Nonmonotone Enhanced Proximal DC Algorithms for a Class of Structured Nonsmooth DC Programming

Lu, Zhaosong; Zhou, Zirui

Published in:
SIAM Journal on Optimization

DOI:
10.1137/18M1214342

Published: 31/10/2019

Document Version:
Publisher's PDF, also known as Version of record

SIAM J. OPTIM.
Vol. 29, No. 4, pp. 2725–2752

NONMONOTONE ENHANCED PROXIMAL DC ALGORITHMS FOR A CLASS OF STRUCTURED NONSMOOTH DC PROGRAMMING*

ZHAOSONG LU† AND ZIRUI ZHOU‡

Abstract. In this paper we consider a class of structured nonsmooth difference-of-convex (DC) minimization in which the first convex component is the sum of a smooth and a nonsmooth function while the second convex component is the supremum of finitely many convex smooth functions. The existing methods for this problem usually have weak convergence guarantees or exhibit slow convergence. Due to this, we propose two nonmonotone enhanced proximal DC algorithms for solving this problem. For possible acceleration, one uses a nonmonotone line-search scheme in which the associated Lipschitz constant is adaptively approximated by some local curvature information of the smooth function in the first convex component, and the other employs an extrapolation scheme. It is shown that every accumulation point of the solution sequence generated by them is a *D-stationary* point of the problem. These methods may, however, become inefficient when the number of convex smooth functions in defining the second convex component is large. To remedy this issue, we propose randomized counterparts for them and show that every accumulation point of the generated solution sequence is a *D-stationary* point of the problem *almost surely*. Some preliminary numerical experiments are conducted to demonstrate the efficiency of the proposed algorithms.

Key words. nonsmooth DC programming, D-stationary point, proximal DCA, nonmonotone line search, extrapolation

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

DOI. 10.1137/18M1214342

1. Introduction.

Difference-of-convex (DC) minimization, which refers to the problem of minimizing the difference of two convex functions, has been widely studied in the literature and has also found rich applications in science and engineering (e.g., see [1, 8, 9, 11, 18]). In this paper we consider a class of nonsmooth DC programming in the form of

$$\min_{x\in\Re^n} \{F(x) := f(x) - g(x)\}, \tag{1.1}$$

where

$$f(x) = f_s(x) + f_n(x), \quad g(x) = \max_{1\le i\le I} \psi_i(x). \tag{1.2}$$

We make the following assumptions for problem (1.1) throughout the paper.

Assumption 1.

(a) f_n is a proper closed convex function with a nonempty domain denoted by $\mathrm{dom}(f_n)$. Moreover, the proximal operator associated with f_n can be evaluated.[1]

*Received by the editors September 27, 2018; accepted for publication (in revised form) June 13, 2019; published electronically October 31, 2019.

https://doi.org/10.1137/18M1214342

Funding: The work of the authors was supported by an NSERC Discovery grant. The work of the second author was also supported by an SFU Alan Mekler postdoctoral fellowship.

†Department of Industrial and Systems Engineering, University of Minnesota, 111 Church St. S.E. Minneapolis, MN 55455 and Department of Mathematics, Simon Fraser University, Burnaby, British Columbia, V5A 1S6 Canada (zhaosong@umn.edu).

‡Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong (zirui-zhou@hkbu.edu.hk).

[1]The proximal operator associated with f_n is defined as $\mathrm{prox}_{f_n}(x) = \mathrm{argmin}_y\{\frac{1}{2}\|y-x\|^2 + f_n(y)\}$.

(b) f_s is convex and continuously differentiable on $\Re^n$, and its gradient ∇f_s is Lipschitz continuous with Lipschitz constant $L > 0$.
(c) For all $i = 1, \ldots, I$, ψ_i is convex and continuously differentiable on $\Re^n$.
(d) The optimal value of (1.1), denoted by F^*, is finite.

Many optimization problems arising in applications can be formulated as (1.1). For example, in the context of sparse regression, the model

$$\min_{x\in\Re^n} \left\{ \frac{1}{2}\|Ax - b\|^2 + \lambda P(x) \right\} \tag{1.3}$$

is often used, where $\|\cdot\|$ is the Euclidean norm, $A \in \Re^{p\times n}$, $b \in \Re^p$, and $\lambda > 0$ are given, and P is a penalty function for promoting sparse solutions. As shown in [2], the model (1.3) can be recast into (1.1) for some popular penalty functions P such as SCAD [6], MCP [22], and K-sparsity [8]. Some other applications of DC minimization (1.1) can be found, for example, in digital communication systems [1] and assignment allocation [18].

The classical difference-of-convex algorithm (DCA) is widely used in DC programming (e.g., see [7, 11, 8]) and can be applied to problem (1.1). Given an iterate x^k, DCA generates the next one by solving the convex optimization problem[2]

$$x^{k+1} \in \operatorname*{Argmin}_{x\in\Re^n} \left\{ f(x) - \langle v^k, x\rangle \right\}$$

for some $v^k \in \partial g(x^k)$. By exploiting the structure of f in (1.2), the proximal DCA (PDCA) has been proposed for solving a class of DC programming (e.g., see [8]). It can be suitably applied to (1.1) for which the new iterate is obtained by solving the proximal subproblem

$$x^{k+1} = \operatorname*{argmin}_{x\in\Re^n} \left\{ f_n(x) + \langle \nabla f_s(x^k) - v^k, x\rangle + \frac{L}{2}\|x - x^k\|^2 \right\} \tag{1.4}$$

for some $v^k \in \partial g(x^k)$. Recently, Tono, Takeda, and Gotoh [19] proposed a proximal DCA with nonmonotone line search (NPDCA) for possible acceleration, which solves almost the same subproblems as (1.4) except that the Lipschitz constant L is adaptively approximated by some local curvature information of f_s. In addition, for possibly accelerating PDCA, Wen, Chen, and Pong [20] recently proposed a proximal DCA with extrapolation (PDCA$_e$) that is also applicable to solving (1.1). In particular, let $\{\beta_t\}_{t\geq 0} \subseteq [0, 1)$ with $\sup_t \beta_t < 1$ be given. The PDCA$_e$ first constructs an extrapolation point $z^k = x^k + \beta_k(x^k - x^{k-1})$, and then computes the next iterate by letting

$$x^{k+1} = \operatorname*{argmin}_{x\in\Re^n} \left\{ f_n(x) + \langle \nabla f_s(z^k) - v^k, x\rangle + \frac{L}{2}\|x - z^k\|^2 \right\} \tag{1.5}$$

for some $v^k \in \partial g(x^k)$. It has been shown that every accumulation point x^∞ of the sequence $\{x^k\}$ generated by DCA, PDCA, NPDCA, and PDCA$_e$ is a *critical point* of problem (1.1), that is, $\partial f(x^\infty) \cap \partial g(x^\infty) \neq \emptyset$.

[2]By convention, the symbol "Argmin" stands for the set of the solutions of the associated minimization problem. When this set is known to be a singleton, we use the symbol "argmin" to stand for it instead.

By exploiting the structure of g in (1.2), Pang, Razaviyayn, and Alvarado [15] recently proposed a novel enhanced DCA (EDCA) for solving (1.1). Given an iterate x^k, EDCA first solves the convex optimization problems

$$x^{k,i} = \operatorname*{argmin}_{x\in\Re^n} \left\{ f(x) - \langle \nabla\psi_i(x^k), x\rangle + \frac{1}{2}\|x - x^k\|^2 \right\} \tag{1.6}$$

for each $i \in \mathcal{A}_\eta(x^k)$, where $\mathcal{A}_\eta(x) = \{i : \psi_i(x) \geq g(x) - \eta,\ 1 \leq i \leq I\}$ for some $\eta > 0$. It then generates the next iterate by letting $x^{k+1} = x^{k,\hat{i}}$ with $\hat{i}$ given by

$$\hat{i} \in \operatorname*{Argmin}_{i\in\mathcal{A}_\eta(x^k)} \left\{ F(x^{k,i}) + \frac{1}{2}\|x^{k,i} - x^k\|^2 \right\}.$$

It is shown in [15] that any accumulation point x^∞ of the sequence $\{x^k\}$ generated by EDCA is a *directional-stationary* (D-stationary) point of problem (1.1), that is, $\partial g(x^\infty) \subseteq \partial f(x^\infty)$, which is generally stronger than the aforementioned critical point.[3]

Given that finding the exact solution of the subproblems (1.6) of EDCA is generally impossible, Lu, Zhou, and Sun [13] recently proposed an enhanced PDCA (EPDCA) for solving problem (1.1), which has much simpler subproblems than EDCA but maintains its strong convergence guarantee. In particular, let $c > 0$ and $\{\beta_t\}_{t\geq 0} \subseteq [0, \sqrt{c/L})$ with $\sup_t \beta_t < \sqrt{c/L}$ be given. Analogous to PDCA$_e$, EPDCA first constructs an extrapolation point $z^k = x^k + \beta_k(x^k - x^{k-1})$. It then solves the convex subproblems

$$x^{k,i} = \operatorname*{argmin}_{x\in\Re^n} \left\{ f_n(x) + \langle \nabla f_s(z^k) - \nabla\psi_i(x^k), x\rangle + \frac{L}{2}\|x - z^k\|^2 + \frac{c}{2}\|x - x^k\|^2 \right\} \tag{1.7}$$

for each $i \in \mathcal{A}_\eta(x^k)$ for some $\eta > 0$. Finally, it generates the next iterate by letting $x^{k+1} = x^{k,\hat{i}}$ with $\hat{i}$ given by

$$\hat{i} \in \operatorname*{Argmin}_{i\in\mathcal{A}_\eta(x^k)} \left\{ F(x^{k,i}) + \frac{c}{2}\|x^{k,i} - x^k\|^2 \right\}.$$

It is shown in [13] that any accumulation point x^∞ of the sequence $\{x^k\}$ generated by EPDCA is an $(\alpha, \tilde{\eta})$-D-stationary point of problem (1.1) for any $\alpha \in (0, (L+c)^{-1}]$ and $\tilde{\eta} \in [0, \eta)$, that is,

$$F(x^\infty) \leq f(x) - \psi_i(x^\infty) - \langle \nabla\psi_i(x^\infty), x - x^\infty\rangle + \frac{1}{2\alpha}\|x - x^\infty\|^2 \quad \forall x \in \Re^n,\ \forall i \in \mathcal{A}_{\tilde{\eta}}(x^\infty),$$

which is generally stronger than the aforementioned D-stationary point.

It is nice that EPDCA has a stronger convergence guarantee than PDCA$_e$ in terms of solution quality, while its subproblems (1.7) are as simple as those of PDCA$_e$ in (1.5). Nevertheless, EPDCA can converge much more slowly than PDCA$_e$. In fact, akin to PDCA$_e$, to make the extrapolation effective on EPDCA, c shall not be chosen too small. Due to the proximal term $c\|x - x^k\|^2/2$ in (1.7), a large c makes the step

[3] The convergence result of EDCA established in [15] can be strengthened. In fact, it is shown in [13] that any accumulation point of the sequence generated by EDCA is an $(\alpha, \tilde{\eta})$-D-stationary point of problem (1.1) for any $\alpha \in (0, 1]$ and $\tilde{\eta} \in [0, \eta)$, which is generally stronger than the aforementioned D-stationary point.

size of EPDCA much smaller than that of PDCA$_e$ and thus renders a slower convergence of EPDCA. To speed up computation while maintaining a similarly strong convergence guarantee to that of EPDCA, we propose in this paper two nonmonotone EPDCAs (NEPDCAs) for solving problem (1.1). The first NEPDCA solves the subproblems (1.7) with $c = 0$, $z^k = x^k$, and L being adaptively approximated by some local curvature information of f_s in a similar vein to in [19]. The second NEPDCA is similar to EPDCA except that (i) it solves the subproblems (1.7) with $c = 0$, and (ii) $\beta_k \in [0, \beta]$ with $\beta \in [0, \sqrt{c/L}) \cup \{0\}$. We show that every accumulation point of the sequence generated by both NEPDCAs is a D-stationary point of problem (1.1). Since both EPDCAs require solving a number of subproblems (1.7) per iteration, they may become computationally inefficient when the I in defining g is large. Inspired by a randomized algorithm proposed in [15, section 5.2], we remedy this issue by proposing two randomized NEPDCAs for solving problem (1.1), which solve subproblems in the form of (1.7) only once per iteration. We also show that any accumulation point of the sequence generated by both randomized NEPDCAs is a D-stationary point of (1.1) *almost surely*. In addition, we conduct some preliminary numerical experiments to compare the performance of the proposed methods with EPDCA and PDCA$_e$. The computational results demonstrate that the proposed methods inherit the advantages of EPDCA and PDCA$_e$. In particular, they are comparable to EPDCA but substantially outperform PDCA$_e$ in terms of solution quality. Moreover, they are comparable to PDCA$_e$ but much faster than EPDCA in terms of speed.

The rest of this paper is organized as follows. In section 2, we present some technical preliminaries. In sections 3 and 4, we propose two nonmonotone EPDCAs and also their randomized counterparts for solving problem (1.1), and establish their convergence. In section 5, we present some numerical results for the proposed algorithms. Finally, we present some concluding remarks in section 6 .

1.1. Notation. Let $\Re^n$ be the n-dimensional Euclidean space, $\langle\cdot,\cdot\rangle$ the standard inner product, and $\|\cdot\|$ the Euclidean norm. For a real number t, let $t_+ = \max\{0, t\}$. Given a function $h : \Re^n \to (-\infty, \infty]$, we use $\mathrm{dom}(h)$ to denote the domain of h, that is, $\mathrm{dom}(h) = \{x \in \Re^n : h(x) < \infty\}$. The directional derivative of h at a point $x \in \mathrm{dom}(h)$ along a direction $d \in \Re^n$ is defined as

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

if the limit exists. Suppose that h is additionally convex. We use ∂h to denote the subdifferential of h (e.g., see [16]). The proximal operator of h, denoted by prox_h, is a mapping from $\Re^n$ to $\Re^n$ defined as

$$\mathrm{prox}_h(z) = \mathop{\mathrm{argmin}}_{x \in \Re^n} \left\{ \frac{1}{2}\|x - z\|^2 + h(x) \right\}.$$

For the function g given in (1.2) and any $\eta \geq 0$, we write $\mathcal{I} = \{1, \ldots, I\}$ and define

$$\mathcal{A}(x) = \{i \in \mathcal{I} \mid \psi_i(x) = g(x)\}, \qquad \mathcal{A}_\eta(x) = \{i \in \mathcal{I} \mid \psi_i(x) \geq g(x) - \eta\}. \tag{1.8}$$

Clearly, $\mathcal{A}(x)$ consists of the associated active indices in defining $g(x)$. Moreover, $\mathcal{A}_0(x) = \mathcal{A}(x)$ and $\mathcal{A}(x) \subseteq \mathcal{A}_\eta(x) \subseteq \mathcal{I}$. Given any $i \in \mathcal{I}$ and $y, z \in \mathrm{dom}(F)$, we define

$$\ell_i(x; y, z) = f_s(y) + \langle \nabla f_s(y), x - y\rangle + f_n(x) - \psi_i(z) - \langle \nabla \psi_i(z), x - z\rangle, \tag{1.9}$$

which is clearly a convex function in x. With a slight abuse of notation, we write $\ell_i(x;y,y)$ as $\ell_i(x;y)$, that is,

$$\ell_i(x;y) = f_s(y) + \langle \nabla f_s(y), x-y\rangle + f_n(x) - \psi_i(y) - \langle \nabla \psi_i(y), x-y\rangle. \tag{1.10}$$

Given $x \in \mathrm{dom}(F)$ and $\eta \geq 0$, we denote the associated η-level set by $\mathcal{L}(x;\eta)$, that is,

$$\mathcal{L}(x;\eta) = \{z \in \Re^n : F(z) \leq F(x) + \eta\}.$$

For simplicity, we let $\mathcal{L}(x) = \mathcal{L}(x;0)$. Also, x is said to be a *critical* point of problem (1.1) if $0 \in \partial f(x) - \partial g(x)$, or equivalently, $\partial f(x) \cap \partial g(x) \neq \emptyset$. In addition, x is called a *directional-stationary* (D-stationary) point of (1.1) if $F'(x;d) \geq 0$ for all $d \in \Re^n$, or equivalently, $\partial g(x) \subset \partial f(x)$. It is known that any local minimizer of problem (1.1) must be a critical point and also a D-stationary point of (1.1). In addition, a D-stationary point of (1.1) must be a critical point of (1.1), but the converse generally does not hold (see [15] for a detailed discussion).

