Bregman Proximal Algorithms Exploiting DC Structure for Nonconvex Optimization and Their Applications

Shota Takahashi

Department of Statistical Science
School of Multidisciplinary Sciences
The Graduate University for Advanced Studies, SOKENDAI

March 2023

Abstract

In nonconvex optimization, it is difficult to obtain a global or even a local optimal solution. It is even challenging to obtain a stationary point quickly. In this thesis, we propose fast algorithms to obtain it for general nonconvex optimization problems. For dealing with nonconvex optimization problems, difference of convex functions (DC) optimization is a general effective approach. Exploiting DC structure, we propose the Bregman proximal DC algorithm (BPDCA), the BPDCA with extrapolation (BPDCAe), which is an acceleration of BPDCA, and the hybrid BPDCA (HBPDCA), which is an alternating minimization based on BPDCA and a convex optimization algorithm. These algorithms are applicable to a wide range of nonconvex optimization problems. In addition, the convergence of our proposed algorithms is theoretically guaranteed under the smooth adaptable property, which is less restrictive than L-smoothness. We establish convergence analysis for BPDCA(e) under the Kurdyka–Łojasiewicz (KL) property or the subanalyticity. For HBPDCA, we establish its global subsequential convergence. We demonstrate the excellent performance of the Bregman proximal algorithms exploiting DC structure through some applications, which include phase retrieval, blind deconvolution, and self-calibration in radio interferometric imaging. We first show that these problems are nonconvex and can be solved with the DC algorithms. Exploiting DC structure, we apply our proposed algorithms to them. For phase retrieval, we obtain new larger step sizes than the existing one. By using these step sizes, we succeed in accelerating BPDCA(e). For blind deconvolution, we obtain an appropriate Bregman distance by exploiting DC structure. We demonstrate BPDCA(e) through numerical experiments on phase retrieval and blind deconvolution. The results show that BPDCAe outperformed other existing algorithms. Especially, in blind deconvolution, our proposed algorithms successfully recovered the original image through numerical experiments on image deblurring. We also obtain the closed-form solution of the subproblem of HBPDCA for self-calibration in radio interferometric imaging.

Acknowledgements

First and foremost, I would like to my deepest gratitude to my supervisor, Prof. Mirai Tanaka. He has provided encouragement, advice, and guidance. This thesis would not have been completed without his thoughtful guidance.

I would like to thank my supervisor in the master's course, Prof. Mituhiro Fukuda. He gave me advice on my doctoral career and collaborated to write a paper. Moreover, I would like to thank my collaborator, Prof. Shiro Ikeda. The research with him was inspiring to me.

I am really grateful to the members of the dissertation committee, Prof. Takashi Takenouchi and Prof. Bruno F. Lourenço for giving valuable comments and judging my Ph.D. degree. I would like to express my special thanks to my colleagues at SOKENDAI and The Institute of Statistical Mathematics.

Last but not the least, I would like to thank my family for their constant support.

Contents

Chapter 1

Introduction

Optimization theory is an important technique in the fields of science and engineering. In optimization theory, the objective function measures the performance of a certain model on a constraint set. For example, in machine learning and signal processing, the objective function measures the goodness of fit between observation data, a model or the prior of a model, and some constraints.

An optimization problem is said to be convex if its objective function and its constraint set are convex. Otherwise, it is said to be nonconvex. In many interesting applications, including those of machine learning and signal processing, the objective function often becomes nonconvex. In this thesis, we propose fast algorithms for general *nonconvex* optimization problems.

Convex optimization has been studied for a long time and is known to be a powerful tool, such as least-squares and linear programming problems (see, for more details, [18, 87, 118]). In a convex optimization problem, any local optimal solution is also a global optimal solution. On the other hand, in a nonconvex optimization problem, a local optimal solution is not always a global optimal solution. There are many local optimal solutions and saddle points, and then optimization algorithms are sometimes trapped in them. In addition, in a nonconvex optimization problem, the convergence of algorithms for convex optimization is not theoretically guaranteed. Therefore, it is generally impossible to obtain even a local optimal solution. From this kind of circumstance, when the objective function is continuously differentiable, instead of finding a local optimal solution, the goal of nonconvex optimization is to obtain a stationary point. When the objective function is not continuously differentiable, the goal is to obtain a limiting stationary point (defined later in Definition 3.5), which is an extension of a stationary point. In general, any local optimal solution is a (limiting) stationary point (see also [102, Theorem 10.1]) and not vice versa.

Nonconvex optimization arises in many fields of science and engineering, such as machine learning [60, 70, 46] and signal processing [39, 47]. For machine learning, nonconvex optimization problems arise in a maximum a posteriori probability (MAP) estimate for image processing [16, 39], ridge regression [46], neural network [120], and support vector machine [130]. For signal processing, nonconvex optimization problems arise in

image deblurring with known blurring [10, 98] and without knowing the blurring kernel [54, 69, 95, 114], background/foreground extraction [17, 123], image compression [39], image separation [40], and communication engineering [109]. Therefore, it is important to study a fast algorithm with a good convergence property.

We first consider algorithms for the following optimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \quad f(\boldsymbol{x}), \tag{1.1}$$

where the function $f : \mathbb{R}^d \to (-\infty, +\infty]$ is convex. In mathematical optimization, most algorithms are iterative. In particular, algorithms that exploit first-order information such as objective function values, gradients, and subgradients (not Hessian) in optimization problems are called *first-order methods*. First-order methods for the convex optimization problem (1.1) have been studied for a long time. The oldest first-order method is the gradient descent method (also called the steepest descent method) by Cauchy [25] in 1847. His motivation was to compute the finite equations that represent the orbit of a heavenly body. Let an initial point $\boldsymbol{x}^0$ and a sequence of an iterative algorithm $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$. The gradient descent method requires that f is continuously differentiable, and the updating step at its iteration is given by

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \lambda^k \nabla f(\boldsymbol{x}^k),$$

where $\lambda^k > 0$ is called the step size at the kth iteration. The step size λ^k is given by, for example, line search (see also [11, 89]). If f is L-smooth, *i.e.*, there exists $L > 0$ such that

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d,$$

the step size λ^k can be set to a constant $\lambda^k = 1/L$. When f is nonsmooth, Shor [108] developed the subgradient method in the 1960s and applied it to network transportation problems. At each iteration of the subgradient method, the updating step is given by

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \lambda^k \boldsymbol{\xi}^k,$$

where $\boldsymbol{\xi}^k \in \partial_{\mathrm{c}} f(\boldsymbol{x}^k)$ is a (classical) subgradient of f at $\boldsymbol{x}^k$ defined by

$$\partial_{\mathrm{c}} f(\boldsymbol{x}) := \{\boldsymbol{\xi} \in \mathbb{R}^d \mid f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \boldsymbol{\xi}, \boldsymbol{y} - \boldsymbol{x}\rangle \geq 0, \forall \boldsymbol{y} \in \mathbb{R}^d\}.$$

The set $\partial_{\mathrm{c}} f(\boldsymbol{x})$ is called a (classical) subdifferential of f at $\boldsymbol{x} \in \mathbb{R}^d$.

The constrained optimization problem is given by the following equation:

$$\min_{\boldsymbol{x}\in C} \quad f(\boldsymbol{x}), \tag{1.2}$$

where the function $f : \mathbb{R}^d \to (-\infty, +\infty]$ is convex and possibly nonsmooth and the constraint set $C \subset \mathbb{R}^d$ is nonempty, closed, and convex. To solve (1.2), Polyak [97]

developed the projected subgradient method in 1987. At each iteration of the projected subgradient method, the subproblem to be solved is given by

$$\boldsymbol{x}^{k+1} = P_C(\boldsymbol{x}^k - \lambda^k \boldsymbol{\xi}^k),$$

where $\boldsymbol{\xi}^k \in \partial_{\mathrm{c}} f(\boldsymbol{x}^k)$ is a (classical) subgradient, and P_C is the orthogonal projection mapping defined by

$$P_C(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{y} \in C} \|\boldsymbol{y} - \boldsymbol{x}\|.$$

To generalize the projected subgradient method, for a function $g : \mathbb{R}^d \to (-\infty, +\infty]$, the proximal mapping is an effective approach that is defined by

$$\operatorname{prox}_g(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{y} \in \mathbb{R}^d} \left\{ g(\boldsymbol{y}) + \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2 \right\}.$$

Moreau [82] showed the properties of the proximal mapping. The proximal mapping is easily computable when g has a simple structure. Let δ_C be the indicator function of a constraint set $C \subset \mathbb{R}^d$, which is $\delta_C(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in C$ and $\delta_C(\boldsymbol{x}) = +\infty$ otherwise. By the definition of the proximal mapping, $\operatorname{prox}_{\delta_C} = P_C$. Thus, the subproblem of the projected (sub)gradient method is represented by

$$\boldsymbol{x}^{k+1} = \operatorname{prox}_{\lambda^k \delta_C}(\boldsymbol{x}^k - \lambda^k \boldsymbol{p}^k),$$

where $\boldsymbol{p}^k = \boldsymbol{\xi}^k$ for $\boldsymbol{\xi}^k \in \partial_{\mathrm{c}} f(\boldsymbol{x}^k)$, and especially $\boldsymbol{p}^k = \nabla f(\boldsymbol{x}^k)$ when f is continuously differentiable.

In applications of machine learning and signal processing, the objective function often includes several terms to avoid over-fitting or impose the structure of the model, called regularization. We consider the following composite optimization problem:

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} \quad f(\boldsymbol{x}) + g(\boldsymbol{x}), \tag{1.3}$$

where the function $f : \mathbb{R}^d \to (-\infty, +\infty]$ is convex and continuously differentiable, and the function $g : \mathbb{R}^d \to (-\infty, +\infty]$ is possibly nonsmooth. By setting $g = \delta_C$, we can write the constrained optimization problem (1.2) as a special case of (1.3). In machine learning and signal processing, f is a loss function to measure the performance of the model, and g is a regularization term, such as ℓ_0 regularization, ℓ_1 regularization, ℓ_2 regularization, and total variation regularization, to avoid over-fitting. The ℓ_0 norm $\|\boldsymbol{x}\|_0$ is the number of nonzero elements of $\boldsymbol{x}$, while it is nonconvex and nonsmooth. Instead of ℓ_0 regularization, convex and nonsmooth ℓ_1 regularization is used in many applications. In particular, the maximum likelihood with regularization can be regarded as a MAP estimate, and the linear regression with ℓ_1 regularization is called the least absolute shrinkage and selection operator (LASSO) [117]. In image processing, ℓ_1 regularization imposes the sparsity of images, and total variation regularization imposes the sparsity of the difference between

adjacent pixels of images. The optimization problem (1.3) has many kinds of applications in signal processing [47] and machine learning [60, 70]. For example, in signal processing, problem (1.3) arises in image up-sampling [4], background/foreground extraction [17, 123], MRI in medical image processing [36, 84], image deblurring [54, 69, 95, 114], image segmentation [67], and communication engineering [109]. In machine learning, regularization is used to avoid over-fitting and arises in the principal component analysis [38, 105], LASSO [117], and the support vector machine [130]. In these applications, the function f would be nonconvex (for example, [54, 69, 95, 114, 123, 130]). In addition, it is also possible to choose a nonconvex regularization g.

For solving the composite optimization problem (1.3), the proximal gradient method was introduced by Bruck [20], Passty [92], and Lions and Mercier [71]. At its iteration, the subproblem can be written as

$$\begin{aligned} \boldsymbol{x}^{k+1} &= \mathrm{prox}_{\lambda^k g}(\boldsymbol{x}^k - \lambda^k \nabla f(\boldsymbol{x}^k)) \\ &= \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^d} \left\{ \langle \nabla f(\boldsymbol{x}^k), \boldsymbol{x} - \boldsymbol{x}^k \rangle + g(\boldsymbol{x}) + \frac{1}{2\lambda^k} \|\boldsymbol{x} - \boldsymbol{x}^k\|_2^2 \right\}, \end{aligned} \tag{1.4}$$

where the step size λ^k is set to a constant $\lambda^k = 1/L$ when f is L-smooth or adaptively determined with line search. Subproblem (1.4) minimizes a first-order approximation of f with the regularization term $g(\boldsymbol{x})$ and the proximal term $\frac{1}{2\lambda^k}\|\boldsymbol{x}-\boldsymbol{x}^k\|_2^2$. The proximal term guarantees the accuracy of the first-order approximation. When $g = \theta\|\cdot\|_1$ for $\theta > 0$, the proximal gradient method is called the iterative shrinkage thresholding algorithm (ISTA), and (1.4) becomes

$$\boldsymbol{x}^{k+1} = \mathcal{S}_{\lambda^k\theta}(\boldsymbol{x}^k - \lambda^k \nabla f(\boldsymbol{x}^k)),$$

where $\mathcal{S}_{\lambda^k\theta}$ is called the soft thresholding operator, given by

$$\mathcal{S}_\theta(\boldsymbol{x}) = [\boldsymbol{x} - \theta \mathbf{1}_d]_+ \odot \mathrm{sgn}(\boldsymbol{x}),$$

where $\mathbf{1}_d \in \mathbb{R}^d$ is the d-dimensional all one vector, $([\boldsymbol{x}]_+)_i = \max\{x_i, 0\}$, $\odot$ denotes the Hadamard (elementwise) product, and sgn is defined by

$$\mathrm{sgn}(\boldsymbol{x})_i = \begin{cases} 1, & \text{if } x_i \geq 0, \\ -1, & \text{if } x_i < 0. \end{cases}$$

In this way, when the iteration of the proximal gradient method is represented in a closed form, the computational burden is reduced. See, for other examples of such proximal mappings, [9].

Let $\{\boldsymbol{x}^k\}_{k=0}^\infty$ be a sequence generated by the proximal gradient method for solving the optimization problem (1.3). Furthermore, when f is L-smooth and g is convex, for any optimal solution $\boldsymbol{x}^*$ and any $k \geq 1$, the rate of convergence is given by

$$(f(\boldsymbol{x}^k) + g(\boldsymbol{x}^k)) - (f(\boldsymbol{x}^*) + g(\boldsymbol{x}^*)) \leq \frac{L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}{2k}, \tag{1.5}$$

which is the rate $O(1/k)$ [9, Theorem 10.21]. This rate of convergence is called a sublinear rate. Beck and Teboulle [10] proposed the fast iterative shrinkage-thresholding algorithm (FISTA), which is an acceleration of ISTA. At each iteration of FISTA for solving the optimization problem (1.3), the updating step is defined by

$$\begin{aligned}\beta_k &= \frac{t_{k-1}-1}{t_k} \quad \text{with} \quad t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2},\\ \boldsymbol{y}^k &= \boldsymbol{x}^k + \beta_k(\boldsymbol{x}^k - \boldsymbol{x}^{k-1}),\\ \boldsymbol{x}^{k+1} &= \operatorname{prox}_{\lambda^k g}(\boldsymbol{y}^k - \lambda^k \nabla f(\boldsymbol{y}^k)),\end{aligned}$$

where $\boldsymbol{x}^{-1} = \boldsymbol{x}^0$ and $\lambda^k = 1/L$. Introducing such $\boldsymbol{y}^k$, β_k, and t_k was proposed by Nesterov [86, 87] when $g \equiv 0$. Let $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ be a sequence generated by FISTA. For any $k \geq 1$, the rate of convergence of FISTA is given by

$$(f(\boldsymbol{x}^k) + g(\boldsymbol{x}^k)) - (f(\boldsymbol{x}^*) + g(\boldsymbol{x}^*)) \leq \frac{2L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}{(k+1)^2}, \tag{1.6}$$

which is the rate $O(1/k^2)$. This means that FISTA is faster than ISTA and the proximal gradient method. Such acceleration technique is called extrapolation. Because this technique uses the momentum term $\boldsymbol{x}^k - \boldsymbol{x}^{k-1}$, it is also called Nesterov's momentum.

While f is sometimes not L-smooth in nonconvex optimization problems, the proximal gradient method requires that f is L-smooth for its global convergence. Bolte *et al.* [15] incorporated the Bregman distance [19], given by $D_\phi(\boldsymbol{x}, \boldsymbol{y}) := \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) - \langle \nabla\phi(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle$ for a convex and continuously differentiable function ϕ, into the proximal gradient method and proposed the Bregman proximal gradient algorithm (BPG). Instead of L-smoothness, BPG requires that the pair (f, ϕ) is L-smooth adaptable (L-smad), *i.e.*, there exists $L > 0$ such that $L\phi - f$ and $L\phi + f$ are convex, defined later in Definition 2.9. Even when f is not L-smooth, (f, ϕ) can be L-smad for some ϕ (see also an example in Remark 2.11). At each iteration of BPG, the subproblem is given by

$$\boldsymbol{x}^{k+1} = \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^d} \left\{ \langle \nabla f(\boldsymbol{x}^k), \boldsymbol{x} - \boldsymbol{x}^k\rangle + g(\boldsymbol{x}) + \frac{1}{\lambda^k} D_\phi(\boldsymbol{x}, \boldsymbol{x}^k) \right\}, \tag{1.7}$$

where the step size λ^k satisfies $0 < \lambda^k L < 1$. When $g \equiv 0$, BPG is called the mirror descent method, introduced by Nemirovski and Yudin [85] in 1983. For $\phi = \frac{1}{2}\|\cdot\|_2^2$, $D_\phi(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$. Thus, for $\phi = \frac{1}{2}\|\cdot\|_2^2$, BPG corresponds to the proximal gradient method. From this point of view, BPG is a generalization of the proximal gradient method, since the L-smad property is less restrictive than L-smoothness. Subproblem (1.7) minimizes a first-order approximation of f with the regularization term $g(\boldsymbol{x})$ and the Bregman proximity $\frac{1}{\lambda^k} D_\phi(\boldsymbol{x}, \boldsymbol{x}^k)$. The Bregman proximity guarantees the accuracy of the first-order approximation. Mukkamala *et al.* [83] proposed a variant of BPG that iteratively estimates small L by backtracking. To accelerate it, Wu *et al.* [122] proposed the inertial BPG, and Zhang *et al.* [128] proposed the BPG with extrapolation (BPGe).

For dealing with nonconvex optimization, difference of convex functions (DC) optimization is a general effective approach. Some researchers call DC optimization the concave-convex procedure [126]. When the objective function is equivalent to a DC function $f_1 - f_2$ with two convex functions $f_1, f_2 : \mathbb{R}^d \to (-\infty, +\infty]$, a DC optimization problem (1.1) becomes

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \quad f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}). \tag{1.8}$$

DC functions have been considered for about 70 years, for example, by Hartman [45] and Landis [61]. A well-known iterative method to solve the DC optimization problem (1.8) is the DC algorithm (DCA) (see also [65]). DCA was introduced by Pham and Souad [96] in 1986. At its iteration of DCA, the subproblem is given by

$$\boldsymbol{x}^{k+1} = \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^d} \left\{ f_1(\boldsymbol{x}) - \langle \boldsymbol{\xi}^k, \boldsymbol{x} - \boldsymbol{x}^k \rangle \right\}, \tag{1.9}$$

where $\boldsymbol{\xi}^k \in \partial_{\mathrm{c}} f_2(\boldsymbol{x}^k)$ is a (classical) subgradient of f_2 at $\boldsymbol{x}^k \in \mathbb{R}^d$. A sequence generated by DCA converges to a critical point $\tilde{\boldsymbol{x}}$ such that $\boldsymbol{0} \in \partial_{\mathrm{c}} f_1(\tilde{\boldsymbol{x}}) - \partial_{\mathrm{c}} f_2(\tilde{\boldsymbol{x}})$ (or equivalently $\partial_{\mathrm{c}} f_1(\tilde{\boldsymbol{x}}) \cap \partial_{\mathrm{c}} f_2(\tilde{\boldsymbol{x}}) \neq \emptyset$). In general, a critical point $\tilde{\boldsymbol{x}}$ is not always a local optimal solution. When f_2 is polyhedral convex and differentiable at $\tilde{\boldsymbol{x}}$, then a critical point $\tilde{\boldsymbol{x}}$ is also a local optimal solution [64, Theorem 1]. DC optimization is a powerful tool to deal with nonconvex optimization problems. However, the computational burden of DCA depends mainly on the resolution of subproblem (1.9). Additionally, solving subproblem (1.9) may be computationally demanding unless f_1 has a simple structure or (1.9) is small-scale.

Next, we consider the following DC optimization problem with a regularization term:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \quad f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}) + g(\boldsymbol{x}), \tag{1.10}$$

where the function $f_1 : \mathbb{R}^d \to (-\infty, +\infty]$ is convex and continuously differentiable, the function $f_2 : \mathbb{R}^d \to (-\infty, +\infty]$ is convex, and the function $g : \mathbb{R}^d \to (-\infty, +\infty]$ is possibly nonsmooth. Le Thi *et al.* [66] considered the DC optimization problem (1.10) with $g(\boldsymbol{x})$ as a penalty term. When f_1 is L-smooth and g is convex, the proximal DC algorithm (pDCA) (see, for example, [121]) is an alternative algorithm based on the proximal gradient method. At each iteration of pDCA, the subproblem is given by

$$\boldsymbol{x}^{k+1} = \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^d} \left\{ \langle \nabla f_1(\boldsymbol{x}^k) - \boldsymbol{\xi}^k, \boldsymbol{x} - \boldsymbol{x}^k \rangle + g(\boldsymbol{x}) + \frac{1}{2\lambda^k} \|\boldsymbol{x} - \boldsymbol{x}^k\|_2^2 \right\}, \tag{1.11}$$

where $\boldsymbol{\xi}^k \in \partial_{\mathrm{c}} f_2(\boldsymbol{x}^k)$ for $\boldsymbol{x}^k \in \mathbb{R}^d$, and the step size $\lambda^k > 0$ satisfies $0 < \lambda^k L < 1$. Wen *et al.* [121] proposed the proximal DC algorithm with extrapolation (pDCAe) in the same way as FISTA [10] and as Nesterov's extrapolation technique [86, 87].

For these first-order methods as above, their convergence behaviors have been studied in [5, 6, 13, 14, 15, 48, 83, 113, 121, 122, 128]. A sequence of these first-order methods converges to a limiting stationary point (defined later in Definition 3.5) under the

Kurdyka–Łojasiewicz (KL) property or subanalyticity, defined later in Section 2.4. The rate of convergence is also revealed under these properties and depends on the KL exponents. Note that the meaning of this rate is different from that of (1.5) and (1.6). Li *et al.* [68] developed calculus rules of the KL exponent, which affects the rate of convergence. The KL exponents for several problems were calculated in [124, 127].

In this thesis, for general nonconvex optimization problems, we propose fast algorithms exploiting the Bregman distance and DC structure (it is based on [113]). We proposed the Bregman proximal DC algorithm (BPDCA) [113], which is pDCA based on the Bregman distance, BPDCA with extrapolation (BPDCAe) [113], which is an acceleration of BPDCA, the hybrid BPDCA (HBPDCA), which minimizes subproblems based on the Bregman distance and a convex optimization problem (their updating steps are given in Sections 3.1, 3.2, and 3.3). In addition, we obtain convergence analysis of our proposed algorithms. The Bregman distance is a generalization of the squared Euclidean distance. Thus, it generalizes algorithms to apply a wide range of optimization problems, including optimization problems that lack L-smoothness. Because DC decomposition is not unique, we have flexibility in the choice of the Bregman distance. Using an appropriate Bregman distance, we accelerate and apply Bregman algorithms to a wider range of applications. Our proposed algorithms have the potential to address a variety of nonconvex optimization problems. BPDCAe adopts the adaptive restart scheme [113] on extrapolation based on the Bregman distance. He *et al.* [48] very recently proposed an alternating minimization algorithm extended from BPDCA. These applications to signal processing are based on [113, 114].

We demonstrate the performance of our proposed algorithms through some applications, which are phase retrieval, blind deconvolution, and self-calibration in radio interferometric imaging. They are known to be ill-posed because their solutions may not be unique. Adding some regularization, we write these problems as nonconvex optimization problems. For phase retrieval, exploiting DC structure, we obtain larger step sizes than the existing one, and then our proposed algorithms outperformed the existing algorithms [113]. For blind deconvolution and self-calibration in radio interferometric imaging, deriving Bregman algorithms were not trivial because their objective functions have the quartic and bilinear terms (see also Remark 4.5). Hence, exploiting DC structure, we obtain an appropriate Bregman distance for these applications and apply our proposed algorithms. Especially in blind deconvolution, through numerical experiments on image deblurring, our proposed algorithms successfully recovered the original image [114]. For self-calibration in radio interferometric imaging, we obtain a closed-form solution to the subproblem of HBPDCA.

Firstly, we applied our methods to phase retrieval. Phase retrieval is the problem of recovering the phase from magnitude measurements. Phase retrieval has a long history. For example, Patterson studied phase retrieval in X-ray crystallography in 1934 [93] and in 1944 [94]. Phase retrieval arises in many fields of science and engineering, such as image processing [22], astronomy [35], X-ray crystallography [78, 93, 94], and optics [104]. Many algorithms for solving phase retrieval have been proposed. Gerchberg and Saxton [43] represented phase retrieval as a nonconvex optimization problem and applied the alter-

nating projection algorithm to it, called the Gerchberg–Saxton algorithm. Fienup [42] proposed a modification of the Gerchberg–Saxton algorithm as the hybrid input-output algorithm. For the nonconvex optimization problem of phase retrieval, the semidefinite programming (SDP) relaxation was proposed as PhaseLift [21, 23] and PhaseCut [119]. Because these SDP approaches require massive memory for high dimensional problems, Candès *et al.* [22] proposed the Wirtinger flow algorithm, which is based on the gradient descent method and the Wirtinger derivatives (see Section 2.5). Sun *et al.* [111] applied the modified trust-region algorithm to phase retrieval. Bolte *et al.* [15] applied BPG to phase retrieval with ℓ_1 and ℓ_0 regularization. Some researchers conducted numerical experiments on phase retrieval with ℓ_1 regularization [113, 128].

Secondly, we solved blind deconvolution with our approach. Blind deconvolution is a technique to recover an original signal without knowing a convolving filter from its convolution. The study of blind deconvolution began in the 1970s, for example, [24, 110]. For non-blind deconvolution, *i.e.*, when a convolving filter is known, such as the point spread function, the Richardson–Lucy method was independently proposed by Richardson [101] and Lucy [73]. Blind deconvolution arises in many fields of science and engineering, such as sensor networks [7], optics [30], astronomy [41, 53], communication engineering [72], medical image processing [69, 129], and image processing [3, 52]. For blind deconvolution, Lane and Bates [62] proposed a classical iterative method. Under the assumption that the filter and the signal belong to the known subspaces, blind deconvolution is naturally formulated as a nonconvex optimization problem. Ahmed *et al.* [3] relaxed it as an SDP. Li *et al.* [69] represented blind deconvolution with smooth regularization and applied the Wirtinger gradient descent method. Because the nonconvex objective function has the quartic and bilinear terms, finding an appropriate Bregman distance is difficult, and the application of the Bregman proximal algorithms was challenging (see also Remark 4.5). Takahashi *et al.* [114] applied BPDCA(e) [113] to blind deconvolution with nonsmooth regularization by exploiting DC decomposition. For blind deconvolution with nonsmooth regularization, alternating minimization [54, 95] was also proposed.

Finally, we show that self-calibration in radio interferometric imaging can be solved with our approach. A radio interferometer has several antennas to observe radio waves. It measures the complex visibilities of Fourier-transformed images with noise. The purpose of calibration is to remove noise in the visibilities arising from measuring instruments and the atmosphere. Self-calibration is a calibration of complex gains given by each antenna. For technical aspects of self-calibration in radio interferometric imaging, see [116]. Self-calibration in radio interferometric imaging is represented as a nonconvex optimization problem. Its objective function is the chi-square of the difference between the observed visibilities and the corresponding values for a model. The model is given by the target image and the gains. Recently, sparse modeling has produced some remarkable results in astronomy, especially, in radio interferometric imaging [26, 56, 100]. In recent major news, the Event Horizon Telescope Collaboration has successfully photographed a black hole using radio interferometric imaging with sparse modeling [115]. Kuramochi *et al.* [56] proposed total squared variation (TSV) regularization for interferometric imaging. Repetti *et al.* [100] dealt with the nonconvex optimization problem with ℓ_1 regularization

and proposed the imaging method in radio interferometry.

In these applications of signal processing, the variables of optimization problems sometimes belong to $\mathbb{C}^d$. Complex optimization problems arise in image processing [2, 28], multiple-input multiple-output radar [44, 125], message passing [75], and sensor array [76]. As a problem with a special structure, complex fractional programming has also been studied in [58, 59]. Complex fractional programming arises in power control, beamforming [106], and uplink scheduling [107]. For other complex optimization problems, complex linear programming [32, 33] and complex-valued LASSO [77] have been studied. Sun *et al.* [112] summarized majorization-minimization algorithms for complex optimization problems. Some complex optimization algorithms require the Wirtinger derivatives, instead of complex derivatives. For details about complex analysis including the Wirtinger derivatives, see also Section 2.5.

1.1 Outline

This thesis is organized as follows. Chapter 2 summarizes the important notions and their examples, such as subdifferentials, the Bregman distance, the L-smad property, the KL property, subanalyticity, and complex analysis.

Chapter 3 introduces the Bregman proximal algorithms exploiting DC structure and establishes their convergence analysis. We first introduce BPDCA [113], which is pDCA based on the Bregman distance. Second, we introduce BPDCAe [113], which is accelerated by the extrapolation technique adapted to the Bregman distance. This extrapolation technique requires fewer computational tasks and is easy to implement. We establish global convergence of BPDCA(e) to a limiting stationary point or a limiting critical point under the KL property or subanalyticity of the objective function (for BPDCAe, the auxiliary function), respectively. Furthermore, we evaluate the rates of convergence of BPDCA(e). Finally, we propose the hybrid Bregman proximal DC algorithm (HBPDCA), which is different from [48]. It alternately minimizes the two subproblems: One is the same as BPDCA, while the other is a convex optimization problem. For HBPDCA, we establish global subsequential convergence.

Chapter 4 shows applications to signal processing, such as phase retrieval, blind deconvolution, and self-calibration in radio interferometric imaging. These applications are represented as nonconvex optimization problems. We reformulate each problem as a DC optimization problem and apply BPDCA, BPDCAe, and HBPDCA. In order to apply these algorithms, we obtain an appropriate Bregman distance and a parameter $L > 0$ that ensure the L-smad property. For phase retrieval, we obtain several smaller L than the existing one. Using these L, we succeed in accelerating BPDCA(e) [113]. For blind deconvolution, although it is difficult to find an appropriate Bregman distance, we obtain an appropriate Bregman distance by exploiting DC structure. We demonstrate BPDCA and BPDCAe through numerical experiments on phase retrieval [113] and blind deconvolution [114]. Especially in blind deconvolution, we provide the stability analysis of the equilibrium points, and our proposed algorithms successfully recovered the original

image through numerical experiments on image deblurring [114]. For self-calibration in radio interferometric imaging, we obtain an appropriate Bregman distance by using DC structure and provide the closed-form solution of the subproblem of HBPDCA.

Chapter 5 summarizes our contributions and discusses future work.

1.2 Notations

In what follows, we use the following notations. Let $\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{R}_{++}$ be the set of real numbers, nonnegative real numbers, and positive real numbers, respectively. Let $\mathbb{R}^d$ and $\mathbb{R}^d_+$ be the real space of d dimension and the positive orthant of the real space, respectively. $\mathbb{C}$ denotes the set of complex numbers and $\mathbb{C}^d$ denotes the complex space of d dimension. Let $\mathbb{R}^{d_1\times d_2}$ and $\mathbb{C}^{d_1\times d_2}$ be the set of $d_1 \times d_2$ real and complex matrices, respectively. $\mathbb{S}^d$ denotes the set of $d \times d$ real symmetric matrices. Vectors and matrices are shown in boldface. The d-dimensional all one vector is $\mathbf{1}_d \in \mathbb{R}^d$, the d-dimensional all zero vector is $\mathbf{0}_d \in \mathbb{R}^d$, the $d\times d$ identity matrix is $\boldsymbol{I}_d \in \mathbb{R}^{d\times d}$, and the $d\times d$ zero matrix is $\boldsymbol{O}_d \in \mathbb{R}^{d\times d}$. Let $|\boldsymbol{z}|$ and $\boldsymbol{z}^2$ be elementwise absolute and squared vectors for $\boldsymbol{z} \in \mathbb{C}^d$, respectively. $\mathrm{Re}(\boldsymbol{z})$, $\bar{\boldsymbol{z}}$, and $\boldsymbol{z}^{\mathsf{H}}$ denote its real part, complex conjugate, and complex conjugate transpose, respectively. The inner product of $\boldsymbol{z}$, $\boldsymbol{w} \in \mathbb{C}^d$ (or $\mathbb{R}^d$) is defined by $\langle \boldsymbol{z}, \boldsymbol{w}\rangle = \boldsymbol{z}^{\mathsf{H}}\boldsymbol{w}$. The operator $\odot$ denotes The Hadamard (elementwise) product. Given a real number $p \geq 1$, the ℓ_p norm is defined by $\|\boldsymbol{z}\|_p = (\sum_{j=1}^d |z_j|^p)^{1/p}$. For a matrix $\boldsymbol{M} \in \mathbb{C}^{d\times d}$ (or $\mathbb{R}^{d\times d}$), the Frobenius norm is defined by $\|\boldsymbol{M}\|_F = \sqrt{\sum_{j,k} |M_{j,k}|^2}$. Let $\lambda_{\min}(\boldsymbol{M})$ and $\lambda_{\max}(\boldsymbol{M})$ be the minimum and maximum eigenvalues of a symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$, respectively.

Let $\operatorname{int} C$ and $\operatorname{cl} C$ be the interior and the closure of a set $C \subset \mathbb{R}^d$, respectively. We also define the distance from a point $\boldsymbol{x} \in \mathbb{R}^d$ to C by $\operatorname{dist}(\boldsymbol{x}, C) := \inf_{\boldsymbol{y}\in C} \|\boldsymbol{x} - \boldsymbol{y}\|_2$. The function $\delta_C(\boldsymbol{x})$ is the indicator function $\delta_C(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in C$ and $\delta_C(\boldsymbol{x}) = +\infty$ otherwise.

Chapter 2

Preliminaries

2.1 Subdifferentials

For an extended-real-valued function $f : \mathbb{R}^d \to [-\infty, +\infty]$, we introduce the set

$$\operatorname{dom} f := \{\boldsymbol{x} \in \mathbb{R}^d \mid f(\boldsymbol{x}) < +\infty\}$$

called the effective domain. The function f is said to be proper if $f(\boldsymbol{x}) > -\infty$ for all $\boldsymbol{x} \in \mathbb{R}^d$ and $\operatorname{dom} f \neq \emptyset$.

Definition 2.1 (Regular and limiting subdifferentials [102, Definition 8.3])**.** *For a proper and lower semicontinuous function $f : \mathbb{R}^d \to (-\infty, +\infty]$, the regular subdifferential of f at $\boldsymbol{x} \in \operatorname{dom} f$ is defined by*

$$\hat{\partial} f(\boldsymbol{x}) = \left\{ \boldsymbol{\xi} \in \mathbb{R}^d \;\middle|\; \liminf_{\boldsymbol{y} \to \boldsymbol{x},\, \boldsymbol{y} \neq \boldsymbol{x}} \frac{f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \boldsymbol{\xi}, \boldsymbol{y} - \boldsymbol{x} \rangle}{\|\boldsymbol{y} - \boldsymbol{x}\|_2} \geq 0 \right\}.$$

The limiting subdifferential of f at $\boldsymbol{x} \in \operatorname{dom} f$ is defined by

$$\partial f(\boldsymbol{x}) = \left\{ \boldsymbol{\xi} \in \mathbb{R}^d \;\middle|\; \exists \boldsymbol{x}^k \xrightarrow{f} \boldsymbol{x},\ \boldsymbol{\xi}^k \to \boldsymbol{\xi} \text{ such that } \boldsymbol{\xi}^k \in \hat{\partial} f(\boldsymbol{x}^k) \text{ for all } k \right\},$$

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

In general, $\hat{\partial} f(\boldsymbol{x}) \subset \partial f(\boldsymbol{x})$ holds for all $\boldsymbol{x} \in \mathbb{R}^d$ [102, Theorem 8.6].

Example 2.2 ([102, p. 304])**.** *Let a function $f : \mathbb{R} \to \mathbb{R}$ be given by*

$$f(x) = \begin{cases} x^2 \sin \frac{1}{x} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

Although $\hat{\partial} f(0) = \{0\}$, we have $\partial f(0) = [-1, 1]$. We have $\hat{\partial} f(0) \subset \partial f(0)$ with $\hat{\partial} f(0) \neq \partial f(0)$.

In the following example, it holds that $\hat{\partial} f(\boldsymbol{x}) = \partial f(\boldsymbol{x})$.

Example 2.3 ([102, p. 303])**.** *Let a function $f : \mathbb{R} \to \mathbb{R}$ be given by*

$$f(x) = \begin{cases} x^2 + x & \text{if } x \leq 0, \\ 1 - x & \text{if } x > 0. \end{cases}$$

We have $\hat{\partial} f(0) = \partial f(0) = [1, +\infty)$.

We also define $\operatorname{dom} \partial f := \{\boldsymbol{x} \in \mathbb{R}^d \mid \partial f(\boldsymbol{x}) \neq \emptyset\}$. Note that when f is convex, the limiting subdifferential coincides with the (classical) subdifferential [102, Proposition 8.12], that is, $\partial f(\boldsymbol{x}) = \partial_{\mathrm{c}} f(\boldsymbol{x})$. For a proper and lower semicontinuous function $f : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to (-\infty, +\infty]$, the partial subdifferential $\partial_{\boldsymbol{x}} f(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ of f at $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ with respect to $\boldsymbol{x}$ is defined by the subdifferential of $f(\cdot, \tilde{\boldsymbol{y}})$ at $\tilde{\boldsymbol{x}}$ for a fixed $\tilde{\boldsymbol{y}}$. Similarly, we define the partial subdifferential of f at $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ with respect to $\boldsymbol{y}$.

In addition, we can define the limiting subdifferential of the gradient. Let a function $f : \mathbb{R}^d \to (-\infty, +\infty]$ be proper and continuously differentiable and ∇f be locally Lipschitz continuous. By using the second-order subdifferential [81, Definition 2.1] and [80, Proposition 1.120], we obtain

$$\partial(\nabla f(\boldsymbol{x}))(\boldsymbol{u}) = \partial \langle \boldsymbol{u}, \nabla f(\boldsymbol{x}) \rangle, \quad \forall \boldsymbol{u} \in \mathbb{R}^d . \tag{2.1}$$

2.2 Bregman Distances

Definition 2.4 (Kernel generating distance [15, Definition 2.1])**.** *Let C be a nonempty open convex subset of $\mathbb{R}^d$. A function $\phi : \mathbb{R}^d \to (-\infty, +\infty]$ is called a kernel generating distance associated with C if it meets the following conditions:*

(i) *ϕ is proper, lower semicontinuous, and convex, with $\operatorname{dom} \phi \subset \operatorname{cl} C$ and $\operatorname{dom} \partial \phi = C$.*

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

We denote the class of kernel generating distances associated with C by $\mathcal{G}(C)$.

Definition 2.5 (Bregman distance [19])**.** *Given $\phi \in \mathcal{G}(C)$, the Bregman distance $D_\phi : \operatorname{dom} \phi \times \operatorname{int} \operatorname{dom} \phi \to \mathbb{R}_+$ is defined by*

$$D_\phi(\boldsymbol{x}, \boldsymbol{y}) := \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) - \langle \nabla \phi(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle.$$

From the gradient inequality, the function ϕ is convex if and only if $D_\phi(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ for any $\boldsymbol{x} \in \operatorname{dom} \phi$ and $\boldsymbol{y} \in \operatorname{int} \operatorname{dom} \phi$. When ϕ is a strictly convex function, the equality holds if and only if $\boldsymbol{x} = \boldsymbol{y}$. When $\phi = \frac{1}{2} \|\cdot\|_2^2$, $D_\phi(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$, which is the squared Euclidean distance. We show other well-known examples of ϕ and D_ϕ below (see also [8] and [37, Table 2.1]).

Example 2.6. *We show examples of $d = 1$:*

- *Boltzmann–Shannon entropy: Let* $\phi(x) = x\log x$, $\operatorname{dom}\phi = \mathbb{R}_+$, *and* $0\log 0 = 0$. *Then, we obtain*

$$D_\phi(x,y) = x\log x - y\log y - (\log y + 1)(x-y) = x\log\frac{x}{y} - x + y.$$

  This D_ϕ *leads to the Kullback–Leibler divergence* [55] *in Example* 2.7.

- *Burg entropy: Let* $\phi(x) = -\log x$ *and* $\operatorname{dom}\phi = \mathbb{R}_{++}$. *Then, we obtain*

$$D_\phi(x,y) = -\log x + \log y + \frac{1}{y}(x-y) = \frac{x}{y} - \log\frac{x}{y} - 1.$$

  This D_ϕ *is called the Itakura–Saito divergence* [51].

