This is a post-peer-review, pre-copyedit version of an article published in Computational Optimization and Applications. The final authenticated version is available online at: http://dx.doi.org/10.1007/s10589-017-9954-1

A Proximal Difference-of-convex Algorithm with Extrapolation

Bo Wen * Xiaojun Chen † Ting Kei Pong ‡

September 20, 2017

Abstract

We consider a class of difference-of-convex (DC) optimization problems whose objective is level-bounded and is the sum of a smooth convex function with Lipschitz gradient, a proper closed convex function and a continuous concave function. While this kind of problems can be solved by the classical difference-of-convex algorithm (DCA) [26], the difficulty of the subproblems of this algorithm depends heavily on the choice of DC decomposition. Simpler subproblems can be obtained by using a specific DC decomposition described in [27]. This decomposition has been proposed in numerous work such as [18], and we refer to the resulting DCA as the proximal DCA. Although the subproblems are simpler, the proximal DCA is the same as the proximal gradient algorithm when the concave part of the objective is void, and hence is potentially slow in practice. In this paper, motivated by the extrapolation techniques for accelerating the proximal gradient algorithm in the convex settings, we consider a proximal difference-of-convex algorithm with extrapolation to possibly accelerate the proximal DCA. We show that any cluster point of the sequence generated by our algorithm is a stationary point of the DC optimization problem for a fairly general choice of extrapolation parameters: in particular, the parameters can be chosen as in FISTA with fixed restart [15]. In addition, by assuming the Kurdyka-Łojasiewicz property of the objective and the differentiability of the concave part, we establish global convergence of the sequence generated by our algorithm and analyze its convergence rate. Our numerical experiments on two difference-of-convex regularized least squares models show that our algorithm usually outperforms the proximal DCA and the general iterative shrinkage and thresholding algorithm proposed in [17].

Keywords: difference-of-convex problems, nonconvex, nonsmooth, extrapolation, Kurdyka-Łojasiewicz inequality

AMS subject classifications. 90C30, 65K05, 90C26

1 Introduction

Difference-of-convex (DC) optimization problems are problems whose objective can be written as the difference of a proper closed convex function and a continuous convex function. They arise in various applications such as digital communication system [2], assignment and power allocation [29] and compressed sensing [35]; we refer the readers to Sections 7.6 to 7.8 of the recent monograph [33] for more applications of DC optimization problems.

*Department of Mathematics, Harbin Institute of Technology, Harbin, P.R. China. Current address: Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, P.R. China. E-mail: bo.wen@connect.polyu.hk.
†Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, P.R. China. This author's work was supported in part by Hong Kong Research Grants Council PolyU153000/15p. E-mail: maxjchen@polyu.edu.hk.
‡Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, P.R. China. This author's work was supported in part by Hong Kong Research Grants Council PolyU253008/15p. E-mail: tk.pong@polyu.edu.hk.

A classical algorithm for solving DC optimization problems is the so-called DC algorithm (DCA), which was proposed by Tao and An [26]; see also [6, 18, 30, 31, 32] for more recent developments.[1] In each iteration, this algorithm replaces the concave part of the objective by a linear majorant and solves the resulting convex optimization problem. The difficulty of the subproblems involved relies heavily on the choice of DC decomposition of the objective function. When the objective can be written as the sum of a smooth convex function with Lipschitz gradient, a proper closed convex function and a continuous concave function, simpler subproblems can be obtained by using a specific DC decomposition described in [27, Eq. 16]. This idea appears in numerous work and is also recently adopted in [18], where they proposed the so-called proximal DCA.[2] This algorithm not only majorizes the concave part in the objective by a linear majorant in each iteration, but also majorizes the smooth convex part by a quadratic majorant. When the proximal mapping of the proper closed convex function is easy to compute, the subproblems of the proximal DCA can be solved efficiently. However, this algorithm may take a lot of iterations: indeed, when the concave part of the objective is void, the proximal DCA reduces to the proximal gradient algorithm for convex optimization problems, which can be slow in practice [15, Section 5].

It is then tempting to incorporate techniques to possibly accelerate the proximal DCA while not significantly increasing the computational cost per iteration. One such technique is to perform extrapolation. More precisely, this means adding *momentum* terms that involve previous iterates for updating the current iterate. Such technique has been adopted for convex optimization problems, dating back to Polyak's heavy ball method [25]. More recent examples of such techniques are Nesterov's extrapolation techniques [21, 22, 23, 24] which have been extensively used for accelerating the proximal gradient algorithm and its variants for convex optimization problems. One representative algorithm that incorporates these techniques is the fast iterative shrinkage-thresholding algorithm (FISTA) [7, 23]. It is known that the function values generated by FISTA converges at a rate of $O(1/k^2)$, which is faster than the $O(1/k)$ convergence rate of the proximal gradient algorithm. We refer the readers to [8, 15] for more examples of such algorithms.

In view of the success of extrapolation techniques in accelerating the proximal gradient algorithm for convex optimization problems, and noting that the proximal gradient algorithm and the proximal DCA are the same when applied to convex problems, in this paper, we incorporate extrapolation techniques to possibly accelerate the proximal DCA in the general DC settings.[3] We call our algorithm the proximal DCA with extrapolation (pDCA$_e$). We prove that, for a fairly general choice of extrapolation parameters, if the objective is level-bounded, then any cluster point of the sequence generated by our algorithm is a stationary point of the DC optimization problem. The choice of parameters is general enough to cover those used in FISTA with fixed restart [15]. Additionally, by assuming that the objective is a level-bounded Kurdyka-Łojasiewicz function (see, for example, [4]) and the concave part is differentiable with a locally Lipschitz gradient, we establish global convergence of the whole sequence generated by our algorithm. We also analyze the convergence rate based on the Kurdyka-Łojasiewicz exponent. Finally, we perform numerical experiments on ℓ_{1-2} [35] and logarithmic [12] regularized least squares problems. Our numerical experiments show that the pDCA$_e$ usually outperforms the proximal DCA and the general iterative shrinkage and thresholding algorithm (GIST) proposed in [17].

The rest of this paper is organized as follows. In Section 2, we introduce notation and discuss some preliminary materials. In Section 3, we describe the DC optimization problem we study in this paper and present our algorithm pDCA$_e$. The convergence of the sequence generated by the algorithm and the convergence rate are studied in Section 4. Finally, we present numerical experiments in Section 5.

[1] We would also like to point to the article "DC programming and DCA" on the person webpage of Le Thi Hoai An: `http://www.lita.univ-lorraine.fr/~lethi/index.php/en/research/dc-programming-and-dca.html`

[2] This algorithm was called "the proximal difference-of-convex decomposition algorithm" in [18]. As noted in [18], their algorithm is the DCA applied to a specific DC decomposition.

[3] It is also discussed at the end of the numerical section of [18] that suitably incorporating extrapolation techniques into the proximal DCA can accelerate the algorithm empirically.

2 Notation and preliminaries

In this paper, we use $\mathbb{R}^n$ to denote the n-dimensional Euclidean space with inner product $\langle\cdot,\cdot\rangle$ and Euclidean norm $\|\cdot\|$, and use $\|\cdot\|_1$ and $\|\cdot\|_\infty$ to denote the ℓ_1 norm and the ℓ_∞ norm, respectively. Given a matrix $A\in\mathbb{R}^{m\times n}$, the transpose of A is denoted by A^T. Moreover, for a symmetric matrix $A\in\mathbb{R}^{n\times n}$, we use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote its largest and smallest eigenvalues, respectively. In addition, for a nonempty closed set $\mathcal{C}\subseteq\mathbb{R}^n$, we denote the distance from a point $x\in\mathbb{R}^n$ to $\mathcal{C}$ by $\operatorname{dist}(x,\mathcal{C}):=\inf_{y\in\mathcal{C}}\|x-y\|$.

For an extended-real-valued function $h:\mathbb{R}^n\to[-\infty,\infty]$, we denote its domain by $\operatorname{dom}h=\{x\in\mathbb{R}^n:\ h(x)<\infty\}$. The function h is said to be proper if it never equals $-\infty$ and $\operatorname{dom}h\neq\emptyset$. Moreover, a proper function is closed if it is lower semicontinuous. A proper closed function h is said to be level-bounded if the lower level sets of h (i.e., $\{x\in\mathbb{R}^n:\ h(x)\le r\}$ for any $r\in\mathbb{R}$) are bounded. Given a proper closed function $h:\mathbb{R}^n\to\mathbb{R}\cup\{\infty\}$, the (limiting) subdifferential of h at $x\in\operatorname{dom}h$ is given by

$$\partial h(x)=\left\{v\in\mathbb{R}^n:\exists\ x^t\xrightarrow{h}x,v^t\to v\text{ with }\liminf_{y\to x^t}\frac{h(y)-h(x^t)-\langle v^t,y-x^t\rangle}{\|y-x^t\|}\ge 0\ \text{ for each }t\right\},\qquad(2.1)$$

where $z\xrightarrow{h}x$ means $z\to x$ and $h(z)\to h(x)$. We also write $\operatorname{dom}\partial h:=\{x\in\mathbb{R}^n:\ \partial h(x)\neq\emptyset\}$. It is known that the above subdifferential reduces to the classical subdifferential in convex analysis when h is convex, i.e.,

$$\partial h(x)=\{v\in\mathbb{R}^n:\ h(u)-h(x)-\langle v,u-x\rangle\ge 0,\ \forall u\in\mathbb{R}^n\};$$

see, for example, [28, Proposition 8.12]. In addition, if h is continuously differentiable, then the subdifferential (2.1) reduces to the gradient of h denoted by ∇h. We also use $\nabla_i h$ to denote the partial gradient of h with respect to x_i, the i-th component of x.

We next recall the Kurdyka-Łojasiewicz (KL) property [3, 4, 5, 10], which is satisfied by a wide variety of functions such as proper closed semialgebraic functions, and plays an important role in the convergence analysis of many first-order methods; see, for example, [4, 5].

Definition 2.1. (KL property) *A proper closed function h is said to satisfy the KL property at $\hat{x}\in\operatorname{dom}\partial h$ if there exist $a\in(0,\infty]$, a neighborhood $\mathcal{O}$ of $\hat{x}$, and a continuous concave function $\phi:[0,a)\to\mathbb{R}_+$ with $\phi(0)=0$ such that:*

(i) *ϕ is continuously differentiable on $(0,a)$ with $\phi'>0$;*

(ii) *For any $x\in\mathcal{O}$ with $h(\hat{x})<h(x)<h(\hat{x})+a$, one has*

$$\phi'(h(x)-h(\hat{x}))\operatorname{dist}(0,\partial h(x))\ge 1.\qquad(2.2)$$

A proper closed function h satisfying the KL property at all points in $\operatorname{dom}\partial h$ is called a KL function.

