New Bregman proximal type algorithms for solving DC optimization problems

Shota Takahashi[1] · Mituhiro Fukuda[2,3,4,5] · Mirai Tanaka[1,6,7]

Abstract

Difference of Convex (DC) optimization problems have objective functions that are differences between two convex functions. Representative ways of solving these problems are the proximal DC algorithms, which require that the convex part of the objective function have *L*-smoothness. In this article, we propose the Bregman Proximal DC Algorithm (BPDCA) for solving large-scale DC optimization problems that do not possess *L*-smoothness. Instead, it requires that the convex part of the objective function has the *L*-smooth adaptable property that is exploited in Bregman proximal gradient algorithms. In addition, we propose an accelerated version, the Bregman Proximal DC Algorithm with extrapolation (BPDCAe), with a new restart scheme. We show the global convergence of the iterates generated by BPDCA(e) to a limiting critical point under the assumption of the Kurdyka-Łojasiewicz property or subanalyticity of the objective function and other weaker conditions than those of the existing methods. We applied our algorithms to phase retrieval, which can be described both as a nonconvex optimization problem and as a DC optimization problem. Numerical experiments showed that BPDCAe outperformed existing Bregman proximal-type algorithms because the DC formulation allows for larger admissible step sizes.

1 Introduction

We are interested in solving Difference of Convex (DC) optimization problems:

✉ Shota Takahashi
    takahashi.shota@ism.ac.jp

Extended author information available on the last page of the article

$$(\mathcal{P}) \quad \min\left\{\Psi(x) := f_1(x) - f_2(x) + g(x) \mid x \in \overline{C}\right\},$$

where f_1 and f_2 are convex functions on $\mathbb{R}^d$, and $\overline{C}$ is the closure of C which is a nonempty, convex, and open set. Note that the function $f_1 - f_2$ is not always convex. Also, g may be nonsmooth, such as the ℓ_1-norm $\|x\|_1$ in [8, 20, 29], or alternatively, f_2 may be nonsmooth [16]. Some interesting examples of $(\mathcal{P})$ can be found in [28]. Although we will place some assumptions on C, it can be regarded as $\mathbb{R}^d$ for simplicity. Applications of DC optimization are summarized in [13, 17, 26].

The DC Algorithm (DCA) (see for instance [17]) is a well-known iterative method for solving the DC optimization problems $(\mathcal{P})$. At each iteration, its computational burden mainly depends on the resolution of the subproblem,

$$x^{k+1} = \operatorname*{argmin}_{x \in \overline{C}} \left\{f_1(x) - \langle \xi^k, x - x^k \rangle + g(x)\right\}, \tag{1}$$

where $\xi^k \in \partial_c f_2(x^k) := \{\xi \in \mathbb{R}^d \mid f_2(y) - f(x^k) - \langle \xi, y - x^k \rangle \geq 0, \forall y \in \mathbb{R}^d\}$ is a (classical) subgradient of f_2 at $x^k \in \overline{C}$. Solving subproblem (1) may be computationally demanding unless f_1 and g have simple structure or $(\mathcal{P})$ is small-scale. When g is convex, the proximal DC Algorithm (pDCA) (see for instance [28]) is an alternative method of solving large-scale DC optimization problems. However, to guarantee global convergence of its iterates to a critical point, f_1 needs to be L-smooth; *i.e.*, its gradient needs to be globally Lipschitz continuous. Each step of pDCA is given by

$$x^{k+1} = \operatorname*{argmin}_{x \in \overline{C}} \left\{g(x) + \langle \nabla f_1(x^k) - \xi^k, x - x^k \rangle + \frac{1}{2\lambda}\|x - x^k\|^2\right\}, \tag{2}$$

where $\xi^k \in \partial_c f_2(x^k)$, $x^k \in \overline{C}$, $\lambda > 0$ satisfies $0 < \lambda L < 1$, and $\|\cdot\|$ denotes the Euclidean norm. Since $\lambda \ (< 1/L)$ plays the role of a step size, finding a larger upper bound $1/L$, *i.e.*, finding a smaller L, is of fundamental importance to achieving fast convergence. Wen *et al.*[28] proposed the proximal DC Algorithm with extrapolation (pDCAe) to accelerate pDCA with the extrapolation technique, which is used, for instance, in the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) of Beck and Teboulle [4] and in the Nesterov's extrapolation technique [21, 22].

Bolte *et al.*[8], who incorporated the kernel generating distance (function) h and Bregman distances [9] into the algorithm framework, came up with the notion of the L-smooth adaptable property (see also [2]). This property is less restrictive than L-smoothness. A variant of the Bregman Proximal Gradient algorithm (BPG) proposed by Mukkamala *et al.*[20] iteratively estimates a small L, while Zhang *et al.*[29] proposed the Bregman Proximal Gradient algorithm with extrapolation (BPGe), which combines BPG with a line search step for extrapolating parameters.

In this paper, we propose two new algorithms, namely, the Bregman Proximal Difference of Convex Algorithm (BPDCA) and the Bregman Proximal Difference of Convex Algorithm with extrapolation (BPDCAe), which are inspired by pDCA(e) and BPG(e). These novel algorithms combine pDCA(e) with the Bregman distances. In the subproblem of BPDCA(e), the use of Bregman distances guarantees the accuracy of a linear approximation of $f_1 - f_2$.

The novelty of our contributions can be better understood by comparing them with the existing work. As already mentioned, Bregman distances allow us to extend the class of functions to be minimized f_1 from L-smooth in pDCAe [28] to the larger class of L-smooth adaptable pairs of functions (f_1, h). In addition, the function g does not need to be convex in the case of BPDCA. By assuming that either f_2 or g are differentiable and that their gradients are locally Lipschitz continuous, the iterates of BPDCA(e) converge globally to a limiting stationary point (Theorems 2 and 7) or a limiting critical point (Theorems 3 and 8), where the definitions of these convergent points are given in Definition 6. This means that either g or f_2 can be nonsmooth.

Compared with BPG-type algorithms [8, 20, 29], BPDCA(e) well exploits the structure of the objective function. When applying these BPG-type algorithms to solve problem $(\mathcal{P})$, we decompose Ψ into two functions. There are two naive ways to decompose Ψ. First, we consider the decomposition $\Psi = f_1 + (g - f_2)$ to apply BPG. In this case, BPG solves its subproblem $\min\{g(x) - f_2(x) + \langle \nabla f_1(x^k), x - x^k \rangle + \frac{1}{\lambda} D_h(x, x^k)\}$ at the kth iteration, where D_h is the Bregman distance associated with a kernel generating distance h (see Definition 2) and $\lambda < 1/L$ is a positive parameter. In general, it is difficult to efficiently solve it because f_2 often does not have simple structure such as separability. This fact is also true even if g is convex and separable, simultaneously. With BPDCA, we only need to solve its subproblem $\min\{g(x) + \langle \nabla f_1(x^k) - \xi^k, x - x^k \rangle + \frac{1}{\lambda} D_{h'}(x, x^k)\}$, where $\xi^k \in \partial_c f_2(x^k)$. If g is additionally convex, the subproblem becomes convex and hence is often efficiently solved. Moreover, if g and h are also separable, the subproblem is reduced to d independent one-dimensional convex optimization problems. Even without separability of h, it often has closed-form solution formulae as we mentioned in Sect. 5. As an alternative way of decomposition of Ψ, we consider $\Psi = (f_1 - f_2) + g$ to apply BPG. In this case, to guarantee global convergence, the L-smooth adaptability of $(f_1 - f_2, h)$ is required (see Definition 3). Meanwhile, for the global convergence of BPDCA(e), the L'-smooth adaptability of (f_1, h') is required. Comparing these constants, $L' \leq L$ in general, and then, we can expect substantial decrease in the objective function at each iteration (Lemma 5 and [8, Lemma 4.1]). This fact has dramatic consequences in practice, as we found in numerical experiments on phrase retrieval (Sect. 5.1).

The convergence of our algorithms and the monotonicity of the objective function are based on standard assumptions. Our new restart scheme (Sect. 3.2) plays an important role in guaranteeing the non-increasing property of the objective functions of BPDCAe without the need for a line search, as in [29]. We show global convergence under local Lipschitz continuity of the gradients and the Kurdyka-Łojasiewicz property or subanalyticity of the objective function. Additionally, we evaluated the rates of convergence of BPDCA(e).

To evaluate the performance of BPDCA(e), we applied them to phase retrieval, a well-known application in nonconvex optimization. Phase retrieval arises in many fields of science and engineering, such as X-ray crystallography and image processing [10, 24]. It can be formulated as a nonconvex optimization problem or DC optimization problem $(\mathcal{P})$, such as in [14]. It cannot be solved via pDCA or proximal algorithms, since the function we want to minimize is not L-smooth. When we formulated phase retrieval as a DC optimization problem, we obtained much smaller L-smooth adaptable parameters than the existing ones [8, Lemma 5.1], [20]. Thus, our algorithms outperformed BPG(e) in

our numerical experiment. Further experiments showed that, under the Gaussian model, BPDCAe had a higher success rate of phrase retrieval than that of Wirtinger flow [10]. Although the kernel generating distance h we utilized does not satisfy Assumption 4 (i), the sequences generated by BPDCA(e) converged in the numerical experiments. Therefore, we conjecture that all convergence analyses can be carried out with weaker conditions.

This paper is organized as follows. Section 2 summarizes notions such as the limiting subdifferential, the Bregman distances, the L-smooth adaptable property, the Kurdyka-Łojasiewicz property, and the subanalytic functions. Section 3 introduces our Bregman proximal-type algorithms under the assumption that (f_1, h) has an L-smooth adaptable property. Section 4 (and Appendix A) establishes the global convergence of BPDCA(e) to a limiting stationary point or a limiting critical point of the problem $(\mathcal{P})$ and analyzes its rate of convergence. Section 5 derives small values for the constant L and compares the performance of our algorithms with that of BPG(e). Section 6 summarizes our contributions and discusses future work.

2 Preliminaries

Here, we review the important notions we will need in the subsequent sections.

2.1 Subdifferentials

Definition 1 (Limiting Subdifferential [23]) For a proper and lower semicontinuous function $f : \mathbb{R}^d \to (-\infty, +\infty]$, the limiting subdifferential [23] of f at $x \in \text{dom} f$ is defined by

$$\partial f(x) = \left\{ \xi \in \mathbb{R}^d \mid \exists x^k \xrightarrow{f} x, \xi^k \to \xi \text{ such that } \liminf_{y \to x^k, y \neq x^k} \frac{f(y) - f(x^k) - \langle \xi^k, y - x^k \rangle}{\|y - x^k\|} \geq 0 \right\},$$

where $x^k \xrightarrow{f} x$ means $x^k \to x$ and $f(x^k) \to f(x)$.

Note that when f is convex, the limiting subdifferential coincides with the (classical) subdifferential [23, Proposition 8.12], that is, $\partial f(x) = \partial_c f(x)$ for all $x \in \mathbb{R}^d$.

2.2 Bregman distances

First, we define kernel generating distances and Bregman distances.

Definition 2 (Kernel Generating Distances [8] and Bregman Distances [9]) Let C be a nonempty, convex, and open subset of $\mathbb{R}^d$. Associated with C, a function $h : \mathbb{R}^d \to (-\infty, +\infty]$ is called a kernel generating distance if it meets the following conditions:

(i) h is proper, lower semicontinuous, and convex, with $\operatorname{dom} h \subset \overline{C}$ and $\operatorname{dom} \partial h = C$.

(ii) h is $\mathcal{C}^1$ on $\operatorname{int} \operatorname{dom} h = C$.

We denote the class of kernel generating distances by $\mathcal{G}(C)$. Given $h \in \mathcal{G}(C)$, the Bregman distance $D_h : \operatorname{dom} h \times \operatorname{int} \operatorname{dom} h \to \mathbb{R}_+$ is defined by

$$D_h(x, y) := h(x) - h(y) - \langle \nabla h(y), x - y \rangle.$$

From the gradient inequality, h is convex if and only if $D_h(x, y) \geq 0$ for any $x \in \operatorname{dom} h$ and $y \in \operatorname{int} \operatorname{dom} h$. When h is a strictly convex function, the equality holds if and only if $x = y$. When $h = \frac{1}{2} \| \cdot \|^2$, $D_h(x, y) = \frac{1}{2} \|x - y\|^2$, which is the squared Euclidean distance.

In addition, the Bregman distances satisfy the three-point identity [8],

$$D_h(x, z) - D_h(x, y) - D_h(y, z) = \langle \nabla h(y) - \nabla h(z), x - y \rangle, \tag{3}$$

for any $y, z \in \operatorname{int} \operatorname{dom} h$, and $x \in \operatorname{dom} h$.

2.3 Smooth adaptable functions

Now let us define the notion of L-smooth adaptable.

Definition 3 (L-smooth adaptable [8]) Consider a pair of functions (f, h) satisfying the following conditions:

(i) $h \in \mathcal{G}(C)$,

(ii) $f : \mathbb{R}^d \to (-\infty, +\infty]$ is a proper and lower semicontinuous function with $\operatorname{dom} h \subset \operatorname{dom} f$, which is $\mathcal{C}^1$ on $C = \operatorname{int} \operatorname{dom} h$.

The pair (f, h) is called L-smooth adaptable (L-smad) on C if there exists $L > 0$ such that $Lh - f$ and $Lh + f$ are convex on C.

Since our focus is on DC optimization, the function f_1 in $(\mathcal{P})$ is always convex. Thus, it will be enough to verify that $Lh - f_1$ is convex on C to have (f_1, h) L-smad on C.

From the L-smooth adaptable property comes the Descent Lemma [8].

Lemma 1 (Full Extended Descent Lemma [8]) *A pair of functions (f, h) is L-smad on $C = \operatorname{int} \operatorname{dom} h$ if and only if:*

$$|f(x) - f(y) - \langle \nabla f(y), x - y \rangle| \leq L D_h(x, y), \quad \forall x, y \in \operatorname{int} \operatorname{dom} h.$$

2.4 Kurdyka-Łojasiewicz property and subanalytic functions

Given $\eta > 0$, let Ξ_η denote the set of all continuous concave functions $\phi : [0, \eta) \to \mathbb{R}_+$ that are $\mathcal{C}^1$ on $(0, \eta)$ with positive derivatives and which satisfy $\phi(0) = 0$. Here, we introduce the Kurdyka-Łojasiewicz property [7, 15], which we need when analyzing our algorithms:

Definition 4 (Kurdyka-Łojasiewicz property) Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function.

(i) f is said to have the Kurdyka-Łojasiewicz (KL) property at $\hat{x} \in \text{dom}\, \partial f$ if there exist $\eta \in (0, +\infty]$, a neighborhood U of $\hat{x}$, and a function $\phi \in \Xi_\eta$ such that the following inequality holds:

$$\phi'(f(x) - f(\hat{x})) \cdot \text{dist}\,(0, \partial f(x)) \geq 1, \tag{4}$$
$$\forall x \in U \cap \{x \in \mathbb{R}^d \mid f(\hat{x}) < f(x) < f(\hat{x}) + \eta\}.$$

(ii) If f has the KL property at each point of dom ∂f, then it is called a KL function.

Lemma 2 (*Uniformized KL property* [7]) *Suppose that* $f : \mathbb{R}^d \to (-\infty, +\infty]$ *is a proper and lower semicontinuous function and let* Γ *be a compact set. If* f *is constant on* Γ *and has the KL property at each point of* Γ, *then there exist positive scalars* $\epsilon, \eta > 0$, *and* $\phi \in \Xi_\eta$ *such that*

$$\phi'(f(x) - f(\hat{x})) \cdot \text{dist}\,(0, \partial f(x)) \geq 1,$$

for any $\hat{x} \in \Gamma$ *and any* x *satisfying* $\text{dist}\,(x, \Gamma) < \epsilon$ *and* $f(\hat{x}) < f(x) < f(\hat{x}) + \eta$.

Next, we describe subanalytic functions.

Definition 5 (Subanalyticity [6])

(i) A subset A of $\mathbb{R}^d$ is called semianalytic if each point of $\mathbb{R}^d$ admits a neighborhood V for which $A \cap V$ assumes the following form:

$$\bigcup_{i=1}^{p} \bigcap_{j=1}^{q} \{x \in V \mid f_{ij}(x) = 0, g_{ij}(x) > 0\},$$

where the functions $f_{ij}, g_{ij} : V \to \mathbb{R}$ are real-analytic for all $1 \leq i \leq p, 1 \leq j \leq q$.

