J. Nonlinear Var. Anal. 10 (2026), No. 1, pp. 139-161
Available online at http://jnva.biemdas.com
https://doi.org/10.23952/jnva.10.2026.1.06

A TWO-STEP INERTIAL BREGMAN ALTERNATING STRUCTURE-ADAPTED PROXIMAL GRADIENT DESCENT ALGORITHM FOR NONCONVEX AND NONSMOOTH PROBLEMS

JING ZHAO[1,*], CHENZHENG GUO[2]

[1]*College of Science, Civil Aviation University of China, Tianjin 300300, China*
[2]*Mathematics and Statistics, Xidian University, Xian 710126, China*

Abstract. In this paper, we propose an accelerated alternating structure-adapted proximal gradient descent algorithm for a class of nonconvex and nonsmooth nonseparable problems. The proposed algorithm is a monotone method which combines two-step inertial extrapolation and generalized Bregman distance. Under some assumptions, we prove that every cluster point of the sequence generated by our algorithm is a critical point. Furthermore, with the help of the Kurdyka–Łojasiewicz property, we establish the convergence of the whole sequence generated by proposed algorithm. In order to make the algorithm more effective and flexible, we also use some strategies to update the extrapolation parameter. Moreover, we report some preliminary numerical results on Poisson linear inverse problems to demonstrate the feasibility and effectiveness of the proposed algorithm.

Keywords. Accelerated methods; Bregman distance; Extrapolation; Kurdyka–Łojasiewicz property; Nonconvex and nonsmooth nonseparable optimization.

1. INTRODUCTION

In this paper, we investigate the problem of solving the following nonconvex and nonsmooth nonseparable optimization problem:

$$\min_{x\in\mathbb{R}^n, y\in\mathbb{R}^m} L(x,y) = f(x) + Q(x,y) + g(y), \tag{1.1}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ and $g:\mathbb{R}^m\to\mathbb{R}$ are proper, lower semicontinuous, and nonconvex functions, $Q:\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}\cup\{\infty\}$ is a proper, lower semicontinuous and biconvex function. Problem (1.1) is used in many application scenarios, such as Poisson linear inverse problems [1, 2, 3], signal recovery [4, 5, 6], nonnegative matrix facorization [7, 8], multi-modal learning for image classification [9], and so on.

A natural method to solve problem (1.1) is the alternating minimization (AM) method (also called block coordinate descent (BCD) method), which, from a given initial point $(x_0,y_0)\in$

*Corresponding author.
E-mail address: zhaojing200103@163.com (J. Zhao).
Received 15 November 2024; Accepted 15 May 2025; Published online 1 October 2025.

$\mathbb{R}^n \times \mathbb{R}^m$, generates the iterative sequence $\{(x_k, y_k)\}$ via the scheme:

$$\begin{cases} x_{k+1} \in \arg\min_{x\in\mathbb{R}^n}\{L(x, y_k)\}, \\ y_{k+1} \in \arg\min_{y\in\mathbb{R}^m}\{L(x_{k+1}, y)\}. \end{cases} \tag{1.2}$$

If $L(x,y)$ is convex and continuously differentiable, and strictly convex in one argument while the other is fixed, then the sequence converges to a critical point [10, 11].

To relax the requirements of AM method and remove the strict convexity assumption, Auslender [12] introduced proximal terms to (1.2) for convex function L:

$$\begin{cases} x_{k+1} \in \arg\min_{x\in\mathbb{R}^n}\{L(x, y_k) + \frac{1}{2\lambda_k}\|x - x_k\|_2^2\}, \\ y_{k+1} \in \arg\min_{y\in\mathbb{R}^m}\{L(x_{k+1}, y) + \frac{1}{2\mu_k}\|y - y_k\|_2^2\}, \end{cases} \tag{1.3}$$

where $\{\lambda_k\}_{k\in\mathbb{N}}$ and $\{\mu_k\}_{k\in\mathbb{N}}$ are positive sequences. The above proximal point method, which is called proximal alternating minimization (PAM) algorithm, was further extended to nonconvex nonsmooth functions. In [13], Attouch et al. applied (1.3) to solve nonconvex problem (1.1) and proved the sequence generated via (1.3) converges to a critical point. More convergence analysis of the proximal point method can be found in [14, 15, 16, 17]. Because the proximal alternating minimization algorithm requires an exact solution at each iteration-step, the subproblems are very expensive if the minimizers of subproblems are not given in a closed form. The linearization technique is one of the effective methods to overcome the absence of an analytic solution to the subproblems. Bolte, Sabach and Teboulle [18] proposed the following proximal alternating linearized minimization (PALM) algorithm under the condition that the coupling term $Q(x,y)$ is continuously differentiable:

$$\begin{cases} x_{k+1} \in \arg\min_{x\in\mathbb{R}^n}\{f(x) + \langle x, \nabla_x Q(x_k, y_k)\rangle + \frac{1}{2\lambda_k}\|x - x_k\|_2^2\}, \\ y_{k+1} \in \arg\min_{y\in\mathbb{R}^m}\{g(y) + \langle y, \nabla_y Q(x_{k+1}, y_k)\rangle + \frac{1}{2\mu_k}\|y - y_k\|_2^2\}. \end{cases} \tag{1.4}$$

For any fixed y_k, $\nabla_x Q(\cdot, y_k)$ is $L_{\nabla_x Q(\cdot,y_k)}$-Lipschitz continuous. Likewise, for any fixed x_k, $\nabla_y Q(x_k, \cdot)$ is $L_{\nabla_y Q(x_k,\cdot)}$-Lipschitz continuous. So the step-size λ_k and μ_k are limited to

$$\lambda_k \in \left(0, \frac{1}{L_{\nabla_x Q(\cdot,y_k)}}\right), \mu_k \in \left(0, \frac{1}{L_{\nabla_y Q(x_{k+1},\cdot)}}\right).$$

In this way, the solution of some subproblems may be expressed by a closed-form or can be easily calculated. The global convergence result was established if $L(x,y)$ satisfied the Kurdyka–Łojasiewicz property.

When f and g are continuously differentiable, a natural idea is to linearize f and g. Nikolova and Tan [19] proposed the corresponding algorithm, called the alternating structure-adapted proximal gradient descent (ASAP) algorithm with the following scheme:

$$\begin{cases} x_{k+1} \in \arg\min\limits_{x\in\mathbb{R}^n}\{Q(x, y_k) + \langle x, \nabla f(x_k)\rangle + \frac{1}{2\tau}\|x - x_k\|_2^2\}, \\ y_{k+1} \in \arg\min\limits_{y\in\mathbb{R}^m}\{Q(x_{k+1}, y) + \langle y, \nabla g(y_k)\rangle + \frac{1}{2\sigma}\|y - y_k\|_2^2\}, \end{cases} \tag{1.5}$$

where $\tau \in (0, \frac{1}{L_{\nabla f}})$ and $\sigma \in (0, \frac{1}{L_{\nabla g}})$. With the help of the Kurdyka–Łojasiewicz property, they established the convergence of the whole sequence generated by (1.5). The key point of the ASAP algorithm is that f and g are assumed to be globally Lipschitz smooth on the entire spaces, which is very restrictive and excludes many applications. In order to attenuate this

assumption, Bauschke, Bolte and Teboulle [1] introduced a NoLips algorithm, avoiding the dependence of the Lipschitz smoothness, which is then extended to the nonconvex case in [20, 21, 22]. In [23], Gao et al. proposed the following alternating structure-adapted Bregman proximal (ASABP) gradient descent algorithm:

$$\begin{cases} x_{k+1} \in \arg\min\limits_{x\in\mathbb{R}^n}\{Q(x,y_k)+\langle x,\nabla f(x_k)\rangle+\frac{1}{\tau_k}D_{\phi_1}(x,x_k)\},\\ y_{k+1} \in \arg\min\limits_{y\in\mathbb{R}^m}\{Q(x_{k+1},y)+\langle y,\nabla g(y_k)\rangle+\frac{1}{\sigma_k}D_{\phi_2}(y,y_k)\}, \end{cases}$$

where they choose the generalized Bregman functions ϕ_1 and ϕ_2 such that the pairs (f,ϕ_1) and (g,ϕ_2) are GL-smad on x and y (see Definition 2.4 and Definition 2.7). By associating the generalized Bregman functions ϕ_1 and ϕ_2 to the objective functions f and g in a suitable way, and merely assuming that the underlying function satisfies the Kurdyka–Łojasiewicz property yet without the Lipschitz smoothness, they established the global convergence to a critical point.

The inertial extrapolation technique has been widely used to accelerate various algorithms for convex and nonconvex optimizations since the cost of each iteration stays basically unchanged [24, 25, 26, 27, 28]. The inertial scheme, starting from the so-called heavy ball method of Polyak [29], was recently proved to be very efficient in accelerating numerical methods, especially the first-order methods. Alvarez and Attouch [30] applied the inertial strategy to the proximal point method and proved that it could improve the rate of convergence. The main feature of the idea is that the new iteration use the previous two or more iterations.

Based on (1.4), Pock and Sabach [31] proposed the following inertial proximal alternating linearized minimization (iPALM) algorithm:

$$\begin{cases} u_{1k}=x_k+\alpha_{1k}(x_k-x_{k-1}), v_{1k}=x_k+\beta_{1k}(x_k-x_{k-1}),\\ x_{k+1}\in\arg\min_{x\in\mathbb{R}^n}\{f(x)+\langle x,\nabla_x Q(v_{1k},y_k)\rangle+\frac{1}{2\lambda_k}\|x-u_{1k}\|_2^2\},\\ u_{2k}=y_k+\alpha_{2k}(y_k-y_{k-1}), v_{2k}=y_k+\beta_{2k}(y_k-y_{k-1}),\\ y_{k+1}\in\arg\min_{y\in\mathbb{R}^m}\{g(y)+\langle y,\nabla_y Q(x_{k+1},v_{2k})\rangle+\frac{1}{2\mu_k}\|y-u_{2k}\|_2^2\}, \end{cases}$$

where $\alpha_{1k},\alpha_{2k},\beta_{1k},\beta_{2k}\in[0,1]$. They proved that the generated sequence globally converges to critical point of the objective function under the condition of the Kurdyka–Łojasiewicz property. When $\alpha_{1k}\equiv\alpha_{2k}\equiv\beta_{1k}\equiv\beta_{2k}\equiv 0$, iPALM reduces to PALM. Then Gao, Cai and Han [32] presented a Gauss–Seidel type inertial proximal alternating linearized minimization (GiPALM) algorithm for solving problem (1.1):

$$\begin{cases} x_{k+1}\in\arg\min_{x\in\mathbb{R}^n}\{f(x)+\langle x,\nabla_x Q(\tilde{x}_k,\tilde{y}_k)\rangle+\frac{1}{2\lambda_k}\|x-\tilde{x}_k\|_2^2\},\\ \tilde{x}_{k+1}=x_{k+1}+\alpha(x_{k+1}-\tilde{x}_k),\alpha\in[0,1),\\ y_{k+1}\in\arg\min_{y\in\mathbb{R}^m}\{g(y)+\langle y,\nabla_y Q(\tilde{x}_{k+1},\tilde{y}_k)\rangle+\frac{1}{2\mu_k}\|y-\tilde{y}_k\|_2^2\},\\ \tilde{y}_{k+1}=y_{k+1}+\beta(y_{k+1}-\tilde{y}_k),\beta\in[0,1). \end{cases}$$

By using inertial extrapolation technique, Yang and Xu [33] proposed the following accelerated alternating structure-adapted proximal gradient descent (aASAP) algorithm:

$$\begin{cases} x_{k+1} \in \arg\min_{x\in\mathbb{R}^n}\{Q(x,\hat{y}_k) + \langle \nabla f(\hat{x}_k), x\rangle + \frac{1}{2\tau}\|x-\hat{x}_k\|_2^2\}, \\ y_{k+1} \in \arg\min_{y\in\mathbb{R}^m}\{Q(x_{k+1},y) + \langle \nabla g(\hat{y}_k), y\rangle + \frac{1}{2\sigma}\|y-\hat{y}_k\|_2^2\}, \\ u_{k+1} = x_{k+1} + \beta_k(x_{k+1}-x_k), v_{k+1} = y_{k+1} + \beta_k(y_{k+1}-y_k), \\ \text{if } L(u_{k+1},v_{k+1}) \le L(x_{k+1},y_{k+1}), \text{ then } \hat{x}_{k+1} = u_{k+1}, \hat{y}_{k+1} = v_{k+1}, \\ \text{else } \hat{x}_{k+1} = x_{k+1}, \hat{y}_{k+1} = y_{k+1}. \end{cases} \tag{1.6}$$

Compared with the traditional extrapolation algorithm, the main difference is to ensure that the algorithm is monotone in terms of objective function values, while general extrapolation algorithms may be nonmonotonic.