- *Fermi–Dirac entropy: Let* $\phi(x) = x\log x + (1-x)\log(1-x)$ *and* $\operatorname{dom}\phi = [0,1]$. *Then, we obtain*

$$\begin{aligned} D_\phi(x,y) &= x\log x + (1-x)\log(1-x) - y\log y - (1-y)\log(1-y) \\ &\quad - (\log y - \log(1-y))(x-y) \\ &= x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}. \end{aligned}$$

- *Hellinger: Let* $\phi(x) = -\sqrt{1-x^2}$ *and* $\operatorname{dom}\phi = [-1,1]$. *Then, we obtain*

$$D_\phi(x,y) = -\sqrt{1-x^2} + \sqrt{1-y^2} - \frac{y}{\sqrt{1-y^2}}(x-y) = \frac{1-xy}{\sqrt{1-y^2}} - \sqrt{1-x^2}.$$

Example 2.7. *We show examples in* $\mathbb{R}^d$. *Kernel generating distances* ϕ *and Bregman distances* D_ϕ *on* $\mathbb{R}^1$ *in Example* 2.6 *are extended to* $\tilde{\phi}(\boldsymbol{x}) = \sum_{j=1}^d \phi(x_j)$ *and* $D_{\tilde{\phi}}(\boldsymbol{x},\boldsymbol{y}) = \sum_{j=1}^d D_\phi(x_j,y_j)$ *on* $\mathbb{R}^d$.

- *Quadratic form: For a positive definite matrix* $\boldsymbol{A} \in \mathbb{S}^d$, *let* $\phi(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\mathsf{T}\boldsymbol{A}\boldsymbol{x}$ *and* $\operatorname{dom}\phi = \mathbb{R}^d$. *Then, we obtain*

$$D_\phi(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2}\boldsymbol{x}^\mathsf{T}\boldsymbol{A}\boldsymbol{x} - \frac{1}{2}\boldsymbol{y}^\mathsf{T}\boldsymbol{A}\boldsymbol{y} - \langle\boldsymbol{A}\boldsymbol{y}, \boldsymbol{x}-\boldsymbol{y}\rangle = \frac{1}{2}(\boldsymbol{x}-\boldsymbol{y})^\mathsf{T}\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{y}).$$

  This D_ϕ *is called the general quadratic distance. In addition, let* P *be a probability distribution,* $\boldsymbol{\Sigma}$ *be its positive definite covariance matrix, and* $\boldsymbol{\mu}$ *be its mean vector. When* $\boldsymbol{A} = \boldsymbol{\Sigma}^{-1}$ *and* $\boldsymbol{y} = \boldsymbol{\mu}$, D_ϕ *is called the Mahalanobis distance* [74] *with respect to* P.

- *Quartic function: Let* $\phi(\boldsymbol{x}) = \frac{1}{4}\|\boldsymbol{x}\|_2^4 + \frac{1}{2}\|\boldsymbol{x}\|_2^2$ *and* $\operatorname{dom}\phi = \mathbb{R}^d$. *Then, we obtain*

$$\begin{aligned} D_\phi(\boldsymbol{x},\boldsymbol{y}) &= \frac{1}{4}\|\boldsymbol{x}\|_2^4 + \frac{1}{2}\|\boldsymbol{x}\|_2^2 - \frac{1}{4}\|\boldsymbol{y}\|_2^4 - \frac{1}{2}\|\boldsymbol{y}\|_2^2 - \langle(\|\boldsymbol{y}\|_2^2 + 1)\boldsymbol{y}, \boldsymbol{x}-\boldsymbol{y}\rangle \\ &= \frac{1}{4}\|\boldsymbol{x}\|_2^4 - \frac{1}{4}\|\boldsymbol{y}\|_2^4 - \langle\|\boldsymbol{y}\|_2^2\boldsymbol{y}, \boldsymbol{x}-\boldsymbol{y}\rangle + \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2. \end{aligned}$$

- *Boltzmann–Shannon entropy: Let* $\phi(\boldsymbol{x}) = \sum_{j=1} x_j \log x_j$ *and* $\operatorname{dom}\phi = \mathbb{R}^d_+$. *We consider* D_ϕ *on the unit simplex* $\{\boldsymbol{x} \in \mathbb{R}^d_+ \mid \sum_{j=1}^d x_j = 1\}$. *Then, we obtain*

$$D_\phi(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^d \left(x_j \log \frac{x_j}{y_j} - x_j + y_j \right) = \sum_{j=1}^d x_j \log \frac{x_j}{y_j}.$$

  This D_ϕ *is called the Kullback–Leibler divergence* [55].

When the function ϕ is separable, the subproblem of Bregman proximal algorithms may be reduced to d independent one-dimensional problems (see also Section 3.1). Note that the first and second ϕ in Example 2.7 are not separable. Even if ϕ is not separable, depending on the combination of ϕ and g, the subproblem can be solved in a closed form.

Furthermore, the Bregman distance satisfies the three-point identity [29, Lemma 3.1],

$$D_\phi(\boldsymbol{x}, \boldsymbol{z}) - D_\phi(\boldsymbol{x}, \boldsymbol{y}) - D_\phi(\boldsymbol{y}, \boldsymbol{z}) = \langle \nabla\phi(\boldsymbol{y}) - \nabla\phi(\boldsymbol{z}), \boldsymbol{x} - \boldsymbol{y} \rangle, \tag{2.2}$$

for any $\boldsymbol{y}, \boldsymbol{z} \in \operatorname{int}\operatorname{dom}\phi$, and $\boldsymbol{z} \in \operatorname{dom}\phi$.

Remark 2.8. *Bregman distances are nonnegative and satisfy the three-point identity. However, Bregman distances are neither symmetric nor satisfy the triangle inequality except, for example,* $\phi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2$ *and* $\phi(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\mathsf{T}\boldsymbol{A}\boldsymbol{x}$. *For example, we consider the Itakura–Saito divergence in Example 2.6 and have*

$$D_\phi(x, y) - D_\phi(y, x) = \frac{x}{y} - \log\frac{x}{y} - \frac{y}{x} + \log\frac{y}{x} = \frac{x^2 - y^2}{xy} - 2\log\frac{x}{y},$$

which implies $D_\phi(x, y) = D_\phi(y, x)$ *if and only if* $x = y$. *Because of* $D_\phi(x, y) = 0$ *for* $x = y$, *the Itakura–Saito divergence is not symmetric. From (2.2), if* $\langle \nabla\phi(\boldsymbol{y}) - \nabla\phi(\boldsymbol{z}), \boldsymbol{x} - \boldsymbol{y} \rangle \leq 0$, *the triangle inequality holds. Otherwise, it does not hold. Therefore, the Bregman distance is not a metric.*

The symmetrized Bregman distance $\tilde{D}_\phi$ *is defined by*

$$\tilde{D}_\phi(\boldsymbol{x}, \boldsymbol{y}) = \frac{D_\phi(\boldsymbol{x}, \boldsymbol{y}) + D_\phi(\boldsymbol{y}, \boldsymbol{x})}{2}.$$

It is used in computing entropic centers [88].

2.3 Smooth Adaptable Functions

Now let us define the notions of the L-smooth adaptable property.

Definition 2.9 (L-smooth adaptable [15])**.** *Consider a pair of functions* (f, ϕ) *satisfying the following conditions:*

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

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

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

When $\phi = \frac{1}{2}\|\cdot\|_2^2$, the L-smad property corresponds to L-smoothness, *i.e.*, the L-smad property is a generalization of L-smoothness.

When f and ϕ are $\mathcal{C}^2$, then $L\phi - f$ is convex on C if and only if there exists $L > 0$ such that $L\nabla^2\phi(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x}) \succeq \boldsymbol{O}_d$ for all $\boldsymbol{x} \in C$, given by [8, Proposition 1]. From this, it is easy to obtain $L > 0$ such that $L\phi - f$ is convex. Examples of the L-smad property are shown in [8, Lemma 7] and [15, Lemma 5.1]. In this thesis, we show the L-smad property for phase retrieval in Propositions 4.1 and 4.3, for blind deconvolution in Theorem 4.6, and for self-calibration in Theorem 4.13.

From the L-smooth adaptable property comes the descent lemma [15].

Lemma 2.10 (Full extended descent lemma [15])**.** *A pair of functions* (f, ϕ) *is* L*-smad on* $C = \operatorname{int}\operatorname{dom}\phi$ *if and only if:*

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

For further properties, see also [8].

Remark 2.11. *From Lemma 2.10, the* L*-smad property is an adequate upper approximation of* f *by* ϕ*. No research has focused on finding an appropriate* ϕ *for the* L*-smad property. It is an empirical way to construct* ϕ *from the basis functions that constitute* f*.*

For example, $f(x) = ax^4 + bx^2 + c$ *for* $a > 0$*,* $b > 0$*, and* $c \in \mathbb{R}$*. This* $f(x)$ *is not* L*-smooth because* $|f'(x) - f'(y)| = |4ax^3 + 2bx - 4ay^3 - 2by|$*. The basis functions that constitute* f *are* x^4*,* x^2*, and* 1*. From these basis functions, let* $\phi(x) = x^4 + x^2$ *(we can ignore a constant term). For this choice, we can prove that* $L\phi - f$ *is convex if and only if* $L = \max\{a, b\}$*.*

How to construct an appropriate ϕ *is an important issue for future research.*

2.4 Kurdyka–Łojasiewicz Property and Subanalyticity

Given $\eta > 0$, let Ξ_η denote the set of all continuous concave functions $\psi : [0, \eta) \to \mathbb{R}_+$ that are $\mathcal{C}^1$ on $(0, \eta)$ with positive derivatives and which satisfy $\psi(0) = 0$. Here, we introduce the Kurdyka–Łojasiewicz property [14, 57], which we need when analyzing our algorithms:

Definition 2.12 (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 $\hat{\boldsymbol{x}} \in \operatorname{dom} \partial f$ if there exist $\eta \in (0, +\infty]$, a neighborhood U of $\hat{\boldsymbol{x}}$, and a function $\psi \in \Xi_\eta$ such that, for all*

$$\boldsymbol{x} \in U \cap \{\boldsymbol{x} \in \mathbb{R}^d \mid f(\hat{\boldsymbol{x}}) < f(\boldsymbol{x}) < f(\hat{\boldsymbol{x}}) + \eta\},$$

the following inequality holds:

$$\psi'(f(\boldsymbol{x}) - f(\hat{\boldsymbol{x}})) \cdot \operatorname{dist}(\mathbf{0}_d, \partial f(\boldsymbol{x})) \geq 1. \tag{2.3}$$

(ii) *If f has the KL property at each point of $\operatorname{dom} \partial f$, then it is called a KL function.*

Lemma 2.13 (Uniformized KL property [14])**.** *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 $\psi \in \Xi_\eta$ such that*

$$\psi'(f(\boldsymbol{x}) - f(\hat{\boldsymbol{x}})) \cdot \operatorname{dist}(\mathbf{0}_d, \partial f(\boldsymbol{x})) \geq 1,$$

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

When the function f is continuously differentiable, instead of the KL property, we consider the Łojasiewicz gradient inequality.

Remark 2.14. *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function. For the sake of simplicity, we also assume that f is $\mathcal{C}^1$. Let $\psi : [0, \eta) \to \mathbb{R}_+$ be given by $\psi(s) = cs^{1-\theta}$ for some $\theta \in [0, 1)$ and $c > 0$. From $\psi'(s) = c(1-\theta)s^{-\theta}$, we obtain (2.3) given by*

$$c(1-\theta)(f(\boldsymbol{x}) - f(\hat{\boldsymbol{x}}))^{-\theta} \|\nabla f(\boldsymbol{x})\|_2 \geq 1.$$

Multiplying $(f(\boldsymbol{x}) - f(\hat{\boldsymbol{x}}))^\theta$ on both sides of the above inequality, we obtain

$$(f(\boldsymbol{x}) - f(\hat{\boldsymbol{x}}))^\theta \leq c(1-\theta)\|\nabla f(\boldsymbol{x})\|_2.$$

This inequality is called the Łojasiewicz gradient inequality [57]. Therefore, it is a special case of (2.3).

We consider the Łojasiewicz gradient inequality for specific functions.

Example 2.15. *For the sake of simplicity, let a function $f : \mathbb{R}^1 \to \mathbb{R}$ be continuously differentiable.*

- *When $f(x) = x^2$ and $\hat{x} = 0$, we obtain*

$$x^{2\theta} \leq 2c(1-\theta)|x|.$$

For $\theta = \frac{1}{2}$ and $c \geq 1$, the Łojasiewicz gradient inequality holds.

- *When* $f(x) = x^2 + x^4$ *and* $\hat{x} = 0$, *we obtain*

$$(x^2 + x^4)^\theta \leq 2c(1-\theta)|x + 2x^3|.$$

For $\theta = \frac{1}{2}$, *it holds that*

$$|x|\sqrt{1+x^2} \leq c|x|(1+2x^2). \tag{2.4}$$

When $x = 0$, (2.4) *holds for any* $c > 0$. *When* $x \neq 0$, *from* (2.4), *we have*

$$\sqrt{1+x^2} \leq c(1+2x^2).$$

By simple calculations, this inequality holds when $c \geq 1$. *Therefore, for* $\theta = \frac{1}{2}$ *and* $c \geq 1$, *the Łojasiewicz gradient inequality holds.*

- *When* $f(x) = (x-a)^{2n}$, $n \geq \frac{1}{2}$, $a \in \mathbb{R}$, *and* $\hat{x} = a$, *we obtain*

$$(x-a)^{2n\theta} \leq 2nc(1-\theta)|x-a|^{2n-1}.$$

For $\theta = \frac{2n-1}{2n}$ *and* $c \geq 1$, *the Łojasiewicz gradient inequality holds. Taking* $n \to +\infty$, $\theta \to 1$ *and the graph of* f *is flat around* $\hat{x} = a$. *The closer* θ *is to 1, the flatter* f *is; the closer* θ *is to 0, the shaper* f *is.*

The parameter θ *controls the sharpness of* f. *However, for general* d, *the parameter* θ *is difficult to obtain.*

Next, we describe subanalytic functions.

Definition 2.16 (Subanalyticity [13])**.**

(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:*

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

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*

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

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* $\mathbb{R}^d \times \mathbb{R}$.

Example 2.17. *We show some important examples below:*

- *Given a subanalytic set S,* $\mathrm{dist}(\boldsymbol{x}, S)$ *is subanalytic* [13, p. 1208].
- *Osgood's example* [90, Theorem 1] *: Let $f : \mathbb{R}^2 \to \mathbb{R}^3$ be given by*

$$f(x, y) = \begin{bmatrix} x \\ xy \\ xye^y \end{bmatrix}.$$

  Then, the set $A = \{f(x, y) \mid x^2 + y^2 \leq 1\}$ is not semianalytic but subanalytic.
- *$f(x) = |x|^{1/r}$ for $r \in \mathbb{N}$ is subanalytic* [91].

Note that every subanalytic function is a KL function. See [12, 13, 90, 91] for further properties of subanalyticity.

2.5 Complex Analysis

We introduce the Wirtinger derivatives for complex functions. In this section, we follow the outline by [22, Section 6], [49], and [103, Appendix 2]. Applications to complex analysis are summarized in [1].

A complex function $f : \mathbb{C} \to \mathbb{C}$ is said to be complex differentiable at $z_0 \in \mathbb{C}$ if there exists $f'(z_0) \in \mathbb{C}$ given by

$$f'(z_0) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}.$$

If the complex function f is complex differentiable, then f is called a holomorphic function. Otherwise, f is called an antiholomorphic function. For a complex number $z \in \mathbb{C}$, $\mathrm{Re}(z)$ and $\mathrm{Im}(z)$ are called the real and imaginary parts of z, respectively. That is, for $z = x + \sqrt{-1}y$, $x = \mathrm{Re}(z)$ and $y = \mathrm{Im}(z)$ hold. For real-valued functions $u, v : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, the function f can be written as

$$f(z) = u(x, y) + \sqrt{-1}v(x, y). \tag{2.5}$$

The function f is complex differentiable at $z_0 \in \mathbb{C}$ if and only if u and v are $\mathcal{C}^1$ and Cauchy–Riemann equations, defined by

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x},$$

hold at z_0.

In optimization problems, we use real-valued functions. However, in general, real-valued functions with complex variables are not complex differentiable because $\frac{\partial v}{\partial y} = 0$ and $\frac{\partial v}{\partial x} = 0$ hold at any points from $v \equiv 0$, *i.e.*, Cauchy–Riemann equations do not hold. For antiholomorphic functions, instead of complex derivatives, the Wirtinger derivatives are used.

Definition 2.18 (Wirtinger derivatives [49]). *Let $f : \mathbb{C} \to \mathbb{C}$ be a complex function, and let $z = x + \sqrt{-1}y$, where $x, y \in \mathbb{R}$, then the Wirtinger derivatives with respect to z and $\bar{z}$ at $z_0 \in \mathbb{C}$ are defined by*

$$\frac{\partial f(z_0)}{\partial z} := \frac{1}{2}\left(\frac{\partial f(z_0)}{\partial x} - \sqrt{-1}\frac{\partial f(z_0)}{\partial y}\right), \tag{2.6}$$

$$\frac{\partial f(z_0)}{\partial \bar{z}} := \frac{1}{2}\left(\frac{\partial f(z_0)}{\partial x} + \sqrt{-1}\frac{\partial f(z_0)}{\partial y}\right). \tag{2.7}$$

Note that the right-hand sides of (2.6) and (2.7) correspond to the derivatives of f when the variables z and $\bar{z}$ are treated as independent variables. We confirm that this fact is true in the following example.

Example 2.19. *Let a function $f : \mathbb{C} \to \mathbb{R}$ be given by $f(z) = |z|^2 = x^2 + y^2$, where $z = x + \sqrt{-1}y$ for $x, y \in \mathbb{R}$. From Definition 2.18, we obtain the Wirtinger derivatives of f at $z_0 \in \mathbb{C}$:*

$$\frac{\partial f(z_0)}{\partial z} = \frac{1}{2}\left(\frac{\partial f(z_0)}{\partial x} - \sqrt{-1}\frac{\partial f(z_0)}{\partial y}\right) = \frac{1}{2}\left(2x_0 - 2\sqrt{-1}y_0\right) = \bar{z}_0,$$

$$\frac{\partial f(z_0)}{\partial \bar{z}} = \frac{1}{2}\left(\frac{\partial f(z_0)}{\partial x} + \sqrt{-1}\frac{\partial f(z_0)}{\partial y}\right) = \frac{1}{2}\left(2x_0 + 2\sqrt{-1}y_0\right) = z_0,$$

where $z_0 = x_0 + \sqrt{-1}y_0$ for $x_0, y_0 \in \mathbb{R}$. Next, we treat the variables z and $\bar{z}$ as independent variables for f, i.e., let a function $g : \mathbb{C}^2 \to \mathbb{R}$ be defined by $g(z, \bar{z}) := z\bar{z} = |z|^2 = f(z)$. The Wirtinger derivatives of f at $z_0 \in \mathbb{C}$ are given by

$$\frac{\partial f(z_0)}{\partial z} = \left.\frac{\partial g(z, \bar{z})}{\partial z}\right|_{z=z_0} = \bar{z}_0, \quad \frac{\partial f(z_0)}{\partial \bar{z}} = \left.\frac{\partial g(z, \bar{z})}{\partial \bar{z}}\right|_{z=z_0} = z.$$

These results correspond to the first results.

The Wirtinger derivatives are also given for holomorphic functions. We show an example of the Wirtinger derivatives for holomorphic functions below.

Example 2.20. *Let a function $f : \mathbb{C} \to \mathbb{C}$ be given by $f(z) = z$. f is a holomorphic function because f is complex differentiable at $z_0 \in \mathbb{C}$, that is, there exists $f'(z_0)$, given by*

$$f'(z_0) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} = \lim_{z \to z_0} \frac{z - z_0}{z - z_0} = 1.$$

On the other hand, the Wirtinger derivatives of $f(z) = z = x + \sqrt{-1}y$ at $z_0 \in \mathbb{C}$ are given by

$$\frac{\partial f(z_0)}{\partial z} = \frac{1}{2}\left(\frac{\partial f(z_0)}{\partial x} - \sqrt{-1}\frac{\partial f(z_0)}{\partial y}\right) = \frac{1}{2}\left(1 - \sqrt{-1}\sqrt{-1}\right) = \frac{1}{2}\left(1 + 1\right) = 1,$$

$$\frac{\partial f(z_0)}{\partial \bar{z}} = \frac{1}{2}\left(\frac{\partial f(z_0)}{\partial x} + \sqrt{-1}\frac{\partial f(z_0)}{\partial y}\right) = \frac{1}{2}\left(1 + \sqrt{-1}\sqrt{-1}\right) = \frac{1}{2}\left(1 - 1\right) = 0.$$

All holomorphic functions are analytic and vice versa from Cauchy's integral expression. For a holomorphic function f, $f'(z_0) = \frac{\partial f(z_0)}{\partial z}$ and $\frac{\partial f(z_0)}{\partial \bar{z}} = 0$.

Next, we consider the Wirtinger derivatives for multivariable functions. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, and $\boldsymbol{z} = \boldsymbol{x} + \sqrt{-1}\boldsymbol{y}$, let $f : \mathbb{C}^d \to \mathbb{C}$ be a complex function

$$f(\boldsymbol{z}) = u(\boldsymbol{x}, \boldsymbol{y}) + \sqrt{-1}v(\boldsymbol{x}, \boldsymbol{y}),$$

where real-valued functions $u, v : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. The Wirtinger derivatives of f at $\boldsymbol{z}_0 \in \mathbb{C}^d$ are defined by

$$\frac{\partial f(\boldsymbol{z}_0)}{\partial \boldsymbol{z}} := \left[\frac{\partial f(\boldsymbol{z}_0)}{\partial z_1}, \cdots, \frac{\partial f(\boldsymbol{z}_0)}{\partial z_d}\right],$$
$$\frac{\partial f(\boldsymbol{z}_0)}{\partial \bar{\boldsymbol{z}}} := \left[\frac{\partial f(\boldsymbol{z}_0)}{\partial \bar{z}_1}, \cdots, \frac{\partial f(\boldsymbol{z}_0)}{\partial \bar{z}_d}\right].$$

From these results, the complex gradient at $\boldsymbol{z}_0 \in \mathbb{C}^d$ is given by

$$\nabla_{\mathbb{C}} f(\boldsymbol{z}_0) = \left[\frac{\partial f(\boldsymbol{z}_0)}{\partial \boldsymbol{z}} \quad \frac{\partial f(\boldsymbol{z}_0)}{\partial \bar{\boldsymbol{z}}}\right]^{\mathsf{H}},$$

and the complex Hessian at $\boldsymbol{z}_0 \in \mathbb{C}^d$ is given by

$$\nabla^2_{\mathbb{C}} f(\boldsymbol{z}_0) := \begin{bmatrix} \boldsymbol{H}_{\boldsymbol{z}\boldsymbol{z}} & \boldsymbol{H}_{\bar{\boldsymbol{z}}\boldsymbol{z}} \\ \boldsymbol{H}_{\boldsymbol{z}\bar{\boldsymbol{z}}} & \boldsymbol{H}_{\bar{\boldsymbol{z}}\bar{\boldsymbol{z}}} \end{bmatrix}, \tag{2.8}$$

where $\boldsymbol{H}_{\boldsymbol{z}\boldsymbol{z}} := \frac{\partial}{\partial \boldsymbol{z}}\left(\frac{\partial f(\boldsymbol{z}_0)}{\partial \boldsymbol{z}}\right)^{\mathsf{H}}$, $\boldsymbol{H}_{\bar{\boldsymbol{z}}\boldsymbol{z}} := \frac{\partial}{\partial \bar{\boldsymbol{z}}}\left(\frac{\partial f(\boldsymbol{z}_0)}{\partial \boldsymbol{z}}\right)^{\mathsf{H}}$, $\boldsymbol{H}_{\boldsymbol{z}\bar{\boldsymbol{z}}} := \frac{\partial}{\partial \boldsymbol{z}}\left(\frac{\partial f(\boldsymbol{z}_0)}{\partial \bar{\boldsymbol{z}}}\right)^{\mathsf{H}}$, and $\boldsymbol{H}_{\bar{\boldsymbol{z}}\bar{\boldsymbol{z}}} := \frac{\partial}{\partial \bar{\boldsymbol{z}}}\left(\frac{\partial f(\boldsymbol{z}_0)}{\partial \bar{\boldsymbol{z}}}\right)^{\mathsf{H}}$. When f is a real-valued function, we can easily prove that

$$\overline{\frac{\partial f(\boldsymbol{z}_0)}{\partial \boldsymbol{z}}} = \frac{\partial f(\boldsymbol{z}_0)}{\partial \bar{\boldsymbol{z}}}.$$

For real-valued functions of complex variables, we define

$$\nabla f(\boldsymbol{z}_0) = \left(\frac{\partial f(\boldsymbol{z}_0)}{\partial \boldsymbol{z}}\right)^{\mathsf{H}}.$$

Therefore, for any $\boldsymbol{w} \in \mathbb{C}^d$, we obtain

$$\left\langle \nabla_{\mathbb{C}} f(\boldsymbol{z}_0), \begin{bmatrix} \boldsymbol{w} \\ \bar{\boldsymbol{w}} \end{bmatrix} \right\rangle = 2\,\mathrm{Re}\langle \nabla f(\boldsymbol{z}_0), \boldsymbol{w} \rangle. \tag{2.9}$$

The following example shows the complex gradient and (2.9).

Example 2.21. *Let a function $f : \mathbb{C}^d \to \mathbb{R}$ be given by $f(\boldsymbol{z}) = \|\boldsymbol{z}\|_2^2 = \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2$ for $\boldsymbol{z} = \boldsymbol{x} + \sqrt{-1}\boldsymbol{y}$, where $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. We obtain the Wirtinger derivatives of f at $\boldsymbol{w} \in \mathbb{C}^d$:*

$$\frac{\partial f(\boldsymbol{z}_0)}{\partial z_j} = \frac{1}{2}\left(\frac{\partial f(\boldsymbol{z}_0)}{\partial x_j} - \sqrt{-1}\frac{\partial f(\boldsymbol{z}_0)}{\partial y_j}\right) = \frac{1}{2}\left(2\operatorname{Re}(z_{0,j}) - 2\sqrt{-1}\operatorname{Im}(z_{0,j})\right) = \bar{z}_{0,j},$$
$$\frac{\partial f(\boldsymbol{z}_0)}{\partial \bar{z}_j} = \frac{1}{2}\left(\frac{\partial f(\boldsymbol{z}_0)}{\partial x_j} + \sqrt{-1}\frac{\partial f(\boldsymbol{z}_0)}{\partial y_j}\right) = \frac{1}{2}\left(2\operatorname{Re}(z_{0,j}) + 2\sqrt{-1}\operatorname{Im}(z_{0,j})\right) = z_{0,j},$$

where $z_{0,j}$ is the jth element of $\boldsymbol{z}_0$. Therefore, we obtain the complex gradient, given by

$$\nabla_{\mathbb{C}} f(\boldsymbol{z}_0) = \begin{bmatrix} \boldsymbol{z}_0 \\ \bar{\boldsymbol{z}}_0 \end{bmatrix}.$$

In addition, we obtain

$$\begin{aligned}\left\langle \nabla_{\mathbb{C}} f(\boldsymbol{z}_0), \begin{bmatrix} \boldsymbol{w} \\ \bar{\boldsymbol{w}} \end{bmatrix} \right\rangle &= \begin{bmatrix} \boldsymbol{z}_0^{\mathsf{H}} & \bar{\boldsymbol{z}}_0^{\mathsf{H}} \end{bmatrix} \begin{bmatrix} \boldsymbol{w} \\ \bar{\boldsymbol{w}} \end{bmatrix} \\ &= \boldsymbol{z}_0^{\mathsf{H}}\boldsymbol{w} + \overline{\boldsymbol{z}_0^{\mathsf{H}}\boldsymbol{w}} \\ &= 2\operatorname{Re}(\boldsymbol{z}_0^{\mathsf{H}}\boldsymbol{w}) \\ &= 2\operatorname{Re}\langle \nabla f(\boldsymbol{z}_0), \boldsymbol{w} \rangle,\end{aligned}$$

where the last equation holds because of $\nabla f(\boldsymbol{z}_0) = \boldsymbol{z}_0$.

Inspired by the Wirtinger derivatives, the complex subdifferential is defined.

Definition 2.22 (Complex subdifferential)**.** *For a proper, lower semicontinuous, and convex function $f : \mathbb{C}^d \to (-\infty, +\infty]$, the complex subdifferential of f at $\boldsymbol{x} \in \operatorname{dom} f$ is defined by*

$$\partial_{\mathrm{c}} f(\boldsymbol{x}) = \left\{ \boldsymbol{\xi} \in \mathbb{C}^d \;\middle|\; f(\boldsymbol{y}) - f(\boldsymbol{x}) - 2\operatorname{Re}\langle \boldsymbol{\xi}, \boldsymbol{y} - \boldsymbol{x} \rangle \geq 0, \forall \boldsymbol{y} \in \mathbb{C}^d \right\}.$$

Chapter 3

Bregman Proximal Algorithms Exploiting DC Structure

3.1 Bregman Proximal DC Algorithm

We are interested in solving the following DC optimization problem with a regularization term:

$$\min_{\boldsymbol{x}\in \mathrm{cl}\, C} \quad \Psi(\boldsymbol{x}) := f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}) + g(\boldsymbol{x}), \tag{3.1}$$

where $f_1, f_2 : \mathbb{R}^d \to (-\infty, +\infty]$ are convex functions on $\mathbb{R}^d$, and $C \subset \mathbb{R}^d$ is a nonempty open convex set. Also, the function $g : \mathbb{R}^d \to (-\infty, +\infty]$ may be nonsmooth, such as the ℓ_1 norm $\|x\|_1$ in [15, 83, 128], or alternatively, f_2 may be nonsmooth [63]. Some interesting examples of (3.1) can be found in [121]. Although we will place some assumptions on C, it can be regarded as $\mathbb{R}^d$ for simplicity.

Recall that $C = \operatorname{int} \operatorname{dom} \phi$.

Assumption 3.1.

(i) $\phi \in \mathcal{G}(C)$ *with* $\mathrm{cl}\, C = \mathrm{cl} \operatorname{dom} \phi$.

(ii) $f_1 : \mathbb{R}^d \to (-\infty, +\infty]$ *is proper and convex with* $\operatorname{dom} \phi \subset \operatorname{dom}(f_1 + g)$, *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* $\operatorname{dom} g \cap C \neq \emptyset$.

(v) $v := \inf_{\boldsymbol{x}\in \mathrm{cl}\, C} \Psi(\boldsymbol{x}) > -\infty$.

(vi) *For any* $\lambda > 0$, $\lambda g + \phi$ *is supercoercive, that is,*

$$\lim_{\|\boldsymbol{u}\|_2 \to \infty} \frac{\lambda g(\boldsymbol{u}) + \phi(\boldsymbol{u})}{\|\boldsymbol{u}\|_2} = \infty.$$

Let $\boldsymbol{x} \in \operatorname{dom}(f_1 + g)$, then $f_2(\boldsymbol{x}) \leq g(\boldsymbol{x}) + f_1(\boldsymbol{x}) - v < +\infty$ due to Assumption 3.1 (v). Thus, $\boldsymbol{x} \in \operatorname{dom} f_2$, *i.e.*, $\operatorname{dom}(f_1 + g) \subset \operatorname{dom} f_2$. From Assumption 3.1 (ii), we have $C \subset \operatorname{dom}(f_1 + g) \subset \operatorname{dom} f_2$. Note that Assumption 3.1 (vi) holds when $\operatorname{cl} C$ is compact [15, p. 2136].

To obtain the Bregman proximal DC algorithm (BPDCA) mapping for some $\lambda > 0$, we recast the objective function of (3.1) via a DC decomposition:

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

and, given $\boldsymbol{x} \in C = \operatorname{int}\operatorname{dom}\phi$ and $\boldsymbol{\xi} \in \partial_{\mathrm{c}} f_2(\boldsymbol{x})$, define the mapping,

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

Additionally, we put the following assumption on (3.1).

Assumption 3.2. *For all $\boldsymbol{x} \in C$ and $\lambda > 0$, we have*

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

Note that Assumption 3.2 is not so restrictive because it holds when $C \equiv \mathbb{R}^d$. Under Assumptions 3.1 and 3.2, we have the following lemma [15, Lemma 3.1].

Lemma 3.3. *Suppose that Assumptions 3.1 and 3.2 hold, and let $\boldsymbol{x} \in C = \operatorname{int}\operatorname{dom}\phi$. Then, the set $\mathcal{T}_\lambda(\boldsymbol{x})$ is a nonempty and compact subset of C for any $\lambda > 0$.*

Note that when the function ϕ is strictly convex, $\mathcal{T}_\lambda(\boldsymbol{x})$ is a singleton. Also, when g and ϕ are separable, this mapping is easily computable, since $\mathcal{T}_\lambda(\boldsymbol{x})$ can be decomposed into a single-valued optimization problem, and often has a closed-form solution. For example, when $\phi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2$, for $g(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$, $\mathcal{T}_\lambda(\boldsymbol{x})$ becomes the soft-thresholding operator or, for $g(\boldsymbol{x}) = \|\boldsymbol{x}\|_0$, the hard-thresholding operator.

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

Algorithm 1 Bregman proximal DC algorithm (BPDCA)

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

$$\boldsymbol{x}^{k+1} = \operatorname*{argmin}_{\boldsymbol{x} \in \operatorname{cl} C} \left\{ \langle \nabla f_1(\boldsymbol{x}^k) - \boldsymbol{\xi}^k, \boldsymbol{x} - \boldsymbol{x}^k \rangle + g(\boldsymbol{x}) + \frac{1}{\lambda} D_\phi(\boldsymbol{x}, \boldsymbol{x}^k) \right\}. \tag{3.2}$$

end for

The parameter λ ($< 1/L$) 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.

As a recurrent example, $D_\phi(\boldsymbol{x}, \boldsymbol{x}^k) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}^k\|_2^2$ when $\phi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2$. In this case, if L is regarded as the Lipschitz constant for the gradient of f_1, subproblem (3.2) reduces to subproblem (1.11). If f_2 is $\mathcal{C}^1$ on C and the pair $(f_1 - f_2, \phi)$ is L-smad, BPDCA reduces to BPG [15].

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

3.1.1 Properties of BPDCA

First, we show the sufficiently decreasing property of BPDCA mapping for $0 < \lambda L < 1$ (the argument is adapted from [15, Lemma 4.1]). We define the sufficiently decreasing property below.

Definition 3.4 (Sufficiently decreasing property)**.** *Let $\Psi : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function. A sequence $\{\boldsymbol{x}^k\}_{k=0}^\infty$ has the sufficient decrease property if there exists a positive scalar κ such that*

$$\kappa D_\phi(\boldsymbol{x}^k, \boldsymbol{x}^{k+1}) < \Psi(\boldsymbol{x}^k) - \Psi(\boldsymbol{x}^{k+1}) \quad \forall k \in \mathbb{N}.$$

When ϕ is σ-strongly convex (see also Assumption 3.7 (i)), we obtain $\frac{\kappa\sigma}{2}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| \leq \kappa D_\phi(\boldsymbol{x}^k, \boldsymbol{x}^{k+1}) < \Psi(\boldsymbol{x}^k) - \Psi(\boldsymbol{x}^{k+1})$, which implies the sufficiently decreasing property by the squared Euclidean distance [15, Definition 4.1].

Lemma 3.5. *Suppose that Assumptions 3.1 and 3.2 hold. For any $\boldsymbol{x} \in C = \operatorname{int}\operatorname{dom}\phi$ and any $\boldsymbol{x}^+ \in C = \operatorname{int}\operatorname{dom}\phi$ defined by*

$$\boldsymbol{x}^+ \in \operatorname*{argmin}_{\boldsymbol{u} \in \operatorname{cl} C} \left\{ \langle \nabla f_1(\boldsymbol{x}) - \boldsymbol{\xi}, \boldsymbol{u} - \boldsymbol{x} \rangle + g(\boldsymbol{u}) + \frac{1}{\lambda} D_\phi(\boldsymbol{u}, \boldsymbol{x}) \right\}, \tag{3.3}$$

where $\boldsymbol{\xi} \in \partial_{\mathrm{c}} f_2(\boldsymbol{x})$ and $\lambda > 0$, it holds that

$$\lambda \Psi(\boldsymbol{x}^+) \leq \lambda \Psi(\boldsymbol{x}) - (1 - \lambda L) D_\phi(\boldsymbol{x}^+, \boldsymbol{x}). \tag{3.4}$$

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 $\boldsymbol{x}^+$ by taking $\boldsymbol{u} = \boldsymbol{x} \in \operatorname{int}\operatorname{dom}\phi$ and $\boldsymbol{\xi} \in \partial_{\mathrm{c}} f_2(\boldsymbol{x})$, we obtain

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

Invoking the full extended descent lemma (Lemma 2.10) for f_1, the definition of the subgradient for f_2, and the above inequality, we have

$$f_1(\boldsymbol{x}^+) - f_2(\boldsymbol{x}^+) + g(\boldsymbol{x}^+)$$

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

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

Proposition 3.6 follows immediately from Lemma 3.5, as in [15].

Proposition 3.6. *Suppose that Assumptions 3.1 and 3.2 hold. Let $\{\boldsymbol{x}^k\}_{k=0}^\infty$ be a sequence generated by BPDCA with $0 < \lambda L < 1$. Then, the following statements hold:*

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

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

(iii) *$\min_{1 \leq k \leq n} D_\phi(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}) \leq \frac{\lambda}{n} \left(\frac{\Psi(\boldsymbol{x}^0) - \Psi_*}{1 - \lambda L} \right)$, where $\Psi_* := v > -\infty$ (by Assumption 3.1 (v).*

3.1.2 Convergence Analysis of BPDCA

Suppose that the following conditions hold.

Assumption 3.7.

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

(ii) *$\nabla \phi$ 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 $\{\boldsymbol{x} \in \mathbb{R}^d \mid \Psi(\boldsymbol{x}) \leq r\}$ are bounded.*

Since $C = \operatorname{int} \operatorname{dom} \phi = \mathbb{R}^d$ under Assumption 3.7 (i), Assumptions 3.2 and 3.18 are automatically fulfilled. For nonconvex functions, we use the limiting subdifferential [102]. We define the limiting critical points and the limiting stationary points of Ψ.

Definition 3.8. *We say that $\tilde{\boldsymbol{x}}$ is a limiting critical point of (3.1) with $C \equiv \mathbb{R}^d$ if*

$$
\mathbf{0}_d \in \nabla f_1(\tilde{\boldsymbol{x}}) - \partial_{\mathrm{c}} f_2(\tilde{\boldsymbol{x}}) + \partial g(\tilde{\boldsymbol{x}}). \tag{3.5}
$$

The set of all limiting critical points of (3.1) is denoted by $\mathcal{X}$. In addition, we say that $\tilde{\boldsymbol{x}}$ is a limiting stationary point of (3.1) with $C \equiv \mathbb{R}^d$ if

$$
\mathbf{0}_d \in \partial \Psi(\tilde{\boldsymbol{x}}). \tag{3.6}
$$

Although the limiting stationary points are sometimes called the limiting critical points in some papers, for example, [14, Definition 1 (iv)], we distinguish these two terms. The reasons are the following: When Ψ is convex, we call $\tilde{\boldsymbol{x}}$ a stationary point if it satisfies $\mathbf{0}_d \in \partial_{\mathrm{c}}\Psi(\tilde{\boldsymbol{x}})$. Because (3.6) is its natural extension by replacing $\partial_{\mathrm{c}}\Psi$ with $\partial\Psi$, we use the terminology "limiting stationary point" after [34, Definition 6.1.4]. We similarly name $\tilde{\boldsymbol{x}}$ satisfying (3.5): When g is convex, we call $\tilde{\boldsymbol{x}}$ a critical point if it satisfies $\mathbf{0}_d \in \nabla f_1(\tilde{\boldsymbol{x}}) - \partial_{\mathrm{c}} f_2(\tilde{\boldsymbol{x}}) + \partial_{\mathrm{c}} g(\tilde{\boldsymbol{x}})$. Because (3.5) is its natural extension by replacing $\partial_{\mathrm{c}} g$ with ∂g, we use the terminology "limiting critical point."

The limiting stationary point is a first-order necessary condition for local optimality. This relation is known as the generalized Fermat's rule [102, Theorem 10.1]. We can deduce $\partial(g - f_2)(\boldsymbol{x}) \subseteq \partial g(\boldsymbol{x}) - \partial_{\mathrm{c}} f_2(\boldsymbol{x})$ from [80, Corollary 3.4]. Plugging it into [102, Corollary 10.9], it generally holds that $\partial\Psi(\boldsymbol{x}) \subseteq \nabla f_1(\boldsymbol{x}) - \partial_{\mathrm{c}} f_2(\boldsymbol{x}) + \partial g(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$. It implies that 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 $\partial\Psi(\boldsymbol{x}) \equiv \nabla f_1(\boldsymbol{x}) - \nabla f_2(\boldsymbol{x}) + \partial g(\boldsymbol{x})$ from [102, Corollary 10.9] or [79, 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 $\mathcal{C}^1$. We show an example of limiting critical points and limiting stationary points.