2. Technical preliminaries. In this section we present some technical preliminaries that will be used subsequently.

PROPOSITION 1. *A point $x \in \mathrm{dom}(F)$ is a D-stationary point of* (1.1) *if and only if*

$$0 \in \nabla f_s(x) + \partial f_n(x) - \nabla \psi_i(x) \quad \forall i \in \mathcal{A}(x).$$

Proof. The conclusion immediately follows from [13, Proposition 1]. ∎

PROPOSITION 2. *Let $h : \Re^n \to (-\infty, \infty]$ be a proper closed convex function, $\{y^k\}$ a sequence of vectors in $\Re^n$ converging to some $y^\infty \in \Re^n$, and $\{\alpha_k\}$ a sequence of positive scalars converging to some $\alpha_\infty > 0$. Suppose that the sequence $\{x^k\}$ is given by*

$$x^k = \operatorname*{argmin}_{x \in \Re^n} \left\{ h(x) + \frac{\alpha_k}{2}\|x - y^k\|^2 \right\}. \tag{2.1}$$

Then $\{x^k\}$ converges to x^∞, where x^∞ is given by

$$x^\infty = \operatorname*{argmin}_{x \in \Re^n} \left\{ h(x) + \frac{\alpha_\infty}{2}\|x - y^\infty\|^2 \right\}. \tag{2.2}$$

Proof. The first-order optimality conditions of (2.1) and (2.2) yield

$$\alpha_k(y^k - x^k) \in \partial h(x^k), \qquad \alpha_\infty(y^\infty - x^\infty) \in \partial h(x^\infty).$$

These together with the monotonicity of ∂h imply that

$$\left\langle x^k - x^\infty,\ \alpha_k(y^k - x^k) - \alpha_\infty(y^\infty - x^\infty)\right\rangle \geq 0.$$

Upon some simple manipulation of this inequality, one has that

$$\begin{aligned}\alpha_k\|x^k - x^\infty\|^2 &\leq \left\langle x^k - x^\infty,\ \alpha_k(y^k - y^\infty) + (\alpha_k - \alpha_\infty)(y^\infty - x^\infty)\right\rangle \\ &\leq \|x^k - x^\infty\|\|\alpha_k(y^k - y^\infty) + (\alpha_k - \alpha_\infty)(y^\infty - x^\infty)\|.\end{aligned}$$

It then follows from this and $\alpha_k > 0$ that

$$\|x^k - x^\infty\| \leq \left\| y^k - y^\infty + \frac{\alpha_k - \alpha_\infty}{\alpha_k}(y^\infty - x^\infty)\right\|.$$

This together with

$$\lim_{k\to\infty} y^k = y^\infty \quad \text{and} \quad \lim_{k\to\infty} \alpha_k = \alpha_\infty > 0$$

implies that $\lim_{k\to\infty} \|x^k - x^\infty\| = 0$ and hence $\lim_{k\to\infty} x^k = x^\infty$. ∎

COROLLARY 1. *Let $\{y^k\}$ and $\{z^k\}$ be two sequences of vectors in $\mathrm{dom}(F)$ converging to some y^∞ and z^∞, respectively, and $\{\alpha_k\}$ a sequence of positive scalars converging to some $\alpha_\infty > 0$. Suppose that the sequence $\{x^k\}$ is given by*

$$x^k = \operatorname*{argmin}_{x\in\Re^n} \left\{ \ell_i(x; y^k, z^k) + \frac{\alpha_k}{2}\|x - y^k\|^2 \right\} \tag{2.3}$$

for some $i \in \mathcal{I}$, where $\ell_i(x; y, z)$ is defined in (1.9). Then $\{x^k\}$ converges to x^∞, where x^∞ is given by

$$x^\infty = \operatorname*{argmin}_{x\in\Re^n} \left\{ \ell_i(x; y^\infty, z^\infty) + \frac{\alpha_\infty}{2}\|x - y^\infty\|^2 \right\}. \tag{2.4}$$

Proof. By (1.9) and (2.3), it is not hard to observe that

$$x^k = \operatorname*{argmin}_{x\in\Re^n} \left\{ f_n(x) + \frac{\alpha_k}{2} \left\| x - \left[y^k - \frac{1}{\alpha_k}\left(\nabla f_s(y^k) - \nabla\psi_i(z^k)\right)\right]\right\|^2 \right\}. \tag{2.5}$$

Due to $\lim_{k\to\infty}\alpha_k = \alpha_\infty > 0$, $\lim_{k\to\infty} y^k = y^\infty$, $\lim_{k\to\infty} z^k = z^\infty$, and the continuity of ∇f_s and $\nabla\psi_i$, we have that

$$\lim_{k\to\infty} \left[y^k - \frac{1}{\alpha_k}\left(\nabla f_s(y^k) - \nabla\psi_i(y^k)\right)\right] = y^\infty - \frac{1}{\alpha_\infty}\left(\nabla f_s(y^\infty) - \nabla\psi_i(y^\infty)\right).$$

It then follows from this, (1.9), (2.4), (2.5), and Proposition 2 that

$$\begin{aligned}\lim_{k\to\infty} x^k &= \operatorname*{argmin}_{x\in\Re^n} \left\{ f_n(x) + \frac{\alpha_\infty}{2} \left\| x - \left[y^\infty - \frac{1}{\alpha_\infty}\left(\nabla f_s(y^\infty) - \nabla\psi_i(z^\infty)\right)\right]\right\|^2 \right\} \\ &= \operatorname*{argmin}_{x\in\Re^n} \left\{ \ell_i(x; y^\infty, z^\infty) + \frac{\alpha_\infty}{2}\|x - y^\infty\|^2 \right\} \;=\; x^\infty. \end{aligned}$$

∎

3. Nonmonotone enhanced proximal DCA with line search. In this section we propose an enhanced PDCA for solving problem (1.1) in which a nonmonotone line-search scheme is applied. In particular, the associated Lipschitz constant is adaptively approximated by some local curvature information of f_s in a similar vein to in [19]. We also propose a randomized counterpart for this method, and establish convergence for both methods.

3.1. A deterministic nonmonotone enhanced PDCA with line search. In this subsection, we present a deterministic nonmonotone enhanced PDCA with line search for solving problem (1.1), and study its convergence properties.

ALGORITHM 1 (a deterministic nonmonotone enhanced PDCA with line search).

(0) Input $x^0 \in \mathrm{dom}(F)$, $\eta > 0$, $\rho > 1$, $0 < c < L/2$, $0 < \underline{\alpha} \le \overline{\alpha}$, and integer $N \ge 0$. Set $k \leftarrow 0$.

(1) Choose $\alpha_{k,0} \in [\underline{\alpha}, \overline{\alpha}]$.

(2) For $m = 0, 1, \ldots$:

(2a) Let $\alpha_k = \alpha_{k,0}\rho^m$.

(2b) For each $i \in \mathcal{A}_\eta(x^k)$, compute

$$x^{k,i}(\alpha_k) = \operatorname*{argmin}_{x\in\Re^n} \left\{ \ell_i(x; x^k) + \frac{\alpha_k}{2}\|x - x^k\|^2 \right\}. \tag{3.1}$$

(2c) Let

$$\hat{i} \in \operatorname*{Argmin}_{i\in\mathcal{A}_\eta(x^k)} \left\{ F(x^{k,i}(\alpha_k)) + \frac{c}{2}\|x^{k,i}(\alpha_k) - x^k\|^2 \right\}.$$

If $x^{k,\hat{i}}(\alpha_k)$ satisfies

$$\begin{aligned} F(x^{k,\hat{i}}(\alpha_k)) \le \max\left\{ f_s(x^k) + f_n(x^k) - \psi_i(x^k), \max_{[k-N]_+\le j\le k} F(x^j) \right\} \\ - \frac{c}{2}\|x^{k,\hat{i}}(\alpha_k) - x^k\|^2 - \frac{c}{2}\|x^{k,i}(\alpha_k) - x^k\|^2 \ \forall i \in \mathcal{A}_\eta(x^k), \end{aligned} \tag{3.2}$$

set $x^{k+1} = x^{k,\hat{i}}(\alpha_k)$ and go to step (3).

(3) Set $k \leftarrow k+1$ and go to step (1).

Remark 1.

(i) For Algorithm 1, $\{F(x^k)\}$ is monotone when $N = 0$. It is, however, generally nonmonotone when $N > 0$. In addition, when $g \equiv 0$, Algorithm 1 is reduced to a similar algorithm, as proposed in [21]. Furthermore, when the number I in defining g is equal to one, Algorithm 1 is reduced to the nonmonotone PDCA proposed in [19].

(ii) A popular choice of $\alpha_{k,0}$ is taken with the following formula proposed by Barzilai and Borwein [3], which adaptively approximates the associated Lipschitz constant via some local curvature information of f_s:

$$\alpha_{k,0} = \begin{cases} \max\left\{\underline{\alpha},\ \min\left\{\overline{\alpha},\ \frac{|\Delta x^T \Delta G|}{\|\Delta x\|^2}\right\}\right\} & \text{if } \Delta x \neq 0, \\ \overline{\alpha} & \text{if } \Delta x = 0 \end{cases} \tag{3.3}$$

for some $0 < \underline{\alpha} < \overline{\alpha}$, where $\Delta x = x^k - x^{k-1}$ and $\Delta G = \nabla f_s(x^k) - \nabla f_s(x^{k-1})$.

In what follows, we conduct convergence analysis for Algorithm 1. In particular, we first show that for each outer loop, its associated inner loops must terminate in a finite number of iterations. We then show that any accumulation point of $\{x^k\}$ is a D-stationary point of problem (1.1).

THEOREM 1. *For any $k \ge 0$, step* (2) *of Algorithm* 1 *terminates at some $\alpha_k \le \tilde{\alpha}$ in at most M iterations, where*

$$\tilde{\alpha} = \max\{\overline{\alpha}, \rho L\}, \qquad M = \left\lfloor \frac{\log(\max\{\overline{\alpha}, \rho L\}) - \log\underline{\alpha}}{\log\rho} \right\rfloor + 1. \tag{3.4}$$

Proof. Claim that (3.2) is satisfied whenever $\alpha_k \ge L$. Indeed, suppose $\alpha_k \ge L$. Notice that the objective function in (3.1) is strongly convex with modulus α_k. It follows from this, (1.10), and (3.1) that for every $i \in \mathcal{A}_\eta(x^k)$,

$$\begin{aligned} \ell_i(x^{k,i}(\alpha_k); x^k) + \frac{\alpha_k}{2}\|x^{k,i}(\alpha_k) - x^k\|^2 \\ \le f_s(x^k) + f_n(x^k) - \psi_i(x^k) - \frac{\alpha_k}{2}\|x^{k,i}(\alpha_k) - x^k\|^2. \end{aligned} \tag{3.5}$$

Using this, $0 < c < L/2$, $\alpha_k \geq L$, the Lipschitz continuity of ∇f_s, and the convexity of ψ_i, we have that for every $i \in \mathcal{A}_\eta(x^k)$,

$$
\begin{aligned}
& f_s(x^k) + f_n(x^k) - \psi_i(x^k) - c\|x^{k,i}(\alpha_k) - x^k\|^2 \\
& \geq f_s(x^k) + f_n(x^k) - \psi_i(x^k) - \frac{\alpha_k}{2}\|x^{k,i}(\alpha_k) - x^k\|^2 \qquad (3.6) \\
& \geq f_s(x^k) + \langle \nabla f_s(x^k), x^{k,i}(\alpha_k) - x^k\rangle + \frac{\alpha_k}{2}\|x^{k,i}(\alpha_k) - x^k\|^2 + f_n(x^{k,i}(\alpha_k)) \\
& \quad - \psi_i(x^k) - \langle \nabla \psi_i(x^k), x^{k,i}(\alpha_k) - x^k\rangle \qquad (3.7) \\
& \geq f_s(x^{k,i}(\alpha_k)) + f_n(x^{k,i}(\alpha_k)) - \psi_i(x^{k,i}(\alpha_k)) \qquad (3.8) \\
& \geq f_s(x^{k,i}(\alpha_k)) + f_n(x^{k,i}(\alpha_k)) - \max_{i\in\mathcal{I}} \psi_i(x^{k,i}(\alpha_k)) = F(x^{k,i}(\alpha_k)), \qquad (3.9)
\end{aligned}
$$

where (3.6) is due to $0 < c < L/2$ and $\alpha_k \geq L$, (3.7) follows from (1.10) and (3.5), and (3.8) follows from $\alpha_k \geq L$, the Lipschitz continuity of ∇f_s, and the convexity of ψ_i. By (3.9) and the choice of $\hat{i}$ in step (2c) of Algorithm 1, one has that for every $i \in \mathcal{A}_\eta(x^k)$,

$$
\begin{aligned}
f_s(x^k) + f_n(x^k) - \psi_i(x^k) - \frac{c}{2}\|x^{k,i}(\alpha_k) - x^k\|^2 &\geq F(x^{k,i}(\alpha_k)) + \frac{c}{2}\|x^{k,i}(\alpha_k) - x^k\|^2 \\
&\geq F(x^{k,\hat{i}}(\alpha_k)) + \frac{c}{2}\|x^{k,\hat{i}}(\alpha_k) - x^k\|^2.
\end{aligned}
$$

The last inequality implies that for all $i \in \mathcal{A}_\eta(x^k)$,

$$
\begin{aligned}
& F(x^{k,\hat{i}}(\alpha_k)) \leq f_s(x^k) + f_n(x^k) - \psi_i(x^k) - \frac{c}{2}\|x^{k,\hat{i}}(\alpha_k) - x^k\|^2 - \frac{c}{2}\|x^{k,i}(\alpha_k) - x^k\|^2 \\
& \leq \max\left\{ f_s(x^k) + f_n(x^k) - \psi_i(x^k), \max_{[k-N]_+ \leq j \leq k} F(x^j) \right\} - \frac{c}{2}\|x^{k,\hat{i}}(\alpha_k) - x^k\|^2 \\
& \quad - \frac{c}{2}\|x^{k,i}(\alpha_k) - x^k\|^2.
\end{aligned}
$$

Hence, (3.2) is satisfied whenever $\alpha_k \geq L$. By this and the update scheme of α_k, it is not hard to see that the inner loops of the kth outer loop must terminate at some $\alpha_k \leq \max\{\overline{\alpha}, \rho L\}$. It then follows that step (2) of Algorithm 1 terminates in at most M iterations, where M is given in (3.4). ∎

We next show that any accumulation point of $\{x^k\}$ is a D-stationary point of problem (1.1). To proceed, let $\iota(k)$ be an integer between $[k-N]_+$ and k such that

$$
F(x^{\iota(k)}) = \max_j \big\{ F(x^j) : j = [k-N]_+, \ldots, k \big\} \quad \forall k \geq 0. \tag{3.10}
$$

THEOREM 2. *Let $\{x^k\}$ be generated by Algorithm 1. Assume that F is uniformly continuous in $\mathcal{L}(x^0)$. Then the following statements hold:*

(i) $\{x^k\} \subseteq \mathcal{L}(x^0)$.