(ii) The set A is called subanalytic if each point of $\mathbb{R}^d$ admits a neighborhood V such that

$$A \cap V = \{x \in \mathbb{R}^d \mid (x, y) \in B\},$$

where B is a bounded semianalytic subset of $\mathbb{R}^d \times \mathbb{R}^m$ for some $m \geq 1$.

(iii) A function $f : \mathbb{R}^d \to (-\infty, +\infty]$ is called subanalytic if its graph is a subanalytic subset of $\mathbb{R}^d \times \mathbb{R}$.

For instance, given a subanalytic set S, dist (x, S) is subanalytic, and every analytic function is subanalytic. Note that subanalytic functions are KL functions. See [5, 6] for further properties of subanalyticity.

3 Proposed methods: Bregman Proximal DC Algorithms

We place the following assumptions on the DC optimization problem $(\mathcal{P})$. Recall that $C = \text{int dom } h$.

Assumption 1

  (i) $h \in \mathcal{G}(C)$ with $\overline{C} = \overline{\text{dom } h}$.
  (ii) $f_1 : \mathbb{R}^d \to (-\infty, +\infty]$ is proper and convex with dom $h \subset \text{dom } (f_1 + g)$, which is $\mathcal{C}^1$ on C.
  (iii) $f_2 : \mathbb{R}^d \to (-\infty, +\infty]$ is proper and convex.
  (iv) $g : \mathbb{R}^d \to (-\infty, +\infty]$ is proper and lower semicontinuous with dom $g \cap C \neq \emptyset$.
  (v) $v(\mathcal{P}) := \inf\{\Psi(x) \mid x \in \overline{C}\} > -\infty$.
  (vi) For any $\lambda > 0$, $\lambda g + h$ is supercoercieve, that is,

$$\lim_{\|u\| \to \infty} \frac{\lambda g(u) + h(u)}{\|u\|} = \infty.$$

Let $x \in \text{dom } (f_1 + g)$, then $f_2(x) \leq g(x) + f_1(x) - v(\mathcal{P}) < +\infty$ due to Assumption 1 (v). Thus, $x \in \text{dom} f_2$, i.e., dom $(f_1 + g) \subset \text{dom} f_2$. From Assumption 1 (ii), we have $C \subset \text{dom} (f_1 + g) \subset \text{dom} f_2$. Note that Assumption 1 (iv) holds when $\overline{C}$ is compact [8].

3.1 Bregman Proximal DC Algorithm (BPDCA)

To obtain the Bregman Proximal DC Algorithm (BPDCA) mapping for some $\lambda > 0$, we recast the objective function of $(\mathcal{P})$ via a DC decomposition:

$$\Psi(u) = f_1(u) - f_2(u) + g(u) = \left(\frac{1}{\lambda}h(u) + g(u)\right) - \left(\frac{1}{\lambda}h(u) - f_1(u) + f_2(u)\right),$$

and, given $x \in C = \text{int dom } h$ and $\xi \in \partial_c f_2(x)$, define the mapping,

$$\mathcal{T}_\lambda(x) := \operatorname*{argmin}_{u \in \overline{C}} \left\{ g(u) + \langle \nabla f_1(x) - \xi, u - x \rangle + \frac{1}{\lambda} D_h(u, x) \right\}.$$

Additionally, we put the following assumption on $(\mathcal{P})$.

Assumption 2 For all $x \in C$ and $\lambda > 0$, we have

$$\mathcal{T}_\lambda(x) \subset C, \quad \forall x \in C.$$

Note that Assumption 2 is not so restrictive because it holds when $C \equiv \mathbb{R}^d$. Under Assumptions 1 and 2, we have the following lemma [8, Lemma 3.1].

Lemma 3 *Suppose that Assumptions 1 and 2 hold, and let $x \in C = $ int dom h. Then, the set $\mathcal{T}_\lambda(x)$ is a nonempty and compact subset of C for any $\lambda > 0$.*

Note that when h is strictly convex, $\mathcal{T}_\lambda(x)$ is a singleton. Also, when g and h are separable, this mapping is easily computable since $\mathcal{T}_\lambda(x)$ can be decomposed into a single-valued optimization problem, and often has a closed-form solution. For instance, when $h(x) = \frac{1}{2}\|x\|^2$, for $g(x) = \|x\|_1$, $\mathcal{T}_\lambda(x)$ becomes the soft-thresholding operator, or for $g(x) = \|x\|_0$, the hard-thresholding operator. Other well-known examples where this mapping has a closed-form solution are when we use an appropriate h such as Burg entropy [2], Shannon entropy [3], or $h(x) = \frac{1}{4}\|x\|^4 + \frac{1}{2}\|x\|^2$ [8] for the corresponding g. Note that this $h(x)$ is not separable. For further examples, see [12, Table 2.1].

The Bregman Proximal DC Algorithm (BPDCA), which we are proposing, is listed as Algorithm 1.

Algorithm 1 Bregman Proximal DC Algorithm (BPDCA)

Input: $h \in \mathcal{G}(C)$ with $C = $ int dom h such that L-smad for the pair (f_1, h) holds on C.
Initialization: $x^0 \in C$ and $0 < \lambda < 1/L$.
for $k = 0, 1, 2, \ldots,$ **do**
    Take any $\xi^k \in \partial_c f_2(x^k)$ and compute

$$x^{k+1} = \operatorname*{argmin}_{x \in \overline{C}} \left\{ g(x) + \langle \nabla f_1(x^k) - \xi^k, x - x^k \rangle + \frac{1}{\lambda} D_h(x, x^k) \right\}. \tag{5}$$

end for

As a recurrent example, $D_h(x, x^k) = \frac{1}{2}\|x - x^k\|^2$ when $h(x) = \frac{1}{2}\|x\|^2$. In this case, if L is regarded as the Lipschitz constant for the gradient of f_1, subproblem (5) reduces to subproblem (2). If f_2 is $\mathcal{C}^1$ on C and the pair $(f_1 - f_2, h)$ is L-smad, BPDCA reduces to BPG [8].

3.2 Bregman Proximal DC Algorithm with extrapolation (BPDCAe)

Algorithm 2, which we are proposing, is an acceleration of BPDCA that uses the extrapolation technique [4, 21, 22] to solve the DC optimization problem $(\mathcal{P})$.

Algorithm 2 Bregman Proximal DC Algorithm with extrapolation (BPDCAe)

Input: $h \in \mathcal{G}(C)$ with $C = \text{int dom } h$ such that L-smad for the pair (f_1, h) holds on $\mathbb{R}^d$.
Initialization: $x^0 = x^{-1} \in \mathbb{R}^d, \theta_{-1} = \theta_0 = 1, \rho \in (0, 1]$, and $0 < \lambda < 1/L$.
for $k = 0, 1, 2, \ldots,$ **do**
    Take any $\xi^k \in \partial_c f_2(x^k)$ and compute

$$\beta_k = \frac{\theta_{k-1} - 1}{\theta_k} \quad \text{with} \quad \theta_{k+1} = \frac{1 + \sqrt{1 + 4\theta_k^2}}{2}, \tag{6}$$
$$y^k = x^k + \beta_k(x^k - x^{k-1}).$$

  if $y^k \notin C$ or $D_h(x^k, y^k) > \rho D_h(x^{k-1}, x^k)$ **then**
    Set $\beta_k = 0$ with $\theta_{k-1} = \theta_k = 1$.
  end if
  Compute $y^k = x^k + \beta_k(x^k - x^{k-1})$ and

$$x^{k+1} = \underset{y \in \overline{C}}{\text{argmin}} \left\{ g(y) + \langle \nabla f_1(y^k) - \xi^k, y - y^k \rangle + \frac{1}{\lambda} D_h(y, y^k) \right\}. \tag{7}$$

end for

When $\beta_k \equiv 0$ for all $k \geq 0$, BPDCAe reduces to BPDCA. Here, we prefer the popular choice for the coefficients β_k (and θ_k) given in [28] for acceleration. Accordingly, (6) guarantees that $\{\beta_k\}_{k=0}^{\infty} \subset [0, 1)$ and $\sup_{k \geq 0} \beta_k < 1$. These properties are needed to prove global subsequential convergence of the iterates (see Theorem 6 (ii)). Algorithm 2 introduces a new *adaptive restart scheme*, which resets θ_k and β_k whenever

$$D_h(x^k, y^k) > \rho D_h(x^{k-1}, x^k), \tag{8}$$

is satisfied for a fixed $\rho \in [0, 1)$. This adaptive restart scheme guarantees the non-increasing property for BPDCAe (see Lemma 6). In addition, we can enforce this resetting every K iterations for a given positive integer K. In our numerical experiments, we set $\{\beta_k\}_{k=0}^{\infty}$ as both the fixed and the adaptive restart schemes.

When $C = \text{int dom } h = \mathbb{R}^d$, y^k always stays in C. However, when $C \neq \mathbb{R}^d$ and $x^k + \beta_k(x^k - x^{k-1}) \notin C$, Algorithm 2 enforces $\beta_k = 0$ and BPDCAe is not accelerated at the kth iteration. This operation guarantees that y^k always stays in C. In practice, however, the extrapolation technique may be valid and accelerates BPDCAe.

We define the following BPDCAe mapping for all $x, y \in C = \text{int dom } h$, and $\lambda \in (0, 1/L)$:

$$\mathcal{S}_\lambda(x, y) := \underset{u \in \overline{C}}{\text{argmin}} \left\{ g(u) + \langle \nabla f_1(y) - \xi, u - y \rangle + \frac{1}{\lambda} D_h(u, y) \right\},$$

where $\xi \in \partial_c f_2(x)$. Similarly to the case of BPDCA, we make an Assumption 3 and can prove Lemma 4 for $\mathcal{S}_\lambda(x, y) \subset \overline{C}$.

Assumption 3 For all $x, y \in C$ and $\lambda > 0$, we have

$$\mathcal{S}_\lambda(x, y) \subset C, \quad \forall x, y \in C.$$

Lemma 4 *Suppose that Assumptions* 1 *and* 3 *hold, and let* $x, y \in C =$ int dom h. *Then, the set* $\mathcal{S}_\lambda(x, y)$ *is a nonempty and compact subset of C for any $\lambda > 0$.*

4 Convergence analysis

Throughout this section, we will assume that the pair of functions (f_1, h) is L-smad on C.

4.1 Properties of BPDCA

First, we show the decreasing property of BPDCA mapping for $0 < \lambda L < 1$ (the argument is adapted from [8, Lemma 4.1]).

Lemma 5 *Suppose that Assumptions* 1 *and* 2 *hold. For any* $x \in C =$ int dom h *and any* $x^+ \in C =$ int dom h *defined by*

$$x^+ \in \operatorname*{argmin}_{u \in \overline{C}} \left\{ g(u) + \langle \nabla f_1(x) - \xi, u - x \rangle + \frac{1}{\lambda} D_h(u, x) \right\}, \tag{9}$$

where $\xi \in \partial_c f_2(x)$ *and* $\lambda > 0$, *it holds that*

$$\lambda \Psi(x^+) \leq \lambda \Psi(x) - (1 - \lambda L) D_h(x^+, x). \tag{10}$$

In particular, the sufficiently decreasing property in the objective function value Ψ is ensured when $0 < \lambda L < 1$.

Proof From the global optimality of x^+ by taking $u = x \in$ int dom h and $\xi \in \partial_c f_2(x)$, we obtain

$$g(x^+) + \langle \nabla f_1(x) - \xi, x^+ - x \rangle + \frac{1}{\lambda} D_h(x^+, x) \leq g(x).$$

Invoking the full Extended Descent Lemma (Lemma 1) for f_1, the definition of the subgradient for f_2, and the above inequality, we have

$$
\begin{aligned}
f_1(x^+) - f_2(x^+) + g(x^+) &\leq f_1(x) - f_2(x) + \langle \nabla f_1(x) - \xi, x^+ - x \rangle + L D_h(x^+, x) + g(x^+) \\
&\leq f_1(x) - f_2(x) + L D_h(x^+, x) + g(x) - \frac{1}{\lambda} D_h(x^+, x) \\
&= f_1(x) - f_2(x) + g(x) - \left(\frac{1}{\lambda} - L \right) D_h(x^+, x),
\end{aligned}
$$

for $\Psi = f_1 - f_2 + g$. The last statement follows with $0 < \lambda L < 1$. $\qquad \square$

Proposition 1 follows immediately from Lemma 5, as in [8].

Proposition 1 *Suppose that Assumptions 1 and 2 hold. Let $\{x^k\}_{k=0}^{\infty}$ be a sequence generated by BPDCA with $0 < \lambda L < 1$. Then, the following statements hold*:

(i) *The sequence $\{\Psi(x^k)\}_{k=0}^{\infty}$ is non-increasing.*

(ii) $\sum_{k=1}^{\infty} D_h(x^k, x^{k-1}) < \infty$; *hence, the sequence $\{D_h(x^k, x^{k-1})\}_{k=0}^{\infty}$ converges to zero.*

(iii) $\min_{1 \le k \le n} D_h(x^k, x^{k-1}) \le \frac{\lambda}{n} \left(\frac{\Psi(x^0) - \Psi_*}{1 - \lambda L} \right)$, *where $\Psi_* = v(\mathcal{P}) > -\infty$ (by Assumption 1 (v)).*

4.2 Convergence analysis of BPDCA

Suppose that the following conditions hold.

Assumption 4

(i) $\operatorname{dom} h = \mathbb{R}^d$ and h is σ-strongly convex on $\mathbb{R}^d$.

(ii) ∇h and ∇f_1 are Lipschitz continuous on any bounded subset of $\mathbb{R}^d$.

(iii) The objective function Ψ is level-bounded; *i.e.*, for any $r \in \mathbb{R}$, the lower level sets $\{x \in \mathbb{R}^d \mid \Psi(x) \le r\}$ are bounded.

Since $C = \operatorname{int} \operatorname{dom} h = \mathbb{R}^d$ under Assumption 4 (i), Assumptions 2 and 3 are automatically fulfilled. For nonconvex functions, we use the limiting subdifferential [23] (Definition 1). Inspired by Fermat's rule [23, Theorem 10.1], we define the limiting critical points and the limiting stationary points of Ψ.

Definition 6 We say that $\tilde{x}$ is a limiting critical point of $(\mathcal{P})$ with $C \equiv \mathbb{R}^d$ if

$$0 \in \nabla f_1(\tilde{x}) - \partial_\circ f_2(\tilde{x}) + \partial g(\tilde{x}). \tag{11}$$

The set of all limiting critical points of $(\mathcal{P})$ is denoted by $\mathcal{X}$. In addition, we say that $\tilde{x}$ is a limiting stationary point of $(\mathcal{P})$ with $C \equiv \mathbb{R}^d$ if

$$0 \in \partial \Psi(\tilde{x}). \tag{12}$$

Although the limiting stationary points are sometimes called the limiting critical points in some papers, for example [7, Definition 1 (iv)], we distinguish these two terms. The reasons are the following: When Ψ is convex, we call $\tilde{x}$ a stationary point if it satisfies $0 \in \partial_c \Psi(\tilde{x})$. Because (12) is its natural extension by replacing $\partial_c \Psi$ with $\partial \Psi$, we use the terminology "limiting stationary point" after [11, Definition 6.1.4]. We similarly name $\tilde{x}$ satisfying (11): When g is convex, we call $\tilde{x}$ a critical point if it satisfies $0 \in \nabla f_1(\tilde{x}) - \partial_\circ f_2(\tilde{x}) + \partial_c g(\tilde{x})$. Because (11) is its natural extension by replacing $\partial_c g$ with ∂g, we use the terminology "limiting critical point."

