - z^{k_j}, \nabla H_s(z^{k_j})\rangle + \frac{L_s}{2}\|x^{k_j,i} - z^{k_j}\|^2 \\
&\quad + H_n(x^{k_j,i}) - g_i(x^{k_j}) - \nabla g_i(x^{k_j})^\top(x^{k_j,i} - x^{k_j}) \\
&\leq H_s(z^{k_j}) + \langle x - z^{k_j}, \nabla H_s(z^{k_j})\rangle + \frac{L_s}{2}\|x - z^{k_j}\|^2 \\
&+ H_n(x) - \langle x - x^{k_j}, \nabla g_i(x^{k_j})\rangle - g_i(x^{k_j}).
\end{aligned}$$

The first inequality is due to (27), the third by strong convexity of g_i and the last by the update rule (26). By taking $j \to +\infty$ on both sides, we obtain x^∞ is an $A_{\varepsilon'}$-stationary for any $\varepsilon' \in (0, \varepsilon)$. (iv) It follows directly from [9, Proposition 8.3.10]. □

Next, we prove the linear convergence of PEDCA$_e$ under Assumption 2. Our proof is different from the proofs in Section 3. We introduce multiple projected vectors for convergence rate analysis. More specifically, suppose $\{x^k\}$ be the sequence generated by PEDCA$_e$ and let $\mathcal{L}$ denote the accumulation set of $\{x^k\}$. We define projection vector $\bar{x}^k_{(i)}$ for each $i \in \mathcal{M}(x^k)$ at iteration k, i.e.,

$$\bar{x}^k_{(i)} \in \arg\min_y \left\{\|y - x^k\| \ : \ y \in \Omega^{(i)}\right\}. \tag{29}$$

Next, we introduce the following index set

$$\mathcal{A}^\infty = \left\{i \in [m] \ : \ i \in \mathcal{M}(x),\ x \in \mathcal{L}\right\},$$

which collects all the indicates active at an accumulation point of $\{x^k\}$. In what follows, we show that $\mathcal{M}(x^k) \subseteq \mathcal{A}^\infty$ when k is sufficiently large.

Lemma 4.1 *Suppose Assumption 1 holds. Let $\{x^k\}$ be the sequence generated by PEDCA$_e$ and $\mathcal{L}$ denote accumulation set of $\{x^k\}$. Then, the following statements hold: (i)There exists an index K_0 such that for all $k \geq K_0$, $\mathcal{M}(x^k) \subseteq \mathcal{A}^\infty$; (ii) $\mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)} \neq \emptyset$ for all $i \in \mathcal{A}^\infty$.*

Proof (i) Suppose not, 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 $i_{k_j} \notin \mathcal{A}^\infty$. There exists a subsequence $i_{k_j} \equiv \bar{i}$ when $j \in \kappa$ where κ denotes the subsequence. Thus, we have $\bar{i} \notin \mathcal{A}^\infty$. On the other hand, $\bar{i} \in \mathcal{A}^\infty$ due to $\bar{i} \in \mathcal{M}(x^{k_j})$ for infinitely many k_j which contradicts $\bar{i} \notin \mathcal{A}^\infty$.

(ii) Take any $\bar{i} \in \mathcal{A}^\infty$. 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 $j \to +\infty$. Obviously, $x^\infty \in \mathcal{L} \cap \Omega^{(\bar{i})} \cap \mathcal{D}^{(\bar{i})}$. □

Remark 4.1 From Lemma 4.1, we see that $\Omega^{(i)}$ is nonempty for all $i \in \mathcal{M}(x^k)$ when $k \geq K_0$. Therefore, the vectors of $\{\bar{x}^k_{(i)}\}_{i \in \mathcal{M}(x^k)}$ are well-defined.

Next, we show that $\bar{x}^k_{(i)}$ eventually settles down at some isocost surface of F_i which was also used for the convergence rate analysis of feasible descent methods [19].

Lemma 4.2 *Suppose Assumption 1 holds. Let $\{x^k\}$ be the sequence generated by PEDCA$_e$. Suppose Assumption 2(b) holds, then there exists an index $K_1 (\geq K_0)$ such that $i \in \mathcal{M}(x^k)$, $F_i(\bar{x}^k_{(i)}) = \hat{F}^*$ when $k \geq K_1$.*

Proof Let $\mathcal{L}$ denote the accumulation set of $\{x^k\}$. We show that for any given $\varepsilon > 0$, there exists an index K such that when $k \geq K$, for any $i \in \mathcal{M}(x^k)$, $\mathrm{dist}(x^k, \mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)}) < \varepsilon$. Suppose not. There exists a $\varepsilon > 0$ and $i_{k_j} \in \mathcal{M}(x^{k_j})$ 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. There is an index $\bar{i}$ appears infinite numbers in the sequence $\{i_{k_j}\}$. Thus, there is a subsequence $\{x^{k_j}\}_{j \in \kappa}$ such that $i_{k_j} \equiv \bar{i}$ for $j \in \kappa$ and $\{x^{k_j}\}_{j \in \kappa}$ converges to $x^\infty \in \mathcal{L}$. Thus, $x^\infty \in \mathcal{L} \cap \Omega^{(\bar{i})} \cap \mathcal{D}^{(\bar{i})}$ which contradicts $\mathrm{dist}(x^{k_j}, \mathcal{L} \cap \Omega^{(i_{k_j})} \cap \mathcal{D}^{(i_{k_j})}) \geq \varepsilon$.