(ii) $\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0$ *and there exists some* $\zeta \in \Re$ *such that*

$$
\lim_{k\to\infty} F(x^k) = \zeta. \tag{3.11}
$$

(iii) *Any accumulation point of $\{x^k\}$ is a D-stationary point of problem* (1.1).

Proof. For notational convenience, let $\bar{\alpha}_k$ be the final value of α_k generated at the kth outer loop of Algorithm 1, and let $x^{k,i} = x^{k,i}(\bar{\alpha}_k)$ for every $i \in \mathcal{A}_\eta(x^k)$. It follows

from Theorem 1 that $\underline{\alpha} \le \bar{\alpha}_k \le \tilde{\alpha}$ for all $k \ge 0$, where $\tilde{\alpha}$ is given in (3.4). Also, by (3.2), (3.10), and the updating scheme on x^{k+1}, one has that for all $i \in \mathcal{A}_\eta(x^k)$,

(3.12)
$$F(x^{k+1}) \le \max\left\{f_s(x^k) + f_n(x^k) - \psi_i(x^k), F(x^{\iota(k)})\right\} - \frac{c}{2}\|x^{k+1} - x^k\|^2 - \frac{c}{2}\|x^{k,i} - x^k\|^2.$$

We are now ready to prove statements (i)–(iii) as follows.

(i) By (3.10), we have that for all $k \ge 0$,

$$F(x^{\iota(k+1)}) = \max_{[k+1-N]_+ \le j \le k+1} F(x^j) = \max\left\{F(x^{k+1}), \max_{[k+1-N]_+ \le j \le k} F(x^j)\right\}$$

$$\le \max\left\{F(x^{k+1}), F(x^{\iota(k)})\right\}. \tag{3.13}$$

Notice that $\mathcal{A}(x^k) \subseteq \mathcal{A}_\eta(x^k)$ and $f_s(x^k) + f_n(x^k) - \psi_i(x^k) = F(x^k)$ for all $i \in \mathcal{A}(x^k)$. By this and (3.12) with $i \in \mathcal{A}(x^k)$, we have that for all $k \ge 0$,

(3.14)
$$F(x^{k+1}) \le \max\left\{F(x^k), F(x^{\iota(k)})\right\} - \frac{c}{2}\|x^{k+1} - x^k\|^2 = F(x^{\iota(k)}) - \frac{c}{2}\|x^{k+1} - x^k\|^2.$$

It then follows from (3.13) and (3.14) that

$$F(x^{\iota(k+1)}) \le \max\left\{F(x^{\iota(k)}) - \frac{c}{2}\|x^{k+1} - x^k\|^2, F(x^{\iota(k)})\right\} = F(x^{\iota(k)}) \quad \forall k \ge 0.$$

Hence, $\{F(x^{\iota(k)})\}$ is nonincreasing. This together with (3.10) and (3.14) implies that

$$F(x^{k+1}) \le F(x^{\iota(k)}) \le F(x^{\iota(0)}) = F(x^0) \quad \forall k \ge 0,$$

and hence statement (i) holds.

(ii) Recall that $\{F(x^{\iota(k)})\}$ is nonincreasing. This together with the assumption that F is bounded below implies that there exists some $\zeta \in \Re$ such that

$$\lim_{k\to\infty} F(x^{\iota(k)}) = \zeta.$$

By this, (3.14), $\{x^k\} \subseteq \mathcal{L}(x^0)$, the uniform continuity of F in $\mathcal{L}(x^0)$, and the same arguments as those in the proof of [21, Lemma 4], one can show that

$$\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0 \quad \text{and} \quad \lim_{k\to\infty} F(x^k) = \zeta.$$

The details of this proof are omitted.

(iii) Let x^∞ be an accumulation point of $\{x^k\}$. Then there exists a subsequence $\mathcal{K}$ such that $\lim_{\mathcal{K}\ni k\to\infty} x^k = x^\infty$. By the assumption that F, f_s, and ψ_i are continuous in $\mathcal{L}(x^0)$, one can see that f_n is also continuous in $\mathcal{L}(x^0)$. Using this, $\{x^k\} \subseteq \mathcal{L}(x^0)$, $\lim_{\mathcal{K}\ni k\to\infty} x^k = x^\infty$, and $\psi_i(x^\infty) = g(x^\infty)\ \forall i \in \mathcal{A}(x^\infty)$, we obtain that

$$\lim_{\mathcal{K}\ni k\to\infty} \left[f_s(x^k) + f_n(x^k) - \psi_i(x^k)\right] = f_s(x^\infty) + f_n(x^\infty) - \psi_i(x^\infty) = F(x^\infty) \tag{3.15}$$

for all $i \in \mathcal{A}(x^\infty)$. Further, by the assumption that F is continuous in $\mathcal{L}(x^0)$, one has $\lim_{\mathcal{K}\ni k\to\infty} F(x^k) = F(x^\infty)$, which together with (3.11) yields $F(x^\infty) = \zeta$. It follows from this and (3.15) that

$$\lim_{\mathcal{K}\ni k\to\infty} \left[f_s(x^k) + f_n(x^k) - \psi_i(x^k)\right] = \zeta \quad \forall i \in \mathcal{A}(x^\infty). \tag{3.16}$$

Using (1.8), $\lim_{\mathcal{K}\ni k\to\infty} x^k = x^\infty$, and the continuity of ψ_i for every $i \in \mathcal{I}$, it is not hard to see that

$$\mathcal{A}(x^\infty) \subseteq \mathcal{A}_\eta(x^k) \quad \text{for all } k \in \mathcal{K} \text{ sufficiently large.} \tag{3.17}$$

Recall from the proof of statement (ii) that we have $\lim_{k\to\infty} F(x^{\iota(k)}) = \zeta$. By this, $\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0$, $\lim_{k\to\infty} F(x^k) = \zeta$, (3.12), (3.16), and (3.17), one has that for any $i \in \mathcal{A}(x^\infty)$,

$$\begin{aligned}
&\zeta + \limsup_{\mathcal{K}\ni k\to\infty} \frac{c}{2}\|x^{k,i} - x^k\|^2 \\
&\quad = \limsup_{\mathcal{K}\ni k\to\infty} \Big[F(x^{k+1}) + \frac{c}{2}\|x^{k+1} - x^k\|^2 + \frac{c}{2}\|x^{k,i} - x^k\|^2\Big] \\
&\quad \le \limsup_{\mathcal{K}\ni k\to\infty} \max\Big\{f_s(x^k) + f_n(x^k) - \psi_i(x^k),\ F(x^{\iota(k)})\Big\} \qquad (3.18)\\
&\quad = \max\Big\{\limsup_{\mathcal{K}\ni k\to\infty} f_s(x^k) + f_n(x^k) - \psi_i(x^k),\ \limsup_{\mathcal{K}\ni k\to\infty} F(x^{\iota(k)})\Big\} \ = \ \zeta, \qquad (3.19)
\end{aligned}$$

where (3.18) and (3.19) follow from (3.12) and (3.16), respectively. It then follows from (3.19) and $c > 0$ that $\lim_{\mathcal{K}\ni k\to\infty} \|x^{k,i} - x^k\| = 0$ for all $i \in \mathcal{A}(x^\infty)$, which together with $\lim_{\mathcal{K}\ni k\to\infty} x^k = x^\infty$ implies that

$$\lim_{\mathcal{K}\ni k\to\infty} x^{k,i} = x^\infty \quad \forall i \in \mathcal{A}(x^\infty). \tag{3.20}$$

Recall that $x^{k,i} = x^{k,i}(\bar\alpha_k)$ and $\underline{\alpha} \le \bar\alpha_k \le \tilde\alpha$ for all $i \in \mathcal{A}_\eta(x^k)$ and $k \ge 0$. These together with (3.1) and (3.17) imply that for all $k \in \mathcal{K}$ sufficiently large,

$$x^{k,i} = x^{k,i}(\bar\alpha_k) = \operatorname*{argmin}_{x\in\Re^n} \Big\{\ell_i(x; x^k) + \frac{\bar\alpha_k}{2}\|x - x^k\|^2\Big\} \quad \forall i \in \mathcal{A}(x^\infty). \tag{3.21}$$

Since $\underline{\alpha} \le \bar\alpha_k \le \tilde\alpha$, by passing to a subsequence of $\mathcal{K}$ if necessary, we can assume for convenience that $\lim_{\mathcal{K}\ni k\to\infty} \bar\alpha_k = \bar\alpha_\infty$ for some $\bar\alpha_\infty \in [\underline{\alpha}, \tilde\alpha]$. In view of this, $\lim_{\mathcal{K}\ni k\to\infty} x^k = x^\infty$, (3.20), (3.21), and Corollary 1, one has that

$$x^\infty = \operatorname*{argmin}_{x\in\Re^n} \Big\{\ell_i(x; x^\infty) + \frac{\bar\alpha_\infty}{2}\|x - x^\infty\|^2\Big\} \quad \forall i \in \mathcal{A}(x^\infty),$$

whose first-order optimality condition implies that

$$0 \in \nabla f_s(x^\infty) + \partial f_n(x^\infty) - \nabla\psi_i(x^\infty) \quad \forall i \in \mathcal{A}(x^\infty).$$

It then follows from Proposition 1 that x^∞ is a D-stationary point of (1.1). ∎

3.2. A randomized nonmonotone enhanced PDCA with line search. It is nice that Algorithm 1 converges subsequentially to a D-stationary point of (1.1). However, it requires solving a number of subproblems in the form of (3.1) per iteration. Although each subproblem is assumed to be cheaply solvable, the method may become inefficient when the I in defining g is large. Inspired by a randomized algorithm proposed in [15, section 5.2], we remedy this issue by proposing a randomized counterpart of Algorithm 1 that solves a single subproblem per iteration.

ALGORITHM 2 (a randomized nonmonotone enhanced PDCA with line search).

(0) Input $x^0 \in \mathrm{dom}(F)$, $\eta > 0$, $\rho > 1$, $0 < c < L$, $0 < \underline{\alpha} \le \overline{\alpha}$, and integer $N \ge 0$. Set $k \leftarrow 0$.

(1) Pick $i_k \in \mathcal{A}_\eta(x^k)$ uniformly at random. Choose $\alpha_{k,0} \in [\underline{\alpha}, \overline{\alpha}]$.
(2) For $m = 0, 1, \ldots$:
  (2a) Let $\alpha_k = \alpha_{k,0}\rho^m$. If $\alpha_k > \alpha_{k,0}$ and $\alpha_k \geq \rho L$, set $x^{k+1} = x^k$ and go to step (3).
  (2b) Compute

$$x^{k,i_k}(\alpha_k) = \operatorname*{argmin}_{x\in\Re^n} \left\{ \ell_{i_k}(x; x^k) + \frac{\alpha_k}{2}\|x - x^k\|^2 \right\}.$$

  (2c) If $x^{k,i_k}(\alpha_k)$ satisfies

$$F(x^{k,i_k}(\alpha_k)) \leq \max_{[k-N]_+ \leq j \leq k} F(x^j) - \frac{c}{2}\|x^{k,i_k}(\alpha_k) - x^k\|^2,$$

  set $x^{k+1} = x^{k,i_k}(\alpha_k)$ and go to step (3).
(3) Set $k \leftarrow k + 1$ and go to step (1).

Similar to Algorithm 1, a popular choice of $\alpha_{k,0}$ is taken with the formula (3.3). Before studying the convergence of Algorithm 2, we introduce some notation as follows. After $k+1$ iterations, Algorithm 2 generates a random output $(x^{k+1}, F(x^{k+1}))$, which depends on the observed realization of the random vector $\xi_k = \{i_0, i_1, \ldots, i_k\}$. For convenience, let

(3.22)

$$x^{k,i}(\alpha) = \operatorname*{argmin}_{x\in\Re^n} \left\{ \ell_i(x; x^k) + \frac{\alpha}{2}\|x - x^k\|^2 \right\} \quad \forall i \in \mathcal{I},\ \alpha > 0,$$

$$\mathcal{M}(x^k) = \left\{ i \in \mathcal{A}_\eta(x^k) \,\middle|\, \begin{array}{l} F(x^{k,i}(\alpha)) \leq F(x^{\iota(k)}) - \dfrac{c}{2}\|x^{k,i}(\alpha) - x^k\|^2 \ \text{ for some} \\ \alpha = \alpha_{k,0} \text{ or } \alpha = \alpha_{k,0}\rho^m < \rho L \text{ with some integer } m \geq 1 \end{array} \right\},$$

where $\iota(k)$ is defined in (3.10). For each scenario $i \in \mathcal{A}_\eta(x^k)$, let $\hat{x}^{k+1,i}$ denote the corresponding x^{k+1} generated by the procedure detailed in step (2) of Algorithm 2 with i_k replaced by i. Let $\hat{i}$ be an arbitrary element in the set

$$\operatorname*{Argmax}_{i\in\mathcal{M}(x^k)} \|\hat{x}^{k+1,i} - x^k\|.$$

Define

$$\hat{d}^k = \hat{x}^{k+1,\hat{i}} - x^k, \qquad d^k = x^{k+1} - x^k. \tag{3.23}$$

Notice that $\|\hat{d}^k\|$ only depends on x^k, but $\|d^k\|$ depends on both x^k and the realization of i_k.