Example 3.9. *Consider functions $f_1, f_2, g : \mathbb{R} \to \mathbb{R}$, and $\Psi = f_1 - f_2 + g$, given by*

$$f_1(x) = x^2, \quad f_2(x) = \max\{-2x, x\}, \quad g(x) = \max\{-x, 2x\}, \text{ and } \Psi(x) = x^2 + x.$$

In this case, since the functions f_1, f_2, g, and Ψ are convex, their limiting subdifferentials correspond to (classical) subdifferentials. Then, we obtain $\nabla f_1(x) = 2x$, $\partial_{\mathrm{c}}\Psi(x) = \{2x + 1\}$,

$$\partial_{\mathrm{c}} f_2(x) = \begin{cases} \{-2\} & \text{if } x < 0, \\ [-2, 1] & \text{if } x = 0, \\ \{1\} & \text{if } x > 0, \end{cases} \qquad \partial_{\mathrm{c}} g(x) = \begin{cases} \{-1\} & \text{if } x < 0, \\ [-1, 2] & \text{if } x = 0, \\ \{2\} & \text{if } x > 0, \end{cases}$$

and hence

$$\nabla f_1(x) - \partial_{\mathrm{c}} f_2(x) + \partial_{\mathrm{c}} g(x) = \begin{cases} \{2x + 1\} & \text{if } x \neq 0, \\ [-2, 4] & \text{if } x = 0. \end{cases}$$

Therefore, we have $\partial_{\mathrm{c}}\Psi(x) \subset \nabla f_1(x) - \partial_{\mathrm{c}} f_2(x) + \partial_{\mathrm{c}} g(x)$ for any $x \in \mathbb{R}$. For $\tilde{x} = -\frac{1}{2}$, because of $0 \in \partial_{\mathrm{c}}\Psi(\tilde{x}) = \{0\}$ and $0 \in \nabla f_1(\tilde{x}) - \partial_{\mathrm{c}} f_2(\tilde{x}) + \partial_{\mathrm{c}} g(\tilde{x}) = \{0\}$, $\tilde{x}$ is a limiting critical point and also a limiting stationary point. However, for $\tilde{x} = 0$, because of $0 \notin \partial_{\mathrm{c}}\Psi(\tilde{x}) = \{1\}$ and $0 \in \nabla f_1(\tilde{x}) - \partial_{\mathrm{c}} f_2(\tilde{x}) + \partial_{\mathrm{c}} g(\tilde{x}) = [-2, 4]$, $\tilde{x}$ is not a limiting stationary point but a limiting critical point.

Next, using Lemma 3.5 and Proposition 3.6, we will show the global subsequential convergence of the iterates to a limiting critical point of the problem (3.1). We can easily see that Theorem 3.10 (i) holds from the level-boundedness of Ψ. Theorem 3.10 (iii) and (vi) will play an essential role in determining global convergence and the rate of convergence of BPDCA.

Theorem 3.10 (Global subsequential convergence of BPDCA)**.** *Suppose that Assumptions 3.1, 3.2, and 3.7 hold. Let $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ be a sequence generated by BPDCA with $0 < \lambda L < 1$ for solving (3.1). Then, the following statements hold:*

(i) *The sequence $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ is bounded.*

(ii) *The sequence $\{\boldsymbol{\xi}^k\}_{k=0}^{\infty}$ is bounded.*

(iii) $\lim_{k\to\infty} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2 = 0$.

(iv) *Any accumulation point of $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ is a limiting critical point of (3.1).*

Proof. (i) From Proposition 3.6, we obtain $\Psi(\boldsymbol{x}^k) \leq \Psi(\boldsymbol{x}^0)$ for all $k \in \mathbb{N}$, which shows that $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ is bounded from Assumption 3.7 (iii).

(ii) From Assumption 3.1 (ii), 3.7 (i), and the convexity of f_2, $\operatorname{dom} f_2 = \mathbb{R}^d$ and $\partial_{\mathrm{c}} f_2(\boldsymbol{x}^k) \neq \emptyset$. Suppose, for the sake of proof by contradiction, that $\{\boldsymbol{\xi}^k\}_{k=0}^{\infty}$ is unbounded, *i.e.*, $\|\boldsymbol{\xi}^k\|_2 \to \infty$ as $k \to \infty$. By the definition of the subgradients of convex functions, we see that for any $\boldsymbol{y} \in \mathbb{R}^d$,

$$f_2(\boldsymbol{y}) \geq f_2(\boldsymbol{x}^k) + \langle \boldsymbol{\xi}^k, \boldsymbol{y} - \boldsymbol{x}^k \rangle. \tag{3.7}$$

Assume for a moment that $\|\boldsymbol{\xi}^k\|_2 \neq 0$. Letting $\{\boldsymbol{d}^k\}_{k=0}^{\infty}$ be the subsequence given by $\boldsymbol{d}^k = \boldsymbol{\xi}^k / \|\boldsymbol{\xi}^k\|_2$ and substituting $\boldsymbol{x}^k + \boldsymbol{d}^k = \boldsymbol{x}^k + \boldsymbol{\xi}^k / \|\boldsymbol{\xi}^k\|_2$ into $\boldsymbol{y}$ in (3.7), we obtain

$$f_2(\boldsymbol{x}^k + \boldsymbol{d}^k) \geq f_2(\boldsymbol{x}^k) + \left\langle \boldsymbol{\xi}^k, \boldsymbol{d}^k \right\rangle = f_2(\boldsymbol{x}^k) + \|\boldsymbol{\xi}^k\|_2,$$

which is also true when $\|\boldsymbol{\xi}^k\|_2 = 0$ by defining $\boldsymbol{d}^k = 0$. By taking $k \to \infty$, we obtain

$$\limsup_{k\to\infty} \|\boldsymbol{\xi}^k\|_2 \leq \limsup_{k\to\infty} \left(f_2(\boldsymbol{x}^k + \boldsymbol{d}^k) - f_2(\boldsymbol{x}^k) \right). \tag{3.8}$$

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

$$\limsup_{k\to\infty} \left(f_2(\boldsymbol{x}^k + \boldsymbol{d}^k) - f_2(\boldsymbol{x}^k) \right) \leq f_2(\bar{\boldsymbol{x}}) - \bar{f}_2 < \infty, \tag{3.9}$$

for some value $\bar{f}_2 \leq f_2(\boldsymbol{x}^k)$, $k \geq 0$. (3.8) and (3.9) contradict $\|\boldsymbol{\xi}^k\|_2 \to \infty$.

(iii) From (3.4), we obtain

$$\begin{aligned} \Psi(\boldsymbol{x}^{k-1}) - \Psi(\boldsymbol{x}^k) &\geq \left(\frac{1}{\lambda} - L \right) D_\phi(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}) \\ &\geq \left(\frac{1}{\lambda} - L \right) \frac{\sigma}{2} \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2^2, \end{aligned} \tag{3.10}$$

where the last inequality holds since ϕ is a σ-strongly convex function from Assumption 3.7 (i). Summing the above inequality from $k = 1$ to ∞, we obtain

$$\left(\frac{1}{\lambda} - L \right) \sum_{k=1}^{\infty} \frac{\sigma}{2} \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2^2 \leq \Psi(\boldsymbol{x}^0) - \liminf_{n\to\infty} \Psi(\boldsymbol{x}^n) \leq \Psi(\boldsymbol{x}^0) - v < \infty,$$

which shows that $\lim_{k\to\infty} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2 = 0$.

(iv) Let $\tilde{\boldsymbol{x}}$ be an accumulation point of $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ and let $\{\boldsymbol{x}^{k_j}\}$ be a subsequence such that $\lim_{j\to\infty} \boldsymbol{x}^{k_j} = \tilde{\boldsymbol{x}}$. Then, from the first-order optimality condition of subproblem (3.2) under Assumption 3.2, we have

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

Therefore,

$$\boldsymbol{\xi}^{k_j} + \frac{1}{\lambda}\left(\nabla\phi(\boldsymbol{x}^{k_j}) - \nabla\phi(\boldsymbol{x}^{k_j+1})\right) \in \partial g(\boldsymbol{x}^{k_j+1}) + \nabla f_1(\boldsymbol{x}^{k_j}). \tag{3.11}$$

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

$$\left\|\frac{1}{\lambda}\left(\nabla\phi(\boldsymbol{x}^{k_j}) - \nabla\phi(\boldsymbol{x}^{k_j+1})\right)\right\|_2 \leq \frac{A_0}{\lambda}\|\boldsymbol{x}^{k_j+1} - \boldsymbol{x}^{k_j}\|_2.$$

Therefore, using $\|\boldsymbol{x}^{k_j+1} - \boldsymbol{x}^{k_j}\|_2 \to 0$, we obtain

$$\frac{1}{\lambda}\left(\nabla\phi(\boldsymbol{x}^{k_j}) - \nabla\phi(\boldsymbol{x}^{k_j+1})\right) \to \mathbf{0}_d. \tag{3.12}$$

Note that the sequence $\{\boldsymbol{\xi}^{k_j}\}$ 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 $\lim_{j\to\infty} \boldsymbol{\xi}^{k_j} =: \tilde{\boldsymbol{\xi}}$ exists, which belongs to $\partial_{\mathrm{c}} f_2(\tilde{\boldsymbol{x}})$ since f_2 becomes continuous due to its convexity on $\mathbb{R}^d$. Using this and (3.12), we can take the limit of (3.11). Setting $\|\boldsymbol{x}^{k_j+1} - \boldsymbol{x}^{k_j}\|_2 \to 0$ and invoking the lower semicontinuity of g and ∇f_1, we obtain $\tilde{\boldsymbol{\xi}} \in \partial g(\tilde{\boldsymbol{x}}) + \nabla f_1(\tilde{\boldsymbol{x}})$. Therefore, $\mathbf{0}_d \in \partial g(\tilde{\boldsymbol{x}}) + \nabla f_1(\tilde{\boldsymbol{x}}) - \partial_{\mathrm{c}} f_2(\tilde{\boldsymbol{x}})$, which shows that $\tilde{\boldsymbol{x}}$ is a limiting critical point of (3.1). □

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

Proposition 3.11. *Suppose that Assumptions 3.1, 3.2, and 3.7 hold. Let $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ be a sequence generated by BPDCA with $0 < \lambda L < 1$ for solving (3.1). Then, the following statements hold:*

(i) $\zeta := \lim_{k\to\infty} \Psi(\boldsymbol{x}^k)$ *exists.*

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

Proof. (i) From Assumption 3.1 (v) and Proposition 3.6 (i), the sequence $\{\Psi(\boldsymbol{x}^k)\}_{k=0}^{\infty}$ is bounded from below and non-increasing. Consequently, $\zeta := \lim_{k\to\infty} \Psi(\boldsymbol{x}^k)$ exists.

(ii) Take any $\hat{\boldsymbol{x}} \in \Omega$, that is $\lim_{j\to\infty} \boldsymbol{x}^{k_j} = \hat{\boldsymbol{x}}$. From (3.2), it follows that

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

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

From the above inequality and the fact that f_1 is convex at $\boldsymbol{x}^k$, we obtain

$$\begin{aligned} f_1(\boldsymbol{x}^k) + g(\boldsymbol{x}^k) \leq & \langle \nabla f_1(\boldsymbol{x}^{k-1}) - \boldsymbol{\xi}^{k-1}, \hat{\boldsymbol{x}} - \boldsymbol{x}^k \rangle + g(\hat{\boldsymbol{x}}) + \frac{1}{\lambda} D_\phi(\hat{\boldsymbol{x}}, \boldsymbol{x}^{k-1}) - \frac{1}{\lambda} D_\phi(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}) \\ & + f_1(\hat{\boldsymbol{x}}) + \langle \nabla f_1(\boldsymbol{x}^k), \boldsymbol{x}^k - \hat{\boldsymbol{x}} \rangle. \end{aligned}$$

Substituting k_j for k and limiting j to ∞, we have, from Proposition 3.6 (ii),

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

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

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

Assumption 3.12. *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 3.13. *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 3.12 is nonrestrictive because many functions in [121], including the f_2 in numerical experiments, satisfy it. Thus, let us discuss the global convergence of Algorithm 1 under Assumption 3.12 by following the argument presented in [121]. Note that every limiting critical point is a limiting stationary point from the differentiability of f_2 under Assumption 3.12.

Theorem 3.14 (Global convergence of BPDCA under the local differentiability of f_2)**.** *Suppose that Assumptions 3.1, 3.2, 3.7, and 3.12 hold and that Ψ is a KL function. Let $\{\boldsymbol{x}^k\}_{k=0}^\infty$ be a sequence generated by BPDCA with $0 < \lambda L < 1$ for solving (3.1). Then, the following statements hold:*

(i) $\lim_{k \to \infty} \operatorname{dist}(\mathbf{0}_d, \partial \Psi(\boldsymbol{x}^k)) = 0$.

(ii) *The sequence $\{\boldsymbol{x}^k\}_{k=0}^\infty$ converges to a limiting stationary point of (3.1); moreover, $\sum_{k=1}^\infty \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2 < \infty$.*

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

$$\lim_{k \to \infty} \operatorname{dist}(\boldsymbol{x}^k, \Omega) = 0. \tag{3.13}$$

From Theorem 3.10 (iv), we also have $\Omega \subseteq \mathcal{X}$. Thus, for any $\mu > 0$, there exists $k_0 > 0$ such that $\operatorname{dist}(\boldsymbol{x}^k, \Omega) < \mu$ and $\boldsymbol{x}^k \in \mathcal{N}_0$ for any $k \geq k_0$, where $\mathcal{N}_0$ is defined in Assumption 3.12. As for $\mathcal{N}_0$, since Ω is compact from the boundedness of $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$, by decreasing μ, if needed, we can suppose without loss of generality that ∇f_2 is globally Lipschitz continuous on $\mathcal{N} := \{\boldsymbol{x} \in \mathcal{N}_0 \mid \operatorname{dist}(\boldsymbol{x}, \Omega) < \mu\}$.

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

$$\partial \Psi(\boldsymbol{x}^k) = \nabla f_1(\boldsymbol{x}^k) - \nabla f_2(\boldsymbol{x}^k) + \partial g(\boldsymbol{x}^k). \tag{3.14}$$

Moreover, considering the first-order optimality condition of subproblem (3.2), we see that, for any $k \geq k_0 + 1$,

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

since f_2 is $\mathcal{C}^1$ on $\mathcal{N}$ and $\boldsymbol{x}^{k-1} \in \mathcal{N}$ for any $k \geq k_0 + 1$. Using the above and (3.14), we see that

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

From the global Lipschitz continuity of $\nabla f_1, \nabla f_2$, and $\nabla \phi$, there exists $A_1 > 0$ such that

$$\operatorname{dist}(\mathbf{0}_d, \partial \Psi(\boldsymbol{x}^k)) \leq A_1 \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2, \tag{3.15}$$

where $k \geq k_0 + 1$. From Theorem 3.10 (iii), we conclude that $\lim_{k\to\infty} \operatorname{dist}(\mathbf{0}_d, \partial \Psi(\boldsymbol{x}^k)) = 0$.

(ii) From Theorem 3.10 (iv), it is sufficient to prove that $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ is convergent. Here, consider the case in which there exists a positive integer $k > 0$ such that $\Psi(\boldsymbol{x}^k) = \zeta$. From Proposition 3.6 (i) and Proposition 3.11 (i), the sequence $\{\Psi(\boldsymbol{x}^k)\}_{k=0}^{\infty}$ is non-increasing and converges to ζ. Hence, for any $\hat{k} \geq 0$, we have $\Psi(\boldsymbol{x}^{k+\hat{k}}) = \zeta$. Recalling (3.10), we conclude that there exists a positive scalar A_2 such that

$$\Psi(\boldsymbol{x}^{k-1}) - \Psi(\boldsymbol{x}^k) \geq A_2 \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2^2, \quad \forall k \in \mathbb{N}. \tag{3.16}$$

From (3.16), we obtain $\boldsymbol{x}^k = \boldsymbol{x}^{k+\hat{k}}$ for any $\hat{k} \geq 0$, which means that $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ is finitely convergent.

Next, consider the case where $\Psi(\boldsymbol{x}^k) > \zeta$ for all $k \geq 0$. Since $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ is bounded, Ω is a compact subset of $\operatorname{dom} \partial \Psi$ and $\Psi \equiv \zeta$ on Ω from Proposition 3.11 (ii). From Lemma 2.13 and since Ψ is a KL function, there exist a positive scalar $\epsilon > 0$ and a continuous concave function $\psi \in \Xi_\eta$ with $\eta > 0$ such that

$$\psi'(\Psi(\boldsymbol{x}) - \zeta) \cdot \operatorname{dist}(\mathbf{0}_d, \partial \Psi(\boldsymbol{x})) \geq 1, \tag{3.17}$$

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

From (3.13), there exists $k_1 > 0$ such that $\operatorname{dist}(\boldsymbol{x}^k, \Omega) < \epsilon$ for any $k \geq k_1$. Since $\{\Psi(\boldsymbol{x}^k)\}_{k=0}^{\infty}$ is non-increasing and converges to ζ, there exists $k_2 > 0$ such that $\zeta <$

$\Psi(\boldsymbol{x}^k) < \zeta + \eta$ for all $k \geq k_2$. Taking $\bar{k} = \max\{k_0 + 1, k_1, k_2\}$, then $\{\boldsymbol{x}^k\}_{k \geq \bar{k}}$ belongs to U. Hence, from (3.17), we obtain

$$\psi'(\Psi(\boldsymbol{x}^k) - \zeta) \cdot \operatorname{dist}(\mathbf{0}_d, \partial\Psi(\boldsymbol{x}^k)) \geq 1, \quad \forall k \geq \bar{k}. \tag{3.18}$$

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

$$\begin{aligned}
&[\psi(\Psi(\boldsymbol{x}^k) - \zeta) - \psi(\Psi(\boldsymbol{x}^{k+1}) - \zeta)] \cdot \operatorname{dist}(\mathbf{0}_d, \partial\Psi(\boldsymbol{x}^k)) \\
&\qquad \geq \psi'(\Psi(\boldsymbol{x}^k) - \zeta) \cdot \operatorname{dist}(\mathbf{0}_d, \partial\Psi(\boldsymbol{x}^k)) \cdot (\Psi(\boldsymbol{x}^k) - \Psi(\boldsymbol{x}^{k+1})) \\
&\qquad \geq \Psi(\boldsymbol{x}^k) - \Psi(\boldsymbol{x}^{k+1}) \\
&\qquad \geq A_2 \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2^2,
\end{aligned}$$

where the second inequality holds from (3.18) and the fact that $\{\Psi(\boldsymbol{x}^k)\}_{k=0}^{\infty}$ is non-increasing, and the last inequality holds from (3.16). From (3.15) and the above inequality, we obtain

$$\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2^2 \leq \frac{A_1}{A_2} \left(\psi(\Psi(\boldsymbol{x}^k) - \zeta) - \psi(\Psi(\boldsymbol{x}^{k+1}) - \zeta)\right) \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2. \tag{3.19}$$

Taking the square root of (3.19) and using the inequality of arithmetic and geometric means, we find that

$$\begin{aligned}
\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2 &\leq \sqrt{\frac{A_1}{A_2} \left(\psi(\Psi(\boldsymbol{x}^k) - \zeta) - \psi(\Psi(\boldsymbol{x}^{k+1}) - \zeta)\right)} \cdot \sqrt{\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2} \\
&\leq \frac{A_1}{2A_2} \left(\psi(\Psi(\boldsymbol{x}^k) - \zeta) - \psi(\Psi(\boldsymbol{x}^{k+1}) - \zeta)\right) + \frac{1}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2.
\end{aligned}$$

This shows that

$$\begin{aligned}
\frac{1}{2}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2 \leq \frac{A_1}{2A_2} &\left(\psi(\Psi(\boldsymbol{x}^k) - \zeta) - \psi(\Psi(\boldsymbol{x}^{k+1}) - \zeta)\right) \\
&+ \frac{1}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2 - \frac{1}{2}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2.
\end{aligned} \tag{3.20}$$

Summing (3.20) from $k = \bar{k}$ to ∞, we have

$$\sum_{k=\bar{k}}^{\infty} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2 \leq \frac{A_1}{A_2}\psi(\Psi(\boldsymbol{x}^{\bar{k}}) - \zeta) + \|\boldsymbol{x}^{\bar{k}} - \boldsymbol{x}^{\bar{k}-1}\|_2 < \infty,$$

which implies that $\sum_{k=1}^{\infty} \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2 < \infty$, *i.e.*, the sequence $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ is a Cauchy sequence. Thus, $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ converges to a limiting critical point of (3.1) from Theorem 3.10 (iv). Because every limiting critical point is a limiting stationary point from the differentiability of f_2, $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ converges to a limiting stationary point of (3.1). □

Next, suppose that Assumption 3.13 holds instead of Assumption 3.12. Here, we can show the global convergence of BPDCA by referring to [63, Theorem 3.4]. We will use subanalyticity instead of the KL property in the proof.

Theorem 3.15 (Global convergence of BPDCA under the local differentiability of g)**.** *Suppose that Assumptions 3.1, 3.2, 3.7, and 3.13 hold and that Ψ is subanalytic. Let $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ be a sequence generated by BPDCA with $0 < \lambda L < 1$ for solving (3.1). Then, the sequence $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ converges to a limiting critical point of (3.1); moreover, $\sum_{k=1}^{\infty} \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2 < \infty$.*

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 $\{\boldsymbol{\xi}^k\}_{k=0}^{\infty}$ on $\mathbb{R}^d$ be a sequence of subgradients of f_2. From Theorem 3.10 (i) and (ii), $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ and $\{\boldsymbol{\xi}^k\}_{k=0}^{\infty}$ are bounded. Let $\tilde{\boldsymbol{x}}$ be a limiting stationary point of $-\Psi$ and $B(\tilde{\boldsymbol{x}}, \epsilon_0)$ be an open ball with center $\tilde{\boldsymbol{x}}$ and radius $\epsilon_0 > 0$. Since ∇g is locally Lipschitz continuous and Assumption 3.7 (ii) holds, for $\lambda > 0$, there exist $\kappa_0 > 0$ and $\epsilon_0 > 0$ such that

$$\left\| \nabla \left(g + \frac{1}{\lambda}\phi \right)(\boldsymbol{u}) - \nabla \left(g + \frac{1}{\lambda}\phi \right)(\boldsymbol{v}) \right\|_2 \leq \kappa_0 \|\boldsymbol{u} - \boldsymbol{v}\|_2, \quad \forall \boldsymbol{u}, \boldsymbol{v} \in B(\tilde{\boldsymbol{x}}, \epsilon_0). \tag{3.21}$$

From Assumption 3.1 (v), $-\Psi$ is finite. Furthermore, recalling the continuity and subanalyticity of $-\Psi$ on $B(\tilde{\boldsymbol{x}}, \epsilon_0)$, we can apply [13, Theorem 3.1] to the subanalytic function $-\Psi$ and obtain $\nu_0 > 0$ and $\theta_0 \in [0, 1)$ such that

$$|\Psi(\boldsymbol{u}) - \zeta|^{\theta_0} \leq \nu_0 \|\hat{\boldsymbol{x}}\|_2, \quad \forall u \in B(\tilde{\boldsymbol{x}}, \epsilon_0), \quad \hat{\boldsymbol{x}} \in \partial(-\Psi)(\boldsymbol{u}), \tag{3.22}$$

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

Let Ω be the set of accumulation points of $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$. Since Ω is compact, Ω can be covered by a finite number of $B(\tilde{\boldsymbol{x}}_j, \epsilon_j)$ with $\tilde{\boldsymbol{x}}_j \in \Omega$ and $\epsilon_j > 0$, $j = 1, \ldots, p$. From Theorem 3.10 (iv), $\tilde{\boldsymbol{x}}_j \in \Omega$, $j = 1, \ldots, p$ are limiting critical points of (3.1). Hence, (3.21) with $\kappa_j > 0$ and $\epsilon_j > 0$ and (3.22) 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

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

From (3.13), there exists $k_1 > 0$ such that $\operatorname{dist}(\boldsymbol{x}^k, \Omega) < \epsilon$ for any $k \geq k_1$; hence, $\boldsymbol{x}^k \in \bigcup_{j=1}^{p} B(\tilde{\boldsymbol{x}}_j, \epsilon_j)$ for any $k \geq k_1$. From Theorem 3.10 (iii), letting $\bar{\epsilon} > 0$ be a sufficiently small constant, there exists $k_2 > 0$ such that $\|\boldsymbol{x}^k - \boldsymbol{x}^{k+1}\|_2 \leq \frac{\bar{\epsilon}}{2}$ for any $k \geq k_2$. Therefore, redefining $\bar{\epsilon}$, ϵ_j, $j = 1, \ldots, p$ and relabeling if necessary, we can assume without loss of generality that

$$\boldsymbol{x}^k \in \bigcup_{j=1}^{p} B\left(\tilde{\boldsymbol{x}}_j, \frac{\epsilon_j}{2}\right) \quad \text{and} \quad \|\boldsymbol{x}^k - \boldsymbol{x}^{k+1}\|_2 \leq \frac{\bar{\epsilon}}{2}, \quad \forall k \geq \bar{k},$$

where $\bar{\epsilon} = \min_{j=1,\ldots,p} \epsilon_j$ and $\bar{k} = \max\{k_1, k_2\}$, which implies $\boldsymbol{x}^k \in B(\tilde{\boldsymbol{x}}_{j_k}, \epsilon_{j_k}/2)$, $j_k \in \{1, \ldots, p\}$ and hence $\boldsymbol{x}^{k+1} \in B(\tilde{\boldsymbol{x}}_{j_k}, \epsilon_{j_k})$. Thus, from (3.21) and (3.22), we have

$$\left\| \nabla \left(g + \frac{1}{\lambda}\phi \right)(\boldsymbol{x}^k) - \nabla \left(g + \frac{1}{\lambda}\phi \right)(\boldsymbol{x}^{k+1}) \right\|_2 \leq \kappa \|\boldsymbol{x}^k - \boldsymbol{x}^{k+1}\|_2, \tag{3.23}$$

$$|\Psi(\boldsymbol{x}^k) - \zeta|^\theta \leq \nu \|\hat{\boldsymbol{x}}^k\|_2, \quad \hat{\boldsymbol{x}}^k \in \partial(-\Psi)(\boldsymbol{x}^k), \quad \forall k \geq \bar{k}, \tag{3.24}$$

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 (3.2), we find that

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

which implies

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

where we have used $\partial(-\Psi)(\boldsymbol{x}^k) = \partial_{\mathrm{c}} f_2(\boldsymbol{x}^k) - \nabla f_1(\boldsymbol{x}^k) - \nabla g(\boldsymbol{x}^k)$, which comes from the convexity of f_2. Using (3.23) and (3.24), we obtain, for all $k \geq \bar{k}$,

$$\begin{aligned} |\Psi(\boldsymbol{x}^k) - \zeta|^\theta &\leq \nu \left\| \nabla\left(g + \frac{1}{\lambda}\phi\right)(\boldsymbol{x}^k) - \nabla\left(g + \frac{1}{\lambda}\phi\right)(\boldsymbol{x}^{k+1}) \right\|_2 \\ &\leq \kappa\nu \|\boldsymbol{x}^k - \boldsymbol{x}^{k+1}\|_2. \end{aligned} \tag{3.25}$$

Since the function $t^{1-\theta}$ is concave on $[0, \infty)$, $\Psi(\boldsymbol{x}^k) - \zeta \geq 0$, (3.10), and (3.25), we find that, for all $k \geq \bar{k}$,

$$\begin{aligned} (\Psi(\boldsymbol{x}^k) - \zeta)^{1-\theta} - (\Psi(\boldsymbol{x}^{k+1}) - \zeta)^{1-\theta} &\geq (1-\theta)(\Psi(\boldsymbol{x}^k) - \zeta)^{-\theta}(\Psi(\boldsymbol{x}^k) - \Psi(\boldsymbol{x}^{k+1})) \\ &\geq \frac{1-\theta}{\kappa\nu\|\boldsymbol{x}^k - \boldsymbol{x}^{k+1}\|_2}\left(\frac{1}{\lambda} - L\right)\frac{\sigma}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^{k+1}\|_2^2 \\ &= \frac{(1-\theta)\sigma}{2\kappa\nu}\left(\frac{1}{\lambda} - L\right)\|\boldsymbol{x}^k - \boldsymbol{x}^{k+1}\|_2. \end{aligned} \tag{3.26}$$

Summing (3.26) from $k = \bar{k}$ to ∞ yields

$$\sum_{k=\bar{k}}^{\infty} \|\boldsymbol{x}^k - \boldsymbol{x}^{k+1}\|_2 \leq \frac{2\kappa\nu}{(1/\lambda - L)(1-\theta)\sigma}(\Psi(\boldsymbol{x}^{\bar{k}}) - \zeta)^{1-\theta} < \infty,$$

which implies that $\sum_{k=1}^{\infty} \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2 < \infty$, *i.e.*, the sequence $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ is a Cauchy sequence. Thus, $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ converges to a limiting critical point of (3.1) from Theorem 3.10 (iv). □

Finally, we show the rate of convergence in the following manner [5, 121].

Theorem 3.16 (Rate of convergence under the local differentiability of f_2)**.** *Suppose that Assumptions 3.1, 3.2, 3.7, and 3.12 hold. Let $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ be a sequence generated by BPDCA with $0 < \lambda L < 1$ for solving (3.1) and suppose that $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ converges to some $\tilde{\boldsymbol{x}} \in \mathcal{X}$. Suppose further that Ψ is a KL function with ϕ in the KL inequality (2.3) taking the form $\psi(s) = cs^{1-\theta}$ 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 $\boldsymbol{x}^k$ 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* $\|\boldsymbol{x}^k - \tilde{\boldsymbol{x}}\|_2 < c_1\eta^k$ *for* $k > k_1$;

(iii) *If* $\theta \in (\frac{1}{2}, 1)$, *then there exist* $c_2 > 0$ *and* $k_2 > 0$ *such that* $\|\boldsymbol{x}^k - \tilde{\boldsymbol{x}}\|_2 < c_2 k^{-\frac{1-\theta}{2\theta-1}}$ *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 $\Psi(\boldsymbol{x}^{k_0}) = \zeta$ by assuming to the contrary that $\Psi(\boldsymbol{x}^k) > \zeta$ for all $k > 0$ and showing a contradiction. The sequence $\{\Psi(\boldsymbol{x}^k)\}_{k=0}^{\infty}$ converges to ζ due to Proposition 3.11 (i). In addition, from the KL inequality (3.18) and $\psi'(\cdot) = c$, we can see that for all sufficiently large k,

$$\operatorname{dist}(\mathbf{0}_d, \partial\Psi(\boldsymbol{x}^k)) \geq \frac{1}{c},$$

which contradicts Theorem 3.14 (i). Therefore, there exists $k_0 > 0$ such that $\Psi(\boldsymbol{x}^{k_0}) = \zeta$. Since $\{\Psi(\boldsymbol{x}^k)\}_{k=0}^{\infty}$ is non-increasing and converges to ζ, we have $\Psi(\boldsymbol{x}^{k_0+\bar{k}}) = \zeta$ for all $\bar{k} \geq 0$. This, together with (3.16), leads us to conclude that there exists $k_0 > 0$ such that $\boldsymbol{x}^k$ is constant for $k > k_0$.

(ii–iii) Next, consider the case $\theta \in (0,1)$. If there exists $k_0 > 0$ such that $\Psi(\boldsymbol{x}^{k_0}) = \zeta$, then we can show that the sequence $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ 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 $\Psi(\boldsymbol{x}^k) > \zeta$ for all $k > 0$.

Define $R_k = \Psi(\boldsymbol{x}^k) - \zeta$ and $S_k = \sum_{j=k}^{\infty} \|\boldsymbol{x}^{j+1} - \boldsymbol{x}^j\|_2$, where S_k is well-defined due to Theorem 3.14 (ii). From (3.20), for any $k \geq \bar{k}$, where $\bar{k}$ is defined in (3.18), we obtain

$$\begin{aligned}
S_k &= 2\sum_{j=k}^{\infty} \frac{1}{2}\|\boldsymbol{x}^{j+1} - \boldsymbol{x}^j\|_2 \\
&\leq 2\sum_{j=k}^{\infty} \left[\frac{A_1}{2A_2}\left(\psi(\Psi(\boldsymbol{x}^j) - \zeta) - \psi(\Psi(\boldsymbol{x}^{j+1}) - \zeta)\right) + \frac{1}{2}\|\boldsymbol{x}^j - \boldsymbol{x}^{j-1}\|_2 - \frac{1}{2}\|\boldsymbol{x}^{j+1} - \boldsymbol{x}^j\|_2\right] \\
&\leq \frac{A_1}{A_2}\psi(\Psi(\boldsymbol{x}^k) - \zeta) + \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2 \\
&= \frac{A_1}{A_2}\psi(R_k) + S_{k-1} - S_k. \qquad (3.27)
\end{aligned}$$

On the other hand, since $\lim_{k\to\infty} \boldsymbol{x}^k = \tilde{\boldsymbol{x}}$ and $\{\Psi(\boldsymbol{x}^k)\}$ is non-increasing and converges to ζ, the KL inequality (3.18) with $\psi(s) = cs^{1-\theta}$ ensures that, for all sufficiently large k,

$$c(1-\theta)R_k^{-\theta}\operatorname{dist}(\mathbf{0}_d, \partial\Psi(\boldsymbol{x}^k)) \geq 1. \qquad (3.28)$$

From the definition of S_k and (3.15), we also have that, for all sufficiently large k,

$$\operatorname{dist}(\mathbf{0}_d, \partial\Psi(\boldsymbol{x}^k)) \leq A_1(S_{k-1} - S_k). \qquad (3.29)$$

Combining (3.28) and (3.29), 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 (3.27) and recalling $\psi(R_k) = cR_k^{1-\theta}$, we find that, for all sufficiently large k,

$$S_k \leq A_3(S_{k-1}-S_k)^{\frac{1-\theta}{\theta}} + S_{k-1} - S_k, \tag{3.30}$$

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 $\lim_{k\to\infty} \|\boldsymbol{x}^{k+1}-\boldsymbol{x}^k\|_2 = 0$ by Theorem 3.10 (iii), $\lim_{k\to\infty} S_{k-1} - S_k = 0$. From these considerations and (3.30), 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 $k \geq k_1$,

$$\|\boldsymbol{x}^k - \tilde{\boldsymbol{x}}\|_2 \leq \sum_{j=k}^{\infty} \|\boldsymbol{x}^{j+1} - \boldsymbol{x}^j\|_2 = S_k \leq S_{k_1-1} \left(\frac{A_3+1}{A_3+2}\right)^{k-k_1+1}.$$

(iii) For $\theta \in \left(\frac{1}{2}, 1\right)$, $\frac{1-\theta}{\theta} < 1$. From (3.30) and $\lim_{k\to\infty} S_{k-1} - S_k = 0$, there exists $k_2 > 0$ such that

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

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 $S_k^{\frac{\theta}{1-\theta}} \leq A_4(S_{k-1}-S_k)$, where $A_4 = (A_3+1)^{\frac{\theta}{1-\theta}}$. From [5, Theorem 2], we find that, for all sufficiently large k, there exists $A_5 > 0$ such that $S_k \leq A_5 k^{-\frac{1-\theta}{2\theta-1}}$. □

In Theorem 3.16, the parameter θ is called the KL exponent. Calculation of the KL exponent for first-order methods has been studied in [68]. Using Theorem 3.15, we can obtain another rate of convergence in the same way as in the proof of [5, Theorem 2] or [63, Theorem 3.5].

Theorem 3.17 (Rate of convergence under the local differentiability of g)**.** *Suppose that Assumptions 3.1, 3.2, 3.7, and 3.13 hold. Let $\{\boldsymbol{x}^k\}_{k=0}^\infty$ be a sequence generated by BPDCA with $0 < \lambda L < 1$ for solving (3.1) and suppose that $\{\boldsymbol{x}^k\}_{k=0}^\infty$ converges to some $\tilde{\boldsymbol{x}} \in \mathcal{X}$. Suppose further that Ψ is subanalytic. Let $\theta \in [0,1)$ be a Łojasiewicz exponent of $\tilde{\boldsymbol{x}}$. Then, the following statements hold:*

(i) *If $\theta = 0$, then there exists $k_0 > 0$ such that $\boldsymbol{x}^k$ 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 $\|\boldsymbol{x}^k - \tilde{\boldsymbol{x}}\|_2 < c_1\eta^k$ for $k > k_1$;*

(iii) *If $\theta \in (\frac{1}{2}, 1)$, then there exist $c_2 > 0$ and $k_2 > 0$ such that $\|\boldsymbol{x}^k - \tilde{\boldsymbol{x}}\|_2 < c_2 k^{-\frac{1-\theta}{2\theta-1}}$ for $k > k_2$.*

In this thesis, we call θ the KL exponent under the KL property, while we call θ the Łojasiewicz exponent under subanalyticity.

3.2 Bregman Proximal DC Algorithm with extrapolation

Algorithm 2, which we are proposing, is an acceleration of BPDCA that uses the extrapolation technique [10, 86, 87] to solve the DC optimization problem (3.1).

Algorithm 2 Bregman proximal DC algorithm with extrapolation (BPDCAe)

Input: $\phi \in \mathcal{G}(C)$ with $C = \operatorname{int}\operatorname{dom}\phi$ such that L-smad for the pair (f_1, ϕ) holds on C.

Initialization: $\boldsymbol{x}^0 = \boldsymbol{x}^{-1} \in C$, $t_{-1} = t_0 = 1$, $\rho \in (0, 1]$, and $0 < \lambda < 1/L$.

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

Compute

$$\begin{aligned} \beta_k &= \frac{t_{k-1} - 1}{t_k} \quad \text{with} \quad t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}, \\ \boldsymbol{y}^k &= \boldsymbol{x}^k + \beta_k(\boldsymbol{x}^k - \boldsymbol{x}^{k-1}). \end{aligned} \tag{3.31}$$

if $\boldsymbol{y}^k \notin C$ or $D_\phi(\boldsymbol{x}^k, \boldsymbol{y}^k) > \rho D_\phi(\boldsymbol{x}^{k-1}, \boldsymbol{x}^k)$ **then**

Set $\beta_k = 0$ with $t_{k-1} = t_k = 1$ and $\boldsymbol{y}^k = \boldsymbol{x}^k$.

end if

Take any $\boldsymbol{\xi}^k \in \partial_{\mathrm{c}} f_2(\boldsymbol{x}^k)$ and compute

$$\boldsymbol{x}^{k+1} = \operatorname*{argmin}_{\boldsymbol{y} \in \operatorname{cl} C} \left\{ \langle \nabla f_1(\boldsymbol{y}^k) - \boldsymbol{\xi}^k, \boldsymbol{y} - \boldsymbol{y}^k \rangle + g(\boldsymbol{y}) + \frac{1}{\lambda} D_\phi(\boldsymbol{y}, \boldsymbol{y}^k) \right\}. \tag{3.32}$$

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 t_k) given in [121] for acceleration. Accordingly, (3.31) guarantees that $\{\beta_k\}_{k=0}^{\infty} \subset [0, 1)$ and $\sup_{k \geq 0} \beta_k < 1$. These properties are needed to prove the global subsequential convergence of the iterates (see Theorem 3.23 (ii)). Algorithm 2 introduces a new *adaptive restart scheme*, which resets t_k and β_k whenever

$$D_\phi(\boldsymbol{x}^k, \boldsymbol{y}^k) > \rho D_\phi(\boldsymbol{x}^{k-1}, \boldsymbol{x}^k), \tag{3.33}$$

is satisfied for a fixed $\rho \in [0, 1)$. This adaptive restart scheme guarantees the non-increasing property for BPDCAe (see Lemma 3.21). In addition, we can enforce this reset every K iterations for a given positive integer K. In numerical experiments, we set $\{\beta_k\}_{k=0}^{\infty}$ as both the fixed and the adaptive restart schemes.

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

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

$$\mathcal{T}_\lambda(\boldsymbol{x}, \boldsymbol{y}) := \underset{\boldsymbol{u}\in \operatorname{cl} C}{\operatorname{argmin}} \left\{ \langle \nabla f_1(\boldsymbol{y}) - \boldsymbol{\xi}, \boldsymbol{u} - \boldsymbol{y}\rangle + g(\boldsymbol{u}) + \frac{1}{\lambda} D_\phi(\boldsymbol{u}, \boldsymbol{y}) \right\},$$

where $\boldsymbol{\xi} \in \partial_{\mathrm{c}} f_2(\boldsymbol{x})$. Similarly to the case of BPDCA, we make an Assumption 3.18 and can prove Lemma 3.19 for $\mathcal{T}_\lambda(\boldsymbol{x}, \boldsymbol{y}) \subset \operatorname{cl} C$.