For the μ given in Assumption 2(b), there exists an index $K_1 (\geq K_0)$ such that

$$\mathrm{dist}(x^k, \mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)}) < \mu/2, \quad \forall\, i \in \mathcal{M}(x^k), \text{ when } k \geq K_1.$$

By definition of $\bar{x}^k_{(i)}$ defined in (29), we have for all $i \in \mathcal{M}(x^k)$, $\|x^k - \bar{x}^k_{(i)}\| = \mathrm{dist}(x^k, \Omega^{(i)}) < \mu/2$. Define $\hat{x}^k \in \mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)}$ be the projection of x^k onto $\mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)}$. By the triangle inequality, one has $\|\hat{x}^k - \bar{x}^k_{(i)}\| < \mu$. Since $\hat{x}^k \in \mathcal{L} \cap \Omega^{(i)} \cap \mathcal{D}^{(i)}$, $i \in \mathcal{M}(\hat{x}^k)$ when $k \geq K_1$, then $F_i(\hat{x}^k) = F(\hat{x}^k)$. In addition, Assumption 2(b) implies that $F_i(\bar{x}^k_{(i)}) = F_i(\hat{x}^k) = F(\hat{x}^k) = \hat{F}^*$ for all $i \in \mathcal{M}(x^k)$ and $k \geq K_1$. □

Now we are ready to prove the linear convergence of PEDCA$_e$.

Theorem 4.2 *Suppose Assumptions 1 and 2 hold, and let $\{x^k\}$ be the sequence generated by PEDCA$_e$, then*

(i) $\{x^k\}$ converges to an $A_{\varepsilon'}$-stationary point R-linearly, where ε' is an arbitrary value in $(0, \varepsilon)$;
(ii) $\{F(x^k) - \hat{F}^*\}$ converges to zero R-linearly.

Proof In the following, we divide into two cases to verify.

Case (I): $\beta_k \not\equiv 0$.

(i) First, we show that the sequence $\{\tilde{\Delta}^{k+1}\}$ converges to zero Q-linearly where $\tilde{\Delta}^{k+1} = \Delta(x^{k+1}, x^k, \bar{\beta}) - \hat{F}^*$. Similar to the proof for Lemma 3.4 and recalling $\beta_{k+1} \leq \bar{\beta} < \sqrt{\frac{\sigma}{L_s}}$, we have for any $i \in \mathcal{M}(x^k)$

$$F_i(\bar{x}^k_{(i)}) + \frac{L_s}{2}\|\bar{x}^k_{(i)} - z^k\|^2 + \frac{\hat{L}}{2}\|\bar{x}^k_{(i)} - x^k\|^2 \geq F(x^{k+1}) + \frac{\bar{\beta}^2 L_s}{2}\|x^{k+1} - x^k\|^2.$$

In addition, $F_i(x^k_{(i)}) = \hat{F}^*$ $(i \in \mathcal{M}(x^k))$ when k is sufficiently large from Lemma 4.2. Using this fact and (17), it follows from the above inequality that for any $i \in \mathcal{M}(x^k)$

$$L_s(\|\bar{x}^k_{(i)} - x^k\|^2 + \beta_k^2\|x^k - x^{k-1}\|^2) + \frac{\hat{L}}{2}\|\bar{x}^k_{(i)} - x^k\|^2 \geq \Delta(x^{k+1}, x^k, \bar{\beta}) - \hat{F}^*.$$

Invoking (20), it further leads to

$$\tilde{\Delta}^{k+1} \le \tilde{c}_1\|x^k - x^{k+1}\|^2 + \tilde{c}_2\|x^k - x^{k-1}\|^2, \tag{30}$$

where $\tilde{c}_1 = 2L_s\tau^2 + \hat{L}\tau^2$ and $\tilde{c}_2 = (8(L_s\tau^2 + \frac{1}{2}\hat{L}\tau^2) + L_s)\bar{\beta}^2$. Next, invoking Proposition 4.1, we have

$$\|x^k - x^{k+1}\|^2 \le \frac{2}{\tilde{\alpha}}\left[\Delta(x^k, x^{k-1}, \bar{\beta}) - \Delta(x^{k+1}, x^k, \bar{\beta})\right], \tag{31}$$

where $\tilde{\alpha} = \sigma - L_s\bar{\beta}^2$. By setting $k = k-1$ in (31), we have

$$\|x^k - x^{k-1}\|^2 \le \frac{2}{\tilde{\alpha}}\left[\Delta(x^{k-1}, x^{k-2}, \bar{\beta}) - \Delta(x^k, x^{k-1}, \bar{\beta})\right]. \tag{32}$$

By substituting (32) and (31) into (30), and with some rearrangements, we further have

$$(1 + \frac{2\tilde{c}_1}{\tilde{\alpha}})\tilde{\Delta}^{k+1} \le (\frac{2\tilde{c}_1}{\tilde{\alpha}} - \frac{2\tilde{c}_2}{\tilde{\alpha}})\tilde{\Delta}^k + \frac{2\tilde{c}_2}{\tilde{\alpha}}\tilde{\Delta}^{k-1}. \tag{33}$$