The Bregman distance gives us alternative ways for more flexibility in the selection of regularization. Bregman distance regularization is an effective way to improve the numerical results of various algorithms. In [34], the authors constructed the following two-step inertial Bregman alternating minimization algorithm by using the information of the previous three iterates:

$$\begin{cases} x_{k+1} \in \arg\min_{x\in\mathbb{R}^n}\{L(x,y_k) + D_{\phi_1}(x,x_k) + \alpha_{1k}\langle x, x_{k-1}-x_k\rangle + \alpha_{2k}\langle x, x_{k-2}-x_{k-1}\rangle\}, \\ y_{k+1} \in \arg\min_{y\in\mathbb{R}^m}\{L(x_{k+1},y) + D_{\phi_2}(y,y_k) + \beta_{1k}\langle y, y_{k-1}-y_k\rangle + \beta_{2k}\langle y, y_{k-2}-y_{k-1}\}, \end{cases}$$

where $D_{\phi_i}(i=1,2)$ denotes the Bregman distance with respect to $\phi_i(i=1,2)$, respectively. The convergence was obtained provided that an appropriate regularization of the objective function satisfies the Kurdyka–Łojasiewicz inequality. Based on alternating minimization algorithm, Cho, Nong and Zhao [35] proposed the following inertial alternating minimization with Bregman distance (BIAM) algorithm:

$$\begin{cases} x_{k+1} \in \arg\min_{x\in\mathbb{R}^n}\{f(x) + Q(x,\hat{y}_k) + \lambda_k D_{\phi_1}(x,\hat{x}_k)\}, \\ \hat{x}_{k+1} = x_{k+1} + \alpha(x_{k+1}-\hat{x}_k), \alpha \in [0,1), \\ y_{k+1} \in \arg\min_{y\in\mathbb{R}^m}\{g(y) + Q(\hat{x}_{k+1},y) + \mu_k D_{\phi_2}(y,\hat{y}_k)\}, \\ \hat{y}_{k+1} = y_{k+1} + \beta(y_{k+1}-\hat{y}_k), \beta \in [0,1). \end{cases}$$

Suppose that the benefit function satisfies the Kurdyka–Łojasiewicz property and the parameters are selected appropriately, they proved the convergence of BIAM algorithm.

In this paper, based on the alternating structure-adapted proximal gradient method, we combine inertial extrapolation technique and a generalized Bregman distance to construct a two-step inertial Bregman alternating structure-adapted proximal gradient descent algorithm. In order to make the proposed algorithm more effective and flexible, we also use some strategies to update the extrapolation parameter. Under some assumptions about the penalty parameter and objective function, the convergence of the proposed algorithm is obtained based on the Kurdyka–Łojasiewicz property yet the underlying function without the Lipschitz smoothness. Moreover, we report some preliminary numerical results on Poisson linear inverse problem to show the feasibility and effectiveness of the proposed method.

The article is organized as follows. In Section 2, we recall some concepts and important lemmas which are used in the proof of main results. In Section 3, we present the two-step inertial Bregman alternating structure-adapted proximal gradient descent algorithm and analyze

its convergence. Finally, in Section 5, the preliminary numerical examples on Poisson linear inverse problem are provided to illustrate the behavior of the proposed algorithm.

2. PRELIMINARIES

Consider the Euclidean space $\mathbb{R}^d$ of dimension $d \geq 1$. The standard inner product and the induced norm on $\mathbb{R}^d$ are denoted by $\langle \cdot, \cdot \rangle$ and $\|\cdot\|_2$, respectively. We use $\omega(x_k) = \{x : \exists x_{k_j} \to x\}$ to stand for the limit set of $\{x_k\}_{k\in\mathbb{N}}$. The *domain* of f is defined as $\mathrm{dom} f := \{x \in \mathbb{R}^d : f(x) < +\infty\}$. We say that f is *proper* if $\mathrm{dom} f \neq \emptyset$, and f is called *lower semicontinuous* at x if $f(x) \leq \liminf_{k\to\infty} f(x_k)$ for every sequence $\{x_k\}$ converging to x. If f is lower semicontinuous in its domain, we say f is a lower semicontinuous function. If $\mathrm{dom} f$ is closed and f is lower semicontinuous over $\mathrm{dom} f$, then f is a closed function. Further we recall some generalized subdifferential notions and the basic properties which are needed in this paper.

2.1. **Subdifferentials.**

Definition 2.1. (Subdifferentials) Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function. For $x \in \mathrm{dom} f$, the Fréchet subdifferential of f at x, written as $\hat{\partial} f(x)$, is the set of vectors $v \in \mathbb{R}^d$ which satisfies

$$\liminf_{y\to x} \frac{1}{\|x-y\|_2}[f(y) - f(x) - \langle v, y-x\rangle] \geq 0.$$

If $x \notin \mathrm{dom} f$, then $\hat{\partial} f(x) = \emptyset$. The limiting-subdifferential [36], or simply the subdifferential for short, of f at $x \in \mathrm{dom} f$, written as $\partial f(x)$, is defined as follows:

$$\partial f(x) := \{v \in \mathbb{R}^d : \exists x_k \to x, f(x_k) \to f(x), v_k \in \hat{\partial} f(x_k), v_k \to v\}.$$

Remark 2.1. (a) The above definition implies that $\hat{\partial} f(x) \subseteq \partial f(x)$ for each $x \in \mathbb{R}^d$, where the first set is convex and closed while the second one is closed (see [37]).

(b) (Closedness of ∂f) Let $\{x_k\}_{k\in\mathbb{N}}$ and $\{v_k\}_{k\in\mathbb{N}}$ be sequences in $\mathbb{R}^d$ such that $v_k \in \partial f(x_k)$ for all $k \in \mathbb{N}$. If $(x_k, v_k) \to (x, v)$ and $f(x_k) \to f(x)$ as $k \to \infty$, then $v \in \partial f(x)$.

(c) If $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous and $h : \mathbb{R}^d \to \mathbb{R}$ is a continuously differentiable function, then $\partial(f + h)(x) = \partial f(x) + \nabla h(x)$ for all $x \in \mathbb{R}^d$.

In what follows, we consider the problem of finding a critical point $(x^*, y^*) \in \mathrm{dom} L$.

Lemma 2.1. (Fermat's rule [37]) *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function. If f has a local minimum at x^*, then $0 \in \partial f(x^*)$.*

We call x^* is a critical point of f if $0 \in \partial f(x^*)$. The set of all critical points of f is denoted by $\mathrm{crit} f$.

2.2. **The Kurdyka–Łojasiewicz property.**

Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function. For η_1, η_2 such that $-\infty < \eta_1 < \eta_2 \leq +\infty$, we set $[\eta_1 < f < \eta_2] = \{x \in \mathbb{R}^d : \eta_1 < f(x) < \eta_2\}$. For $\eta > 0$, we denote by Φ_η the class of continuous concave function $\varphi : [0, \eta) \to \mathbb{R}_+$ such that $\varphi(0) = 0$, φ is C^1 on $(0, \eta)$ and $\varphi'(s) > 0$, $\forall s \in (0, \eta)$.

Definition 2.2. (Kurdyka–Łojasiewicz property [13]) Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function.

(i) $f:\mathbb{R}^d\to(-\infty,+\infty]$ is said to have the Kurdyka–Łojasiewicz (KŁ) property at $x^\star\in\mathrm{dom}f$ if there exist $\eta\in(0,+\infty]$, a neighborhood U of $x^\star$, and a function $\varphi\in\Phi_\eta$ such that, for all x in $U\cap[f(x^\star)<f<f(x^\star)+\eta]$, the Kurdyka–Łojasiewicz inequality holds,

$$\varphi'(f(x)-f(x^\star))\mathrm{dist}(0,\partial f(x))\geq 1.$$

(ii) Proper lower semicontinuous functions which have the Kurdyka–Łojasiewicz property at each point of its domain are called KŁ functions.

Lemma 2.2. (Uniformized KŁ property [37]) *Let Ψ be a compact set and let $f:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$ be a proper and lower semicontinuous function. Assume that f is constant on Ψ and satisfies the KŁ property at each point of Ψ. Then, there exist $\varepsilon>0,\eta>0$ and $\varphi\in\Phi_\eta$ such that, for all $x^\star\in\Psi$ and for all $x\in\{x\in\mathbb{R}^d:\mathrm{dist}(x,\Psi)<\varepsilon\}\cap[f(x^\star)<f<f(x^\star)+\eta]$,*

$$\varphi'(f(x)-f(x^\star))\mathrm{dist}(0,\partial f(x))\geq 1.$$

There is a broad class of functions satisfying the KŁ property, such as strongly convex functions, real analytic functions, semi-algebraic functions [13], subanalytic functions [38], log-exp functions, and so on.

2.3. **Generalized Bregman function and Bregman distance.**

Definition 2.3. A function f is said convex if $\mathrm{dom}f$ is a convex set and if, for all $x,y\in\mathrm{dom}f$, $\alpha\in[0,1]$, $f(\alpha x+(1-\alpha)y)\leq\alpha f(x)+(1-\alpha)f(y)$. f is said θ-strongly convex with $\theta>0$ if $f-\frac{\theta}{2}\|\cdot\|^2$ is convex, i.e.,

$$f(\alpha x+(1-\alpha)y)\leq\alpha f(x)+(1-\alpha)f(y)-\frac{1}{2}\theta\alpha(1-\alpha)\|x-y\|^2$$

for all $x,y\in\mathrm{dom}f$ and $\alpha\in[0,1]$.

Suppose that f is differentiable. Then f is convex if and only if $\mathrm{dom}f$ is a convex set and $f(x)\geq f(y)+\langle\nabla f(y),x-y\rangle$ holds for all $x,y\in\mathrm{dom}f$. Moreover, f is θ-strongly convex with $\theta>0$ if and only if $f(x)\geq f(y)+\langle\nabla f(y),x-y\rangle+\frac{\theta}{2}\|x-y\|^2$ for all $x,y\in\mathrm{dom}f$. To define the Bregman distance, we first give the definition of generalized Bregman function.

Definition 2.4. (Generalized Bregman function [23]) Let C be a nonempty and convex open subset of $\mathbb{R}^d$. Associated with C, a function $\phi:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$ is called a generalized Bregman function if it satisfies the following:

(i) ϕ is a proper lower semicontinuous, continuously differentiable, and strictly convex function.

(ii) $\mathrm{dom}\partial\phi=C=int\ \mathrm{dom}\phi$.

(iii) If $\{x_k\}\in C$ converges to $\bar{x}\in\partial C$ (the boundary of C), then $\lim_{k\to+\infty}\langle\nabla\phi(x_k),u-x_k\rangle=-\infty$ for all $u\in C$.