Remark 2.
(i) Using an argument similar to the proof of Theorem 1, one can show that $\mathcal{A}(x^k) \subseteq \mathcal{M}(x^k)$ and hence $\mathcal{M}(x^k) \neq \emptyset$.
(ii) One can observe that if $i \in \mathcal{A}_\eta(x^k) \setminus \mathcal{M}(x^k)$, then $\hat{x}^{k+1,i} = x^k$. Otherwise, if $i \in \mathcal{M}(x^k)$, by the definition of $\mathcal{M}(x^k)$ and $\hat{x}^{k+1,i}$, there exists some $\tilde{\alpha}_k \in [\underline{\alpha}, \tilde{\alpha}]$ such that $\hat{x}^{k+1,i} = x^{k,i}(\tilde{\alpha}_k)$ and

$$F(x^{k,i}(\tilde{\alpha}_k)) \leq F(x^{\iota(k)}) - c\|x^{k,i}(\tilde{\alpha}_k) - x^k\|^2/2,$$

where $\tilde{\alpha}$ is defined in (3.4). Combining these two cases, we can see that

$$F(\hat{x}^{k+1,i}) \leq F(x^{\iota(k)}) - \frac{c}{2}\|\hat{x}^{k+1,i} - x^k\|^2 \quad \forall i \in \mathcal{A}_\eta(x^k). \tag{3.24}$$

THEOREM 3. *Let $\{x^k\}$ be generated by Algorithm 2, and let $\{d^k\}$ and $\{\hat{d}^k\}$ be as defined in* (3.23). *Assume that F is uniformly continuous in $\mathcal{L}(x^0;\eta)$. Then the following statements hold:*

(i) $\{x^k\} \subseteq \mathcal{L}(x^0)$.

(ii) $\lim_{k\to\infty} \|d^k\| = 0$ *and*

$$\lim_{k\to\infty} F(x^k) = \lim_{k\to\infty} F(x^{\iota(k)}) = F^*_{\xi_\infty} \tag{3.25}$$

for some $F^*_{\xi_\infty} \in \Re$, where $\xi_\infty = \{i_0, i_1, \ldots\}$.

(iii) $\lim_{k\to\infty} \mathbf{E}_{\xi_k}[\|d^k\|^2] = 0$ *and*

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k)] = \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{\iota(k)})] = \mathbf{E}_{\xi_\infty}[F^*_{\xi_\infty}]. \tag{3.26}$$

(iv) $\lim_{k\to\infty} \|\hat{d}^k\| = 0$ *almost surely.*

(v) *Any accumulation point of $\{x^k\}$ is a D-stationary point of problem* (1.1) *almost surely.*

Proof. (i) By (3.10), (3.23), and the update scheme of x^{k+1} in Algorithm 2, one can observe that

$$F(x^{k+1}) \le F(x^{\iota(k)}) - \frac{c}{2}\|d^k\|^2 \quad \forall k \ge 0. \tag{3.27}$$

This together with (3.10) yields that

$$F(x^{\iota(k+1)}) = \max_{[k+1-N]_+ \le j \le k+1} F(x^j) \le \max\{F(x^{k+1}), F(x^{\iota(k)})\} \le F(x^{\iota(k)}) \quad \forall k \ge 0. \tag{3.28}$$

Hence, $\{F(x^{\iota(k)})\}$ is nonincreasing. It then follows from this, (3.10), and (3.27) that

$$F(x^{k+1}) \le F(x^{\iota(k)}) \le F(x^{\iota(0)}) = F(x^0) \quad \forall k \ge 0,$$

and hence statement (i) holds.

(ii) We know from above that $\{F(x^{\iota(k)})\}$ is nonincreasing. By the assumption that F is uniformly continuous on $\mathcal{L}(x^0;\eta)$, so is F on $\mathcal{L}(x^0)$. Using these, (3.27), and the same arguments as those in the proof of Theorem 2(ii), one can conclude that statement (ii) holds.

(iii) By (3.27), one has

$$\mathbf{E}_{\xi_k}[F(x^{k+1})] \le \mathbf{E}_{\xi_{k-1}}[F(x^{\iota(k)})] - \frac{c}{2}\mathbf{E}_{\xi_k}[\|d^k\|^2] \quad \forall k \ge 0. \tag{3.29}$$

Also, it follows from (3.28) that $\mathbf{E}_{\xi_k}[F(x^{\iota(k+1)})] \le \mathbf{E}_{\xi_{k-1}}[F(x^{\iota(k)})]$ for all $k \ge 0$. Hence, $\{\mathbf{E}_{\xi_{k-1}}[F(x^{\iota(k)})]\}$ is nonincreasing. By this and the assumption that F is bounded below, there exists some $\hat{F}^* \in \Re$ such that $\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{\iota(k)})] = \hat{F}^*$. Using this, (3.25), (3.29), $\{x^k\} \subseteq \mathcal{L}(x^0)$, the uniform continuity of F on $\mathcal{L}(x^0)$, and the same arguments as those in the proof of [12, Theorem 2.6], we have that $\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k)] = \lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^{\iota(k)})]$ and $\hat{F}^* = \mathbf{E}_{\xi_\infty}[F^*_{\xi_\infty}]$. It then follows that (3.26) holds. By (3.26) and (3.29), one has $\lim_{k\to\infty} \mathbf{E}_{\xi_k}[\|d^k\|^2] = 0$.

(iv) Let $p_k = 1/|\mathcal{A}_\eta(x^k)|$. Since $\mathcal{A}_\eta(x^k) \subseteq \mathcal{I}$, we have $p_k \ge 1/I$. Recall from (3.23) that $\hat{d}^k = \hat{x}^{k+1,\hat{i}} - x^k$ for some $\hat{i} \in \mathrm{Argmax}_{i\in\mathcal{M}(x^k)} \|\hat{x}^{k+1,i} - x^k\| \subseteq \mathcal{A}_\eta(x^k)$.

Notice from (3.24) that $F(\hat{x}^{k+1,i}) \le F(x^{\iota(k)})$ for all $i \in \mathcal{A}_\eta(x^k)$. Using these, (3.23), and (3.24), we obtain that

$$\mathbf{E}_{i_k}[F(x^{k+1})|x^k] = \sum_{i\in\mathcal{A}_\eta(x^k)} p_k F(\hat{x}^{k+1,i}) = \sum_{i\in\mathcal{A}_\eta(x^k)\setminus\{\hat{i}\}} p_k F(\hat{x}^{k+1,i}) + p_k F(\hat{x}^{k+1,\hat{i}})$$
$$\le (1-p_k)F(x^{\iota(k)}) + p_k\left(F(x^{\iota(k)}) - \frac{c}{2}\|\hat{d}^k\|^2\right) \le F(x^{\iota(k)}) - \frac{c}{2I}\|\hat{d}^k\|^2.$$

Since $x^{\iota(k)}$ is independent of i_k, one has $\mathbf{E}_{i_k}[F(x^{\iota(k)})|x^k] = F(x^{\iota(k)})$. It then follows that

$$\frac{c}{2I}\|\hat{d}^k\|^2 \le \mathbf{E}_{i_k}[F(x^{\iota(k)}) - F(x^{k+1})|x^k]. \tag{3.30}$$

By statement (i), we have $F^* \le F(x^k) \le F(x^0)$ for all k, where $F^* = \min_x F(x) > -\infty$. Hence,

$$|F(x^{\iota(k)}) - F(x^{k+1})| \le 2\max\{|F(x^0)|, |F^*|\} \quad \forall k.$$

In addition, by statement (ii), we have $\lim_{k\to\infty} F(x^{\iota(k)}) - F(x^{k+1}) = 0$. Using these and [5, Theorem 9.4.8], we obtain that

$$\lim_{k\to\infty} \mathbf{E}_{i_k}[F(x^{\iota(k)}) - F(x^{k+1})|x^k] = 0 \text{ almost surely.}$$

This together with (3.30) implies that $\lim_{k\to\infty}\|\hat{d}^k\| = 0$ almost surely.

(v) In view of statement (iv), it suffices to show that if $\lim_{k\to\infty}\|\hat{d}^k\| = 0$, any accumulation point of $\{x^k\}$ is a D-stationary point of problem (1.1). To this end, we assume that $\lim_{k\to\infty}\|\hat{d}^k\| = 0$. Let x^∞ be an arbitrary accumulation point of $\{x^k\}$. By passing to a subsequence if necessary, we can assume for convenience that $\lim_{k\to\infty} x^k = x^\infty$. Arguing for a contradiction, suppose that x^∞ is not a D-stationary point of (1.1). It then follows from Proposition 1 that there exists some $i \in \mathcal{A}(x^\infty)$ such that

$$0 \notin \nabla f_s(x^\infty) + \partial f_n(x^\infty) - \nabla\psi_i(x^\infty). \tag{3.31}$$

By $i \in \mathcal{A}(x^\infty)$, $\lim_{k\to\infty} x^k = x^\infty$, and the continuity of ψ_i, it is not hard to see that $i \in \mathcal{A}_\eta(x^k)$ for all k sufficiently large.

Claim that $i \in \mathcal{M}(x^k)$ for all k sufficiently large. Indeed, arguing for a contradiction, suppose that this claim does not hold. Recall that $i \in \mathcal{A}_\eta(x^k)$ for all k sufficiently large. Then there exists a subsequence $\mathcal{K}$ such that

$$i \in \mathcal{A}_\eta(x^k) \text{ but } i \notin \mathcal{M}(x^k) \text{ for all } k \in \mathcal{K}. \tag{3.32}$$

We first show that for all $k \in \mathcal{K}$, there exists some $\hat{\alpha}_k \in [L, \tilde{\alpha}]$ such that

$$F(x^{k,i}(\hat{\alpha}_k)) > F(x^{\iota(k)}) - \frac{c}{2}\|x^{k,i}(\hat{\alpha}_k) - x^k\|^2, \tag{3.33}$$

where $\tilde{\alpha}$ and $x^{k,i}(\alpha)$ are defined in (3.4) and (3.22), respectively. For the proof of (3.33), we consider two separate cases as follows.

Case 1 ($\alpha_{k,0} \ge L$). Notice that $\alpha_{k,0} \in [\underline{\alpha}, \overline{\alpha}]$. By these and the definition of $\tilde{\alpha}$ in (3.4), one has $\alpha_{k,0} \in [L, \tilde{\alpha}]$. This together with (3.32) and the definition of $\mathcal{M}(x^k)$ implies that (3.33) holds for $\hat{\alpha}_k = \alpha_{k,0}$.

Case 2 ($\alpha_{k,0} < L$). It follows from this, (3.32), and the definition of $\mathcal{M}(x^k)$ that there exists some $\hat{\alpha} = \alpha_{k,0}\rho^m \in [L, \rho L)$ for some integer $m \geq 1$ such that $F(x^{k,i}(\hat{\alpha})) > F(x^{\iota(k)}) - c\|x^{k,i}(\hat{\alpha}) - x^k\|^2/2$. By $\hat{\alpha} \in [L, \rho L)$ and the definition of $\tilde{\alpha}$ in (3.4), one has $\hat{\alpha} \in [L, \tilde{\alpha}]$. Hence, (3.33) holds for $\hat{\alpha}_k = \hat{\alpha}$.

Since $[L, \tilde{\alpha}]$ is compact, by passing to a subsequence of $\mathcal{K}$ if necessary, we can assume for convenience that $\lim_{\mathcal{K}\ni k\to\infty} \hat{\alpha}_k = \hat{\alpha}_\infty$ for some $\hat{\alpha}_\infty \in [L, \tilde{\alpha}]$. By this, $\lim_{k\to\infty} x^k = x^\infty$, (3.22) with $\alpha = \hat{\alpha}_k$, and Corollary 1, one has

$$\lim_{\mathcal{K}\ni k\to\infty} x^{k,i}(\hat{\alpha}_k) = x^{\infty,i}, \tag{3.34}$$

where

$$x^{\infty,i} = \operatorname*{argmin}_{x\in\Re^n} \left\{ \ell_i(x; x^\infty) + \frac{\hat{\alpha}_\infty}{2}\|x - x^\infty\|^2 \right\}. \tag{3.35}$$

Using $\{x^k\} \subseteq \mathcal{L}(x^0)$, the continuity of F on $\mathcal{L}(x^0; \eta)$, $\lim_{k\to\infty} x^k = x^\infty$ and (3.25), we obtain that

$$\lim_{k\to\infty} F(x^{\iota(k)}) = \lim_{k\to\infty} F(x^k) = F(x^\infty). \tag{3.36}$$

In addition, by (1.2), (1.10), (3.22), $\hat{\alpha}_k \geq L$ for all $k \in \mathcal{K}$, the convexity of f_s and ψ_i, and the Lipschitz continuity of ∇f_s, one has that for all $k \in \mathcal{K}$,

(3.37)

$$\begin{aligned}
F(x^{k,i}(\hat{\alpha}_k)) &\leq f_s(x^{k,i}(\hat{\alpha}_k)) + f_n(x^{k,i}(\hat{\alpha}_k)) - \psi_i(x^{k,i}(\hat{\alpha}_k)) \\
&\leq f_s(x^k) + \langle \nabla f_s(x^k), x^{k,i}(\hat{\alpha}_k) - x^k\rangle + \frac{L}{2}\|x^{k,i}(\hat{\alpha}_k) - x^k\|^2 + f_n(x^{k,i}(\hat{\alpha}_k)) \\
&\quad - \psi_i(x^k) - \langle \nabla\psi_i(x^k), x^{k,i}(\hat{\alpha}_k) - x^k\rangle && (3.38) \\
&= \ell_i(x^{k,i}(\hat{\alpha}_k); x^k) + \frac{L}{2}\|x^{k,i}(\hat{\alpha}_k) - x^k\|^2 && (3.39) \\
&\leq \ell_i(x^{k,i}(\hat{\alpha}_k); x^k) + \frac{\hat{\alpha}_k}{2}\|x^{k,i}(\hat{\alpha}_k) - x^k\|^2 && (3.40) \\
&\leq \ell_i(x^k; x^k) = f_s(x^k) + f_n(x^k) - \psi_i(x^k) && (3.41) \\
&\leq F(x^k) + \eta \leq F(x^0) + \eta, && (3.42)
\end{aligned}$$

where (3.37) is due to (1.2), (3.38) follows from the convexity of ψ_i and the Lipschitz continuity of ∇f_s, (3.39) follows from (1.10), (3.40) is due to $\hat{\alpha}_k \geq L$ for all $k \in \mathcal{K}$, (3.41) follows from (3.22), and (3.42) is due to $i \in \mathcal{A}_\eta(x^k)$ for all $k \in \mathcal{K}$ and $\{x^k\} \subseteq \mathcal{L}(x^0)$. Hence, $\{x^{k,i}(\hat{\alpha}_k)\}_{k\in\mathcal{K}} \subseteq \mathcal{L}(x^0; \eta)$, which together with the continuity of F on $\mathcal{L}(x^0; \eta)$ and (3.34) implies that

$$\lim_{\mathcal{K}\ni k\to\infty} F(x^{k,i}(\hat{\alpha}_k)) = F(x^{\infty,i}).$$

Using this, (3.34) and (3.36), and taking limit on both sides of (3.33) as $\mathcal{K} \ni k \to \infty$, we obtain that

$$F(x^{\infty,i}) \geq F(x^\infty) - \frac{c}{2}\|x^{\infty,i} - x^\infty\|^2. \tag{3.43}$$

On the other hand, by the same arguments as those for deriving (3.39), one has

$$F(x^{\infty,i}) \leq \ell_i(x^{\infty,i}; x^\infty) + \frac{L}{2}\|x^{\infty,i} - x^\infty\|^2. \tag{3.44}$$

Also, by (3.35) and the fact that the objective function in (3.35) is strongly convex with modulus $\hat{\alpha}_\infty$, we have

$$\ell_i(x^{\infty,i};x^\infty)+\frac{\hat{\alpha}_\infty}{2}\|x^{\infty,i}-x^\infty\|^2\le \ell_i(x^\infty;x^\infty)-\frac{\hat{\alpha}_\infty}{2}\|x^{\infty,i}-x^\infty\|^2 = F(x^\infty)-\frac{\hat{\alpha}_\infty}{2}\|x^{\infty,i}-x^\infty\|^2,$$

which together with (3.44) and $\hat{\alpha}_\infty \ge L$ yields

$$F(x^{\infty,i})\le F(x^\infty)-\frac{\hat{\alpha}_\infty}{2}\|x^{\infty,i}-x^\infty\|^2. \tag{3.45}$$

It then follows from this inequality, (3.43), and $\hat{\alpha}_\infty \ge L > c$ that $x^{\infty,i}=x^\infty$. Combining this with (3.35) and using the first-order optimality condition of (3.35), we have

$$0\in \nabla f_s(x^\infty)+\partial f_n(x^\infty)-\nabla\psi_i(x^\infty),$$

which contradicts (3.31). Therefore, the above claim holds as desired, that is, $i\in\mathcal{M}(x^k)$ for all k sufficiently large.