Next, we divide into two cases to verify: (a) $\tilde{c}_1 \ge \tilde{c}_2$; (b) $\tilde{c}_1 < \tilde{c}_2$.
Case (a): $\tilde{c}_1 \ge \tilde{c}_2$. It follows from the above inequality that $(1 + \frac{2\tilde{c}_1}{\tilde{\alpha}})\tilde{\Delta}^{k+1} \le \frac{2\tilde{c}_1}{\tilde{\alpha}}\tilde{\Delta}^{k-1}$ due to $\tilde{\Delta}^k \le \tilde{\Delta}^{k-1}$. Thus, $\tilde{\Delta}^{k+1} \le \frac{2\tilde{c}_1}{\tilde{\alpha}+2\tilde{c}_1}\tilde{\Delta}^{k-1}$.
Case (b): $\tilde{c}_1 < \tilde{c}_2$. It leads to $(1 + \frac{2\tilde{c}_1}{\alpha})\tilde{\Delta}^{k+1} + \frac{2\tilde{c}_2 - 2\tilde{c}_1}{\alpha}\tilde{\Delta}^k \le \frac{2\tilde{c}_2}{\alpha}\tilde{\Delta}^{k-1}$. By using $\tilde{\Delta}^{k+1} \le \tilde{\Delta}^k$, we get $\tilde{\Delta}^{k+1} \le \frac{2\tilde{c}_2}{\alpha+2\tilde{c}_2}\tilde{\Delta}^{k-1}$. For both cases, we have $\tilde{\Delta}^k \le \max(\tilde{\Delta}^0, \frac{\tilde{\Delta}^1}{\tilde{q}})\tilde{q}^k$ by letting $\tilde{q} = \max(\tilde{q}_1, \tilde{q}_2)$ where $\tilde{q}_1 = \sqrt{\frac{2\tilde{c}_1}{\tilde{\alpha}+2\tilde{c}_1}} \in (0,1)$ and $\tilde{q}_2 = \sqrt{\frac{2\tilde{c}_2}{\tilde{\alpha}+2\tilde{c}_2}} \in (0,1)$. The sequence $\{\tilde{\Delta}^k\}$ converges R-linearly to zero. The remaining proof for (i) and (ii) are similar to Theorem 3.2, thus omitted here.

Case (II): $\beta_k \equiv 0$.
(i) First, we show that the sequence $\{\hat{\Delta}^k\}$ converges to zero Q-linearly where $\hat{\Delta}^k = F(x^k) + \frac{\sigma}{2}\|x^k - x^{k-1}\|^2 - \hat{F}^*$. Analogous to (28), for $i \in \mathcal{M}(x^k)$, one has

$$F(x^k) = F_i(x^k) \ge F(x^{k+1}) + \frac{\sigma}{2}\|x^{k+1} - x^k\|^2 + \frac{L_s}{2}\|x^{k,i} - x^k\|^2.$$

Invoking the definition of $\hat{\Delta}^k$, we have

$$\hat{\Delta}^k \ge \hat{\Delta}^{k+1} + \frac{L_s}{2}\|x^{k,i} - x^k\|^2 + \frac{\sigma}{2}\|x^k - x^{k-1}\|^2, \quad i \in \mathcal{M}(x^k). \tag{34}$$

Similar to the proof for Lemma 3.4, we obtain that

$$F(x^{k,i}) + \frac{\sigma}{2}\|x^{k,i} - x^k\|^2 \le F(x) + \frac{L_s + \hat{L}}{2}\|x - x^k\|^2, \quad i \in \mathcal{M}(x^k).$$

Set $x = \bar{x}^k_{(i)}$ in the above inequality and using Lemma 4.2, it leads to

$$F(x^{k,i}) + \frac{\sigma}{2}\|x^{k,i} - x^k\|^2 \le \hat{F}^* + \frac{L_s + \hat{L}}{2}\|\bar{x}^k_{(i)} - x^k\|^2, \quad i \in \mathcal{M}(x^k). \tag{35}$$

By Theorem 4.1, we have $\lim_{k\to+\infty}\|x^k - x^{k+1}\| = 0$. Thus, by using the ith error bound condition in Assumption 2(a), and invoking $x^{k,i} = \mathcal{T}^{(i)}(x^k)$, it leads to

$$F(x^{k,i}) + \frac{\sigma}{2}\|x^{k,i} - x^k\|^2 \le \hat{F}^* + \frac{L_s + \hat{L}}{2}\tau^2\|x^k - x^{k,i}\|^2.$$

Furthermore, $\hat{\Delta}^{k+1} \le \frac{L_s+\hat{L}}{2}\tau^2\|x^k - x^{k,i}\|^2$. Combining the above inequality with (34), and we have that $\hat{\Delta}^{k+1} \le \frac{L_s+\hat{L}}{L_s}\tau^2(\hat{\Delta}^k - \hat{\Delta}^{k+1})$. By rearranging terms, it further leads to,

$$\hat{\Delta}^{k+1} \le \frac{\hat{M}}{\hat{M}+1}\hat{\Delta}^k,$$

where $\hat{M} = \frac{L_s+\hat{L}}{L_s}\tau^2$. Therefore, the sequence of $\{\hat{\Delta}^k\}$ converges to zero Q-linearly. The remaining proof for (i) and (ii) are similar to Case (I), thus omitted here. □

To end this section, we show that the linear convergence of ε-DCA [4, Algorithm 2] under Assumption 2. By setting $L_s = 0$, $H_n(x) = H(x) + \iota_X$, $H_s(x) = 0$ and $\beta_k \equiv 0$ in PEDCA$_e$, it reduces to ε-DCA.

Theorem 4.3 *Suppose Assumptions 1 and 2 hold, and let $\{x^k\}$ be the sequence generated by ε-DCA, then*

(i) $\{F(x^k) - \hat{F}^*\}$ converges to zero R-linearly;
(ii) $\{x^k\}$ converges to an $A_{\varepsilon'}$-stationary point R-linearly, where ε' is an arbitrary value in $(0, \varepsilon)$.