Definition 2.5. Let $\phi:\mathbb{R}^d\to(-\infty,+\infty]$ be a convex differentiable function. The function $D_\phi:\mathrm{dom}\phi\times\mathrm{dom}\phi\to[0,+\infty)$, defined by $D_\phi(x,y)=\phi(x)-\phi(y)-\langle\nabla\phi(y),x-y\rangle$, is called the Bregman distance with respect to ϕ.

From Definition 2.3, it follows that $D_\phi(x,y)\geq\frac{\theta}{2}\|x-y\|^2$ if ϕ is θ-strongly convex.

Remark 2.2. Note that the structural form of D_ϕ is also useful when ϕ is not convex, and still enjoys the following two simple but remarkable properties:

(i) The three point identity: For any $y,\ z \in int\ \mathrm{dom}\phi$ and $x \in \mathrm{dom}\phi$,

$$D_\phi(x,z) - D_\phi(x,y) - D_\phi(y,z) = \langle \nabla\phi(y) - \nabla\phi(z), x-y\rangle.$$

(ii) Linear additivity: For any $\alpha,\beta \in \mathbb{R}$, and any functions ϕ_1 and ϕ_2,

$$D_{\alpha\phi_1+\beta\phi_2}(x,y) = \alpha D_{\phi_1}(x,y) + \beta D_{\phi_2}(x,y),$$

for any $(x,y) \in (\mathrm{dom}\phi_1 \cap \mathrm{dom}\phi_2)^2$ such that both ϕ_1 and ϕ_2 are differentiable at y.

It is obvious that the Bregman distance is, in general, not symmetric. It is thus natural to introduce a measure for the symmetry of D_ϕ.

Definition 2.6. (Symmetry coefficient [1]) Given $\phi \in \mathscr{G}(C)$, its symmetry coefficient is defined by

$$\alpha(\phi) = \inf\left\{D_\phi(x,y)/D_\phi(y,x) \mid (x,y) \in int\ \mathrm{dom}\phi \times int\ \mathrm{dom}\phi,\ x \neq y\right\} \in [0,1].$$

Remark 2.3. The symmetry coefficient has the following properties:

(i) Clearly, the closer $\alpha(\phi)$ gets to 1, the more symmetric D_ϕ is.

(ii) For any $x,\ y \in int\ \mathrm{dom}\phi$, $\alpha(\phi)D_\phi(x,y) \le D_\phi(y,x) \le \frac{1}{\alpha(\phi)}D_\phi(x,y)$, where we have adopted the convention that $\frac{1}{0} = +\infty$ and $+\infty \times r = +\infty$ for any $r > 0$.

Lemma 2.3. *Given $\phi \in \mathscr{G}(C)$, for any proper, lower semicontinuous, and convex function $\Gamma : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ and any $z \in int\ \mathrm{dom}\phi$, if $z_+ = \arg\min_{x\in\overline{C}}\{\Gamma(x) + D_\phi(x,z)\}$, then, for any $x \in \mathrm{dom}\phi$, $\Gamma(z_+) + D_\phi(z_+,z) \le \Gamma(x) + D_\phi(x,z) - D_\phi(x,z_+)$.*

2.4. **Generalized L-smooth adaptable and extended descent lemma.** We denote $\mathscr{G}(f,\phi)$ the set of pair of functions (f,ϕ) satisfying

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

(ii) $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is proper lower semicontinuous nonconvex with $\mathrm{dom}\phi \subset \mathrm{dom}f$, which is continuously differentiable on C.

Definition 2.7. (Generalized L-smooth adaptable) A pair of functions $(f,\phi) \in \mathscr{G}(f,\phi)$ is called generalized L-smooth adaptable (GL-smad) on C if there exists $L > 0$ such that $Lh + f$ and $Lh - f$ are convex on C.

From the above definition, we immediately obtain the two-sided descent lemma, which complements and extends the NoLips descent lemma derived in [1].

Lemma 2.4. (Extended descent lemma) *The pair of functions $(f,\phi) \in \mathscr{G}(f,\phi)$ is GL-smad on C if and only if*

$$|f(x) - f(y) - \langle \nabla f(y), x-y\rangle| \le LD_\phi(x,y),\ \forall x \in \mathrm{dom}\phi,\ \forall \mathrm{y} \in \mathrm{int\ dom}\phi.$$

Due to the structural form of D_ϕ, the extended descent lemma reads equivalently as

$$\left|f(x) - f(z) - \langle \nabla f(y), x-z\rangle + D_f(z,y)\right| \le LD_\phi(x,y),\ \forall x,\ z \in \mathrm{dom}\phi,\ \forall \mathrm{y} \in \mathrm{int\ dom}\phi. \quad (2.1)$$

When f is assumed to be convex, the convexity condition of $Lh + f$ naturally holds. In this case, the NoLips descent lemma given in [1] is recovered. When $\phi(z) = \frac{1}{2}\|z\|^2$ and consequently

$D_\phi(x,y) = \frac{1}{2}\|x-y\|^2$, Lemma 2.4 would be reduced to the descent lemma [39],

$$|f(x) - f(y) - \langle \nabla f(y), x-y\rangle| \le \frac{L}{2}\|x-y\|^2,\ \forall x \in \mathrm{dom}\phi,\ \forall \mathrm{y} \in \mathrm{int\ dom}\phi.$$

3. Two-step Inertial Bregman ASAP Algorithm

Assumption 3.1. (i) $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ is lower bounded.

(ii) $\phi_1 \in \mathscr{G}(X)$ and $\phi_2 \in \mathscr{G}(Y)$ such that the pairs $(f,\phi_1) \in \mathscr{G}(f,\phi_1)$ and $(g,\phi_2) \in \mathscr{G}(g,\phi_2)$ are GL-smad on X and Y with coefficients L_{ϕ_1} and L_{ϕ_2}, respectively.

(iii) For the sequence $\{x_k\} \in int\ \mathrm{dom}\phi_1$ and $x \in \mathrm{dom}\phi_1$, $\|x_k - x\| \to 0 \Leftrightarrow D_{\phi_1}(x,x_k) \to 0$. Similarly, $\|y_k - y\| \to 0 \Leftrightarrow D_{\phi_2}(y,y_k) \to 0$ for any sequence $\{y_k\} \in int\ \mathrm{dom}\phi_2$ and $y \in \mathrm{dom}\phi_2$.

(iv) $Q : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ is a proper, lower semicontinuous, and biconvex function.

(v) $\phi_i (i = 1,2)$ is μ_{ϕ_i}-strongly convex and differentiable function. And the gradient $\nabla\phi_i$ is L_i-Lipschitz continuous, i.e.,

$$\|\nabla\phi_1(x) - \nabla\phi_1(\hat{x})\| \le L_1\|x-\hat{x}\|,\ \|\nabla\phi_2(y) - \nabla\phi_2(\hat{y})\| \le L_2\|y-\hat{y}\|,\ \forall y,\hat{y} \in \mathbb{R}^m. \tag{3.1}$$

Remark 3.1. Combining linear additivity of the Bregman distance (see Remark 2.2 (ii)) and Assumption 3.1 (ii) and (v), ∇f is Lipschitz continuous with coefficient $L_{\phi_1}L_1$ on any bounded subset. Similarly, ∇g is Lipschitz continuous with coefficient $L_{\phi_2}L_2$ on any bounded subset.

Algorithm 1 TiBASAP: Two-step inertial Bregman alternating structure-adapted proximal gradient descent algorithm

---

Require: Take $(x_0,y_0) \in \mathbb{R}^n \times \mathbb{R}^m, (\hat{x}_0,\hat{y}_0) = (x_0,y_0), \alpha_k \in [0,\alpha_{\max}], \beta_k \in [0,\beta_{\max}], \alpha_{\max} + \beta_{\max} < 1$, and $k = 0$.

1. **Compute**

$$\begin{cases} x_{k+1} \in \arg\min\limits_{x\in\mathbb{R}^n}\{Q(x,\hat{y}_k) + \langle \nabla f(\hat{x}_k), x-\hat{x}_k\rangle + \dfrac{1}{\tau_k}D_{\phi_1}(x,\hat{x}_k)\}, \\ y_{k+1} \in \arg\min\limits_{y\in\mathbb{R}^m}\{Q(x_{k+1},y) + \langle \nabla g(\hat{y}_k), y-\hat{y}_k\rangle + \dfrac{1}{\sigma_k}D_{\phi_2}(y,\hat{y}_k)\}. \end{cases} \tag{3.2}$$

2.
$$\begin{pmatrix} u_{k+1} \\ v_{k+1} \end{pmatrix} = \begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} + \alpha_k \begin{pmatrix} x_{k+1}-x_k \\ y_{k+1}-y_k \end{pmatrix} + \beta_k \begin{pmatrix} x_k - x_{k-1} \\ y_k - y_{k-1} \end{pmatrix}.$$

3. **If** $L(u_{k+1},v_{k+1}) \le L(x_{k+1},y_{k+1})$, **then**

$$\hat{x}_{k+1} = u_{k+1},\ \hat{y}_{k+1} = v_{k+1}, \tag{3.3}$$

else

$$\hat{x}_{k+1} = x_{k+1},\ \hat{y}_{k+1} = y_{k+1}. \tag{3.4}$$

4. **Set** $k \leftarrow k+1$, go to step1.

Remark 3.2. We discuss the relation of Algorithm 1 to the other existing algorithms from the literature.

(i) If we take $\phi_1(x) = \frac{1}{2}\|x\|_2^2$, $\phi_2(y) = \frac{1}{2}\|y\|_2^2$, $\tau_k \equiv \tau$ and $\sigma_k \equiv \sigma$ for all $x \in \mathbb{R}^n$, and $y \in \mathbb{R}^m$, then Algorithm 1 becomes the following iterative method:

$$\begin{cases} x_{k+1} \in \arg\min_{x\in\mathbb{R}^n}\{Q(x,\hat{y}_k) + \langle \nabla f(\hat{x}_k), x\rangle + \frac{1}{2\tau}\|x-\hat{x}_k\|_2^2\}, \\ y_{k+1} \in \arg\min_{y\in\mathbb{R}^m}\{Q(x_{k+1},y) + \langle \nabla g(\hat{y}_k), y\rangle + \frac{1}{2\sigma}\|y-\hat{y}_k\|_2^2\}, \\ u_{k+1} = x_{k+1} + \alpha_k(x_{k+1}-x_k) + \beta_k(x_k - x_{k-1}), \\ v_{k+1} = y_{k+1} + \alpha_k(y_{k+1}-y_k) + \beta_k(y_k - y_{k-1}), \\ \text{if } L(u_{k+1},v_{k+1}) \le L(x_{k+1},y_{k+1}), \text{ then } \hat{x}_{k+1} = u_{k+1}, \hat{y}_{k+1} = v_{k+1}, \\ \text{else } \hat{x}_{k+1} = x_{k+1}, \hat{y}_{k+1} = y_{k+1}. \end{cases} \tag{3.5}$$

(ii) Letting $\beta_k \equiv 0$ for all $k \ge 0$, one sees that (3.5) becomes the accelerated alternating structure-adapted proximal gradient descent (aASAP) algorithm (1.6).

(iii) Letting $\alpha_k \equiv \beta_k \equiv 0$ for all $k \ge 0$, one sees that (3.5) becomes the alternating structure-adapted proximal gradient descent (ASAP) algorithm (1.5).

Remark 3.3. Compared with the traditional extrapolation algorithm, the main difference is Step 3 which ensures the algorithm is a monotone method in terms of objective function value, while general extrapolation algorithms may be nonmonotonic.

For extrapolation parameters α_k and β_k, there are at least two ways to choose them, either as costant or by dynamic update. For example, in [40, 41], it was defined as

$$\begin{cases} \alpha_k = \beta_k = \frac{t_{k-1}-1}{2t_k}, \\ t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}, \end{cases} \tag{3.6}$$

where $t_{-1} = t_0 = 1$. In order to make Algorithm 1 more effective, we present an adaptive method to update α_k and β_k, which are given in Algorithm 2.