We also recall the following result proved in [11, Lemma 6] concerning the uniformized KL property. For notational simplicity, we use Ξ_a to denote the set of all concave continuous functions $\phi:[0,a)\to\mathbb{R}_+$ that are continuously differentiable on $(0,a)$ with positive derivatives and satisfy $\phi(0)=0$.

Lemma 2.1. (Uniformized KL property) *Suppose that h is a proper closed function and let Γ be a compact set. If h is a constant on Γ and satisfies the KL property at each point of Γ, then there exist $\epsilon,a>0$ and $\phi\in\Xi_a$ such that*

$$\phi'(h(x)-h(\hat{x}))\operatorname{dist}(0,\partial h(x))\ge 1$$

for any $\hat{x}\in\Gamma$ and any x satisfying $\operatorname{dist}(x,\Gamma)<\epsilon$ and $h(\hat{x})<h(x)<h(\hat{x})+a$.

3 Problem formulation and the proximal difference-of-convex algorithm with extrapolation

In this section, we describe the optimization problem we study in this paper and present our proximal difference-of-convex algorithm with extrapolation (pDCA$_e$).

We focus on problems of the following form:

$$v := \min_{x\in\mathbb{R}^n} F(x) := \; f(x) + P(x), \tag{3.1}$$

where f is a smooth convex function with a Lipschitz continuous gradient whose Lipschitz continuity modulus is $L > 0$, and

$$P(x) = P_1(x) - P_2(x),$$

with P_1 being a proper closed convex function and P_2 being a *continuous* convex function. We assume in addition that F is level-bounded. This latter assumption implies that $v > -\infty$ and that the set of global minimizers of (3.1) is nonempty. Problem (3.1) arises in applications such as compressed sensing, where f is typically the data fitting term such as the least squares loss function, and P is a nonsmooth regularizer for inducing desirable structures in the solution. We refer the readers to [1, 9, 16, 35, 36, 37] for concrete examples.

It is clear that problem (3.1) is a DC optimization problem and can be solved by the renowned DCA. However, as noted in the introduction, the difficulty of the subproblems involved in the DCA depends on the DC decomposition used. Indeed, when decomposing F naturally as the difference of $f + P_1$ and P_2, the subproblems of the corresponding DCA take the following form:

$$x^{t+1} \in \operatorname*{Arg\,min}_{x\in\mathbb{R}^n} \left\{ f(x) + P_1(x) - \langle \xi^t, x\rangle \right\}, \tag{3.2}$$

where $\xi^t \in \partial P_2(x^t)$. Although these problems are convex, they do not necessarily have closed form/simple solutions. On the other hand, simpler subproblems can be obtained via a specific DC decomposition described in [27, Eq. 16] and many other related papers such as [18], i.e.,

$$F(x) = \left(\frac{L}{2}\|x\|^2 + P_1(x)\right) - \left(\frac{L}{2}\|x\|^2 - f(x) + P_2(x)\right),$$

and we refer to the resulting DCA as the proximal DCA. When applied to solving (3.1), the subproblems of the proximal DCA take the following form:

$$\begin{aligned} x^{t+1} &= \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \left\{ \langle \nabla f(x^t) - \xi^t, x\rangle + \frac{L}{2}\|x - x^t\|^2 + P_1(x) \right\} \\ &= \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \left\{ \frac{L}{2} \left\| x - \left(x^t - \frac{1}{L}[\nabla f(x^t) - \xi^t] \right) \right\|^2 + P_1(x) \right\}, \end{aligned} \tag{3.3}$$

where $\xi^t \in \partial P_2(x^t)$, and x^{t+1} is uniquely defined because P_1 is proper closed convex. In contrast to (3.2), solving the subproblem (3.3) amounts to evaluating the so-called proximal operator of $\frac{1}{L}P_1$, and this proximal operator is easy to compute for a wide variety of P_1; see, for example, [14, Tables 10.1 and 10.2].

Despite having simple subproblems for many commonly used P_1, the proximal DCA is potentially slow: this is because the proximal DCA is the same as the proximal gradient algorithm when $P_2 = 0$ and the proximal gradient algorithm can take a lot of iterations in practice [15, Section 5]. Fortunately, the proximal gradient algorithm for convex problems (i.e., when $P_2 = 0$) has been successfully accelerated by various extrapolation techniques [21, 22, 23, 24]. Thus, it is tempting to incorporate extrapolation techniques into the proximal DCA to possibly accelerate the algorithm. Specifically, we consider the following algorithm for solving the DC optimization problem (3.1):

Proximal difference-of-convex algorithm with extrapolation (pDCA$_e$):

Input: $x^0 \in \operatorname{dom} P_1$, $\{\beta_t\} \subseteq [0,1)$ with $\sup_t \beta_t < 1$. Set $x^{-1} = x^0$.

for $t = 0, 1, 2, \cdots$

Take any $\xi^t \in \partial P_2(x^t)$ and set

$$
\begin{aligned}
y^t &= x^t + \beta_t(x^t - x^{t-1}),\\
x^{t+1} &= \operatorname*{arg\,min}_{y \in \mathbb{R}^n} \left\{ \langle \nabla f(y^t) - \xi^t, y \rangle + \frac{L}{2}\|y - y^t\|^2 + P_1(y) \right\}.
\end{aligned}
\tag{3.4}
$$

end for

In view of the algorithmic framework of pDCA$_e$ and the subproblem (3.3) in the proximal DCA, it is not hard to see that pDCA$_e$ reduces to the proximal DCA when $\beta_t \equiv 0$. Hence, the proximal DCA is a special case of pDCA$_e$. In addition, we would like to point out that the conditions on $\{\beta_t\}$ in pDCA$_e$ (i.e., $\{\beta_t\} \subseteq [0,1)$ and $\sup_t \beta_t < 1$) are general enough to cover many popular choices of extrapolation parameters including those used in FISTA with fixed restart or FISTA with both fixed and adaptive restart for solving (3.1) with $P_2 = 0$ [15]. In detail, in these schemes, one starts with $\theta_{-1} = \theta_0 = 1$, recursively defines for $t \geq 0$ that

$$
\beta_t = \frac{\theta_{t-1} - 1}{\theta_t} \quad \text{with} \quad \theta_{t+1} = \frac{1 + \sqrt{1 + 4\theta_t^2}}{2}, \tag{3.5}
$$

and resets $\theta_{t-1} = \theta_t = 1$ for some $t > 0$ under suitable conditions: in the fixed restart scheme, one fixes a positive number $\bar{T}$ and resets $\theta_{t-1} = \theta_t = 1$ every $\bar{T}$ iterations, while the adaptive restart scheme amounts to resetting $\theta_{t-1} = \theta_t = 1$ whenever $\langle y^{t-1} - x^t, x^t - x^{t-1} \rangle > 0$. From these definitions, one can readily show by induction that the $\{\beta_t\}$ chosen as in FISTA with fixed restart or FISTA with both fixed and adaptive restart satisfies $\{\beta_t\} \subseteq [0,1)$ and $\sup_t \beta_t < 1$.[4] The choice of $\{\beta_t\}$ as in FISTA with both fixed and adaptive restart will be used in our numerical experiments in Section 5.

4 Convergence analysis

In this section, we study the convergence behavior of pDCA$_e$. We first establish the global subsequential convergence of pDCA$_e$. Then, by making an additional differentiability assumption on P_2 and assuming that the Kurdyka-Łojasiewicz property holds for an auxiliary function, we prove the global convergence of the whole sequence generated by pDCA$_e$ and analyze the rate of convergence.

4.1 Convergence analysis I: Global subsequential convergence of pDCA$_e$

We start with the following definition of stationary points; see, for example, [17, Remark 1]. It is routine to show that any local minimizer of F is a stationary point of F; see [26, Theorem 2(i)].

Definition 4.1. *Let F be given in* (3.1)*. We say that $\bar{x}$ is a stationary point of F if*

$$
0 \in \nabla f(\bar{x}) + \partial P_1(\bar{x}) - \partial P_2(\bar{x}).
$$

The set of all stationary points of F is denoted by $\mathcal{X}$.

[4] Indeed, when $P_2 = 0$, FISTA with fixed restart and FISTA with both fixed and adaptive restart are special cases of pDCA$_e$.

We are now ready to prove a global subsequential convergence result for pDCA_e applied to solving (3.1). Recall that F in (3.1) is level-bounded, and the extrapolation parameters $\{\beta_t\}$ in pDCA_e satisfy $\sup_t \beta_t < 1$ and $\{\beta_t\} \subseteq [0, 1)$.

Theorem 4.1. (Global subsequential convergence of pDCA_e**)** *Let $\{x^t\}$ be a sequence generated by* pDCA_e *for solving* (3.1). *Then the following statements hold.*

(i) *The sequence $\{x^t\}$ is bounded.*

(ii) $\lim_{t\to\infty} \|x^{t+1} - x^t\| = 0$.