Assumption 3.18. *For all $\boldsymbol{x}, \boldsymbol{y} \in C$ and $\lambda > 0$, we have*

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

Lemma 3.19. *Suppose that Assumptions 3.1 and 3.18 hold, and let $\boldsymbol{x}, \boldsymbol{y} \in C = \operatorname{int}\operatorname{dom}\phi$. Then, the set $\mathcal{T}_\lambda(\boldsymbol{x}, \boldsymbol{y})$ is a nonempty and compact subset of C for any $\lambda > 0$.*

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

3.2.1 Properties of BPDCAe

Inspired by [128], we introduce the auxiliary function,

$$H_M(\boldsymbol{x}, \boldsymbol{y}) = \Psi(\boldsymbol{x}) + M D_\phi(\boldsymbol{y}, \boldsymbol{x}), \quad M > 0.$$

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

Assumption 3.20. *The function g is convex.*

Under the adaptive restart scheme (3.33), we show the decreasing property of H_M.

Lemma 3.21. *Suppose that Assumptions 3.1, 3.18, and 3.20 hold. For any $\boldsymbol{x}^k$, $\boldsymbol{y}^k \in C = \operatorname{int}\operatorname{dom}\phi$ and any $\boldsymbol{x}^{k+1} \in C = \operatorname{int}\operatorname{dom}\phi$ defined by*

$$\boldsymbol{x}^{k+1} \in \underset{\boldsymbol{y}\in \operatorname{cl} C}{\operatorname{argmin}} \left\{ \langle \nabla f_1(\boldsymbol{y}^k) - \boldsymbol{\xi}^k, \boldsymbol{y} - \boldsymbol{y}^k\rangle + g(\boldsymbol{y}) + \frac{1}{\lambda} D_\phi(\boldsymbol{y}, \boldsymbol{y}^k) \right\}, \tag{3.34}$$

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

$$\lambda\Psi(\boldsymbol{x}^{k+1}) \le \lambda\Psi(\boldsymbol{x}^k) + D_\phi(\boldsymbol{x}^k, \boldsymbol{y}^k) - D_\phi(\boldsymbol{x}^k, \boldsymbol{x}^{k+1}) - (1 - \lambda L) D_\phi(\boldsymbol{x}^{k+1}, \boldsymbol{y}^k). \tag{3.35}$$

Furthermore, when $0 < \lambda L < 1$ and the sequence $\{\beta_k\}_{k=0}^\infty$ is given by the adaptive restart scheme (3.33),

$$\begin{aligned} H_M(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) \le H_M(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}) &- \left(\frac{1}{\lambda} - M\right) D_\phi(\boldsymbol{x}^k, \boldsymbol{x}^{k+1}) \\ &- \left(M - \frac{\rho}{\lambda}\right) D_\phi(\boldsymbol{x}^{k-1}, \boldsymbol{x}^k) - \left(\frac{1}{\lambda} - L\right) D_\phi(\boldsymbol{x}^{k+1}, \boldsymbol{y}^k). \end{aligned} \tag{3.36}$$

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

Proof. From the first-order optimality condition for (3.34), we obtain

$$\mathbf{0}_d \in \nabla f_1(\boldsymbol{y}^k) - \boldsymbol{\xi}^k + \partial_{\mathrm{c}} g(\boldsymbol{x}^{k+1}) + \frac{1}{\lambda}(\nabla\phi(\boldsymbol{x}^{k+1}) - \nabla\phi(\boldsymbol{y}^k)).$$

From the convexity of g, we find that

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

Using the three-point identity (2.2) of the Bregman distances,

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

we have

$$\begin{aligned} f_1(\boldsymbol{x}^k) - f_1(\boldsymbol{x}^{k+1}) + g(\boldsymbol{x}^k) - g(\boldsymbol{x}^{k+1}) &\geq f_1(\boldsymbol{x}^k) - f_1(\boldsymbol{x}^{k+1}) \\ &+ \langle -\nabla f_1(\boldsymbol{y}^k) + \boldsymbol{\xi}^k, \boldsymbol{x}^k - \boldsymbol{x}^{k+1}\rangle - \frac{1}{\lambda}(D_\phi(\boldsymbol{x}^k, \boldsymbol{y}^k) - D_\phi(\boldsymbol{x}^k, \boldsymbol{x}^{k+1}) - D_\phi(\boldsymbol{x}^{k+1}, \boldsymbol{y}^k)). \end{aligned}$$

From the convexity of f_1 and Lemma 2.10, we find that

$$\begin{aligned} &f_1(\boldsymbol{x}^k) - f_1(\boldsymbol{x}^{k+1}) - \langle \nabla f_1(\boldsymbol{y}^k), \boldsymbol{x}^k - \boldsymbol{x}^{k+1}\rangle \\ &= f_1(\boldsymbol{x}^k) - f_1(\boldsymbol{y}^k) - \langle \nabla f_1(\boldsymbol{y}^k), \boldsymbol{x}^k - \boldsymbol{y}^k\rangle - f_1(\boldsymbol{x}^{k+1}) + f_1(\boldsymbol{y}^k) + \langle \nabla f_1(\boldsymbol{y}^k), \boldsymbol{x}^{k+1} - \boldsymbol{y}^k\rangle \\ &\geq -LD_\phi(\boldsymbol{x}^{k+1}, \boldsymbol{y}^k). \end{aligned}$$

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

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

which implies inequality (3.35). If $\beta_k = 0$, then $\boldsymbol{y}^k = \boldsymbol{x}^k$ and $D_\phi(\boldsymbol{x}^k, \boldsymbol{y}^k) = 0$. If $\beta_k \neq 0$, since we chose the adaptive restart scheme, there is a $\rho \in [0, 1)$ satisfying $D_\phi(\boldsymbol{x}^k, \boldsymbol{y}^k) \leq \rho D_\phi(\boldsymbol{x}^{k-1}, \boldsymbol{x}^k)$. From the definition of $H_M(\boldsymbol{x}^k, \boldsymbol{x}^{k-1})$ and $0 < \lambda L < 1$, we have

$$\begin{aligned} H_M(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) &\leq H_M(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}) + \frac{1}{\lambda}D_\phi(\boldsymbol{x}^k, \boldsymbol{y}^k) - \left(\frac{1}{\lambda} - M\right) D_\phi(\boldsymbol{x}^k, \boldsymbol{x}^{k+1}) \\ &\quad - MD_\phi(\boldsymbol{x}^{k-1}, \boldsymbol{x}^k) - \left(\frac{1}{\lambda} - L\right) D_\phi(\boldsymbol{x}^{k+1}, \boldsymbol{y}^k) \\ &\leq H_M(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}) - \left(\frac{1}{\lambda} - M\right) D_\phi(\boldsymbol{x}^k, \boldsymbol{x}^{k+1}) \\ &\quad - \left(M - \frac{\rho}{\lambda}\right) D_\phi(\boldsymbol{x}^{k-1}, \boldsymbol{x}^k) - \left(\frac{1}{\lambda} - L\right) D_\phi(\boldsymbol{x}^{k+1}, \boldsymbol{y}^k), \end{aligned} \tag{3.37}$$

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

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

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

We can use Lemma 3.21 to prove Proposition 3.22.

Proposition 3.22. *Suppose that Assumptions 3.1, 3.18, and 3.20 hold. Let* $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ *be a sequence generated by BPDCAe with* $0 < \lambda L < 1$. *Assume that the auxiliary function* $H_M(\boldsymbol{x}^k, \boldsymbol{x}^{k-1})$ *satisfies* $\frac{\rho}{\lambda} \leq M \leq \frac{1}{\lambda}$ *for* $\rho \in [0, 1)$. *Then, the following statements hold:*

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

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

(iii) $\min_{1\leq k\leq n} D_\phi(\boldsymbol{x}^{k-1}, \boldsymbol{x}^k) \leq \frac{\lambda}{n(1-\rho)} (\Psi(\boldsymbol{x}^0) - \Psi_*)$, *where* $\Psi_* := v > -\infty$ (by Assumption 3.1 (v).

Proof. (i) The statement was proved in Lemma 3.21.

(ii) Modify (3.37) into

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

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

$$\begin{aligned}\sum_{k=1}^{n} D_\phi(\boldsymbol{x}^{k-1}, \boldsymbol{x}^k) = \sum_{k=0}^{n} D_\phi(\boldsymbol{x}^{k-1}, \boldsymbol{x}^k) &\leq \frac{\lambda\left(H_M(\boldsymbol{x}^0, \boldsymbol{x}^{-1}) - H_M(\boldsymbol{x}^{n+1}, \boldsymbol{x}^n)\right)}{1-\rho} \\ &\leq \frac{\lambda\left(\Psi(\boldsymbol{x}^0) - \Psi(\boldsymbol{x}^{n+1})\right)}{1-\rho} \\ &\leq \frac{\lambda\left(\Psi(\boldsymbol{x}^0) - \Psi_*\right)}{1-\rho},\end{aligned} \tag{3.38}$$

where the second inequality comes from $D_\phi(\boldsymbol{x}^{-1}, \boldsymbol{x}^0) = 0$, $\boldsymbol{x}^{-1} = \boldsymbol{x}^0$, and $D_\phi(\boldsymbol{x}^n, \boldsymbol{x}^{n+1}) \geq 0$. Note that $\boldsymbol{x}^{n+1} \in C$ by Assumption 3.18. Taking the limit as $n \to \infty$, we arrive at the former statement (ii). The latter statement follows directly from the former.

(iii) From (3.38), we immediately have

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

□

3.2.2 Convergence Analysis of BPDCAe

They follow arguments that are similar to their BPDCA counterparts.

Theorem 3.23 (Global subsequential convergence of BPDCAe)**.** *Suppose that Assumptions 3.1, 3.18, 3.7, and 3.20 hold. Let $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ be a sequence generated by BPDCAe with $0 < \lambda L < 1$ for solving (3.1). Assume that the auxiliary function $H_M(\boldsymbol{x}^k, \boldsymbol{x}^{k-1})$ satisfies $\frac{\rho}{\lambda} \leq M \leq \frac{1}{\lambda}$ for $\rho \in [0, 1)$. Then, the following statements hold:*

(i) *The sequence $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ is bounded.*

(ii) $\lim_{k\to\infty} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2 = 0$.

(iii) *Any accumulation point of $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ is a limiting critical point of (3.1).*

Proof. (i) Since $H_M(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}) \leq H_M(\boldsymbol{x}^0, \boldsymbol{x}^{-1})$ for all $k \in \mathbb{N}$ from Proposition 3.22 (i), with $\boldsymbol{x}^0 = \boldsymbol{x}^{-1}$, we obtain

$$\Psi(\boldsymbol{x}^k) \leq \Psi(\boldsymbol{x}^k) + MD_\phi(\boldsymbol{x}^{k-1}, \boldsymbol{x}^k) = H_M(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}) \leq H_M(\boldsymbol{x}^0, \boldsymbol{x}^{-1}) = \Psi(\boldsymbol{x}^0),$$

which shows that $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ is bounded due to Assumption 3.7 (iii).

(ii) From (3.36), we obtain

$$\begin{aligned} H_M(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}) - H_M(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) &\geq \left(\frac{1}{\lambda} - M\right) D_\phi(\boldsymbol{x}^k, \boldsymbol{x}^{k+1}) + \left(M - \frac{\rho}{\lambda}\right) D_\phi(\boldsymbol{x}^{k-1}, \boldsymbol{x}^k) \\ &\quad + \left(\frac{1}{\lambda} - L\right) D_\phi(\boldsymbol{x}^{k+1}, \boldsymbol{y}^k) \\ &\geq \frac{\sigma(1-\lambda L)}{2\lambda} \left(\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2^2 - \beta_k \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2^2\right), \end{aligned}$$

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

$$\begin{aligned} &\frac{\sigma(1-\lambda L)}{2\lambda} \left(\sum_{k=0}^{\infty} (1 - \beta_{k+1}) \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2^2 - \beta_1 \|\boldsymbol{x}^0 - \boldsymbol{x}^1\|_2^2\right) \\ &\leq H_M(\boldsymbol{x}^0, \boldsymbol{x}^{-1}) - \liminf_{n\to\infty} H_M(\boldsymbol{x}^{n+1}, \boldsymbol{x}^n) \\ &= \Psi(\boldsymbol{x}^0) - \liminf_{n\to\infty} \left(\Psi(\boldsymbol{x}^{n+1}) + MD_\phi(\boldsymbol{x}^n, \boldsymbol{x}^{n+1})\right) \\ &\leq \Psi(\boldsymbol{x}^0) - v < \infty, \end{aligned}$$

which shows that $\lim_{k\to\infty} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2 = 0$ due to $\frac{1}{\lambda} - L > 0$ and $\sup_{k>0} \beta_k < 1$.

(iii) Let $\tilde{\boldsymbol{x}}$ be an accumulation point of $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ and let $\{\boldsymbol{x}^{k_j}\}$ be a subsequence such that $\lim_{j\to\infty} \boldsymbol{x}^{k_j} = \tilde{\boldsymbol{x}}$. Then, from the first-order optimality condition of subproblem (3.32) under Assumption 3.18, we have

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

Therefore, we obtain

$$\boldsymbol{\xi}^{k_j} + \nabla f_1(\boldsymbol{x}^{k_j+1}) - \nabla f_1(\boldsymbol{y}^{k_j}) + \frac{1}{\lambda}\left(\nabla\phi(\boldsymbol{y}^{k_j}) - \nabla\phi(\boldsymbol{x}^{k_j+1})\right) \in \partial_c g(\boldsymbol{x}^{k_j+1}) + \nabla f_1(\boldsymbol{x}^{k_j+1}). \tag{3.39}$$

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

$$\left\|\nabla f_1(\boldsymbol{x}^{k_j+1}) - \nabla f_1(\boldsymbol{y}^{k_j}) + \frac{1}{\lambda}\left(\nabla\phi(\boldsymbol{y}^{k_j}) - \nabla\phi(\boldsymbol{x}^{k_j+1})\right)\right\|_2 \leq A_0\|\boldsymbol{x}^{k_j+1} - \boldsymbol{y}^{k_j}\|_2.$$

Therefore, using $\|\boldsymbol{x}^{k_j+1} - \boldsymbol{x}^{k_j}\|_2 \to 0$ and $\|\boldsymbol{x}^{k_j} - \boldsymbol{x}^{k_j-1}\|_2 \to 0$, we obtain

$$\nabla f_1(\boldsymbol{x}^{k_j+1}) - \nabla f_1(\boldsymbol{y}^{k_j}) + \frac{1}{\lambda}\left(\nabla\phi(\boldsymbol{y}^{k_j}) - \nabla\phi(\boldsymbol{x}^{k_j+1})\right) \to \mathbf{0}_d. \tag{3.40}$$

Note that the sequence $\{\boldsymbol{\xi}^{k_j}\}$ is bounded as shown in Theorem 3.10 (ii), and the sequence $\{\boldsymbol{x}^{k_j}\}$ is bounded and converges to $\tilde{\boldsymbol{x}}$. Thus, by taking the limit as $j \to \infty$ or, more precisely, its subsequence, we can assume without loss of generality that $\lim_{j\to\infty}\boldsymbol{\xi}^{k_j} =: \tilde{\boldsymbol{\xi}}$ exists, which belongs to $\partial_{\mathrm{c}} f_2(\tilde{\boldsymbol{x}})$ since f_2 is continuous. Using this and (3.40), we take the limit of (3.39). Invoking $\|\boldsymbol{x}^{k_j+1} - \boldsymbol{x}^{k_j}\|_2 \to 0$ and the continuity of g and ∇f_1, we obtain $\tilde{\boldsymbol{\xi}} \in \partial_{\mathrm{c}} g(\tilde{\boldsymbol{x}}) + \nabla f_1(\tilde{\boldsymbol{x}})$. Therefore, $\mathbf{0}_d \in \partial_{\mathrm{c}} g(\tilde{\boldsymbol{x}}) + \nabla f_1(\tilde{\boldsymbol{x}}) - \partial_{\mathrm{c}} f_2(\tilde{\boldsymbol{x}})$, which shows that $\tilde{\boldsymbol{x}}$ is a limiting critical point of (3.1). □

Proposition 3.24. *Suppose that Assumptions 3.1, 3.18, 3.7, and 3.20 hold. Let $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ be a sequence generated by BPDCAe with $0 < \lambda L < 1$ for solving (3.1) and $\frac{\rho}{\lambda} \leq M \leq \frac{1}{\lambda}$ for $\rho \in [0, 1)$. Then, the following statements hold:*

(i) $\zeta := \lim_{k\to\infty}\Psi(\boldsymbol{x}^k)$ *exists.*

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

Proof. (i) From Assumption 3.1 (v) and Proposition 3.22 (i), $\{H_M(\boldsymbol{x}^k, \boldsymbol{x}^{k-1})\}_{k=0}^{\infty}$ is bounded from below and non-increasing. Consequently, using $\lim_{k\to\infty} D_\phi(\boldsymbol{x}^{k-1}, \boldsymbol{x}^k) = 0$ from Proposition 3.22 (ii), we obtain $\lim_{k\to\infty} H_M(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}) = \lim_{k\to\infty}\Psi(\boldsymbol{x}^k) =: \zeta$.

(ii) Take any $\hat{\boldsymbol{x}} \in \Omega$, that is $\lim_{j\to\infty}\boldsymbol{x}^{k_j} = \hat{\boldsymbol{x}}$. From (3.32), it follows that

$$\begin{aligned} &g(\boldsymbol{x}^k) + \langle\nabla f_1(\boldsymbol{y}^{k-1}) - \boldsymbol{\xi}^{k-1}, \boldsymbol{x}^k - \boldsymbol{y}^{k-1}\rangle + \frac{1}{\lambda}D_\phi(\boldsymbol{x}^k, \boldsymbol{y}^{k-1}) \\ &\quad\leq g(\hat{\boldsymbol{x}}) + \langle\nabla f_1(\boldsymbol{y}^{k-1}) - \boldsymbol{\xi}^{k-1}, \hat{\boldsymbol{x}} - \boldsymbol{y}^{k-1}\rangle + \frac{1}{\lambda}D_\phi(\hat{\boldsymbol{x}}, \boldsymbol{y}^{k-1}). \end{aligned}$$

From the above inequality and the fact that f_1 is convex at $\boldsymbol{x}^k$, we obtain

$$\begin{aligned} g(\boldsymbol{x}^k) + f_1(\boldsymbol{x}^k) &\leq g(\hat{\boldsymbol{x}}) + \langle\nabla f_1(\boldsymbol{y}^{k-1}) - \boldsymbol{\xi}^{k-1}, \hat{\boldsymbol{x}} - \boldsymbol{x}^k\rangle + \frac{1}{\lambda}D_\phi(\hat{\boldsymbol{x}}, \boldsymbol{y}^{k-1}) - \frac{1}{\lambda}D_\phi(\boldsymbol{x}^k, \boldsymbol{y}^{k-1}) \\ &\quad + f_1(\hat{\boldsymbol{x}}) + \langle\nabla f_1(\boldsymbol{x}^k), \boldsymbol{x}^k - \hat{\boldsymbol{x}}\rangle \\ &\leq g(\hat{\boldsymbol{x}}) + \langle\nabla f_1(\boldsymbol{y}^{k-1}) - \boldsymbol{\xi}^{k-1}, \hat{\boldsymbol{x}} - \boldsymbol{x}^k\rangle + \frac{1}{\lambda}D_\phi(\hat{\boldsymbol{x}}, \boldsymbol{y}^{k-1}) + \frac{1}{\lambda}D_\phi(\boldsymbol{y}^{k-1}, \hat{\boldsymbol{x}}) \\ &\quad + f_1(\hat{\boldsymbol{x}}) + \langle\nabla f_1(\boldsymbol{x}^k), \boldsymbol{x}^k - \hat{\boldsymbol{x}}\rangle, \end{aligned} \tag{3.41}$$

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

$$\lim_{j\to\infty}\left(D_\phi(\hat{\boldsymbol{x}}, \boldsymbol{y}^{k_j-1}) + D_\phi(\boldsymbol{y}^{k_j-1}, \hat{\boldsymbol{x}})\right) \leq \lim_{j\to\infty} \|\nabla\phi(\boldsymbol{y}^{k_j-1}) - \nabla\phi(\hat{\boldsymbol{x}})\|_2 \|\boldsymbol{y}^{k_j-1} - \hat{\boldsymbol{x}}\|_2 = 0.$$

Substituting k_j for k in (3.41) and limiting j to ∞, we have, from Proposition 3.22 (ii),

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

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

Since $H_M(\boldsymbol{x}, \boldsymbol{y})$ has a Bregman distance term, the subdifferential of $H_M(\boldsymbol{x}, \boldsymbol{y})$ has a $\nabla\phi$ term. To prove Theorem 3.26, we should additionally suppose that there is a bounded subdifferential of the gradient $\nabla\phi$ [128].

Assumption 3.25. *For any bounded subset $S \subset \mathbb{R}^d$ and any point $\boldsymbol{x} \in S$, there exists $A > 0$ such that $\|\mathcal{T}(\boldsymbol{u})\|_2 \leq A\|\boldsymbol{u}\|_2$ for some $\mathcal{T} : \mathbb{R}^d \to \mathbb{R}^d$ such that $\mathcal{T}(\boldsymbol{u}) \in \partial(\nabla\phi(\boldsymbol{x}))(\boldsymbol{u})$ for any $\boldsymbol{u} \in S$.*

The subdifferential $\partial(\nabla\phi(\boldsymbol{x}))(\boldsymbol{u})$ is given by (2.1) (see also [80, Section 1.3.5]). We can prove the following theorems by supposing the KL property or the subanalyticity of the auxiliary function $H_M(\boldsymbol{x}, \boldsymbol{y})$ in relation to $\boldsymbol{x}$ and $\boldsymbol{y}$.

Theorem 3.26 (Global convergence of BPDCAe under the local differentiability of f_2)**.** *Suppose that Assumptions 3.1, 3.18, 3.7, 3.12, 3.20, and 3.25 hold and that the auxiliary function $H_M(\boldsymbol{x}, \boldsymbol{y})$ is a KL function satisfying $\frac{\rho}{\lambda} \leq M \leq \frac{1}{\lambda}$ for $\rho \in [0, 1)$. Let $\{\boldsymbol{x}^k\}_{k=0}^\infty$ be a sequence generated by BPDCAe with $0 < \lambda L < 1$ for solving (3.1). Then, the following statements hold:*

(i) $\lim_{k\to\infty} \operatorname{dist}((\mathbf{0}_d, \mathbf{0}_d), \partial H_M(\boldsymbol{x}^k, \boldsymbol{x}^{k-1})) = 0$.

(ii) *The set of accumulation points of $\{(\boldsymbol{x}^k, \boldsymbol{x}^{k-1})\}_{k=0}^\infty$ is $\Upsilon := \{(\boldsymbol{x}, \boldsymbol{x}) \mid x \in \Omega\}$ and $H_M \equiv \zeta$ on Υ, where Ω is the set of accumulation points of $\{\boldsymbol{x}^k\}_{k=0}^\infty$.*

(iii) *The sequence $\{\boldsymbol{x}^k\}_{k=0}^\infty$ converges to a limiting stationary point of (3.1); moreover, $\sum_{k=1}^\infty \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2 < \infty$.*

Proof. (i) Let $\mu > 0$, $k_0 > 0$, $\mathcal{N}_0$, and $\mathcal{N} := \{x \in \mathcal{N}_0 \mid \operatorname{dist}(\boldsymbol{x}, \Omega) < \mu\}$ as defined in the proof of Theorem 3.14 (i).

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

$$\partial H_M(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}) = \nabla f_1(\boldsymbol{x}^k) - \nabla f_2(\boldsymbol{x}^k) + \partial_c g(\boldsymbol{x}^k) - M\partial(\nabla\phi(\boldsymbol{x}^k))(\boldsymbol{x}^{k-1} - \boldsymbol{x}^k). \quad (3.42)$$

Moreover, considering the first-order optimality condition of subproblem (3.32), for any $k \geq k_0 + 1$, we have

$$\frac{1}{\lambda}\left(\nabla\phi(\boldsymbol{y}^{k-1}) - \nabla\phi(\boldsymbol{x}^k)\right) - \nabla f_1(\boldsymbol{y}^{k-1}) + \nabla f_2(\boldsymbol{x}^{k-1}) \in \partial_{\mathrm{c}} g(\boldsymbol{x}^k),$$

since f_2 is $\mathcal{C}^1$ on $\mathcal{N}$ and $\boldsymbol{x}^{k-1} \in \mathcal{N}$ whenever $k \geq k_0 + 1$. Using the above relation and (3.42), for $\mathcal{U}^k(\boldsymbol{x}^k - \boldsymbol{x}^{k-1}) \in \partial(\nabla\phi(\boldsymbol{x}^k))(\boldsymbol{x}^k - \boldsymbol{x}^{k-1})$ with some $\mathcal{U}^k : \mathbb{R}^d \to \mathbb{R}^d$ by Assumption 3.25, we also obtain

$$\begin{aligned}\frac{1}{\lambda}\left(\nabla\phi(\boldsymbol{y}^{k-1}) - \nabla\phi(\boldsymbol{x}^k)\right) + \nabla f_1(\boldsymbol{x}^k) - \nabla f_1(\boldsymbol{y}^{k-1}) \\ + \nabla f_2(\boldsymbol{x}^{k-1}) - \nabla f_2(\boldsymbol{x}^k) + M\mathcal{U}^k(\boldsymbol{x}^k - \boldsymbol{x}^{k-1}) \in \partial H_M(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}).\end{aligned}$$

Due to the global Lipschitz continuity of ∇f_1, ∇f_2, and $\nabla\phi$ on $\mathcal{N}_0$, and Assumption 3.25, we see that there exist $A_0 > 0$, $A_1 > 0$, and $A_2 > 0$ such that

$$\begin{aligned}\operatorname{dist}((\mathbf{0}_d, \mathbf{0}_d), \partial H_M(\boldsymbol{x}^k, \boldsymbol{x}^{k-1})) &\leq A_0\|\boldsymbol{x}^k - \boldsymbol{y}^{k-1}\|_2 + A_1\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2 \\ &\leq A_2\left(\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2 + \|\boldsymbol{x}^{k-1} - \boldsymbol{x}^{k-2}\|_2\right),\end{aligned}$$

where $k \geq k_0 + 1$. Since $\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2 \to 0$ and $\|\boldsymbol{x}^{k-1} - \boldsymbol{x}^{k-2}\|_2 \to 0$, we conclude the claim (i).

(ii) Suppose that $\hat{\boldsymbol{x}} \in \Omega$, $\boldsymbol{x}^{k_j} \to \hat{\boldsymbol{x}}$, and $\boldsymbol{x}^{k_j-1} \to \hat{\boldsymbol{x}}$ as in Proposition 3.24 (ii). Therefore, the set of accumulation points of $\{(\boldsymbol{x}^k, \boldsymbol{x}^{k-1})\}_{k=0}^{\infty}$ is Υ. From Propositions 3.22 and 3.24,

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

Additionally, from Proposition 3.24 (ii), for any $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}) \in \Upsilon$, $\hat{\boldsymbol{x}} \in \Omega$, we have $H_M(\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}) = \Psi(\hat{\boldsymbol{x}}) = \zeta$. Since $\hat{\boldsymbol{x}}$ is arbitrary, we conclude that $H_M \equiv \zeta$ on Υ.

(iii) The proof is similar to Theorem 3.14 (ii). □

Theorem 3.27 (Global convergence of BPDCAe under the local differentiability of g)**.** *Suppose that Assumptions 3.1, 3.18, 3.7, 3.13, 3.20, and 3.25 hold and that the auxiliary function $H_M(\boldsymbol{x}, \boldsymbol{y})$ is subanalytic satisfying $\frac{\rho}{\lambda} \leq M \leq \frac{1}{\lambda}$ for $\rho \in [0, 1)$. Let $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ be a sequence generated by BPDCAe with $0 < \lambda L < 1$ for solving (3.1). Then, the sequence $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ converges to a limiting critical point of (3.1); moreover, $\sum_{k=1}^{\infty}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2 < \infty$.*

Proof. Let k_1, κ_i, ν_i, and θ_i be defined similarly to the proof of Theorem 3.15. Using the differentiability of g and [13, Theorem 3.1], we have

$$\left\|\nabla g(\boldsymbol{x}^k) - \nabla g(\boldsymbol{x}^{k+1})\right\|_2 \leq \kappa\|\boldsymbol{x}^k - \boldsymbol{x}^{k+1}\|_2, \tag{3.43}$$

$$|H_M(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}) - \zeta|^\theta \leq \nu\|\hat{\boldsymbol{x}}^k\|_2, \quad \hat{\boldsymbol{x}}^k \in \partial(-H)(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}), \quad \forall k \geq k_1 + 1, \tag{3.44}$$

where $\zeta = H_M(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}) = \Psi(\tilde{\boldsymbol{x}})$, $\tilde{\boldsymbol{x}} \in \Omega$, $\kappa = \max_{j=1,\ldots,p}\kappa_i$, $\nu = \max_{j=1,\ldots,p}\nu_i$, and $\theta = \max_{j=1,\ldots,p}\theta_i$. From (3.32), we obtain

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

which implies, for $\hat{\boldsymbol{x}}^k \in \partial(-H_M)(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}) = \partial_{\mathrm{c}} f_2(\boldsymbol{x}^k) + M\partial(\nabla\phi(\boldsymbol{x}^k))(\boldsymbol{x}^{k-1} - \boldsymbol{x}^k) - \nabla f_1(\boldsymbol{x}^k) - \nabla g(\boldsymbol{x}^k)$ and some $\mathcal{U}^k(\boldsymbol{x}^{k-1} - \boldsymbol{x}^k) \in \partial(\nabla\phi(\boldsymbol{x}^k))(\boldsymbol{x}^{k-1} - \boldsymbol{x}^k)$ with $\mathcal{U}^k : \mathbb{R}^d \to \mathbb{R}^d$,

$$\begin{aligned}\hat{\boldsymbol{x}}^k &= \boldsymbol{\xi}^k + M\mathcal{U}^k(\boldsymbol{x}^{k-1} - \boldsymbol{x}^k) - \nabla f_1(\boldsymbol{x}^k) - \nabla g(\boldsymbol{x}^k)\\ &= \nabla g(\boldsymbol{x}^{k+1}) - \nabla g(\boldsymbol{x}^k) + \nabla f_1(\boldsymbol{y}^k) - \nabla f_1(\boldsymbol{x}^k)\\ &\quad + \frac{1}{\lambda}\left(\nabla\phi(\boldsymbol{x}^{k+1}) - \nabla\phi(\boldsymbol{y}^k)\right) + M\mathcal{U}^k(\boldsymbol{x}^{k-1} - \boldsymbol{x}^k).\end{aligned}$$

Using (3.43), (3.44), Assumptions 3.7, and 3.25, we obtain $C > 0$ such that

$$\begin{aligned}|H_M(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}) - \zeta|^\theta &\le \nu\|\hat{\boldsymbol{x}}^k\|_2\\ &\le C(\|\boldsymbol{x}^k - \boldsymbol{x}^{k+1}\|_2 + \|\boldsymbol{x}^{k-1} - \boldsymbol{x}^k\|_2), \quad \forall k \ge k_1 + 1,\end{aligned}$$

where the second inequality comes from $\nabla\phi(\boldsymbol{x}^{k+1}) - \nabla\phi(\boldsymbol{y}^k) = \nabla\phi(\boldsymbol{x}^{k+1}) - \nabla\phi(\boldsymbol{x}^k) + \nabla\phi(\boldsymbol{x}^k) - \nabla\phi(\boldsymbol{y}^k)$. The rest of the proof is similar to Theorem 3.15. □

Finally, we have theorems regarding the convergence rate of BPDCAe, whose proof is almost identical to Theorems 3.16 and 3.17. Note that the KL exponent (or the Łojasiewicz exponent) of the auxiliary function H_M is equal to that of the objective function Ψ from [124, Lemma 5.1].

Theorem 3.28 (Rate of convergence under the local differentiability of f_2)**.** *Suppose that Assumptions 3.1, 3.18, 3.7, 3.12, 3.20, and 3.25 hold. Let $\{\boldsymbol{x}^k\}_{k=0}^\infty$ be a sequence generated by BPDCAe with $0 < \lambda L < 1$ for solving (3.1) and suppose that $\{\boldsymbol{x}^k\}_{k=0}^\infty$ converges to some $\tilde{\boldsymbol{x}} \in \mathcal{X}$. Suppose further that the auxiliary function $H_M(\boldsymbol{x}, \boldsymbol{y})$ satisfying $\frac{\rho}{\lambda} \le M \le \frac{1}{\lambda}$ for $\rho \in [0, 1)$ is a KL function with ψ in the KL inequality (2.3) taking the form $\psi(s) = cs^{1-\theta}$ 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 $\boldsymbol{x}^k$ 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 $\|\boldsymbol{x}^k - \tilde{\boldsymbol{x}}\|_2 < c_1\eta^k$ for $k > k_1$;*

(iii) *If $\theta \in (\frac{1}{2}, 1)$, then there exist $c_2 > 0$ and $k_2 > 0$ such that $\|\boldsymbol{x}^k - \tilde{\boldsymbol{x}}\|_2 < c_2 k^{-\frac{1-\theta}{2\theta-1}}$ for $k > k_2$.*

Theorem 3.29 (Rate of convergence under the local differentiability of g)**.** *Suppose that Assumptions 3.1, 3.18, 3.7, 3.13, 3.20, and 3.25 hold. Let $\{\boldsymbol{x}^k\}_{k=0}^\infty$ be a sequence generated by BPDCAe with $0 < \lambda L < 1$ for solving (3.1) and suppose that $\{\boldsymbol{x}^k\}_{k=0}^\infty$ converges to some $\tilde{\boldsymbol{x}} \in \mathcal{X}$. Suppose further that the auxiliary function $H_M(\boldsymbol{x}, \boldsymbol{y})$ 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 $\tilde{\boldsymbol{x}}$. Then, the following statements hold:*

(i) *If $\theta = 0$, then there exists $k_0 > 0$ such that $\boldsymbol{x}^k$ 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 $\|\boldsymbol{x}^k - \tilde{\boldsymbol{x}}\|_2 < c_1\eta^k$ for $k > k_1$;*

(iii) *If $\theta \in (\frac{1}{2}, 1)$, then there exist $c_2 > 0$ and $k_2 > 0$ such that $\|\boldsymbol{x}^k - \tilde{\boldsymbol{x}}\|_2 < c_2 k^{-\frac{1-\theta}{2\theta-1}}$ for $k > k_2$.*

3.3 Hybrid Bregman Proximal DC Algorithm

Instead of (3.1), we consider the following block DC optimization problem:

$$\min_{(\boldsymbol{x},\boldsymbol{y})\in \operatorname{cl} C_1\times \operatorname{cl} C_2} \quad \Psi_{\mathrm{B}}(\boldsymbol{x},\boldsymbol{y}) := f_1(\boldsymbol{x},\boldsymbol{y}) - f_2(\boldsymbol{x},\boldsymbol{y}) + g_1(\boldsymbol{x}) - g_2(\boldsymbol{x}) + h(\boldsymbol{y}), \tag{3.45}$$

where $f_1, f_2 : \mathbb{C}^{d_1}\times\mathbb{R}^{d_2} \to (-\infty,+\infty]$ are real-valued functions of complex variables $\boldsymbol{x}\in\operatorname{cl} C_1$ and real variables $\boldsymbol{y}\in\operatorname{cl} C_2$, $g_1, g_2 : \mathbb{C}^{d_1}\to(-\infty,+\infty]$ are real-valued convex functions of complex variables $\boldsymbol{x}\in\operatorname{cl} C_1$, $h : \mathbb{R}^{d_2}\to(-\infty,+\infty]$ is a convex function of real variables $\boldsymbol{y}\in\operatorname{cl} C_2$, and $C_1\subset\mathbb{C}^{d_1}$ and $C_2\subset\mathbb{R}^{d_2}$ are nonempty open convex sets. We assume the following assumption.

Assumption 3.30.

(i) $\phi\in\mathcal{G}(C_1)$ *with* $\operatorname{cl} C_1 = \operatorname{cl}\operatorname{dom}\phi$.

(ii) $f_1, f_2 : \mathbb{C}^{d_1}\times\mathbb{R}^{d_2}\to(-\infty,+\infty]$ *are proper with* $\operatorname{dom}\phi\subset\operatorname{dom}(f_1+g_1)$, *which are* $\mathcal{C}^1$ *on* $C_1\times C_2$. *Additionally,* $f_1(\cdot,\boldsymbol{y})$, $f_2(\cdot,\boldsymbol{y})$, *and* $f(\boldsymbol{x},\cdot) := f_1(\boldsymbol{x},\cdot) - f_2(\boldsymbol{x},\cdot)$ *are convex.*

(iii) $g_1, g_2 : \mathbb{C}^{d_1}\to(-\infty,+\infty]$ *and* $h : \mathbb{R}^{d_2}\to(-\infty,+\infty]$ *are proper, lower semicontinuous, and convex with* $\operatorname{dom} g_1\cap C_1\neq\emptyset$ *and* $\operatorname{dom} h\cap C_2\neq\emptyset$, *respectively.*

(iv) $v_{\mathrm{B}} := \inf_{(\boldsymbol{x},\boldsymbol{y})\in\operatorname{cl} C_1\times\operatorname{cl} C_2}\Psi_{\mathrm{B}}(\boldsymbol{x},\boldsymbol{y}) > -\infty$.

(v) *For any* $\lambda>0$, $\lambda g_1+\phi$ *is supercoercieve, that is,*

$$\lim_{\|\boldsymbol{u}\|_2\to\infty}\frac{\lambda g_1(\boldsymbol{u})+\phi(\boldsymbol{u})}{\|\boldsymbol{u}\|_2} = \infty.$$

Under Assumption 3.30 (ii), $\nabla f = \nabla f_1 - \nabla f_2$ for $f := f_1 - f_2$ and $\Psi_{\mathrm{B}}(\boldsymbol{x},\cdot)$ is convex. The hybrid Bregman proximal DC algorithm (HBPDA) is listed as Algorithm 3.

Algorithm 3 Hybrid Bregman proximal DC algorithm (HBPDCA)

Input: $\phi\in\mathcal{G}(C_1)$ with $C_1 = \operatorname{int}\operatorname{dom}\phi$ such that the $L(\boldsymbol{y})$-smad property for the pair (f_1,ϕ) holds on C_1 for $\boldsymbol{y}\in\operatorname{cl} C_2$.
Initialization: $(\boldsymbol{x}^0,\boldsymbol{y}^0)\in C_1\times C_2$.
for $k=0,1,2,\ldots,$ **do**
Take any $\boldsymbol{\xi}^k\in\partial_{\mathrm{c}} g_2(\boldsymbol{x}^k)$ and compute $\lambda^k = 1/L(\boldsymbol{y}^k)$ and

$$\boldsymbol{x}^{k+1} = \operatorname*{argmin}_{\boldsymbol{x}\in\operatorname{cl} C_1}\left\{2\operatorname{Re}\langle\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^k,\boldsymbol{y}^k) - \boldsymbol{\xi}^k, \boldsymbol{x}-\boldsymbol{x}^k\rangle + g_1(\boldsymbol{x}) + \frac{1}{\lambda^k}D_\phi(\boldsymbol{x},\boldsymbol{x}^k)\right\}, \tag{3.46}$$

$$\boldsymbol{y}^{k+1} = \operatorname*{argmin}_{\boldsymbol{y}\in\operatorname{cl} C_2}\left\{f(\boldsymbol{x}^{k+1},\boldsymbol{y}) + h(\boldsymbol{y})\right\}. \tag{3.47}$$

end for

Note that the $L(\boldsymbol{y})$-smad property depends on $\boldsymbol{y} \in \operatorname{cl} C_2$. For all fixed $\boldsymbol{y}^k \in \operatorname{cl} C_2$, HBPDCA corresponds to BPDCA. Since $\Psi_{\mathrm{B}}(\boldsymbol{x}, \cdot)$ is convex, (3.47) is a convex optimization problem. Since HBPDCA has the convex subproblem, it is different from the unified Bregman alternating minimization algorithm [48].

Remark 3.31. *Because we use existing properties and algorithms for convex optimization, the variable $\boldsymbol{y}$ belongs to $C_2 \subset \mathbb{R}^{d_2}$. It would be possible to extend $\boldsymbol{y}$ to complex variables.*

3.3.1 Properties of HBPDCA

We obtain the sufficiently decreasing property of HBPDCA.