Algorithm 2 Two-step inertial Bregman alternating structure-adapted proximal gradient descent with adaptive extrapolation parameter algorithm

Require: Take $(x_0,y_0) \in \mathbb{R}^n \times \mathbb{R}^m$, $(\hat{x}_0,\hat{y}_0) = (x_0,y_0)$, $\alpha_0 \in [0,\alpha_{\max}]$, $\beta_0 \in [0,\beta_{\max}]$, $\alpha_{\max} + \beta_{\max} < 1$, $t > 1$, and $k = 0$.

1. **Compute**

$$\begin{cases} x_{k+1} \in \arg\min\limits_{x\in\mathbb{R}^n}\{Q(x,\hat{y}_k) + \langle \nabla f(\hat{x}_k), x-\hat{x}_k\rangle + \dfrac{1}{\tau_k}D_{\phi_1}(x,\hat{x}_k)\}, \\ y_{k+1} \in \arg\min\limits_{y\in\mathbb{R}^m}\{Q(x_{k+1},y) + \langle \nabla g(\hat{y}_k), y-\hat{y}_k\rangle + \dfrac{1}{\sigma_k}D_{\phi_2}(y,\hat{y}_k)\}. \end{cases}$$

2.
$$\begin{pmatrix} u_{k+1} \\ v_{k+1} \end{pmatrix} = \begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} + \alpha_k \begin{pmatrix} x_{k+1}-x_k \\ y_{k+1}-y_k \end{pmatrix} + \beta_k \begin{pmatrix} x_k - x_{k-1} \\ y_k - y_{k-1} \end{pmatrix}.$$

3. **If** $L(u_{k+1},v_{k+1}) \le L(x_{k+1},y_{k+1})$, **then**

$$\hat{x}_{k+1} = u_{k+1},\ \hat{y}_{k+1} = v_{k+1},\ \alpha_{k+1} = \min\{t\alpha_k, \alpha_{\max}\},\ \beta_{k+1} = \min\{t\beta_k, \beta_{\max}\},$$

else

$$\hat{x}_{k+1} = x_{k+1},\ \hat{y}_{k+1} = y_{k+1},\ \alpha_{k+1} = \alpha_k/t,\ \beta_{k+1} = \beta_k/t.$$

4. **Set** $k \leftarrow k+1$, go to Step1.

Remark 3.4. Compared with constant or dynamic update by (3.6), the adaptive extrapolation parameters α_k and β_k can make Algorithm 2 more effective and flexible. The numerical results presented in this paper verify the effectiveness of the adaptive strategy in Section 5.

4. Convergence Analysis

In this section, we prove the convergence of Algorithm 1. Note that the bound of α_k and β_k is no more than $\alpha_{\max}$ and $\beta_{\max}$ in Algorithm 2, respectively. So the convergence properties of Algorithm 1 are also applicable for Algorithm 2.

Under Assumption 3.1, some convergence results are proved (see Lemma 4.1). We also consider the following additional assumptions to establish stronger convergence results.

Assumption 4.1. (i) L is coercive and the domain of Q is closed.

(ii) The subdifferential of Q obeys:

$$\forall (x,y) \in \operatorname{dom} Q, \partial_x Q(x,y) \times \partial_y Q(x,y) \subset \partial Q(x,y).$$

(iii) $Q : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ has the following form $Q(x,y) = q(x,y) + h(x)$, where $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is continuous on its domain; $q : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ is a continuous function on dom Q such that, for any y, the partial function $q(\cdot, y)$ is continuously differentiable about x. Besides, for each bounded subset $D_1 \times D_2 \subset \operatorname{dom} Q$, there exists $\xi > 0$, such that, for any $\bar{x} \in D_1$, $(y, \bar{y}) \in D_2 \times D_2$, it holds that $\|\nabla_x q(\bar{x}, y) - \nabla_x q(\bar{x}, \bar{y})\| \leq \xi \|y - \bar{y}\|$.

Remark 4.1. (i) Assumption 4.1(i) ensures that the sequences generated by our proposed algorithms is bounded which plays an important role in the proof of convergence.

(ii) Assumption 4.1(ii) is a generic assumption for the convergence of alternating schemes. Because f and g are continuously differentiable, we have $\partial_x L(x,y) \times \partial_y L(x,y) \subset \partial L(x,y)$ by Remark 2.1 (c).

(iii) From Assumptions 3.1 and 4.1, $L(x,y)$ is continuous on its domain, equal to dom Q, which is nonempty and closed. For any sequence $\{(x_k, y_k)\}$ converges to $(\bar{x}, \bar{y})$, it holds that $\{L(x_k, y_k)\}$ converges to $L(\bar{x}, \bar{y})$.

For more convenient notation, we introduce the following useful notations:

$$\underline{\tau} \leq \inf \tau_k \leq \sup \tau_k \leq \overline{\tau}, \ \underline{\sigma} \leq \inf \sigma_k \leq \sup \sigma_k \leq \overline{\sigma}, \ \forall k \in \mathbb{N}.$$

Defining $\rho_k = \min\{\frac{1+\alpha(\phi_1)}{\tau_k} - L_{\phi_1}, \frac{1+\alpha(\phi_2)}{\sigma_k} - L_{\phi_2}\}$, together with $0 < \tau_k < \frac{1+\alpha(\phi_1)}{L_{\phi_1}}$ and $0 < \sigma_k < \frac{1+\alpha(\phi_2)}{L_{\phi_2}}$, we have $\rho_k > 0$.

Lemma 4.1. *Let Assumption 3.1 hold. Let $\{z_k = (x_k, y_k)\}$ and $\{\hat{z}_k = (\hat{x}_k, \hat{y}_k)\}$ be the sequences generated by Algorithm 1. Then the following assertions hold.*

(i) *$\{L(z_k)\}$ is monotonically nonincreasing. In particular, there exists $\rho > 0$ such that*

$$L(z_k) - L(z_{k+1}) \geq \rho \left[D_{\phi_1}(x_{k+1}, \hat{x}_k) + D_{\phi_2}(y_{k+1}, \hat{y}_k) \right]. \tag{4.1}$$

where $\rho = \min\{\frac{1+\alpha(\phi_1)}{\overline{\tau}} - L_{\phi_1}, \frac{1+\alpha(\phi_2)}{\overline{\sigma}} - L_{\phi_2}\}$. Moreover, $\{L(z_k)\}$ converges to some finite value, denoted by L^*.

(ii) *It holds that $\sum_{k=0}^{\infty} \left(D_{\phi_1}(x_{k+1}, \hat{x}_k) + D_{\phi_2}(y_{k+1}, \hat{y}_k) \right) < +\infty$, $\lim_{k \to +\infty} D_{\phi_1}(x_{k+1}, \hat{x}_k) = 0$ and $\lim_{k \to +\infty} D_{\phi_2}(y_{k+1}, \hat{y}_k) = 0$. Moreover, $\lim_{k \to +\infty} \|z_{k+1} - \hat{z}_k\| = 0$.*

(iii) (Convergence rate) *For any $K \geq 0$, it holds that*

$$\min_{0\leq k\leq K}\left\{D_{\phi_1}(x_{k+1},\hat{x}_k)+D_{\phi_2}(y_{k+1},\hat{y}_k)\right\} \leq \frac{1}{\rho(K+1)}\left(L(z_0)-L^*\right).$$

Proof. (i) Applying Lemma 2.3 with $\Gamma(x)=\tau_k\left(Q(x,\hat{y}_k)+\langle\nabla f(\hat{x}_k),x-\hat{x}_k\rangle\right)$, it yields that, for any $x\in \mathrm{dom}\phi_1$,

$$\begin{aligned}
&Q(x_{k+1},\hat{y}_k)-Q(x,\hat{y}_k)\\
&\leq \langle\nabla f(\hat{x}_k),x-x_{k+1}\rangle+\frac{1}{\tau_k}D_{\phi_1}(x,\hat{x}_k)-\frac{1}{\tau_k}D_{\phi_1}(x,x_{k+1})-\frac{1}{\tau_k}D_{\phi_1}(x_{k+1},\hat{x}_k)\\
&\leq f(x)-f(x_{k+1})+L_{\phi_1}D_{\phi_1}(x,\hat{x}_k)+D_f(x_{k+1},\hat{x}_k)+\frac{1}{\tau_k}D_{\phi_1}(x,\hat{x}_k)-\frac{1}{\tau_k}D_{\phi_1}(x,x_{k+1})\\
&\quad-\frac{1}{\tau_k}D_{\phi_1}(x_{k+1},\hat{x}_k)\\
&\leq f(x)-f(x_{k+1})+\left(\frac{1}{\tau_k}+L_{\phi_1}\right)D_{\phi_1}(x,\hat{x}_k)-\frac{1}{\tau_k}D_{\phi_1}(x,x_{k+1})-\left(\frac{1}{\tau_k}-L_{\phi_1}\right)D_{\phi_1}(x_{k+1},\hat{x}_k),
\end{aligned}$$

where the second inequality follows from Lemma 2.4, and the last inequality follows from the GL-smad property of (f,ϕ). Particularly, taking $x=\hat{x}_k$, together with the nonnegativity of Bregman distance D_{ϕ_1} and Remark 2.3 (ii), we obtain

$$\begin{aligned}
f(x_{k+1})+Q(x_{k+1},\hat{y}_k)\leq& f(\hat{x}_k)+Q(\hat{x}_k,\hat{y}_k)-\frac{1}{\tau_k}D_{\phi_1}(\hat{x}_k,x_{k+1})-\left(\frac{1}{\tau_k}-L_{\phi_1}\right)D_{\phi_1}(x_{k+1},\hat{x}_k)\\
\leq& f(\hat{x}_k)+Q(\hat{x}_k,\hat{y}_k)-\left(\frac{1+\alpha(\phi_1)}{\tau_k}-L_{\phi_1}\right)D_{\phi_1}(x_{k+1},\hat{x}_k).
\end{aligned}\tag{4.2}$$

Similarly, one has

$$Q(x_{k+1},y_{k+1})+g(y_{k+1})\leq Q(x_{k+1},\hat{y}_k)+g(\hat{y}_k)-\left(\frac{1+\alpha(\phi_2)}{\sigma_k}-L_{\phi_2}\right)D_{\phi_2}(y_{k+1},\hat{y}_k).\tag{4.3}$$

Adding (4.2) and (4.3), we have

$$\begin{aligned}
&L(x_{k+1},y_{k+1})\\
\leq& L(\hat{x}_k,\hat{y}_k)-\left(\frac{1+\alpha(\phi_1)}{\tau_k}-L_{\phi_1}\right)D_{\phi_1}(x_{k+1},\hat{x}_k)-\left(\frac{1+\alpha(\phi_2)}{\sigma_k}-L_{\phi_2}\right)D_{\phi_2}(y_{k+1},\hat{y}_k)\\
\leq& L(x_k,y_k)-\left(\frac{1+\alpha(\phi_1)}{\overline{\tau}}-L_{\phi_1}\right)D_{\phi_1}(x_{k+1},\hat{x}_k)-\left(\frac{1+\alpha(\phi_2)}{\overline{\sigma}}-L_{\phi_2}\right)D_{\phi_2}(y_{k+1},\hat{y}_k)\\
\leq& L(x_k,y_k)-\rho\left[D_{\phi_1}(x_{k+1},\hat{x}_k)+D_{\phi_2}(y_{k+1},\hat{y}_k)\right],
\end{aligned}$$

where $\rho=\min\{\frac{1+\alpha(\phi_1)}{\overline{\tau}}-L_{\phi_1},\frac{1+\alpha(\phi_2)}{\overline{\sigma}}-L_{\phi_2}\}$, which can be abbreviated as

$$L(z_k)-L(z_{k+1})\geq\rho\left[D_{\phi_1}(x_{k+1},\hat{x}_k)+D_{\phi_2}(y_{k+1},\hat{y}_k)\right].\tag{4.4}$$

According to Assumption 3.1, we see L is lower bounded. Hence $\{L(z_k)\}$ converges to some real number L^*.