(iii) *Any accumulation point of $\{x^t\}$ is a stationary point of F.*

Proof. First we prove (i). We note from (3.4) that x^{t+1} is the global minimizer of a strongly convex function. Using this and comparing the objective values of this strongly convex function at x^{t+1} and x^t, we see immediately that

$$\begin{aligned}
&\langle \nabla f(y^t) - \xi^t, x^{t+1}\rangle + \frac{L}{2}\|x^{t+1} - y^t\|^2 + P_1(x^{t+1}) \\
&\le \langle \nabla f(y^t) - \xi^t, x^t\rangle + \frac{L}{2}\|x^t - y^t\|^2 + P_1(x^t) - \frac{L}{2}\|x^{t+1} - x^t\|^2.
\end{aligned} \tag{4.1}$$

On the other hand, using the fact that ∇f is Lipschitz continuous with a modulus of $L > 0$, we have

$$\begin{aligned}
&f(x^{t+1}) + P(x^{t+1}) \le f(y^t) + \langle \nabla f(y^t), x^{t+1} - y^t\rangle + \frac{L}{2}\|x^{t+1} - y^t\|^2 + P(x^{t+1}) \\
&= f(y^t) + \langle \nabla f(y^t), x^{t+1} - y^t\rangle + \frac{L}{2}\|x^{t+1} - y^t\|^2 + P_1(x^{t+1}) - P_2(x^{t+1}) \\
&\le f(y^t) + \langle \nabla f(y^t), x^{t+1} - y^t\rangle + \frac{L}{2}\|x^{t+1} - y^t\|^2 + P_1(x^{t+1}) - P_2(x^t) - \langle \xi^t, x^{t+1} - x^t\rangle \\
&\le f(y^t) + \langle \nabla f(y^t), x^t - y^t\rangle + \frac{L}{2}\|x^t - y^t\|^2 + P_1(x^t) - P_2(x^t) - \frac{L}{2}\|x^{t+1} - x^t\|^2 \\
&\le f(x^t) + P(x^t) + \frac{L}{2}\|x^t - y^t\|^2 - \frac{L}{2}\|x^{t+1} - x^t\|^2,
\end{aligned} \tag{4.2}$$

where the second inequality follows from the subgradient inequality and the fact that $\xi^t \in \partial P_2(x^t)$, the third inequality follows from (4.1), while the last inequality follows from the convexity of f and the definition of P. Now, invoking the definition of y^t, we obtain further from (4.2) that

$$f(x^{t+1}) + P(x^{t+1}) \le f(x^t) + P(x^t) + \frac{L}{2}\beta_t^2\|x^t - x^{t-1}\|^2 - \frac{L}{2}\|x^{t+1} - x^t\|^2.$$

Consequently, we have upon rearranging terms that

$$\frac{L}{2}(1 - \beta_t^2)\|x^t - x^{t-1}\|^2 \le \left[f(x^t) + P(x^t) + \frac{L}{2}\|x^t - x^{t-1}\|^2\right] - \left[f(x^{t+1}) + P(x^{t+1}) + \frac{L}{2}\|x^{t+1} - x^t\|^2\right]. \tag{4.3}$$

Since $\{\beta_t\} \subset [0, 1)$, we deduce from (4.3) that the sequence $\{f(x^t) + P(x^t) + \frac{L}{2}\|x^t - x^{t-1}\|^2\}$ is nonincreasing. This together with the fact that $x^0 = x^{-1}$ gives

$$f(x^t) + P(x^t) \le f(x^t) + P(x^t) + \frac{L}{2}\|x^t - x^{t-1}\|^2 \le f(x^0) + P(x^0)$$

for all $t \ge 0$, which shows that $\{x^t\}$ is bounded, thanks to the level-boundedness of $f + P$. This proves (i).

Next we prove (ii). Summing both sides of (4.3) from $t = 0$ to ∞, we obtain that

$$\frac{L}{2}\sum_{t=0}^{\infty}(1-\beta_t^2)\|x^t - x^{t-1}\|^2 \le f(x^0) + P(x^0) - \liminf_{t\to\infty}\left[f(x^{t+1}) + P(x^{t+1}) + \frac{L}{2}\|x^{t+1} - x^t\|^2\right]$$
$$\le f(x^0) + P(x^0) - v < \infty.$$

Since $\sup_t \beta_t < 1$, we deduce immediately from the above relation that $\lim_{t\to\infty}\|x^{t+1} - x^t\| = 0$. This proves (ii).

Finally, let $\bar{x}$ be an accumulation point of $\{x^t\}$ and let $\{x^{t_i}\}$ be a subsequence such that $\lim_{i\to\infty} x^{t_i} = \bar{x}$. Then, from the first-order optimality condition of the subproblem (3.4), we have

$$-L(x^{t_i+1} - y^{t_i}) \in \partial P_1(x^{t_i+1}) + \nabla f(y^{t_i}) - \xi^{t_i}.$$

Using this together with the fact that $y^{t_i} = x^{t_i} + \beta_{t_i}(x^{t_i} - x^{t_i-1})$, we obtain further that

$$-L[(x^{t_i+1} - x^{t_i}) - \beta_{t_i}(x^{t_i} - x^{t_i-1})] \in \partial P_1(x^{t_i+1}) + \nabla f(y^{t_i}) - \xi^{t_i}. \tag{4.4}$$

In addition, note that the sequence $\{\xi^{t_i}\}$ is bounded due to the continuity and convexity of P_2 and the boundedness of $\{x^{t_i}\}$. Thus, by passing to a further subsequence if necessary, we may assume without loss of generality that $\lim_{i\to\infty}\xi^{t_i}$ exists, which belongs to $\partial P_2(\bar{x})$ due to the closedness of ∂P_2. Using this and invoking $\|x^{t_i+1} - x^{t_i}\| \to 0$ from (ii) together with the closedness of ∂P_1 and the continuity of ∇f, we have upon passing to the limit in (4.4) that

$$0 \in \partial P_1(\bar{x}) + \nabla f(\bar{x}) - \partial P_2(\bar{x}).$$

This completes the proof. □

We next study the behavior of $\{F(x^t)\}$ for a sequence $\{x^t\}$ generated by pDCA$_e$. The result will subsequently be used in establishing global convergence of the whole sequence $\{x^t\}$ under additional assumptions in the next subsection.

Proposition 4.1. *Let $\{x^t\}$ be a sequence generated by* pDCA$_e$ *for solving* (3.1). *Then the following statements hold.*

(i) $\zeta := \lim_{t\to\infty} F(x^t)$ *exists.*

(ii) $F \equiv \zeta$ *on Ω, where Ω is the set of accumulation points of $\{x^t\}$.*

Proof. Since $\{\beta_t\} \subseteq [0, 1)$, we see immediately from (4.3) that the sequence $\{F(x^t) + \frac{L}{2}\|x^t - x^{t-1}\|^2\}$ is nonincreasing. In addition, this sequence is also bounded below by v. Furthermore, we recall from Theorem 4.1(ii) that $\|x^{t+1}-x^t\| \to 0$. The conclusion that $\zeta := \lim_{t\to\infty} F(x^t)$ exists now follows immediately from the aforementioned facts. This proves (i).

Now we prove (ii). We first note from Theorem 4.1(i) and (iii) that $\emptyset \ne \Omega \subseteq \mathcal{X}$. Take any $\hat{x} \in \Omega$. By the definition of accumulation point, there exists a convergent subsequence $\{x^{t_i}\}$ such that $\lim_{i\to\infty} x^{t_i} = \hat{x}$. Since x^{t_i} is the minimizer of the subproblem (3.4), we see that

$$P_1(x^{t_i}) + \langle \nabla f(y^{t_i-1}) - \xi^{t_i-1}, x^{t_i}\rangle + \frac{L}{2}\|x^{t_i} - y^{t_i-1}\|^2 \le P_1(\hat{x}) + \langle \nabla f(y^{t_i-1}) - \xi^{t_i-1}, \hat{x}\rangle + \frac{L}{2}\|\hat{x} - y^{t_i-1}\|^2.$$

Rearranging terms, we obtain further that

$$P_1(x^{t_i}) + \langle \nabla f(y^{t_i-1}) - \xi^{t_i-1}, x^{t_i} - \hat{x}\rangle + \frac{L}{2}\|x^{t_i} - y^{t_i-1}\|^2 \le P_1(\hat{x}) + \frac{L}{2}\|\hat{x} - y^{t_i-1}\|^2. \tag{4.5}$$

On the other hand, observe that

$$\|\hat{x} - y^{t_i-1}\| = \|\hat{x} - x^{t_i} + x^{t_i} - y^{t_i-1}\| \le \|\hat{x} - x^{t_i}\| + \|x^{t_i} - y^{t_i-1}\| \tag{4.6}$$

and that

$$\begin{aligned} \|x^{t_i} - y^{t_i-1}\| &= \|x^{t_i} - x^{t_i-1} - \beta_{t_i-1}(x^{t_i-1} - x^{t_i-2})\| \\ &\le \|x^{t_i} - x^{t_i-1}\| + \|x^{t_i-1} - x^{t_i-2}\|, \end{aligned} \tag{4.7}$$

where we made use of the fact that $y^{t_i-1} = x^{t_i-1} + \beta_{t_i-1}(x^{t_i-1} - x^{t_i-2})$ for the equality. Since $\|x^{t+1} - x^t\| \to 0$ from Theorem 4.1(ii) and $\lim_{i\to\infty} x^{t_i} = \hat{x}$, we have by passing to the limits in (4.6) and (4.7) that

$$\|\hat{x} - y^{t_i-1}\| \to 0 \quad \text{and} \quad \|x^{t_i} - y^{t_i-1}\| \to 0. \tag{4.8}$$

In addition, notice that the sequence $\{\xi^{t_i}\}$ is bounded, thanks to the convexity and continuity of P_2 and the fact that $\lim_{i\to\infty} x^{t_i} = \hat{x}$. Using this and (4.8), we obtain further that

$$\begin{aligned} \zeta &= \lim_{i\to\infty} f(x^{t_i}) + P(x^{t_i}) \\ &= \lim_{i\to\infty} f(x^{t_i}) + P(x^{t_i}) + \langle \nabla f(y^{t_i-1}) - \xi^{t_i-1}, x^{t_i} - \hat{x}\rangle + \frac{L}{2}\|x^{t_i} - y^{t_i-1}\|^2 \\ &\le \limsup_{i\to\infty} f(x^{t_i}) + P_1(\hat{x}) - P_2(x^{t_i}) + \frac{L}{2}\|\hat{x} - y^{t_i-1}\|^2 = F(\hat{x}), \end{aligned}$$

where the inequality follows from (4.5) and the definition of P. Finally, since F is lower semicontinuous, we also have

$$F(\hat{x}) \le \liminf_{i\to\infty} F(x^{t_i}) = \lim_{i\to\infty} F(x^{t_i}) = \zeta.$$

Consequently, $F(\hat{x}) = \lim_{i\to\infty} F(x^{t_i}) = \zeta$. Since $\hat{x} \in \Omega$ is arbitrary, we conclude that $F \equiv \zeta$ on Ω. This completes the proof. □

4.2 Convergence analysis II: Global convergence and convergence rate of the pDCA$_e$

In this subsection, we consider the global convergence property of the whole sequence $\{x^t\}$ generated by pDCA$_e$ for solving (3.1) and establish the convergence rate of $\{x^t\}$ under suitable conditions. We start by introducing the following assumption.