The limiting stationary point is a first-order necessary condition for the local optimality. This relation is known as the generalized Fermat's rule [23, Theorem 10.1]. We can deduce $\partial(g - f_2)(x) \subseteq \partial g(x) - \partial_c f_2(x)$ from [19, Corollary 3.4]. Plugging it into [23, Corollary 10.9], it generally holds that $\partial \Psi(x) \subseteq \nabla f_1(x) - \partial_c f_2(x) + \partial g(x)$ for all $x \in \mathbb{R}^d$. It implies the limiting critical point is weaker than the limiting stationary point. When f_2 is $\mathcal{C}^1$ on $\mathbb{R}^d$, it holds that $\partial \Psi(x) \equiv \nabla f_1(x) - \nabla f_2(x) + \partial g(x)$ from [23, Corollary 10.9] or [18, Proposition 1.107 (ii)] and the definition of the limiting subdifferentials of f_2 and g. Thus, every limiting critical point is a limiting stationary point when f_2 is $\mathcal{C}^1$.

Next, using Lemma 5 and Proposition 1, we will show global subsequential convergence of the iterates to a limiting critical point of the problem ($\mathcal{P}$). We can easily see that Theorem 1 (i) holds from the level-boundedness of Ψ. Theorem 1 (iii) and (iv) will play an essential role in determining the global convergence and the rate of convergence of BPDCA.

Theorem 1 (Global subsequential convergence of BPDCA) *Suppose that Assumptions* 1, 2, *and* 4 *hold. Let* $\{x^k\}_{k=0}^\infty$ *be a sequence generated by BPDCA with* $0 < \lambda L < 1$ *for solving* ($\mathcal{P}$). *Then, the following statements hold*:

(i) *The sequence* $\{x^k\}_{k=0}^\infty$ *is bounded.*
(ii) *The sequence* $\{\xi^k\}_{k=0}^\infty$ *is bounded.*
(iii) $\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0$.
(iv) *Any accumulation point of* $\{x^k\}_{k=0}^\infty$ *is a limiting critical point of* ($\mathcal{P}$).

Proof (i) From Proposition 1, we obtain $\Psi(x^k) \leq \Psi(x^0)$ for all $k \in \mathbb{N}$, which shows that $\{x^k\}_{k=0}^\infty$ is bounded from Assumption 4 (iii).

(ii) From Assumption 1 (ii), 4 (i), and the convexity of f_2, $\mathrm{dom} f_2 = \mathbb{R}^d$ and $\partial_c f_2(x^k) \neq \emptyset$. Suppose, for the sake of proof by contradiction, that $\{\xi^k\}_{k=0}^\infty$ is unbounded, *i.e.*, $\|\xi^k\| \to \infty$ as $k \to \infty$. By the definition of the subgradients of convex functions, we have that, for any $y \in \mathbb{R}^d$,

$$f_2(y) \geq f_2(x^k) + \langle \xi^k, y - x^k \rangle. \tag{13}$$

Assume for a moment that $\|\xi^k\| \neq 0$. Letting $\{d^k\}_{k=0}^\infty$ be the subsequence given by $d^k = \xi^k / \|\xi^k\|$ and substituting $x^k + d^k = x^k + \xi^k / \|\xi^k\|$ into y in (13), we obtain

$$f_2(x^k + d^k) \geq f_2(x^k) + \langle \xi^k, d^k \rangle = f_2(x^k) + \|\xi^k\|,$$

which is also true when $\|\xi^k\| = 0$ by defining $d^k = 0$. By taking $k \to \infty$, we obtain

$$\limsup_{k\to\infty} \|\xi^k\| \leq \limsup_{k\to\infty} \left(f_2(x^k + d^k) - f_2(x^k) \right). \tag{14}$$

We can take a compact set S such that $\{x^k + d^k\}_{k=0}^{\infty} \subset S$, since $\{x^k + d^k\}_{k=0}^{\infty}$ is bounded. For $\bar{x} \in \text{argmax}_{x \in S} f_2(x)$, since f_2 is continuous because of its convexity on $\mathbb{R}^d$ and $\{x^k\}_{k=0}^{\infty}$ is bounded, it holds that

$$\limsup_{k \to \infty} \left(f_2(x^k + d^k) - f_2(x^k) \right) \le f_2(\bar{x}) - \bar{f}_2 < \infty, \tag{15}$$

for some value $\bar{f}_2 \le f_2(x^k), k \ge 0$. (14) and (15) contradict to $\|\xi^k\| \to \infty$.

(iii) From (10), we obtain

$$\begin{aligned} \Psi(x^{k-1}) - \Psi(x^k) &\ge \left(\frac{1}{\lambda} - L \right) D_h(x^k, x^{k-1}) \\ &\ge \left(\frac{1}{\lambda} - L \right) \frac{\sigma}{2} \|x^k - x^{k-1}\|^2, \end{aligned} \tag{16}$$

where the last inequality holds since h is a σ-strongly convex function from Assumption 4 (i). Summing the above inequality from $k = 1$ to ∞, we obtain

$$\left(\frac{1}{\lambda} - L \right) \sum_{k=1}^{\infty} \frac{\sigma}{2} \|x^k - x^{k-1}\|^2 \le \Psi(x^0) - \liminf_{n \to \infty} \Psi(x^n) \le \Psi(x^0) - v(\mathcal{P}) < \infty,$$

which shows that $\lim_{k \to \infty} \|x^{k+1} - x^k\| = 0$.

(iv) Let $\tilde{x}$ be an accumulation point of $\{x^k\}_{k=0}^{\infty}$ and let $\{x^{k_j}\}$ be a subsequence such that $\lim_{j \to \infty} x^{k_j} = \tilde{x}$. Then, from the first-order optimality condition of subproblem (5) under Assumption 2, we have

$$0 \in \partial g(x^{k_j+1}) + \nabla f_1(x^{k_j}) - \xi^{k_j} + \frac{1}{\lambda} \left(\nabla h(x^{k_j+1}) - \nabla h(x^{k_j}) \right).$$

Therefore,

$$\xi^{k_j} + \frac{1}{\lambda} \left(\nabla h(x^{k_j}) - \nabla h(x^{k_j+1}) \right) \in \partial g(x^{k_j+1}) + \nabla f_1(x^{k_j}). \tag{17}$$

From the boundedness of $\{x^{k_j}\}$ and the Lipschitz continuity of ∇h on a bounded subset of $\mathbb{R}^d$, there exists $A_0 > 0$ such that

$$\left\| \frac{1}{\lambda} \left(\nabla h(x^{k_j}) - \nabla h(x^{k_j+1}) \right) \right\| \le \frac{A_0}{\lambda} \|x^{k_j+1} - x^{k_j}\|.$$

Therefore, using $\|x^{k_j+1} - x^{k_j}\| \to 0$, we obtain

$$\frac{1}{\lambda} \left(\nabla h(x^{k_j}) - \nabla h(x^{k_j+1}) \right) \to 0. \tag{18}$$

Note that the sequence $\{\xi^{k_j}\}$ is bounded due to (ii). Thus, by taking the limit as $j \to \infty$ or more precisely, its subsequence, we can assume without loss of generality that $\lim_{j \to \infty} \xi^{k_j} =: \tilde{\xi}$ exists, which belongs to $\partial_c f_2(\tilde{x})$ since f_2 becomes continuous due to its convexity on $\mathbb{R}^d$. Using this and (18), we can take the limit of (17). Setting $\|x^{k_j+1} - x^{k_j}\| \to 0$ and invoking the lower semicontinuity of g and ∇f_1, we obtain

$\tilde{\xi} \in \partial g(\tilde{x}) + \nabla f_1(\tilde{x})$. Therefore, $0 \in \partial g(\tilde{x}) + \nabla f_1(\tilde{x}) - \partial_c f_2(\tilde{x})$, which shows that $\tilde{x}$ is a limiting critical point of $(\mathcal{P})$. $\qquad\square$

We can estimate the objective value at an accumulation point from $\liminf_{j\to\infty} \Psi(x^{k_j})$ and $\limsup_{j\to\infty} \Psi(x^{k_j})$. Consequently, we can prove that Ψ is constant on the set of accumulation points of BPDCA.

Proposition 2 *Suppose that Assumptions* 1, 2, *and* 4 *hold. Let* $\{x^k\}_{k=0}^{\infty}$ *be a sequence generated by BPDCA with* $0 < \lambda L < 1$ *for solving* $(\mathcal{P})$. *Then, the following statements hold*:

(i) $\zeta := \lim_{k\to\infty} \Psi(x^k)$ *exists.*
(ii) $\Psi \equiv \zeta$ *on* Ω, *where* Ω *is the set of accumulation points of* $\{x^k\}_{k=0}^{\infty}$.

Proof (i) From Assumption 1 (v) and Proposition 1 (i), the sequence $\{\Psi(x^k)\}_{k=0}^{\infty}$ is bounded from below and non-increasing. Consequently, $\zeta := \lim_{k\to\infty} \Psi(x^k)$ exists.

(ii) Take any $\hat{x} \in \Omega$, that is $\lim_{j\to\infty} x^{k_j} = \hat{x}$. From (5), it follows that

$$g(x^k) + \langle \nabla f_1(x^{k-1}) - \xi^{k-1}, x^k - x^{k-1} \rangle + \frac{1}{\lambda} D_h(x^k, x^{k-1})$$
$$\leq g(\hat{x}) + \langle \nabla f_1(x^{k-1}) - \xi^{k-1}, \hat{x} - x^{k-1} \rangle + \frac{1}{\lambda} D_h(\hat{x}, x^{k-1}).$$

From the above inequality and the fact that f_1 is convex at x^k, we obtain

$$g(x^k) + f_1(x^k) \leq g(\hat{x}) + \langle \nabla f_1(x^{k-1}) - \xi^{k-1}, \hat{x} - x^k \rangle + \frac{1}{\lambda} D_h(\hat{x}, x^{k-1}) - \frac{1}{\lambda} D_h(x^k, x^{k-1})$$
$$+ f_1(\hat{x}) + \langle \nabla f_1(x^k), x^k - \hat{x} \rangle.$$

Substituting k_j for k and limiting j to ∞, we have, from Proposition 1 (ii),

$$\limsup_{j\to\infty} \left(g(x^{k_j}) + f_1(x^{k_j}) \right) \leq g(\hat{x}) + f_1(\hat{x}),$$

which provides $\limsup_{j\to\infty} \Psi(x^{k_j}) \leq \Psi(\hat{x})$ from the continuity of $-f_2$ since f_2 is convex. Combining this and the lower semicontinuity of Ψ yields $\Psi(x^{k_j}) \to \Psi(\hat{x}) =: \zeta$ as $j \to \infty$. Since $\hat{x} \in \Omega$ is arbitrary, we conclude that $\Psi \equiv \zeta$ on Ω. $\qquad\square$

To discuss the global convergence of BPDCA, we will suppose either of the following two assumptions.

Assumption 5 f_2 is continuously differentiable on an open set $\mathcal{N}_0 \subset \mathbb{R}^d$ that contains the set of all limiting critical points of Ψ, *i.e.*, $\mathcal{X}$. Furthermore, ∇f_2 is locally Lipschitz continuous on $\mathcal{N}_0$.

Assumption 6 g is differentiable on $\mathbb{R}^d$ and ∇g is locally Lipschitz continuous on an open set $\mathcal{N}_0 \subset \mathbb{R}^d$ that contains the set of all limiting stationary points of $-\Psi$.

Assumption 5 is nonrestrictive because many functions in [28], including the f_2 in our numerical experiments, satisfy it. Thus, let us discuss the global convergence of Algorithm 1 under Assumption 5 by following the argument presented in [28]. Note that every limiting critical point is a limiting stationary point from the differentiability of f_2 under Assumption 5.

Theorem 2 (Global convergence of BPDCA under the local differentiability of f_2) *Suppose that Assumptions* 1, 2, 4, *and* 5 *hold and that* Ψ *is a KL function. Let* $\{x^k\}_{k=0}^{\infty}$ *be a sequence generated by BPDCA with* $0 < \lambda L < 1$ *for solving* ($\mathcal{P}$). *Then, the following statements hold*:

(i) $\lim_{k \to \infty}$ dist $(0, \partial \Psi(x^k)) = 0$.
(ii) *The sequence* $\{x^k\}_{k=0}^{\infty}$ *converges to a limiting stationary point of* ($\mathcal{P}$); *moreover*, $\sum_{k=1}^{\infty} \|x^k - x^{k-1}\| < \infty$.

Proof (i) Since $\{x^k\}_{k=0}^{\infty}$ is bounded and Ω is the set of accumulation points of $\{x^k\}_{k=0}^{\infty}$, we have

$$\lim_{k \to \infty} \text{dist}\,(x^k, \Omega) = 0. \tag{19}$$

From Theorem 1 (iv), we also have $\Omega \subseteq \mathcal{X}$. Thus, for any $\mu > 0$, there exists $k_0 > 0$ such that dist $(x^k, \Omega) < \mu$ and $x^k \in \mathcal{N}_0$ for any $k \geq k_0$, where $\mathcal{N}_0$ is defined in Assumption 5. As for $\mathcal{N}_0$, since Ω is compact from the boundedness of $\{x^k\}_{k=0}^{\infty}$, by decreasing μ, if needed, we can suppose without loss of generality that ∇f_2 is globally Lipschitz continuous on $\mathcal{N} := \{x \in \mathcal{N}_0 \mid \text{dist}\,(x, \Omega) < \mu\}$.

The subdifferential of Ψ at x^k for $k \geq k_0$ is

$$\partial \Psi(x^k) = \nabla f_1(x^k) - \nabla f_2(x^k) + \partial g(x^k). \tag{20}$$

Moreover, considering the first-order optimality condition of subproblem (5), we have that, for any $k \geq k_0 + 1$,

$$\frac{1}{\lambda} \left(\nabla h(x^{k-1}) - \nabla h(x^k) \right) - \nabla f_1(x^{k-1}) + \nabla f_2(x^{k-1}) \in \partial g(x^k),$$

since f_2 is $\mathcal{C}^1$ on $\mathcal{N}$ and $x^{k-1} \in \mathcal{N}$ for any $k \geq k_0 + 1$. Using the above and (20), we see that

$$\frac{1}{\lambda} \left(\nabla h(x^{k-1}) - \nabla h(x^k) \right) + \nabla f_1(x^k) - \nabla f_1(x^{k-1}) + \nabla f_2(x^{k-1}) - \nabla f_2(x^k) \in \partial \Psi(x^k).$$

From the global Lipschitz continuity of ∇f_1, ∇f_2, and ∇h, there exists $A_1 > 0$ such that

$$\text{dist}\,(0, \partial \Psi(x^k)) \leq A_1 \|x^k - x^{k-1}\|, \tag{21}$$

where $\quad k \geq k_0 + 1$. $\quad$ From $\quad$ Theorem $\quad$ 1 $\quad$ (iii), $\quad$ we $\quad$ conclude $\quad$ that $\lim_{k \to \infty} \operatorname{dist}(0, \partial \Psi(x^k)) = 0$.

(ii) From Theorem 1 (iv), it is sufficient to prove that $\{x^k\}_{k=0}^{\infty}$ is convergent. Here, consider the case in which there exists a positive integer $k > 0$ such that $\Psi(x^k) = \zeta$. From Proposition 1 (i) and Proposition 2 (i), the sequence $\{\Psi(x^k)\}_{k=0}^{\infty}$ is non-increasing and converges to ζ. Hence, for any $\hat{k} \geq 0$, we have $\Psi(x^{k+\hat{k}}) = \zeta$. By recalling (16), we conclude that there exists a positive scalar A_2 such that

$$\Psi(x^{k-1}) - \Psi(x^k) \geq A_2 \|x^k - x^{k-1}\|^2, \quad \forall k \in \mathbb{N}. \tag{22}$$

From (22), we obtain $x^k = x^{k+\hat{k}}$ for any $\hat{k} \geq 0$, meaning that $\{x^k\}_{k=0}^{\infty}$ is finitely convergent.