(ii) Using inequality (4.4), we have, for all $k\geq 0$,

$$D_{\phi_1}(x_{k+1},\hat{x}_k)+D_{\phi_2}(y_{k+1},\hat{y}_k)\leq\frac{1}{\rho}\left(L(z_k)-L(z_{k+1})\right).$$

For $K \geq 0$, summing from $k = 0$ to K and using the statement (i), we obtain

$$\sum_{k=0}^{K} \left(D_{\phi_1}(x_{k+1}, \hat{x}_k) + D_{\phi_2}(y_{k+1}, \hat{y}_k)\right) \leq \frac{1}{\rho}\left(L(z_0) - L(z_{K+1})\right) \leq \frac{1}{\rho}\left(L(z_0) - L^*\right) < +\infty. \tag{4.5}$$

Taking the limit as $k \to \infty$ leads to

$$\sum_{k=0}^{\infty} \left(D_{\phi_1}(x_{k+1}, \hat{x}_k) + D_{\phi_2}(y_{k+1}, \hat{y}_k)\right) < +\infty,$$

which deduces that $\lim_{k\to+\infty} D_{\phi_1}(x_{k+1}, \hat{x}_k) = 0$ and $\lim_{k\to+\infty} D_{\phi_2}(y_{k+1}, \hat{y}_k) = 0$. Moreover, one has $\lim_{k\to+\infty} \|z_{k+1} - \hat{z}_k\| = 0$.

(iii) Using (4.5) yields

$$\sum_{k=0}^{K} \left(D_{\phi_1}(x_{k+1}, \hat{x}_k) + D_{\phi_2}(y_{k+1}, \hat{y}_k)\right) \leq \frac{1}{\rho}\left(L(z_0) - L^*\right).$$

Hence, we obtain

$$\min_{0\leq k\leq K} \left\{D_{\phi_1}(x_{k+1}, \hat{x}_k) + D_{\phi_2}(y_{k+1}, \hat{y}_k)\right\} \leq \frac{1}{\rho(K+1)}\left(L(z_0) - L^*\right).$$

□

Lemma 4.2. *Let Assumption 3.1 and Assumption 4.1 hold. Let $\{z_k = (x_k, y_k)\}$ and $\{\hat{z}_k = (\hat{x}_k, \hat{y}_k)\}$ be the sequences generated by Algorithm 1. For any integer $k \geq 1$, set*

$$p_x^{k+1} = \nabla_x q(x_{k+1}, y_{k+1}) - \nabla_x q(x_{k+1}, \hat{y}_k) + q_x^{k+1}, \tag{4.6}$$

where $q_x^{k+1} = \nabla f(x_{k+1}) - \nabla f(\hat{x}_k) - \frac{1}{\tau_k}\left(\nabla\phi_1(x_{k+1}) - \nabla\phi_1(\hat{x}_k)\right)$. Then there exists $\rho > 0$ such that

$$\operatorname{dist}\left(0, \partial L(z_{k+1})\right) \leq \rho \left\|z_{k+1} - \widehat{z}_k\right\|. \tag{4.7}$$

Proof. From iterative scheme (3.2), we see that

$$x_{k+1} \in \arg\min_{x\in\mathbb{R}^n} \{Q(x, \hat{y}_k) + \langle \nabla f(\hat{x}_k), x\rangle + \frac{1}{\tau_k} D_{\phi_1}(x, \hat{x}_k)\}.$$

By Fermat's rule, we have

$$0 \in \partial_x Q(x_{k+1}, \hat{y}_k) + \nabla f(\hat{x}_k) + \frac{1}{\tau_k}\left(\nabla\phi_1(x_{k+1}) - \nabla\phi_1(\hat{x}_k)\right),$$

which implies that

$$\nabla f(x_{k+1}) - \nabla f(\hat{x}_k) - \frac{1}{\tau_k}\left(\nabla\phi_1(x_{k+1}) - \nabla\phi_1(\hat{x}_k)\right) \in \partial_x Q(x_{k+1}, \hat{y}_k) + \nabla f(x_{k+1}) = \partial_x L(x_{k+1}, \hat{y}_k). \tag{4.8}$$

Similarly, in view of the y-subproblem in iterative scheme (3.2), we have

$$0 \in \partial_y Q(x_{k+1}, y_{k+1}) + \nabla g(\hat{y}_k) + \frac{1}{\sigma_k}\left(\nabla\phi_2(y_{k+1}) - \nabla\phi_2(\hat{y}_k)\right),$$

which implies that

$$\nabla g(y_{k+1}) - \nabla g(\hat{y}_k) - \frac{1}{\sigma_k}\left(\nabla\phi_2(y_{k+1}) - \nabla\phi_2(\hat{y}_k)\right) \in \partial_y Q(x_{k+1}, y_{k+1}) + \nabla g(y_{k+1}) = \partial_y L(z_{k+1}). \tag{4.9}$$

From Assumption 4.1 (iii), we see that $\partial_x Q(x,y) = \nabla_x q(x,y) + \partial h(x)$. According to (4.8), we have

$$q_x^{k+1} \in \partial_x L(x_{k+1},\hat{y}_k) = \partial_x Q(x_{k+1},\hat{y}_k) + \nabla f(x_{k+1}) = \nabla_x q(x_{k+1},\hat{y}_k) + \partial h(x_{k+1}) + \nabla f(x_{k+1}).$$

It follows from (4.6) that

$$\begin{aligned} p_x^{k+1} &= \nabla_x q(x_{k+1},y_{k+1}) - \nabla_x q(x_{k+1},\hat{y}_k) + q_x^{k+1} \\ &\in \nabla_x q(x_{k+1},y_{k+1}) + \partial h(x_{k+1}) + \nabla f(x_{k+1}) \\ &= \partial_x Q(x_{k+1},y_{k+1}) + \nabla f(x_{k+1}) \\ &= \partial_x L(z_{k+1}). \end{aligned} \tag{4.10}$$

Hence,

$$(p_x^{k+1}, p_y^{k+1}) \in \partial_x L(z_{k+1}) \times \partial_y L(z_{k+1}) \subset \partial L(z_{k+1}).$$

Now, we estimate the norms of p_x^{k+1} and p_y^{k+1}. Under Assumption 4.1 (i) that L is coercive, we deduce that $\{z_{k+1}\}$ is a bounded set. In view of Assumption 3.1 (ii) and (3.1), we have

$$\begin{aligned} \left\|p_x^{k+1}\right\| &\le \|\nabla_x q(x_{k+1},y_{k+1}) - \nabla_x q(x_{k+1},\hat{y}_k)\| + \left\|q_x^{k+1}\right\| \\ &\le \xi \|y_{k+1} - \hat{y}_k\| + \|\nabla f(x_{k+1}) - \nabla f(\hat{x}_k)\| + \frac{1}{\tau_k}\|\nabla\phi_1(\hat{x}_k) - \nabla\phi_1(x_{k+1})\| \\ &\le \xi \|y_{k+1} - \hat{y}_k\| + L_{\phi_1} L_1 \|x_{k+1} - \hat{x}_k\| + \frac{L_1}{\tau_k}\|x_{k+1} - \hat{x}_k\| \\ &\le \xi \|y_{k+1} - \hat{y}_k\| + L_1\left(L_{\phi_1} + \frac{1}{\underline{\tau}}\right)\|x_{k+1} - \hat{x}_k\| \end{aligned}$$

and

$$\begin{aligned} \left\|p_y^{k+1}\right\| &\le \|\nabla g(y_{k+1}) - \nabla g(\hat{y}_k)\| + \frac{1}{\sigma_k}\|\nabla\phi_2(\hat{y}_k) - \nabla\phi_2(y_{k+1})\| \\ &\le L_{\phi_2} L_2 \|y_{k+1} - \hat{y}_k\| + \frac{L_2}{\sigma_k}\|y_{k+1} - \hat{y}_k\| \\ &\le L_2\left(L_{\phi_2} + \frac{1}{\underline{\sigma}}\right)\|y_{k+1} - \hat{y}_k\|, \end{aligned}$$

and hence

$$\begin{aligned} \left\|(p_x^{k+1}, p_y^{k+1})\right\| &\le \left\|p_x^{k+1}\right\| + \left\|p_y^{k+1}\right\| \\ &\le L_1\left(L_{\phi_1} + \frac{1}{\underline{\tau}}\right)\|x_{k+1} - \hat{x}_k\| + \left(\xi + L_2\left(L_{\phi_2} + \frac{1}{\underline{\sigma}}\right)\right)\|y_{k+1} - \hat{y}_k\| \\ &\le \rho\,\|z_{k+1} - \hat{z}_k\|, \end{aligned}$$

where $\rho = \sqrt{2}\max\{L_1\left(L_{\phi_1} + \frac{1}{\underline{\tau}}\right), \xi + L_2\left(L_{\phi_2} + \frac{1}{\underline{\sigma}}\right)\}$. □

Below, we summarize some properties about cluster points and prove every cluster point of a sequence generated by Algorithm 1 is the critical point of L. Let $\{z_k\}$ be the sequence generated by Algorithm 1 with initial point z_0. Under Assumption 4.1 (i) that L is coercive, we deduce

that $\{z_k\}$ is a bounded set, and it has at least one cluster point. The set of all cluster points is denoted by Ω, i.e.,

$$\Omega := \{\hat{z} = (\hat{x}, \hat{y}) \in \mathbb{R}^n \times \mathbb{R}^m : \exists \text{ strictly increasing } \{k_j\}_{j\in\mathbb{N}} \text{ such that } z_{k_j} \to \hat{z}, j \to \infty\}.$$

Lemma 4.3. *Let Assumption 3.1 and Assumption 4.1 hold. Let $\{z_k\}$ be a sequence generated by Algorithm 1 with initial point z_0. Then the following results hold.*

(i) *Ω is a nonempty compact set, and L is finite and constant on Ω;*

(ii) $\Omega \subset \operatorname{crit} L$;

(iii) $\lim_{k\to\infty} \operatorname{dist}(z_k, \Omega) = 0$.

Proof. (i) The fact that $\{z_k\}$ is bounded yields the nonemptyness of Ω. In addition, Ω can be reformulated as an intersection of compact sets $\Omega = \bigcap_{s\in\mathbb{N}} \overline{\bigcup_{k\geq s} z_k}$, which illustrates that Ω is a compact set.

For any $\hat{z} = (\hat{x}, \hat{y}) \in \Omega$, there exists a subsequence $\{z_{k_j}\}$ such that $\lim_{j\to\infty} z_{k_j} = \hat{z}$. Since L is continuous, we have $\lim_{j\to\infty} L(z_{k_j}) = L(\hat{z})$. According to Lemma 4.1, we see that $\{L(z_k)\}$ converges to L^* globally. Hence

$$\lim_{j\to\infty} L(z_{k_j}) = \lim_{k\to\infty} L(z_k) = L(\hat{z}) = L^*. \tag{4.11}$$

which means L is a constant on Ω.

(ii) Letting $\hat{z} \in \Omega$, one sees that $\exists\ z_{k_j}$ such that $z_{k_j} \to \hat{z}$. According to Lemma 4.1 (ii), one obtain $\lim_{j\to\infty} \|z_{k_j} - \hat{z}_{k_{j-1}}\| = 0$, and hence, $\lim_{j\to\infty} z_{k_j} = \lim_{j\to\infty} \hat{z}_{k_{j-1}} = \hat{z}$. By (4.7), one arrives at

$$\left\|(p_x^{k_j}, p_y^{k_j})\right\| \leq \rho \left\|z_{k_j} - \hat{z}_{k_{j-1}}\right\|.$$

Thus $(p_x^{k_j}, p_y^{k_j}) \to (0,0)$ as $j \to \infty$. Based on the result $\lim_{j\to\infty} L(z_{k_j}) = L(\hat{z})$, $(p_x^{k_j}, p_y^{k_j}) \in \partial L(z_{k_j})$ and the closedness property of ∂L, we conclude that $(0,0) \in \partial L(\hat{z})$, which means $\hat{z} = (\hat{x}, \hat{y})$ is a critical point of L, and $\Omega \subset \operatorname{crit} L$.