Assumption 4.1. *The function P_2 in* (3.1) *is continuously differentiable on an open set $\mathcal{N}_0$ that contains $\mathcal{X}$. Moreover, the gradient ∇P_2 is locally Lipschitz continuous on $\mathcal{N}_0$.*

While Assumption 4.1 may look restrictive at first glance, it is satisfied by many DC regularizers $P(x)$ that arise in applications. We present some concrete examples below.

Example 4.1. We consider the least squares problem with ℓ_{1-2} regularization [35], which takes the following form

$$\min_{x\in\mathbb{R}^n} F_{\ell_{1-2}}(x) = \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1 - \lambda\|x\|, \tag{4.9}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and $\lambda > 0$. We also assume that A does not have zero columns so that $F_{\ell_{1-2}}$ is level-bounded (see [35, Lemma 3.1] and [20, Example 4.1(b)]). This model corresponds to (3.1) with $f(x) = \frac{1}{2}\|Ax - b\|^2$, $P_1(x) = \lambda\|x\|_1$ and $P_2(x) = \lambda\|x\|$.

We claim that if $2\lambda < \|A^T b\|_\infty$, then 0 is not a stationary point of $F_{\ell_{1-2}}$. Suppose to the contrary that $0 \in \mathcal{X}$, then we have from the definition of stationary point that $A^T b \in \lambda\partial\|0\|_1 - \lambda\partial\|0\|$, which is equivalent to

$$A^T b \in \lambda[-1,1]^n - \lambda B(0,1),$$

where $B(0,1) = \{x \in \mathbb{R}^n : \|x\| \le 1\}$. From this, we see that $\|A^T b\|_\infty \le 2\lambda$, which is a contradiction.

Hence, if $\lambda < \frac{1}{2}\|A^T b\|_\infty$, then $\mathcal{X}$ does not contain 0. Since $\mathcal{X}$ is closed, one can then construct an open set $\mathcal{N}_0$ containing $\mathcal{X}$ so that P_2 is continuously differentiable with locally Lipschitz gradient on $\mathcal{N}_0$. Thus, Assumption 4.1 is satisfied for (4.9) when $\lambda < \frac{1}{2}\|A^T b\|_\infty$.

Example 4.2. We consider the minmax concave penalty (MCP) regularization [36], whose DC decomposition is given in [17]:

$$P(x) = \lambda \sum_{i=1}^n \int_0^{|x_i|} \left[1 - \frac{x}{\theta\lambda}\right]_+ dx = \lambda\|x\|_1 - \underbrace{\lambda \sum_{i=1}^n \int_0^{|x_i|} \min\left\{1, \frac{x}{\theta\lambda}\right\} dx}_{P_2(x)},$$

where $\theta > 0$ is a constant, $\lambda > 0$ is the regularization parameter and $[x]_+ = \max\{0, x\}$. It is routine to show that P_2 is continuously differentiable and

$$\nabla_i P_2(x) = \lambda\, \mathrm{sign}(x_i) \min\{1, |x_i|/(\theta\lambda)\}.$$

Moreover, the gradient ∇P_2 is Lipschitz continuous with modulus $\frac{1}{\theta}$.

Example 4.3. We consider the smoothly clipped absolute deviation (SCAD) regularization [16], whose DC decomposition is given in [17]:

$$P(x) = \lambda \sum_{i=1}^n \int_0^{|x_i|} \min\left\{1, \frac{[\theta\lambda - x]_+}{(\theta-1)\lambda}\right\} dx = \lambda\|x\|_1 - \underbrace{\lambda \sum_{i=1}^n \int_0^{|x_i|} \frac{[\min\{\theta\lambda, x\} - \lambda]_+}{(\theta-1)\lambda} dx}_{P_2(x)},$$

where $\lambda > 0$ is the regularization parameter and $\theta > 2$ is a constant. It is routine to show that P_2 is continuously differentiable with

$$\nabla_i P_2(x) = \mathrm{sign}(x_i) \frac{[\min\{\theta\lambda, |x_i|\} - \lambda]_+}{\theta - 1}.$$

Thus it is routine to show that $\frac{1}{\theta-1}$ is a Lipschitz continuity modulus of ∇P_2.

Example 4.4. We consider the transformed ℓ_1 regularization [37], whose DC decomposition is given in [1]:

$$P(x) = \sum_{i=1}^n \frac{(a+1)|x_i|}{a+|x_i|} = \frac{a+1}{a}\|x\|_1 - \underbrace{\sum_{i=1}^n \left[\frac{a+1}{a}|x_i| - \frac{(a+1)|x_i|}{a+|x_i|}\right]}_{P_2(x)},$$

where $a > 0$. It was shown in [1, Section 5.4] that $P_2(x)$ is continuously differentiable with a Lipschitz continuous gradient whose Lipschitz continuity modulus is $\frac{2(a+1)}{a^2}$.

Example 4.5. The last regularization function we consider is the logarithmic penalty function [12], whose DC decomposition is given in [17]:

$$P(x) = \sum_{i=1}^n [\lambda \log(|x_i| + \epsilon) - \lambda \log \epsilon] = \frac{\lambda}{\epsilon}\|x\|_1 - \underbrace{\sum_{i=1}^n \lambda \left[\frac{|x_i|}{\epsilon} - \log(|x_i| + \epsilon) + \log \epsilon\right]}_{P_2(x)},$$

where λ and ϵ are positive numbers. One can see that $P_2(x)$ is continuously differentiable with a Lipschitz continuous gradient whose Lipschitz continuity modulus is $\frac{\lambda}{\epsilon^2}$.

We next present our global convergence analysis. We will show that the sequence $\{x^t\}$ generated by pDCA$_e$ is convergent to a stationary point of F under suitable assumptions. Our analysis follows a similar line of arguments to other convergence analysis based on KL property (see, for example, [3, 4, 5, 6]), but has to make extensive use of the following auxiliary function:

$$E(x,y) = f(x) + P(x) + \frac{L}{2}\|x-y\|^2. \tag{4.10}$$

Theorem 4.2. (Global convergence of pDCA$_e$**)** *Suppose that Assumption 4.1 holds and E is a KL function. Let $\{x^t\}$ be a sequence generated by* pDCA$_e$ *for solving* (3.1)*. Then the following statements hold.*

(i) $\lim\limits_{t\to\infty} \operatorname{dist}((0,0), \partial E(x^t, x^{t-1})) = 0.$

(ii) *The sequence $\{E(x^t, x^{t-1})\}$ is nonincreasing and $\lim\limits_{t\to\infty} E(x^t, x^{t-1}) = \zeta$, where ζ is given in Proposition 4.1.*

(iii) *The set of accumulation points of $\{(x^t, x^{t-1})\}$ is $\Upsilon := \{(x,x):\ x \in \Omega\}$ and $E \equiv \zeta$ on Υ, where Ω is the set of accumulation points of $\{x^t\}$.*

(iv) *The sequence $\{x^t\}$ converges to a stationary point of F; moreover, $\sum_{t=1}^{\infty} \|x^t - x^{t-1}\| < \infty$.*

Proof. From Theorem 4.1(i), we see that $\{x^t\}$ is bounded. This together with the definition of Ω implies that $\lim\limits_{t\to\infty} \operatorname{dist}(x^t, \Omega) = 0$. Also recall from Theorem 4.1(iii) that $\Omega \subseteq \mathcal{X}$. Thus, for any $\nu > 0$, there exists $T_0 > 0$ so that $\operatorname{dist}(x^t, \Omega) < \nu$ and $x^t \in \mathcal{N}_0$ whenever $t \ge T_0$, where $\mathcal{N}_0$ is the open set from Assumption 4.1. Moreover, since Ω is compact due to the boundedness of $\{x^t\}$, by shrinking ν if necessary, we may assume without loss of generality that ∇P_2 is globally Lipschitz continuous on the bounded set $\mathcal{N} := \{x \in \mathcal{N}_0 :\ \operatorname{dist}(x, \Omega) < \nu\}$.

Next, considering the subdifferential of the function E in (4.10) at the point (x^t, x^{t-1}) for $t \ge T_0$, we have

$$\partial E(x^t, x^{t-1}) = [\{\nabla f(x^t) - \nabla P_2(x^t) + L(x^t - x^{t-1})\} + \partial P_1(x^t)] \times \{-L(x^t - x^{t-1})\}, \tag{4.11}$$

where we made use of the definition of P, the facts that P_2 is continuously differentiable in $\mathcal{N}$ and that $x^t \in \mathcal{N}$ for $t \ge T_0$.

On the other hand, using the first-order optimality condition of the subproblem (3.4) in pDCA$_e$, we have for any $t \ge T_0 + 1$ that

$$-L(x^t - y^{t-1}) - \nabla f(y^{t-1}) + \nabla P_2(x^{t-1}) \in \partial P_1(x^t),$$

since P_2 is continuously differentiable in $\mathcal{N}$ and $x^{t-1} \in \mathcal{N}$ whenever $t \ge T_0 + 1$. Using this relation, we see further that

$$\begin{aligned}
&- L(x^{t-1} - y^{t-1}) + \nabla f(x^t) - \nabla f(y^{t-1}) + \nabla P_2(x^{t-1}) - \nabla P_2(x^t)\\
&= \nabla f(x^t) - \nabla P_2(x^t) + L(x^t - x^{t-1}) - L(x^t - y^{t-1}) - \nabla f(y^{t-1}) + \nabla P_2(x^{t-1})\\
&\in \nabla f(x^t) - \nabla P_2(x^t) + L(x^t - x^{t-1}) + \partial P_1(x^t).
\end{aligned}$$

Combining this with (4.11), we obtain

$$(-L(x^{t-1} - y^{t-1}) + \nabla f(x^t) - \nabla f(y^{t-1}) + \nabla P_2(x^{t-1}) - \nabla P_2(x^t), -L(x^t - x^{t-1})) \in \partial E(x^t, x^{t-1}).$$

Using this, the definition of y^t and the global Lipschitz continuity of ∇f and ∇P_2 on $\mathcal{N}$, we see that there exists $C > 0$ such that

$$\operatorname{dist}((0,0), \partial E(x^t, x^{t-1})) \le C(\|x^t - x^{t-1}\| + \|x^{t-1} - x^{t-2}\|) \tag{4.12}$$

whenever $t \geq T_0 + 1$. Since $\|x^{t+1} - x^t\| \to 0$ according to Theorem 4.1(ii), we conclude that

$$\lim_{t\to\infty} \operatorname{dist}((0,0), \partial E(x^t, x^{t-1})) = 0,$$

which proves (i).