Next, consider the case in which $\Psi(x^k) > \zeta$ for all $k \geq 0$. Since $\{x^k\}_{k=0}^{\infty}$ is bounded, Ω is a compact subset of $\operatorname{dom} \partial \Psi$ and $\Psi \equiv \zeta$ on Ω from Proposition 2 (ii). From Lemma 2 and since Ψ is a KL function, there exist a positive scalar $\epsilon > 0$ and a continuous concave function $\phi \in \Xi_\eta$ with $\eta > 0$ such that

$$\phi'(\Psi(x) - \zeta) \cdot \operatorname{dist}(0, \partial \Psi(x)) \geq 1, \tag{23}$$

for all $x \in U$, where

$$U = \left\{ x \in \mathbb{R}^d \mid \operatorname{dist}(x, \Omega) < \epsilon \right\} \cap \left\{ x \in \mathbb{R}^d \mid \zeta < \Psi(x) < \zeta + \eta \right\}.$$

From (19), there exists $k_1 > 0$ such that $\operatorname{dist}(x^k, \Omega) < \epsilon$ for any $k \geq k_1$. Since $\{\Psi(x^k)\}_{k=0}^{\infty}$ is non-increasing and converges to ζ, there exists $k_2 > 0$ such that $\zeta < \Psi(x^k) < \zeta + \eta$ for all $k \geq k_2$. Taking $\bar{k} = \max\{k_0 + 1, k_1, k_2\}$, then $\{x^k\}_{k \geq \bar{k}}$ belongs to U. Hence, from (23), we obtain

$$\phi'(\Psi(x^k) - \zeta) \cdot \operatorname{dist}(0, \partial \Psi(x^k)) \geq 1, \quad \forall k \geq \bar{k}. \tag{24}$$

Since ϕ is a concave function, we see that for any $k \geq \bar{k}$,

$$
\begin{aligned}
\left[\phi(\Psi(x^k) - \zeta) - \phi(\Psi(x^{k+1}) - \zeta) \right] & \cdot \operatorname{dist}(0, \partial \Psi(x^k)) \\
& \geq \phi'(\Psi(x^k) - \zeta) \cdot \operatorname{dist}(0, \partial \Psi(x^k)) \cdot \left(\Psi(x^k) - \Psi(x^{k+1}) \right) \\
& \geq \Psi(x^k) - \Psi(x^{k+1}) \\
& \geq A_2 \|x^{k+1} - x^k\|^2,
\end{aligned}
$$

where the second inequality holds from (24) and the fact that $\{\Psi(x^k)\}_{k=0}^{\infty}$ is non-increasing, and the last inequality holds from (22). From the above inequality and (21), we obtain

$$\|x^{k+1} - x^k\|^2 \leq \frac{A_1}{A_2} \left(\phi(\Psi(x^k) - \zeta) - \phi(\Psi(x^{k+1}) - \zeta) \right) \|x^k - x^{k-1}\|. \tag{25}$$

Taking the square root of (25) and using the inequality of the arithmetic and geometric means, we find that

$$\|x^{k+1} - x^k\| \leq \sqrt{\frac{A_1}{A_2} \left(\phi(\Psi(x^k) - \zeta) - \phi(\Psi(x^{k+1}) - \zeta) \right)} \cdot \sqrt{\|x^k - x^{k-1}\|}$$

$$\leq \frac{A_1}{2A_2} \left(\phi(\Psi(x^k) - \zeta) - \phi(\Psi(x^{k+1}) - \zeta) \right) + \frac{1}{2} \|x^k - x^{k-1}\|.$$

This shows that

$$\frac{1}{2} \|x^{k+1} - x^k\| \leq \frac{A_1}{2A_2} \left(\phi(\Psi(x^k) - \zeta) - \phi(\Psi(x^{k+1}) - \zeta) \right) + \frac{1}{2} \|x^k - x^{k-1}\| - \frac{1}{2} \|x^{k+1} - x^k\|. \tag{26}$$

Summing (26) from $k = \bar{k}$ to ∞, we have

$$\sum_{k=\bar{k}}^{\infty} \|x^{k+1} - x^k\| \leq \frac{A_1}{A_2} \phi(\Psi(x^{\bar{k}}) - \zeta) + \|x^{\bar{k}} - x^{\bar{k}-1}\| < \infty,$$

which implies that $\sum_{k=1}^{\infty} \|x^k - x^{k-1}\| < \infty$, *i.e.*, the sequence $\{x^k\}_{k=0}^{\infty}$ is a Cauchy sequence. Thus, $\{x^k\}_{k=0}^{\infty}$ converges to a limiting critical point of $(\mathcal{P})$ from Theorem 1 (iv). Because every limiting critical point is a limiting stationary point from the differentiability of f_2, $\{x^k\}_{k=0}^{\infty}$ converges to a limiting stationary point of $(\mathcal{P})$. □

Next, suppose that Assumption 6 holds instead of Assumption 5. Here, we can show the global convergence of BPDCA by referring to [16, Theorem 3.4]. We will use subanalyticity instead of the KL property in the proof.

Theorem 3 (Global convergence of BPDCA under the local differentiability of g) *Suppose that Assumptions* 1, 2, 4, *and* 6 *hold and that* Ψ *is subanalytic. Let* $\{x^k\}_{k=0}^{\infty}$ *be a sequence generated by BPDCA with* $0 < \lambda L < 1$ *for solving* $(\mathcal{P})$. *Then, the sequence* $\{x^k\}_{k=0}^{\infty}$ *converges to a limiting critical point of* $(\mathcal{P})$; *moreover,* $\sum_{k=1}^{\infty} \|x^k - x^{k-1}\| < \infty$.

Proof Since g is differentiable, g is continuous on $\mathbb{R}^d$. Since the convexity of f_1 and f_2 implies their continuity, Ψ is continuous on $\mathbb{R}^d$.

Let $\{\xi^k\}_{k=0}^{\infty}$ on $\mathbb{R}^d$ be a sequence of subgradients of f_2. From Theorem 1 (i) and (ii), $\{x^k\}_{k=0}^{\infty}$ and $\{\xi^k\}_{k=0}^{\infty}$ are bounded. Let $\tilde{x}$ be a limiting stationary point of $-\Psi$ and $B(\tilde{x}, \epsilon_0)$ be an open ball with center $\tilde{x}$ and radius $\epsilon_0 > 0$. Since ∇g is locally Lipschitz continuous, and Assumption 4 (ii) holds, for $\lambda > 0$, there exist $\kappa_0 > 0$ and $\epsilon_0 > 0$ such that

$$\left\| \nabla \left(g + \frac{1}{\lambda} h \right)(u) - \nabla \left(g + \frac{1}{\lambda} h \right)(v) \right\| \leq \kappa_0 \|u - v\|, \quad \forall u, v \in B(\tilde{x}, \epsilon_0). \tag{27}$$

From Assumption 1 (v), $-\Psi$ is finite. In addition, by recalling the continuity and subanalyticity of $-\Psi$ on $B(\tilde{x}, \epsilon_0)$, we can apply [6, Theorem 3.1] to the subanalytic function $-\Psi$ and obtain $\nu_0 > 0$ and $\theta_0 \in [0, 1)$ such that

$$|\Psi(u) - \zeta|^{\theta_0} \leq \nu_0 \|\hat{x}\|, \quad \forall u \in B(\tilde{x}, \epsilon_0), \quad \hat{x} \in \partial(-\Psi)(u), \tag{28}$$

where $\zeta = \Psi(\tilde{x})$.

Let Ω be the set of accumulation points of $\{x^k\}_{k=0}^\infty$. Since Ω is compact, Ω can be covered by a finite number of $B(\tilde{x}_j, \epsilon_j)$ with $\tilde{x}_j \in \Omega$ and $\epsilon_j > 0$, $j = 1, \ldots, p$. From Theorem 1 (iv), $\tilde{x}_j \in \Omega, j = 1, \ldots, p$ are limiting critical points of $(\mathcal{P})$. Hence, (27) with $\kappa_j > 0$ and $\epsilon_j > 0$ and (28) with $\nu_j > 0$ and $\theta_j \in [0, 1)$ hold for $j = 1, \ldots, p$. Letting $\epsilon > 0$ be a sufficiently small constant, we obtain

$$\left\{x \in \mathbb{R}^d \mid \text{dist}(x, \Omega) < \epsilon\right\} \subset \bigcup_{j=1}^p B(\tilde{x}_j, \epsilon_j).$$

From (19), there exists $k_1 > 0$ such that $\text{dist}(x^k, \Omega) < \epsilon$ for any $k \geq k_1$; hence, $x^k \in \bigcup_{j=1}^p B(\tilde{x}_j, \epsilon_j)$ for any $k \geq k_1$. From Theorem 1 (iii), letting $\bar{\epsilon} > 0$ be a sufficiently small constant, there exists $k_2 > 0$ such that $\|x^k - x^{k+1}\| \leq \frac{\bar{\epsilon}}{2}$ for any $k \geq k_2$. Therefore, redefining $\bar{\epsilon}, \epsilon_j, j = 1, \ldots, p$ and relabeling if necessary, we can assume without loss of generality that

$$x^k \in \bigcup_{j=1}^p B\left(\tilde{x}_j, \frac{\epsilon_j}{2}\right) \quad \text{and} \quad \|x^k - x^{k+1}\| \leq \frac{\bar{\epsilon}}{2}, \quad \forall k \geq \bar{k},$$

where $\bar{\epsilon} = \min_{j=1,\ldots,p} \epsilon_j$ and $\bar{k} = \max\{k_1, k_2\}$, which implies $x^k \in B(\tilde{x}_{j_k}, \epsilon_{j_k}/2), j_k \in \{1, \ldots, p\}$ and hence $x^{k+1} \in B(\tilde{x}_{j_k}, \epsilon_{j_k})$. Thus, from (27) and (28), we have

$$\left\|\nabla\left(g + \frac{1}{\lambda}h\right)(x^k) - \nabla\left(g + \frac{1}{\lambda}h\right)(x^{k+1})\right\| \leq \kappa\|x^k - x^{k+1}\|, \tag{29}$$

$$|\Psi(x^k) - \zeta|^\theta \leq \nu\|\hat{x}^k\|, \quad \hat{x}^k \in \partial(-\Psi)(x^k), \quad \forall k \geq \bar{k}, \tag{30}$$

where $\kappa = \max_{j=1,\ldots,p} \kappa_j, \nu = \max_{j=1,\ldots,p} \nu_j$, and $\theta = \max_{j=1,\ldots,p} \theta_j$. From (5), we find that

$$0 = \nabla g(x^{k+1}) + \nabla f_1(x^k) - \xi^k + \frac{1}{\lambda}\left(\nabla h(x^{k+1}) - \nabla h(x^k)\right),$$

which implies

$$\nabla g(x^{k+1}) - \nabla g(x^k) + \frac{1}{\lambda}\left(\nabla h(x^{k+1}) - \nabla h(x^k)\right) = \xi^k - \nabla f_1(x^k) - \nabla g(x^k) \in \partial(-\Psi)(x^k),$$

where we have used $\partial(-\Psi)(x^k) = \partial_\circ f_2(x^k) - \nabla f_1(x^k) - \nabla g(x^k)$, which comes from the convexity of f_2. Using (29) and (30), we obtain

$$|\Psi(x^k) - \zeta|^\theta \leq \nu\left\|\nabla\left(g + \frac{1}{\lambda}h\right)(x^k) - \nabla\left(g + \frac{1}{\lambda}h\right)(x^{k+1})\right\| \leq \kappa\nu\|x^k - x^{k+1}\|, \quad \forall k \geq \bar{k}. \tag{31}$$

Since the function $t^{1-\theta}$ is concave on $[0, \infty)$, $\Psi(x^k) - \zeta \geq 0$, (16), and (31), we find that, for all $k \geq \bar{k}$,

$$
\begin{aligned}
(\Psi(x^k) - \zeta)^{1-\theta} - (\Psi(x^{k+1}) - \zeta)^{1-\theta} &\geq (1-\theta)(\Psi(x^k) - \zeta)^{-\theta}(\Psi(x^k) - \Psi(x^{k+1})) \\
&\geq \frac{1-\theta}{\kappa v \|x^k - x^{k+1}\|}\left(\frac{1}{\lambda} - L\right)\frac{\sigma}{2}\|x^k - x^{k+1}\|^2 \\
&= \frac{(1-\theta)\sigma}{2\kappa v}\left(\frac{1}{\lambda} - L\right)\|x^k - x^{k+1}\|.
\end{aligned} \tag{32}
$$

Summing (32) from $k = \bar{k}$ to ∞ yields

$$
\sum_{k=\bar{k}}^{\infty} \|x^k - x^{k+1}\| \leq \frac{2\kappa v}{(1/\lambda - L)(1-\theta)\sigma}(\Psi(x^{\bar{k}}) - \zeta)^{1-\theta} < \infty,
$$

which implies that $\sum_{k=1}^{\infty} \|x^k - x^{k-1}\| < \infty$, *i.e.*, the sequence $\{x^k\}_{k=0}^{\infty}$ is a Cauchy sequence. Thus, $\{x^k\}_{k=0}^{\infty}$ converges to a limiting critical point of $(\mathcal{P})$ from Theorem 1 (iv). $\qquad\square$

Finally, we will show the rate of convergence in a manner following [1, 28].

Theorem 4 (Rate of convergence under the local differentiability of f_2) *Suppose that Assumptions* 1, 2, 4, *and* 5 *hold. Let* $\{x^k\}_{k=0}^{\infty}$ *be a sequence generated by BPDCA with* $0 < \lambda L < 1$ *for solving* $(\mathcal{P})$ *and suppose that* $\{x^k\}_{k=0}^{\infty}$ *converges to some* $\tilde{x} \in \mathcal{X}$. *Suppose further that* Ψ *is a KL function with* ϕ *in the KL inequality* (4) *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* $k_0 > 0$ *such that* x^k *is constant for* $k > k_0$;

(ii) *If* $\theta \in (0, \frac{1}{2}]$, *then there exist* $c_1 > 0, k_1 > 0$, *and* $\eta \in (0, 1)$ *such that* $\|x^k - \tilde{x}\| < c_1 \eta^k$ *for* $k > k_1$;

(iii) *If* $\theta \in (\frac{1}{2}, 1)$, *then there exist* $c_2 > 0$ *and* $k_2 > 0$ *such that* $\|x^k - \tilde{x}\| < c_2 k^{-\frac{1-\theta}{2\theta-1}}$ *for* $k > k_2$.

Proof (i) For the case of $\theta = 0$, we will prove that there exists an integer $k_0 > 0$ such that $\Psi(x^{k_0}) = \zeta$ by assuming to the contrary that $\Psi(x^k) > \zeta$ for all $k > 0$ and showing a contradiction. The sequence $\{\Psi(x^k)\}_{k=0}^{\infty}$ converges to ζ due to Proposition 2 (i). In addition, from the KL inequality (24) and $\phi'(\cdot) = c$, we can see that for all sufficiently large k,

$$
\text{dist}\,(0, \partial\Psi(x^k)) \geq \frac{1}{c},
$$

which contradicts Theorem 2 (i). Therefore, there exists $k_0 > 0$ such that $\Psi(x^{k_0}) = \zeta$. Since $\{\Psi(x^k)\}_{k=0}^{\infty}$ is non-increasing and converges to ζ, we have $\Psi(x^{k_0 + \bar{k}}) = \zeta$ for all $\bar{k} \geq 0$. This, together with (22), lead us to conclude that there exists $k_0 > 0$ such that x^k is constant for $k > k_0$.

(ii–iii) Next, consider the case $\theta \in (0,1)$. If there exists $k_0 > 0$ such that $\Psi(x^{k_0}) = \zeta$, then we can show that the sequence $\{x^k\}_{k=0}^{\infty}$ is finitely convergent in the same way as in the proof of (i). Therefore, for $\theta \in (0,1)$, we only need to consider the case that $\Psi(x^k) > \zeta$ for all $k > 0$.