(iii) We prove the assertion by contradiction. Assume that $\lim_{k\to\infty} \operatorname{dist}(z_k, \Omega) \neq 0$. Then, there exists a subsequence $\{z_{k_m}\}$ and a constant $M > 0$ such that

$$\|z_{k_m} - \hat{z}\| \geq \operatorname{dist}(z_{k_m}, \Omega) > M, \quad \forall \hat{z} \in \Omega.$$

On the other hand, $\{z_{k_m}\}$ is bounded and has a subsequence $\{z_{k_{m_j}}\}$ converging to a point in Ω. Thus, $\lim_{j\to\infty} \operatorname{dist}(z_{k_{m_j}}, \Omega) = 0$, which is a contradiction to (4.12). □

Now, we can prove the main convergence results of proposed algorithms under the KŁ property.

Theorem 4.1. *Let Assumptions 3.1 hold, and let $\{z_k = (x_k, y_k)\}$ and $\{\hat{z}_k = (\hat{x}_k, \hat{y}_k)\}$ be the bounded sequences generated by Algorithm 1 with initial point z_0. Assume that L is KŁ function. Then the following results hold:*

(i) *$\{z_k\}$ has finite length, i.e.,*

$$\sum_{k=1}^{\infty} \|z_{k+1} - z_k\| < +\infty, \tag{4.12}$$

(ii) *$\{z_k\}$ converges to a critical point of L.*

Proof. In the process of our proof, we always assume $L(z_k) \neq L(z^*)$ for $z^* \in \Omega$. Otherwise, there exists an integer $\hat{k}$ such that $L(z_{\hat{k}}) = L(z^*)$, so $L(z_k) \equiv L(z^*)$ for $k \geq \widehat{k}$. For $k \geq \widehat{k}$, by (4.1), we have

$$\rho \|z_{k+1} - \hat{z}_k\|^2 \leq L(z_k) - L(z_{k+1}) \leq L(z_{\widehat{k}}) - L(z^*),$$

where the last inequality follows from the nonincreasing of $L(\cdot)$. Therefore, for any $k \geq \widehat{k}$, we have $z_{k+1} = z_k$ and the assertions (4.12) holds trivially.

(i) Since $\{L(z_k)\}$ is a nonincreasing sequence, we assume that $L(z_k) > L(z^*)$ for all $k \geq 1$. From (4.11), we find $\lim_{k\to\infty} L(z_k) = L(z^*)$. For any $\eta > 0$, there exists a positive integer k_0 such that $L(z^*) < L(z_k) < L(z^*) + \eta$ for all $k > k_0$. From Lemma 4.3 (iii), we have that, for any $\varepsilon > 0$, there exists a positive integer k_1 such that $\operatorname{dist}(z_k, \Omega) < \varepsilon$ for all $k > k_1$. Consequently, for any $\eta, \varepsilon > 0$, there exists a positive integer $l = \max\{k_0, k_1\}$ such that

$$\operatorname{dist}(z_k, \Omega) < \varepsilon \ \text{ and } \ L(z^*) < L(z_k) < L(z^*) + \eta$$

for $k > l$. Since Ω is a nonempty and compact set, and $L(\cdot)$ is a constant on Ω, we can apply the Lemma 2.2 with $\Psi = \Omega$. Therefore, there exists a concave function $\varphi \in \Phi_\eta$ such that

$$\varphi'(L(z_k) - L(z^*)) \operatorname{dist}(0, \partial L(z_k)) \geq 1, \quad \forall k > l. \tag{4.13}$$

From Lemma 4.2, we obtain that

$$\operatorname{dist}(0, \partial L(z_k)) \leq \left\| (p_x^k, p_y^k) \right\| \leq \rho \|z_k - \hat{z}_{k-1}\|. \tag{4.14}$$

Substituting (4.14) into (4.13), we conclude

$$\varphi'(L(z_k) - L(z^*)) \geq \frac{1}{\operatorname{dist}(0, L(z_k))} \geq \frac{1}{\rho \|z_k - \hat{z}_{k-1}\|}.$$

Now, we define $\triangle_{p,q} = \varphi(L(z_p) - L(z^*)) - \varphi(L(z_q) - L(z^*))$. From the concavity of φ, we have

$$\varphi(L(z_{k+1}) - L(z^*)) \leq \varphi(L(z_k) - L(z^*)) + \varphi'(L(z_k) - L(z^*))(L(z_{k+1}) - L(z_k)).$$

According to the strongly convexity of ϕ_1 and ϕ_2 and (4.1), we have

$$L(z_k) - L(z_{k+1}) \geq \rho \left[D_{\phi_1}(x_{k+1}, \hat{x}_k) + D_{\phi_2}(y_{k+1}, \hat{y}_k) \right] \geq \overline{\rho} \|z_{k+1} - \hat{z}_k\|^2,$$

where $\overline{\rho} = \rho \min\left\{ \frac{\mu_{\phi_1}}{2}, \frac{\mu_{\phi_2}}{2} \right\}$. So (5.1) is equivalent to the following inequality

$$\triangle_{k,k+1} \geq \varphi'(L(z_k) - L(z^*))(L(z_{k+1}) - L(z_k)) \geq \frac{\overline{\rho} \|z_{k+1} - \hat{z}_k\|^2}{\rho \|z_k - \hat{z}_{k-1}\|}. \tag{4.15}$$

Let $C = \frac{\rho}{\overline{\rho}}$. Then (4.15) can be simplified as $\|z_{k+1} - \hat{z}_k\|^2 \leq C \triangle_{k,k+1} \|z_k - \hat{z}_{k-1}\|$. Using the fact that $2\sqrt{ab} \leq a + b$ for $a, b \geq 0$, we infer

$$2\|z_{k+1} - \hat{z}_k\| \leq C \triangle_{k,k+1} + \|z_k - \hat{z}_{k-1}\|. \tag{4.16}$$

Summing up (4.16) for $k = l+1, \ldots, K$ yields

$$2 \sum_{k=l+1}^{K} \|z_{k+1} - \hat{z}_k\| \leq C \triangle_{l+1,K+1} + \|z_{l+1} - \hat{z}_l\| - \|z_{K+1} - \hat{z}_K\| + \sum_{k=l+1}^{K} \|z_{k+1} - \hat{z}_k\|.$$

Eliminating the same terms of the inequality, we have

$$\sum_{k=l+1}^{K} \|z_{k+1} - \hat{z}_k\| \leq C \triangle_{l+1,K+1} + \|z_{l+1} - \hat{z}_l\| - \|z_{K+1} - \hat{z}_K\| < \infty.$$

Letting $K \to \infty$, we see that $\sum_{k=l+1}^{K} \|z_{k+1} - \hat{z}_k\| < \infty$. Note that $z_{k+1} - \hat{z}_k = (x_{k+1} - \hat{x}_k, y_{k+1} - \hat{y}_k)$ and investigate the iterative point $\hat{z}_k = (\hat{x}_k, \hat{y}_k)$ in Algorithm 1. If $\hat{z}_k$ is generated by (3.3), then $z_{k+1} - \hat{z}_k = z_{k+1} - z_k - \alpha_k(z_k - z_{k-1}) - \beta_k(z_{k-1} - z_{k-2})$. If $\hat{z}_k$ is generated by (3.4), then $z_{k+1} - \hat{z}_k = (x_{k+1} - x_k, y_{k+1} - y_k) = z_{k+1} - z_k$. No matter how $\hat{z}_k$ is generated, we always have

$$\|z_{k+1} - \hat{z}_k\| \geq \|z_{k+1} - z_k\| - \alpha_k \|z_k - z_{k-1}\| - \beta_k \|z_{k-1} - z_{k-2}\|. \tag{4.17}$$

Summing up (4.17) for $k = l+1, \ldots, K$ yields

$$\sum_{k=l+1}^{K} \|z_{k+1} - z_k\| - \sum_{k=l+1}^{K} \alpha_k \|z_k - z_{k-1}\| - \sum_{k=l+1}^{K} \beta_k \|z_{k-1} - z_{k-2}\| \leq \sum_{k=l+1}^{K} \|z_{k+1} - \hat{z}_k\| < \infty. \tag{4.18}$$

Note that $\alpha_k \in [0, \alpha_{\max}], \beta_k \in [0, \beta_{\max}]$. Letting $\overline{\alpha} = \sup_k \{\alpha_k\}, \overline{\beta} = \sup_k \{\beta_k\}$, one has $0 \leq \overline{\alpha} + \overline{\beta} \leq \alpha_{\max} + \beta_{\max} < 1$, and

$$\begin{aligned} &\sum_{k=l+1}^{K} \|z_{k+1} - z_k\| - \overline{\alpha} \sum_{k=l+1}^{K} \|z_k - z_{k-1}\| - \overline{\beta} \sum_{k=l+1}^{K} \|z_{k-1} - z_{k-2}\| \\ &\leq \sum_{k=l+1}^{K} \|z_{k+1} - z_k\| - \sum_{k=l+1}^{K} \alpha_k \|z_k - z_{k-1}\| - \sum_{k=l+1}^{K} \beta_k \|z_{k-1} - z_{k-2}\|, \end{aligned} \tag{4.19}$$

and

$$\begin{aligned} &\sum_{k=l+1}^{K} \|z_{k+1} - z_k\| - \overline{\alpha} \sum_{k=l+1}^{K} \|z_k - z_{k-1}\| - \overline{\beta} \sum_{k=l+1}^{K} \|z_{k-1} - z_{k-2}\| \\ &= (1 - \overline{\alpha} - \overline{\beta}) \sum_{k=l+1}^{K} \|z_{k+1} - z_k\| - (\overline{\alpha} + \overline{\beta})(\|z_{l+1} - z_l\| - \|z_{K+1} - z_K\|) \\ &\quad - \overline{\beta}(\|z_l - z_{l-1}\| - \|z_K - z_{K-1}\|). \end{aligned} \tag{4.20}$$

Combining (4.18), (4.19), and (4.20), we have

$$\begin{aligned} &(1 - \overline{\alpha} - \overline{\beta}) \sum_{k=l+1}^{K} \|z_{k+1} - z_k\| \\ \leq &(\overline{\alpha} + \overline{\beta})(\|z_{l+1} - z_l\| - \|z_{K+1} - z_K\|) + \overline{\beta}(\|z_l - z_{l-1}\| - \|z_K - z_{K-1}\|) + \sum_{k=l+1}^{K} \|z_{k+1} - z_k\| \\ &- \sum_{k=l+1}^{K} \alpha_k \|z_k - z_{k-1}\| - \sum_{k=l+1}^{K} \beta_k \|z_{k-1} - z_{k-2}\| \\ \leq &(\overline{\alpha} + \overline{\beta})(\|z_{l+1} - z_l\| - \|z_{K+1} - z_K\|) + \overline{\beta}(\|z_l - z_{l-1}\| - \|z_K - z_{K-1}\|) + \sum_{k=l+1}^{K} \|z_{k+1} - \hat{z}_k\| \\ < &\infty. \end{aligned}$$

Taking the limit as $K \to \infty$ and using the fact $\overline{\alpha} + \overline{\beta} < 1$, we obtain $\sum_{k=l+1}^{\infty} \|z_{k+1} - z_k\| < \infty$, which shows that $\sum_{k=0}^{\infty} \|z_{k+1} - z_k\| < \infty$.