Since $i\in\mathcal{M}(x^k)$ for all k sufficiently large, it follows from Remark 2(ii) that there exists some $\tilde{\alpha}_k\in[\underline{\alpha},\tilde{\alpha}]$ such that $\hat{x}^{k+1,i}=x^{k,i}(\tilde{\alpha}_k)$. Since $[\underline{\alpha},\tilde{\alpha}]$ is compact, by passing to a subsequence if necessary, we can assume for convenience that $\lim_{k\to\infty}\tilde{\alpha}_k=\tilde{\alpha}_\infty$ for some $\tilde{\alpha}_\infty\in[\underline{\alpha},\tilde{\alpha}]$. By this, $\lim_{k\to\infty}x^k=x^\infty$, $\hat{x}^{k+1,i}=x^{k,i}(\tilde{\alpha}_k)$, and (3.22), we obtain from Corollary 1 that

$$\lim_{k\to\infty}\hat{x}^{k+1,i}=\lim_{k\to\infty}x^{k,i}(\tilde{\alpha}_k)=\tilde{x}^{\infty,i}, \tag{3.46}$$

where

$$\tilde{x}^{\infty,i}=\operatorname*{argmin}_{x\in\Re^n}\left\{\ell_i(x;x^\infty)+\frac{\tilde{\alpha}_\infty}{2}\|x-x^\infty\|^2\right\}. \tag{3.47}$$

Recall that $i\in\mathcal{M}(x^k)$ for all k sufficiently large. By this, (3.23), and the assumption that $\lim_{k\to\infty}\|\hat{d}^k\|=0$, one has $\lim_{k\to\infty}\|\hat{x}^{k+1,i}-x^k\|=0$. This together with (3.46) and $\lim_{k\to\infty}x^k=x^\infty$ implies that $\tilde{x}^{\infty,i}=x^\infty$. Combining this with (3.47), and using the first-order optimality condition of (3.47) and $\tilde{\alpha}_\infty\ge\underline{\alpha}>0$, we obtain that

$$0\in \nabla f_s(x^\infty)+\partial f_n(x^\infty)-\nabla\psi_i(x^\infty),$$

which contradicts (3.31). Hence, if $\lim_{k\to\infty}\|\hat{d}^k\|=0$, any accumulation point of $\{x^k\}$ is a D-stationary point of (1.1). This along with statement (iv) leads to the conclusion in (v). ∎

4. Nonmonotone enhanced proximal DCA with extrapolation. In this section we propose a nonmonotone enhanced PDCA for solving problem (1.1) in which an extrapolation scheme is applied. We also propose a randomized counterpart for this method, and establish convergence for both methods.

4.1. A deterministic nonmonotone enhanced PDCA with extrapolation. In this subsection we present a deterministic nonmonotone enhanced PDCA with extrapolation for solving problem (1.1), and study its convergence properties.

ALGORITHM 3 (a deterministic nonmonotone enhanced PDCA with extrapolation).

(0) Input $x^0 \in \mathrm{dom}(F)$, $\eta > 0$, $0 \le c < L$, and $\beta \in [0, \sqrt{c/L}) \cup \{0\}$. Set $x^{-1} = x^0$ and $k \leftarrow 0$.

(1) Choose $\beta_k \in [0, \beta]$ arbitrarily. Set $z^k = x^k + \beta_k(x^k - x^{k-1})$.

(2) For each $i \in \mathcal{A}_\eta(x^k)$, compute $x^{k,i}$ as

$$x^{k,i} = \mathop{\mathrm{argmin}}_{x \in \Re^n} \left\{ \ell_i(x; z^k, x^k) + \frac{L}{2}\|x - z^k\|^2 \right\}. \tag{4.1}$$

(3) Let

$$\hat{i} \in \mathop{\mathrm{Argmin}}_{i \in \mathcal{A}_\eta(x^k)} \left\{ F(x^{k,i}) + \frac{c}{2}\|x^{k,i} - x^k\|^2 \right\}.$$

Set $x^{k+1} = x^{k,\hat{i}}$.

(4) Set $k \leftarrow k + 1$ and go to step (1).

THEOREM 4. *Let $\{x^k\}$ be generated by Algorithm 3. Then the following statements hold:*

(i) $\{x^k\} \subseteq \mathcal{L}(x^0)$.

(ii) $\lim_{k\to\infty} \|x^k - x^{k-1}\| = 0$ *if* $c > 0$.

(iii) $\lim_{k\to\infty} F(x^k)$ *exists and* $\lim_{k\to\infty} F(x^k) = F(x^\infty)$ *for any accumulation point* x^∞ *of* $\{x^k\}$.

(iv) *Any accumulation point of* $\{x^k\}$ *is a D-stationary point of problem* (1.1).

Proof. (i) Notice that the objective function in (4.1) is strongly convex with modulus L. Hence, for all $k \ge 0$ and $i \in \mathcal{A}_\eta(x^k)$, we obtain that

$$\ell_i(x^k; z^k, x^k) + \frac{L}{2}\|x^k - z^k\|^2 \ge \ell_i(x^{k,i}; z^k, x^k) + \frac{L}{2}\|x^{k,i} - z^k\|^2 + \frac{L}{2}\|x^{k,i} - x^k\|^2. \tag{4.2}$$

By this, (1.9), the convexity of f_s and ψ_i, the Lipschitz continuity of ∇f_s, and the update scheme of x^{k+1}, we have that for all $k \ge 0$ and $i \in \mathcal{A}_\eta(x^k)$,

$$\begin{aligned}
&f_s(x^k) + f_n(x^k) - \psi_i(x^k) \\
&\ge\ f_s(z^k) + \langle \nabla f_s(z^k), x^k - z^k\rangle + f_n(x^k) - \psi_i(x^k) = \ell_i(x^k; z^k, x^k) && (4.3)\\
&\ge \ell_i(x^{k,i}; z^k, x^k) + \frac{L}{2}\|x^{k,i} - z^k\|^2 + \frac{L}{2}\|x^{k,i} - x^k\|^2 - \frac{L}{2}\|x^k - z^k\|^2 && (4.4)\\
&= f_s(z^k) + \langle \nabla f_s(z^k), x^{k,i} - z^k\rangle + f_n(x^{k,i}) - \psi_i(x^k) - \langle \nabla \psi_i(x^k), x^{k,i} - x^k\rangle \\
&\quad + \frac{L}{2}\|x^{k,i} - z^k\|^2 + \frac{L}{2}\|x^{k,i} - x^k\|^2 - \frac{L}{2}\|x^k - z^k\|^2 && (4.5)\\
&\ge f_s(x^{k,i}) + f_n(x^{k,i}) - \psi_i(x^{k,i}) + \frac{L}{2}\|x^{k,i} - x^k\|^2 - \frac{L}{2}\|x^k - z^k\|^2 && (4.6)\\
&\ge F(x^{k,i}) + \frac{L}{2}\|x^{k,i} - x^k\|^2 - \frac{L}{2}\|x^k - z^k\|^2 && (4.7)\\
&\ge F(x^{k+1}) + \frac{c}{2}\|x^{k+1} - x^k\|^2 + \frac{L-c}{2}\|x^{k,i} - x^k\|^2 - \frac{L}{2}\|x^k - z^k\|^2 && (4.8)\\
&\ge F(x^{k+1}) + \frac{c}{2}\|x^{k+1} - x^k\|^2 - \frac{L}{2}\|x^k - z^k\|^2, && (4.9)
\end{aligned}$$

where (4.3) is due to (1.9) and the convexity of f_s, (4.4) and (4.5) follow respectively from (4.2) and (1.9), (4.6) is due to the Lipschitz continuity of ∇f_s and the convexity

of ψ_i, (4.8) follows from the update scheme of x^{k+1}, and (4.9) is due to $L > c$. Notice that $\mathcal{A}(x^k) \subseteq \mathcal{A}_\eta(x^k)$. It follows from (4.9) with $i \in \mathcal{A}(x^k)$ that

$$F(x^k) = f_s(x^k) + f_n(x^k) - \psi_i(x^k) \geq F(x^{k+1}) + \frac{c}{2}\|x^{k+1} - x^k\|^2 - \frac{L}{2}\|x^k - z^k\|^2.$$

This together with $\beta_k \in [0, \beta]$ and $z^k = x^k + \beta_k(x^k - x^{k-1})$ implies that

$$F(x^{k+1}) + \frac{c}{2}\|x^{k+1} - x^k\|^2 \leq F(x^k) + \frac{L}{2}\|x^k - z^k\|^2 \leq F(x^k) + \frac{L\beta^2}{2}\|x^k - x^{k-1}\|^2. \tag{4.10}$$

Since $\beta \in [0, \sqrt{c/L}) \cup \{0\}$, we have $\beta^2 \leq c/L$. This together with (4.10) yields

$$F(x^{k+1}) + \frac{c}{2}\|x^{k+1} - x^k\|^2 \leq F(x^k) + \frac{c}{2}\|x^k - x^{k-1}\|^2 \quad \forall k \geq 0,$$

that is, $\{F(x^k) + c\|x^k - x^{k-1}\|^2/2\}$ is nonincreasing. This implies that

$$F(x^k) \leq F(x^k) + \frac{c}{2}\|x^k - x^{k-1}\|^2 \leq F(x^0) + \frac{c}{2}\|x^0 - x^{-1}\|^2 = F(x^0) \quad \forall k \geq 0.$$

Hence, statement (i) holds.

(ii) Suppose $c > 0$. By this and $\beta \in [0, \sqrt{c/L}) \cup \{0\}$, one has $c - L\beta^2 > 0$. In addition, by (4.10), we have that for all $k \geq 0$,

$$\begin{aligned} &\frac{c - L\beta^2}{2}\|x^{k+1} - x^k\|^2 \\ &\quad \leq \left(F(x^k) + \frac{L\beta^2}{2}\|x^k - x^{k-1}\|^2\right) - \left(F(x^{k+1}) + \frac{L\beta^2}{2}\|x^{k+1} - x^k\|^2\right), \end{aligned}$$

which together with $x^0 = x^{-1}$ yields that

$$\frac{c - L\beta^2}{2}\sum_{k=0}^{j}\|x^k - x^{k-1}\|^2 \leq F(x^0) - F(x^j) \leq F(x^0) - F^* \quad \forall j \geq 0.$$

It then follows from this and $c - L\beta^2 > 0$ that $\lim_{k\to\infty}\|x^k - x^{k-1}\| = 0$.

(iii) Recall that $\{F(x^k) + c\|x^k - x^{k-1}\|^2/2\}$ is nonincreasing and F is bounded below. Hence, $\lim_{k\to\infty}\{F(x^k) + c\|x^k - x^{k-1}\|^2/2\}$, denoted by ζ, exists. We next show that

$$\lim_{k\to\infty} F(x^k) = \lim_{k\to\infty}\left\{F(x^k) + \frac{c}{2}\|x^k - x^{k-1}\|^2\right\} = \zeta. \tag{4.11}$$

Indeed, if $c = 0$, (4.11) clearly holds. On the other hand, if $c > 0$, we know from statement (ii) that $\lim_{k\to\infty}\|x^k - x^{k-1}\| = 0$, which yields (4.11).