5 Guarantees for Key Assumptions

We focus on discussing the guarantees for the key assumptions used to prove linear convergence of two algorithms in Sections 3 and 4. We first consider the error bound condition, i.e., Assumption 2(a). For the DC program (7), we show that the equivalence between the subregularity of the subdifferential of the objective function and the error bound condition in item (ii) of Proposition 5.1. Second, we show that two types of error bound condition described in items (i) and (ii) of Proposition 5.2 are also equivalent. Equipped with these results, we illustrate that Assumption 2(a) does not depend on a particular DC decomposition. Therefore, we can apply the existing results on the error bound condition [18,17,19,25,30] to identify Assumption 2(a).

Proposition 5.1 *Consider DC program (7) with the triplet (H_n, H_s, g) satisfying Assumption 1 with parameters (σ, L_s, L_g). The stationary set Ω is defined in (8). Suppose that $\Omega \neq \emptyset$ and $\bar{x} \in \Omega$. Consider the following two conditions:*

(i) The multi-function $\partial(H_n + H_s - g)(\cdot)$ is subregular at $(\bar{x}, 0)$;
(ii) There exists an open neighborhood $\mathbb{O}$ of $\bar{x}$ such that $dist(x, \Omega) \leq \hat{\ell} \left\| x - \mathcal{T}^f(x) \right\|$ for all $x \in \mathbb{O}$ and the operator $\mathcal{T}^f$ is defined as i.e.,

$$\mathcal{T}^f(x) = \operatorname*{argmin}_y \left\{ H_s(x) + \langle \nabla H_s(x) - \nabla g(x), y - x \rangle + H_n(y) - g(x) + \frac{L_s}{2} \|x - y\|^2 \right\}.$$

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

Proof The proof is similar to [4, Proposition 6.1], thus is omitted here. □

The next lemma is to reveal the equivalence between two types of error bounds. The difference between them lies in the "valid regions".

Proposition 5.2 *Consider DC program (7) with the triplet (H_n, H_s, g) satisfying Assumption 1 with parameters (σ, L_s, L_g). Suppose that $\Omega \neq \emptyset$ and Ω is defined in (8). Let $\mathcal{X}$ be a compact set such that $\Omega \subseteq \mathcal{X}$ (Suppose such compact set $\mathcal{X}$ exists). The following two conditions are equivalent:*

(i) For each $\bar{x} \in \Omega$, there exists an open neighborhood $\mathbb{O}_{\bar{x}}$ of $\bar{x}$ such that $dist(x, \Omega) \leq \hat{\ell} \left\| x - \mathcal{T}^f(x) \right\|$ for all $x \in \mathbb{O}_{\bar{x}} \cap \mathcal{X}$;
(ii) There exist $\ell, \varepsilon > 0$ such that $dist(x, \Omega) \leq \ell \left\| x - \mathcal{T}^f(x) \right\|$ when $x \in \mathcal{X}$ and $\left\| x - \mathcal{T}^f(x) \right\| \leq \varepsilon$.

Proof For (ii)⇒(i), we only need to show that

$$x \in \mathbb{O}_{\bar{x}} \cap \mathcal{X} \Rightarrow x \in \mathcal{X} \text{ and } \left\| x - \mathcal{T}^f(x) \right\| \leq \varepsilon.$$

It holds obviously due to $\bar{x} = \mathcal{T}^f(\bar{x})$ and the Lipschitz continuity of the mapping of $(x - \mathcal{T}^f(x))$.

To prove (i)⇒(ii), suppose condition (i) holds. For each $\bar{x} \in \Omega$, there exists an open ball $\mathcal{B}(\bar{x}, \bar{r})$ (without loss of generality, we assume $\mathcal{B}(\bar{x}, \bar{r}) \subseteq \mathcal{X}$) such that $\operatorname{dist}(x, \Omega) \leq \hat{\ell} \left\| x - \mathcal{T}^f(x) \right\|$ for all $x \in \mathcal{B}(\bar{x}, \bar{r})$. These open balls $\mathcal{B}(\bar{x}, \bar{r}/2)$ form an open cover of Ω, hence there exists a finite subcover, i.e., $\Omega \subseteq \cup_{i \in \mathcal{S}} \mathcal{B}(x_i, r_i/2)$, where $\mathcal{S}$ is a finite index set. By setting $\delta = \min_{i \in \mathcal{S}} r_i/2$, we can show that $\Omega_\delta = \{x : \operatorname{dist}(x, \Omega) < \delta/2\}$ belongs to $\cup_{i \in \mathcal{S}} \mathcal{B}(x_i, r_i)$.