We now prove (ii) and (iii). Using the fact that $\sup_t \beta_t < 1$, the definition of E and (4.3), we see that there exists a positive number D such that

$$E(x^t, x^{t-1}) - E(x^{t+1}, x^t) \geq D\|x^t - x^{t-1}\|^2 \tag{4.13}$$

for all t. In particular, the sequence $\{E(x^t, x^{t-1})\}$ is nonincreasing. Since this sequence is also bounded below by v, it is convergent. Next, in view of Theorem 4.1(ii) which says that $\|x^t - x^{t-1}\| \to 0$, it is not hard to show that the set of accumulation points of $\{(x^t, x^{t-1})\}_{t\geq 1}$ is Υ. Moreover,

$$\lim_{t\to\infty} E(x^t, x^{t-1}) = \zeta,$$

thanks to Proposition 4.1(i). Furthermore, for any $(\hat{x}, \hat{x}) \in \Upsilon$ so that $\hat{x} \in \Omega$, we have $E(\hat{x}, \hat{x}) = F(\hat{x}) = \zeta$, where the last equality follows from Proposition 4.1(ii). Since $\hat{x} \in \Omega$ is arbitrary, we conclude that $E \equiv \zeta$ on Υ. This proves (ii) and (iii).

Finally, we prove (iv). In view of Theorem 4.1(iii), it suffices to show that $\{x^t\}$ is convergent. We first consider the case that there exists a $t > 0$ such that $E(x^t, x^{t-1}) = \zeta$. Since $\{E(x^t, x^{t-1})\}$ is nonincreasing and convergent to ζ due to (ii), we conclude that for any $\bar{t} \geq 0$, $E(x^{t+\bar{t}}, x^{t+\bar{t}-1}) = \zeta$. Hence, we have from (4.13) that $x^t = x^{t+\bar{t}}$ for any $\bar{t} \geq 0$, meaning that $\{x^t\}$ converges finitely.

We next consider the case that $E(x^t, x^{t-1}) > \zeta$ for all t. Since E is a KL function, Υ is a compact subset of $\operatorname{dom} \partial E$ and $E \equiv \zeta$ on Υ, by Lemma 2.1, there exist an $\epsilon > 0$ and a continuous concave function $\phi \in \Xi_a$ with $a > 0$ such that

$$\phi'(E(x, y) - \zeta)\operatorname{dist}((0,0), \partial E(x, y)) \geq 1 \tag{4.14}$$

for all $(x, y) \in U$, where

$$U = \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n : \operatorname{dist}((x, y), \Upsilon) < \epsilon\} \cap \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n : \zeta < E(x, y) < \zeta + a\}.$$

Since Υ is the set of accumulation points of $\{(x^t, x^{t-1})\}_{t\geq 1}$ by (iii), and $\{x^t\}$ is bounded due to Theorem 4.1(i), we have

$$\lim_{t\to\infty} \operatorname{dist}((x^t, x^{t-1}), \Upsilon) = 0.$$

Hence, there exists $T_1 > 0$ such that $\operatorname{dist}((x^t, x^{t-1}), \Upsilon) < \epsilon$ whenever $t \geq T_1$. In addition, since the sequence $\{E(x^t, x^{t-1})\}$ is nonincreasing and convergent to ζ by (ii), there exists $T_2 > 0$ such that $\xi < E(x^t, x^{t-1}) < \xi + a$ for all $t \geq T_2$. Taking $\bar{T} = \max\{T_0 + 1, T_1, T_2\}$, then the sequence $\{(x^t, x^{t-1})\}_{t\geq \bar{T}}$ belongs to U. Hence we deduce from (4.14) that

$$\phi'(E(x^t, x^{t-1}) - \zeta) \cdot \operatorname{dist}((0,0), \partial E(x^t, x^{t-1})) \geq 1, \quad \text{for all } t \geq \bar{T}. \tag{4.15}$$

From the concavity of ϕ, we see further that for any $t \geq \bar{T}$,

$$\begin{aligned}
&\left[\phi(E(x^t, x^{t-1}) - \zeta) - \phi(E(x^{t+1}, x^t) - \zeta)\right] \cdot \operatorname{dist}((0,0), \partial E(x^t, x^{t-1})) \\
&\geq \phi'(E(x^t, x^{t-1}) - \zeta)) \cdot \operatorname{dist}((0,0), \partial E(x^t, x^{t-1})) \cdot (E(x^t, x^{t-1}) - E(x^{t+1}, x^t)) \\
&\geq E(x^t, x^{t-1}) - E(x^{t+1}, x^t),
\end{aligned}$$

where the last inequality holds due to (4.15) and the fact that $\{E(x^t, x^{t-1})\}$ is nonincreasing. Combining this with (4.12) and (4.13) and rearranging terms, we obtain that for any $t \geq \bar{T}$,

$$\|x^t - x^{t-1}\|^2 \leq \frac{C}{D}\left(\phi(E(x^t, x^{t-1}) - \zeta) - \phi(E(x^{t+1}, x^t) - \zeta)\right) \cdot \left(\|x^t - x^{t-1}\| + \|x^{t-1} - x^{t-2}\|\right). \tag{4.16}$$

Taking square root on both sides of (4.16) and using the AM-GM inequality, we have

$$\begin{aligned}\|x^t - x^{t-1}\| &\le \sqrt{\frac{2C}{D}\left(\phi(E(x^t,x^{t-1})-\zeta) - \phi(E(x^{t+1},x^t)-\zeta)\right)} \cdot \sqrt{\frac{\|x^t - x^{t-1}\| + \|x^{t-1}-x^{t-2}\|}{2}} \\ &\le \frac{C}{D}\left(\phi(E(x^t,x^{t-1})-\zeta) - \phi(E(x^{t+1},x^t)-\zeta)\right) + \frac{1}{4}\|x^t - x^{t-1}\| + \frac{1}{4}\|x^{t-1}-x^{t-2}\|,\end{aligned}$$

which implies that

$$\frac{1}{2}\|x^t - x^{t-1}\| \le \frac{C}{D}\left(\phi(E(x^t,x^{t-1})-\zeta) - \phi(E(x^{t+1},x^t)-\zeta)\right) + \frac{1}{4}(\|x^{t-1}-x^{t-2}\| - \|x^t - x^{t-1}\|). \quad (4.17)$$

Summing the above relation from $t = \bar{T}$ to ∞, we have

$$\sum_{t=\bar{T}}^{\infty} \|x^t - x^{t-1}\| \le \frac{2C}{D}\phi(E(x^{\bar{T}}, x^{\bar{T}-1}) - \zeta) + \frac{1}{2}\|x^{\bar{T}-1} - x^{\bar{T}-2}\| < \infty,$$

which implies the convergence of $\{x^t\}$ as well as the summability of $\{\|x^{t+1} - x^t\|\}_{t\ge 0}$. This completes the proof. □

Remark 4.1. *If the objective is not level bounded but we still have $v > -\infty$ (which can be true for least squares with regularizers in Examples 4.2, 4.3 and 4.4), we can still show that $\|x^t - x^{t-1}\| \to 0$ by following the same arguments as in the proof of Theorem 4.1(ii). Consequently, if the sequence $\{x^t\}$ also has an accumulation point, then using a similar proof as Theorem 4.1(iii), this accumulation point can be shown to be a stationary point of* (3.1).

We next consider the convergence rate of the sequence $\{x^t\}$ under the assumption that the auxiliary function E is a KL function whose $\phi \in \Xi_a$ (see Definition 2.1) takes the form $\phi(s) = cs^{1-\theta}$ for some $\theta \in [0, 1)$. This kind of convergence rate analysis has also been performed for other optimization algorithms; see, for example, [3]. Our analysis is similar to theirs but makes use of the auxiliary function E in (4.10).

Theorem 4.3. *Suppose that Assumption 4.1 holds. Let $\{x^t\}$ be a sequence generated by* pDCA$_e$ *for solving* (3.1) *and suppose that $\{x^t\}$ converges to some $\bar{x}$. Suppose further that E is a KL function with ϕ in the KL inequality* (2.2) *taking the form $\phi(s) = cs^{1-\theta}$ for some $\theta \in [0, 1)$ and $c > 0$. Then the following statements hold.*

(i) *If $\theta = 0$, then there exists $t_0 > 0$ so that x^t is constant for $t > t_0$;*

(ii) *If $\theta \in (0, \frac{1}{2}]$, then there exist $c_1 > 0$, $t_1 > 0$ and $\eta \in (0, 1)$ such that $\|x^t - \bar{x}\| < c_1\eta^t$ for $t > t_1$;*

(iii) *If $\theta \in (\frac{1}{2}, 1)$, then there exist $c_2 > 0$ and $t_2 > 0$ such that $\|x^t - \bar{x}\| < c_2 t^{-\frac{1-\theta}{2\theta-1}}$ for $t > t_2$.*

Proof. First, we prove (i). If $\theta = 0$, we claim that there must exist $t_0 > 0$ such that $E(x^{t_0}, x^{t_0-1}) = \zeta$. Suppose to the contrary that $E(x^t, x^{t-1}) > \zeta$ for all $t > 0$. Since $\lim\limits_{t\to\infty} x^t = \bar{x}$ and the sequence $\{E(x^t, x^{t-1})\}$ is nonincreasing and convergent to ζ by Theorem 4.2(ii), we have from $\phi(s) = cs$ and the KL inequality (4.15) that for all sufficiently large t,

$$\operatorname{dist}((0,0), \partial E(x^t, x^{t-1})) \ge \frac{1}{c},$$

which contradicts Theorem 4.2(i). Thus, there exists $t_0 > 0$ so that $E(x^{t_0}, x^{t_0-1}) = \zeta$. Since $\{E(x^t, x^{t-1})\}$ is nonincreasing and convergent to ζ, it must then hold that $E(x^{t_0+\bar{t}}, x^{t_0+\bar{t}-1}) = \zeta$ for any $\bar{t} \ge 0$. Thus, we conclude from (4.13) that $x^{t_0} = x^{t_0+\bar{t}}$ for any $\bar{t} \ge 0$. This proves (i).