Let x^∞ be any accumulation point of $\{x^k\}$. Suppose that $\mathcal{K}$ is a subsequence such that $\lim_{\mathcal{K}\ni k\to\infty} x^k = x^\infty$. By this, (1.8), and the continuity of ψ_i for every $i \in \mathcal{I}$, it is not hard to see that

$$\mathcal{A}(x^\infty) \subseteq \mathcal{A}_\eta(x^k) \quad \text{for all } k \in \mathcal{K} \text{ sufficiently large.} \tag{4.12}$$

Claim that

$$\lim_{\mathcal{K}\ni k\to\infty} z^k = \lim_{\mathcal{K}\ni k\to\infty} x^k = x^\infty. \tag{4.13}$$

Indeed, if $c = 0$, it follows from $\beta \in [0, \sqrt{c/L}) \cup \{0\}$ that $\beta = 0$, which together with $z^k = x^k + \beta_k(x^k - x^{k-1})$ and $\beta_k \in [0, \beta]$ yields that $z^k = x^k$ for all k. Hence, (4.13) holds for $c = 0$. On the other hand, if $c > 0$, we have from statement (ii) that $\lim_{k\to\infty} \|x^k - x^{k-1}\| = 0$. By this, $z^k = x^k + \beta_k(x^k - x^{k-1})$, and $\beta_k \in [0, \beta]$, one can easily see that (4.13) also holds for $c > 0$. We next show that

$$\lim_{\mathcal{K}\ni k\to\infty} x^{k,i} = \lim_{\mathcal{K}\ni k\to\infty} x^k = x^\infty \quad \forall i \in \mathcal{A}(x^\infty). \tag{4.14}$$

To this end, let $i \in \mathcal{A}(x^\infty)$ be arbitrarily chosen. By $\lim_{\mathcal{K}\ni k\to\infty} x^k = x^\infty$, the continuity of ψ_i, and $i \in \mathcal{A}(x^\infty)$, one can see that $\lim_{\mathcal{K}\ni k\to\infty} \psi_i(x^k) = \psi_i(x^\infty) = g(x^\infty)$. This together with (4.11) and the continuity of g implies that

$$\begin{aligned}\lim_{\mathcal{K}\ni k\to\infty} \big\{f_s(x^k) + f_n(x^k)\big\} = \lim_{\mathcal{K}\ni k\to\infty} \big\{F(x^k) + g(x^k)\big\} &= \zeta + g(x^\infty)\\ &= \zeta + \lim_{\mathcal{K}\ni k\to\infty} \psi_i(x^k),\end{aligned}$$

which leads to

$$\lim_{\mathcal{K}\ni k\to\infty} \big\{f_s(x^k) + f_n(x^k) - \psi_i(x^k)\big\} = \zeta. \tag{4.15}$$

By $i \in \mathcal{A}(x^\infty)$ and (4.12), one has that $i \in \mathcal{A}_\eta(x^k)$ for all $k \in \mathcal{K}$ sufficiently large. Using this, (4.8), (4.11), (4.13), and (4.15), we obtain that

$$\begin{aligned}\zeta &= \lim_{\mathcal{K}\ni k\to\infty} \big\{f_s(x^k) + f_n(x^k) - \psi_i(x^k)\big\}\\ &\ge \limsup_{\mathcal{K}\ni k\to\infty} \bigg\{F(x^{k+1}) + \frac{c}{2}\|x^{k+1} - x^k\|^2\\ &\qquad + \frac{L-c}{2}\|x^{k,i} - x^k\|^2 - \frac{L}{2}\|x^k - z^k\|^2\bigg\} \qquad (4.16)\\ &= \zeta + \limsup_{\mathcal{K}\ni k\to\infty} \frac{L-c}{2}\|x^{k,i} - x^k\|^2, \qquad (4.17)\end{aligned}$$

where (4.16) is due to (4.8), and (4.17) follows from (4.11) and (4.13). It then follows from (4.17) and $L > c$ that $\lim_{\mathcal{K}\ni k\to\infty} \|x^{k,i} - x^k\| = 0$, which along with $\lim_{\mathcal{K}\ni k\to\infty} x^k = x^\infty$ and the arbitrariness of $i \in \mathcal{A}(x^\infty)$ implies that (4.14) holds.

We are now ready to show that $F(x^\infty) = \lim_{k\to\infty} F(x^k)$. Indeed, by (4.1), (4.5), and (4.9), we have that for all $i \in \mathcal{A}_\eta(x^k)$,

$$\begin{aligned}&F(x^{k+1}) + \frac{c}{2}\|x^k - x^{k+1}\|^2\\ &\le f_s(z^k) + f_n(x^{k,i}) + \langle \nabla f_s(z^k), x^{k,i} - z^k\rangle - \psi_i(x^k) - \langle \nabla\psi_i(x^k), x^{k,i} - x^k\rangle\\ &\quad + \frac{L}{2}\|x^{k,i} - z^k\|^2 + \frac{L}{2}\|x^{k,i} - x^k\|^2 \qquad (4.18)\\ &\le f_s(z^k) + f_n(x) + \langle \nabla f_s(z^k), x - z^k\rangle - \psi_i(x^k) - \langle \nabla\psi_i(x^k), x - x^k\rangle\\ &\quad + \frac{L}{2}\|x - z^k\|^2 + \frac{L}{2}\|x^{k,i} - x^k\|^2 \quad \forall x \in \Re^n, \qquad (4.19)\end{aligned}$$

where (4.18) follows from (4.5) and (4.9), and (4.19) is due to (4.1). Recall from Assumption 1 that f_s, ∇f_s, ψ_i, and $\nabla\psi_i$ are continuous. Using this, (4.11), (4.13),

(4.14), and taking the limit of both sides of (4.19) as $\mathcal{K} \ni k \to \infty$, we obtain that

$$\begin{aligned}\zeta &\le f_s(x^\infty) + f_n(x) - \psi_i(x^\infty) + \langle \nabla f_s(x^\infty) - \nabla \psi_i(x^\infty), x - x^\infty \rangle + \frac{L}{2}\|x - x^\infty\|^2 \\ &\le f(x) - \psi_i(x^\infty) - \langle \nabla \psi_i(x^\infty), x - x^\infty \rangle + \frac{L}{2}\|x - x^\infty\|^2 \quad \forall x \in \Re^n,\ \forall i \in \mathcal{A}(x^\infty),\end{aligned} \tag{4.20}$$

where (4.20) follows from the convexity of f_s and $f = f_s + f_n$. Letting $x = x^\infty$ in (4.20), we have $\zeta \le f(x^\infty) - \psi_i(x^\infty)$ for all $i \in \mathcal{A}(x^\infty)$, which along with (1.8) yields $\zeta \le F(x^\infty)$. On the other hand, by $\lim_{\mathcal{K} \ni k \to \infty} x^k = x^\infty$, $\zeta = \lim_{k\to\infty} F(x^k)$, and the lower-semicontinuity of F, one has $F(x^\infty) \le \zeta$. Hence, $\lim_{k\to\infty} F(x^k) = \zeta = F(x^\infty)$.

(iv) By $\zeta = F(x^\infty)$, it follows from (4.20) that

$$F(x^\infty) \le f(x) - \psi_i(x^\infty) - \langle \nabla \psi_i(x^\infty), x - x^\infty \rangle + \frac{L}{2}\|x - x^\infty\|^2 \quad \forall x \in \Re^n,\ \forall i \in \mathcal{A}(x^\infty),$$

which together with $F(x^\infty) = f(x^\infty) - \psi_i(x^\infty)$ for all $i \in \mathcal{A}(x^\infty)$ implies that

$$x^\infty = \operatorname*{argmin}_{x \in \Re^n} \left\{ f(x) - \langle \nabla \psi_i(x^\infty), x - x^\infty \rangle + \frac{L}{2}\|x - x^\infty\|^2 \right\} \quad \forall i \in \mathcal{A}(x^\infty).$$

Its first-order optimality condition yields that

$$0 \in \nabla f_s(x^\infty) + \partial f_n(x^\infty) - \nabla \psi_i(x^\infty) \quad \forall i \in \mathcal{A}(x^\infty).$$

It then follows from Proposition 1 that x^∞ is a D-stationary point of (1.1). ∎

Before ending this subsection, we make some remarks on the difference between Algorithm 3 and EPDCA [13, Algorithm 2].

(i) The main difference between Algorithm 3 and EPDCA is that instead of (4.1), EPDCA computes $x^{k,i}$ as

$$x^{k,i} = \operatorname*{argmin}_{x \in \Re^n} \left\{ \ell_i(x; z^k, x^k) + \frac{c}{2}\|x - x^k\|^2 + \frac{L}{2}\|x - z^k\|^2 \right\}. \tag{4.21}$$

Compared to (4.1), subproblem (4.21) has an extra proximal term $\frac{c}{2}\|x - x^k\|^2$ in (4.21), which can lead to a slow convergence for EPDCA. In fact, to make the extrapolation scheme effective, $\{\beta_k\}$ shall not be chosen too small, which together with $\beta_k \le \beta \le \sqrt{c/L}$ implies that c shall not be chosen too small. One can observe from (4.21) that a large c typically results in a small step size of EPDCA and thus renders a slow convergence for EPDCA. Nevertheless, one can see from (4.1) that the step size of Algorithm 3 does not depend on c and a large c can be chosen in Algorithm 3 to make the extrapolation scheme effective. Due to these, EPDCA generally has a slower convergence than Algorithm 3.

(ii) The proof of subsequential convergence to a D-stationary point of problem (1.1) for Algorithm 3 is very different from that for EPDCA [13]. In particular, the property (4.14) is a new observation and plays a crucial role in the proof of Theorem 4.

(iii) From a theoretical point of view, the convergence property of Algorithm 3 is weaker than that of EPDCA. In particular, it is shown in [13, Theorem 2] that any accumulation point of the solution sequence generated by EPDCA

is an $(\alpha,\tilde{\eta})$-D-stationary point of problem (1.1) for any $\alpha\in(0,(L+c)^{-1}]$ and $\tilde{\eta}\in[0,\eta)$, which is generally stronger than a D-stationary point. It is not clear, however, whether or not such a result holds for Algorithm 3. We shall leave this for our future research.

4.2. A randomized nonmonotone enhanced PDCA with extrapolation. For the same reason as mentioned in section 3.2, Algorithm 3 may become inefficient when the I in defining g is large. Inspired by a randomized algorithm proposed in [15, section 5.2], we remedy this issue by proposing its randomized counterpart as follows.

ALGORITHM 4 (a randomized nonmonotone enhanced PDCA with extrapolation).

(0) Input $x^0\in\mathrm{dom}(F)$, $0<c<L$, and $0\le\beta<\sqrt{c/L}$. Set $x^{-1}=x^0$ and $k\leftarrow 0$.

(1) Choose $\beta_k\in[0,\beta]$ arbitrarily. Set $z^k=x^k+\beta_k(x^k-x^{k-1})$.

(2) Pick $i_k\in\mathcal{A}_\eta(x^k)$ uniformly at random. Compute

$$x^{k,i_k}=\mathop{\mathrm{argmin}}_{x\in\Re^n}\left\{\ell_{i_k}(x;z^k,x^k)+\frac{L}{2}\|x-z^k\|^2\right\}.$$

(3) Set

$$x^{k+1}=\begin{cases}x^{k,i_k} & \text{if } F(x^{k,i_k})+\frac{c}{2}\|x^{k,i_k}-x^k\|^2\le F(x^k)+\frac{L}{2}\|x^k-z^k\|^2,\\ x^k & \text{otherwise.}\end{cases}\tag{4.22}$$

(4) Set $k\leftarrow k+1$ and go to step (1).

Before studying the convergence of Algorithm 4, we introduce the following notation. After $k+1$ iterations, Algorithm 4 generates a random output $(x^{k+1},F(x^{k+1}))$, which depends on the observed realization of the random vector $\xi_k=\{i_0,i_1,\ldots,i_k\}$. For each scenario $i\in\mathcal{A}_\eta(x^k)$, let $x^{k,i}$ and $\hat{x}^{k+1,i}$ respectively denote the corresponding x^{k,i_k} and x^{k+1} generated by the procedure detailed in steps (2) and (3) of Algorithm 4 with i_k replaced by i. Define

$$\widehat{\mathcal{M}}(x^k)=\left\{i\in\mathcal{A}_\eta(x^k)\,\middle|\,F(x^{k,i})+\frac{c}{2}\|x^{k,i}-x^k\|^2\le F(x^k)+\frac{L}{2}\|x^k-z^k\|^2\right\}.\tag{4.23}$$

Let $\hat{i}$ be an arbitrary element in the set $\mathrm{Argmax}_{i\in\widehat{\mathcal{M}}(x^k)}\|\hat{x}^{k+1,i}-x^k\|$. Define

$$\hat{d}^k=\hat{x}^{k+1,\hat{i}}-x^k,\qquad d^k=x^{k+1}-x^k.\tag{4.24}$$

Notice that $\|\hat{d}^k\|$ only depends on x^k, but $\|d^k\|$ depends on both x^k and the realization of i_k.

Remark 3.

(i) By the same arguments as those for deriving (4.7), one can show that

$$F(x^k)\ge F(x^{k,i})+\frac{L}{2}\|x^{k,i}-x^k\|^2-\frac{L}{2}\|x^k-z^k\|^2\quad\forall i\in\mathcal{A}(x^k).$$

It then follows from this and $c<L$ that $\mathcal{A}(x^k)\subseteq\widehat{\mathcal{M}}(x^k)$ and therefore $\widehat{\mathcal{M}}(x^k)\neq\emptyset$.

(ii) One can observe that if $i \in \mathcal{A}_\eta(x^k) \setminus \widehat{\mathcal{M}}(x^k)$, then $\hat{x}^{k+1,i} = x^k$. Otherwise, if $i \in \widehat{\mathcal{M}}(x^k)$, by the definition of $\widehat{\mathcal{M}}(x^k)$ and $\hat{x}^{k+1,i}$, one can observe that $\hat{x}^{k+1,i} = x^{k,i}$ and $F(x^{k,i}) + c\|x^{k,i} - x^k\|^2/2 \le F(x^k) + L\|x^k - z^k\|^2/2$. Combining these two cases, one can see that

$$F(\hat{x}^{k+1,i}) + \frac{c}{2}\|\hat{x}^{k+1,i} - x^k\|^2 \le F(x^k) + \frac{L}{2}\|x^k - z^k\|^2 \quad \forall i \in \mathcal{A}_\eta(x^k). \tag{4.25}$$

THEOREM 5. *Let $\{x^k\}$ be generated by Algorithm 4, and let $\{d^k\}$ and $\{\hat{d}^k\}$ be defined in* (4.24). *Assume that F is continuous on $\mathcal{L}(x^0;\eta)$. Then the following statements hold:*

(i) $\{x^k\} \subseteq \mathcal{L}(x^0)$.

(ii) $\lim_{k\to\infty} \|d^k\| = 0$ *and* $\lim_{k\to\infty} F(x^k) = F^*_{\xi_\infty}$ *for some* $F^*_{\xi_\infty} \in \Re$, *where* $\xi_\infty = \{i_0, i_1, \ldots\}$.

(iii) $\lim_{k\to\infty} \mathbf{E}_{\xi_k}[\|d^k\|^2] = 0$ *and* $\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k)] = \mathbf{E}_{\xi_\infty}[F^*_{\xi_\infty}]$.

(iv) $\lim_{k\to\infty} \|\hat{d}^k\| = 0$ *almost surely.*

(v) *Any accumulation point of $\{x^k\}$ is a D-stationary point of problem* (1.1) *almost surely.*

Proof. (i) By (4.22), one can observe that

$$F(x^{k+1}) + \frac{c}{2}\|x^{k+1} - x^k\|^2 \le F(x^k) + \frac{L}{2}\|x^k - z^k\|^2,$$

which along with (4.24), $z^k = x^k + \beta_k(x^k - x^{k-1})$, and $\beta_k \in [0, \beta]$ yields

$$F(x^{k+1}) + \frac{c}{2}\|d^k\|^2 \le F(x^k) + \frac{L\beta_k^2}{2}\|x^k - x^{k-1}\|^2 \le F(x^k) + \frac{L\beta^2}{2}\|d^{k-1}\|^2.$$

It then follows that

$$\frac{c - L\beta^2}{2}\|d^k\|^2 \le \left(F(x^k) + \frac{L\beta^2}{2}\|d^{k-1}\|^2\right) - \left(F(x^{k+1}) + \frac{L\beta^2}{2}\|d^k\|^2\right) \quad \forall k \ge 0, \tag{4.26}$$

which along with $\beta \in [0, \sqrt{c/L})$ implies that $\{F(x^k) + L\beta^2\|d^{k-1}\|^2/2\}$ is nonincreasing. By this and $d^{-1} = x^0 - x^{-1} = 0$, one has that

$$F(x^k) \le F(x^k) + \frac{L\beta^2}{2}\|d^{k-1}\|^2 \le F(x^0) + \frac{L\beta^2}{2}\|d^{-1}\|^2 = F(x^0) \quad \forall k \ge 0, \tag{4.27}$$

and hence $\{x^k\} \subseteq \mathcal{L}(x^0)$.