(ii) For any $m > n$, we have

$$\|z_m - z_n\| = \left\| \sum_{k=n}^{m-1} (z_{k+1} - z_k) \right\| \leq \sum_{k=n}^{m-1} \|z_{k+1} - z_k\| < \sum_{k=n}^{\infty} \|z_{k+1} - z_k\|.$$

Note that $\|z_m - z_n\| \to 0$, which means that $\{z_k\}$ is a Cauchy sequence. Hence $\{z_k\}$ is a convergent sequence. We also know that $\{z_k\}$ converges to a critical point of L from Lemma 4.3(ii). □

Since the KŁ property is also a very useful tool in establishing the convergence rate of first-order methods. Based on the KŁ inequality, Attouch and Bolte [42] first established convergence rate results which are related to the desingularizing function for proximal algorithms. Similar to the derivation process of [42], we can obtain convergence rate results as following.

Theorem 4.2. *(Convergence rate) Let Assumption 3.1 and 4.1 hold and let $\{z_k\}$ be a sequence generated by Algorithm 1 with $z_0 = (x_0, y_0)$ as initial point. Assume that L is a KŁ function and the desingularizing function has the form of $\varphi(t) = \frac{C}{\theta}t^\theta$ with $\theta \in (0,1]$, $C > 0$. Let $L^* = L(z)$ for all $z \in \mathscr{L}(z_0)$. Then the following assertions hold.*

(i) *If $\theta = 1$, then Algorithm 1 terminates in finite steps.*

(ii) *If $\theta \in [\frac{1}{2}, 1)$, then there exist $\omega > 0$ and $k_0 \in \mathbb{N}$ such that*

$$L(z_k) - L^* \leq \mathscr{O}\left(\exp\left(-\frac{\omega}{\rho}\right)\right), \quad \forall k > k_0.$$

(iii) *If $\theta \in (0, \frac{1}{2})$, then there exist $\omega > 0$ and $k_0 \in \mathbb{N}$ such that*

$$L(z_k) - L^* \leq \mathscr{O}\left(\left(\frac{k-k_0}{\rho}\right)^{\frac{-1}{1-2\theta}}\right), \quad \forall k > k_0.$$

The result is almost the same as it was mentioned in [40, 41]. We omit the proof here.

5. Numerical Experiments

We consider the Poisson linear inverse problems [1, 23], which can be conveniently described as follows. Given a matrix $A \in \mathbb{R}_+^{m\times n}$ modeling the experimental protocol, and $b \in \mathbb{R}_+^m$, the vector of measurement, the goal is to reconstruct the signal or image $x \in \mathbb{R}_+^n$ from the noisy measurement b such that $Ax \simeq b$. Moreover, since the dimension of x is often much larger than the number of observations, there is a need to regularize the problem through an appropriate choice of a regularizer freflecting desired features of the solution. Thus, given some adequate convex proximity measure $d(\cdot,\cdot)$ that quantifies the "error" between b and Ax, the task of recovering x can be represented as a minimize problem like

$$\min_x \left\{ d(b, Ax) + \lambda h(x) : x \in \mathbb{R}_+^n \right\}, \tag{5.1}$$

where $\lambda > 0$ plays the role of a regularizing parameter controlling the trade-off between matching the data fidelity criteria and the weight given to its regularizer.

By introducing an auxiliary variable $y \in \mathbb{R}_+^n$, we can solve (5.1) approximately according to the following optimization problem

$$\min_{x\in\mathbb{R}_+^n,\ y\in\mathbb{R}^n} \quad f(x) + Q(x,y) + g(y) \tag{5.2}$$

by defining

$$f(x) = d(b, Ax),\ g(y) = \lambda h(y),\ Q(x,y) = \frac{\mu}{2}\|x - y\|^2.$$

where λ is a positive penalization parameter.

Based on model (5.2), we give two choices for the first component $f(x)$ in the objective function:

(i) Burg's entropy: $d(b,Ax)=\sum_{i=1}^{m}\{(Ax)_i-b_i\log(Ax)_i\}$.

It is easy to find that $f(x)=d(b,Ax)$ has no globally Lipschitz continuous gradient [1], but satisfies GL-samd condition with a generalized Bregman function called Burg's entropy, denoted as

$$\varphi_1(x)=-\sum_{j=1}^{n}\log x_j,\ \mathrm{dom}\phi_1=\mathbb{R}^n_{++},$$

so the Bregman distance is now given by

$$D_{\varphi_1}(x,z)=\sum_{j=1}^{n}\left\{\frac{x_j}{z_j}-\log\left(\frac{x_j}{z_j}\right)-1\right\},$$

and for any L_{φ_1} satisfying $L_{\varphi_1}\geq\|b\|_1=\sum_{i=1}^{m}b_i$, the function $L_{\varphi_1}\varphi_1-f$ is convex on $\mathbb{R}^n_+$.

(ii) Boltzmann-Shannon entropy: $d(b,Ax)=\sum_{i=1}^{m}\{(Ax)_i\log(Ax)_i-(\log b_i+1)(Ax)_i+b_i\}$.

In this case, $f(x)=d(b,Ax)$ also has no globally Lipschitz continuous gradient [1], but satisfies GL-samd condition with a generalized Bregman function called Boltzmann-Shannon entropy, denoted as

$$\varphi_2(x)=\sum_{j=1}^{n}x_j\log x_j,\ \mathrm{dom}\phi_1=\mathbb{R}^n_{++},$$

so the Bregman distance is now given by

$$D_{\varphi_2}(x,z)=\sum_{j=1}^{n}\left\{x_j\log\left(\frac{x_j}{z_j}\right)+z_j-x_j\right\},$$

and for any L_{φ_2} satisfying $L_{\varphi_2}\geq\max_{1\leq j\leq n}\sum_{i=1}^{m}a_{ij}$, the function $L_{\varphi_2}\varphi_2-f$ is convex on $\mathbb{R}^n_+$.

We consider Tikhonov regularization in model (5.2), i.e., the regularizer is $g(y)=\lambda h(y)=\frac{\lambda}{2}\|y\|^2$. Take Energy $\phi_2(y)=\frac{1}{2}\|y\|^2$, which corresponding Bregman distance is $D_{\phi_2}(y,z)=\frac{1}{2}\|y-z\|^2$, and for any L_{ϕ_2} satisfying $L_{\phi_2}\geq\lambda$, the function $L_{\phi_2}\phi_2-g$ is convex on $\mathbb{R}^n$.

In numerical experiments, we set $A=D+D^T\in\mathbb{R}^{n\times n}$, where D is a matrix generated by i.i.d. standard Gaussian entries. The vector b is also generated by i.i.d. standard Gaussian entries. The parameters of the problem are set as $\lambda=1$, $L_{\varphi_1}=\|b\|_1$, $L_{\varphi_2}=\max_{1\leq j\leq n}\sum_{i=1}^{m}a_{ij}$, $L_{\phi_2}=\lambda$ and the penalization parameter μ as large as the conditions permit, then we can set $\sigma=\frac{1}{2\lambda}$ and $\tau=\frac{1}{2L_{\varphi_i}}(i=1,2)$. We take $m=n=500$ and select the starting point randomly, and use

$$\mathrm{Error}=\frac{\|x_k-x_{k-1}\|}{\max\{1,\|x_k\|\}}\leq 10^{-6}$$

as the stopping criteria. In the numerical results, "Iter." denotes the number of iterations, "Time" denotes the CPU time, "Extrapolation" records the number of taking extrapolation step, i.e., the number of adopting (3.3). In order to show the effectiveness of the proposed algorithms, we compare Algorithm 1, Algorithm 2 with ASABP [23], IASABP [23] and aASAP [33] for different Bregman distance. Note that, when $\alpha_k\equiv\beta_k\equiv 0$, Algorithm 1 and Algorithm 2 correspond to ASABP. The main parameters in IASABP are set as follows: η is a random number between 0.90 and 0.95 and $\theta=0.99$; $\alpha_k^0=\beta_k^0=1$ for all k. For aASAP, we take $\alpha_k=0.3$. For Algorithm 1, we set $\alpha_k=0.3$ and $\beta_k=0.2$. And we also take extrapolation parameter dynamically updating with $\alpha_k=\beta_k=\frac{k-1}{k+2}$. Even if the theoretical bound of extrapolation parameter

$\alpha_k+\beta_k$ with dynamically updating dost not permit to go beyond 1, for the convergence is also obtained for this case with a better performance. For Algorithm 2, we set $\alpha_0=0.3,\beta_0=0.2$ as the initial extrapolation parameter and $t=1.2,\beta_{\max}=0.499$. We use "Alg. 1-i" and "Alg. 2-i" to denote Algorithm 1 and Algorithm 2 with $\phi_1(x)=\varphi_i(x)(1\leq i\leq 2)$, respectively, where extrapolation parameter $\alpha_k=0.3,\beta_k=0.2$. We use "Alg. 1-i(F)" and "Alg. 2-i(F)" to denote Algorithm 1 and Algorithm 2 with $\phi_1(x)=\varphi_i(x)(1\leq i\leq 2)$, respectively, where extrapolation parameter $\alpha_k=\beta_k=\frac{k-1}{k+2}$.

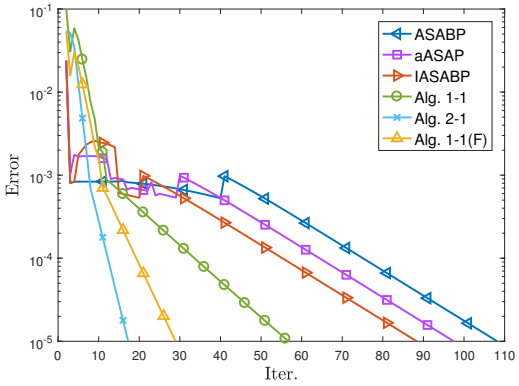
In Table 1, we list the iterations, CPU time and extrapolation step of the above algorithm for different Bregman distance. In Table 2, we report on more results for comparing the above algorithms for $m=200$ and $n=1000$, the corresponding graphical results are displayed in Figure 2. In Figure 1, (a) and (b) reports the result of different extrapolation parameter, respectively, (c) reports the result of different Bregman distance. It can be seen that the Burg's entropy have computational advantage than the Boltzmann-Shannon entropy for Algorithm 1 and Algorithm 2 in terms of number of iteration and CPU time. Compared with one-step extrapolation and original algorithm, two-step extrapolation performs much better. It shows that Algorithm 2 with adaptive extrapolation parameters performs the best among all algorithms.