We next turn to the case that $\theta \in (0,1)$. If there exists $t_0 > 0$ such that $E(x^{t_0}, x^{t_0-1}) = \zeta$, then one can show that $\{x^t\}$ is finitely convergent as above, and the desired conclusions hold trivially. Hence, for $\theta \in (0,1)$, we only need to consider the case when $E(x^t, x^{t-1}) > \zeta$ for all $t > 0$.

Define $H_t = E(x^t, x^{t-1}) - \zeta$ and $S_t = \sum_{i=t}^{\infty} \|x^{i+1} - x^i\|$, where S_t is well-defined due to Theorem 4.2(iv). Then, using (4.17), we have for any $t \geq \bar{T}$ (where $\bar{T}$ is defined as in (4.15)) that

$$
\begin{aligned}
S_t &= 2\sum_{i=t}^{\infty} \frac{1}{2}\|x^{i+1} - x^i\| \leq 2\sum_{i=t}^{\infty} \frac{1}{2}\|x^i - x^{i-1}\| \\
&\leq 2\sum_{i=t}^{\infty} \left[\frac{C}{D}\left(\phi(E(x^i, x^{i-1}) - \zeta) - \phi(E(x^{i+1}, x^i) - \zeta)\right) + \frac{1}{4}(\|x^{i-1} - x^{i-2}\| - \|x^i - x^{i-1}\|)\right] \\
&\leq \frac{2C}{D}\phi(E(x^t, x^{t-1}) - \zeta) + \frac{1}{2}\|x^{t-1} - x^{t-2}\| = \frac{2C}{D}\phi(H_t) + \frac{1}{2}(S_{t-2} - S_{t-1}).
\end{aligned}
$$

Using this and the fact that $\{S_t\}$ is nonincreasing, we obtain further that

$$
S_t \leq \frac{2C}{D}\phi(H_t) + \frac{1}{2}(S_{t-2} - S_t) \tag{4.18}
$$

for all $t \geq \bar{T}$. On the other hand, since $\lim_{t\to\infty} x^t = \bar{x}$ and the sequence $\{E(x^t, x^{t-1})\}$ is nonincreasing and convergent to ζ by Theorem 4.2(ii), we have from the KL inequality (4.15) with $\phi(s) = cs^{1-\theta}$ that for all sufficiently large t,

$$
c(1-\theta)(H_t)^{-\theta}\mathrm{dist}((0,0), \partial E(x^t, x^{t-1})) \geq 1. \tag{4.19}
$$

In addition, using (4.12) and the definition of S_t, we see that for all sufficiently large t,

$$
\mathrm{dist}((0,0), \partial E(x^t, x^{t-1})) \leq C(S_{t-2} - S_t). \tag{4.20}
$$

Combining (4.19) and (4.20), we have for all sufficiently large t that

$$
(H_t)^{\theta} \leq C \cdot c(1-\theta) \cdot (S_{t-2} - S_t).
$$

Raising to a power of $\frac{1-\theta}{\theta}$ to both sides of the above inequality and scaling both sides by c, we obtain that

$$
c(H_t)^{1-\theta} \leq c \cdot (C \cdot c(1-\theta) \cdot (S_{t-2} - S_t))^{\frac{1-\theta}{\theta}}.
$$

Combining this with (4.18) and recalling that $\phi(H_t) = c(H_t)^{1-\theta}$, we see that for all sufficiently large t,

$$
S_t \leq C_1(S_{t-2} - S_t)^{\frac{1-\theta}{\theta}} + \frac{1}{2}(S_{t-2} - S_t) \leq C_1(S_{t-2} - S_t)^{\frac{1-\theta}{\theta}} + S_{t-2} - S_t, \tag{4.21}
$$

where $C_1 = \frac{2C}{D}c \cdot (C \cdot c(1-\theta))^{\frac{1-\theta}{\theta}}$.

We now consider two cases: $\theta \in (0, \frac{1}{2}]$ or $\theta \in (\frac{1}{2}, 1)$.

Suppose first that $\theta \in (0, \frac{1}{2}]$. Then $\frac{1-\theta}{\theta} \geq 1$. Since $\|x^{t+1} - x^t\| \to 0$ from Theorem 4.1(ii), it holds that $S_{t-2} - S_t \to 0$. From these and (4.21), we conclude that there exists $t_1 > 0$ so that for all $t \geq t_1$, we have

$$
S_t \leq (C_1 + 1)(S_{t-2} - S_t),
$$

which implies that $S_t \leq \frac{C_1+1}{C_1+2}S_{t-2}$. Hence,

$$
\|x^t - \bar{x}\| \leq \sum_{i=t}^{\infty} \|x^{i+1} - x^i\| = S_t \leq S_{t_1-2}\left(\sqrt{\frac{C_1+1}{C_1+2}}\right)^{t-t_1+1}
$$

for all $t \geq t_1$. This proves (ii).

Finally, we consider the case that $\theta \in (\frac{1}{2}, 1)$. In this case, we have $\frac{1-\theta}{\theta} < 1$. Combining this with (4.21) and the fact that $S_{t-2} - S_t \to 0$, we see that there exists $t_2 > 0$ such that for all $t \geq t_2$, we have

$$\begin{aligned} S_t &\leq C_1(S_{t-2} - S_t)^{\frac{1-\theta}{\theta}} + S_{t-2} - S_t \\ &\leq C_1(S_{t-2} - S_t)^{\frac{1-\theta}{\theta}} + (S_{t-2} - S_t)^{\frac{1-\theta}{\theta}} \\ &= (C_1 + 1)(S_{t-2} - S_t)^{\frac{1-\theta}{\theta}}. \end{aligned}$$

Raising to a power of $\frac{\theta}{1-\theta}$ to both sides of the above inequality, we see further that,

$$S_t^{\frac{\theta}{1-\theta}} \leq C_2(S_{t-2} - S_t)$$

whenever $t \geq t_2$, where $C_2 = (C_1 + 1)^{\frac{\theta}{1-\theta}}$. Consider the sequence $\Delta_t := S_{2t}$. Then for any $t \geq \lceil \frac{t_2}{2} \rceil$, we have

$$\Delta_t^{\frac{\theta}{1-\theta}} \leq C_2(\Delta_{t-1} - \Delta_t).$$

Proceeding as in the proof of [3, Theorem 2] starting from [3, Equation (13)], one can show similarly that for all sufficiently large t,

$$\Delta_t \leq C_3 t^{-\frac{1-\theta}{2\theta-1}}$$

for some $C_3 > 0$; see the first equation on [3, Page 15]. This implies that for all sufficiently large t, we have

$$\|x^t - \bar{x}\| \leq S_t \begin{cases} = \Delta_{\frac{t}{2}} \leq 2^\rho C_3 t^{-\rho} & \text{if } t \text{ is even,} \\ \leq S_{t-1} = \Delta_{\frac{t-1}{2}} \leq 2^\rho C_3 (t-1)^{-\rho} \leq 4^\rho C_3 t^{-\rho} & \text{if } t \text{ is odd and } t \geq 2, \end{cases}$$

where $\rho := \frac{1-\theta}{2\theta-1}$. This completes the proof. □

Remark 4.2. *We recall that there are many concrete examples of functions f satisfying the KL property at all points in $\operatorname{dom} \partial f$ with $\phi(s) = cs^{1-\theta}$ for some $\theta \in [0, 1)$ and $c > 0$. Indeed, all proper closed semialgebraic functions satisfy this property; see, for example, [10, section 2] and [4, section 4.3]. We refer the readers to [4, 19] for more examples. In particular, one can show that if $f(x) = \frac{1}{2}\|Ax - b\|^2$ for some matrix A and vector b, P is given as in any one of the five examples at the beginning of this subsection, then the function E in* (4.10) *is a KL function with $\phi(s) = cs^{1-\theta}$ for some $\theta \in [0, 1)$ and $c > 0$.*

5 Numerical experiments

In this section, we perform numerical experiments to illustrate the efficiency of our algorithm pDCA$_e$ for solving problem (3.1). All experiments are performed in Matlab 2015b on a 64-bit PC with an Intel(R) Core(TM) i7-4790 CPU (3.60GHz) and 32GB of RAM.

In our numerical tests, we focus on the following DC regularized least squares problem:

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2 + P_1(x) - P_2(x), \tag{5.1}$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, P_1 is a proper closed convex function and P_2 is a continuous convex function. We consider two different classes of regularizers: the ℓ_{1-2} regularizer discussed in Example 4.1 and the logarithmic regularizer presented in Example 4.5. We compare three algorithms for solving (5.1) with these regularizers: our algorithm pDCA$_e$, the proximal DCA (pDCA) studied in various work such as [27] and [18], and the GIST proposed in [17]. We discuss the implementation details of these algorithms below.

pDCA$_e$. For this algorithm, we set $L = \lambda_{\max}(A^T A)$,[5] choose the extrapolation parameters $\{\beta_t\}$ as in (3.5), and perform both the fixed restart (with $\bar{T} = 200$) and the adaptive restart strategies as described in Section 3. We initialize the algorithm at the origin and terminate it when

$$\frac{\|x^t - x^{t-1}\|}{\max\{1, \|x^t\|\}} < 10^{-5}. \tag{5.2}$$

pDCA. This is a special case of pDCA$_e$ with $\beta_t \equiv 0$. We set $L = \lambda_{\max}(A^T A)$, initialize the algorithm at the origin and terminate it when (5.2) holds. In our experiments below, this algorithm turns out to be very slow, and so we also terminate this algorithm when the iteration number hits 5000.

GIST. This algorithm was proposed in [17], and is the same as the nonmonotone proximal gradient algorithm described in [34] (see also [13, Appendix A, Algorithm 1]) applied to $f(x) = \frac{1}{2}\|Ax - b\|^2$ and $P(x) = P_1(x) - P_2(x)$. Following the notation in [13, Appendix A, Algorithm 1], in our implementation, we set $c = 10^{-4}$, $\tau = 2$, $M = 4$, $L_0^0 = 1$, and

$$L_t^0 = \min\left\{\max\left\{\frac{\|A(x^t - x^{t-1})\|^2}{\|x^t - x^{t-1}\|^2}, 10^{-8}\right\}, 10^8\right\}$$

for $t \geq 1$. We would like to point out that the subproblem in [13, Appendix A, A.4] now becomes

$$\min_{x \in \mathbb{R}^n} \left\{\langle A^T(Ax^t - b), x - x^t\rangle + \frac{L_t}{2}\|x - x^t\|^2 + P_1(x) - P_2(x)\right\},$$

which has closed form solutions for the two regularizers used in our experiments below; see the appendices of [17] and [20]. We initialize this algorithm at the origin and terminate it when (5.2) holds.