(ii) Recall that $\{F(x^k) + L\beta^2\|d^{k-1}\|^2/2\}$ is nonincreasing and F is bounded below. Hence, there exists some $F^*_{\xi_\infty} \in \Re$ such that

$$\lim_{k\to\infty} \left\{F(x^k) + \frac{L\beta^2}{2}\|d^{k-1}\|^2\right\} = F^*_{\xi_\infty}. \tag{4.28}$$

Using this, (4.26), and $\beta \in [0, \sqrt{c/L})$, we have that $\lim_{k\to\infty} \|d^k\| = 0$, which together with (4.28) implies that $\lim_{k\to\infty} F(x^k) = F^*_{\xi_\infty}$.

(iii) It follows from (4.26) that

$$\begin{aligned} &\frac{c - L\beta^2}{2}\mathbf{E}_{\xi_k}[\|d^k\|^2] \\ &\quad \le \mathbf{E}_{\xi_{k-1}}\left[F(x^k) + \frac{L\beta^2}{2}\|d^{k-1}\|^2\right] - \mathbf{E}_{\xi_k}\left[F(x^{k+1}) + \frac{L\beta^2}{2}\|d^k\|^2\right] \quad \forall k \ge 0. \end{aligned} \tag{4.29}$$

By this and $\beta \in [0, \sqrt{c/L})$, one can see that $\{\mathbf{E}_{\xi_{k-1}}[F(x^k) + L\beta^2\|d^{k-1}\|^2/2]\}$ is nonincreasing. Since F is bounded below, there exists some $\hat{F}^* \in \Re$ such that

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}\left[F(x^k) + \frac{L\beta^2}{2}\|d^{k-1}\|^2\right] = \hat{F}^*. \tag{4.30}$$

It then follows from this, (4.29), and $\beta \in [0, \sqrt{c/L})$ that $\lim_{k\to\infty} \mathbf{E}_{\xi_k}[\|d^k\|^2] = 0$. By this and (4.30), we obtain that

$$\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}\left[F(x^k)\right] = \hat{F}^*. \tag{4.31}$$

To prove $\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k)] = \mathbf{E}_{\xi_\infty}[F^*_{\xi_\infty}]$, it thus suffices to show $\hat{F}^* = \mathbf{E}_{\xi_\infty}[F^*_{\xi_\infty}]$. Indeed, recall from statement (i) that $\{x^k\} \subseteq \mathcal{L}(x^0)$. Hence, $F(x^0) \geq F(x^k) \geq F^*$ for all $k \geq 0$. It then follows that

$$|F(x^k)| \leq \max\{|F(x^0)|, |F^*|\} \quad \forall k \geq 0.$$

Using this and the dominated convergence theorem, we have

$$\lim_{k\to\infty} \mathbf{E}_{\xi_\infty}[F(x^k)] = \mathbf{E}_{\xi_\infty}\left[\lim_{k\to\infty} F(x^k)\right] = \mathbf{E}_{\xi_\infty}[F^*_{\xi_\infty}],$$

which along with $\lim_{k\to\infty} \mathbf{E}_{\xi_{k-1}}[F(x^k)] = \lim_{k\to\infty} \mathbf{E}_{\xi_\infty}[F(x^k)]$ and (4.31) implies that $\hat{F}^* = \mathbf{E}_{\xi_\infty}[F^*_{\xi_\infty}]$.

(iv) Let $p_k = 1/|\mathcal{A}_\eta(x^k)|$. Since $\mathcal{A}_\eta(x^k) \subseteq \mathcal{I}$, we have $p_k \geq 1/I$. By (4.24), (4.25), $z^k = x^k + \beta_k(x^k - x^{k-1})$, and $\beta_k \in [0, \beta]$, one has

$$F(\hat{x}^{k+1,i}) + \frac{c}{2}\|\hat{x}^{k+1,i} - x^k\|^2 \leq F(x^k) + \frac{L\beta_k^2}{2}\|x^k - x^{k-1}\|^2 \leq F(x^k) + \frac{L\beta^2}{2}\|d^{k-1}\|^2$$

for all $i \in \mathcal{A}_\eta(x^k)$. It then follows that

$$F(\hat{x}^{k+1,i}) + \frac{L\beta^2}{2}\|\hat{x}^{k+1,i} - x^k\|^2 \leq F(x^k) + \frac{L\beta^2}{2}\|d^{k-1}\|^2 - \frac{c - L\beta^2}{2}\|\hat{x}^{k+1,i} - x^k\|^2 \tag{4.32}$$

for all $i \in \mathcal{A}_\eta(x^k)$. Recall from (4.24) that $\hat{d}^k = \hat{x}^{k+1,\hat{i}} - x^k$ for some

$$\hat{i} \in \operatorname*{Argmax}_{i\in\widehat{\mathcal{M}}(x^k)} \|\hat{x}^{k+1,i} - x^k\| \subseteq \mathcal{A}_\eta(x^k).$$

By this, $\beta \in [0, \sqrt{c/L})$, the update scheme on x^{k+1}, (4.24), and (4.32), we obtain that

$$\begin{aligned}
&\mathbf{E}_{i_k}\left[F(x^{k+1}) + \frac{L\beta^2}{2}\|d^k\|^2 \;\middle|\; x^k\right] \\
&= \sum_{i\in\mathcal{A}_\eta(x^k)} p_k\left(F(\hat{x}^{k+1,i}) + \frac{L\beta^2}{2}\|\hat{x}^{k+1,i} - x^k\|^2\right) \\
&\leq (1-p_k)\left(F(x^k) + \frac{L\beta^2}{2}\|d^{k-1}\|^2\right) + p_k\left(F(\hat{x}^{k+1,\hat{i}}) + \frac{L\beta^2}{2}\|\hat{x}^{k+1,\hat{i}} - x^k\|^2\right) \qquad (4.33)\\
&\leq (1-p_k)\left(F(x^k) + \frac{L\beta^2}{2}\|d^{k-1}\|^2\right) \\
&\quad + p_k\left(F(x^k) + \frac{L\beta^2}{2}\|d^{k-1}\|^2 - \frac{c - L\beta^2}{2}\|\hat{x}^{k+1,\hat{i}} - x^k\|^2\right) \qquad (4.34)\\
&\leq F(x^k) + \frac{L\beta^2}{2}\|d^{k-1}\|^2 - \frac{c - L\beta^2}{2I}\|\hat{d}^k\|^2, \qquad (4.35)
\end{aligned}$$

where (4.33) and (4.34) follow from (4.32), and (4.35) is due to $p_k \geq 1/I$ and (4.24). Subtracting F^* from both sides of (4.35), where $F^* = \min_x F(x) > -\infty$, we obtain

$$\mathbf{E}_{i_k}\left[F(x^{k+1}) - F^* + \frac{L\beta^2}{2}\|d^k\|^2 \;\Big|\; x^k\right] \leq F(x^k) - F^* + \frac{L\beta^2}{2}\|d^{k-1}\|^2 - \frac{c - L\beta^2}{2I}\|\hat{d}^k\|^2. \tag{4.36}$$

By this and the fact that $F(x^k) \geq F^*$ for all k, one can observe that

$$\left\{F(x^k) - F^* + \frac{L\beta^2}{2}\|d^{k-1}\|^2\right\}$$

is a nonnegative supermartingale. It then follows from (4.36), $\beta \in [0, \sqrt{c/L})$, and the Robbins–Siegmund theorem [17, Theorem 1] that $\sum_{k=0}^{\infty}\|\hat{d}^k\|^2 < \infty$ almost surely. Hence, we conclude that $\lim_{k\to\infty}\|\hat{d}^k\| = 0$ almost surely.

(v) In view of statement (iv), it suffices to show that if $\lim_{k\to\infty}\|\hat{d}^k\| = 0$, any accumulation point of $\{x^k\}$ is a D-stationary point of problem (1.1). To this end, we assume that $\lim_{k\to\infty}\|\hat{d}^k\| = 0$. Let x^∞ be an accumulation point of $\{x^k\}$ and $\mathcal{K}$ be a subsequence such that $\lim_{\mathcal{K}\ni k\to\infty} x^k = x^\infty$. Arguing for a contradiction, suppose that x^∞ is not a D-stationary point of (1.1). It then follows from Proposition 1 that there exists some $i \in \mathcal{A}(x^\infty)$ such that

$$0 \notin \nabla f_s(x^\infty) + \partial f_n(x^\infty) - \nabla\psi_i(x^\infty). \tag{4.37}$$

By $i \in \mathcal{A}(x^\infty)$, $\lim_{\mathcal{K}\ni k\to\infty} x^k = x^\infty$, and the continuity of ψ_i, it is not hard to see that $i \in \mathcal{A}_\eta(x^k)$ for all $k \in \mathcal{K}$ sufficiently large. Recall that $x^{k,i}$ denotes the corresponding x^{k,i_k} generated by step (2) of Algorithm 4 if i_k is chosen to be i. It thus follows that

$$x^{k,i} = \operatorname*{argmin}_{x\in\Re^n}\left\{\ell_i(x; z^k, x^k) + \frac{L}{2}\|x - z^k\|^2\right\}. \tag{4.38}$$

By $z^k = x^k + \beta_k(x^k - x^{k-1})$, $\beta_k \in [0, \beta]$, and $\lim_{k\to\infty}\|x^k - x^{k-1}\| = 0$, one has $\lim_{k\to\infty}\|z^k - x^k\| = 0$. This together with $\lim_{\mathcal{K}\ni k\to\infty} x^k = x^\infty$ implies that $\lim_{\mathcal{K}\ni k\to\infty} z^k = x^\infty$. Using this and (4.38), we obtain from Corollary 1 that

$$\lim_{\mathcal{K}\ni k\to\infty} x^{k,i} = x^{\infty,i}, \tag{4.39}$$

where

$$x^{\infty,i} = \operatorname*{argmin}_{x\in\Re^n}\left\{\ell_i(x; x^\infty) + \frac{L}{2}\|x - x^\infty\|^2\right\}. \tag{4.40}$$

By this, $i \in \mathcal{A}(x^\infty)$, and a similar argument as that for deriving (3.45), one has

$$F(x^{\infty,i}) \leq F(x^\infty) - \frac{L}{2}\|x^{\infty,i} - x^\infty\|^2. \tag{4.41}$$

Claim that $i \in \widehat{\mathcal{M}}(x^k)$ for all $k \in \mathcal{K}$ sufficiently large. Indeed, arguing for a contradiction, suppose that this claim does not hold. Recall that $i \in \mathcal{A}_\eta(x^k)$ for all $k \in \mathcal{K}$ sufficiently large. Then there exists a subsequence $\mathcal{K}_1 \subset \mathcal{K}$ such that

$$i \in \mathcal{A}_\eta(x^k) \text{ but } i \notin \widehat{\mathcal{M}}(x^k) \text{ for all } k \in \mathcal{K}_1. \tag{4.42}$$

This together with (4.23) implies that

$$F(x^{k,i}) + \frac{c}{2}\|x^{k,i} - x^k\|^2 > F(x^k) + \frac{L}{2}\|x^k - z^k\|^2 \quad \forall k \in \mathcal{K}_1. \tag{4.43}$$

By (4.38) and similar arguments to those used for deriving (4.7), one has

$$f_s(x^k) + f_n(x^k) - \psi_i(x^k) \geq F(x^{k,i}) + \frac{L}{2}\|x^{k,i} - x^k\|^2 - \frac{L}{2}\|x^k - z^k\|^2. \tag{4.44}$$

By this, (1.8), (4.27), $i \in \mathcal{A}_\eta(x^k)$ for all $k \in \mathcal{K}_1$, $z^k = x^k + \beta_k(x^k - x^{k-1})$, and $\beta_k \in [0, \beta]$, we obtain that for all $k \in \mathcal{K}_1$,

$$F(x^{k,i}) + \frac{L}{2}\|x^{k,i} - x^k\|^2 \leq f_s(x^k) + f_n(x^k) - \psi_i(x^k) + \frac{L}{2}\|x^k - z^k\|^2 \tag{4.45}$$

$$\leq F(x^k) + \frac{L}{2}\|x^k - z^k\|^2 + \eta \;=\; F(x^k) + \frac{L\beta_k^2}{2}\|d^{k-1}\|^2 + \eta \tag{4.46}$$

$$\leq F(x^k) + \frac{L\beta^2}{2}\|d^{k-1}\|^2 + \eta \;\leq\; F(x^0) + \eta, \tag{4.47}$$

where (4.45) is due to (4.44), (4.46) follows from (1.8), $i \in \mathcal{A}_\eta(x^k)$, $z^k = x^k + \beta_k(x^k - x^{k-1})$, and $d^{k-1} = x^k - x^{k-1}$, and (4.47) is due to $\beta_k \in [0, \beta]$ and (4.27). Hence, we have $\{x^{k,i}\}_{k\in\mathcal{K}_1} \subseteq \mathcal{L}(x^0; \eta)$. Recall from above that $\lim_{k\to\infty}\|x^k - z^k\| = 0$, $\{x^k\} \subseteq \mathcal{L}(x^0; \eta)$, $\lim_{\mathcal{K}\ni k\to\infty} x^k = x^\infty$, and $\mathcal{K}_1 \subset \mathcal{K}$. By these, (4.39), the assumption that F is continuous on $\mathcal{L}(x^0; \eta)$, and taking the limit on both sides of (4.43) as $\mathcal{K}_1 \ni k \to \infty$, we obtain that

$$F(x^{\infty,i}) + \frac{c}{2}\|x^{\infty,i} - x^\infty\|^2 \geq F(x^\infty). \tag{4.48}$$

It then follows from (4.41), (4.48), and $c < L$ that $x^{\infty,i} = x^\infty$. Combining this with (4.40) and using the first-order optimality condition of (4.40), one has

$$0 \in \nabla f_s(x^\infty) + \partial f_n(x^\infty) - \nabla\psi_i(x^\infty),$$

which contradicts (4.37). Hence, the above claim holds as desired, that is, $i \in \widehat{\mathcal{M}}(x^k)$ for all $k \in \mathcal{K}$ sufficiently large.