TABLE 1. Numerical results of different Bregman distance with different extrapolation parameter ($m=n=500$)

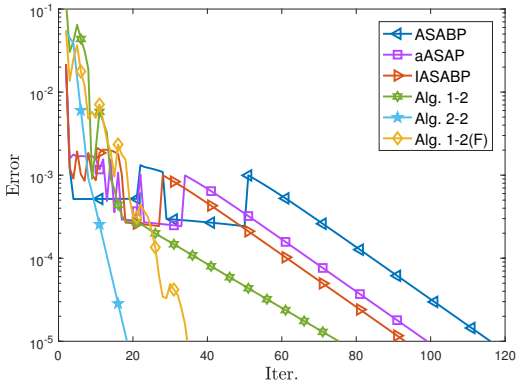
Burg's entropy				Boltzmann-Shannon entropy			
Algorithm	Iter.	Time(s)	Extrapolation	Algorithm	Iter.	Time(s)	Extrapolation
ASABP	109	0.2231	108	ASABP	117	0.2725	116
aASAP	98	0.1563	97	aASAP	100	0.1706	99
IASABP	89	0.1394	88	IASABP	94	0.1438	93
Alg. 1-1	58	0.0898	55	Alg. 1-2	76	0.1072	72
Alg. 2-1	18	0.0129	16	Alg. 2-2	19	0.0136	17
Alg. 1-1(F)	29	0.0156	24	Alg. 1-2(F)	35	0.0234	29

TABLE 2. Numerical results of different Bregman distance with different extrapolation parameter ($m=200,\ n=1000$)

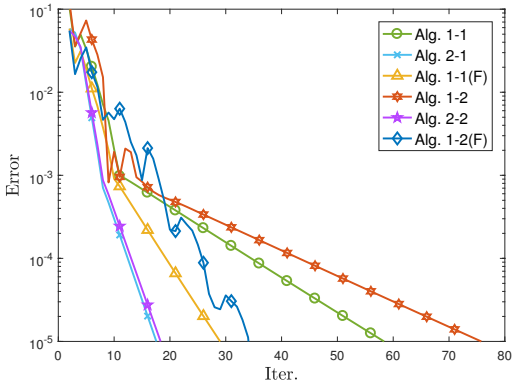
Burg's entropy				Boltzmann-Shannon entropy			
Algorithm	Iter.	Time(s)	Extrapolation	Algorithm	Iter.	Time(s)	Extrapolation
ASABP	120	0.3975	119	ASABP	129	0.3827	128
aASAP	105	0.2013	104	aASAP	113	0.2236	112
IASABP	98	0.1892	97	IASABP	116	0.2578	115
Alg. 1-1	76	0.1196	73	Alg. 1-2	93	0.1508	88
Alg. 2-1	25	0.0172	22	Alg. 2-2	37	0.0278	35
Alg. 1-1(F)	45	0.0632	40	Alg. 1-2(F)	51	0.0713	45



(A) Burg's entropy



(B) Boltzmann-Shannon entropy

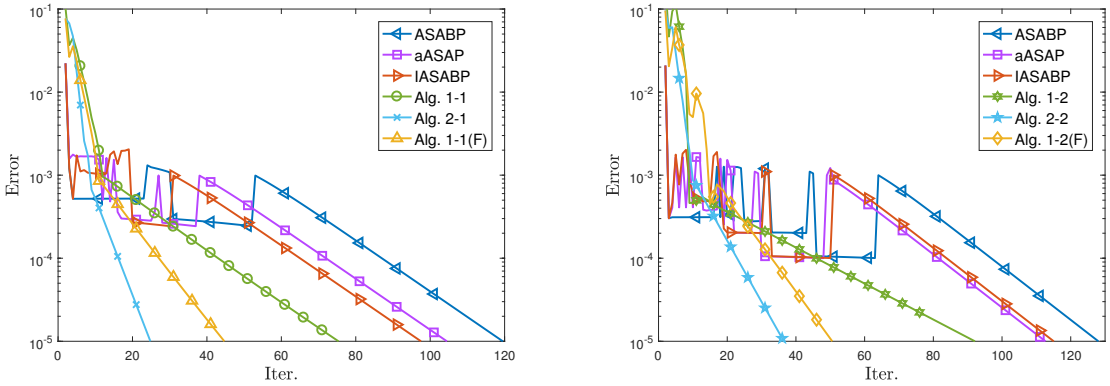


(C) Burg's entropy and Boltzmann-Shannon entropy

FIGURE 1. The value of $\frac{\|x_k - x_{k-1}\|}{\max\{1, \|x_k\|\}}$ versus the iteration numbers for different Bregman distance with different extrapolation parameter ($m = n = 500$).

6. CONCLUSION

In this paper, we introduced a two-step inertial Bregman alternating structure-adapted proximal gradient descent algorithm for solving a nonconvex and nonsmooth nonseparable optimization problem. Under some assumptions, we proved that our algorithm is a descent method in sense of objective function values, and every cluster point is a critical point of the objective function. The convergence of the proposed algorithm is proved by assuming that the underlying function satisfies the Kurdyka–Łojasiewicz property yet without the Lipschitz smoothness, Furthermore, if the desingularizing function has the special form, we also established the linear and sub-linear convergence rates of the function value sequence generated by the algorithm. In numerical experiments, based on different Bregman distance, we investigated Poisson linear inverse problems. Numerical results are reported to support the effectiveness of the proposed algorithm.



(A) Burg's entropy (B) Boltzmann-Shannon entropy

FIGURE 2. The value of $\frac{\|x_k - x_{k-1}\|}{\max\{1, \|x_k\|\}}$ versus the iteration numbers for different Bregman distance with different extrapolation parameter ($m = 200,\ n = 1000$).

REFERENCES

[1] H. Bauschke, J. Bolte, M. Teboulle, A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications, Math. Oper. Res. 42 (2016), 330-348.

[2] M. Bertero, P. Boccacci, G. Desiderà, G. Vicidomini, Image deblurring with Poisson data: from cells to galaxies, Inverse Probl. 25 (2009), 123006.

[3] M. Mukkamala, P. Ochs, T. Pock, S. Sabach, Convex-concave backtracking for inertial Bregman proximal gradient algorithms in nonconvex optimization, SIAM J. Math. Data Sci. 2 (2020), 658-682.

[4] M. Nikolova, M.K. Ng, S. Zhang, W.K. Ching, Efficient reconstruction of piecewise constant images using nonsmooth nonconvex minimization, SIAM J. Imaging Sci. 1 (2008), 2-25.

[5] S. Gu, L. Zhang, W. Zuo, X. Feng, Weighted nuclear norm minimization with application to image denoising, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2862-2869, 2014.

[6] W. Bian, X. Chen, Linearly constrained non-Lipschitz optimization for image restoration, SIAM J. Imaging Sci. 8 (2015), 2294-2322.

[7] P. Paatero, U. Tapper, Positive matrix factorization: a nonnegative factor model with optimal utilization of error estimates of data values, Environmetrics 5 (1994), 111-126.

[8] D.D. Lee, H.S. Seung, Learning the parts of objects by nonnegative matrix factorization, Nature 401 (1999), 788-791.

[9] X. Zhang, X. Zhang, X. Li, Z. Li, S. Wang, Classify social image by integrating multi-modal content, Multimed. Tools. Appl. 77 (2018), 7469-7485.

[10] D.P. Bertsekas, Nonlinear programming, J. Oper. Res. Soc. 48 (1977), 334-334.

[11] A. Beck, L. Tetruashvili, On the convergence of block coordinate descent type methods, SIAM J. Optim. 23 (2013), 2037-2060.

[12] A. Auslender, Asymptotic properties of the Fenchel dual functional and applications to decomposition problems, J. Optim. Theory Appl. 73 (1992), 427-449.

[13] H. Attouch, J. Bolte, P. Redont, A. Soubeyran, Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka–Łojasiewicz inequality, Math. Oper. Res. 35 (2010), 438-457.

[14] H. Attouch, P. Redont, A. Soubeyran, A new class of alternating proximal minimization algorithms with costs-to-move, SIAM J. Optim. 118 (2007), 1061-1081.

[15] Y. Xu, W. Yin, A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion, SIAM J. Imaging Sci. 6 (2013), 1758-1789.

[16] X. Qin, A weakly convergent method for splitting problems with nonexnpansive mappings, J. Nonlinear Convex Anal. 24 (2023), 1033-1043.
[17] X. Zuo, S. Osher and W. Li, Primal-dual damping algorithms for optimization, Ann. Math. Sci. Appl. 9 (2024), 467–504.
[18] J. Bolte, S. Sabach, M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, Math. Program. 146 (2014), 459-494.
[19] M. Nikolova, P. Tan, Alternating structure-adapted proximal gradient descent for nonconvex block-regularised problems, SIAM J. Optim. 29 (2019), 2053-2078.
[20] H. Bauschke, J. Bolte, J. Chen, M. Teboulle, X. Wang, On linear convergence of non-Euclidean gradient methods without strong convexity and Lipschitz gradient continuity, J. Optim. Theory Appl. 182 (2019), 1068-1087.
[21] J. Bolte, S. Sabach, M. Teboulle, Y. Vaisbourd, First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems, SIAM J. Optim. 28 (2018), 2131-2151.
[22] H. Lu, R. Freund, Y. Nesterov, Relatively smooth convex optimization by first-order methods, and applications, SIAM J. Optim. 28 (2018), 333-354.
[23] X. Gao, X. Cai, X. Wang, D. Han, An alternating structure-adapted Bregman proximal gradient descent algorithm for constrained nonconvex nonsmooth optimization problems and its inertial variant, J. Glob. Optim. 87 (2023), 277-300.
[24] L. Liu, S.Y. Cho, A Bregman projection algorithm with self adaptive step sizes for split variational inequality problems involving non-Lipschitz operators, J. Nonlinear Var. Anal. 8 (2024), 396-417.
[25] B. Tan, X. Qin, S.Y. Cho, Revisiting subgradient extragradient methods for solving variational inequalities, Numer. Algo. 90 (2022), 1593-1615.
[26] R. I. Boţ, E.R. Csetnek, An inertial Tseng's type proximal algorithm for nonsmooth and nonconvex optimization problems, J. Optim. Theory Appl. 171 (2016), 600-616.
[27] B. Tan, X. Qin, On relaxed inertial projection and contraction algorithms for solving monotone inclusion problems, Adv. Comput. Math. 50 (2024), 59.
[28] X. Qin, An inertial Krasnosel'skiĭ-Mann iterative algorithm for accretive and nonexnpansive mappings, J. Nonlinear Convex Anal. 26 (2025), 407-414.
[29] B.T. Polyak, Some methods of speeding up the convergence of iteration methods, USSR Comput. Math. Math. Phys. 4 (1964), 1-17.
[30] F. Alvarez, H. Attouch, An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping, Set-Valued Anal. 9 (2001), 3-11.
[31] T. Pock, S. Sabach, Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems, SIAM J. Imaging Sci. 9 (2017), 1756-1787.
[32] X. Gao, X. Cai, D. Han , A Gauss-Seidel type inertial proximal alternating linearized minimization for a class of nonconvex optimization problems, J. Glob. Optim. 76 (2020), 863-887.
[33] X. Yang, L. Xu, Some accelerated alternating proximal gradient algorithms for a class of nonconvex nonsmooth problems, J. Glob. Optim. 87 (2023), 939-964.
[34] J. Zhao, Q.L. Dong, T.R. Michael, F. Wang, Two-step inertial Bregman alternating minimization algorithm for nonconvex and nonsmooth problems, J. Glob. Optim. 84 (2022), 941-966.
[35] M. Chao, F. Nong, M. Zhao, An inertial alternating minimization with Bregman distance for a class of nonconvex and nonsmooth problems, J. Appl. Math. Comput. 69 (2023), 1559-1581.
[36] B. Mordukhovich, Variational Analysis and Generalized Differentiation, I: Basic Theory. Grundlehren der Mathematischen Wissenschaften, Vol. 330. Springer-Verlag, Berlin, 2006.
[37] R.T. Rockafellar, J.B. Wets, Variational Analysis, Springer, New York, 1998.
[38] F. Wang, W. Cao, Z. Xu, Convergence of multi-block Bregman ADMM for nonconvex composite problems, Sci. China Inf. Sci. 61 (2018), 122101.
[39] D.P. Bertsekas, J.N. Tsitsiklis, Parallel and Distributed Computation: Numerical Methods, Prentice hall, Englewood Cliffs, NJ, 1989.
[40] H. Li, Z. Lin, Accelerated proximal gradient methods for nonconvex programming, In: Advances in Neural Information Processing Systems, pp. 379-387, 2015.

[41] Q. Li, Y. Zhou, Y. Liang, P.K. Varshney, Convergence analysis of proximal gradient with momentum for nonconvex optimization. In: Proceedings of the 34th International Conference on Machine Learning, pp. 2111-2119, 2017.

[42] H. Attouch, J. Bolte, On the convergence of the proximal algorithm for nonsmooth functions involving analytic features, Math. Program. Ser. B, 116 (2009), 5-16.