Lemma 3.32. *Suppose that Assumptions 3.2 and 3.30 hold. For any $\boldsymbol{x} \in C_1 = \operatorname{int} \operatorname{dom} \phi$ and $\boldsymbol{y} \in C_2$ and any $(\boldsymbol{x}^+, \boldsymbol{y}^+) \in C_1 \times C_2$ defined by*

$$\boldsymbol{x}^+ \in \operatorname*{argmin}_{\boldsymbol{u} \in \operatorname{cl} C_1} \left\{ 2 \operatorname{Re} \langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{\xi}, \boldsymbol{u} - \boldsymbol{x} \rangle + g_1(\boldsymbol{u}) + \frac{1}{\lambda} D_\phi(\boldsymbol{u}, \boldsymbol{x}) \right\}, \tag{3.48}$$

$$\boldsymbol{y}^+ \in \operatorname*{argmin}_{\boldsymbol{v} \in \operatorname{cl} C_2} \left\{ f(\boldsymbol{x}^+, \boldsymbol{v}) + h(\boldsymbol{v}) \right\}, \tag{3.49}$$

where $\boldsymbol{\xi} \in \partial_{\mathrm{c}} g_2(\boldsymbol{x})$ and $\lambda > 0$, it holds that

$$\lambda \Psi_{\mathrm{B}}(\boldsymbol{x}^+, \boldsymbol{y}^+) \leq \lambda \Psi_{\mathrm{B}}(\boldsymbol{x}, \boldsymbol{y}) - (1 - \lambda L) D_\phi(\boldsymbol{x}^+, \boldsymbol{x}). \tag{3.50}$$

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

Proof. From Lemma 3.5, we obtain

$$\lambda \Psi_{\mathrm{B}}(\boldsymbol{x}^+, \boldsymbol{y}) \leq \lambda \Psi_{\mathrm{B}}(\boldsymbol{x}, \boldsymbol{y}) - (1 - \lambda L) D_\phi(\boldsymbol{x}^+, \boldsymbol{x}),$$

for $\Psi_{\mathrm{B}}(\cdot, \boldsymbol{y}) = f_1(\cdot, \boldsymbol{y}) - f_2(\cdot, \boldsymbol{y}) + g_1(\cdot) - g_2(\cdot) + h(\boldsymbol{y})$. From the global optimality of $\boldsymbol{y}^+$, we have

$$f(\boldsymbol{x}^+, \boldsymbol{y}^+) + h(\boldsymbol{y}^+) \leq f(\boldsymbol{x}^+, \boldsymbol{y}) + h(\boldsymbol{y}), \tag{3.51}$$

for $f = f_1 - f_2$. Using the above inequalities, it holds that

$$\lambda \Psi_{\mathrm{B}}(\boldsymbol{x}^+, \boldsymbol{y}^+) \leq \lambda \Psi_{\mathrm{B}}(\boldsymbol{x}^+, \boldsymbol{y}) \leq \lambda \Psi_{\mathrm{B}}(\boldsymbol{x}, \boldsymbol{y}) - (1 - \lambda L) D_\phi(\boldsymbol{x}^+, \boldsymbol{x}).$$

The last statement follows from $0 < \lambda L < 1$. □

Remark 3.33. *In practice, even if (3.49) is not solved exactly, Lemma 3.32 is guaranteed by the point $\boldsymbol{y}^+$ satisfying $f(\boldsymbol{x}, \boldsymbol{y}^+) + h(\boldsymbol{y}^+) \leq f(\boldsymbol{x}, \boldsymbol{y}) + h(\boldsymbol{y})$ for any $\boldsymbol{x} \in C_1$ and $\boldsymbol{y} \in C_2$. For example, $\boldsymbol{y}^+$ is generated by a certain number of inner iterations of the proximal gradient method under the L-smoothness of $f(\boldsymbol{x}, \cdot)$ until $\boldsymbol{y}^+$ satisfies (3.51).*

3.3.2 Convergence Analysis of HBPDCA

In the same way as BPDCA, suppose that the following conditions hold.

Assumption 3.34.

(i) $\operatorname{dom}\phi = \mathbb{C}^{d_1}$ *and* ϕ *is* σ*-strongly convex on* $\mathbb{C}^{d_1}$.

(ii) $\nabla\phi$ *and* ∇f_1 *are Lipschitz continuous on any bounded subset of* $\mathbb{C}^{d_1}$.

(iii) *The objective function* Ψ_{B} *is level-bounded.*

From Definition 3.8, $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ is a critical point of (3.45) with $C_1 \equiv \mathbb{C}^{d_1}$ and $C_2 \equiv \mathbb{R}^{d_2}$ if and only if

$$\mathbf{0}_{d_1+d_2} \in \nabla f(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) + \begin{bmatrix} \partial_{\mathrm{c}} g_1(\tilde{\boldsymbol{x}}) - \partial_{\mathrm{c}} g_2(\tilde{\boldsymbol{x}}) \\ \partial_{\mathrm{c}} h(\tilde{\boldsymbol{y}}) \end{bmatrix}. \tag{3.52}$$

From Definition 3.8, $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ is a limiting stationary point of (3.45) with $C_1 \equiv \mathbb{C}^{d_1}$ and $C_2 \equiv \mathbb{R}^{d_2}$ if

$$\mathbf{0}_{d_1+d_2} \in \partial\Psi_{\mathrm{B}}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}). \tag{3.53}$$

Theorem 3.35 (Global subsequential convergence of HBPDCA)**.** *Suppose that Assumptions 3.2, 3.30, and 3.34 hold. Let* $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}_{k=0}^{\infty}$ *be a sequence generated by HBPDCA with* $0 < \lambda^k L(\boldsymbol{y}^k) < 1$ *for solving (3.45). Then, the following statements hold:*

(i) *The sequence* $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}_{k=0}^{\infty}$ *is bounded.*

(ii) *The sequence* $\{\boldsymbol{\xi}^k\}_{k=0}^{\infty}$ *is bounded.*

(iii) $\lim_{k\to\infty} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2 = 0$.

(iv) *Any accumulation point of* $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}_{k=0}^{\infty}$ *is a critical point of (3.45).*

Proof. (i) From Lemma 3.32, we obtain $\Psi_{\mathrm{B}}(\boldsymbol{x}^k, \boldsymbol{y}^k) \leq \Psi_{\mathrm{B}}(\boldsymbol{x}^0, \boldsymbol{y}^0)$ for all $k \in \mathbb{N}$, which shows that $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}_{k=0}^{\infty}$ is bounded from Assumption 3.34 (iii).

(ii) We can prove (ii) in a manner similar to Theorem 3.10 (ii).

(iii) From $1/\lambda^k - L(\boldsymbol{y}^k) > 0$ for any k, there exists $A_0 > 0$ such that $A_0 = \inf_k\{1/\lambda^k - L(\boldsymbol{y}^k)\}$. From (3.50), we obtain

$$\begin{aligned} \Psi_{\mathrm{B}}(\boldsymbol{x}^{k-1}, \boldsymbol{y}^{k-1}) - \Psi_{\mathrm{B}}(\boldsymbol{x}^k, \boldsymbol{y}^k) &\geq \left(\frac{1}{\lambda^k} - L(\boldsymbol{y}^k)\right) D_\phi(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}) \\ &\geq \left(\frac{1}{\lambda^k} - L(\boldsymbol{y}^k)\right) \frac{\sigma}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2^2 \\ &\geq \frac{\sigma A_0}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2^2, \end{aligned} \tag{3.54}$$

where the second inequality holds since ϕ is a σ-strongly convex function from Assumption 3.34 (i). Summing the above inequality from $k = 1$ to ∞, we obtain, for v_{B} from Assumption 3.30 (iv),

$$\sum_{k=1}^{\infty} \frac{\sigma A_0}{2} \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2^2 \leq \Psi_{\mathrm{B}}(\boldsymbol{x}^0, \boldsymbol{y}^0) - \liminf_{n\to\infty} \Psi_{\mathrm{B}}(\boldsymbol{x}^n, \boldsymbol{y}^n) \leq \Psi_{\mathrm{B}}(\boldsymbol{x}^0, \boldsymbol{y}^0) - v_{\mathrm{B}} < \infty,$$

which shows that $\lim_{k\to\infty} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2 = 0$.

(iv) Let $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ be an accumulation point of $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}_{k=0}^{\infty}$ and let $\{(\boldsymbol{x}^{k_j}, \boldsymbol{y}^{k_j})\}$ be a subsequence such that $\lim_{j\to\infty} \boldsymbol{x}^{k_j} = \tilde{\boldsymbol{x}}$ and $\lim_{j\to\infty} \boldsymbol{y}^{k_j} = \tilde{\boldsymbol{y}}$. Then, from the first-order optimality condition of subproblem (3.46) under Assumption 3.2, we have

$$\mathbf{0}_{d_1} \in \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{k_j}, \boldsymbol{y}^{k_j}) - \boldsymbol{\xi}^{k_j} + \partial_{\mathrm{c}} g_1(\boldsymbol{x}^{k_j+1}) + \frac{1}{\lambda}\left(\nabla\phi(\boldsymbol{x}^{k_j+1}) - \nabla\phi(\boldsymbol{x}^{k_j})\right).$$

Therefore,

$$\boldsymbol{\xi}^{k_j} + \frac{1}{\lambda}\left(\nabla\phi(\boldsymbol{x}^{k_j}) - \nabla\phi(\boldsymbol{x}^{k_j+1})\right) \in \partial g_1(\boldsymbol{x}^{k_j+1}) + \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{k_j}, \boldsymbol{y}^{k_j}). \tag{3.55}$$

From the boundedness of $\{(\boldsymbol{x}^{k_j}, \boldsymbol{y}^{k_j})\}$ and the Lipschitz continuity of $\nabla\phi$ on a bounded subset of $\mathbb{C}^{d_1}$, there exists $A_1 > 0$ such that

$$\left\| \frac{1}{\lambda}\left(\nabla\phi(\boldsymbol{x}^{k_j}) - \nabla\phi(\boldsymbol{x}^{k_j+1})\right) \right\|_2 \leq \frac{A_1}{\lambda} \|\boldsymbol{x}^{k_j+1} - \boldsymbol{x}^{k_j}\|_2.$$

Therefore, using $\|\boldsymbol{x}^{k_j+1} - \boldsymbol{x}^{k_j}\|_2 \to 0$, we obtain

$$\frac{1}{\lambda}\left(\nabla\phi(\boldsymbol{x}^{k_j}) - \nabla\phi(\boldsymbol{x}^{k_j+1})\right) \to \mathbf{0}_{d_1}. \tag{3.56}$$

Note that the sequence $\{\boldsymbol{\xi}^{k_j}\}$ 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 $\lim_{j\to\infty} \boldsymbol{\xi}^{k_j} =: \tilde{\boldsymbol{\xi}}$ exists, which belongs to $\partial_{\mathrm{c}} g_2(\tilde{\boldsymbol{x}})$ since g_2 becomes continuous due to its convexity on $\mathbb{C}^{d_1}$. Using this and (3.56), we can take the limit of (3.55). Setting $\|\boldsymbol{x}^{k_j+1} - \boldsymbol{x}^{k_j}\|_2 \to 0$ and invoking the lower semicontinuity of g_1 and $\nabla_{\boldsymbol{x}} f$, we obtain $\tilde{\boldsymbol{\xi}} \in \nabla_{\boldsymbol{x}} f(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) + \partial_{\mathrm{c}} g_1(\tilde{\boldsymbol{x}})$. Therefore, $\mathbf{0}_{d_1} \in \nabla_{\boldsymbol{x}} f(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) + \partial_{\mathrm{c}} g_1(\tilde{\boldsymbol{x}}) - \partial_{\mathrm{c}} g_2(\tilde{\boldsymbol{x}})$. Then, from the first-order optimality condition of subproblem (3.47), we have

$$\mathbf{0}_{d_2} \in \nabla_{\boldsymbol{y}} f(\boldsymbol{x}^{k_j+1}, \boldsymbol{y}^{k_j+1}) + \partial_{\mathrm{c}} h(\boldsymbol{y}^{k_j+1}),$$

which implies $\mathbf{0}_{d_2} \in \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) + \partial_{\mathrm{c}} h(\tilde{\boldsymbol{y}})$ as $j \to \infty$ from the lower semicontinuity of h and $\nabla_{\boldsymbol{y}} f$. Therefore, we obtain

$$\mathbf{0}_{d_1+d_2} \in \nabla f(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) + \begin{bmatrix} \partial_{\mathrm{c}} g_1(\tilde{\boldsymbol{x}}) - \partial_{\mathrm{c}} g_2(\tilde{\boldsymbol{x}}) \\ \partial_{\mathrm{c}} h(\tilde{\boldsymbol{y}}) \end{bmatrix},$$

which shows that $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ is a critical point of (3.45). □

If the KL property is extended to complex variables, under the KL property on complex variables, we expect that HBPDCA converges to a limiting stationary point. It remains as future work.

Chapter 4

Applications

4.1 Application of Bregman Proximal Algorithms Exploiting DC Structure

In this chapter, we consider applications to signal processing, such as phase retrieval, blind deconvolution, and self-calibration in radio interferometric imaging. They are known to be ill-posed because their solutions may not be unique. Adding a regularization term that may be nonsmooth, we can write these applications as the following nonconvex optimization problem:

$$\min_{\boldsymbol{x}\in\operatorname{cl} C} \quad f(\boldsymbol{x}) + g(\boldsymbol{x}), \tag{4.1}$$

where $f : \mathbb{R}^d \to (-\infty, +\infty]$ is a nonconvex loss function, $g : \mathbb{R}^d \to (-\infty, +\infty]$ is a regularization term, $C \subset \mathbb{R}^d$ is a nonempty open convex set. In self-calibration in radio interferometric imaging, we replace $\mathbb{R}^d$ with $\mathbb{C}^d$.

It is non-trivial to apply our proposed algorithms to these problems in signal processing. In practice, when we apply Bregman DC proximal algorithms to (4.1), we find an appropriate DC structure $f = f_1 - f_2$ and an appropriate kernel generating distance $\phi \in \mathcal{G}(C)$ that satisfy the following conditions at the same time:

(i) The functions $f_1, f_2 : \mathbb{R}^d \to (-\infty, +\infty]$ are convex, and f_1 is $\mathcal{C}^1$.

(ii) The pair (f_1, ϕ) is L-smad. Because f_1 is a convex function, we only need to obtain ϕ and L such that the function $L\phi - f_1$ is convex.

(iii) Subproblems (3.2), (3.32), or (3.46) are efficiently solved. For example, they are solved in closed forms.

The above conditions are followed in corresponding sections and remarks:

(i) Sections 4.2.2, 4.3.2, and 4.4.2.

(ii) Sections 4.2.3, 4.2.4, 4.3.3, and 4.4.3.

(iii) Remarks 4.2, 4.9, and 4.14.

For other Bregman proximal algorithms without DC decomposition, although they do not require the condition (i), they have less flexibility on the choice of ϕ. Because DC decomposition is not unique, we can choose tractable f_1 and ϕ. The function ϕ has the role of approximating f_1 (see also Lemma 2.10 and Remark 2.11). When ϕ approximates f_1 well, L is smaller, and then smaller L accelerates Bregman DC proximal algorithms. It is desirable to choose ϕ that approximates f_1 well, while empirically such ϕ makes it difficult to solve subproblems with a small computational burden. Therefore, there is a trade-off between the tractability of subproblems and how well ϕ approximates f_1. From this kind of circumstance, we need to choose ϕ that approximates f_1 as well as possible, and we must be able to solve subproblems with D_ϕ with a small computational burden. In the following sections, we find an appropriate DC decomposition $f = f_1 - f_2$ and an appropriate ϕ satisfying these conditions.

In particular, for phase retrieval, exploiting DC structure, we obtain smaller parameters L than the existing one. By using a smaller L, we succeed in accelerating BPDCA(e). For blind deconvolution, although it is difficult to find an appropriate ϕ without DC structure (see also Remark 4.5), exploiting DC structure, we obtain an appropriate ϕ.

4.2 Phase Retrieval

4.2.1 Problem Description

We are interested in finding a vector $\boldsymbol{x} \in \mathbb{R}^d$ that approximately satisfies

$$\langle \boldsymbol{a}_r, \boldsymbol{x} \rangle^2 \simeq b_r, \quad r = 1, \dots, m, \tag{4.2}$$

where the vectors $\boldsymbol{a}_r \in \mathbb{R}^d$ describe the model and $\boldsymbol{b} = (b_1, b_2, \dots, b_m)$ is a vector of (usually) noisy measurements. As described in [15, 22], the system (4.2) can be formulated as a nonconvex optimization problem:

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} \quad \Psi(\boldsymbol{x}) := \frac{1}{4} \sum_{r=1}^{m} \left(\langle \boldsymbol{a}_r, \boldsymbol{x} \rangle^2 - b_r \right)^2 + g(\boldsymbol{x}), \tag{4.3}$$

where the function $g : \mathbb{R}^d \to \mathbb{R}$ is a regularizer, in particular $g(\boldsymbol{x}) = \theta \|\boldsymbol{x}\|_1$ for a trade-off parameter $\theta \geq 0$ between the data fidelity criteria and the regularizer g. In this case, the underlying space of (3.1) is $C \equiv \mathbb{R}^d$. Define $f : \mathbb{R}^d \to \mathbb{R}$ as $f(\boldsymbol{x}) = \frac{1}{4} \sum_{r=1}^{m} \left(\langle \boldsymbol{a}_r, \boldsymbol{x} \rangle^2 - b_r \right)^2$, which is a nonconvex differentiable function that does not admit a global Lipschitz continuous gradient.

BPG [15] is one of the methods to solve (4.3). For BPG, assuming L-smad for the pair (f, ϕ) using $\phi(\boldsymbol{x}) = \frac{1}{4}\|\boldsymbol{x}\|_2^4 + \frac{1}{2}\|\boldsymbol{x}\|_2^2$, the parameter L satisfies the following inequality [15, Lemma 5.1]:

$$L \geq \sum_{r=1}^{m} \left(3\|\boldsymbol{a}_r \boldsymbol{a}_r^\mathsf{T}\|_2^2 + \|\boldsymbol{a}_r \boldsymbol{a}_r^\mathsf{T}\|_2 |b_r| \right). \tag{4.4}$$

4.2.2 DC Decomposition

The function f in the optimization problem (4.3) can also be reformulated as a difference between two convex functions such as in [50]. That is, $f(\boldsymbol{x}) = f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})$, where

$$f_1(\boldsymbol{x}) = \frac{1}{4}\sum_{r=1}^{m}\langle \boldsymbol{a}_r, \boldsymbol{x}\rangle^4 + \frac{1}{4}\|\boldsymbol{b}\|_2^2, \quad f_2(\boldsymbol{x}) = \frac{1}{2}\sum_{r=1}^{m} b_r\langle \boldsymbol{a}_r, \boldsymbol{x}\rangle^2. \tag{4.5}$$

Since $f = f_1 - f_2$ holds, (4.3) is equivalent to the following DC optimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \quad \Psi(\boldsymbol{x}) = f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}) + g(\boldsymbol{x}). \tag{4.6}$$

4.2.3 L-smooth Adaptable Parameters

For problem (4.6), we define $\phi : \mathbb{R}^d \to \mathbb{R}$ as

$$\phi(\boldsymbol{x}) = \frac{1}{4}\|\boldsymbol{x}\|_2^4, \tag{4.7}$$

which is simpler than the original nonconvex formulation. Since this $\phi(\boldsymbol{x})$ is not strongly convex, it does not satisfy Assumption 3.7 (i).

Proposition 4.1. *Let f_1 and ϕ be as defined above. Then, for any L satisfying*

$$L \geq 3\left\|\sum_{r=1}^{m}\|\boldsymbol{a}_r\|_2^2\boldsymbol{a}_r\boldsymbol{a}_r^\mathsf{T}\right\|_F, \tag{4.8}$$

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

Proof. Let $\boldsymbol{x} \in \mathbb{R}^d$. Suppose that L satisfies (4.8), in order to guarantee the convexity of $L\phi - f_1$, it is sufficient to show $L\lambda_{\min}(\nabla^2\phi(\boldsymbol{x})) \geq \lambda_{\max}(\nabla^2 f_1(\boldsymbol{x}))$ since f_1 and ϕ are $\mathcal{C}^2$ on $\mathbb{R}^d$. Now, we have the Hessian for f_1 and ϕ:

$$\nabla^2 f_1(\boldsymbol{x}) = 3\sum_{r=1}^{m}\langle \boldsymbol{a}_r, \boldsymbol{x}\rangle^2\boldsymbol{a}_r\boldsymbol{a}_r^\mathsf{T}, \quad \nabla^2\phi(\boldsymbol{x}) = \|\boldsymbol{x}\|_2^2\boldsymbol{I}_d + 2\boldsymbol{x}\boldsymbol{x}^\mathsf{T}.$$

Since $\nabla^2\phi(\boldsymbol{x}) \succeq \|\boldsymbol{x}\|_2^2\boldsymbol{I}_d$, we obtain $\lambda_{\min}\left(\nabla^2\phi(\boldsymbol{x})\right) \geq \|\boldsymbol{x}\|_2^2$. From the well-known fact, $\lambda_{\max}(\boldsymbol{M}) \leq \|\boldsymbol{M}\|_F$, we have the following inequality:

$$\begin{aligned}
\lambda_{\max}\left(\nabla^2 f_1(\boldsymbol{x})\right) &\leq 3\left\|\sum_{r=1}^{m}\langle \boldsymbol{a}_r, \boldsymbol{x}\rangle^2\boldsymbol{a}_r\boldsymbol{a}_r^\mathsf{T}\right\|_F \\
&\leq 3\left\|\sum_{r=1}^{m}\|\boldsymbol{a}_r\|_2^2\boldsymbol{a}_r\boldsymbol{a}_r^\mathsf{T}\right\|_F\|\boldsymbol{x}\|_2^2 \\
&\leq L\|\boldsymbol{x}\|_2^2 \leq L\lambda_{\min}(\nabla^2\phi(\boldsymbol{x})).
\end{aligned}$$

Therefore, we obtain the desired result. □

Comparing the right-hand side of (4.4) and that of (4.8), we can see that

$$3\left\|\sum_{r=1}^{m}\|\boldsymbol{a}_r\|_2^2\boldsymbol{a}_r\boldsymbol{a}_r^\mathsf{T}\right\|_F \leq \sum_{r=1}^{m}(3\|\boldsymbol{a}_r\boldsymbol{a}_r^\mathsf{T}\|_F^2 + \|\boldsymbol{a}_r\boldsymbol{a}_r^\mathsf{T}\|_F|b_r|). \tag{4.9}$$

The constant L has the important role of defining the step size and thereby affects the performance of the algorithms. Note that even if $\left\|\sum_{r=1}^{m}\|\boldsymbol{a}_r\|_2^2\boldsymbol{a}_r\boldsymbol{a}_r^\mathsf{T}\right\|_F = \sum_{r=1}^{m}\|\boldsymbol{a}_r\boldsymbol{a}_r^\mathsf{T}\|_F^2$, the left-hand side of (4.9) is always smaller than the right-hand side because $\|\boldsymbol{a}_r\boldsymbol{a}_r^\mathsf{T}\|_F|b_r|$ is nonnegative. When $\phi(\boldsymbol{x}) = \frac{1}{4}\|\boldsymbol{x}\|_2^4 + \frac{1}{2}\|\boldsymbol{x}\|_2^2$, the subproblems of BPG(e) have a closed-form solution formula [15, Proposition 5.1]. When $\phi(\boldsymbol{x}) = \frac{1}{4}\|\boldsymbol{x}\|_2^4$, subproblems (3.2) and (3.32) also have a closed-form solution formula, which is obtained by slight modifications of those in BPG(e).

Remark 4.2. *Let ϕ be defined by $\phi(\boldsymbol{x}) = \frac{1}{4}\|\boldsymbol{x}\|_2^4$. Let $\boldsymbol{u}^k = \mathcal{S}_{\lambda\theta}(\lambda\nabla f(\boldsymbol{x}^k) - \nabla\phi(\boldsymbol{x}^k))$. We can prove from* [15, Proposition 5.1] *that $\boldsymbol{x}^{k+1} = -t^*\boldsymbol{u}^k$ solves subproblem* (3.2)*, where t^* is the unique positive real root of the cubic equation $t^3\|\boldsymbol{u}^k\|_2^2 - 1 = 0$, i.e., $t^* = \sqrt[3]{\|\boldsymbol{u}^k\|_2^2}$. It is also true for subproblem* (3.32)*.*

In this application, the functions f_1, f_2, g, and ϕ satisfy Assumptions from 3.1 to 3.25 excepting Assumption 3.7 (i) and 3.13. In particular, Assumption 3.7 (i) is not satisfied for our choice $\phi(\boldsymbol{x}) = \frac{1}{4}\|\boldsymbol{x}\|_2^4$, but is satisfied if we replace it by $\phi(x) = \frac{1}{4}\|\boldsymbol{x}\|_2^4 + \frac{1}{2}\|\boldsymbol{x}\|_2^2$. Finally, Ψ and H_M are KL functions due to their semi-algebraicity [5]. Therefore, in this application, Assumption 3.13 is not required for the global convergence of BPDCAe. Moreover, the KL exponent θ for phase retrieval is known to be at least $\frac{1}{4}$ [127]. From Theorems 3.16 and 3.26, BPDCA(e) for phase retrieval linearly converges to a limiting stationary point of (4.2).

4.2.4 L-smooth Adaptable Parameters in a Gaussian Model

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

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

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

Since the elements of $\boldsymbol{a}_r$ are independently generated from the standard Gaussian distribution, the jth diagonal element of the above matrix is given by

$$\mathbb{E}\left[\langle\boldsymbol{a}_r, \boldsymbol{x}\rangle^2 a_{r,j}^2\right] = \mathbb{E}\left[a_{r,j}^4 x_j^2 + \sum_{k=1,k\neq j}^{d} a_{r,j}^2 a_{r,k}^2 x_k^2\right] = 3x_j^2 + \sum_{k=1,k\neq j}^{d} x_k^2 = 2x_j^2 + \|\boldsymbol{x}\|_2^2.$$

The non-diagonal (j,k) elements are

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

Moreover, noting that $\phi(\boldsymbol{x}) = \frac{1}{4}\|\boldsymbol{x}\|_2^4$, we obtain $\mathbb{E}\left[\langle \boldsymbol{a}_r, \boldsymbol{x}\rangle^2 \boldsymbol{a}_r \boldsymbol{a}_r^\mathsf{T}\right] = \|\boldsymbol{x}\|_2^2 \boldsymbol{I}_d + 2\boldsymbol{x}\boldsymbol{x}^\mathsf{T} = \nabla^2\phi(\boldsymbol{x})$. Therefore, the Hessian expectation of f_1 is given by $\mathbb{E}[\nabla^2 f_1(\boldsymbol{x})] = 3m\nabla^2\phi(\boldsymbol{x})$.

Under a Gaussian model, we can reduce the lower bound of L given in Proposition 4.1 with high probability by applying [22, Lemma 7.4] as shown in the following proposition.

Proposition 4.3. *Let the functions f_1 and ϕ be given by (4.5) and (4.7), respectively. Moreover, assume that the vectors $\boldsymbol{a}_r$ are independently distributed according to a 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*

$$L \geq 9\left\|\sum_{r=1}^m \boldsymbol{a}_r \boldsymbol{a}_r^\mathsf{T}\right\|_F + \delta, \tag{4.10}$$

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

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

$$\boldsymbol{y}^\mathsf{T}\mathbb{E}\left[\sum_{r=1}^m \boldsymbol{a}_r \boldsymbol{a}_r^\mathsf{T}\right]\boldsymbol{y} = \sum_{r=1}^m \mathbb{E}\left[\langle \boldsymbol{a}_r, \boldsymbol{y}\rangle^2\right] = \sum_{r=1}^m \sum_{j=1}^d y_j^2 = \sum_{r=1}^m \|\boldsymbol{y}\|_2^2. \tag{4.11}$$

From (4.11), for any $\boldsymbol{y} \in \mathbb{R}^d$, we have

$$\begin{aligned}
\boldsymbol{y}^\mathsf{T}\mathbb{E}[\nabla^2 f_1(\boldsymbol{x})]\boldsymbol{y} &= 3\sum_{r=1}^m \left(\|\boldsymbol{x}\|_2^2\|\boldsymbol{y}\|_2^2 + 2\langle \boldsymbol{x}, \boldsymbol{y}\rangle^2\right) \\
&\leq 9\|\boldsymbol{x}\|_2^2 \sum_{r=1}^m \|\boldsymbol{y}\|_2^2 \\
&= 9\|\boldsymbol{x}\|_2^2 \boldsymbol{y}^\mathsf{T}\mathbb{E}\left[\sum_{r=1}^m \boldsymbol{a}_r \boldsymbol{a}_r^\mathsf{T}\right]\boldsymbol{y}.
\end{aligned} \tag{4.12}$$

We can easily find that

$$9\sum_{r=1}^m \boldsymbol{a}_r \boldsymbol{a}_r^\mathsf{T} \preceq 9\left\|\sum_{r=1}^m \boldsymbol{a}_r \boldsymbol{a}_r^\mathsf{T}\right\|_F \boldsymbol{I}_d,$$

which implies that

$$9\mathbb{E}\left[\sum_{r=1}^{m} \boldsymbol{a}_r \boldsymbol{a}_r^\top\right] \preceq 9\left\|\sum_{r=1}^{m} \boldsymbol{a}_r \boldsymbol{a}_r^\top\right\|_F \boldsymbol{I}_d. \tag{4.13}$$

From (4.12) and (4.13), we have

$$\mathbb{E}[\nabla^2 f_1(\boldsymbol{x})] \preceq 9\|\boldsymbol{x}\|_2^2 \left\|\sum_{r=1}^{m} \boldsymbol{a}_r \boldsymbol{a}_r^\top\right\|_F \boldsymbol{I}_d. \tag{4.14}$$

From [22, Lemma 7.4], (4.10), and (4.14), we conclude that

$$\nabla^2 f_1(\boldsymbol{x}) \preceq \mathbb{E}[\nabla^2 f_1(\boldsymbol{x})] + \delta\|\boldsymbol{x}\|_2^2 \boldsymbol{I}_d \preceq L\|\boldsymbol{x}\|_2^2 \boldsymbol{I}_d \tag{4.15}$$

with probability at least $1 - 5e^{-\gamma d} - 4/d^2$. From $\nabla^2\phi(\boldsymbol{x}) \succeq \|\boldsymbol{x}\|_2^2 \boldsymbol{I}_d$ and (4.15), we have $\nabla^2 f_1(\boldsymbol{x}) \preceq L\nabla^2\phi(\boldsymbol{x})$, which proves that $L\phi - f_1$ is convex with probability at least $1 - 5e^{-\gamma d} - 4/d^2$. Therefore, the pair (f_1, ϕ) is L-smad on $\mathbb{R}^d$. □

Remark 4.4. *Since each element of $\boldsymbol{a}_r$ independently follows the standard Gaussian distribution, $\|\boldsymbol{a}_r\|_2^2$ follows the chi-square distribution with d degrees of freedom. Thus, we can show $\|\boldsymbol{a}_r\|_2^2 \geq 3$ with a high probability for sufficiently large d. It implies that the bound given in Proposition 4.3 is smaller than that given in Proposition 4.1.*

4.2.5 Numerical Experiments

Here, we summarize the results of a 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.

The codes of BPDCA(e) and the datasets generated during and/or analyzed in Section 4.2 are available in the GitHub repository, https://github.com/ShotaTakahashi/bregman-proximal-dc-algorithm.

First, let us examine the results for the Bregman proximal algorithms, *i.e.*, BPG [15], BPGe [128], 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 4.1 and 4.2). Let $\hat{\boldsymbol{x}}$ be a recovered solution and $\boldsymbol{x}^\circ$ be the ground truth generated according to the method described in Section 4.2.4. In order to compare the objective function values, we took the difference $\log_{10}|\Psi(\hat{\boldsymbol{x}}) - \Psi(\boldsymbol{x}^\circ)|$ to be the accuracy. In numerical experiments, $\Psi(\hat{\boldsymbol{x}}) > \Psi(\boldsymbol{x}^\circ)$. The termination criterion was defined as $\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2 / \max\{1, \|\boldsymbol{x}^k\|_2\} \leq 10^{-6}$. The equation numbers under each algorithm in Tables 4.1 and 4.2 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 (4.3). We forcibly stopped the algorithms when they reached the maximum number of iterations (50,000). Table 4.2 compares the results of BPGe and BPDCAe under the same settings

Table 4.1: Average numbers of iterations, CPU time, and accuracy for BPG [15] 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 [15] (4.4)	10000	10	3757	1.638	2.901
		50	50000	37.761	1.977
		100	50000	46.920	5.312
		200	50000	91.925	7.737
	20000	10	3689	2.539	−2.569
		50	50000	76.020	2.007
		100	50000	121.966	5.523
		200	50000	191.780	8.057
	30000	10	3764	3.698	−2.387
		50	50000	104.947	2.257
		100	50000	175.143	5.678
		200	50000	287.735	8.227
BPDCA (4.8)	10000	10	265	0.102	−4.374
		50	1415	0.520	−3.212
		100	3274	2.129	−2.656
		200	8111	10.416	−2.061
	20000	10	255	0.157	−4.350
		50	1299	1.182	−3.193
		100	2833	4.283	−2.642
		200	6572	18.198	−2.057
	30000	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 (4.10)	10000	10	68	0.025	−5.127
		50	92	0.034	−4.627
		100	115	0.075	−4.380
		200	152	0.192	−4.108
	20000	10	65	0.040	−5.137
		50	84	0.077	−4.691
		100	98	0.149	−4.476
		200	121	0.335	−4.229
	30000	10	65	0.059	−5.166
		50	81	0.115	−4.728
		100	93	0.223	−4.515
		200	110	0.465	−4.285

Table 4.2: Average numbers of iterations, CPU time, and accuracy for BPGe [128] 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 [128] (4.4)	10000	10	297	0.124	−3.904
		50	2614	1.209	−0.428
		100	6214	5.949	0.974
		200	23940	44.218	2.426
	20000	10	285	0.198	−3.653
		50	1941	2.871	−0.375
		100	6054	15.376	1.250
		200	21138	82.086	2.734
	30000	10	294	0.290	−3.362
		50	1880	3.826	−0.199
		100	6002	21.271	1.411
		200	21434	123.504	2.806
BPDCAe (4.8)	10000	10	67	0.025	−5.205
		50	203	0.075	−3.802
		100	332	0.218	−3.451
		200	581	0.740	−2.941
	20000	10	62	0.038	−5.071
		50	179	0.165	−4.152
		100	302	0.458	−3.694
		200	501	1.394	−3.110
	30000	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 (4.10)	10000	10	32	0.013	−5.649
		50	42	0.015	−5.371
		100	49	0.032	−5.087
		200	61	0.078	−5.135
	20000	10	29	0.018	−5.550
		50	38	0.035	−5.317
		100	43	0.065	−4.919
		200	52	0.144	−5.051
	30000	10	29	0.026	−5.558
		50	38	0.056	−5.446
		100	41	0.098	−4.908
		200	50	0.210	−5.115

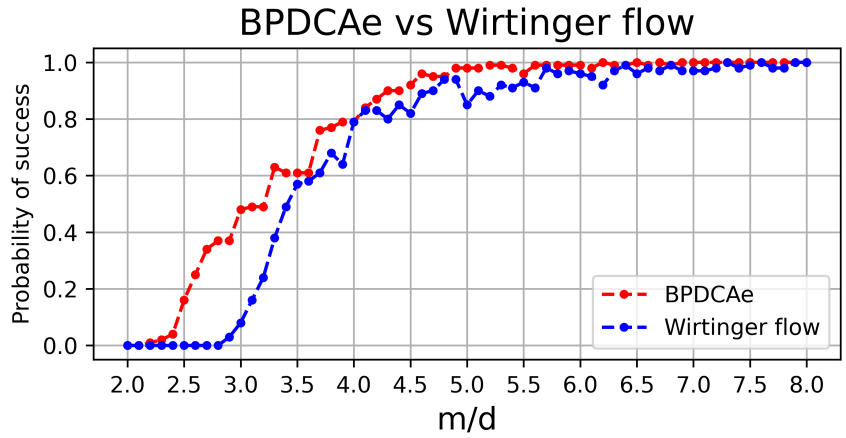


Figure 4.1: The empirical probability of success based on 100 trials for BPDCAe and the Wirtinger flow [22] using the same initialization step (of the Wirtinger flow). We set $d = 128$ and varied the number m of measurements.

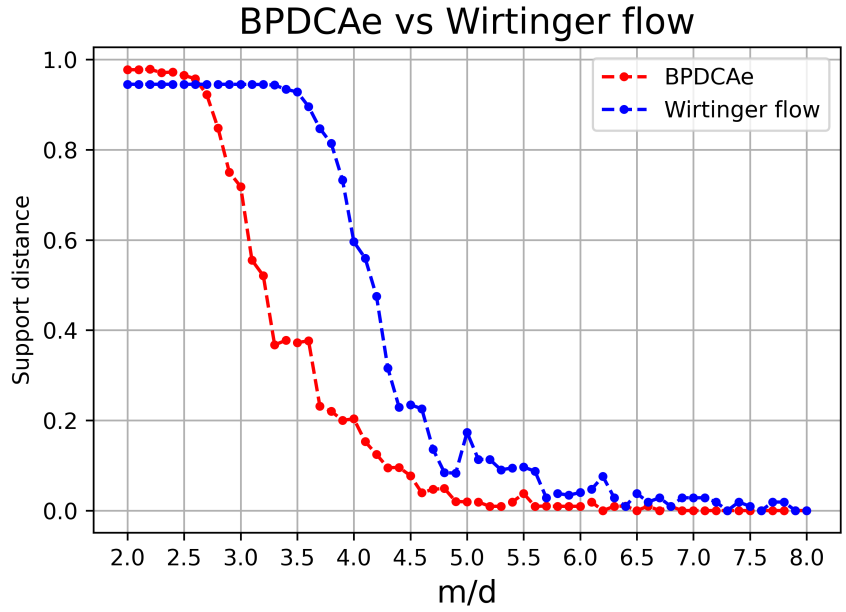


Figure 4.2: The average of support distances based on 100 trials for BPDCAe and the Wirtinger flow [22].

as the results in Table 4.1. BPDCA with (4.10) was the fastest among the algorithms without extrapolation (Table 4.1). On the other hand, the extrapolation method makes each algorithm faster (Table 4.2).

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 (4.5), which permits choosing a smaller L in (4.8) instead of (4.4). In particular, for the Gaussian model, we can use a smaller L in (4.10) with high probability. The extrapolation technique can further enhance performance. Also, we can see that the sequences of BPDCA(e) globally converge to their optimal solutions despite that the kernel generating distance ϕ (4.7) does not satisfy Assumption 3.7 (i). This suggests that this condition may be relaxed in some cases.

Next, we compared the empirical probability of success for BPDCAe and the Wirtinger flow [22], which is a well-known algorithm for phase retrieval. Here we took the initial point $\boldsymbol{x}^0$ in BPDCAe to be the value calculated in the initialization step of the Wirtinger flow. The empirical probability of success and the average of support distances in Figures 4.1 and 4.2 are on 100 trials, respectively. We regard that the algorithms succeeded if the relative error $\|\hat{\boldsymbol{x}} - \boldsymbol{x}^\circ\|_2/\|\boldsymbol{x}^\circ\|_2$ falls below 10^{-5} after 2,500 iterations. The support $S(\boldsymbol{x})$ and the support distance [39, p. 47] are defined by $S(\boldsymbol{x}) = \{j \mid x_j \neq 0\}$ and

$$\operatorname{dist}(S(\hat{\boldsymbol{x}}), S(\boldsymbol{x}^\circ)) = \frac{\max\{|S(\hat{\boldsymbol{x}})|, |S(\boldsymbol{x}^\circ)|\} - |S(\hat{\boldsymbol{x}}) \cap S(\boldsymbol{x}^\circ)|}{\max\{|S(\hat{\boldsymbol{x}})|, |S(\boldsymbol{x}^\circ)|\}},$$

respectively. When $\operatorname{dist}(S(\hat{\boldsymbol{x}}), S(\boldsymbol{x}^\circ))$ is 0, the index set of nonzero elements of $\hat{\boldsymbol{x}}$ corresponds to that of $\boldsymbol{x}^\circ$. 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 (4.3) 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 the Wirtinger flow.