Define $R_k = \Psi(x^k) - \zeta$ and $S_k = \sum_{j=k}^{\infty} \|x^{j+1} - x^j\|$, where S_k is well-defined due to Theorem 2 (ii). From (26), for any $k \geq \bar{k}$, where $\bar{k}$ is defined in (24), we obtain

$$
\begin{aligned}
S_k &= 2 \sum_{j=k}^{\infty} \frac{1}{2} \|x^{j+1} - x^j\| \\
&\leq 2 \sum_{j=k}^{\infty} \left[\frac{A_1}{2A_2} \left(\phi(\Psi(x^j) - \zeta) - \phi(\Psi(x^{j+1}) - \zeta) \right) + \frac{1}{2} \|x^j - x^{j-1}\| - \frac{1}{2} \|x^{j+1} - x^j\| \right] \\
&\leq \frac{A_1}{A_2} \phi(\Psi(x^k) - \zeta) + \|x^k - x^{k-1}\| \\
&= \frac{A_1}{A_2} \phi(R_k) + S_{k-1} - S_k.
\end{aligned}
\tag{33}
$$

On the other hand, since $\lim_{k\to\infty} x^k = \tilde{x}$ and $\{\Psi(x^k)\}$ is non-increasing and converges to ζ, the KL inequality (24) with $\phi(s) = cs^{1-\theta}$ ensures that, for all sufficiently large k,

$$
c(1-\theta)R_k^{-\theta} \operatorname{dist}(0, \partial\Psi(x^k)) \geq 1.
\tag{34}
$$

From the definition of S_k and (21), we also have that, for all sufficiently large k,

$$
\operatorname{dist}(0, \partial\Psi(x^k)) \leq A_1(S_{k-1} - S_k).
\tag{35}
$$

Combining (34) and (35), we have $R_k^{\theta} \leq A_1 \cdot c(1-\theta)(S_{k-1} - S_k)$ for all sufficiently large k. Raising the above inequality to the power of $\frac{1-\theta}{\theta}$ and scaling both sides by c, we find that $cR_k^{1-\theta} \leq c(A_1 \cdot c(1-\theta)(S_{k-1} - S_k))^{\frac{1-\theta}{\theta}}$. Combining this with (33) and recalling $\phi(R_k) = cR_k^{1-\theta}$, we find that, for all sufficiently large k,

$$
S_k \leq A_3(S_{k-1} - S_k)^{\frac{1-\theta}{\theta}} + S_{k-1} - S_k,
\tag{36}
$$

where $A_3 = \frac{A_1}{A_2} c(A_1 \cdot c(1-\theta))^{\frac{1-\theta}{\theta}}$.

(ii) When $\theta \in (0, \frac{1}{2}]$, we have $\frac{1-\theta}{\theta} \geq 1$. Since $\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0$ by Theorem 1 (iii), $\lim_{k\to\infty} S_{k-1} - S_k = 0$. From these considerations and (36), we conclude that there exists $k_1 > 0$ such that for all $k \geq k_1$, $S_k \leq (A_3 + 1)(S_{k-1} - S_k)$, which implies $S_k \leq \frac{A_3+1}{A_3+2} S_{k-1}$. Therefore, for all $k \geq k_1$,

$$
\|x^k - \tilde{x}\| \leq \sum_{j=k}^{\infty} \|x^{j+1} - x^j\| = S_k \leq S_{k_1-1} \left(\frac{A_3+1}{A_3+2} \right)^{k-k_1+1}.
$$

(iii) For $\theta \in (\frac{1}{2}, 1)$, $\frac{1-\theta}{\theta} < 1$. From (36) and $\lim_{k\to\infty} S_{k-1} - S_k = 0$, there exists $k_2 > 0$ such that

$$S_k \leq A_3(S_{k-1} - S_k)^{\frac{1-\theta}{\theta}} + S_{k-1} - S_k \leq A_3(S_{k-1} - S_k)^{\frac{1-\theta}{\theta}} + (S_{k-1} - S_k)^{\frac{1-\theta}{\theta}}$$
$$\leq (A_3 + 1)(S_{k-1} - S_k)^{\frac{1-\theta}{\theta}},$$

for all $k \geq k_2$. Raising the above inequality to the power of $\frac{\theta}{1-\theta}$, for any $k \geq k_2$ we find that $S_k^{\frac{\theta}{1-\theta}} \leq A_4(S_{k-1} - S_k)$, where $A_4 = (A_3 + 1)^{\frac{\theta}{1-\theta}}$. From [1, Theorem 2], we find that, for all sufficiently large k, there exists $A_5 > 0$ such that $S_k \leq A_5 k^{-\frac{1-\theta}{2\theta-1}}$. $\qquad \square$

Using Theorem 3, we can obtain another rate of convergence in the same way as in the proof of [1, Theorem 2] or [16, Theorem 3.5].

Theorem 5 (Rate of convergence under the local differentiability of g) *Suppose that Assumptions* 1, 2, 4, *and* 6 *hold. Let* $\{x^k\}_{k=0}^{\infty}$ *be a sequence generated by BPDCA with* $0 < \lambda L < 1$ *for solving* $(\mathcal{P})$ *and suppose that* $\{x^k\}_{k=0}^{\infty}$ *converges to some* $\tilde{x} \in \mathcal{X}$. *Suppose further that* Ψ *is subanalytic. Let* $\theta \in [0, 1)$ *be a Łojasiewicz exponent of* $\tilde{x}$. *Then, the following statements hold:*

(i) *If* $\theta = 0$, *then there exists* $k_0 > 0$ *such that* x^k *is constant for* $k > k_0$;
(ii) *If* $\theta \in (0, \frac{1}{2}]$, *then there exist* $c_1 > 0, k_1 > 0$, *and* $\eta \in (0, 1)$ *such that* $\|x^k - \tilde{x}\| < c_1 \eta^k$ *for* $k > k_1$;
(iii) *If* $\theta \in (\frac{1}{2}, 1)$, *then there exist* $c_2 > 0$ *and* $k_2 > 0$ *such that* $\|x^k - \tilde{x}\| < c_2 k^{-\frac{1-\theta}{2\theta-1}}$ *for* $k > k_2$.

4.3 Properties of BPDCAe

Inspired by [29], we introduce the auxiliary function,

$$H_M(x, y) = \Psi(x) + MD_h(y, x), \quad M > 0.$$

To show the decreasing property of H_M, instead of Ψ, with respect to $\{x^k\}_{k=0}^{\infty}$, we further assume the convexity of g.

Assumption 7 The function g is convex.

Under the adaptive restart scheme (see (8)), we show the decreasing property of H_M.

Lemma 6 *Suppose that Assumptions* 1, 3, *and* 7 *hold. For any* $x^k, y^k \in C = \text{int dom } h$ *and any* $x^{k+1} \in C = \text{int dom } h$ *defined by*

$$x^{k+1} \in \underset{y \in \mathbb{R}^d}{\text{argmin}} \left\{ g(y) + \langle \nabla f_1(y^k) - \xi^k, y - y^k \rangle + \frac{1}{\lambda} D_h(y, y^k) \right\}, \qquad (37)$$

where $\xi^k \in \partial_c f_2(x^k)$, $y^k = x^k + \beta_k(x^k - x^{k-1})$, $\lambda > 0$, and $\{\beta_k\}_{k=0}^{\infty} \subset [0, 1)$, it holds that

$$\lambda \Psi(x^{k+1}) \leq \lambda \Psi(x^k) + D_h(x^k, y^k) - D_h(x^k, x^{k+1}) - (1 - \lambda L)D_h(x^{k+1}, y^k). \quad (38)$$

Furthermore, when $0 < \lambda L < 1$ and $\{\beta_k\}_{k=0}^{\infty}$ is given by the adaptive restart scheme (8),

$$H_M(x^{k+1}, x^k) \leq H_M(x^k, x^{k-1}) - \left(\frac{1}{\lambda} - M\right)D_h(x^k, x^{k+1})$$
$$- \left(M - \frac{\rho}{\lambda}\right)D_h(x^{k-1}, x^k) - \left(\frac{1}{\lambda} - L\right)D_h(x^{k+1}, y^k). \quad (39)$$

In addition, when $\frac{\rho}{\lambda} \leq M \leq \frac{1}{\lambda}$ for $\rho \in [0, 1)$, the auxiliary function H_M is ensured to be non-increasing.

Proof From the first-order optimality condition for (37), we obtain

$$0 \in \partial_c g(x^{k+1}) + \nabla f_1(y^k) - \xi^k + \frac{1}{\lambda}(\nabla h(x^{k+1}) - \nabla h(y^k)).$$

From the convexity of g, we find that

$$g(x^k) - g(x^{k+1}) \geq \langle -\nabla f_1(y^k) + \xi^k - \frac{1}{\lambda}(\nabla h(x^{k+1}) - \nabla h(y^k)), x^k - x^{k+1} \rangle.$$

Using the three-point identity (3) of the Bregman distances,

$$\frac{1}{\lambda}\langle \nabla h(x^{k+1}) - \nabla h(y^k), x^k - x^{k+1} \rangle = \frac{1}{\lambda}(D_h(x^k, y^k) - D_h(x^k, x^{k+1}) - D_h(x^{k+1}, y^k)),$$

we have

$$g(x^k) - g(x^{k+1}) + f_1(x^k) - f_1(x^{k+1}) \geq f_1(x^k) - f_1(x^{k+1}) + \langle -\nabla f_1(y^k) + \xi^k, x^k - x^{k+1} \rangle$$
$$- \frac{1}{\lambda}\left(D_h(x^k, y^k) - D_h(x^k, x^{k+1}) - D_h(x^{k+1}, y^k)\right).$$

From the convexity of f_1 and Lemma 1, we find that

$$f_1(x^k) - f_1(x^{k+1}) - \langle \nabla f_1(y^k), x^k - x^{k+1} \rangle$$
$$= f_1(x^k) - f_1(y^k) - \langle \nabla f_1(y^k), x^k - y^k \rangle - f_1(x^{k+1}) + f_1(y^k) + \langle \nabla f_1(y^k), x^{k+1} - y^k \rangle$$
$$\geq -LD_h(x^{k+1}, y^k).$$

The above inequalities and the definition of the subgradient for f_2 lead us to

$$\Psi(x^{k+1}) \leq \Psi(x^k) + \frac{1}{\lambda}D_h(x^k, y^k) - \frac{1}{\lambda}D_h(x^k, x^{k+1}) - \left(\frac{1}{\lambda} - L\right)D_h(x^{k+1}, y^k),$$

which implies inequality (38). If $\beta_k = 0$, then $y^k = x^k$ and $D_h(x^k, y^k) = 0$. If $\beta_k \neq 0$, since we chose the adaptive restart scheme, there is a $\rho \in [0, 1)$ satisfying

$D_h(x^k, y^k) \leq \rho D_h(x^{k-1}, x^k)$. From the definition of $H_M(x^k, x^{k-1})$ and $0 < \lambda L < 1$, we have

$$
\begin{aligned}
H_M(x^{k+1}, x^k) &\leq H_M(x^k, x^{k-1}) + \frac{1}{\lambda} D_h(x^k, y^k) - \left(\frac{1}{\lambda} - M\right) D_h(x^k, x^{k+1}) \\
&\quad - M D_h(x^{k-1}, x^k) - \left(\frac{1}{\lambda} - L\right) D_h(x^{k+1}, y^k) \\
&\leq H_M(x^k, x^{k-1}) - \left(\frac{1}{\lambda} - M\right) D_h(x^k, x^{k+1}) \\
&\quad - \left(M - \frac{\rho}{\lambda}\right) D_h(x^{k-1}, x^k) - \left(\frac{1}{\lambda} - L\right) D_h(x^{k+1}, y^k),
\end{aligned}
\tag{40}
$$

where the second inequality comes from $D_h(x^k, y^k) \leq \rho D_h(x^{k-1}, x^k)$. When $\frac{\rho}{\lambda} \leq M \leq \frac{1}{\lambda}$, we have

$$
H_M(x^{k+1}, x^k) \leq H_M(x^k, x^{k-1}), \quad \forall k \geq 0,
$$

which shows that the sequence $\{H_M\}_{k=0}^{\infty}$ is non-increasing. $\qquad\square$

We can use Lemma 6 to prove Proposition 3.

Proposition 3 *Suppose that Assumptions* 1, 3, *and* 7 *hold. Let* $\{x^k\}_{k=0}^{\infty}$ *be a sequence generated by BPDCAe with* $0 < \lambda L < 1$. *Assume that the auxiliary function* $H_M(x^k, x^{k-1})$ *satisfies* $\frac{\rho}{\lambda} \leq M \leq \frac{1}{\lambda}$ *for* $\rho \in [0, 1)$. *Then, the following statements hold*:

(i) *The sequence* $\{H_M(x^k, x^{k-1})\}_{k=0}^{\infty}$ *is non-increasing.*

(ii) $\sum_{k=1}^{\infty} D_h(x^{k-1}, x^k) < \infty$; *hence, the sequence* $\{D_h(x^{k-1}, x^k)\}_{k=0}^{\infty}$ *converges to zero.*

(iii) $\min_{1 \leq k \leq n} D_h(x^{k-1}, x^k) \leq \frac{\lambda}{n(1-\rho)}\left(\Psi(x^0) - \Psi_*\right)$, *where* $\Psi_* = v(\mathcal{P}) > -\infty$ (*by Assumption* 1 (*v*)).

Proof (i) The statement was proved in Lemma 6.

(ii) Modify (40) into

$$
\begin{aligned}
\lambda(H_M(x^{k+1}, x^k) - H_M(x^k, x^{k-1})) &\leq -(1 - \lambda M) D_h(x^k, x^{k+1}) - (\lambda M - \rho) D_h(x^{k-1}, x^k) \\
&\quad - (1 - \lambda L) D_h(x^{k+1}, y^k) \\
&\leq -(1 - \lambda M) D_h(x^k, x^{k+1}) - (\lambda M - \rho) D_h(x^{k-1}, x^k),
\end{aligned}
$$

where the last inequality comes from $(1 - \lambda L) D_h(x^{k+1}, y^k) \geq 0$. Let n be a positive integer. Summing the above inequality from $k = 0$ to n and letting $\Psi_* = v(\mathcal{P}) > -\infty$, we find that

$$\sum_{k=1}^{n} D_h(x^{k-1}, x^k) = \sum_{k=0}^{n} D_h(x^{k-1}, x^k) \leq \frac{\lambda\left(H_M(x^0, x^{-1}) - H_M(x^{n+1}, x^n)\right)}{1 - \rho}$$

$$\leq \frac{\lambda\left(\Psi(x^0) - \Psi(x^{n+1})\right)}{1 - \rho} \tag{41}$$

$$\leq \frac{\lambda\left(\Psi(x^0) - \Psi_*\right)}{1 - \rho},$$

where the second inequality comes from $D_h(x^{-1}, x^0) = 0$, $x^{-1} = x^0$, and $D_h(x^n, x^{n+1}) \geq 0$. Note that $x^{n+1} \in C$ by Assumption 3. By taking the limit as $n \to \infty$, we arrive at the former statement (ii). The latter statement directly follows from the former.

(iii) From (41), we immediately have

$$n \min_{1 \leq k \leq n} D_h(x^{k-1}, x^k) \leq \sum_{k=1}^{n} D_h(x^{k-1}, x^k) \leq \frac{\lambda\left(\Psi(x^0) - \Psi_*\right)}{1 - \rho}.$$

$\square$

4.4 Convergence analysis of BPDCAe

The proofs of Theorems 6, 7, 8, and Proposition 4 are given in the Appendix. They follow arguments that are similar to their BPDCA counterparts.

Theorem 6 (Global subsequential convergence of BPDCAe) *Suppose that Assumptions* 1, 3, 4, *and* 7 *hold. Let* $\{x^k\}_{k=0}^{\infty}$ *be a sequence generated by BPDCAe with* $0 < \lambda L < 1$ *for solving* $(\mathcal{P})$. *Assume that the auxiliary function* $H_M(x^k, x^{k-1})$ *satisfies* $\frac{\rho}{\lambda} \leq M \leq \frac{1}{\lambda}$ *for* $\rho \in [0, 1)$. *Then, the following statements hold*:

(i)   *The sequence* $\{x^k\}_{k=0}^{\infty}$ *is bounded.*
(ii) $\lim_{k \to \infty} \|x^{k+1} - x^k\| = 0$.
(iii) *Any accumulation point of* $\{x^k\}_{k=0}^{\infty}$ *is a limiting critical point of* $(\mathcal{P})$.

Proposition 4 *Suppose that Assumptions* 1, 3, 4, *and* 7 *hold. Let* $\{x^k\}_{k=0}^{\infty}$ *be a sequence generated by BPDCAe with* $0 < \lambda L < 1$ *for solving* $(\mathcal{P})$ *and* $\frac{\rho}{\lambda} \leq M \leq \frac{1}{\lambda}$ *for* $\rho \in [0, 1)$. *Then, the following statements hold*:

(i) $\zeta := \lim_{k \to \infty} \Psi(x^k)$ *exists.*
(ii) $\Psi \equiv \zeta$ *on* Ω, *where* Ω *is the set of accumulation points of* $\{x^k\}_{k=0}^{\infty}$.