Let $\ell = \max_{i\in\mathcal{S}} \hat{\ell}$. Therefore, $\mathrm{dist}(x,\Omega) \le \hat{\ell}\left\|x - \mathcal{T}^f(x)\right\|$ for all $x \in \Omega_\delta$. Next, we show that there exists a scale $\varepsilon > 0$ such that

$$x \in \mathcal{X} \text{ and } \left\|x - \mathcal{T}^f(x)\right\| \le \varepsilon \Rightarrow x \in \Omega_\delta \cap \mathcal{X}.$$

Suppose not. There exists a sequence $\{x^k\} \subseteq \mathcal{X}$ such that $\|x^k - \mathcal{T}^f(x^k)\| \le 1/k$ and $x^k \notin \Omega_\delta$. There exists a subsequence x^{k_j} converges to x^∞. Thus, $x^\infty \notin \Omega_\delta$. On the other hand, x^∞ satisfying $x^\infty = \mathcal{T}^f(x^\infty)$, and thus $x^\infty \in \Omega$ which contradicts $x^\infty \notin \Omega_\delta$. Therefore, (i)⇒(ii) holds. □

Since F is level bounded, for any $\zeta \ge \inf_x F(x)$, the set of $\{x : F(x) \le \zeta\}$ is a compact set. By setting $f = F_i$ and $\mathcal{X} = \{x : F(x) \le \zeta\}$, the assertion (ii) of Proposition 5.2 is exactly Assumption 2(a). The following proposition summarizes the equivalence between the subregularity of the subdifferential mapping ∂F_i and two types of error bounds.

Proposition 5.3 *Let $(H_n, H_s, \{g_i\}_{i=1}^m)$ be a triplet satisfying Assumption 1 with scalars $\big(\sigma, L_s, \{L_i\}_{i=1}^m\big)$. Suppose $i \in \mathcal{I}$. The following three conditions are equivalent: for any $\zeta \ge \inf_x F(x)$,*

(i) for each $\bar{x} \in \Omega^{(i)}$, the multi-function $\partial F_i(\cdot)$ is subregular at $(\bar{x}, 0)$;
(ii) for each $\bar{x} \in \Omega^{(i)}$, there exists an open neighborhood $\mathbb{O}_{\bar{x}}$ of $\bar{x}$ such that $dist(x, \Omega^{(i)}) \le \hat{\ell}\left\|x - \mathcal{T}^{(i)}(x)\right\|$ whenever $x \in \mathbb{O}_{\bar{x}}$ and $F(x) \le \zeta$;
(iii) there exist $\ell, \varepsilon > 0$ such that $dist(x,\Omega) \le \ell\left\|x - \mathcal{T}^{(i)}(x)\right\|$ when $\left\|x - \mathcal{T}^{(i)}(x)\right\| \le \varepsilon$ and $F(x) \le \zeta$.

Therefore, for each $i \in \mathcal{I}$, F_i satisfies the assumptions in [4, Proposition 6.3], then Assumption 2(a) holds. On the other hand, Assumption 2(b) holds whenever F_i takes only finitely many different values on $\Omega^{(i)}$ under Assumption 1; see, e.g., [4, Proposition 6.5].

6 Discussions on Concrete Models

Consider the least squares problems, the objective function of these problems takes the form:

$$\min_{x\in\mathbb{R}^n} \tilde{F}(x) = \frac{1}{2}\|Ax - b\|^2 + P(x), \tag{36}$$

where $A \in \mathbb{R}^{m\times n}$ is measurement matrix and $b \in \mathbb{R}^n$, and P is nonsmooth regularizer. Next, we present two concrete examples with different types of P, and both of them satisfy Assumptions 2(a) and 2(b).

Example 1 For the least squares problems (36) with the smoothly clipped absolute deviation (SCAD) [6] or the minimax concave penalty (MCP) [29] regularization, the function P usually takes the form: $P(x) = \lambda\sum_{i=1}^n p_i(x_i)$. The corresponding function p_i is defined as (with $\lambda > 0, \theta > 0$):

$$p_{\mathrm{MCP}}(t) = \begin{cases} \lambda \mid t \mid -\frac{t^2}{2\theta}, & \text{if } \mid t \mid \le \theta\lambda, \\ \frac{\theta\lambda^2}{2}, & \text{if } \mid t \mid \ge \theta\lambda, \end{cases}$$

for MCP regularization, and

$$p_{\mathrm{SCAD}}(t) = \begin{cases} \lambda \mid t \mid, & \text{if } \mid t \mid \le \lambda, \\ \frac{-t^2 + 2\theta\lambda|t| - \lambda^2}{2(\theta-1)}, & \text{if } \lambda < \mid t \mid \le \theta\lambda, \\ \frac{(\theta+1)\lambda^2}{2}, & \text{if } \mid t \mid \ge \theta\lambda, \end{cases}$$

for SCAD regularization where $\lambda > 0,\ \theta > 2$, respectively. Use a similar technique to the arguments in [10, Section 5.2], (36) with MCP or SCAD regularization can be reformulated as

$$\tilde{F}(x) = \min_{j\in\mathcal{J}} \left(\frac{1}{2}\|Ax - b\|^2 + \sum_{i=1}^n [p_{i,j_i}(x_i) + \delta_{C_i,j_i}(x_i)] \right),$$

where each of $p_{i,s}$ is quadratic (or linear) function, and each of $C_{i,s}$ is a closed interval, and $\mathcal{J} = \{j = (j_1, ..., j_n) \in \mathbb{N}^n \,|1 \le j_i \le m_i,\ 1 \le i \le n\}$ is a finite set. Then, by defining $\tilde{F}_j(x) = \frac{1}{2}\|Ax-b\|^2 + \sum_{i=1}^n [p_{i,j_i}(x_i) + \delta_{C_i,j_i}(x_i)]$ for each $j \in \mathcal{J}$. Consequently, $\tilde{F}(x) = \min_{j\in\mathcal{J}} F_j(x)$. Each F_i is satisfying Assumption 2(a) and 2(b) where each F_i comes from the least squares problem (36) with MCP or SCAD regularizer due to [4, Proposition 6.3].