4.3 Blind Deconvolution with Nonsmooth Regularization

4.3.1 Problem Description

We consider the convolution of a filter $\boldsymbol{f} \in \mathbb{R}^m$ and a signal $\boldsymbol{g} \in \mathbb{R}^m$, given by

$$\tilde{\boldsymbol{y}} = \boldsymbol{f} * \boldsymbol{g}. \tag{4.16}$$

Our goal is to recover $\boldsymbol{g}$ from $\tilde{\boldsymbol{y}}$ without knowing $\boldsymbol{f}$. This problem is known as blind deconvolution. Without any assumptions, blind deconvolution is ill-posed because its solution may not be unique. A common approach is to assume that $\boldsymbol{f}$ and $\boldsymbol{g}$ belong to known subspaces [3]. Concretely, for known linear operators $\tilde{\boldsymbol{B}} : \mathbb{R}^{d_1} \to \mathbb{R}^m$ and

$\tilde{\boldsymbol{A}} : \mathbb{R}^{d_2} \to \mathbb{R}^m$, we assume that there exist the true vectors $\boldsymbol{h}^\circ \in \mathbb{R}^{d_1}$ and $\boldsymbol{x}^\circ \in \mathbb{R}^{d_2}$ such that $\boldsymbol{f} = \tilde{\boldsymbol{B}}\boldsymbol{h}^\circ$ and $\boldsymbol{g} = \tilde{\boldsymbol{A}}\boldsymbol{x}^\circ$, where $d_1, d_2 < m$. Moreover, we consider blind deconvolution in the Fourier domain. Applying the discrete Fourier transform (DFT) to both sides of (4.16) and letting $\boldsymbol{F} \in \mathbb{C}^{m\times m}$ be the unitary DFT matrix, we obtain

$$\sqrt{m}\boldsymbol{F}\tilde{\boldsymbol{y}} = \sqrt{m}\boldsymbol{F}(\boldsymbol{f} * \boldsymbol{g}) = \sqrt{m}\boldsymbol{F}\boldsymbol{f} \odot \sqrt{m}\boldsymbol{F}\boldsymbol{g} = m\boldsymbol{F}\tilde{\boldsymbol{B}}\boldsymbol{h}^\circ \odot \boldsymbol{F}\tilde{\boldsymbol{A}}\boldsymbol{x}^\circ,$$

where the second equality holds from the convolution theorem [99, Section 4.4.2], and $\odot$ denotes the Hadamard (elementwise) product. Thus, (4.16) is rewritten in the Fourier domain as $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{h}^\circ \odot \overline{\boldsymbol{A}\boldsymbol{x}^\circ}$, where $\boldsymbol{y} := \frac{1}{\sqrt{m}}\boldsymbol{F}\tilde{\boldsymbol{y}}$, $\boldsymbol{B} := \boldsymbol{F}\tilde{\boldsymbol{B}}$, and $\overline{\boldsymbol{A}} := \boldsymbol{F}\tilde{\boldsymbol{A}}$.

Now, the goal of our problem is to estimate $\boldsymbol{h}^\circ$ and $\boldsymbol{x}^\circ$ from $\boldsymbol{y}$. To evaluate the fidelity, we consider the squared error function $f(\boldsymbol{h}, \boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{B}\boldsymbol{h} \odot \overline{\boldsymbol{A}\boldsymbol{x}} - \boldsymbol{y}\|_2^2$. In addition, in order to incorporate the image characteristics, we also consider a regularization term $g : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to (-\infty, +\infty]$. It may not be differentiable. Commonly used regularization terms include non-differentiable functions, such as the ℓ_1 norm and the total variation. To compute $\boldsymbol{h}^\circ$ and $\boldsymbol{x}^\circ$, we minimize the sum of these two functions and a constraint set $\operatorname{cl} C$ for a nonempty open convex set $C \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$. That is, we consider the following optimization problem:

$$\min_{(\boldsymbol{h},\boldsymbol{x})\in\operatorname{cl} C} \quad \Psi(\boldsymbol{h}, \boldsymbol{x}) := f(\boldsymbol{h}, \boldsymbol{x}) + g(\boldsymbol{h}, \boldsymbol{x}). \tag{4.17}$$

Note that f is a quartic function because it has a quartic term $h_i h_j x_k x_l$ with respect to $(\boldsymbol{h}, \boldsymbol{x})$, and thus it does not have a Lipschitz continuous gradient. Hence, we cannot rely on the convergence analysis of existing first-order methods, such as FISTA [10], because their convergence analysis depends on the existence of Lipschitz continuous gradients. Instead, we try to resort to Bregman proximal gradient algorithms [15, 122]. These algorithms generalize the proximal gradient method by replacing the squared Euclidean distance with the Bregman distance D_ϕ associated with a kernel generating distance ϕ. The algorithms generate a sequence converging to a limiting stationary point under the L-smad property of (f, ϕ) defined later. For our problem, however, finding an appropriate ϕ is difficult because of the bilinear term of f.

Remark 4.5. *We obtain the Hessian of f as follows:*

$$\nabla^2 f(\boldsymbol{h}, \boldsymbol{x}) = \operatorname{Re}\begin{bmatrix} \boldsymbol{G}_{11} & \boldsymbol{G}_{12} \\ \boldsymbol{G}_{12}^{\mathsf{T}} & \boldsymbol{G}_{22} \end{bmatrix},$$

where

$$\begin{aligned}
\boldsymbol{G}_{11} &:= \boldsymbol{B}^{\mathsf{H}} \operatorname{diag}(|\boldsymbol{A}\boldsymbol{x}|^2)\boldsymbol{B}, \\
\boldsymbol{G}_{12} &:= \boldsymbol{B}^{\mathsf{H}} \operatorname{diag}(\boldsymbol{B}\boldsymbol{h} \odot \boldsymbol{A}\boldsymbol{x})\overline{\boldsymbol{A}} + \boldsymbol{B}^{\mathsf{H}} \operatorname{diag}(\boldsymbol{B}\boldsymbol{h} \odot \overline{\boldsymbol{A}\boldsymbol{x}} - \boldsymbol{y})\boldsymbol{A}, \\
\boldsymbol{G}_{22} &:= \boldsymbol{A}^{\mathsf{H}} \operatorname{diag}(|\boldsymbol{B}\boldsymbol{h}|^2)\overline{\boldsymbol{A}}.
\end{aligned}$$

To have the L-smad property of (f, ϕ), we consider $\nabla^2 f(\boldsymbol{h}, \boldsymbol{x}) \preceq L\nabla^2\phi(\boldsymbol{h}, \boldsymbol{x})$ for any $\boldsymbol{h} \in \mathbb{R}^{d_1}$ and $\boldsymbol{x} \in \mathbb{R}^{d_2}$. Separate $\boldsymbol{w} = (\boldsymbol{u}, \boldsymbol{v})$ using $\boldsymbol{u} \in \mathbb{R}^{d_1}$ and $\boldsymbol{v} \in \mathbb{R}^{d_2}$. We obtain

$$\langle \boldsymbol{w}, \nabla^2 f(\boldsymbol{h}, \boldsymbol{x})\boldsymbol{w}\rangle = \operatorname{Re}\langle \boldsymbol{u}, \boldsymbol{G}_{11}\boldsymbol{u}\rangle + \operatorname{Re}\langle \boldsymbol{v}, \boldsymbol{G}_{22}\boldsymbol{v}\rangle + 2\operatorname{Re}\langle \boldsymbol{u}, \boldsymbol{G}_{12}\boldsymbol{v}\rangle$$

$$= \langle |\boldsymbol{A}\boldsymbol{x}|^2, |\boldsymbol{B}\boldsymbol{u}|^2 \rangle + \left\langle |\boldsymbol{B}\boldsymbol{h}|^2, |\boldsymbol{A}\boldsymbol{v}|^2 \right\rangle + 2\,\mathrm{Re}\langle \boldsymbol{u}, \boldsymbol{G}_{12}\boldsymbol{v} \rangle$$

Then, we have the following inequality:

$$\begin{aligned}
\mathrm{Re}\langle \boldsymbol{u}, \boldsymbol{G}_{12}\boldsymbol{v} \rangle &= \mathrm{Re}\langle \boldsymbol{u}, \boldsymbol{B}^{\mathsf{H}} \operatorname{diag}(\boldsymbol{B}\boldsymbol{h} \odot \boldsymbol{A}\boldsymbol{x})\overline{\boldsymbol{A}}\boldsymbol{v} \rangle + \mathrm{Re}\langle \boldsymbol{u}, \boldsymbol{B}^{\mathsf{H}} \operatorname{diag}(\boldsymbol{B}\boldsymbol{h} \odot \overline{\boldsymbol{A}\boldsymbol{x}} - \boldsymbol{y})\boldsymbol{A}\boldsymbol{v} \rangle \\
&= \mathrm{Re}\langle \boldsymbol{B}\boldsymbol{u} \odot \boldsymbol{A}\boldsymbol{v}, \boldsymbol{B}\boldsymbol{h} \odot \boldsymbol{A}\boldsymbol{x} \rangle + \mathrm{Re}\langle \boldsymbol{B}\boldsymbol{u} \odot \overline{\boldsymbol{A}\boldsymbol{v}}, \boldsymbol{B}\boldsymbol{h} \odot \overline{\boldsymbol{A}\boldsymbol{x}} - \boldsymbol{y} \rangle \\
&\leq 2 \sum_{j=1}^{m} \|\boldsymbol{b}_j\|_2^2 \|\boldsymbol{a}_j\|_2^2 \|\boldsymbol{h}\|_2 \|\boldsymbol{u}\|_2 \|\boldsymbol{x}\|_2 \|\boldsymbol{v}\|_2 + |\,\mathrm{Re}\langle \boldsymbol{B}\boldsymbol{u} \odot \overline{\boldsymbol{A}\boldsymbol{v}}, \boldsymbol{y} \rangle|,
\end{aligned}$$

where the last inequality holds by the Cauchy–Schwarz inequality. Because of the term $\|\boldsymbol{h}\|_2 \|\boldsymbol{u}\|_2 \|\boldsymbol{x}\|_2 \|\boldsymbol{v}\|_2$ *in the above inequality, it is difficult to find appropriate* ϕ *and* L *such that* $\langle \boldsymbol{w}, \nabla^2 f(\boldsymbol{h}, \boldsymbol{x})\boldsymbol{w} \rangle \leq \langle \boldsymbol{w}, \nabla^2 \phi(\boldsymbol{h}, \boldsymbol{x})\boldsymbol{w} \rangle$*. This is because* f *has a bilinear term.*

4.3.2 DC Decomposition

We first reformulate f in (4.17) into a DC function. Let us define convex functions $f_1, f_2 : \mathbb{R}^{d_1 \times d_2} \to (-\infty, +\infty]$ as follows:

$$\begin{aligned}
f_1(\boldsymbol{h}, \boldsymbol{x}) &= \frac{1}{4}\|\boldsymbol{B}\boldsymbol{h}\|_4^4 + \frac{1}{4}\|\boldsymbol{A}\boldsymbol{x}\|_4^4 + \frac{1}{2}\left(\|\boldsymbol{B}\boldsymbol{h} \odot \boldsymbol{A}\boldsymbol{x}\|_2^2 + \|\boldsymbol{y} \odot \boldsymbol{B}\boldsymbol{h}\|_2^2 + \|\boldsymbol{A}\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2\right), \\
f_2(\boldsymbol{h}, \boldsymbol{x}) &= \frac{1}{4}\|\boldsymbol{B}\boldsymbol{h}\|_4^4 + \frac{1}{4}\|\boldsymbol{A}\boldsymbol{x}\|_4^4 + \frac{1}{2}\|\bar{\boldsymbol{y}} \odot \boldsymbol{B}\boldsymbol{h} + \overline{\boldsymbol{A}\boldsymbol{x}}\|_2^2.
\end{aligned}$$

f_2 is convex because a composite function of a linear transform and a convex function is convex [87, Theorem 3.1.6]. Let $\boldsymbol{b}_j$ and $\boldsymbol{a}_j$ be the jth column vectors of $\boldsymbol{B}^{\mathsf{H}}$ and $\boldsymbol{A}^{\mathsf{H}}$, respectively, we have

$$f_1(\boldsymbol{h}, \boldsymbol{x}) = \frac{1}{4}\sum_{j=1}^{m}\left(|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}|^2 + |\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}|^2\right)^2 + \frac{1}{2}\left(\|\boldsymbol{y} \odot \boldsymbol{B}\boldsymbol{h}\|_2^2 + \|\boldsymbol{A}\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2\right),$$

which proves the convexity of f_1. Since $f = f_1 - f_2$ holds, (4.17) is equivalent to the following DC optimization problem:

$$\min_{(\boldsymbol{h}, \boldsymbol{x}) \in \mathrm{cl}\, C} \quad \Psi(\boldsymbol{h}, \boldsymbol{x}) = f_1(\boldsymbol{h}, \boldsymbol{x}) - f_2(\boldsymbol{h}, \boldsymbol{x}) + g(\boldsymbol{h}, \boldsymbol{x}). \tag{4.18}$$

4.3.3 L-smooth Adaptable Parameters

The following theorem provides an appropriate kernel generating distance ϕ and an appropriate parameter L for Algorithms 1 and 2.

Theorem 4.6. *Let a function* ϕ *be defined by*

$$\phi(\boldsymbol{h}, \boldsymbol{x}) = \frac{1}{4}\left(\|\boldsymbol{h}\|_2^2 + \|\boldsymbol{x}\|_2^2\right)^2 + \frac{1}{2}\left(\|\boldsymbol{h}\|_2^2 + \|\boldsymbol{x}\|_2^2\right). \tag{4.19}$$

Then, for any L satisfying

$$L \geq \sum_{j=1}^{m} (3\|\boldsymbol{b}_j\|_2^4 + 3\|\boldsymbol{a}_j\|_2^4 + \|\boldsymbol{b}_j\|_2^2\|\boldsymbol{a}_j\|_2^2 + |y_j|^2\|\boldsymbol{b}_j\|_2^2 + \|\boldsymbol{a}_j\|_2^2), \tag{4.20}$$

the function $L\phi - f_1$ is convex on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$.

Proof. We obtain the Hessian of ϕ and f_1 as follows:

$$\nabla^2\phi(\boldsymbol{z}) = (\|\boldsymbol{z}\|_2^2 + 1)\boldsymbol{I}_{d_1+d_2} + 2\boldsymbol{z}\boldsymbol{z}^{\mathsf{T}}, \quad \nabla^2 f_1(\boldsymbol{h}, \boldsymbol{x}) = \mathrm{Re}\begin{bmatrix} \boldsymbol{H}_{11} & \boldsymbol{H}_{12} \\ \boldsymbol{H}_{12}^{\mathsf{T}} & \boldsymbol{H}_{22} \end{bmatrix},$$

where $\boldsymbol{z} = (\boldsymbol{h}, \boldsymbol{x}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ and

$$\begin{aligned}
\boldsymbol{H}_{11} &:= \boldsymbol{B}^{\mathsf{H}} \operatorname{diag}(2|\boldsymbol{B}\boldsymbol{h}|^2 + |\boldsymbol{A}\boldsymbol{x}|^2 + |\boldsymbol{y}|^2)\boldsymbol{B} + \boldsymbol{B}^{\mathsf{H}} \operatorname{diag}((\boldsymbol{B}\boldsymbol{h})^2)\overline{\boldsymbol{B}}, \\
\boldsymbol{H}_{12} &:= \boldsymbol{B}^{\mathsf{H}} \operatorname{diag}(\boldsymbol{B}\boldsymbol{h} \odot \boldsymbol{A}\boldsymbol{x})\overline{\boldsymbol{A}} + \boldsymbol{B}^{\mathsf{H}} \operatorname{diag}(\boldsymbol{B}\boldsymbol{h} \odot \overline{\boldsymbol{A}\boldsymbol{x}})\boldsymbol{A}, \\
\boldsymbol{H}_{22} &:= \boldsymbol{A}^{\mathsf{H}} \operatorname{diag}(|\boldsymbol{B}\boldsymbol{h}|^2 + 2|\boldsymbol{A}\boldsymbol{x}|^2 + \boldsymbol{1}_m)\boldsymbol{A} + \boldsymbol{A}^{\mathsf{H}} \operatorname{diag}((\boldsymbol{A}\boldsymbol{x})^2)\overline{\boldsymbol{A}}.
\end{aligned}$$

Since the sum of a complex number and its conjugate is real, $\nabla^2 f_1$ is real. To prove the convexity of $L\phi - f_1$, it is sufficient to show that $\langle \boldsymbol{w}, \nabla^2 f_1(\boldsymbol{h}, \boldsymbol{x})\boldsymbol{w}\rangle \leq L\langle \boldsymbol{w}, \nabla^2\phi(\boldsymbol{h}, \boldsymbol{x})\boldsymbol{w}\rangle$ for any $\boldsymbol{w} \in \mathbb{R}^{d_1+d_2}$. Separate $\boldsymbol{w} = (\boldsymbol{u}, \boldsymbol{v})$ using $\boldsymbol{u} \in \mathbb{R}^{d_1}$ and $\boldsymbol{v} \in \mathbb{R}^{d_2}$. We obtain $\langle \boldsymbol{w}, \nabla^2\phi(\boldsymbol{z})\boldsymbol{w}\rangle = (\|\boldsymbol{z}\|_2^2 + 1)\|\boldsymbol{w}\|_2^2 + 2\langle \boldsymbol{z}, \boldsymbol{w}\rangle^2$ and

$$\langle \boldsymbol{w}, \nabla^2 f_1(\boldsymbol{h}, \boldsymbol{x})\boldsymbol{w}\rangle = \mathrm{Re}\langle \boldsymbol{u}, \boldsymbol{H}_{11}\boldsymbol{u}\rangle + \mathrm{Re}\langle \boldsymbol{v}, \boldsymbol{H}_{22}\boldsymbol{v}\rangle + 2\,\mathrm{Re}\langle \boldsymbol{u}, \boldsymbol{H}_{12}\boldsymbol{v}\rangle.$$

Each term of $\langle \boldsymbol{w}, \nabla^2 f_1(\boldsymbol{h}, \boldsymbol{x})\boldsymbol{w}\rangle$ is bounded as follows:

$$\begin{aligned}
\mathrm{Re}\langle \boldsymbol{u}, \boldsymbol{H}_{11}\boldsymbol{u}\rangle &= \langle 2|\boldsymbol{B}\boldsymbol{h}|^2 + |\boldsymbol{A}\boldsymbol{x}|^2 + |\boldsymbol{y}|^2, |\boldsymbol{B}\boldsymbol{u}|^2\rangle + \mathrm{Re}\langle (\boldsymbol{B}\boldsymbol{u})^2, (\boldsymbol{B}\boldsymbol{h})^2\rangle \\
&\leq \langle |\boldsymbol{B}\boldsymbol{h}|^2 + |\boldsymbol{A}\boldsymbol{x}|^2 + |\boldsymbol{y}|^2, |\boldsymbol{B}\boldsymbol{u}|^2\rangle + 2\langle |\boldsymbol{B}\boldsymbol{h}|^2, |\boldsymbol{B}\boldsymbol{u}|^2\rangle, \\
\mathrm{Re}\langle \boldsymbol{v}, \boldsymbol{H}_{22}\boldsymbol{v}\rangle &= \left\langle |\boldsymbol{B}\boldsymbol{h}|^2 + 2|\boldsymbol{A}\boldsymbol{x}|^2 + \boldsymbol{1}_m, |\boldsymbol{A}\boldsymbol{v}|^2 \right\rangle + \mathrm{Re}\langle (\boldsymbol{A}\boldsymbol{v})^2, (\boldsymbol{A}\boldsymbol{x})^2\rangle \\
&\leq \left\langle |\boldsymbol{B}\boldsymbol{h}|^2 + |\boldsymbol{A}\boldsymbol{x}|^2 + \boldsymbol{1}_m, |\boldsymbol{A}\boldsymbol{v}|^2 \right\rangle + 2\left\langle |\boldsymbol{A}\boldsymbol{x}|^2, |\boldsymbol{A}\boldsymbol{v}|^2 \right\rangle, \\
\mathrm{Re}\langle \boldsymbol{u}, \boldsymbol{H}_{12}\boldsymbol{v}\rangle &= \mathrm{Re}\langle \boldsymbol{u}, \boldsymbol{B}^{\mathsf{H}} \operatorname{diag}(\boldsymbol{B}\boldsymbol{h} \odot \boldsymbol{A}\boldsymbol{x})\overline{\boldsymbol{A}}\boldsymbol{v}\rangle + \mathrm{Re}\langle \boldsymbol{u}, \boldsymbol{B}^{\mathsf{H}} \operatorname{diag}(\boldsymbol{B}\boldsymbol{h} \odot \overline{\boldsymbol{A}\boldsymbol{x}})\boldsymbol{A}\boldsymbol{v}\rangle \\
&= \mathrm{Re}\langle \boldsymbol{B}\boldsymbol{u} \odot \boldsymbol{A}\boldsymbol{v}, \boldsymbol{B}\boldsymbol{h} \odot \boldsymbol{A}\boldsymbol{x}\rangle + \mathrm{Re}\langle \boldsymbol{B}\boldsymbol{u} \odot \overline{\boldsymbol{A}\boldsymbol{v}}, \boldsymbol{B}\boldsymbol{h} \odot \overline{\boldsymbol{A}\boldsymbol{x}}\rangle \\
&= \mathrm{Re}\sum_{j=1}^{m} \overline{\langle \boldsymbol{b}_j, \boldsymbol{u}\rangle\langle \boldsymbol{a}_j, \boldsymbol{v}\rangle}\langle \boldsymbol{b}_j, \boldsymbol{h}\rangle\langle \boldsymbol{a}_j, \boldsymbol{x}\rangle + \mathrm{Re}\sum_{j=1}^{m} \overline{\langle \boldsymbol{b}_j, \boldsymbol{u}\rangle}\langle \boldsymbol{a}_j, \boldsymbol{v}\rangle\langle \boldsymbol{b}_j, \boldsymbol{h}\rangle\overline{\langle \boldsymbol{a}_j, \boldsymbol{x}\rangle} \\
&\leq 2\sum_{j=1}^{m} \|\boldsymbol{b}_j\|_2^2\|\boldsymbol{a}_j\|_2^2\|\boldsymbol{h}\|_2\|\boldsymbol{u}\|_2\|\boldsymbol{x}\|_2\|\boldsymbol{v}\|_2,
\end{aligned}$$

where all the inequalities hold by $\mathrm{Re}(\cdot) \leq |\cdot|$, and the last inequality holds by the Cauchy–Schwarz inequality. From the above relation, we obtain

$$\langle |\boldsymbol{B}\boldsymbol{h}|^2, |\boldsymbol{B}\boldsymbol{u}|^2\rangle + \langle |\boldsymbol{A}\boldsymbol{x}|^2, |\boldsymbol{A}\boldsymbol{v}|^2\rangle + \mathrm{Re}\langle \boldsymbol{u}, \boldsymbol{H}_{12}\boldsymbol{v}\rangle$$

$$
\begin{aligned}
&\leq \sum_{j=1}^{m} \left(\|\boldsymbol{b}_j\|_2^4 \|\boldsymbol{h}\|_2^2 \|\boldsymbol{u}\|_2^2 + \|\boldsymbol{a}_j\|_2^4 \|\boldsymbol{x}\|_2^2 \|\boldsymbol{v}\|_2^2 + 2\|\boldsymbol{b}_j\|_2^2 \|\boldsymbol{a}_j\|_2^2 \|\boldsymbol{h}\|_2 \|\boldsymbol{u}\|_2 \|\boldsymbol{x}\|_2 \|\boldsymbol{v}\|_2 \right) \\
&= \sum_{j=1}^{m} \left(\|\boldsymbol{b}_j\|_2^2 \|\boldsymbol{h}\|_2 \|\boldsymbol{u}\|_2 + \|\boldsymbol{a}_j\|_2^2 \|\boldsymbol{x}\|_2 \|\boldsymbol{v}\|_2 \right)^2 \\
&\leq \sum_{j=1}^{m} \left(\|\boldsymbol{b}_j\|_2^4 + \|\boldsymbol{a}_j\|_2^4 \right) \left(\|\boldsymbol{h}\|_2^2 \|\boldsymbol{u}\|_2^2 + \|\boldsymbol{x}\|_2^2 \|\boldsymbol{v}\|_2^2 \right),
\end{aligned}
$$

where both inequalities hold by the Cauchy–Schwarz inequality. Thus, we obtain

$$
\begin{aligned}
&\langle \boldsymbol{w}, \nabla^2 f_1(\boldsymbol{h}, \boldsymbol{x}) \boldsymbol{w} \rangle \\
&\leq \langle |\boldsymbol{B}\boldsymbol{h}|^2 + |\boldsymbol{A}\boldsymbol{x}|^2 + |\boldsymbol{y}|^2, |\boldsymbol{B}\boldsymbol{u}|^2 \rangle + \langle |\boldsymbol{B}\boldsymbol{h}|^2 + |\boldsymbol{A}\boldsymbol{x}|^2 + \mathbf{1}_m, |\boldsymbol{A}\boldsymbol{v}|^2 \rangle \\
&\quad + 2\langle |\boldsymbol{B}\boldsymbol{h}|^2, |\boldsymbol{B}\boldsymbol{u}|^2 \rangle + 2\langle |\boldsymbol{A}\boldsymbol{x}|^2, |\boldsymbol{A}\boldsymbol{v}|^2 \rangle + 2\,\mathrm{Re}\langle \boldsymbol{u}, \boldsymbol{H}_{12}\boldsymbol{v} \rangle \\
&\leq \sum_{j=1}^{m} \big(\|\boldsymbol{b}_j\|_2^2 \|\boldsymbol{u}\|_2^2 (\|\boldsymbol{b}_j\|_2^2 \|\boldsymbol{h}\|_2^2 + \|\boldsymbol{a}_j\|_2^2 \|\boldsymbol{x}\|_2^2 + |y_j|^2) \\
&\quad + \|\boldsymbol{a}_j\|_2^2 \|\boldsymbol{v}\|_2^2 (\|\boldsymbol{b}_j\|_2^2 \|\boldsymbol{h}\|_2^2 + \|\boldsymbol{a}_j\|_2^2 \|\boldsymbol{x}\|_2^2 + 1) \\
&\quad + 2(\|\boldsymbol{b}_j\|_2^4 + \|\boldsymbol{a}_j\|_2^4)(\|\boldsymbol{h}\|_2^2 \|\boldsymbol{u}\|_2^2 + \|\boldsymbol{x}\|_2^2 \|\boldsymbol{v}\|_2^2) \big) \\
&\leq \sum_{j=1}^{m} (3\|\boldsymbol{b}_j\|_2^4 + 3\|\boldsymbol{a}_j\|_2^4 + \|\boldsymbol{b}_j\|_2^2 \|\boldsymbol{a}_j\|_2^2 + |y_j|^2 \|\boldsymbol{b}_j\|_2^2 + \|\boldsymbol{a}_j\|_2^2)(\|\boldsymbol{z}\|_2^2 + 1) \|\boldsymbol{w}\|_2^2 \\
&\leq L \langle \boldsymbol{w}, \nabla^2 \phi(\boldsymbol{h}, \boldsymbol{x}) \boldsymbol{w} \rangle,
\end{aligned}
$$

which proves $L\phi - f_1$ is convex. □

From Theorem 4.6, we obtain the following.

Corollary 4.7. *Let a function $\phi_C \in \mathcal{G}(C)$ be defined by $\phi_C(\boldsymbol{z}) = \frac{1}{4}\|\boldsymbol{z}\|_2^4 + \frac{1}{2}\|\boldsymbol{z}\|_2^2 + \delta_C(\boldsymbol{z})$. For any L satisfying (4.20), the pair (f_1, ϕ_C) is L-smad on C.*

Proof. Because C is an open set, ϕ_C and $\nabla^2 \phi_C$ are the same as these of ϕ given by (4.19) on C. From the convexity of C and Theorem 4.6, the pair (f_1, ϕ_C) is L-smad on C. □

From Corollary 4.7, we can immediately prove the following corollary by using Theorems 3.14 and 3.26.

Corollary 4.8. *Let $\{\boldsymbol{z}^k\}_{k=0}^{\infty}$ be a sequence generated by BPDCA(e) with $0 < \lambda L < 1$ for (4.18). Assume that Assumption 3.1 (iv) and (vi) holds for g. For BPDCAe, assume that Assumption 3.20 holds. Then, $\{\boldsymbol{z}^k\}_{k=0}^{\infty}$ converges to a limiting stationary point of (4.18).*

Proof. Since Assumption 3.1 (iv) and (vi) hold, for the function f_1, f_2, g, and ϕ, Assumptions 3.1, 3.7, 3.12, and 3.25 hold. Assumptions 3.2 and 3.18 hold for ϕ_C, instead of ϕ. Moreover, even if we restrict to the set C, we can assume without loss of generality that Assumption 3.7 (i) holds. Therefore, from Corollary 4.7, Theorem 3.14 holds. For BPDCAe, from the convexity of g, Theorem 3.26 holds. □

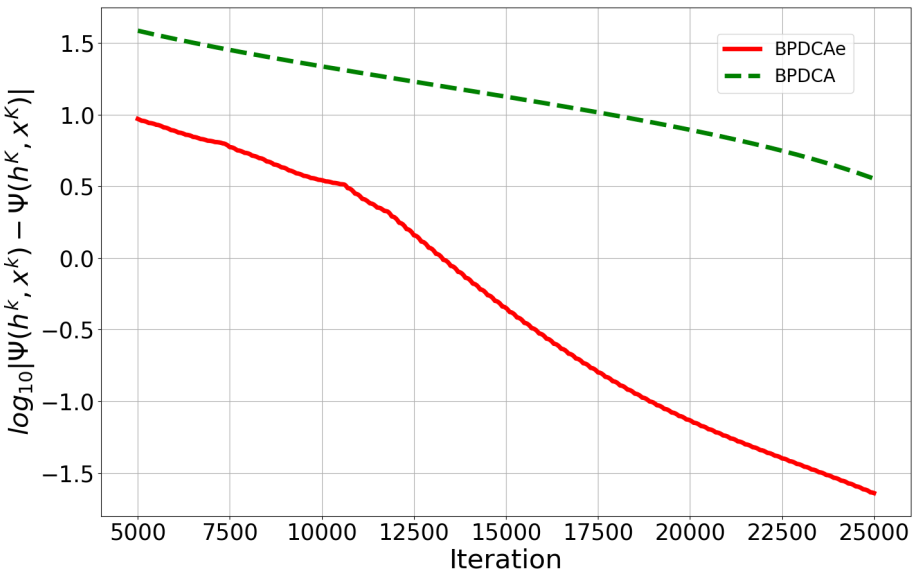


Figure 4.3: Plots of $\{\log_{10}|\Psi(\boldsymbol{h}^k, \boldsymbol{x}^k) - \Psi(\boldsymbol{h}^K, \boldsymbol{x}^K)|\}$ at each iteration.

In practice, we can obtain a closed-form solution of (3.2) and (3.32) from [15, Proposition 5.1] for $g(\boldsymbol{h}, \boldsymbol{x}) = \theta_1\|\boldsymbol{h}\|_1 + \theta_2\|\boldsymbol{x}\|_1$.

Remark 4.9. *For instance, we solve subproblem (3.2) to obtain the sparse signal and filter when $g(\boldsymbol{h}, \boldsymbol{x}) = \theta_1\|\boldsymbol{h}\|_1 + \theta_2\|\boldsymbol{x}\|_1$ for $\theta_1, \theta_2 \geq 0$ and ϕ is given by (4.19). Let $\boldsymbol{u}^k = \mathcal{S}_{\lambda\theta_1}(\lambda\nabla_{\boldsymbol{h}} f(\boldsymbol{z}^k) - \nabla_{\boldsymbol{h}}\phi(\boldsymbol{z}^k))$ and $\boldsymbol{v}^k = \mathcal{S}_{\lambda\theta_2}(\lambda\nabla_{\boldsymbol{x}} f(\boldsymbol{z}^k) - \nabla_{\boldsymbol{x}}\phi(\boldsymbol{z}^k))$. We can prove from [15, Proposition 5.1] that $\boldsymbol{z}^{k+1} = (-t^*\boldsymbol{u}^k, -t^*\boldsymbol{v}^k)$ solves subproblem (3.2), where t^* is the unique positive real root of the cubic equation $t^3(\|\boldsymbol{u}^k\|_2^2 + \|\boldsymbol{v}^k\|_2^2) + t - 1 = 0$. Note that every cubic equation has a closed-form solution via Cardano's formula. This fact indicates the solution of (3.2) is expressed in closed form. It is also true for subproblem (3.32).*

The ℓ_1 regularization term in numerical experiments satisfies Assumption 3.1 (iv) and (vi) and Assumption 3.20. In regard to the rate of convergence, from the proof of Corollary 4.8, Theorems 3.16 and 3.28 hold. However, the KL exponent of blind deconvolution has not been yet known. The difference between the objective value at each iteration and that at the convergence point $\{\log_{10}|\Psi(\boldsymbol{h}^k, \boldsymbol{x}^k) - \Psi(\boldsymbol{h}^K, \boldsymbol{x}^K)|\}$ is plotted in Figure 4.3 in log-scale, where we recall $\Psi = f + g$ and $K = 30000$ (see, for other settings, Section 4.3.5). Here, the first and the last 5000 iterations are trimmed because the difference in the iterations is too large. Figure 4.3 shows that BPDCA(e) converged linearly. Hence, we expect the KL exponent of blind deconvolution to belong to $(0, \frac{1}{2}]$. Calculation of the exact value of the KL exponent is left for future work.

Note that a limiting stationary point of (4.18) is a point $\boldsymbol{z} \in C$ satisfying $\mathbf{0}_{d_1+d_2} \in \partial\Psi(\boldsymbol{z}) = \nabla f_1(\boldsymbol{z}) - \nabla f_2(\boldsymbol{z}) + \partial g(\boldsymbol{z})$ from the smoothness of f_2 and Definition 3.8. Note that under the convexity of g, it is theoretically guaranteed that a sequence generated by FISTA would converge to the limiting stationary point if f had a Lipschitz continuous gradient. In our problem, the convergence of FISTA is not theoretically guaranteed

because f does not have it. Although FISTA is not applicable, the convergent points of BPDCA(e) and FISTA share the same stationary points in theory.

Remark 4.10. *An appropriate value of λ can be evaluated by backtracking, i.e., decrease λ until $f_1(\boldsymbol{z}^{k+1}) - f_1(\boldsymbol{z}^k) - \langle \nabla f_1(\boldsymbol{z}^k), \boldsymbol{z}^{k+1} - \boldsymbol{z}^k \rangle \leq \frac{1}{\lambda} D_\phi(\boldsymbol{z}^{k+1}, \boldsymbol{z}^k)$ is satisfied, because, if this inequality holds, the theoretical convergence is guaranteed. For n_k backtracking procedures at the kth iteration, we need one calculation of $\nabla f_1(\boldsymbol{z}^k)$ and n_k calculations of $f_1(\boldsymbol{z}^{k+1})$. These calculations are sometimes expensive. Thus, we did not use backtracking.*

4.3.4 Stability Analysis

We show the stability analysis of BPDCA(e), the proximal gradient method, and the alternating minimization (AM) [27, 54, 95] around the equilibrium points, which are the fixed points of the update formula of each algorithm.

Definition 4.11. *Let $\mathcal{T}$ be a mapping of some algorithm. The point $\boldsymbol{x} \in \mathbb{R}^d$ is called an equilibrium point if $\boldsymbol{x} = \mathcal{T}(\boldsymbol{x})$.*

For example, the mapping $\mathcal{T}$ of BPDCA is $\mathcal{T}_\lambda$, defined by

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

For the property of $\mathcal{T}_\lambda$, see also Section 3.1. Because an equilibrium point $\boldsymbol{z} = (\boldsymbol{h}, \boldsymbol{x})$ of BPDCA satisfies

$$\mathbf{0}_{d_1+d_2} \in \nabla f(\boldsymbol{z}) + \partial g(\boldsymbol{z}) + \nabla \phi(\boldsymbol{z}) - \nabla \phi(\boldsymbol{z}) = \nabla f(\boldsymbol{z}) + \partial g(\boldsymbol{z}),$$

an equilibrium point of BPDCA corresponds to a limiting stationary point of (4.17). Note that an equilibrium point of AM does not always correspond to a limiting stationary point of (4.17). See also Remark 4.12.

Here, we define $\Phi(\boldsymbol{z}) = \langle \nabla f(\boldsymbol{z}^k), \boldsymbol{z} \rangle + g(\boldsymbol{z}) + \frac{1}{\lambda} D_\phi(\boldsymbol{z}, \boldsymbol{z}^k)$. Then, the first-order condition $\mathbf{0}_{d_1+d_2} \in \partial \Phi(\boldsymbol{z}^{k+1}) = \nabla f(\boldsymbol{z}^k) + \partial g(\boldsymbol{z}^{k+1}) + \frac{1}{\lambda} \left(\nabla \phi(\boldsymbol{z}^{k+1}) - \nabla \phi(\boldsymbol{z}^k) \right)$ is approximated as follows:

$$\mathbf{0}_{d_1+d_2} \simeq \nabla f(\boldsymbol{z}^k) + \boldsymbol{\zeta}^{k+1} + \frac{1}{\lambda} \nabla^2 \phi(\boldsymbol{z}^k)(\boldsymbol{z}^{k+1} - \boldsymbol{z}^k),$$

where $\boldsymbol{\zeta}^{k+1} \in \partial g(\boldsymbol{z}^{k+1})$. Note that $\nabla^2 \phi(\boldsymbol{z}^k)$ is nonsingular. In fact, its inverse is explicitly written as

$$\nabla^2 \phi(\boldsymbol{z})^{-1} = \frac{1}{\|\boldsymbol{z}\|_2^2 + 1} \left(\boldsymbol{I}_{d_1+d_2} - \frac{2 \boldsymbol{z} \boldsymbol{z}^\mathsf{T}}{3\|\boldsymbol{z}\|_2^2 + 1} \right).$$

By multiplying it, we obtain

$$\boldsymbol{z}^{k+1} - \boldsymbol{z}^k \simeq -\lambda \nabla^2 \phi(\boldsymbol{z}^k)^{-1} (\nabla f(\boldsymbol{z}^k) + \boldsymbol{\zeta}^{k+1}), \tag{4.21}$$

which indicates that $\boldsymbol{z}^{k+1} - \boldsymbol{z}^k$ is greatly affected by $\nabla^2\phi(\boldsymbol{z})^{-1}$. Thus, ϕ is important for the performance of BPDCA. For BPDCAe, this fact is also true by substituting $\boldsymbol{w}^k$ for $\boldsymbol{z}^k$.

To simplify the stability analysis of FISTA, we consider the proximal gradient method. The proximal gradient method is called FISTA when $g = \|\cdot\|_1$ with extrapolation. Each iteration of the proximal gradient method computes the following subproblem:

$$\boldsymbol{z}^{k+1} = \operatorname*{argmin}_{\boldsymbol{z}\in \mathrm{cl}\, C} \left\{ \langle \nabla f(\boldsymbol{z}^k), \boldsymbol{z} - \boldsymbol{z}^k \rangle + g(\boldsymbol{z}) + \frac{1}{\lambda}\|\boldsymbol{z} - \boldsymbol{z}^k\|_2^2 \right\}. \tag{4.22}$$

The right-hand side of (4.22) corresponds with the mapping of the proximal gradient method. Setting $\phi(\boldsymbol{z}) = \frac{1}{2}\|\boldsymbol{z}\|_2^2 =: \phi_1(\boldsymbol{z})$, *i.e.*, $\nabla^2\phi_1(\boldsymbol{z}) = \boldsymbol{I}_{d_1+d_2}$, we obtain

$$\boldsymbol{z}^{k+1} - \boldsymbol{z}^k \simeq -\lambda(\nabla f(\boldsymbol{z}^k) + \boldsymbol{\zeta}^{k+1}).$$

Since f does not have a Lipschitz continuous gradient, λ is close to 0, *i.e.*, $\boldsymbol{z}^k \simeq \boldsymbol{z}^{k+1}$. This implies that convergence of the proximal gradient method and FISTA is slow.