Since $H_M(x, y)$ has a Bregman distance term, the subdifferential of $H_M(x, y)$ has a ∇h term. To prove Theorem 7, we should additionally suppose that there is a bounded subdifferential of the gradient ∇h [29].

Assumption 8 There exists a bounded u such that $u \in \partial(\nabla h)$ on any bounded subset of $\mathbb{R}^d$.

We can prove the following theorems by supposing the KL property or the subanalyticity of the auxiliary function $H_M(x, y)$ in relation to x and y.

Theorem 7 (Global convergence of BPDCAe under the local differentiability of f_2) *Suppose that Assumptions* 1, 3, 4, 5, 7, *and* 8 *hold and that the auxiliary function* $H_M(x, y)$ *is a KL function satisfying* $\frac{\rho}{\lambda} \leq M \leq \frac{1}{\lambda}$ *for* $\rho \in [0, 1)$. *Let* $\{x^k\}_{k=0}^{\infty}$ *be a sequence generated by BPDCAe with* $0 < \lambda L < 1$ *for solving* $(\mathcal{P})$. *Then, the following statements hold*:

  (i) $\lim_{k \to \infty} \operatorname{dist}((0, 0), \partial H_M(x^k, x^{k-1})) = 0$.
  (ii) *The set of accumulation points of* $\{(x^k, x^{k-1})\}_{k=0}^{\infty}$ *is* $Y := \{(x, x) \mid x \in \Omega\}$ *and* $H_M \equiv \zeta$ *on* Y, *where* Ω *is the set of accumulation points of* $\{x^k\}_{k=0}^{\infty}$.
  (iii) *The sequence* $\{x^k\}_{k=0}^{\infty}$ *converges to a limiting stationary point of* $(\mathcal{P})$; *moreover,* $\sum_{k=1}^{\infty} \|x^k - x^{k-1}\| < \infty$.

Theorem 8 (*Global convergence of BPDCAe under the local differentiability of g*) *Suppose that Assumptions* 1, 3, 4, 6, 7, *and* 8 *hold and that the auxiliary function* $H_M(x, y)$ *is subanalytic satisfying* $\frac{\rho}{\lambda} \leq M \leq \frac{1}{\lambda}$ *for* $\rho \in [0, 1)$. *Let* $\{x^k\}_{k=0}^{\infty}$ *be a sequence generated by BPDCAe with* $0 < \lambda L < 1$ *for solving* $(\mathcal{P})$. *Then, the sequence* $\{x^k\}_{k=0}^{\infty}$ *converges to a limiting critical point of* $(\mathcal{P})$; *moreover,* $\sum_{k=1}^{\infty} \|x^k - x^{k-1}\| < \infty$.

Finally, we have theorems regarding the convergence rate of BPDCAe, whose proof is almost identical to Theorems 4 and 5.

Theorem 9 (Rate of convergence under the local differentiability of f_2) *Suppose that Assumptions* 1, 3, 4, 5, 7, *and* 8 *hold. Let* $\{x^k\}_{k=0}^{\infty}$ *be a sequence generated by BPD-CAe with* $0 < \lambda L < 1$ *for solving* $(\mathcal{P})$ *and suppose that* $\{x^k\}_{k=0}^{\infty}$ *converges to some* $\tilde{x} \in \mathcal{X}$. *Suppose further that the auxiliary function* $H_M(x, y)$ *satisfying* $\frac{\rho}{\lambda} \leq M \leq \frac{1}{\lambda}$ *for* $\rho \in [0, 1)$ *is a KL function with* ϕ *in the KL inequality* (4) *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* $k_0 > 0$ *such that* x^k *is constant for* $k > k_0$;
  (ii) *If* $\theta \in (0, \frac{1}{2}]$, *then there exist* $c_1 > 0, k_1 > 0$, *and* $\eta \in (0, 1)$ *such that* $\|x^k - \tilde{x}\| < c_1 \eta^k$ *for* $k > k_1$;
  (iii) *If* $\theta \in (\frac{1}{2}, 1)$, *then there exist* $c_2 > 0$ *and* $k_2 > 0$ *such that* $\|x^k - \tilde{x}\| < c_2 k^{-\frac{1-\theta}{2\theta-1}}$ *for* $k > k_2$.

Theorem 10 (*Rate of convergence under the local differentiability of g*) *Suppose that Assumptions* 1, 3, 4, 6, 7, *and* 8 *hold. Let* $\{x^k\}_{k=0}^{\infty}$ *be a sequence generated by BPD-CAe with* $0 < \lambda L < 1$ *for solving* ($\mathcal{P}$) *and suppose that* $\{x^k\}_{k=0}^{\infty}$ *converges to some* $\tilde{x} \in \mathcal{X}$. *Suppose further that the auxiliary function* $H_M(x, y)$ *satisfying* $\frac{\rho}{\lambda} \le M \le \frac{1}{\lambda}$ *for* $\rho \in [0, 1)$ *is subanalytic. Let* $\theta \in [0, 1)$ *be a Łojasiewicz exponent of* $\tilde{x}$. *Then, the following statements hold*:

(i) *If* $\theta = 0$, *then there exists* $k_0 > 0$ *such that* x^k *is constant for* $k > k_0$;

(ii) *If* $\theta \in (0, \frac{1}{2}]$, *then there exist* $c_1 > 0, k_1 > 0$, *and* $\eta \in (0, 1)$ *such that* $\|x^k - \tilde{x}\| < c_1 \eta^k$ *for* $k > k_1$;

(iii) *If* $\theta \in (\frac{1}{2}, 1)$, *then there exist* $c_2 > 0$ *and* $k_2 > 0$ *such that* $\|x^k - \tilde{x}\| < c_2 k^{-\frac{1-\theta}{2\theta-1}}$ *for* $k > k_2$.

5 Applications

5.1 Application to phase retrieval

In phase retrieval, we are interested in finding a (parameter) vector $x \in \mathbb{R}^d$ that approximately solves the system,

$$\langle a_r, x \rangle^2 \simeq b_r, \quad r = 1, 2, \dots, m, \tag{42}$$

where the vectors $a_r \in \mathbb{R}^d$ describe the model and $b = (b_1, b_2, \dots, b_m)^{\mathrm{T}}$ is a vector of (usually) noisy measurements. As described in [8, 10], the system (42) can be formulated as a nonconvex optimization problem:

$$\min \left\{ \Psi(x) := \frac{1}{4} \sum_{r=1}^{m} \left(\langle a_r, x \rangle^2 - b_r \right)^2 + \theta g(x) \,\middle|\, x \in \mathbb{R}^d \right\}, \tag{43}$$

where $\theta \ge 0$ is a trade-off parameter between the data fidelity criteria and the regularizer g. We define $g : \mathbb{R}^d \to \mathbb{R}$, in particular $g(x) = \|x\|_1$.

In this case, the underlying space of ($\mathcal{P}$) is $C \equiv \mathbb{R}^d$. Define $f : \mathbb{R}^d \to \mathbb{R}$ as $f(x) = \frac{1}{4} \sum_{r=1}^{m} \left(\langle a_r, x \rangle^2 - b_r \right)^2$, which is a nonconvex differentiable function that does not admit a global Lipschitz continuous gradient. The objective function of the phase retrieval problem can be also reformulated as a difference between two convex functions such as in [14]. That is, $f(x) = f_1(x) - f_2(x)$, where

$$f_1(x) = \frac{1}{4} \sum_{r=1}^{m} \langle a_r, x \rangle^4 + \frac{1}{4} \|b\|^2 \quad \text{and} \quad f_2(x) = \frac{1}{2} \sum_{r=1}^{m} b_r \langle a_r, x \rangle^2. \tag{44}$$

When we do not regard the phase retrieval (43) as a DC optimization problem, the Bregman Proximal Gradient algorithm (BPG) can be used instead [8]. Enhancements using the extrapolation technique were proposed: the Bregman Proximal

Gradient algorithm with extrapolation (BPGe) [29] and Convex-Concave Inertial BPG [20] for estimating L. For BPG(e), assuming L-smad for the pair $(f_1 - f_2, h)$ using $h(x) = \frac{1}{4}\|x\|^4 + \frac{1}{2}\|x\|^2$, L satisfies the following inequality [8, Lemma 5.1]:

$$L \geq \sum_{r=1}^{m} \left(3\|a_r a_r^{\mathrm{T}}\|^2 + \|a_r a_r^{\mathrm{T}}\|\|b_r|\right). \tag{45}$$

On the other hand, for DC optimization problems, we define $h : \mathbb{R}^d \to \mathbb{R}$ as

$$h(x) = \frac{1}{4}\|x\|^4. \tag{46}$$

This function is simpler than the original nonconvex formulation. The function $h(x) = \frac{1}{4}\|x\|^4$ is not σ-strongly convex. Therefore, this function does not satisfy Assumption 4 (i).

Proposition 5 *Let f_1 and h be as defined above. Then, for any L satisfying*

$$L \geq 3\left\|\sum_{r=1}^{m} \|a_r\|^2 a_r a_r^{\mathrm{T}}\right\|, \tag{47}$$

the function $Lh - f_1$ is convex on $\mathbb{R}^d$. Therefore, the pair (f_1, h) is L-smad on $\mathbb{R}^d$.

Proof Let $x \in \mathbb{R}^d$. Since f_1 and h are $\mathcal{C}^2$ on $\mathbb{R}^d$, to guarantee the convexity of $Lh - f_1$, it is sufficient to find $L > 0$ such that $L\lambda_{\min}(\nabla^2 h(x)) \geq \lambda_{\max}(\nabla^2 f_1(x))$, where $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ denote the minimal and maximal eigenvalues of a matrix M, respectively. Now, we have the Hessian for f_1 and h:

$$\nabla^2 f_1(x) = 3\sum_{r=1}^{m} \langle a_r, x\rangle^2 a_r a_r^{\mathrm{T}} \quad \text{and} \quad \nabla^2 h(x) = \|x\|^2 I_d + 2xx^{\mathrm{T}}.$$

Since $\nabla^2 h(x) \geq \|x\|^2 I_d$, we obtain $\lambda_{\min}\left(\nabla^2 h(x)\right) \geq \|x\|^2$. From the well-known fact, $\lambda_{\max}(M) \leq \|M\|$, we have the following inequality:

$$\lambda_{\max}\left(\nabla^2 f_1(x)\right) \leq 3\left\|\sum_{r=1}^{m} \langle a_r, x\rangle^2 a_r a_r^{\mathrm{T}}\right\| \leq 3\left\|\sum_{r=1}^{m} \|a_r\|^2 a_r a_r^{\mathrm{T}}\right\|\|x\|^2 \leq L\|x\|^2 \leq L\lambda_{\min}(\nabla^2 h(x)).$$

Therefore, we obtain the desired result. $\qquad\square$

Comparing the right hand side of (45) and that of (47), we can see that

$$3\left\|\sum_{r=1}^{m} \|a_r\|^2 a_r a_r^{\mathrm{T}}\right\| \leq \sum_{r=1}^{m} \left(3\|a_r a_r^{\mathrm{T}}\|^2 + \|a_r a_r^{\mathrm{T}}\|\|b_r|\right). \tag{48}$$

The constant L has the important role of defining the step size, and thereby affects the performance of the algorithms. Note that even if $\left\|\sum_{r=1}^{m} \|a_r\|^2 a_r a_r^{\mathrm{T}}\right\| = \sum_{r=1}^{m} \|a_r a_r^{\mathrm{T}}\|^2$, the left-hand side of (48) is always smaller than the right-hand side because

$\sum_{r=1}^{m} \|a_r a_r^T\| |b_r| \geq 0$. When $h(x) = \frac{1}{4}\|x\|^4 + \frac{1}{2}\|x\|^2$, the subproblems of BPG(e) have a closed-form solution formula [8, Proposition 5.1]. When $h(x) = \frac{1}{4}\|x\|^4$, subproblems (5) and (7) also have a closed-form solution formula, which is obtained by slightly modifications of those in BPG(e).

In this application, the functions f_1, f_2, g, and h satisfy Assumptions from 1 to 8 excepting Assumption 4 (i) and 6. In particular, Assumption 4 (i) is not satisfied for our choice $h(x) = \frac{1}{4}\|x\|^4$, but it is satisfied if we replace it by $h(x) = \frac{1}{4}\|x\|^4 + \frac{1}{2}\|x\|^2$. Finally, Ψ and H_M are KL functions due to their semi-algebraicity [1]. Therefore, in this application, Assumption 6 is not required for the global convergence of BPDCAe.

5.2 Lower bound on the *L*-smooth adaptable parameter in the Gaussian model

We dealt with the following Gaussian model. We generated the elements of m vectors $a_r \in \mathbb{R}^d$ and the ground truth $\tilde{x} \in \mathbb{R}^d$, which was a sparse vector (sparsity of 5%), independently from the standard Gaussian distribution. Then, we generated $b_r = \langle a_r, \tilde{x} \rangle^2, r = 1, 2, \ldots, m$ from a_r and $\tilde{x}$.

From the linearity of the expectation, we consider the expectation of $\nabla^2 f_1$,

$$\mathbb{E}\left[\nabla^2 f_1(x)\right] = 3 \sum_{r=1}^{m} \mathbb{E}\left[\langle a_r, x \rangle^2 a_r a_r^T\right].$$

Since the elements of a_r are independently generated from the standard Gaussian distribution, the j-th diagonal element of the above matrix is given by

$$\mathbb{E}\left[\langle a_r, x \rangle^2 a_{r,j}^2\right] = \mathbb{E}\left[a_{r,j}^4 x_j^2 + \sum_{k=1,k\neq j}^{d} a_{r,j}^2 a_{r,k}^2 x_k^2\right] = 3x_j^2 + \sum_{k=1,k\neq j}^{d} x_k^2 = 2x_j^2 + \|x\|^2.$$

The non-diagonal (j, k) elements are

$$\mathbb{E}\left[\langle a_r, x \rangle^2 a_{r,j} a_{r,k}\right] = \mathbb{E}\left[2a_{r,j}^2 a_{r,k}^2 x_j x_k\right] = 2x_j x_k.$$

Moreover, noting that $h(x) = \frac{1}{4}\|x\|^4$, we obtain $\mathbb{E}\left[\langle a_r, x \rangle^2 a_r a_r^T\right] = \|x\|^2 I_d + 2xx^T = \nabla^2 h(x)$. The expectation of the Hessian of $f_1(x)$ is thus given by $\mathbb{E}[\nabla^2 f_1(x)] = 3m\nabla^2 h(x)$.

Under the Gaussian model, we can reduce the lower bound of L given in Proposition 5 with high probability by applying [10, Lemma 7.4] as shown in the following proposition.

Proposition 6 *Let the functions f_1 and h be given by* (44) *and* (46), *respectively. Moreover, assume that the vectors a_r are independently distributed according to the Gaussian model with a sufficiently large number of measurements. Let γ and δ be a fixed positive numerical constant and $c(\cdot)$ be a sufficiently large numerical constant*

that depends on δ; this means that the number of samples obeys $m \geq c(\delta) \cdot d \log d$ in the Gaussian model. Then, for any L satisfying

$$L \geq 9 \left\| \sum_{r=1}^{m} a_r a_r^{\mathrm{T}} \right\| + \delta, \tag{49}$$

the function $Lh - f_1$ is convex on $\mathbb{R}^d$ and hence the pair (f_1, h) is L-smad on $\mathbb{R}^d$ with probability at least $1 - 5e^{-\gamma d} - 4/d^2$.