Example 2 For the least squares problems (36) with the truncated ℓ_1 regularizer [26], the corresponding function $P(x) = \lambda\|x\|_1 - \lambda\mu\sum_{i=1}^p |x_{[i]}|$, where $|x_{[i]}|$ denotes the ith largest element in magnitude, $\lambda > 0$, $\mu \in (0, 1]$ and $p < n$ is a positive integer. Analogous to Example 1, we can show that it also satisfies Assumption 2(a) and 2(b).

7 Numerical results

We compare numerical performances of sPDCA$_e$, PEDCA$_e$, PEDCA (setting $\beta_k \equiv 0$ in PEDCA$_e$) with GIST ([20, Algorithm 2], [7, Algorithm 1]), NPG$_{\rm major}$ ([13, Appendix A, Algorithm 2]) and mAPG [11, Algorithm 1] for solving some sparse regression problems. We will verify the following assertions:

(1) Either PEDCA$_e$ or PEDCA converges to an $A_{\varepsilon'}$-stationary point ($0 < \varepsilon' < \varepsilon$), which is much stronger than d-stationary solution. It can be empirically verified by showing the final objective function values of either PEDCA$_e$ or PEDCA are much lower than the others in most scenarios.
(2) The extrapolation step can accelerate the speed of PEDCA$_e$.

All of these algorithms were coded in MATLAB R2016a, and all of the experiments were performed on a desktop with Windows 10 and an Intel Core i7-7600U CPU processor (2.80 GHz) with 16 GB memory.

Consider the least squares problem with the truncated ℓ_1 regularizer:

$$\min_{x\in\Re^n}\left\{F(x) = \frac{1}{2}\|Ax-b\|^2 + \lambda\left(\|x\|_1 - \sum_{i=1}^p |x_{[i]}|\right)\right\}, \tag{37}$$

where $0 \le p < n$, $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$ and $\lambda > 0$ is the balance parameter. $x_{[i]}$ denotes the ith largest component of x in magnitude. Problem (37) is indeed Example 2 by setting $\mu = 1$, and has been used in [8]. Set $H_s(x) = \|Ax-b\|^2/2$, $H_n(x) = \lambda\|x\|_1 + \frac{\sigma}{2}\|x\|^2$ and $G(x) = \lambda\sum_{i=1}^p |x_{[i]}| + \frac{\sigma}{2}\|x\|^2$.

We generate the model parameters of (n, p, λ, A, b) as follows. Given positive integers n, p ($p < n$) and $\lambda > 0$, we generate a matrix A, a vector b, and a critical but not a d-stationary point $\widetilde{x}$. More specifically, we generate a vector $v \in \mathbb{R}^p$ such that $|v_1| \ge |v_2| \ge \cdots \ge |v_p|$. Then, define a vector $\widetilde{x} \in \mathbb{R}^n$ by letting $\widetilde{x} = v_i + \text{sign}(v_i)$ for $i = 1, \ldots, p$, $\widetilde{x}_{p+1} = \widetilde{x}_{p+2} = v_p + \text{sign}(v_p)$ and $\widetilde{x}_i = 0$ for $i = p+3, \ldots, n$. Thus, $|\widetilde{x}_1| \ge |\widetilde{x}_2| \ge \cdots \ge |\widetilde{x}_p| = |\widetilde{x}_{p+1}| = |\widetilde{x}_{p+2}| > |\widetilde{x}_{p+3}| = \cdots = |\widetilde{x}_n| = 0$. Next, we obtain a vector $d \in \mathbb{R}^n$ by reordering the vector $\widetilde{d} \in \mathbb{R}^n$ in descend order where the entries of $\widetilde{d}$ is generated randomly from the uniform distribution on $\left[-\sqrt{\lambda}, \sqrt{\lambda}\right]$. With the vector d, we generate the matrix A by letting $A = \text{Diag}(d) + 0.01\widetilde{A}$, where $\text{Diag}(d)$ is the diagonal matrix with the vector d on the main diagonal and the entries of $\widetilde{A}$ are subjected to the uniform distribution on $\left[-\frac{1}{n}, \frac{1}{n}\right]$. Finally, we compute the vector b by solving the linear equation $A^\top b = A^\top A\widetilde{x} + \xi^1 - \xi^2$ for some $\xi^1 \in \partial H_n(\widetilde{x})$, $\xi^2 \in \partial G(\widetilde{x})$.

For sPDCA$_e$, PEDCA$_e$, we set the step size β_k via a restart strategy, i.e., let $\beta_k = \tau(\theta_{k-1} - 1)/\theta_k$, where $\theta_{-1} = \theta_0 = 1$, $\theta_{k+1} = \frac{1+\sqrt{1+4\theta_k^2}}{2}$, and reset $\theta_{k-1} = \theta_k = 1$ when $\texttt{mod}(k, 200) = 0$ or $(z^k - x^{k+1})^\top(x^{k+1} - x^k) > 0$. The notation of mod denotes coresidual. For sPDCA$_e$, we set $\tau = 1$ as recommended in [27,16]. For PEDCA$_e$ and PEDCA, we take $\varepsilon = 0.01$. Choose $\tau = 0.99$, $\sigma = \tau^2 L_s$ in PEDCA$_e$ and PEDCA. Obviously, $\sup_k \beta_k < \sqrt{\sigma/L_s}$. For GIST ([20, Algorithm 2]), NPG$_{\rm major}$ ([13, Appendix A, Algorithm 2]) and mAPG [11, Algorithm 1], we all use their defaulted setting. More specifically, we set $\tau = 1.5$, $M = 5$ $c = 1$ and $L_{\min} = 1$, $L_{\max} = \lambda_{\max}(A^\top A) + 10^3$ in