When $g(\boldsymbol{x}, \boldsymbol{h})$ is convex, (4.17) is convex with respect to $\boldsymbol{h}$ for fixed $\boldsymbol{x}$ and vice versa. AM is a method to update $\boldsymbol{h}$ and $\boldsymbol{x}$ alternately, *i.e.*,

$$\begin{aligned}
\boldsymbol{h}^{k+1} &= \operatorname*{argmin}_{\boldsymbol{h}\in \mathrm{cl}\, C_{\boldsymbol{h}}} \left\{ f(\boldsymbol{h}, \boldsymbol{x}^k) + g(\boldsymbol{h}, \boldsymbol{x}^k) \right\}, \\
\boldsymbol{x}^{k+1} &= \operatorname*{argmin}_{\boldsymbol{x}\in \mathrm{cl}\, C_{\boldsymbol{x}}} \left\{ f(\boldsymbol{h}^{k+1}, \boldsymbol{x}) + g(\boldsymbol{h}^{k+1}, \boldsymbol{x}) \right\},
\end{aligned}$$

where $C_{\boldsymbol{h}} = \{\boldsymbol{h} \in \mathbb{R}^{d_1} \mid (\boldsymbol{h}, \boldsymbol{x}) \in C\}$ and $C_{\boldsymbol{x}} = \{\boldsymbol{x} \in \mathbb{R}^{d_2} \mid (\boldsymbol{h}, \boldsymbol{x}) \in C\}$. The first-order conditions around the equilibrium points are

$$\begin{aligned}
\boldsymbol{0}_{d_1} &\in \nabla_{\boldsymbol{h}} f(\boldsymbol{h}^{k+1}, \boldsymbol{x}^k) + \partial_{\boldsymbol{h}} g(\boldsymbol{h}^{k+1}, \boldsymbol{x}^k) \\
&\simeq \nabla_{\boldsymbol{h}} f(\boldsymbol{h}^k, \boldsymbol{x}^k) + \nabla^2_{\boldsymbol{hh}} f(\boldsymbol{h}^k, \boldsymbol{x}^k)(\boldsymbol{h}^{k+1} - \boldsymbol{h}^k) + \partial_{\boldsymbol{h}} g(\boldsymbol{h}^{k+1}, \boldsymbol{x}^k), \\
\boldsymbol{0}_{d_2} &\in \nabla_{\boldsymbol{x}} f(\boldsymbol{h}^{k+1}, \boldsymbol{x}^{k+1}) + \partial_{\boldsymbol{x}} g(\boldsymbol{h}^{k+1}, \boldsymbol{x}^{k+1}) \\
&\simeq \nabla_{\boldsymbol{x}} f(\boldsymbol{h}^{k+1}, \boldsymbol{x}^k) + \nabla^2_{\boldsymbol{xx}} f(\boldsymbol{h}^k, \boldsymbol{x}^k)(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) + \partial_{\boldsymbol{x}} g(\boldsymbol{h}^{k+1}, \boldsymbol{x}^{k+1}),
\end{aligned}$$

where the last approximation holds from $\boldsymbol{h}^{k+1} \simeq \boldsymbol{h}^k$ in $\nabla^2_{\boldsymbol{xx}} f(\boldsymbol{h}^k, \boldsymbol{x}^k)$. Assuming that the Hessians $\nabla^2_{\boldsymbol{hh}} f(\boldsymbol{h}^k, \boldsymbol{x}^k)$ and $\nabla^2_{\boldsymbol{xx}} f(\boldsymbol{h}^k, \boldsymbol{x}^k)$ are regular, for $\boldsymbol{\zeta}_{\boldsymbol{h}}^{k+1} \in \partial_{\boldsymbol{h}} g(\boldsymbol{h}^{k+1}, \boldsymbol{x}^k)$ and $\boldsymbol{\zeta}_{\boldsymbol{x}}^{k+1} \in \partial_{\boldsymbol{x}} g(\boldsymbol{h}^{k+1}, \boldsymbol{x}^{k+1})$, we obtain

$$\begin{aligned}
\boldsymbol{h}^{k+1} - \boldsymbol{h}^k &\simeq -\nabla^2_{\boldsymbol{hh}} f(\boldsymbol{h}^k, \boldsymbol{x}^k)^{-1}(\nabla_{\boldsymbol{h}} f(\boldsymbol{h}^k, \boldsymbol{x}^k) + \boldsymbol{\zeta}_{\boldsymbol{h}}^{k+1}), \\
\boldsymbol{x}^{k+1} - \boldsymbol{x}^k &\simeq -\nabla^2_{\boldsymbol{xx}} f(\boldsymbol{h}^k, \boldsymbol{x}^k)^{-1}(\nabla_{\boldsymbol{x}} f(\boldsymbol{h}^{k+1}, \boldsymbol{x}^k) + \boldsymbol{\zeta}_{\boldsymbol{x}}^{k+1}).
\end{aligned}$$

$\nabla^2\phi(\boldsymbol{z}^k)$ in the approximation of BPDCA is a block matrix that contains the cross derivative terms $\nabla^2_{\boldsymbol{hx}}\phi(\boldsymbol{z}^k)$ and $\nabla^2_{\boldsymbol{xh}}\phi(\boldsymbol{z}^k)$, while the perturbation around the equilibrium points of AM is approximated only with $\nabla^2_{\boldsymbol{xx}} f(\boldsymbol{h}^k, \boldsymbol{x}^k)$ and $\nabla^2_{\boldsymbol{hh}}\phi(\boldsymbol{h}^k, \boldsymbol{x}^k)$ regardless of the cross derivatives. Thus, an equilibrium point of AM is not necessarily that of BPDCA. This implies that BPDCA is not trapped at the points where AM is stuck. On the other hand, every equilibrium point of BPDCA is an equilibrium point of AM under the Clarke regularity as we mention below.

Remark 4.12. *Here, we assume that g is Clarke regular at $\boldsymbol{z} = (\boldsymbol{h}, \boldsymbol{x}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$. In this case, because g is convex, g is Clarke regular. Then, it holds that $\partial_{\mathrm{c}} g(\boldsymbol{h}, \boldsymbol{x}) \subset \partial_{\boldsymbol{h}} g(\boldsymbol{h}, \boldsymbol{x}) \times \partial_{\boldsymbol{x}} g(\boldsymbol{h}, \boldsymbol{x})$* [31, Proposition 2.3.15]. *Hence, for an equilibrium point $\boldsymbol{z} = (\boldsymbol{h}, \boldsymbol{x})$ of BPDCA, we have*

$$\begin{aligned} \mathbf{0}_{d_1+d_2} &\in \nabla f(\boldsymbol{z}) + \partial_{\mathrm{c}} g(\boldsymbol{z}) + \nabla \phi(\boldsymbol{z}) - \nabla \phi(\boldsymbol{z}) = \nabla f(\boldsymbol{z}) + \partial_{\mathrm{c}} g(\boldsymbol{z}) \\ &\subset (\nabla_{\boldsymbol{h}} f(\boldsymbol{h}, \boldsymbol{x}), \nabla_{\boldsymbol{x}} f(\boldsymbol{h}, \boldsymbol{x})) + \partial_{\boldsymbol{h}} g(\boldsymbol{h}, \boldsymbol{x}) \times \partial_{\boldsymbol{x}} g(\boldsymbol{h}, \boldsymbol{x}). \end{aligned}$$

Therefore, an equilibrium point of BPDCA is also an equilibrium point of AM. Additionally, an equilibrium point of BPDCA is always a limiting stationary point, while that of AM is not always a limiting stationary point. They are demonstrated in numerical experiments (Figure 4.7).

4.3.5 Numerical Experiments: Setting

We demonstrated the efficiency of our proposed method by image deblurring via solving problem (4.17). We set $d_1 = 2304$, $d_2 = 65536$, and $m = 262144$ and appropriately took a ground truth of $(\boldsymbol{h}^\circ, \boldsymbol{x}^\circ)$. Using them, we generated a blurring kernel $\boldsymbol{f} = \tilde{\boldsymbol{B}}\boldsymbol{h}^\circ$ and an original image $\boldsymbol{g} = \tilde{\boldsymbol{A}}\boldsymbol{x}^\circ$, where $\tilde{\boldsymbol{B}} : \mathbb{R}^{d_1} \to \mathbb{R}^m$ is an operator reshaping $\boldsymbol{h} \in \mathbb{R}^{d_1}$ into a $\sqrt{m} \times \sqrt{m}$ image and $\tilde{\boldsymbol{A}} : \mathbb{R}^{d_2} \to \mathbb{R}^m$ is an inverse discrete Meyer wavelet transform operator. Figure 4.4 depicts $\boldsymbol{f}$ and $\boldsymbol{g}$ used in our experiments: $\boldsymbol{f}$ corresponds to a diagonal blurring and $\boldsymbol{g}$ approximates a natural image. $\boldsymbol{g}$ is generated from a grayscale image of the original images by ESA/Hubble.[1] We also set $C = \{(\boldsymbol{h}, \boldsymbol{x}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \mid \boldsymbol{h} > \mathbf{0}_{d_1}, \boldsymbol{x} > \mathbf{0}_{d_2}\}$ and $g(\boldsymbol{h}, \boldsymbol{x}) = \theta \|\boldsymbol{h}\|_1$ with $\theta = 0.01$ (the nonsmooth ℓ_1 regularizer) because $\boldsymbol{h}$ is supposed to be sparse in the practice of image deblurring.

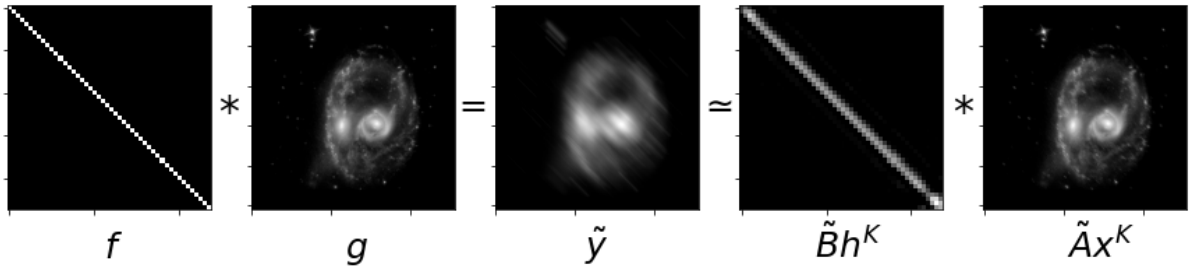


Figure 4.4: The ground truth $\boldsymbol{f}$ and $\boldsymbol{g}$, the blurred image $\tilde{\boldsymbol{y}}$, and $\boldsymbol{B}\boldsymbol{h}^K$ and $\boldsymbol{A}\boldsymbol{x}^K$ recovered by BPDCAe.

4.3.6 Numerical Experiments: Comparison of ℓ_1 and ℓ_2 Regularization

We solved problem (4.17) corresponding to the setting above with BPDCA(e), FISTA, and AM. For all methods, the initial points $\boldsymbol{h}^0$ and $\boldsymbol{x}^0$ are set to be the left and right

[1] The original images are available in https://hubblesite.org/contents/media/images/2019/51/4574-Image and https://hubblesite.org/contents/media/images/2009/25/2616-Image.

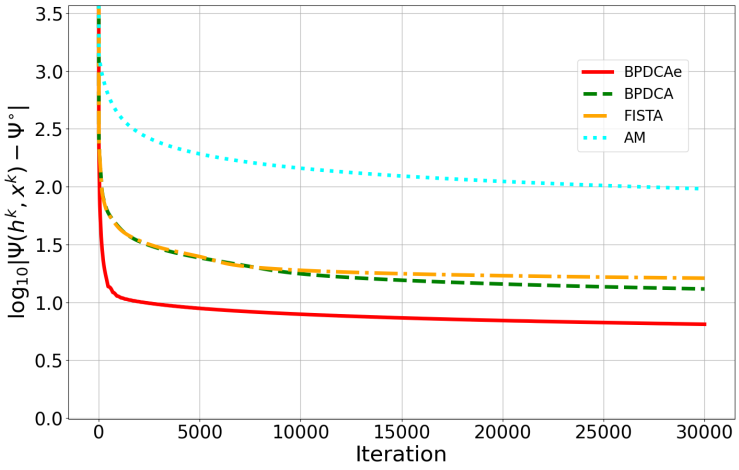
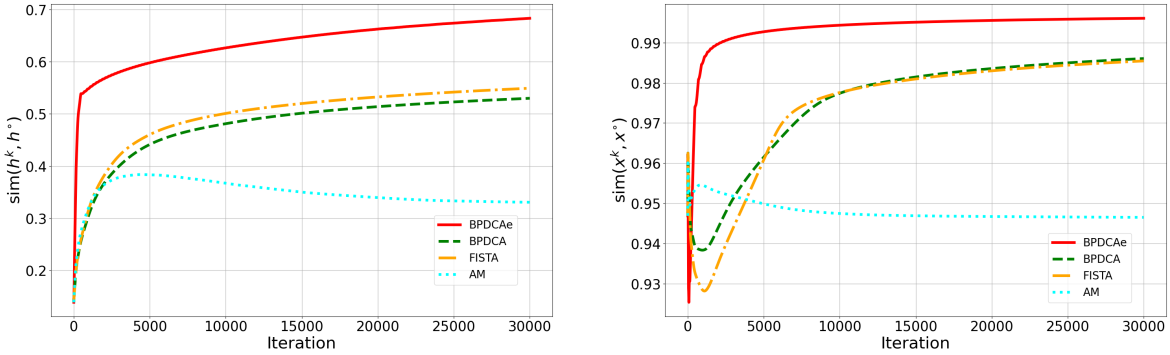


Figure 4.5: Plots of $\{\log_{10}|\Psi(\boldsymbol{h}^k, \boldsymbol{x}^k) - \Psi^\circ|\}$ at each iteration.

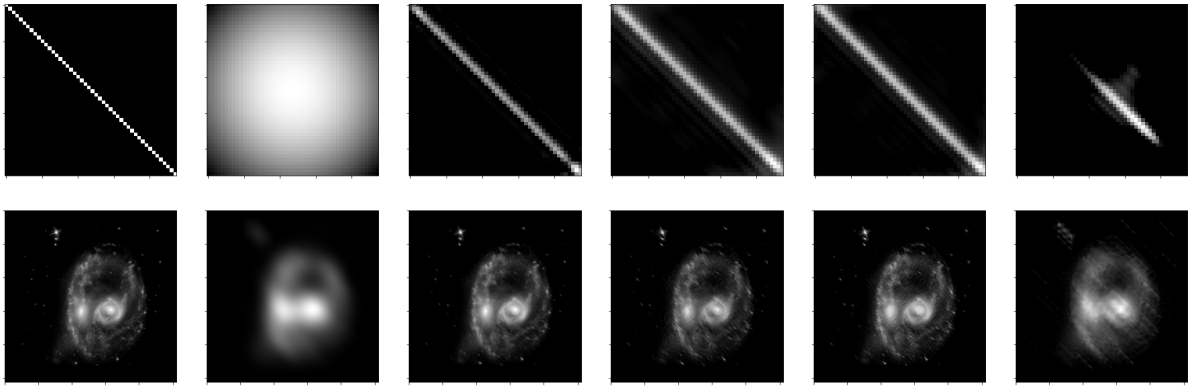


(a) Plots of $\{\mathrm{sim}(\boldsymbol{h}^k, \boldsymbol{h}^\circ)\}$.

(b) Plots of $\{\mathrm{sim}(\boldsymbol{x}^k, \boldsymbol{x}^\circ)\}$.

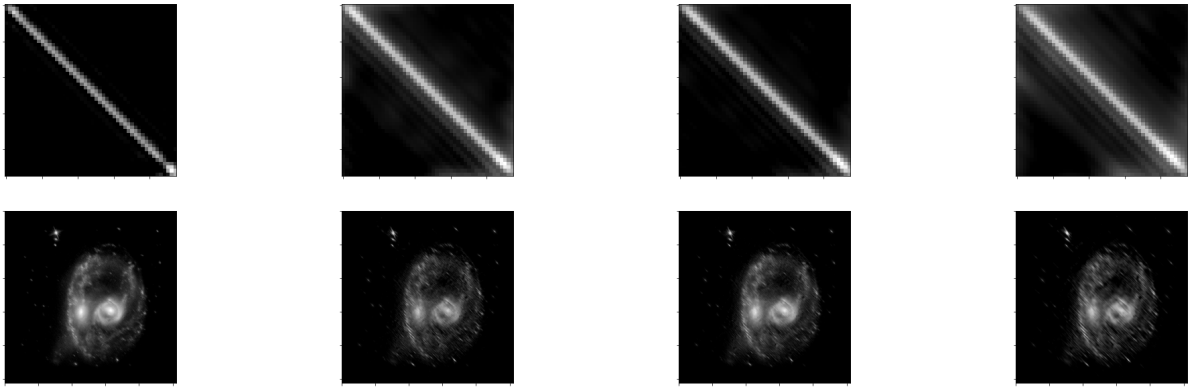
Figure 4.6: Plots of the cosine similarities between the kth point and the ground truth.

singular vectors corresponding to the leading singular value of $\boldsymbol{B}^{\mathsf{H}}\,\mathrm{diag}(\boldsymbol{y})\overline{\boldsymbol{A}}$, respectively, which is proposed in [69]. For BPDCA(e), we adjusted L that satisfies (4.20) and used it as a fixed step size. Step sizes in all iterations of FISTA were obtained by backtracking. Note that the subproblems of BPDCA(e) and FISTA (without backtracking procedures) are solved in closed-form formulae, whose computational cost is almost the same. The maximum number of iterations for BPDCA(e) and FISTA was 30000, and that for AM was 3000 because the subproblems of AM were solved approximately by 10 iterations of FISTA at each iteration. In the following figures, we plot the results of AM every 10 iterations. The difference between the objective value at each iteration and that at the ground truth $\{\log_{10}|\Psi(\boldsymbol{h}^k, \boldsymbol{x}^k) - \Psi^\circ|\}$ is plotted in Figure 4.5 in log-scale, where we recall $\Psi = f + g$ and $\Psi^\circ := \Psi(\boldsymbol{h}^\circ, \boldsymbol{x}^\circ)$. Figure 4.6a shows the cosine similarity between $\boldsymbol{h}^k$ and $\boldsymbol{h}^\circ$ defined by $\{\mathrm{sim}(\boldsymbol{h}^k, \boldsymbol{h}^\circ)(:= \langle \boldsymbol{h}^k, \boldsymbol{h}^\circ\rangle/(\|\boldsymbol{h}^k\|_2\|\boldsymbol{h}^\circ\|_2))\}$, and Figure 4.6b shows that $\mathrm{sim}(\boldsymbol{x}^k, \boldsymbol{x}^\circ)$. As we can see from Figures 4.5 and 4.6, BPDCAe outperformed the other algorithms. Its convergence was the fastest, and $\Psi(\boldsymbol{h}^K, \boldsymbol{x}^K)$, $\mathrm{sim}(\boldsymbol{h}^K, \boldsymbol{h}^\circ)$, and $\mathrm{sim}(\boldsymbol{x}^K, \boldsymbol{x}^\circ)$ were also the best, where $K := 30000$ (for BPDCA(e) and FISTA) or



(a) $\boldsymbol{h}^\circ$ and $\boldsymbol{x}^\circ$ (b) $\boldsymbol{h}^0$ and $\boldsymbol{x}^0$ (c) BPDCAe (d) BPDCA (e) FISTA (f) AM

Figure 4.7: The upper row shows $\boldsymbol{h}^K$, and the lower row shows $\tilde{\boldsymbol{A}}\boldsymbol{x}^K$.



(a) $\theta\|\boldsymbol{h}\|_1$ (b) $\theta\|\boldsymbol{h}\|_2^2$ (c) $\theta\|\boldsymbol{h}\|_1$ (d) $\theta\|\boldsymbol{h}\|_2^2$

Figure 4.8: (a–b) BPDCAe and (c–d) BPDCA.

$K := 3000$ (for AM). Figure 4.7 shows the recovered images. Figure 4.7c shows that there is almost no difference between $\tilde{\boldsymbol{A}}\boldsymbol{x}^\circ$ and $\tilde{\boldsymbol{A}}\boldsymbol{x}^k$, while $\boldsymbol{h}^k$ was not completely recovered. Figures 4.7c and 4.7f show that the sequences generated by AM converged to a different stationary point (see also Section 4.3.4). Figures 4.7c and 4.7d imply that BPDCA(e) might converge to a stronger point, such as a local optimal point or a directional stationary point (see also [34, Definition 6.1.1 and Proposition 6.1.8]), than a limiting stationary point.

We also solved the deblurring problem with the ℓ_2 regularization term $g(\boldsymbol{h}, \boldsymbol{x}) = \theta\|\boldsymbol{h}\|_2^2$ with $\theta = 0.01$. The comparison between the results from these two regularizers is shown in Figure 4.8. It shows the superiority of the nonsmooth ℓ_1 regularization term over the ℓ_2 one, which did not recover the sparse blurring kernel.

4.3.7 Numerical Experiments: Comparisons under Several Situations

We first demonstrate that the efficiency of our proposed methods is independent of the choice of the initial point. To do so, we generated the initial points $\boldsymbol{h}^0$ and $\boldsymbol{x}^0$ from the uniform distribution on $[0, 0.1]$. From these initial points, we generated $\{\boldsymbol{h}^k\}$ and $\{\boldsymbol{x}^k\}$ by each algorithm. Figures 4.9, 4.10a, and 4.10b show $\{\log_{10}|\Psi(\boldsymbol{h}^k, \boldsymbol{x}^k) - \Psi^\circ|\}$, $\{\mathrm{sim}(\boldsymbol{h}^k, \boldsymbol{h}^\circ)\}$, and $\{\mathrm{sim}(\boldsymbol{x}^k, \boldsymbol{x}^\circ)\}$, respectively. Figure 4.11 shows the recovered images, and Figure 4.11c shows that there is almost no difference between $\tilde{\boldsymbol{A}}\boldsymbol{x}^\circ$ and $\tilde{\boldsymbol{A}}\boldsymbol{x}^K$ at this case. As we can see from these figures, BPDCAe also outperformed the other algorithms even when $\boldsymbol{h}^0$ and $\boldsymbol{x}^0$ are generated from the uniform distribution.

We next demonstrate the efficiency of our proposed methods with noisy data. Here, we consider $\tilde{\boldsymbol{y}}$ containing Poisson noise, *i.e.*, $\tilde{\boldsymbol{y}} = \boldsymbol{f} * \boldsymbol{g} + \boldsymbol{n}$, where $\boldsymbol{n} \in \mathbb{R}^m$ is Poisson noise (see Figure 4.13a). By changing the noise level, *i.e.*, the standard deviation, we solved the image deblurring problem with each algorithm. Figure 4.12 shows the objective value

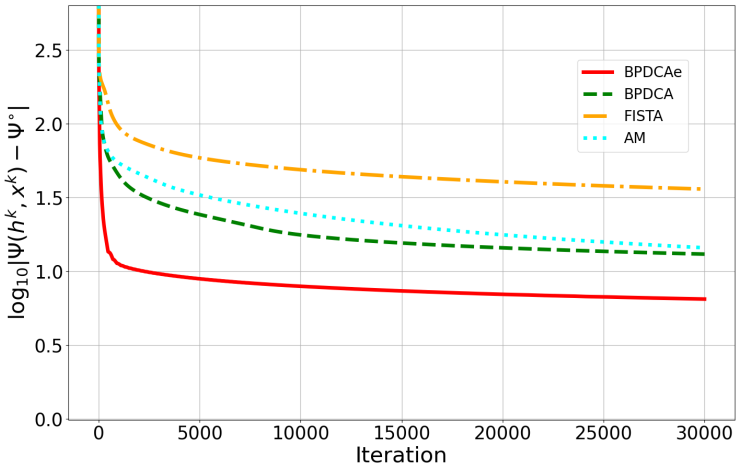
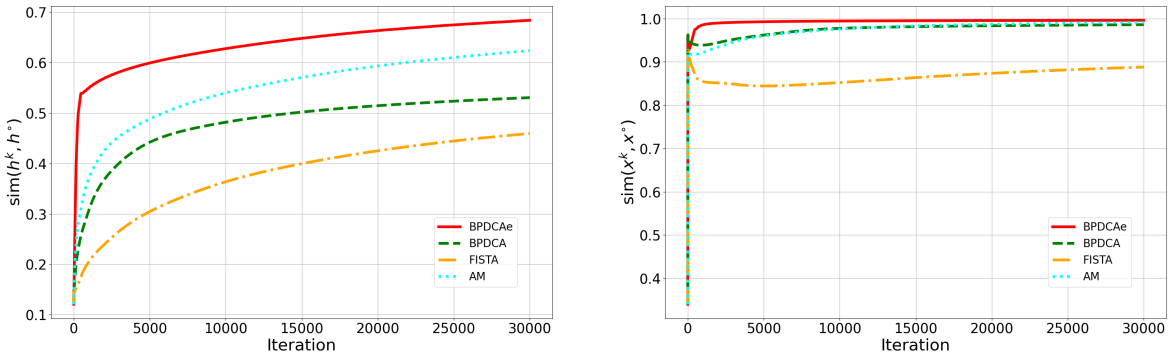


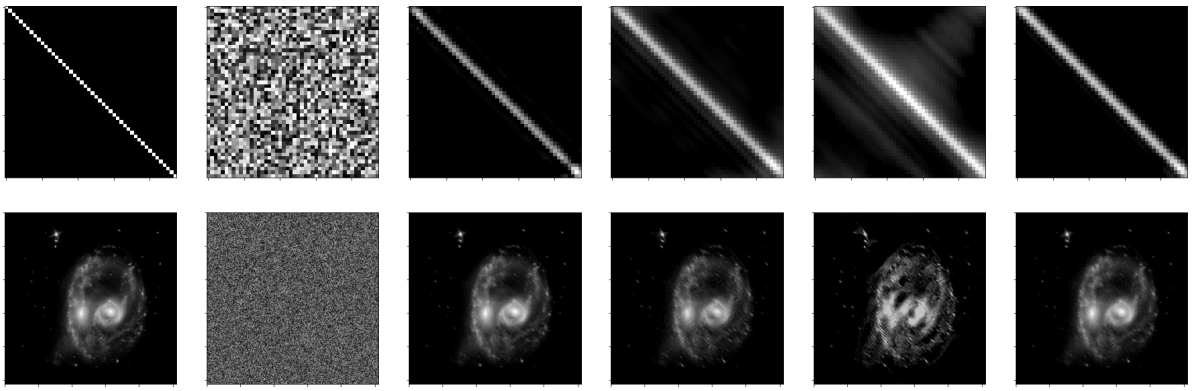
Figure 4.9: Plots of $\{\log_{10}|\Psi(\boldsymbol{h}^k, \boldsymbol{x}^k) - \Psi^\circ|\}$ when $\boldsymbol{z}^0$ is from the uniform distribution.



(a) Plots of $\{\mathrm{sim}(\boldsymbol{h}^k, \boldsymbol{h}^\circ)\}$.

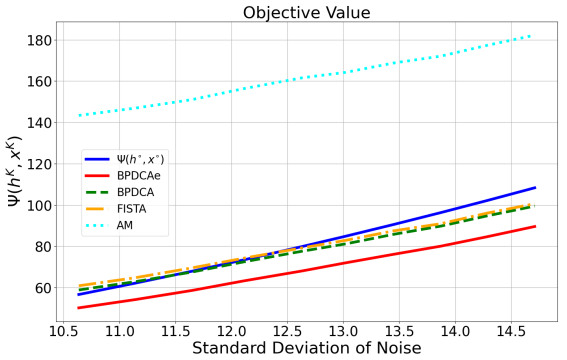
(b) Plots of $\{\mathrm{sim}(\boldsymbol{x}^k, \boldsymbol{x}^\circ)\}$.

Figure 4.10: Plots of the cosine similarities when $\boldsymbol{z}^0$ is from the uniform distribution.

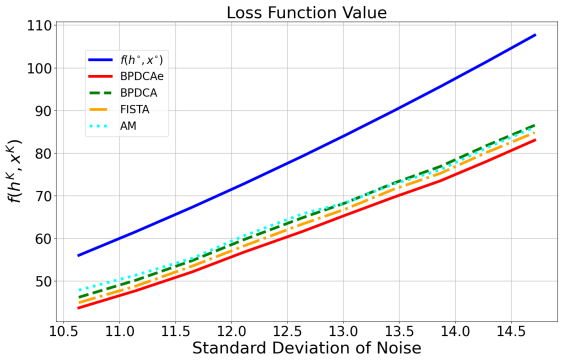


(a) $\boldsymbol{h}^\circ$ and $\boldsymbol{x}^\circ$ (b) $\boldsymbol{h}^0$ and $\boldsymbol{x}^0$ (c) BPDCAe (d) BPDCA (e) FISTA (f) AM

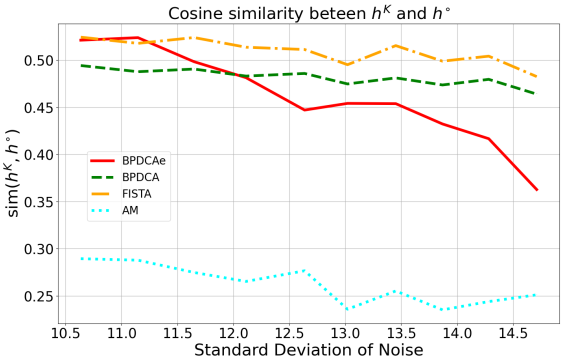
Figure 4.11: The upper row shows $\boldsymbol{h}^K$, and the lower row shows $\tilde{\boldsymbol{A}}\boldsymbol{x}^K$ when $\boldsymbol{z}^0$ is generated from the uniform distribution.



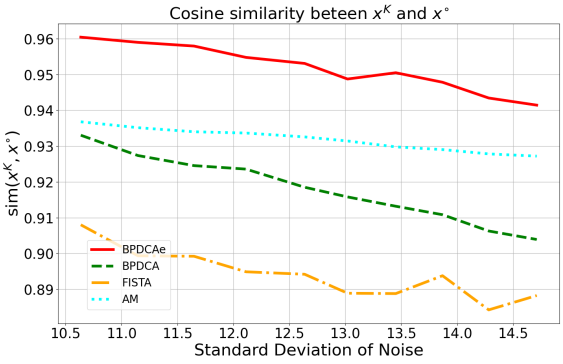
(a) Plots of the objective value $\Psi(\boldsymbol{h}^K, \boldsymbol{x}^K)$.



(b) Plots of the loss function value $f(\boldsymbol{h}^K, \boldsymbol{x}^K)$.



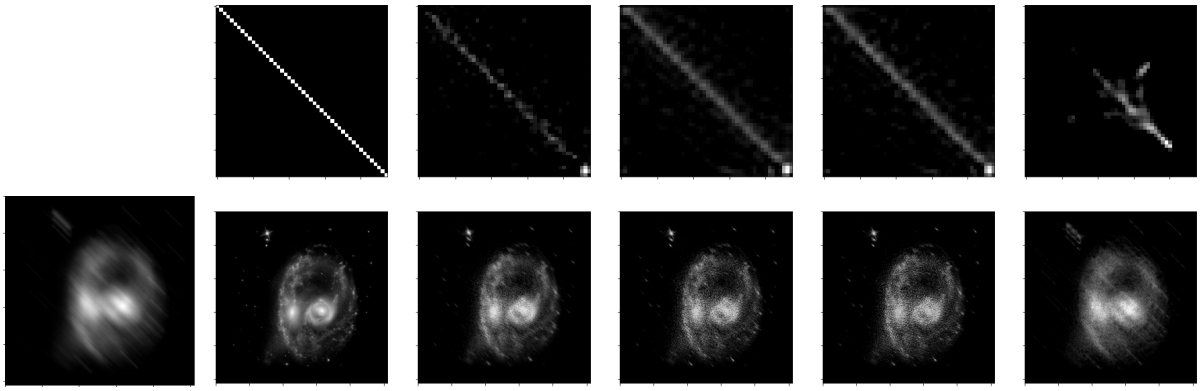
(c) Plots of $\mathrm{sim}(\boldsymbol{h}^K, \boldsymbol{h}^\circ)$.



(d) Plots of $\mathrm{sim}(\boldsymbol{x}^K, \boldsymbol{x}^\circ)$.

Figure 4.12: Plots of the objective value, the loss function value, and the cosine similarities recovered by each algorithm when $\tilde{\boldsymbol{y}}$ contains Poisson noise.

$\Psi(\boldsymbol{h}^K, \boldsymbol{x}^K)$, the loss function value $f(\boldsymbol{h}^K, \boldsymbol{x}^K)$, and the cosine similarities $\mathrm{sim}(\boldsymbol{h}^K, \boldsymbol{h}^\circ)$ and $\mathrm{sim}(\boldsymbol{x}^K, \boldsymbol{x}^\circ)$, where $\boldsymbol{h}^K$ and $\boldsymbol{x}^K$ are recovered by each algorithm after $K = 30000$ (for BPDCA(e) and FISTA) or $K = 3000$ (for AM) iterations for each noise level. Whereas



(a) $\tilde{\boldsymbol{y}}$ (b) $\boldsymbol{h}^\circ$ and $\boldsymbol{x}^\circ$ (c) BPDCAe (d) BPDCA (e) FISTA (f) AM

Figure 4.13: (a) the noisy data; (b–f) the upper row shows $\boldsymbol{h}^K$, and the lower row shows $\tilde{\boldsymbol{A}}\boldsymbol{x}^K$ when $\tilde{\boldsymbol{y}}$ contains Poisson noise.

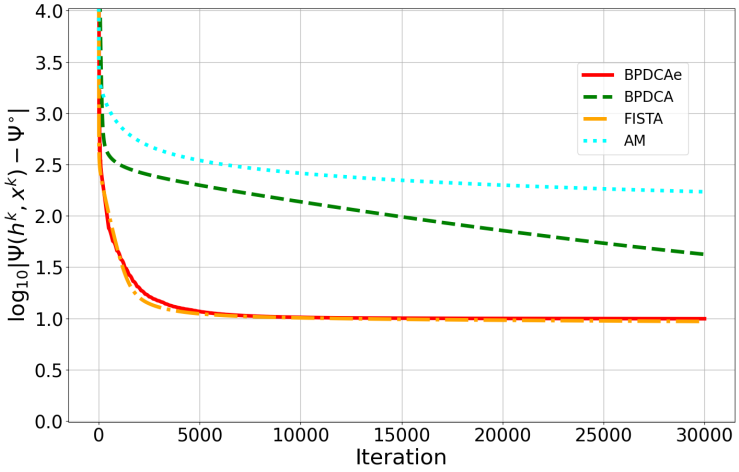
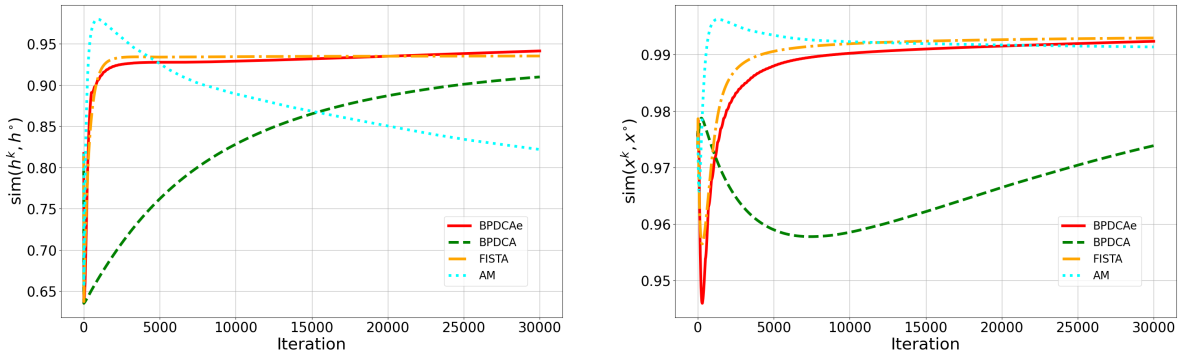


Figure 4.14: Plots of $\{\log_{10}|\Psi(\boldsymbol{h}^k, \boldsymbol{x}^k) - \Psi^\circ|\}$ at each iteration when $\boldsymbol{f}$ is a Gaussian blur.

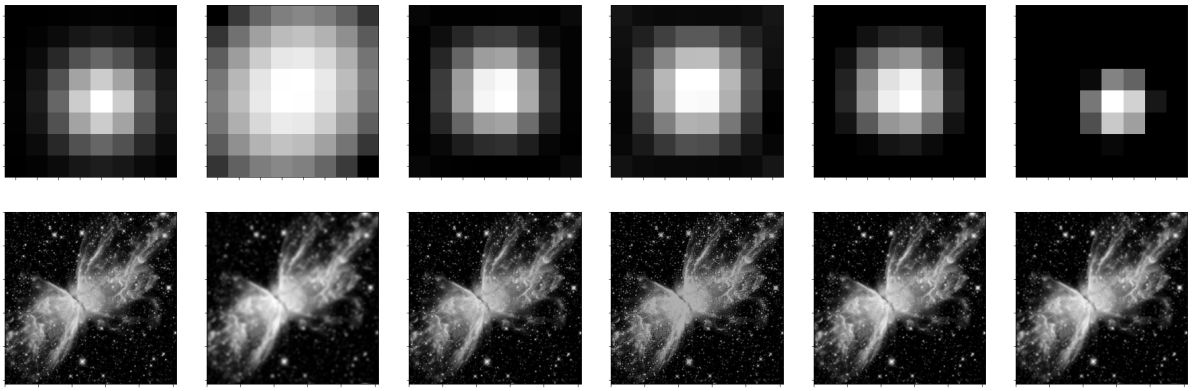
the cosine similarity $\mathrm{sim}(\boldsymbol{h}^K, \boldsymbol{h}^\circ)$ recovered by FISTA was the best among Figure 4.12c, the cosine similarity $\mathrm{sim}(\boldsymbol{x}^K, \boldsymbol{x}^\circ)$ recovered by BPDCAe was the best among Figure 4.12d. Figure 4.13 shows the recovered images when the standard deviation of $\boldsymbol{n}$ is 10.6. The point $\boldsymbol{x}^K$ recovered by BPDCAe is the best, and the objective value by BPDCAe is the smallest in Figure 4.12. Thus, BPDCA(e) outperformed the other algorithms with the noisy data.

Finally, we demonstrate that BPDCA(e) is effective with another blur kernel. We executed similar experiments using a Gaussian blur $\boldsymbol{f}$ and another image $\boldsymbol{g}$. Figures 4.14, 4.15a, and 4.15b show $\{\log_{10}|\Psi(\boldsymbol{h}^k, \boldsymbol{x}^k) - \Psi^\circ|\}$, $\{\mathrm{sim}(\boldsymbol{h}^k, \boldsymbol{h}^\circ)\}$, and $\{\mathrm{sim}(\boldsymbol{x}^k, \boldsymbol{x}^\circ)\}$, respectively, and Figure 4.16 shows the recovered images. The performance in the sense of the objective values and the cosine similarities of BPDCAe is almost the same as that of FISTA. The images recovered by BPDCAe, FISTA, and AM have almost no difference from Figures 4.16c, 4.16e, and 4.16f. Thus, depending on the types of the blur kernel $\boldsymbol{f}$



(a) Plots of $\{\text{sim}(\boldsymbol{h}^k, \boldsymbol{h}^\circ)\}$. (b) Plots of $\{\text{sim}(\boldsymbol{x}^k, \boldsymbol{x}^\circ)\}$.

Figure 4.15: Plots of the cosine similarities between the kth point and the ground truth when $\boldsymbol{f}$ is a Gaussian blur.



(a) $\boldsymbol{h}^\circ$ and $\boldsymbol{x}^\circ$ (b) $\boldsymbol{h}^0$ and $\boldsymbol{x}^0$ (c) BPDCAe (d) BPDCA (e) FISTA (f) AM

Figure 4.16: The upper row shows $\boldsymbol{h}^K$, and the lower row shows $\tilde{\boldsymbol{A}}\boldsymbol{x}^K$ when $\boldsymbol{f}$ is a Gaussian blur.

and the image $\boldsymbol{g}$, BPDCA(e) has the same results as the other algorithms. As we saw here, the performance of BPDCA(e) is almost the same as or superior to that of the other algorithms.