Proof Consider the expectation of $\sum_{r=1}^{m} a_r a_r^{\mathrm{T}}$. Since the elements of a_r are independently generated from the standard Gaussian distribution, for any $y \in \mathbb{R}^d$, we have

$$y^{\mathrm{T}} \mathbb{E} \left[\sum_{r=1}^{m} a_r a_r^{\mathrm{T}} \right] y = \sum_{r=1}^{m} \mathbb{E} \left[\langle a_r, y \rangle^2 \right] = \sum_{r=1}^{m} \sum_{j=1}^{d} y_j^2 = \sum_{r=1}^{m} \|y\|^2. \tag{50}$$

From (50), for any $y \in \mathbb{R}^d$, we have

$$y^{\mathrm{T}} \mathbb{E}[\nabla^2 f_1(x)] y = 3 \sum_{r=1}^{m} \left(\|x\|^2 \|y\|^2 + 2\langle x, y \rangle^2 \right) \leq 9\|x\|^2 \sum_{r=1}^{m} \|y\|^2 = 9\|x\|^2 y^{\mathrm{T}} \mathbb{E} \left[\sum_{r=1}^{m} a_r a_r^{\mathrm{T}} \right] y. \tag{51}$$

We can easily find that

$$9 \sum_{r=1}^{m} a_r a_r^{\mathrm{T}} \preceq 9 \left\| \sum_{r=1}^{m} a_r a_r^{\mathrm{T}} \right\| I_d,$$

which implies that

$$9 \mathbb{E} \left[\sum_{r=1}^{m} a_r a_r^{\mathrm{T}} \right] \preceq 9 \left\| \sum_{r=1}^{m} a_r a_r^{\mathrm{T}} \right\| I_d. \tag{52}$$

From (51) and (52), we have

$$\mathbb{E}[\nabla^2 f_1(x)] \preceq 9\|x\|^2 \left\| \sum_{r=1}^{m} a_r a_r^{\mathrm{T}} \right\| I_d. \tag{53}$$

From [10, Lemma 7.4], (49), and (53), we conclude that

$$\nabla^2 f_1(x) \preceq \mathbb{E}[\nabla^2 f_1(x)] + \delta\|x\|^2 I_d \preceq L\|x\|^2 I_d \tag{54}$$

with probability at least $1 - 5e^{-\gamma d} - 4/d^2$. From $\nabla^2 h(x) \succeq \|x\|^2 I_d$ and (54), we have $\nabla^2 f_1(x) \preceq L\nabla^2 h(x)$, which proves that $Lh - f_1$ is convex with probability at least $1 - 5e^{-\gamma d} - 4/d^2$. Therefore, the pair (f_1, h) is L-smad on $\mathbb{R}^d$. □

Remark 1 Since each element of a_r independently follows the standard Gaussian distribution, $\|a_r\|^2$ follows the chi-squared distribution with d degrees of freedom.

Thus, we can show $\|a_r\|^2 \geq 3$ with high probability for sufficiently large d. It implies that the bound given in Proposition 6 is smaller than that given in Proposition 5.

5.3 Performance results for phase retrieval with the Gaussian model

Here, we summarize the results for the Gaussian model. All numerical experiments were performed in Python 3.7 on an iMac with a 3.3 GHz Intel Core i5 Processor and 8 GB 1867 MHz DDR3 memory.

First, let us examine the results for Bregman proximal-type algorithms, *i.e.*, BPG [8], BPGe [29], BPDCA (Algorithm 1), and BPDCAe (Algorithm 2). We compared the averages of 100 random instances in terms of the number of iterations, CPU time, and accuracy (Tables 1 and 2). Let $\hat{x}$ be a recovered solution and $\tilde{x}$ be the ground truth generated according to the method described in Sect. 5.2. In order to compare the objective function values, we took the difference $\log_{10} |\Psi(\hat{x}) - \Psi(\tilde{x})|$ to be the accuracy. In the numerical experiments, $\Psi(\hat{x}) > \Psi(\tilde{x})$. The termination criterion was defined as $\|x^k - x^{k-1}\| / \max\{1, \|x^k\|\} \leq 10^{-6}$. The equation numbers under each algorithm in Tables 1 and 2 indicate the value of λ; that is, we set $\lambda = 1/L$ for L satisfying the equations. For the restart schemes, we used the adaptive restart scheme with $\rho = 0.99$ and the fixed restart scheme with $K = 200$. We set $\theta = 1$ for the regularizer g in (43). We forcibly stopped the algorithms when they reached the maximum number of iterations (50,000). Table 2 compares the results of BPGe and BPDCAe under the same settings as the results in Table 1. BPDCA with (49) was the fastest among the algorithms without extrapolation (Table 1). On the other hand, the extrapolation method makes each algorithm faster (Table 2).

We can conclude that, at least for phase retrieval, BPDCA has a clear advantage over BPG because of its reformulation as a nonconvex DC optimization problem (44), which permits choosing a smaller L in (47) instead of (45). In particular, for the Gaussian model, we can use a smaller L in (49) with high probability. The extrapolation technique can further enhance performance. Also, we can see that the iterates of BPDCA(e) globally converge to their optimal solutions despite that the kernel generating distance h (46) does not satisfy Assumption 4 (i). This suggests that this condition may be relaxed in some cases.

Next, we compared the empirical probability of success for BPDCAe and Wirtinger flow [10], which is a well-known algorithm for phase retrieval. Here we took x_0 in BPDCAe to be the value calculated in the initialization step of the Wirtinger flow. The empirical probability of success in Fig. 1 is an average over 100 trials. We regard that the algorithms succeeded if the relative error $\|\hat{x} - \tilde{x}\| / \|\tilde{x}\|$ falls below 10^{-5} after 2,500 iterations. The dimension d was fixed at 128, and we varied the number of measurements m. We used the adaptive restart scheme with $\rho = 0.99$ and the fixed restart scheme with $K = 200$. We set $\theta = 0$; *i.e.*, we solved (43) without its regularizer. From the figure, we can see that BPDCAe with the initialization step of the Wirtinger flow achieved almost 100% success rate when $m/d \geq 6$ and obtained more stable results than those of Wirtinger flow.

Table 1 Average number of iterations, CPU time, and accuracy for BPG [8] and BPDCA using 100 random instances of phase retrieval (over the Gaussian model) for different values of L

Algorithm	m	d	Iteration	CPU-Time (s)	Accuracy
BPG [8] (45)	10,000	10	3757	1.638	2.901
		50	50,000	37.761	1.977
		100	50,000	46.920	5.312
		200	50,000	91.925	7.737
	20,000	10	3689	2.539	−2.569
		50	50,000	76.020	2.007
		100	50,000	121.966	5.523
		200	50,000	191.780	8.057
	30,000	10	3764	3.698	−2.387
		50	50,000	104.947	2.257
		100	50,000	175.143	5.678
		200	50,000	287.735	8.227
BPDCA (47)	10,000	10	265	0.102	−4.374
		50	1415	0.520	−3.212
		100	3274	2.129	−2.656
		200	8111	10.416	−2.061
	20,000	10	255	0.157	−4.350
		50	1299	1.182	−3.193
		100	2833	4.283	−2.642
		200	6572	18.198	−2.057
	30,000	10	256	0.233	−4.335
		50	1257	1.790	−3.156
		100	2696	6.484	−2.596
		200	6012	25.666	−2.010
BPDCA (49)	10,000	10	68	0.025	−5.127
		50	92	0.034	−4.627
		100	115	0.075	−4.380
		200	152	0.192	−4.108
	20,000	10	65	0.040	−5.137
		50	84	0.077	−4.691
		100	98	0.149	−4.476
		200	121	0.335	−4.229
	30,000	10	65	0.059	−5.166
		50	81	0.115	−4.728
		100	93	0.223	−4.515
		200	110	0.465	−4.285

6 Conclusions

We proposed two Bregman proximal-type algorithms for solving DC optimization problems $(\mathcal{P})$. One is the Bregman Proximal DC Algorithm (BPDCA), the other is BPDCA with extrapolation (BPDCAe). Proximal-type algorithms including ours

Table 2 Average number of iterations, CPU time, and accuracy for BPGe [29] and BPDCAe using 100 random instances of phase retrieval (over the Gaussian model) for different values of L

Algorithm	m	d	Iteration	CPU-Time (s)	Accuracy
BPGe [29] (45)	10,000	10	297	0.124	−3.904
		50	2614	1.209	−0.428
		100	6214	5.949	0.974
		200	23,940	44.218	2.426
	20,000	10	285	0.198	−3.653
		50	1941	2.871	−0.375
		100	6054	15.376	1.250
		200	21,138	82.086	2.734
	30,000	10	294	0.290	−3.362
		50	1880	3.826	−0.199
		100	6002	21.271	1.411
		200	21,434	123.504	2.806
BPDCAe (47)	10,000	10	67	0.025	−5.205
		50	203	0.075	−3.802
		100	332	0.218	−3.451
		200	581	0.740	−2.941
	20,000	10	62	0.038	−5.071
		50	179	0.165	−4.152
		100	302	0.458	−3.694
		200	501	1.394	−3.110
	30,000	10	59	0.054	−4.852
		50	169	0.242	−4.054
		100	278	0.670	−3.448
		200	446	1.891	−2.987
BPDCAe (49)	10,000	10	32	0.013	−5.649
		50	42	0.015	−5.371
		100	49	0.032	−5.087
		200	61	0.078	−5.135
	20,000	10	29	0.018	−5.550
		50	38	0.035	−5.317
		100	43	0.065	−4.919
		200	52	0.144	−5.051
	30,000	10	29	0.026	−5.558
		50	38	0.056	−5.446
		100	41	0.098	−4.908
		200	50	0.210	−5.115

are effective on large-scale problems. In addition, our algorithms assume that the function f_1 has the L-smooth adaptable property in relation to the kernel generating distance h, instead of L-smoothness. The restart condition for our adaptive restart scheme is different from the existing ones.
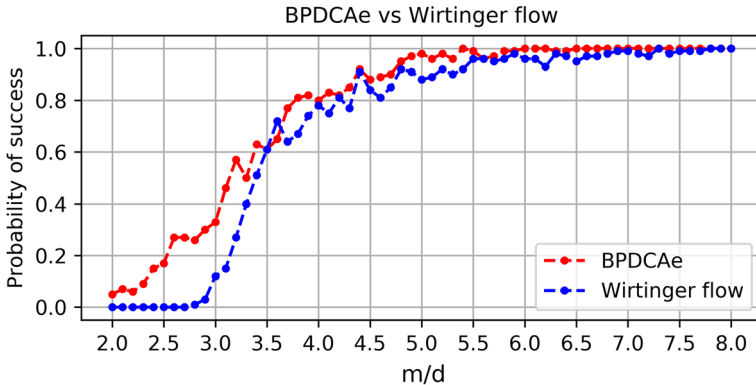
Fig. 1 Empirical probability of success based on 100 trials for BPDCAe and Wirtinger flow [10] using the same initialization step (of the Wirtinger flow). We set $d = 128$ and varied the number m of measurements

We conducted convergence analyses of our algorithms. Assuming the Kurdyka-Łojasiewicz property or subanalyticity of the objective function together with some standard assumptions, we established that the iterates generated by BPDCA(e) globally converge to a limiting stationary point or a limiting critical point and derived their convergence rates.

We applied our algorithms to phase retrieval. The numerical experiments demonstrated that BPDCAe is faster than the other Bregman-type algorithms. For the Gaussian model, BPDCAe offered more stable results than Wirtinger flow [10]. We conclude that BPDCAe is a powerful method for solving large-scale and structured DC optimization problems. Although the kernel generating distance h (46) does not satisfy Assumption 4 (i), the sequences generated by BPDCA(e) converged in the numerical experiments. Therefore, we conjecture that most of the convergent results can be demonstrated under weaker conditions. As future work, since g in BPDCA does not need to be convex, we will attempt to prove the monotonicity of the auxiliary function of BPD-CAe (Lemma 6) without assuming Assumption 7.

Other Bregman proximal-type algorithms have been proposed. Mukkamala *et al.*[20] chose the L-smad parameters by using a line search. As this parameter is generally difficult to estimate accurately, we can utilize this line search in our algorithms.

For constrained problems, Wang *et al.*[27] proposed the Bregman alternating direction methods with multipliers. Tu *et al.*[25] also developed a Bregman-type algorithm for solving linearly constrained DC optimization problems. These variational methods may inspire further improvements and extensions.

Appendix: proof of convergence theorems for BPDCAe

Proof of Theorem 6

(i) Since $H_M(x^k, x^{k-1}) \leq H_M(x^0, x^{-1})$ for all $k \in \mathbb{N}$ from Proposition 3 (i), with $x^0 = x^{-1}$, we obtain

$$\Psi(x^k) \leq \Psi(x^k) + M D_h(x^{k-1}, x^k) = H_M(x^k, x^{k-1}) \leq H_M(x^0, x^{-1}) = \Psi(x^0),$$

which shows that $\{x^k\}_{k=0}^{\infty}$ is bounded due to Assumption 4 (iii).

(ii) From (39), we obtain

$$H_M(x^k, x^{k-1}) - H_M(x^{k+1}, x^k) \geq \left(\frac{1}{\lambda} - M\right) D_h(x^k, x^{k+1}) + \left(M - \frac{\rho}{\lambda}\right) D_h(x^{k-1}, x^k)$$
$$+ \left(\frac{1}{\lambda} - L\right) D_h(x^{k+1}, y^k)$$
$$\geq \frac{\sigma(1 - \lambda L)}{2\lambda} \left(\|x^{k+1} - x^k\|^2 - \beta_k \|x^k - x^{k-1}\|^2\right),$$

where the last inequality holds because h is a σ-strongly convex function and the first two terms are nonnegative. Summing the above inequality from $k = 0$ to ∞, we obtain

$$\frac{\sigma(1 - \lambda L)}{2\lambda} \left(\sum_{k=0}^{\infty} (1 - \beta_{k+1}) \|x^{k+1} - x^k\|^2 - \beta_1 \|x^0 - x^1\|^2 \right)$$
$$\leq H_M(x^0, x^{-1}) - \liminf_{n \to \infty} H_M(x^{n+1}, x^n)$$
$$= \Psi(x^0) - \liminf_{n \to \infty} \left(\Psi(x^{n+1}) + M D_h(x^n, x^{n+1}) \right)$$
$$\leq \Psi(x^0) - v(\mathcal{P}) < \infty,$$

which shows that $\lim_{k \to \infty} \|x^{k+1} - x^k\| = 0$ due to $\frac{1}{\lambda} - L > 0$ and $\sup_{k>0} \beta_k < 1$.

(iii) Let $\tilde{x}$ be an accumulation point of $\{x^k\}_{k=0}^{\infty}$ and let $\{x^{k_j}\}$ be a subsequence such that $\lim_{j \to \infty} x^{k_j} = \tilde{x}$. Then, from the first-order optimality condition of subproblem (7) under Assumption 3, we have

$$0 \in \partial_c g(x^{k_j+1}) + \nabla f_1(y^{k_j}) - \xi^{k_j} + \frac{1}{\lambda} \left(\nabla h(x^{k_j+1}) - \nabla h(y^{k_j}) \right).$$

Therefore, we obtain

$$\xi^{k_j} + \nabla f_1(x^{k_j+1}) - \nabla f_1(y^{k_j}) + \frac{1}{\lambda} \left(\nabla h(y^{k_j}) - \nabla h(x^{k_j+1}) \right) \in \partial_c g(x^{k_j+1}) + \nabla f_1(x^{k_j+1}).$$
$$(55)$$

From the boundedness of $\{x^{k_j}\}$ and the Lipschitz continuity of ∇h and ∇f_1 on a bounded subset of $\mathbb{R}^d$, there exists $A_0 > 0$ such that

$$\left\| \nabla f_1(x^{k_j+1}) - \nabla f_1(y^{k_j}) + \frac{1}{\lambda}\left(\nabla h(y^{k_j}) - \nabla h(x^{k_j+1})\right)\right\| \leq A_0 \|x^{k_j+1} - y^{k_j}\|.$$

Therefore, using $\|x^{k_j+1} - x^{k_j}\| \to 0$ and $\|x^{k_j} - x^{k_j-1}\| \to 0$, we obtain

$$\nabla f_1(x^{k_j+1}) - \nabla f_1(y^{k_j}) + \frac{1}{\lambda}\left(\nabla h(y^{k_j}) - \nabla h(x^{k_j+1})\right) \to 0. \tag{56}$$

Note that the sequence $\{\xi^{k_j}\}$ is bounded as shown in Theorem 1 (ii), and the sequence $\{x^{k_j}\}$ is bounded and converges to $\tilde{x}$. Thus, by taking the limit as $j \to \infty$ or more precisely, its subsequence, we can assume without loss of generality that $\lim_{j\to\infty} \xi^{k_j} =: \tilde{\xi}$ exists, which belongs to $\partial_c f_2(\tilde{x})$ since f_2 is continuous. Using this and (56), we take the limit of (55). Invoking $\|x^{k_j+1} - x^{k_j}\| \to 0$ and the continuity of g and ∇f_1, we obtain $\tilde{\xi} \in \partial_c g(\tilde{x}) + \nabla f_1(\tilde{x})$. Therefore, $0 \in \partial_c g(\tilde{x}) + \nabla f_1(\tilde{x}) - \partial_c f_2(\tilde{x})$, which shows that $\tilde{x}$ is a limiting critical point of $(\mathcal{P})$. $\qquad\square$

Proof of Proposition 4

(i) From Assumption 1 (v) and Proposition 3 (i), the sequence $\{H_M(x^k, x^{k-1})\}_{k=0}^{\infty}$ is bounded from below and non-increasing. Consequently, using $\lim_{k\to\infty} D_h(x^{k-1}, x^k) = 0$ from Proposition 3 (ii), we obtain $\lim_{k\to\infty} H_M(x^k, x^{k-1}) = \lim_{k\to\infty} \Psi(x^k) =: \zeta$.