In our numerical experiments below, we compare our algorithm pDCA$_e$ with pDCA and GIST for solving (5.1) on random instances generated as follows. We first generate an $m \times n$ matrix A with i.i.d. standard Gaussian entries, and then normalize this matrix so that the columns of A have unit norms. A subset T of size s is then chosen uniformly at random from $\{1, 2, 3, \ldots, n\}$ and an s-sparse vector y having i.i.d. standard Gaussian entries on T is generated. Finally, we set $b = Ay + 0.01 \cdot \hat{n}$, where $\hat{n} \in \mathbb{R}^m$ is a random vector with i.i.d. standard Gaussian entries.

We next present the DC models we use in our numerical tests and the numerical results.

5.1 Least squares problems with ℓ_{1-2} regularizer

In this subsection, we consider the ℓ_{1-2} regularized least squares problem:

$$\min_{x \in \mathbb{R}^n} F_{\ell_{1-2}}(x) = \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1 - \lambda\|x\|, \tag{5.3}$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and $\lambda > 0$ is the regularization parameter. This problem takes the form of (5.1) with $P_1(x) = \lambda\|x\|_1$ and $P_2(x) = \lambda\|x\|$. We assume *in addition* that the A in (5.3) does not have zero columns. Using this assumption, Example 4.1, Theorem 4.2 and Remark 4.2, we see that $F_{\ell_{1-2}}$ is level-bounded, and that if we choose $\lambda < \frac{1}{2}\|A^T b\|_\infty$, then the sequence $\{x^t\}$ generated by pDCA$_e$ is globally convergent.

In our numerical experiments below, we consider $(m, n, s) = (720i, 2560i, 80i)$ for $i = 1, 2, \ldots, 10$. For each triple (m, n, s), we generate 30 instances randomly as described above. The computational results are presented in Tables 1 and 2, which correspond to problem (5.3) with $\lambda = 5 \times 10^{-4}$ and $\lambda = 1 \times 10^{-3}$ respectively.[6] We report the time for computing $\lambda_{\max}(A^T A)$ ($\mathsf{t}_{\lambda_{\max}}$), the number of iterations (iter),[7] CPU times in seconds (CPU time),[8] and the function values at termination (fval), averaged over the 30 random instances. We can see that pDCA$_e$ always outperforms pDCA and GIST.

[5] $\lambda_{\max}(A^T A)$ is computed via the MATLAB code `lambda = norm(A*A');` when $m \leq 2000$, and by `opts.issym = 1; lambda= eigs(A*A',1,'LM',opts);` otherwise.

[6] These λ satisfy $\lambda < \frac{1}{2}\|A^T b\|_\infty$ for all our random instances.

[7] In the tables, "max" means the number of iterations hits 5000.

[8] The CPU time reported for pDCA$_e$ does not include the time for computing $\lambda_{\max}(A^T A)$.

Table 1: Solving (5.3) on random instances, $\lambda = 5 \times 10^{-4}$

problem size				iter			CPU time			fval		
n	m	s	$\mathbf{t}_{\lambda_{\max}}$	GIST	pDCA$_e$	pDCA	GIST	pDCA$_e$	pDCA	GIST	pDCA$_e$	pDCA
2560	720	80	0.1	1736	915	max	3.0	1.2	6.1	2.9757e-02	2.9743e-02	4.7049e-02
5120	1440	160	0.7	1726	895	max	14.2	5.4	29.6	6.1497e-02	6.1472e-02	9.5797e-02
7680	2160	240	0.7	1747	929	max	31.0	12.1	64.7	9.3836e-02	9.3799e-02	1.4394e-01
10240	2880	320	1.3	1754	949	max	54.8	21.8	114.6	1.2500e-01	1.2495e-01	1.9063e-01
12800	3600	400	2.4	1767	935	max	86.8	33.9	180.6	1.5956e-01	1.5949e-01	2.4367e-01
15360	4320	480	3.7	1757	955	max	120.5	48.5	253.2	1.8982e-01	1.8975e-01	2.8811e-01
17920	5040	560	6.0	1778	982	max	166.6	67.5	343.7	2.2481e-01	2.2472e-01	3.4110e-01
20480	5760	640	7.6	1780	982	max	215.6	87.5	444.4	2.5908e-01	2.5897e-01	3.9319e-01
23040	6480	720	10.7	1782	982	max	269.8	110.4	561.4	2.9150e-01	2.9137e-01	4.4057e-01
25600	7200	800	14.3	1799	995	max	341.4	140.1	704.0	3.2831e-01	3.2816e-01	4.9679e-01

Table 2: Solving (5.3) on random instances, $\lambda = 1 \times 10^{-3}$

problem size				iter			CPU time			fval		
n	m	s	$\mathbf{t}_{\lambda_{\max}}$	GIST	pDCA$_e$	pDCA	GIST	pDCA$_e$	pDCA	GIST	pDCA$_e$	pDCA
2560	720	80	0.1	925	600	max	1.7	0.8	6.1	5.9909e-02	5.9903e-02	7.2646e-02
5120	1440	160	0.7	908	602	max	7.4	3.6	29.5	1.2002e-01	1.2001e-01	1.4286e-01
7680	2160	240	0.6	928	602	max	16.4	7.9	65.3	1.8679e-01	1.8677e-01	2.2359e-01
10240	2880	320	1.3	941	602	max	29.0	13.8	114.4	2.5185e-01	2.5182e-01	3.0125e-01
12800	3600	400	2.4	946	602	max	45.8	21.8	179.9	3.1906e-01	3.1903e-01	3.8187e-01
15360	4320	480	3.8	949	602	max	64.6	30.6	253.5	3.8418e-01	3.8414e-01	4.6012e-01
17920	5040	560	6.0	943	602	max	87.4	41.7	345.8	4.4659e-01	4.4654e-01	5.3129e-01
20480	5760	640	7.7	946	602	max	112.4	53.6	444.4	5.1037e-01	5.1031e-01	6.0884e-01
23040	6480	720	10.6	943	602	max	141.8	68.2	562.0	5.8029e-01	5.8022e-01	6.9129e-01
25600	7200	800	14.1	946	602	max	179.0	84.9	703.7	6.4830e-01	6.4822e-01	7.7247e-01

To illustrate the ability of recovering the original sparse solution by our method, we plot in Figure 1 the true solution and the solution obtained by pDCA$_e$ for solving (5.3) with $\lambda = 5 \times 10^{-4}$ (the plot on the left) and $\lambda = 10^{-3}$ (the plot on the right) on a random instance $(m, n, s) = (720, 2560, 80)$. The true solution y is represented by asterisks, while the circles are the estimates obtained by pDCA$_e$. We see that the estimates obtained by pDCA$_e$ are quite close to the true values.

5.2 Least squares problems with logarithmic regularizer

In this subsection, we consider the least squares problem with logarithmic regularization function:

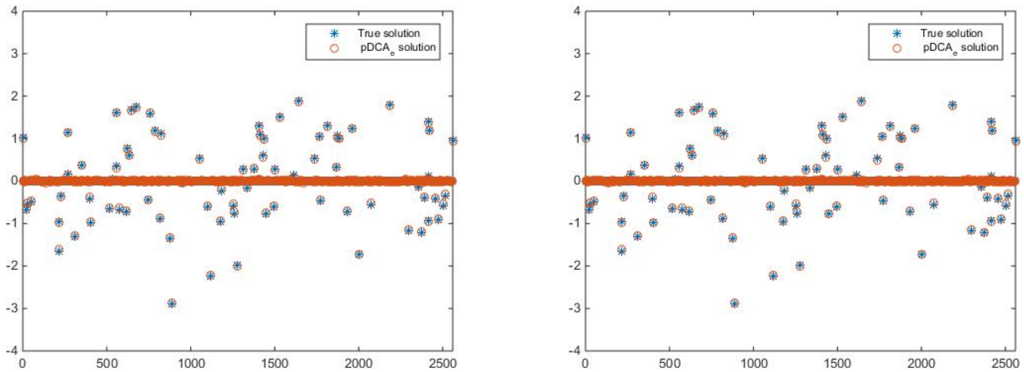
$$\min_{x\in\mathbb{R}^n} F_{\log}(x) = \frac{1}{2}\|Ax - b\|^2 + \sum_{i=1}^{n} \left[\lambda \log(|x_i| + \epsilon) - \lambda \log \epsilon\right], \tag{5.4}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $\epsilon > 0$ is a constant, and $\lambda > 0$ is the regularization parameter. From the discussion in Example 4.5, it is easy to show that $F_{\log}$ takes the form of (5.1) with $P_1(x) = \frac{\lambda}{\epsilon}\|x\|_1$ and $P_2(x) = \sum_{i=1}^n \lambda \left[\frac{|x_i|}{\epsilon} - \log(|x_i| + \epsilon) + \log \epsilon\right]$. In addition, it is not hard to show that $F_{\log}$ is level-bounded. This together with Theorem 4.2 and Remark 4.2 shows that the sequence $\{x^t\}$ generated by pDCA$_e$ is globally convergent to a stationary point of (5.4).

In our experiments below, we consider $(m, n, s) = (720i, 2560i, 80i)$, $i = 1, 2, \ldots, 10$. For each triple, we generate 30 instances randomly as described above. The computational results are presented in Tables 3 and 4, which correspond to problem (5.4) with $\lambda = 5 \times 10^{-4}$ and $\lambda = 1 \times 10^{-3}$ respectively.[9]

[9] We set $\epsilon = 0.5$ in (5.4).

Figure 1: The true solution and the solution obtained by solving (5.3) with $\lambda = 5 \times 10^{-4}$ (left) and $\lambda = 10^{-3}$ (right).



In these tables, we report the time for computing $\lambda_{\max}(A^T A)$ ($\mathbf{t}_{\lambda_{\max}}$), the number of iterations (iter),[10] CPU times in seconds (CPU time),[11] and the function values at termination (fval), averaged over the 30 random instances. We see from the tables that pDCA$_e$ always outperforms pDCA and GIST.