4.4 Self-calibration in Radio Interferometric Imaging

4.4.1 Problem Description

We want to recover the image $\boldsymbol{x} \in \mathbb{R}^d$ from the complex visibilities $\boldsymbol{v} \in \mathbb{C}^m$. Because the complex visibilities contain noise from measuring instruments and the atmosphere, the goal of calibration is to remove noise in the visibilities. Self-calibration is a calibration

of complex gains $\boldsymbol{g} \in \mathbb{C}^n$ given by each antenna α. For further information on radio interferometric imaging, see [116]. Each element v_i is associated with the observation time t_{l_i} and a pair of two stations α_i, β_i for $i = 1, \ldots, m$. In this subsection, we recover the image $\boldsymbol{x} \in \mathbb{R}^d$ with self-calibration of a gain vector $\boldsymbol{g} \in \mathbb{C}^n$. Therefore, we minimize the following chi-square error given by

$$f(\boldsymbol{g}, \boldsymbol{x}) = \sum_{i=1}^{m} \frac{1}{\sigma_i^2} |v_i g_{l_i,\alpha_i} \bar{g}_{l_i,\beta_i} - \mathcal{F}_i(\boldsymbol{x})|^2,$$

where $\mathcal{F}$ is the Fourier transformation and variance $\sigma_i > 0$ for $i = 1, \ldots, m$. We add a regularization term $r(\boldsymbol{g})$ for $\boldsymbol{g}$ given by

$$\begin{aligned} r(\boldsymbol{g}) = \rho \sum_{\alpha} \left| \sum_{l=1}^{L_\alpha} g_{l,\alpha} - L_\alpha \right|^2 \\ + \theta_1 \sum_{\alpha} \sum_{l=2}^{L_\alpha} \frac{1}{s_\alpha^2 (t_l - t_{l-1})} |g_{l,\alpha} - g_{l-1,\alpha}|^2 \\ + \theta_2 \sum_{\alpha} \sum_{l=2}^{L_\alpha} \frac{1}{s_\alpha^2 (t_l - t_{l-1})} (|g_{l,\alpha}| - |g_{l-1,\alpha}|)^2, \end{aligned}$$

where the length L_α of the time sequences depends on a station α, $\rho > 0$, $\theta_1 > 0$, $\theta_2 > 0$, $s_l > 0$, and $t_l > 0$. The first term of $r(\boldsymbol{g})$ imposes the gain having averages of 1, the second term imposes the sparsity of the difference of the amplitude and the phase, and the third term imposes that of the amplitude. We also add the regularization term for $\boldsymbol{x}$ given by $h(\boldsymbol{x}) = \theta_3 \|\boldsymbol{x}\|_1 + \theta_4 \, \mathrm{TSV}(\boldsymbol{x})$ with $\theta_3 > 0$ and $\theta_4 > 0$. The total square variation (TSV) regularization term [56] is defined by $\mathrm{TSV}(\boldsymbol{x}) = \sum_{i,j} ((X_{i+1,j} - X_{i,j})^2 + (X_{i,j+1} - X_{i,j})^2)$, where the matrix $\boldsymbol{X}$ is reshapen from $\boldsymbol{x}$. The TSV regularization term brings the recovered image to be edge-smoothed [56]. Therefore, we consider the following block optimization problem:

$$\min_{(\boldsymbol{g},\boldsymbol{x}) \in \mathbb{C}^n \times \mathbb{R}^d} \quad \Psi_{\mathrm{B}}(\boldsymbol{g}, \boldsymbol{x}) := f(\boldsymbol{g}, \boldsymbol{x}) + r(\boldsymbol{g}) + h(\boldsymbol{x}). \tag{4.23}$$

We define

$$\boldsymbol{A} = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 1 & \cdots & 1 \end{bmatrix}, \quad \boldsymbol{b} = (L_\alpha)_\alpha,$$

and then it holds that

$$\rho \sum_{\alpha} \left| \sum_{l=1}^{L_\alpha} g_{l,\alpha} - L_\alpha \right|^2 = \rho \|\boldsymbol{A}\boldsymbol{g} - \boldsymbol{b}\|_2^2.$$

4.4.2 DC Decomposition

Let $\boldsymbol{y} = \mathcal{F}(\boldsymbol{x})$. We reformulate f into a DC function $f_1 - f_2$. In the same way as blind deconvolution, we have

$$\begin{aligned}|v_i g_{l_i,\alpha_i}\bar{g}_{l_i,\beta} - y_i|^2 &= |v_i g_{l_i,\alpha_i}\bar{g}_{l_i,\beta_i}|^2 + |y_i|^2 - v_i g_{l_i,\alpha_i}\bar{g}_{l_i,\beta_i}\bar{y}_i - \bar{v}_i\bar{g}_{l_i,\alpha_i}g_{l_i,\beta_i}y_i \\ &= \frac{1}{2}\left(|v_i g_{l_i,\alpha_i}|^2 + |g_{l_i,\beta_i}|^2\right)^2 + |v_i g_{l_i,\alpha_i}|^2 + |g_{l_i,\beta_i}y_i|^2 + |y_i|^2 \\ &\quad - \left(\frac{1}{2}|v_i g_{l_i,\alpha_i}|^4 + \frac{1}{2}|g_{l_i,\beta_i}|^4 + |v_i g_{l_i,\alpha_i} + g_{l_i,\beta_i}y_i|^2\right).\end{aligned}$$

Let us define convex functions f_1 and f_2 as follows:

$$\begin{aligned}f_1(\boldsymbol{g},\boldsymbol{x}) &= \sum_{i=1}^{m}\frac{1}{\sigma_i^2}\left(\frac{1}{2}\left(|v_i g_{l_i,\alpha_i}|^2 + |g_{l_i,\beta_i}|^2\right)^2 + |v_i g_{l_i,\alpha_i}|^2 + |g_{l_i,\beta_i}\mathcal{F}_i(\boldsymbol{x})|^2 + |\mathcal{F}_i(\boldsymbol{x})|^2\right), \\ f_2(\boldsymbol{g},\boldsymbol{x}) &= \sum_{i=1}^{m}\frac{1}{\sigma_i^2}\left(\frac{1}{2}|v_i g_{l_i,\alpha_i}|^4 + \frac{1}{2}|g_{l_i,\beta_i}|^4 + |v_i g_{l_i,\alpha_i} + g_{l_i,\beta_i}\mathcal{F}_i(\boldsymbol{x})|^2\right).\end{aligned}$$

Moreover, we reformulate r into a DC function $r_1 - r_2$, given by

$$\begin{aligned}r_1(\boldsymbol{g}) &= \rho\|\boldsymbol{A}\boldsymbol{g} - \boldsymbol{b}\|_2^2 \\ &\quad + \theta_1\sum_{\alpha}\sum_{l=2}^{L_\alpha}\frac{1}{s_\alpha^2(t_l - t_{l-1})}|g_{l,\alpha} - g_{l-1,\alpha}|^2 \\ &\quad + \theta_2\sum_{\alpha}\sum_{l=2}^{L_\alpha}\frac{2}{s_\alpha^2(t_l - t_{l-1})}(|g_{l,\alpha}|^2 + |g_{l-1,\alpha}|^2), \\ r_2(\boldsymbol{g}) &= \theta_2\sum_{\alpha}\sum_{l=2}^{L_\alpha}\frac{1}{s_\alpha^2(t_l - t_{l-1})}(|g_{l,\alpha}| + |g_{l-1,\alpha}|)^2.\end{aligned}$$

The function r_1 is continuously differentiable, whereas the function r_2 is nonsmooth. Therefore, (4.23) is equivalent to the following block DC optimization problem:

$$\min_{(\boldsymbol{g},\boldsymbol{x})\in\mathbb{C}^n\times\mathbb{R}^d} \quad \Psi_{\mathrm{B}}(\boldsymbol{g},\boldsymbol{x}) = f_1(\boldsymbol{g},\boldsymbol{x}) - f_2(\boldsymbol{g},\boldsymbol{x}) + r_1(\boldsymbol{g}) - r_2(\boldsymbol{g}) + h(\boldsymbol{x}). \tag{4.24}$$

4.4.3 L-smooth Adaptable Parameters

The following theorem provides an appropriate kernel generating distance ϕ and an appropriate parameter L_ρ in use of Algorithm 3.

Theorem 4.13. *Let a function ϕ be defined by*

$$\phi(\boldsymbol{g}) = \frac{1}{2}\sum_{l=1}^{\hat{L}}\left(\sum_{\alpha}|g_{l,\alpha}|^2\right)^2 + \sum_{\alpha}\sum_{l=1}^{L_\alpha}|g_{l,\alpha}|^2, \tag{4.25}$$

where $\hat{L} = \max_\alpha L_\alpha$ *and* $g_{l,\alpha} = 0$ *for* $l > L_\alpha$. *Let* $I = \{(l_i, \alpha_i), (l_i, \beta_i) \mid (l_i, \alpha_i, \beta_i), i = 1, \dots, m\}$ *be the set of the pairs* (l, α) *associated with* i *and* $\mathcal{N}(l, \alpha)$ *be a mapping from* (l, α) *to the corresponding* i. *Then, for any* $L_\rho(\cdot)$ *satisfying*

$$L_\rho(\boldsymbol{x}) \geq \rho\lambda_{\max}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A}) + \max_{\alpha, l=2,\dots,L_\alpha} L_l + \max_{(l,\alpha)\in I} \sum_{i\in\mathcal{N}(l,\alpha)} L_i(\boldsymbol{x}), \tag{4.26}$$

where

$$L_l = \frac{4(\theta_1+\theta_2)}{s_\alpha^2(t_l - t_{l-1})}, \quad L_i(\boldsymbol{x}) = \frac{1}{\sigma_i^2}(3|v_i|^4 + 3 + 2|v_i|^2 + |\mathcal{F}_i(\boldsymbol{x})|^2),$$

the function $L_\rho(\boldsymbol{x})\phi - f_1(\cdot, \boldsymbol{x}) - r_1$ *is convex for* $\boldsymbol{x} \in \mathbb{R}^d$.

Proof. From $\|\boldsymbol{g}\|_2^2 = \sum_\alpha \sum_{l=1}^{L_\alpha} |g_{l,\alpha}|^2$ and $\rho\|\boldsymbol{A}\boldsymbol{g} - \boldsymbol{b}\|_2^2$, the function $\rho\lambda_{\max}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A})\|\boldsymbol{g}\|_2^2 - \rho\|\boldsymbol{A}\boldsymbol{g} - \boldsymbol{b}\|_2^2$ is convex. For the fix α and l, we consider the convexity of the function given by

$$\begin{aligned}
&\frac{L_l}{2}(|g_{l-1,\alpha}|^2 + |g_{l,\alpha}|^2) - \frac{\theta_1}{s_\alpha^2(t_l - t_{l-1})}|g_{l,\alpha} - g_{l-1,\alpha}|^2 - \frac{2\theta_2}{s_\alpha^2(t_l - t_{l-1})}(|g_{l,\alpha}|^2 + |g_{l-1,\alpha}|^2) \\
&= \frac{\theta_1}{s_\alpha^2(t_l - t_{l-1})}|g_{l,\alpha} + g_{l-1,\alpha}|^2 + \left(\frac{L_l}{2} - \frac{2\theta_1}{s_\alpha^2(t_l - t_{l-1})} - \frac{2\theta_2}{s_\alpha^2(t_l - t_{l-1})}\right)(|g_{l-1,\alpha}|^2 + |g_{l,\alpha}|^2).
\end{aligned}$$

Therefore, for any L_l satisfying

$$L_l \geq \frac{4\theta_1}{s_\alpha^2(t_l - t_{l-1})} + \frac{4\theta_2}{s_\alpha^2(t_l - t_{l-1})},$$

the function $(\rho\lambda_{\max}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A}) + \max_\alpha \max_{l=2,\dots,L_\alpha} L_l)\phi - r_1$ is convex.

Next, for fixed $\boldsymbol{x} \in \mathbb{R}^d$, we consider the smooth adaptable property for

$$f_{1,i}(g_{l_i,\alpha_i}, g_{l_i,\beta_i}) = \frac{1}{\sigma_i^2}\left(\frac{1}{2}\left(|v_i g_{l_i,\alpha_i}|^2 + |g_{l_i,\beta_i}|^2\right)^2 + |v_i g_{l_i,\alpha_i}|^2 + |g_{l_i,\beta_i}\mathcal{F}_i(\boldsymbol{x})|^2 + |\mathcal{F}_i(\boldsymbol{x})|^2\right).$$

Using from Theorem 4.6, for any $L_i(\cdot)$ satisfying

$$L_i(\boldsymbol{x}) \geq \frac{1}{\sigma_i^2}(3|v_i|^4 + 3 + 2|v_i|^2 + |\mathcal{F}_i(\boldsymbol{x})|^2), \tag{4.27}$$

the function $L_i(\boldsymbol{x})\phi - f_{1,i}$ is convex. Therefore, for any $L_\rho(\cdot)$ satisfying

$$L_\rho(\boldsymbol{x}) \geq \rho\lambda_{\max}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A}) + \max_{\alpha, l=2,\dots,L_\alpha} L_l + \max_{(l,\alpha)\in I} \sum_{i\in\mathcal{N}(l,\alpha)} L_i(\boldsymbol{x}),$$

the function $L_\rho(\boldsymbol{x})\phi - f_1(\cdot, \boldsymbol{x}) - r_1$ is convex. □

We recall $C_1 = \operatorname{int}\operatorname{dom}\phi = \mathbb{C}^n$ and $C_2 = \mathbb{R}^d$. HBPDCA for self-calibration in radio interferometric imaging, which we are proposing, is listed as Algorithm 4.

Algorithm 4 HBPDCA for self-calibration

Input: $\phi \in \mathcal{G}(C_1)$ with $C_1 = \mathbb{C}^n$ such that the $L_\rho(\cdot)$-smad property for the pair $(f_1 + r_1, \phi)$ holds for $\boldsymbol{x} \in \mathbb{R}^d$.
Initialization: $(\boldsymbol{g}^0, \boldsymbol{x}^0) \in C_1 \times C_2$.
for $k = 0, 1, 2, \ldots,$ **do**
Take any $\boldsymbol{\xi}^k \in \partial_{\mathrm{c}} r_2(\boldsymbol{g}^k)$ and compute $\lambda^k = 1/L_\rho(\boldsymbol{x}^k)$ and

$$\boldsymbol{g}^{k+1} = \operatorname*{argmin}_{\boldsymbol{g}\in\mathbb{C}^n} \left\{ 2\operatorname{Re}\langle \nabla_{\boldsymbol{g}} f(\boldsymbol{g}^k, \boldsymbol{x}^k) + \nabla r_1(\boldsymbol{g}^k) - \boldsymbol{\xi}^k, \boldsymbol{g} - \boldsymbol{g}^k \rangle + \frac{1}{\lambda^k} D_\phi(\boldsymbol{g}, \boldsymbol{g}^k) \right\}, \tag{4.28}$$

$$\boldsymbol{x}^{k+1} = \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^d} \left\{ f(\boldsymbol{g}^{k+1}, \boldsymbol{x}) + h(\boldsymbol{x}) \right\}. \tag{4.29}$$

end for

Assumptions 3.30 and 3.34 holds. Especially, since (4.25) is strongly convex, Assumption 3.34 (i) holds. Because of $C_1 = \mathbb{C}^n$, Assumption 3.2 holds for (4.28).

Subproblem (4.28) is solved in a closed-form expression.

Remark 4.14. *Let* $\boldsymbol{u}^k := \lambda^k(\nabla_{\boldsymbol{g}} f(\boldsymbol{g}^k, \boldsymbol{x}^k) + \nabla r_1(\boldsymbol{g}^k) - \boldsymbol{\xi}^k - \nabla\phi(\boldsymbol{g}^k))$. *By the first-order optimality condition of* (4.28), *for each* l, α, *we obtain*

$$u^k_{l,\alpha} + \left(\sum_\alpha |g^{k+1}_{l,\alpha}| + 1 \right) g^{k+1}_{l,\alpha} = 0, \tag{4.30}$$

which implies $g^{k+1}_{l,\alpha} = -\tau_l u^k_{l,\alpha}$ *for* $\tau_l > 0$, *and*

$$\left(\tau_l^3 \sum_\alpha |u^k_{l,\alpha}| + \tau_l - 1 \right) u^k_{l,\alpha} = 0.$$

This cubic equation is solved by using Cardano's formula. Therefore, $g^{k+1}_{l,\alpha} = -\tau_l u^k_{l,\alpha}$, *where* τ_l *is the unique positive real root of*

$$\tau_l^3 \sum_\alpha |u^k_{l,\alpha}| + \tau_l - 1 = 0, \quad l = 1, \ldots, \hat{L}.$$

Because subproblem (4.29) *is a convex optimization problem, it is solved, for example, by FISTA. Thus, the iteration of HBPDCA is easily computable.*

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this thesis, for general nonconvex optimization problems, we have proposed fast algorithms, called Bregman proximal DC algorithm (BPDCA) [113], the BPDCA with extrapolation (BPDCAe) [113], and the hybrid BPDCA (HBPDCA). Besides, we have established convergence analysis of these algorithms. Because our proposed algorithms exploit the Bregman distance and DC structure, they are applicable to a wide range of optimization problems, including optimization problems that lack L-smoothness. Exploiting DC structure, we have flexibility on the choice of the Bregman distance for our proposed algorithms.

Moreover, we have also applied our proposed algorithms to problems in signal processing, such as phase retrieval, blind deconvolution, and self-calibration in radio interferometric imaging. For phase retrieval, exploiting DC structure, we have obtained larger step sizes than the existing one. Using these step sizes, we have succeeded in accelerating BPDCA(e). Then, BPDCAe outperformed the existing algorithms [113]. For blind deconvolution and self-calibration in radio interferometric imaging, exploiting DC structure, we have obtained an appropriate Bregman distance. Especially in blind deconvolution, through numerical experiments on image deblurring, our proposed algorithms successfully recovered the original image [114].

In Chapter 2, we have summarized the important notions and their examples. The Bregman distance is a core idea for our proposed algorithms. Subdifferentials, the L-smad property, the KL property, and subanalyticity have played an important role in convergence analysis. Complex analysis is especially used for HBPDCA and self-calibration.

In Chapter 3, we have proposed the Bregman proximal algorithms exploiting DC structure and established their convergence analysis. First, we have proposed BPDCA [113], which is based on pDCA and the Bregman distance. Bregman proximal algorithms require, instead of L-smoothness, the L-smad property, which is a generalization of L-smoothness. Second, we have proposed BPDCAe [113], which is accelerated by the extrapolation technique adapted to the Bregman distance. The adaptive restart scheme of BPDCAe requires fewer computational tasks and is easy to implement. We have also

established the global convergence of BPDCA(e) to a limiting stationary point or a limiting critical point under the KL property or subanalyticity, respectively. In addition, the rate of convergence with the KL (or Łojasiewicz) exponent has been established. Finally, we have proposed HBPDCA, which minimizes a subproblem based on the Bregman distance and a convex optimization problem. For HBPDCA, we have established its global subsequential convergence.

In Chapter 4, we have applied BPDCA, BPDCAe, and HBPDCA to phase retrieval, blind deconvolution, and self-calibration in radio interferometric imaging. These problems are reformulated as nonconvex optimization problems and also as DC optimization problems. For phase retrieval, exploiting DC structure, we have found several smaller L for the L-smad property than the existing one. Using these L, we have succeeded in accelerating BPDCA(e). Numerical experiments on phase retrieval have demonstrated that BPDCAe with the parameter is faster than the other Bregman proximal algorithms [113]. For phase retrieval under a Gaussian model, BPDCAe offered more stable results than the Wirtinger flow [21]. For blind deconvolution, by exploiting DC structure, we have found an appropriate ϕ. We have conducted numerical experiments on image deblurring. They demonstrated that BPDCAe outperformed other existing algorithms and successfully recovered the original image [114]. The numerical success of image deblurring is also because the regularization term is represented sparsely in the wavelet domain. For self-calibration in radio interferometric imaging, by exploiting DC structure, we have found an appropriate ϕ. Using this ϕ, we obtained a closed-form solution to the subproblem of HBPDCA for self-calibration in radio interferometric imaging.

We conclude that our proposed algorithms are fast for various nonconvex optimization problems by exploiting the Bregman distance and DC structure.

5.2 Future Work

From (4.21), we have shown the choice of the kernel generating distance ϕ affects the performance of BPDCA(e). This fact is also true for algorithms using the Bregman distance. The choice of the kernel generating distance ϕ affects the convergence speed and the convergent point. First, the effective way to choose ϕ and the calculation of the L-smad parameter have not yet been established. Developing a method to compute the L-smad parameter that accelerates the Bregman proximal algorithms is a topic for the future. If choosing the Bregman distance and calculation of the L-smad parameter are established for nonconvex optimization problems, the Bregman proximal algorithms could deal with a wider range of nonconvex optimization problems in practice. Second, the relationship between ϕ and convergent points, such as a limiting stationary point and a limiting critical point. Depending on the choice of ϕ, Bregman proximal algorithms converge to the stronger points, such as a directional stationary point, a local optimal solution, and a global optimal solution. Finally, we conjecture that most 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 BPDCAe (Lemma 3.21) without Assumption 3.20. Although the kernel generating distance ϕ (4.7) does not satisfy Assumption 3.7 (i), the sequences generated by BPDCA(e) converged in numerical experiments. It may be possible to weaken Assumption 3.7 (i).

For HBPDCA, acceleration has not yet been established. Acceleration of HBPDCA is important in practice. HBPDCA would be accelerated by extrapolation in the same way as BPDCAe. In addition, convergence analysis of HBPDCA has not yet been established. If the KL property is extended to complex variables, we expect that HBPDCA converges to a limiting stationary point and the rate of convergence would also be established. Further convergence analysis of HBPDCA is left for future work.

The Bregman proximal algorithms have the potential to be applied to a wide variety of nonconvex optimization problems. For example, they could be applied to a problem whose objective function includes the Bregman distance such as computing entropic centers [88]. In this thesis, it is essential for future work to conduct numerical experiments on self-calibration in radio interferometric imaging. Furthermore, applying HBPDCA to practical data in radio interferometry is for future work. In practice, calculation of the KL exponent is often difficult. In particular, the KL exponent of some applications, such as blind deconvolution and self-calibration, has not been computed and is future work.

Bibliography

[1] T. Adali and P. J. Schreier, *Optimization and estimation of complex-valued signals: Theory and applications in filtering and blind source separation*, IEEE Signal Processing Magazine, 31 (2014), pp. 112–128.

[2] H. S. Aghamiry, A. Gholami, and S. Operto, *Complex-valued imaging with total variation regularization: An application to full-waveform inversion in visco-acoustic media*, SIAM Journal on Imaging Sciences, 14 (2021), pp. 58–91.

[3] A. Ahmed, B. Recht, and J. Romberg, *Blind deconvolution using convex programming*, IEEE Transactions on Information Theory, 60 (2014), pp. 1711–1732.

[4] H. A. Aly and E. Dubois, *Image up-sampling using total-variation regularization with a new observation model*, IEEE Transactions on Image Processing, 14 (2005), pp. 1647–1659.

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

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

[7] L. Balzano and R. Nowak, *Blind calibration of sensor networks*, in Proceedings of the 6th International Conference on Information Processing in Sensor Networks, 2007, pp. 79–88.

[8] H. H. Bauschke, J. Bolte, and M. Teboulle, *A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications*, Mathematics of Operations Research, 42 (2017), pp. 330–348.

[9] A. Beck, *First-Order Methods in Optimization*, vol. 25 of MOS-SIAM Series on Optimization, SIAM, 2017.

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

[11] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, third ed., 2016.

[12] E. Bierstone and P. D. Milman, *Semialgebraic and subanalytic sets*, Publications Mathématiques de l'I.H.É.S., 67 (1988), pp. 5–42.

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

[14] J. Bolte, S. Sabach, and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Mathematical Programming, 146 (2014), pp. 459–494.

[15] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd, *First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems*, SIAM Journal on Optimization, 28 (2018), pp. 2131–2151.

[16] C. Bouman and K. Sauer, *A generalized Gaussian image model for edge-preserving MAP estimation*, IEEE Transactions on Image Processing, 2 (1993), pp. 296–310.

[17] T. Bouwmans, A. Sobral, S. Javed, S. K. Jung, and E.-H. Zahzah, *Decomposition into low-rank plus additive matrices for background/foreground separation: A review for a comparative evaluation with a large-scale dataset*, Computer Science Review, 23 (2017), pp. 1–71.

[18] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[19] L. M. Bregman, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR Computational Mathematics and Mathematical Physics, 7 (1967), pp. 200–217.

[20] R. E. Bruck, *An iterative solution of a variational inequality for certain monotone operators in Hilbert space*, Bulletin of the American Mathematical Society, 81 (1975), pp. 890–892.

[21] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski, *Phase retrieval via matrix completion*, SIAM Journal on Imaging Sciences, 6 (2013), pp. 199–225.

[22] E. J. Candès, X. Li, and M. Soltanolkotabi, *Phase retrieval via Wirtinger flow: Theory and algorithms*, IEEE Transactions on Information Theory, 61 (2015), pp. 1985–2007.

[23] E. J. Candès, T. Strohmer, and V. Voroninski, *PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming*, Communications on Pure and Applied Mathematics, 66 (2013), pp. 1241–1274.

[24] M. Cannon, *Blind deconvolution of spatially invariant image blurs with phase*, IEEE Transactions on Acoustics, Speech, and Signal Processing, 24 (1976), pp. 58–63.

[25] A. Cauchy, *Méthode générale pour la résolution des systèmes d'équations simultanées*, Comptes Rendus de l'Académie des Sciences, 25 (1847), pp. 536–538.

[26] A. A. Chael, M. D. Johnson, K. L. Bouman, L. L. Blackburn, K. Akiyama, and R. Narayan, *Interferometric imaging directly with closure phases and closure amplitudes*, The Astrophysical Journal, 857 (2018), p. 23.

[27] T. F. Chan and C. K. Wong, *Convergence of the alternating minimization algorithm for blind deconvolution*, Linear Algebra and its Applications, 316 (2000), pp. 259–285.

[28] R. Chartrand, *Exact reconstruction of sparse signals via nonconvex minimization*, IEEE Signal Processing Letters, 14 (2007), pp. 707–710.

[29] G. Chen and M. Teboulle, *Convergence analysis of a proximal-like minimization algorithm using Bregman functions*, SIAM Journal on Optimization, 3 (1993), pp. 538–543.

[30] J. Chen, R. Lin, H. Wang, J. Meng, H. Zheng, and L. Song, *Blind-deconvolution optical-resolution photoacoustic microscopy in vivo*, Optics Express, 21 (2013), pp. 7316–7327.

[31] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Classics in Applied Mathematics, SIAM, 1990.

[32] B. D. Craven and B. Mond, *On duality in complex linear programming*, Journal of the Australian Mathematical Society, 16 (1973), pp. 172–175.

[33] ——, *Linear programming with matrix variables*, Linear Algebra and its Applications, 38 (1981), pp. 73–80.

[34] Y. Cui and J.-S. Pang, *Modern Nonconvex Nondifferentiable Optimization*, vol. 29 of MOS-SIAM Series on Optimization, SIAM, 2021.

[35] J. C. Dainty and J. R. Fienup, *Phase retrieval and image reconstruction for astronomy*, in Image Recovery: Theory and Application, Academic Press, 1987, pp. 231–275.

[36] F. Dell'Acqua, P. Scifo, G. Rizzo, M. Catani, A. Simmons, G. Scotti, and F. Fazio, *A modified damped Richardson–Lucy algorithm to reduce isotropic background effects in spherical deconvolution*, Neuroimage, 49 (2010), pp. 1446–1458.

[37] I. Dhillon and J. Tropp, *Matrix nearness problems with Bregman divergences*, SIAM Journal on Matrix Analysis and Applications, 29 (2008), pp. 1120–1146.

[38] C. Ding, D. Zhou, X. He, and H. Zha, *R_1-PCA: rotational invariant L_1-norm principal component analysis for robust subspace factorization*, in Proceedings of the 23rd International Conference on Machine Larning, 2006, pp. 281–288.

[39] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, Springer, 2010.

[40] M. Elad, J.-L. Starck, P. Querre, and D. L. Donoho, *Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA)*, Applied and Computational Harmonic Analysis, 19 (2005), pp. 340–358.

[41] R. J. Fétick, L. M. Mugnier, T. Fusco, and B. Neichel, *Blind deconvolution in astronomy with adaptive optics: The parametric marginal approach*, Monthly Notices of the Royal Astronomical Society, 496 (2020), pp. 4209–4220.

[42] J. R. Fienup, *Phase retrieval algorithms: A comparison*, Applied Optics, 21 (1982), pp. 2758–2769.

[43] R. W. Gerchberg and W. O. Saxton, *A practical algorithm for the determination of the phase from image and diffraction plane pictures*, Optik, 35 (1972), pp. 237–246.

[44] S. Gogineni and A. Nehorai, *Target estimation using sparse modeling for distributed MIMO radar*, IEEE Transactions on Signal Processing, 59 (2011), pp. 5315–5325.

[45] P. Hartman, *On functions representable as a difference of convex functions*, Pacific Journal of Mathematics, 9 (1959), pp. 707–713.

[46] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer, second ed., 2009.

[47] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC Press, 2015.

[48] H. He and Z. Zhang, *A unified Bregman alternating minimization algorithm for generalized DC programming with applications to image processing*, arXiv preprint arXiv:2209.07323, (2022).

[49] A. Hjørungnes, *Complex-Valued Matrix Derivatives: With Applications in Signal Processing and Communications*, Cambridge University Press, 2011.

[50] M. Huang, M.-J. Lai, A. Varghese, and Z. Xu, *On DC based methods for phase retrieval*, in Approximation Theory XVI, Springer, 2019, pp. 87–121.

[51] F. Itakura, *Analysis synthesis telephony based on the maximum likelihood method*, in Proceedings of the 6th International Congress on Acoustics, 1968.

[52] C. Jain, A. Kumar, A. Chugh, and N. Charaya, *Efficient image deblurring application using combination of blind deconvolution method and blur parameters estimation method*, ECS Transactions, 107 (2022), pp. 3695–3704.

[53] S. M. Jefferies and J. C. Christou, *Restoration of astronomical images by iterative blind deconvolution*, The Astrophysical Journal, 415 (1993), pp. 862–874.

[54] D. Krishnan, T. Tay, and R. Fergus, *Blind deconvolution using a normalized sparsity measure*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 233–240.

[55] S. Kullback and R. A. Leibler, *On information and sufficiency*, The annals of mathematical statistics, 22 (1951), pp. 79–86.

[56] K. Kuramochi, K. Akiyama, S. Ikeda, F. Tazaki, V. L. Fish, H.-Y. Pu, K. Asada, and M. Honma, *Superresolution interferometric imaging with sparse modeling using total squared variation: Application to imaging the black hole shadow*, The Astrophysical Journal, 858 (2018), p. 56.

[57] K. Kurdyka, *On gradients of functions definable in o-minimal structures*, Annales de l'Institut Fourier, 48 (1998), pp. 769–783.

[58] H. C. Lai and J. C. Liu, *Complex fractional programming involving generalized quasi/pseudo convex functions*, Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik, 82 (2002), pp. 159–166.

[59] H. C. Lai, J. C. Liu, and S. Schaible, *Complex minimax fractional programming of analytic functions*, Journal of Optimization Theory and Applications, 137 (2008), pp. 171–184.

[60] G. Lan, *First-order and Stochastic Optimization Methods for Machine Learning*, Springer Series in the Data Sciences, Springer, 2020.

[61] E. M. Landis, *On functions representable as the difference of two convex functions*, Doklady Akademii Nauk SSSR, 80 (1951), pp. 9–11.

[62] R. G. Lane and R. H. T. Bates, *Automatic multidimensional deconvolution*, Journal of the Optical Society of America A, 4 (1987), pp. 180–188.

[63] H. A. Le Thi, V. N. Huynh, and T. Pham Dinh, *Convergence analysis of difference-of-convex algorithm with subanalytic data*, Journal of Optimization Theory and Applications, 179 (2018), pp. 103–126.

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

[65] H. A. Le Thi and T. Pham Dinh, *DC programming and DCA: Thirty years of developments*, Mathematical Programming, 169 (2018), pp. 5–68.

[66] H. A. Le Thi, T. Pham Dinh, and M. Le Dung, *Exact penalty in DC programming*, Vietnam Journal of Mathematics, 27 (1999), pp. 169–178.

[67] C. Li, C. Xu, C. Gui, and M. D. Fox, *Distance regularized level set evolution and its application to image segmentation*, IEEE Transactions on Image Processing, 19 (2010), pp. 3243–3254.

[68] G. Li and T. K. Pong, *Calculus of the exponent of Kurdyka–Łojasiewicz inequality and its applications to linear convergence of first-order methods*, Foundations of Computational Mathematics, 18 (2018), pp. 1199–1232.

[69] X. Li, S. Ling, T. Strohmer, and K. Wei, *Rapid, robust, and reliable blind deconvolution via nonconvex optimization*, Applied and Computational Harmonic Analysis, 47 (2019), pp. 893–934.

[70] Z. Lin, H. Li, and C. Fang, *Accelerated Optimization for Machine Learning*, Springer, 2020.

[71] P.-L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM Journal on Numerical Analysis, 16 (1979), pp. 964–979.

[72] J. S. Liu and R. Chen, *Blind deconvolution via sequential imputations*, Journal of the American Statistical Association, 90 (1995), pp. 567–576.

[73] L. B. Lucy, *An iterative technique for the rectification of observed distributions*, The Astronomical Journal, 79 (1974), p. 745.

[74] P. C. Mahalanobis, *On the generalized distance in statistics*, in Proceedings of the National Institute of Sciences, 1936.

[75] A. Maleki, L. Anitori, Z. Yang, and R. G. Baraniuk, *Asymptotic analysis of complex LASSO via complex approximate message passing (CAMP)*, IEEE Transactions on Information Theory, 59 (2013), pp. 4290–4308.

[76] D. Malioutov, M. Cetin, and A. S. Willsky, *A sparse signal reconstruction perspective for source localization with sensor arrays*, IEEE Transactions on Signal Processing, 53 (2005), pp. 3010–3022.

[77] C. F. Mecklenbräuker, P. Gerstoft, and E. Zöchmann, *c–LASSO and its dual for sparse signal estimation from array data*, Signal Processing, 130 (2017), pp. 204–216.

[78] R. P. Millane, *Phase retrieval in crystallography and optics*, Journal of the Optical Society of America A, 7 (1990), pp. 394–411.

[79] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation I: Basic Theory*, vol. 330 of Grundlehren der mathematischen Wissenschaften, Springer, 2006.

[80] B. S. Mordukhovich, N. M. Nam, and N. D. Yen, *Fréchet subdifferential calculus and optimality conditions in nondifferentiable programming*, Optimization, 55 (2006), pp. 685–708.

[81] B. S. Mordukhovich and R. T. Rockafellar, *Second-order subdifferential calculus with applications to tilt stability in optimization*, SIAM Journal on Optimization, 22 (2012), pp. 953–986.

[82] J. J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bulletin de la Société Mathématique de France, 93 (1965), pp. 273–299.

[83] M. C. Mukkamala, P. Ochs, T. Pock, and S. Sabach, *Convex-concave backtracking for inertial Bregman proximal gradient algorithms in nonconvex optimization*, SIAM Journal on Mathematics of Data Science, 2 (2020), pp. 658–682.

[84] A. Myronenko, *3D MRI brain tumor segmentation using autoencoder regularization*, in Proceedings of the International MICCAI Brainlesion Workshop, 2018, pp. 311–320.

[85] A. S. Nemirovski and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, Wiley Series in Discrete Mathematics, Wiley, 1983.

[86] Y. Nesterov, *A method of solving a convex programming problem with convergence rate $O(1/k^2)$*, Soviet Mathematics Doklady, 27 (1983), pp. 372–376.

[87] ——, *Lectures on Convex Optimization*, vol. 137 of Springer Optimization and Its Applications, Springer, second ed., 2018.

[88] F. Nielsen and R. Nock, *Sided and symmetrized Bregman centroids*, IEEE Transactions on Information Theory, 55 (2009), pp. 2882–2904.

[89] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering, Springer, second ed., 2006.

[90] W. F. Osgood, *On functions of several complex variables*, Transactions of the American Mathematical Society, 17 (1916), pp. 1–8.

[91] A. Parusiński, *Subanalytic functions*, Transactions of the American Mathematical Society, (1994), pp. 583–595.

[92] G. B. Passty, *Ergodic convergence to a zero of the sum of monotone operators in Hilbert space*, Journal of Mathematical Analysis and Applications, 72 (1979), pp. 383–390.

[93] A. L. Patterson, *A Fourier series method for the determination of the components of interatomic distances in crystals*, Physical Review, 46 (1934), pp. 372–376.

[94] ——, *Ambiguities in the X-ray analysis of crystal structures*, Physical Review, 65 (1944), pp. 195–201.

[95] D. Perrone and P. Favaro, *Total variation blind deconvolution: The devil is in the details*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2909–2916.

[96] T. Pham Dinh and E. B. Souad, *Algorithms for solving a class of nonconvex optimization problems. methods of subgradients*, in North-Holland Mathematics Studies, vol. 129, Elsevier, 1986, pp. 249–271.

[97] B. T. Polyak, *Introduction to Optimization*, Optimization Software, 1987.

[98] M. Prato, R. Cavicchioli, L. Zanni, P. Boccacci, and M. Bertero, *Efficient deconvolution methods for astronomical imaging: Algorithms and IDL-GPU codes*, Astronomy & Astrophysics, 539 (2012), p. A133.

[99] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles Algorithms and Applications*, Pearson, 2006.

[100] A. Repetti, J. Birdi, A. Dabbech, and Y. Wiaux, *Non-convex optimization for self-calibration of direction-dependent effects in radio interferometric imaging*, Monthly Notices of the Royal Astronomical Society, 470 (2017), pp. 3981–4006.

[101] W. H. Richardson, *Bayesian-based iterative method of image restoration*, Journal of the Optical Society of America, 62 (1972), pp. 55–59.

[102] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, vol. 317 of Grundlehren der Mathematischen Wissenschaften, Springer, 1998.

[103] P. J. Schreier and L. L. Schaef, *Statistical Signal Processing of Complex-Valued Data: The Theory of Improper and Noncircular Signals*, Cambridge University Press, 2010.

[104] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, *Phase retrieval with application to optical imaging: A contemporary overview*, IEEE Signal Processing Magazine, 32 (2015), pp. 87–109.

[105] H. Shen and J. Z. Huang, *Sparse principal component analysis via regularized low rank matrix approximation*, Journal of Multivariate Analysis, 99 (2008), pp. 1015–1034.

[106] K. Shen and W. Yu, *Fractional programming for communication systems—part I: Power control and beamforming*, IEEE Transactions on Signal Processing, 66 (2018), pp. 2616–2630.

[107] ——, *Fractional programming for communication systems—part II: Uplink scheduling via matching*, IEEE Transactions on Signal Processing, 66 (2018), pp. 2631–2644.

[108] N. Z. Shor, *Application of the gradient method for the solution of network transportation problems*, in Scientific Seminar on Theory and Application of Cybernetics and Operations Research, Academy of Sciences U.S.S.R., 1962.

[109] V. Smith, S. Forte, M. Chenxin, M. Takáč, M. I. Jordan, and M. Jaggi, *CoCoA: A general framework for communication-efficient distributed optimization*, Journal of Machine Learning Research, 18 (2018), pp. 1–49.

[110] T. G. Stockham, T. M. Cannon, and R. B. Ingebretsen, *Blind deconvolution through digital signal processing*, Proceedings of the IEEE, 63 (1975), pp. 678–692.

[111] J. Sun, Q. Qu, and J. Wright, *A geometric analysis of phase retrieval*, Foundations of Computational Mathematics, 18 (2018), pp. 1131–1198.

[112] Y. Sun, P. Babu, and D. P. Palomar, *Majorization-minimization algorithms in signal processing, communications, and machine learning*, IEEE Transactions on Signal Processing, 65 (2017), pp. 794–816.

[113] S. Takahashi, M. Fukuda, and M. Tanaka, *New Bregman proximal type algorithms for solving DC optimization problems*, Computational Optimization and Applications, 83 (2022), pp. 893–931.

[114] S. Takahashi, M. Tanaka, and S. Ikeda, *Blind deconvolution with non-smooth regularization via Bregman proximal DCAs*, Signal Processing, 202 (2023), p. 108734.

[115] The Event Horizon Telescope Collaboration, *First M87 event horizon telescope results. IV. Imaging the central supermassive black hole*, The Astrophysical Journal Letters, 875 (2019), p. L4.

[116] A. R. Thompson, J. M. Moran, and G. W. Swenson, *Interferometry and Synthesis in Radio Astronomy*, Astronomy and Astrophysics Library, Springer, third ed., 2017.

[117] R. Tibshirani, *Regression shrinkage and selection via the lasso: A retrospective*, Journal of the Royal Statistical Society: Series B, 58 (1996), pp. 267–288.

[118] R. J. Vanderbei, *Linear Programming: Foundations and Extensions*, vol. 196 of International Series in Operations Research & Management Science, Springer, 2020.

[119] I. Waldspurger, A. d'Aspremont, and S. Mallat, *Phase recovery, maxcut and complex semidefinite programming*, Mathematical Programming, 149 (2015), pp. 47–81.

[120] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, and R. Fergus, *Regularization of neural networks using DropConnect*, in Proceedings of the International Conference on Machine Learning, 2013, pp. 1058–1066.

[121] B. Wen, X. Chen, and T. K. Pong, *A proximal difference-of-convex algorithm with extrapolation*, Computational Optimization and Applications, 69 (2018), pp. 297–324.

[122] Z. Wu, C. Li, M. Li, and A. Lim, *Inertial proximal gradient methods with Bregman regularization for a class of nonconvex optimization problems*, Journal of Global Optimization, 79 (2021), pp. 617–644.

[123] L. Yang, T. K. Pong, and X. Chen, *Alternating direction method of multipliers for a class of nonconvex and nonsmooth problems with applications to background/foreground extraction*, SIAM Journal on Imaging Sciences, 10 (2017), pp. 74–110.

[124] P. Yu, G. Li, and T. K. Pong, *Kurdyka–Łojasiewicz exponent via inf-projection*, Foundations of Computational Mathematics, 22 (2022), pp. 1171–1217.

[125] W. Yu and T. Lan, *Transmitter optimization for the multi-antenna downlink with per-antenna power constraints*, IEEE Transactions on signal processing, 55 (2007), pp. 2646–2660.

[126] A. L. Yuille and A. Rangarajan, *The concave-convex procedure (CCCP)*, in Proceedings of the Advances in Neural Information Processing Systems, vol. 14, 2001.

[127] H. Zhang, A. Milzarek, Z. Wen, and W. Yin, *On the geometric analysis of a quartic-quadratic optimization problem under a spherical constraint*, Mathematical Programming, (2021), pp. 1–53.

[128] X. Zhang, R. Barrio, M. A. Martinez, H. Jiang, and L. Cheng, *Bregman proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems*, IEEE Access, 7 (2019), pp. 126515–126529.

[129] N. Zhao, Q. Wei, A. Basarab, D. Kouamé, and J.-Y. Tourneret, *Blind deconvolution of medical ultrasound images using a parametric model for the point spread function*, in Proceedings of the IEEE International Ultrasonics Symposium, 2016, pp. 1–4.

[130] J. Zhu, S. Rosset, R. Tibshirani, and T. Hastie, *1-norm support vector machines*, Advances in Neural Information Processing Systems, 16 (2003).