(ii) Take any $\hat{x} \in \Omega$, that is $\lim_{j\to\infty} x^{k_j} = \hat{x}$. From (7), it follows that

$$g(x^k) + \langle \nabla f_1(y^{k-1}) - \xi^{k-1}, x^k - y^{k-1}\rangle + \frac{1}{\lambda}D_h(x^k, y^{k-1})$$

$$\leq g(\hat{x}) + \langle \nabla f_1(y^{k-1}) - \xi^{k-1}, \hat{x} - y^{k-1}\rangle + \frac{1}{\lambda}D_h(\hat{x}, y^{k-1}).$$

From the above inequality and the fact that f_1 is convex at x^k, we obtain

$$g(x^k) + f_1(x^k) \leq g(\hat{x}) + \langle \nabla f_1(y^{k-1}) - \xi^{k-1}, \hat{x} - x^k\rangle + \frac{1}{\lambda}D_h(\hat{x}, y^{k-1}) - \frac{1}{\lambda}D_h(x^k, y^{k-1})$$

$$+ f_1(\hat{x}) + \langle \nabla f_1(x^k), x^k - \hat{x}\rangle$$

$$\leq g(\hat{x}) + \langle \nabla f_1(y^{k-1}) - \xi^{k-1}, \hat{x} - x^k\rangle + \frac{1}{\lambda}D_h(\hat{x}, y^{k-1}) + \frac{1}{\lambda}D_h(y^{k-1}, \hat{x})$$

$$+ f_1(\hat{x}) + \langle \nabla f_1(x^k), x^k - \hat{x}\rangle, \tag{57}$$

where the second inequality comes from $-\frac{1}{\lambda}D_h(x^k, y^{k-1}) \leq 0$ and $\frac{1}{\lambda}D_h(y^{k-1}, \hat{x}) \geq 0$. Since ∇h is continuous, we have

$$\lim_{j\to\infty}\left(D_h(\hat{x}, y^{k_j-1}) + D_h(y^{k_j-1}, \hat{x})\right) \leq \lim_{j\to\infty}\|\nabla h(y^{k_j-1}) - \nabla h(\hat{x})\|\|y^{k_j-1} - \hat{x}\| = 0.$$

Substituting k_j for k in (57) and limiting j to ∞, we have, from Proposition 3 (ii),

$$\limsup_{j\to\infty}\left(g(x^{k_j}) + f_1(x^{k_j})\right) \leq g(\hat{x}) + f_1(\hat{x}),$$

which provides $\limsup_{j\to\infty} \Psi(x^{k_j}) \le \Psi(\hat{x})$ from the continuity of $-f_2$. Combining this and the lower semicontinuity of Ψ yields $\Psi(x^{k_j}) \to \Psi(\hat{x}) =: \zeta$ as $j \to \infty$. Since $\hat{x} \in \Omega$ is arbitrary, we conclude that $\Psi \equiv \zeta$ on Ω.   □

Proof of Theorem 7

(i) Let $\mu > 0, k_0 > 0, \mathcal{N}_0$, and $\mathcal{N} := \{x \in \mathcal{N}_0 \mid \text{dist}\,(x, \Omega) < \mu\}$ as defined in the proof of Theorem 2 (i).

We begin by considering the subdifferential of H_M at x^k for $k \ge k_0 + 1$, and obtain

$$\partial H_M(x^k, x^{k-1}) = \nabla f_1(x^k) - \nabla f_2(x^k) + \partial_c g(x^k) - M\partial(\nabla h(x^k))(x^{k-1} - x^k). \quad (58)$$

Moreover, considering the first-order optimality condition of subproblem (7), for any $k \ge k_0 + 1$, we have

$$\frac{1}{\lambda}\left(\nabla h(y^{k-1}) - \nabla h(x^k)\right) - \nabla f_1(y^{k-1}) + \nabla f_2(x^{k-1}) \in \partial_c g(x^k),$$

since f_2 is $\mathcal{C}^1$ on $\mathcal{N}$ and $x^{k-1} \in \mathcal{N}$ whenever $k \ge k_0 + 1$. Using the above relation and (58), for a bounded $U^k \in \partial(\nabla h(x^k))$ which exists by Assumption 8, we also obtain

$$\frac{1}{\lambda}\left(\nabla h(y^{k-1}) - \nabla h(x^k)\right) + \nabla f_1(x^k) - \nabla f_1(y^{k-1})$$
$$+ \nabla f_2(x^{k-1}) - \nabla f_2(x^k) + MU^k(x^k - x^{k-1}) \in \partial H_M(x^k, x^{k-1}).$$

Due to the global Lipschitz continuity of $\nabla f_1, \nabla f_2$, and ∇h on $\mathcal{N}_0$, and the boundedness of U^k from Assumption 8, we see that there exist $A_0 > 0$, $A_1 > 0$, and $A_2 > 0$ such that

$$\text{dist}\,((0, 0), \partial H_M(x^k, x^{k-1})) \le A_0\|x^k - y^{k-1}\| + A_1\|x^k - x^{k-1}\|$$
$$\le A_2\left(\|x^k - x^{k-1}\| + \|x^{k-1} - x^{k-2}\|\right),$$

where $k \ge k_0 + 1$. Since $\|x^k - x^{k-1}\| \to 0$ and $\|x^{k-1} - x^{k-2}\| \to 0$, we conclude the claim (i).

(ii) Suppose that $\hat{x} \in \Omega$, $x^{k_j} \to \hat{x}$, and $x^{k_j-1} \to \hat{x}$ as in Proposition 4 (ii). Therefore, the set of accumulation points of $\{(x^k, x^{k-1})\}_{k=0}^{\infty}$ is Y. From Propositions 3 and 4,

$$\lim_{k\to\infty} H_M(x^k, x^{k-1}) = \lim_{k\to\infty} \Psi(x^k) + M\lim_{k\to\infty} D_h(x^{k-1}, x^k) = \zeta.$$

Additionally, from Proposition 4 (ii), for any $(\hat{x}, \hat{x}) \in Y, \hat{x} \in \Omega$, we have $H_M(\hat{x}, \hat{x}) = \Psi(\hat{x}) = \zeta$. Since $\hat{x}$ is arbitrary, we conclude that $H_M \equiv \zeta$ on Y.

(iii) The proof is similar to Theorem 2 (ii).   □

Proof of Theorem 8

Let k_1, κ_i, ν_i, and θ_i be defined similarly to the proof of Theorem 3. Using the differentiability of g and [6, Theorem 3.1], we have

$$\left\| \nabla g(x^k) - \nabla g(x^{k+1}) \right\| \leq \kappa \| x^k - x^{k+1} \|, \tag{59}$$

$$|H_M(x^k, x^{k-1}) - \zeta|^\theta \leq \nu \| \hat{x}^k \|, \quad \hat{x}^k \in \partial(-H)(x^k, x^{k-1}), \quad \forall k \geq k_1 + 1, \tag{60}$$

where $\zeta = H_M(\tilde{x}, \tilde{x}) = \Psi(\tilde{x}), \tilde{x} \in \Omega$, $\kappa = \max_{j=1,\ldots,p} \kappa_i$, $\nu = \max_{j=1,\ldots,p} \nu_i$, and $\theta = \max_{j=1, \ldots, p} \theta_i$. From (7), we obtain

$$0 = \nabla g(x^{k+1}) + \nabla f_1(y^k) - \xi^k + \frac{1}{\lambda} \left(\nabla h(x^{k+1}) - \nabla h(y^k) \right),$$

which implies

$$\nabla g(x^{k+1}) - \nabla g(x^k) + \nabla f_1(y^k) - \nabla f_1(x^k) + \frac{1}{\lambda} \left(\nabla h(x^{k+1}) - \nabla h(y^k) \right) + MU^k(x^{k-1} - x^k)$$
$$= \xi^k + MU^k(x^{k-1} - x^k) - \nabla f_1(x^k) - \nabla g(x^k) \in \partial(-H_M)(x^k, x^{k-1}),$$

for some bounded $U^k \in \partial(\nabla h(x^k))$ and $\partial(-H_M)(x^k, x^{k-1}) = \partial_c f_2(x^k) + M\partial(\nabla h(x^k))$ $(x^{k-1} - x^k) - \nabla f_1(x^k) - \nabla g(x^k)$. Using (59), (60), Assumption 4, and the boundedness of $\partial(\nabla h(x^k))$ from Assumption 8, we obtain $C > 0$ such that

$$|H_M(x^k, x^{k-1}) - \zeta|^\theta$$
$$\leq \nu \left\| \nabla g(x^{k+1}) - \nabla g(x^k) + \nabla f_1(y^k) - \nabla f_1(x^k) + \frac{1}{\lambda} \left(\nabla h(x^{k+1}) - \nabla h(y^k) \right) + MU^k(x^k - x^{k-1}) \right\|$$
$$\leq C(\| x^k - x^{k+1} \| + \| x^{k-1} - x^k \|), \quad \forall k \geq k_1 + 1,$$

where the second inequality comes from $\nabla h(x^{k+1}) - \nabla h(y^k) = \nabla h(x^{k+1}) - \nabla h(x^k)$ $+\nabla h(x^k) - \nabla h(y^k)$. The rest of the proof is similar to Theorem 3 $\qquad \square$

Data Availibility Statement The datasets generated during and/or analysed during the current study are available in the Github repository, https://github.com/ShotaTakahashi/bregman-proximal-dc-algorithm.

References

1. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Program. **116**(1), 5–16 (2009)
2. Bauschke, H.H., Bolte, J., Teboulle, M.: A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. Math. Op. Res. **42**(2), 330–348 (2017)
3. Beck, A.: First-Order Methods in Optimization, Volume 25 of MOS-SIAM Series on Optimization. SIAM (2017)
4. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imag. Sci. **2**(1), 183–202 (2009)
5. Bierstone, E., Milman, P.D.: Semialgebraic and subanalytic sets. *Publications mathématiques de l'I.H.É.S.*, 67:5–42, (1988)
6. Bolte, J., Daniilidis, A., Lewis, A.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical system. SIAM J. Optim. **17**(4), 1205–1223 (2007)
7. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. **146**(1), 459–494 (2014)
8. Bolte, J., Sabach, S., Teboulle, M., Vaisbourd, Y.: First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. SIAM J. Optim. **28**(3), 2131–2151 (2018)
9. Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Math. Phys. **7**(3), 200–217 (1967)
10. Candès, E.J., Li, X., Soltanolkotabi, M.: Phase retrieval via Wirtinger flow: theory and algorithms. IEEE Trans. Inf. Theory **61**(4), 1985–2007 (2015)
11. Cui, Y., Pang, J.-S.: Modern Nonconvex Nondifferentiable Optimization, volume 29 of *MOS-SIAM Series on Optimization*. SIAM, (2021)
12. Dhillon, I., Tropp, J.: Matrix nearness problems with Bregman divergences. SIAM J. Matrix Anal. Appl. **29**(4), 1120–1146 (2008)
13. Horst, R., Thoai, N.V.: DC programming: overview. J. Optim. Theory Appl. **103**(1), 1–43 (1999)
14. Huang, M., Lai, M.-J., Varghese, A., Xu, Z.: On DC based methods for phase retrieval. In: Approximation Theory XVI, pp. 87–121 (2019)
15. Kurdyka, K.: On gradients of functions definable in o-minimal structures. Annales de l'Institut Fourier **48**(3), 769–783 (1998)
16. Le Thi, H.A., Huynh, V.N., Tao, P.D.: Convergence analysis of difference-of-convex algorithm with subanalytic data. J. Optim. Theory Appl. **179**(1), 103–126 (2018)
17. Le Thi, H.A., Tao, P.D.: DC programming and DCA: thirty years of developments. Math. Program. **169**(1), 5–68 (2018)
18. Mordukhovich, B.S.: Variational Analysis and Generalized Differentiation I: Basic Theory. Springer (2006)
19. Mordukhovich, B.S., Nam, N.M., Yen, N.D.: Fréchet subdifferential calculus and optimality conditions in nondifferentiable programming. Optimization **55**(5–6), 685–708 (2006)
20. Mukkamala, M.C., Ochs, P., Pock, T., Sabach, S.: Convex-concave backtracking for inertial Bregman proximal gradient algorithms in nonconvex optimization. SIAM J. Math. Data Sci. **2**(3), 658–682 (2020)
21. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Math. Doklady **27**, 372–376 (1983)
22. Nesterov, Y.: Lectures on Convex Optimization. Springer Optimization and its Applications, 2nd edn. Springer (2018)
23. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis Volume of 372, Grundlehren der Mathematischen Wissenschaften. Springer (1998)
24. Shechtman, Y., Eldar, Y.C., Cohen, O., Chapman, H.N., Miao, J., Segev, M.: Phase retrieval with application to optical imaging: a contemporary overview. IEEE Signal Process Magazine **32**(3), 87–109 (2015)
25. Tu, K., Zhang, H., Gao, H., Feng, J.: A hybrid Bregman alternating direction method of multipliers for the linearly constrained difference-of-convex problems. J. Global Optim. **76**(4), 665–693 (2020)

26. Tuy, H.: D. C. optimization: theory, methods and algorithms. In: Handbook of Global Optimization, vol. 25, pp. 149–216 (1995)
27. Wang, H., Banerjee, A.: Bregman alternating direction method of multipliers. Adv. Neural Inf. Process. Syst. **84**, 2816–2824 (2014)
28. Wen, B., Chen, X., Pong, T.K.: A proximal difference-of-convex algorithm with extrapolation. Comput. Optim. Appl. **69**(2), 297–324 (2018)
29. Zhang, X., Barrio, R., Martinez, M.A., Jiang, H., Cheng, L.: Bregman proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. IEEE Access **7**, 126515–126529 (2019)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Shota Takahashi[1] · **Mituhiro Fukuda**[2,3,4,5] · **Mirai Tanaka**[1,6,7]

Mituhiro Fukuda
mituhiro@is.titech.ac.jp

Mirai Tanaka
mirai@ism.ac.jp

[1] Department of Statistical Science, The Graduate University for Advanced Studies, 10–3 Midori-cho, Tachikawa, Tokyo 190–8562, Japan

[2] Department of Computer Science, Institute of Mathematics and Statistics, University of São Paulo, Rua do Matão, 1010, Cidade Universitária, São Paulo, SP 05508–090, Brazil

[3] Department of Mathematical and Computing Science, Tokyo Institute of Technology, 2–12–1 Oh-okayama, Meguro-ku, Tokyo 152–8552, Japan

[4] São Paulo State Technological College of Praia Grande, Praça 19 de Janeiro, 144, Boqueirão, Praia Grande, SP 11700–100, Brazil

[5] Center for Mathematics, Computing and Cognition, Federal University of ABC, Avenida dos Estados, 5001, Bangu, Santo André, SP 09210-580, Brazil

[6] Department of Statistical Inference and Mathematics, The Institute of Statistical Mathematics, 10–3 Midori-cho, Tachikawa, Tokyo 190–8562, Japan

[7] Continuous Optimization Team, RIKEN Center for Advanced Intelligence Project, Nihonbashi 1–chome Mitsui Building, 15th floor, 1–4–1 Nihonbashi, Chuo-ku, Tokyo 103–0027, Japan