Table 3: Solving (5.4) on random instances, $\lambda = 5 \times 10^{-4}$

problem size				iter			CPU time			fval		
n	m	s	$\mathbf{t}_{\lambda_{\max}}$	GIST	pDCA$_e$	pDCA	GIST	pDCA$_e$	pDCA	GIST	pDCA$_e$	pDCA
2560	720	80	0.1	863	601	max	1.9	0.8	6.1	3.8020e-02	3.8013e-02	5.3479e-02
5120	1440	160	0.7	866	602	max	7.4	3.6	29.4	7.5865e-02	7.5852e-02	1.0691e-01
7680	2160	240	0.7	878	602	max	16.0	7.8	64.9	1.1419e-01	1.1417e-01	1.6253e-01
10240	2880	320	1.3	866	602	max	27.2	13.8	113.8	1.5219e-01	1.5217e-01	2.1442e-01
12800	3600	400	2.4	869	602	max	43.1	22.0	181.9	1.8917e-01	1.8914e-01	2.6717e-01
15360	4320	480	3.7	869	602	max	59.9	30.9	256.0	2.2823e-01	2.2819e-01	3.2213e-01
17920	5040	560	6.0	866	602	max	80.7	41.8	346.6	2.6594e-01	2.6589e-01	3.7583e-01
20480	5760	640	7.7	874	602	max	104.9	53.8	446.4	3.0510e-01	3.0505e-01	4.3300e-01
23040	6480	720	10.7	873	602	max	132.0	67.9	563.1	3.4211e-01	3.4205e-01	4.8604e-01
25600	7200	800	14.3	871	602	max	164.7	85.0	705.0	3.8055e-01	3.8049e-01	5.4107e-01

Similarly as in the previous subsection, we plot in Figure 2 the true solution and the solution obtained by pDCA$_e$ for solving (5.4) with $\lambda = 5 \times 10^{-4}$ (the plot on the left) and $\lambda = 10^{-3}$ (the plot on the right) on a random instance $(m, n, s) = (720, 2560, 80)$. The true solution y is represented by asterisks, and the solution obtained by pDCA$_e$ is marked by circles. We again observe that the estimates are close to the true solution.

6 Conclusion

In this paper, we propose a proximal difference-of-convex algorithm with extrapolation (pDCA$_e$) for solving (3.1), which reduces to the proximal DCA when $\beta_t \equiv 0$. Our algorithmic framework allows a wide range of choices of the extrapolation parameters $\{\beta_t\}$, including those used in FISTA with fixed restart [15]. We establish global subsequential convergence of the sequence generated by pDCA$_e$. In addition, by assuming the Kurdyka-Łojasiewicz property of the objective and the locally Lipschitz

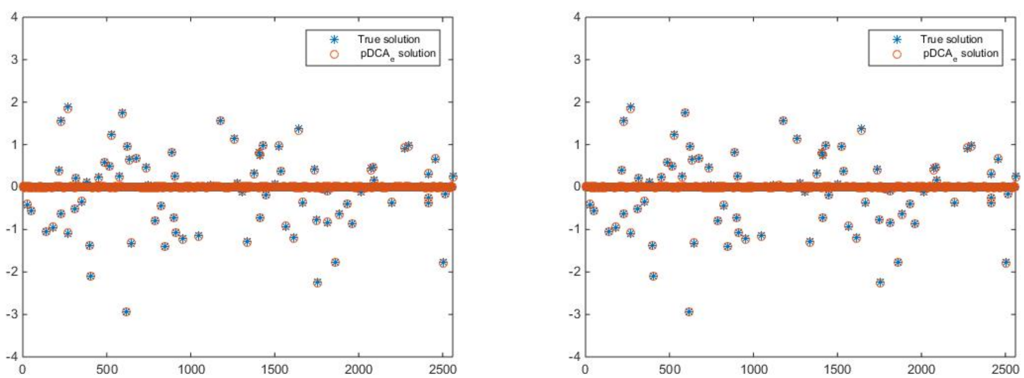
[10] In the tables, "max" means the number of iterations hits 5000.

[11] The CPU time reported for pDCA$_e$ does not include the time for computing $\lambda_{\max}(A^T A)$.

Table 4: Solving (5.4) on random instances, $\lambda = 1 \times 10^{-3}$

problem size				iter			CPU time			fval		
n	m	s	$\mathbf{t}_{\lambda_{\max}}$	GIST	pDCA$_e$	pDCA	GIST	pDCA$_e$	pDCA	GIST	pDCA$_e$	pDCA
2560	720	80	0.1	473	380	4531	1.0	0.5	5.7	7.6101e-02	7.6099e-02	7.6125e-02
5120	1440	160	0.7	473	400	4540	4.1	2.4	27.1	1.5200e-01	1.5200e-01	1.5204e-01
7680	2160	240	0.7	467	402	4546	8.4	5.3	59.7	2.2691e-01	2.2691e-01	2.2696e-01
10240	2880	320	1.3	475	402	4549	14.6	9.2	103.4	3.0374e-01	3.0373e-01	3.0381e-01
12800	3600	400	2.4	470	401	4519	22.7	14.5	162.7	3.7530e-01	3.7529e-01	3.7538e-01
15360	4320	480	3.8	471	402	4539	31.9	20.5	230.6	4.5451e-01	4.5450e-01	4.5461e-01
17920	5040	560	6.1	471	402	4564	42.9	27.6	312.5	5.2941e-01	5.2939e-01	5.2953e-01
20480	5760	640	7.8	475	402	4554	56.1	35.9	406.5	6.0388e-01	6.0385e-01	6.0401e-01
23040	6480	720	10.7	476	402	4593	70.9	45.3	516.9	6.8519e-01	6.8516e-01	6.8534e-01
25600	7200	800	14.3	475	402	4559	88.2	56.8	642.5	7.5684e-01	7.5681e-01	7.5701e-01

Figure 2: The true solution and the solution obtained by solving (5.4) with $\lambda = 5 \times 10^{-4}$ (left) and $\lambda = 10^{-3}$ (right).



differentiability of $P_2(x)$ in (3.1), we establish global convergence of the sequence generated by our algorithm and analyze its convergence rate. Our numerical experiments show that our algorithm usually outperforms the proximal DCA and GIST for two classes of DC regularized least squares problems.

Acknowledgement. The authors would like to thank the two anonymous referees for their helpful comments.

References

[1] M. Ahn, J.S. Pang, and J. Xin. Difference-of-convex learning I: directional stationarity, optimality, and sparsity. *preprint*, 2016.

[2] A. Alvarado, G. Scutari, and J.S. Pang. A new decomposition method for multiuser DC-programming and its applications. *IEEE Transactions on Signal Processing*, 62: 2984–2998, 2014.

[3] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming, Series B*, 116: 5–16, 2009.

[4] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35: 438–457, 2010.

[5] H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Mathematical Programming, Series A*, 137: 91–129, 2013.

[6] S. Banert and R.I. Boţ. A general double-proximal gradient algorithm for d.c. programming. *arXiv preprint arXiv*:1610.06538v1.

[7] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2: 183–202, 2009.

[8] S. Becker, E.J. Candès, and M.C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3: 165–218, 2011.

[9] W. Bian and X. Chen. Optimality and complexity for constrained optimization problems with nonconvex regularization. To appear in *Mathematics of Operations Research*.

[10] J. Bolte, A. Daniilidis, and A. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17: 1205–1223, 2007.

[11] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming, Series A*, 146: 459–494, 2014.

[12] E.J. Candès, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14: 877–905, 2008.

[13] X. Chen, Z. Lu, and T.K. Pong. Penalty methods for a class of non-Lipschitz optimization problems. *SIAM Journal on Optimization*, 26: 1465–1492, 2016.

[14] P.L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, 49: 185–212, 2011.

[15] B. O'Donoghue and E.J. Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15: 715–732, 2015.

[16] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96: 1348–1360, 2001.

[17] P. Gong, C. Zhang, Z. Lu, J. Z. Huang, and J. Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. *ICML*, 2013.

[18] J. Gotoh, A. Takeda, and K. Tono. DC formulations and algorithms for sparse optimization problems. Preprint, *METR 2015-27*, Department of Mathematical Informatics, University of Tokyo. Available at `http://www.keisu.t.u-tokyo.ac.jp/research/techrep/index.html`

[19] G. Li and T.K. Pong. Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods. To appear in *Foundations of Computational Mathematics*, DOI: 10.1007/s10208-017-9366-8.

[20] T. Liu and T.K. Pong. Further properties of the forward-backward envelope with applications to difference-of-convex programming. *Computational Optimization and Applications*, 67: 489–520, 2017.

[21] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(\frac{1}{k^2})$. *Soviet Mathematics Doklady*, 27: 372–376, 1983.

[22] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Kluwer Academic Publishers, Boston, 2004.

[23] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming, Series B*, 140: 125–161, 2013.

[24] Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming, Series B*, 109: 319–344, 2007.

[25] B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4: 1–17, 1964.

[26] D.T. Pham and H.A. Le Thi. Convex analysis approach to D.C. programming: theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22: 289–355, 1997.

[27] D.T. Pham and H.A. Le Thi. A D.C. optimization algorithm for solving the trust-region subproblem. *SIAM Journal on Optimization*, 8: 476–505, 1998.

[28] R.T. Rockafellar and R.J.-B. Wets. *Variational Analysis.* Springer, 1998.

[29] M. Sanjabi, M. Razaviyayn, and Z.-Q. Luo. Optimal joint base station assignment and beamforming for heterogeneous networks. *IEEE Transactions on Signal Processing*, 62: 1950–1961,2014.

[30] H.A. Le Thi and D.T. Pham. The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of Operations Research*, 133: 23–46, 2005.

[31] H.A. Le Thi, D.T. Pham, and V.N. Huynh. Exact penalty and error bounds in DC programming. *Journal of Global Optimization*, 52: 509–535, 2012.

[32] H.A. Le Thi, D.T. Pham, and D.M. Le. Exact penalty in D.C. programming. *Vietnam Journal of Mathematics*, 27: 169–178, 1999.

[33] H. Tuy. *Convex Analysis and Global Optimization, Second Edition.* Springer, 2016.

[34] S.J. Wright, R. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57: 2479–2493, 2009.

[35] P. Yin, Y. Lou, Q. He, and J. Xin. Minimization of ℓ_{1-2} for compressed sensing. *SIAM Journal on Scientific Computing*, 37: A536–A563, 2015.

[36] C. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38: 894–942, 2010.

[37] S. Zhang and J. Xin. Minimization of transformed L_1 penalty: theory, difference of convex function algorithm, and robust application in compressed sensing. *arXiv preprint arXiv*:1411.5735v3.