Since $i \in \widehat{\mathcal{M}}(x^k)$ for all $k \in \mathcal{K}$ sufficiently large, it follows from Remark 3(ii) that $\hat{x}^{k+1,i} = x^{k,i}$ for all $k \in \mathcal{K}$ sufficiently large. Moreover, by (4.24) and (4.23), we have that $\|\hat{x}^{k+1,i} - x^k\| \leq \|\hat{d}^k\|$ when $k \in \mathcal{K}$ is sufficiently large. These together with (4.39), $\lim_{k\to\infty}\|\hat{d}^k\| = 0$, and $\lim_{\mathcal{K}\ni k\to\infty} x^k = x^\infty$ imply that $x^{\infty,i} = x^\infty$. Using this and the first-order optimality condition of (4.40), we have

$$0 \in \nabla f_s(x^\infty) + \partial f_n(x^\infty) - \nabla\psi_i(x^\infty),$$

which contradicts (4.37). Therefore, if $\lim_{k\to\infty}\|\hat{d}^k\| = 0$, any accumulation point of $\{x^k\}$ is a D-stationary point of (1.1). This together with statement (iv) leads to the conclusion in (v). ∎

5. Numerical results. In this section we conduct some preliminary numerical experiments to test the performance of our proposed algorithms, namely, Algorithms 1–4. From a theoretical point of view, our algorithms are generally weaker

than a closely related algorithm EPDCA [13, Algorithm 2] but stronger than another related algorithm PDCA$_e$ [20] in terms of solution quality (see section 1 for the discussion). We will compare these algorithms numerically below. All the algorithms are coded in MATLAB and all the computations are performed on a Dell desktop with a 3.40 GHz Intel Core i7-3770 processor and 16 GB of RAM.

In our experiments, the parameters of the aforementioned algorithms are set as follows. For Algorithms 1 and 2, we set $\eta = 0.01$, $\rho = 2$, $c = 10^{-4}$, $\underline{\alpha} = 10^{-8}$, $\overline{\alpha} = 10^8$, and $N = 5$. Also, we choose $\alpha_{0,0} = 1$ and update $\alpha_{k,0}$ via formula (3.3). For Algorithms 3 and 4, we set $\eta = 0.01$ and $c = \tau^2 L$, where $\tau = 0.99$ and L is the Lipschitz constant of ∇f_s. Moreover, we choose $\{\beta_k\}$ via a similar strategy to that in [14, 20]. In particular, we set $\beta_k = \tau(\theta_{k-1} - 1)/\theta_k$, where

$$\theta_{-1} = \theta_0 = 1, \quad \theta_{k+1} = \frac{1 + \sqrt{1 + 4\theta_k^2}}{2},$$

and reset $\theta_{k-1} = \theta_k = 1$ when $k = 200, 400, 600, \ldots$ or $\langle z^k - x^{k+1}, x^{k+1} - x^k \rangle > 0$. It is not hard to verify that $\sup_k \beta_k < \sqrt{c/L}$, and hence such a $\{\beta_k\}$ satisfies the conditions stated in Algorithms 3 and 4. For the algorithms EPDCA [13] and PDCA$_e$ [20], we use almost the same $\{\beta_k\}$ as above except $\tau = 0.5$ and $\tau = 1$, respectively.[4] In addition, we set $\eta = 0.01$ for EPDCA.

We compare the performance of the above algorithms for solving the problem

$$\min_{x \in \Re^n} \left\{ \frac{1}{2}\|Ax - b\|^2 + \lambda \left(\|x\|_1 - \sum_{i=1}^{K} |x_{[i]}| \right) \right\} \tag{5.1}$$

for some $0 \le K < n$, $A \in \Re^{n \times n}$, $b \in \Re^n$, and $\lambda > 0$, where $x_{[i]}$ denotes the ith largest component of x in magnitude. Problem (5.1) has wide applications in sparse learning (e.g., see [8, 2]). It is not hard to observe that (5.1) is a special case of problem (1.1) with $f_s(x) = \|Ax - b\|^2/2$, $f_n(x) = \lambda\|x\|_1$, and $g(x) = \lambda \sum_{i=1}^{K} |x_{[i]}|$. Therefore, the above algorithms can be suitably applied to problem (5.1).

Given positive integers n, K, and a positive number λ, we generate a matrix A, a vector b, and a critical but not D-stationary point $\tilde{x}$ of (5.1) as follows. In particular, we first generate a vector $\tilde{v} \in \Re^K$ with entries randomly chosen from the standard normal distribution, and obtain a vector $v \in \Re^K$ by reordering the entries of $\tilde{v}$ such that $|v_1| \ge |v_2| \ge \cdots \ge |v_K|$. A vector $\tilde{x} \in \Re^n$ is then generated by letting $\tilde{x}_i = v_i + \mathrm{sign}(v_i)$ for $i = 1, \ldots, K$, $\tilde{x}_{K+1} = \tilde{x}_{K+2} = v_K + \mathrm{sign}(v_K)$, and $\tilde{x}_i = 0$ for $i = K + 3, \ldots, n$, where

$$\mathrm{sign}(a) = \begin{cases} 1 & \text{if } a \ge 0, \\ -1 & \text{if } a < 0 \end{cases} \qquad \forall a \in \Re.$$

It then follows that $|\tilde{x}_1| \ge |\tilde{x}_2| \ge \cdots \ge |\tilde{x}_K| = |\tilde{x}_{K+1}| = |\tilde{x}_{K+2}| > |\tilde{x}_{K+3}| = \cdots = |\tilde{x}_n| = 0$ and hence $g(x)$ is not differentiable at $\tilde{x}$. We next generate a vector $\tilde{d} \in \Re^n$ with entries randomly chosen from the uniform distribution on $[-\sqrt{\lambda}, \sqrt{\lambda}]$, and obtain a vector $d \in \Re^n$ by reordering the entries of $\tilde{d}$ such that $d_1 \ge d_2 \ge \cdots \ge d_n$. We then obtain the matrix A by letting $A = \mathrm{Diag}(d) + 0.01\tilde{A}$, where $\mathrm{Diag}(d)$ is the diagonal matrix with the elements of d on the diagonal and the entries of $\tilde{A}$ are randomly

[4] As observed in our experiments, EPDCA appears to perform best with $\tau = 0.5$ among the five choices 0, 0.25, 0.5, 0.75, and 1.

chosen from the uniform distribution on $[-1/n, 1/n]$. Finally, we compute the vector b by solving the linear equation

$$A^T b = A^T A\tilde{x} + w^1 - w^2$$

for some $w^1 \in \partial f_n(\tilde{x})$ and $w^2 \in \partial g(\tilde{x})$. For such A and b, it is not hard to verify that $\tilde{x}$ is a critical but not D-stationary point of problem (5.1).

In our experiments, we choose $(n, K, \lambda) = (500j, 150j, 5j)$ for $j = 1, 2, \ldots, 10$. For each triple (n, K, λ), we first generate an instance of problem (5.1) and a critical but not D-stationary point $\tilde{x}$ of it as described above. We then perform 20 runs of all the above algorithms. In each run, we choose randomly the same initial point $x^0 = \tilde{x} + 0.01\xi$ for all the algorithms, where the entries of $\xi \in \Re^n$ are chosen randomly from the uniform distribution on $[-1, 1]$, and terminate all the algorithms once

$$|f(x^k) - f(x^{k-1})| \le 10^{-8}.$$

The computational results averaged over each group of 20 runs with the same (n, K, λ) are presented in Table 1, which consists of two subtables. In detail, the parameters n, K, and λ are listed in the first three columns, respectively. For each triple (n, K, λ), the objective value of problem (5.1) and the CPU time for these algorithms averaged over 20 runs are given in Tables 1(a) and 1(b), respectively. One can observe that the objective values found by the proposed methods (Algorithms 1–4) are comparable to those by EPDCA but much lower than those by PDCA$_e$, which is not surprising as the proposed methods and EPDCA generally converge to a stronger stationary point than PDCA$_e$. In terms of CPU time, the proposed methods and PDCA$_e$ substantially outperform EPDCA. Note that EPDCA generally converges to a stronger stationary point than Algorithms 1–4 and PDCA$_e$. It is therefore reasonable that EPDCA has slower convergence than the other methods. In addition, the line-search type of algorithms (Algorithms 1 and 2) slightly outperform the extrapolation-type algorithms (Algorithms 3 and 4) and PDCA$_e$. Finally, we observe that the randomized algorithms (Algorithms 2 and 4) slightly outperform their respective deterministic counterparts (Algorithms 1 and 3).

6. Concluding remarks. In this paper we considered a class of structured nonsmooth DC minimization described in (1.1) and (1.2). The existing methods [15, 20, 13] for this problem usually have weak convergence guarantees or exhibit slow convergence. Due to this, we proposed two nonmonotone enhanced proximal DC algorithms for solving this problem. For possible acceleration, one of our algorithms uses a nonmonotone line-search scheme in which the involved Lipschitz constant is adaptively approximated by some local curvature information of the associated smooth function, and the other one employs an extrapolation scheme. We proved that every accumulation point of the solution sequence generated by them is a *D-stationary* point of the problem. These methods may, however, become inefficient when the number of convex smooth functions involved in the second convex component of the objective function is large. To remedy this issue, we proposed randomized counterparts for them and showed that every accumulation point of the generated solution sequence is a *D-stationary* point of the problem *almost surely*.

We also conducted preliminary numerical experiments to compare the performance of the proposed methods with two closely related algorithms, EPDCA and PDCA$_e$. The computational results demonstrated that the proposed methods are comparable to EPDCA but substantially outperform PDCA$_e$ in terms of solution

TABLE 1
Computational results for solving problem (5.1).

Parameter			Objective value					
n	K	λ	EPDCA	Algo. 1	Algo. 2	Algo. 3	Algo. 4	PDCA$_e$
500	150	5	0.80	0.80	0.81	0.80	0.81	18.44
1000	300	10	1.23	1.23	1.23	1.23	1.23	24.52
1500	450	15	2.09	2.09	2.11	2.09	2.10	19.57
2000	600	20	2.85	2.85	2.87	2.85	2.86	11.98
2500	750	25	3.66	3.63	3.68	3.66	3.67	37.47
3000	900	30	4.73	4.71	4.75	4.73	4.73	29.76
3500	1050	35	5.45	5.44	5.48	5.45	5.49	34.95
4000	1200	40	6.33	6.29	6.36	6.33	6.34	22.84
4500	1350	45	7.21	7.16	7.21	7.21	7.23	16.43
5000	1500	50	7.95	7.90	7.97	7.95	7.94	18.28

(a) *Results for objective value.*

Parameter			CPU time (in seconds)					
n	K	λ	EPDCA	Algo. 1	Algo. 2	Algo. 3	Algo. 4	PDCA$_e$
500	150	5	1.47	0.03	0.03	0.08	0.07	0.05
1000	300	10	5.20	0.12	0.11	0.26	0.22	0.21
1500	450	15	21.28	0.50	0.44	1.09	0.88	1.06
2000	600	20	44.19	1.08	1.02	2.84	2.27	2.24
2500	750	25	65.34	1.66	1.38	3.72	3.00	3.19
3000	900	30	81.97	1.86	1.85	5.49	4.46	4.56
3500	1050	35	111.37	2.74	2.52	7.66	6.26	5.99
4000	1200	40	141.57	3.42	3.08	9.59	7.77	7.52
4500	1350	45	180.69	4.39	3.98	11.01	8.95	10.59
5000	1500	50	225.51	5.56	5.26	15.19	12.44	12.11

(b) *Results for CPU time.*

quality, and moreover, they are comparable to PDCA$_e$ but much faster than EPDCA in terms of speed. Therefore, the practical performance of these methods is worthy of further numerical study.

REFERENCES

[1] A. ALVARADO, G. SCUTARI, AND J.-S. PANG, *A new decomposition method for multiuser DC-programming and its applications*, IEEE Trans. Signal Process., 62 (2014), pp. 2984–2998.

[2] M. AHN, J.-S. PANG, AND J. XIN, *Difference-of-convex learning: Directional stationarity, optimality, and sparsity*, SIAM J. Optim., 27 (2017), pp. 1637–1665.

[3] J. BARZILAI AND J. M. BORWEIN, *Two-point step size gradient methods*, IMA J. Numer. Anal., 8 (1988), pp. 141–148.

[4] E. CANDÈS AND J. ROMBERG, *ℓ_1-magic: Recovery of sparse signals via convex programming*, User guide, Applied & Computational Mathematics, California Institute of Technology, Pasadena, CA, 2005; available at https://statweb.stanford.edu/~candes/l1magic/#code.

[5] K. L. CHUNG, *A Course in Probability Theory*, Academic Press, New York, 2001.

[6] J. FAN AND R. LI, *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Amer. Statist. Assoc., 96 (2001), pp. 1348–1360.

[7] P. GONG, C. ZHANG, Z. LU, J. HUANG, AND J. YE, *A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems*, in Proceedings of the 30th International Conference on Machine Learning, Proc. Mach. Learn. Res. 28, 2013, pp. 37–45; available at http://proceedings.mlr.press/v28/gong13a.html.

[8] J.-Y. GOTOH, A. TAKEDA, AND K. TONO, *DC formulations and algorithms for sparse optimization problems*, Math. Program., 169 (2018), pp. 141–176.

[9] R. HORST AND N. V. THOAI, *DC programming: Overview*, J. Optim. Theory Appl., 103 (1999), pp. 1–43.

[10] K. KOH, S.-J. KIM, AND S. BOYD, *An interior-point method for large-scale ℓ_1-regularized*

logistic regression, J. Mach. Learn. Res., 8 (2017), pp. 1519–1555.
[11] H. A. Le Thi and T. Pham Dinh, *DC programming and DCA: Thirty years of developments*, Math. Program., 169 (2018), pp. 5–68.
[12] Z. Lu and L. Xiao, *A randomized nonmonotone block proximal gradient method for a class of structured nonlinear programming*, SIAM J. Numer. Anal., 55 (2017), pp. 2930–2955.
[13] Z. Lu, Z. Zhou, and Z. Sun, *Enhanced proximal DC algorithms with extrapolation for a class of structured nonsmooth DC minimization*, Math. Program., 176 (2019), pp. 369–401.
[14] B. O'Donoghue and E. Candès, *Adaptive restart for accelerated gradient schemes*, Found. Comput. Math., 15 (2015), pp. 715–732.
[15] J.-S. Pang, M. Razaviyayn, and A. Alvarado, *Computing B-stationary points of nonsmooth DC programs*, Math. Oper. Res., 42 (2016), pp. 95–118.
[16] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[17] H. Robbins and D. Siegmund, *A convergence theorem for non negative almost supermartingales and some applications*, in Optimizing Methods in Statistics, Academic Press, New York, 1971, pp. 233–257.
[18] M. Sanjabi, M. Razaviyayn, and Z.-Q. Luo, *Optimal joint base station assignment and beamforming for heterogeneous networks*, IEEE Trans. Signal Process., 62 (2014), pp. 1950–1961.
[19] K. Tono, A. Takeda, and J. Gotoh, *Efficient DC Algorithm for Constrained Sparse Optimization*, preprint, https://arxiv.org/abs/1701.08498, 2017.
[20] B. Wen, X. Chen, and T. K. Pong, *A proximal difference-of-convex algorithm with extrapolation*, Comput. Optim. Appl., 69 (2018), pp. 297–324.
[21] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, *Sparse reconstruction by separable approximation*, IEEE Trans. Signal Process., 57 (2009), pp. 2479–2493.
[22] C.-H. Zhang, *Nearly unbiased variable selection under minimax concave penalty*, Ann. Statist., 38 (2010), pp. 894–942.