Table 1: Algorithmic comparison in terms of objective function value

(n,p,λ)	PEDCAe	PEDCA	sPDCAe	$\text{NPG}_{\text{major}}$	GIST	mAPG
(500,150,5)	0.8809	0.8809	4.5016	6.2571	0.8658	0.8970
(1000,300,10)	1.8839	1.8839	9.4522	1.8792	1.8878	1.8943
(1500,450,15)	2.6582	2.6582	2.6511	2.7036	2.8663	2.6874
(2000,600,20)	3.3976	3.3976	3.3976	3.4014	3.5143	3.4256
(2500,750,25)	3.8463	3.8463	47.2171	3.9517	4.0776	3.8898
(3000,900,30)	4.7082	4.7082	4.7082	4.8472	4.7293	4.7984
(3500,1050,35)	5.6131	5.6131	5.6133	5.6533	5.6628	5.6261
(4000,1200,40)	6.3537	6.3537	40.1344	6.5069	6.4385	6.4911
(4500,1350,45)	7.1181	7.1181	44.5239	7.1458	7.2472	7.1505
(5000,1500,50)	7.9636	7.9636	7.9838	7.9923	7.9353	8.0402

Table 2: Algorithmic comparison in terms of CPU Time

(n,p,λ)	PEDCAe	PEDCA	sPDCAe	$\text{NPG}_{\text{major}}$	GIST	mAPG
(500,150,5)	0.0703	0.1047	0.0437	0.1031	0.2219	0.0406
(1000,300,10)	0.2437	0.4391	0.1844	0.7172	1.4594	0.2109
(1500,450,15)	0.5219	1.0906	0.4688	4.5625	5.0406	0.6344
(2000,600,20)	0.8375	1.7656	0.8750	11.1563	12.5531	1.0781
(2500,750,25)	1.9312	4.1594	1.3531	24.7156	33.2219	1.9781
(3000,900,30)	2.1969	5.5406	2.1500	30.1844	35.4688	3.2656
(3500,1050,35)	3.1313	7.4656	3.0469	36.7375	39.6844	5.1063
(4000,1200,40)	4.7250	10.5563	4.8250	36.7625	61.7531	6.4594
(4500,1350,45)	4.9719	11.4563	4.7500	85.9437	95.1063	8.1656
(5000,1500,50)	6.8469	15.4156	6.9313	59.1563	62.8250	13.1563

$\text{NPG}_{\text{major}}$; Take $\tau = 1.5$, $r = 5$, $\sigma = 0.2$ and $L_{\min} = 1$, $L_{\max} = \lambda_{\max}(A^\top A) + 10^3$ in GIST; Let $\alpha_x = \alpha_y = \frac{1}{\lambda_{\max}(A^\top A)+1e-6}$ in mAPG.

We set the initial point as $x_0 = \widetilde{x} + 0.01\xi$ for all these algorithms where each entry of $\xi \in \mathbb{R}^n$ is randomly generated from the uniform distribution on $[-1, 1]$. We terminated all these algorithms by the stopping criterion of $\left|H(x^k) - H(x^{k-1})\right| \le 10^{-8}$ which was also used in [16].

In our experiments, we generate the data pair of $(n, p, \lambda, A, b, \tilde{x})$ as above where $\tilde{x}$ is a critical point but not a d-stationary point. By setting the parameters of the model (37) with the constructed pair, we test six algorithms for the same instance. More specifically, we test for different scenarios of $(n, p, \lambda) = (500j, 150j, 50j)$ with $j = 1, 2, \cdots, 10$. For each instance, we run 20 trials for all of these algorithms and record the average results. We report the average performance in Tables 1 and 2, in terms of objective function value and CPU time (s), respectively. From Tables 1 and 2, we see that PEDCA_e, PEDCA, GIST and mAPG all converge to stronger stationary points than the others in the sense of achieving much lower objective function values. Among these four, PEDCA_e and PEDCA always achieve a much lower objective function value than that of GIST and mAPG. According to the analysis in [20], GIST and mAPG also converge to d-stationary points. Nevertheless, either PEDCA_e or PEDCA converges to an $A_{\varepsilon'}$-stationary points ($0 < \varepsilon' < \varepsilon$) which is much stronger than a d-stationary point. We have observed this result by finding that the final objective function values from PEDCA_e and PEDCA are always much lower than that from GIST and mAPG. Second, PEDCA_e costs less time than mAPG, and significantly reduces the CPU time of GIST especially when the dimension is large ($n \ge 2500$). In contrast to PEDCA, PEDCA_e cuts half of its time and ends with the same objective function values. This shows that the extrapolation step can improve the speed of PEDCA_e.

8 Conclusions

In this paper, we consider a class of structured nonsmooth DC minimization. By exploiting its particular structure, we propose two extrapolation proximal difference-of-convex based algorithms for potential acceleration to compute a weak/standard d-stationary point. Indeed, we sharpen the convergence of the latter algorithm to an $A_{\varepsilon'}$-stationary point. We further establish the linear convergence of these algorithms under piecewise error bound and piecewise isocost condition. As

a corollary, the linear convergence of sDCA and ε-DCA is obtained under the same conditions without the LLR condition [4]. Furthermore, we discuss some sufficient conditions to ensure the key assumptions and provide several sparse recovery models satisfying all these assumptions. Some elementary numerical results verify our theoretical results.

References

1. M. Ahn, J. S. Pang, and J. Xin. Difference-of-convex learning: Directional stationarity, optimality, and sparsity. *SIAM J. Optim.*, 27(3):1637–1665, 2017.
2. J. Bolte, A. Daniilidis, and A. Lewis. The Łojasiewicz Inequality for Nonsmooth Subanalytic Functions with Applications to Subgradient Dynamical Systems. *SIAM J. Optim.*, 17(4):1205–1223, 2007.
3. T. Pham Dinh and H. A. Le Thi. Convex analysis approach to DC programming: theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22:289–355, 1997.
4. H. B. Dong, M. Tao. On the linear convergence to weak/standard d-stationary points of DCA-based algorithms for structured nonsmooth DC programming. *J. Optim. Theory Appl.* 189(1):190–220, 2021.
5. D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Math. Oper. Res.*, 2018.
6. J. Q. Fan and R. Z. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, 96(456):1348–1360, 2001.
7. P. Gong, C. Zhang, Z. Lu, J. Z. Huang, and J. Ye. A general iterative shinkage and thresholding algorithm for non-convex regularized optimization problems. *Proc Int Conf Mach Learn.* 28(2):37–45, 2013.
8. J. Y. Gotoh, A. Takeda, and K. Tono. DC formulations and algorithms for sparse optimization problems. *Math. Program.*, 169(1):141–176, 2018.
9. P. T. Harker and J. S. Pang. Finite-Dimensional Variational Inequalities and Complementarity Problems, Vol. II, Springer, New York, 2003.
10. G. Y. Li and T. K. Pong. Calculus of the exponent of Kurdyka–Łojasiewicz inequality and its applications to linear convergence of first-order methods. *Found. Comput. Math.*, 18(5):1199–1232, 2018.
11. H. Li and Z. Lin. Accelerated proximal gradient methods for nonconvex programming. Advances in neural information processing systems, 28, 2015.
12. T. X. Liu, T. K. Pong, and A. Takeda. A refined convergence analysis of pDCA$_e$ with applications to simultaneous sparse recovery and outlier detection. *Comput. Optim. Appl.*, 73(1):69–100, 2019.
13. T. Liu, T. K. Pong, and A. A. Takeda. Successive difference-of-convex approximation method for a class of nonconvex nonsmooth optimization problems. *Math. Program.*, 176:339–367, 2019.
14. T. Liu and T. K. Pong. Further properties of the forward-backward envelope with applications to difference-of-convex programming. *Comput. Optim. Appl.* 67:489–520, 2017.
15. Z. S. Lu, Z. R. Zhou, and Z. Sun. Enhanced Proximal DC Algorithms with Extrapolation for a Class of Structured Nonsmooth DC Minimization. *Math. Program.*, Ser. B, 176(1-2):369–401, 2018.
16. Z. S. Lu, Z. R. Zhou, and Z. Sun. Nonmonotone enhanced proximal DC algorithms for a class of structured nonsmooth DC programming. *SIAM J. Optim.*, 29(4):2725–2752, 2019.
17. Z. Q. Luo and P. Tseng. Error bound and convergence analysis of matrix splitting algorithms for the affine variational inequality problem. *SIAM J. Optim.*, 2(1):43–54, 1992.
18. Z. Q. Luo and P. Tseng. On linear convergence of descent methods for convex essentially smooth minimization. *SIAM J. Control Optim.*, 30(2):408–425, 1992.
19. Z. Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Ann. Oper. Res.*, 46–47(1):157–178, 1993.
20. S. Nakayama, J. Y. Gotoh. On the superiority of PGMs to PDCAs in nonsmooth nonconvex sparse regression. *Optim. Lett.*, 15:2831–2860, 2021.
21. J. S. Pang, M. Razaviyayn, and A. Alvarado. Computing B-stationary points of nonsmooth DC programs. *Math. Oper. Res.*, 42(1):95–118, 2017.
22. H. A. Le Thi and T. Pham Dinh. The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Ann. Oper. Res.*, 133(1-4):23–46, 2005.
23. H. A. Le Thi and T. Pham Dinh. DC programming and DCA: thirty years of developments. *Math. Program. Ser. B*, 169(1):5–68, 2018.
24. H. A. Le Thi, V. N. Huynh, and T. Pham Dinh. Convergence analysis of DC algorithm for DC programming with subanalytic data. *J. Optim. Theory Appl.*, 179(1):103–126, 2018.
25. P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Math. Program. Ser. B*, 117(1-2):387–423, 2009.
26. Y. Wang, Z. Luo and X. Zhang. New improved penalty methods for sparse reconstruction based on difference of two norms. https://doi.org/10.13140/RG.2.1.3256.3369.
27. B. Wen, X. Chen, and T. K. Pong. Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. *SIAM J. Optim.*, 27(1):124–145, 2017.
28. B. Wen, X. Chen, and T. K. Pong. A proximal difference-of-convex algorithm with extrapolation. *Comput. Optim. Appl.*, 69(2):297–324, 2018.
29. C. H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942, 2010.
30. Z. Zhou and A. M. C. So. A unified approach to error bounds for structured convex optimization problems. *Math. Program., Ser. A*, 165(2):689–728, 2017.
31. R. T. Rockafellar and R. J.-B. Wets. Variational Analysis, Springer, 1998.
32. M. Razaviyayn, M. Hong and Z. Q. Luo, A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM J. Optim.*, 23(2):1126–1153, 2013.