Computational Optimization and Applications (2025) 90:227–256
https://doi.org/10.1007/s10589-024-00618-z

Approximate bregman proximal gradient algorithm for relatively smooth nonconvex optimization

Shota Takahashi[1] · **Akiko Takeda**[1,2]

Received: 11 December 2023 / Accepted: 9 October 2024 / Published online: 18 November 2024
© The Author(s) 2024

Abstract

In this paper, we propose the approximate Bregman proximal gradient algorithm (ABPG) for solving composite nonconvex optimization problems. ABPG employs a new distance that approximates the Bregman distance, making the subproblem of ABPG simpler to solve compared to existing Bregman-type algorithms. The subproblem of ABPG is often expressed in a closed form. Similarly to existing Bregman-type algorithms, ABPG does not require the global Lipschitz continuity for the gradient of the smooth part. Instead, assuming the smooth adaptable property, we establish the global subsequential convergence under standard assumptions. Additionally, assuming that the Kurdyka–Łojasiewicz property holds, we prove the global convergence for a special case. Our numerical experiments on the ℓ_p regularized least squares problem, the ℓ_p loss problem, and the nonnegative linear system show that ABPG outperforms existing algorithms especially when the gradient of the smooth part is not globally Lipschitz or even locally Lipschitz continuous.

Keywords Composite nonconvex nonsmooth optimization · Bregman proximal gradient algorithms · Kurdyka–Łojasiewicz property · ℓ_p regularized least squares problem · ℓ_p loss problem · nonnegative linear system · Kullback–Leibler divergence

Mathematics Subject Classification 90C26 · 49M37 · 65K05

✉ Shota Takahashi
shota@mist.i.u-tokyo.ac.jp

Akiko Takeda
takeda@mist.i.u-tokyo.ac.jp

1 Graduate School of Information Science and Technology, The University of Tokyo, 7–3–1 Hongo, Bunkyo-ku 113-8656, Tokyo, Japan

2 Center for Advanced Intelligence Project, RIKEN, 1–4–1 Nihonbashi, Chuo-ku 103-0027, Tokyo, Japan

1 Introduction

In applications to machine learning and signal processing, regularization is used to avoid over-fitting or to impose the model structure. An optimization problem with a regularization term can be formulated as a composite nonconvex optimization problem:

$$\min_{\boldsymbol{x}\in \mathrm{cl}\, C} \quad \Psi(\boldsymbol{x}) := f(\boldsymbol{x}) + g(\boldsymbol{x}), \tag{1.1}$$

where $f : \mathbb{R}^n \to (-\infty, +\infty]$ is a continuously differentiable function, $g : \mathbb{R}^n \to (-\infty, +\infty]$ is a convex function, and $\mathrm{cl}\, C$ denotes the closure of $C \subset \mathbb{R}^n$ that is a nonempty open convex set (see Assumption 9 and Remark 10 for a more precise statement). In the context of machine learning or signal processing, g is called a regularizer or a penalty function. Problem (1.1) includes many applications, such as a maximum a posteriori probability (MAP) estimate [1, 2], ridge regression [3], the least absolute shrinkage and selection operator (LASSO) [4], phase retrieval [5, 6], and blind deconvolution [7–9].

The most well-known algorithm for solving (1.1) would be a proximal algorithm. The proximal algorithm uses the proximal mapping for its updates, that can be easily computed for a certain g (see, for example, [10]) in a closed-form expression. The proximal algorithm has been actively studied since the 1970s (see, e.g., [11, 12]); The proximal algorithm includes many algorithms such as the proximal gradient method [13–15], the fast iterative shrinkage-thresholding algorithm (FISTA) [16], the proximal Newton method [17], and the proximal quasi-Newton method [18, 19]. On the other hand, these algorithms require global Lipschitz continuity for ∇f, i.e., there exists $L > 0$ such that $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. This property is sometimes restrictive, and in some applications, the global Lipschitz continuity does not hold.

Bolte et al. [5] proposed the Bregman proximal gradient algorithm (BPG)[1], which does not require the global Lipschitz continuity for ∇f:

$$\boldsymbol{x}^{k+1} = \operatorname*{argmin}_{\boldsymbol{x}\in \mathrm{cl}\, C} \left\{ \langle \nabla f(\boldsymbol{x}^k), \boldsymbol{x} - \boldsymbol{x}^k \rangle + g(\boldsymbol{x}) + \frac{1}{\lambda} D_\phi(\boldsymbol{x}, \boldsymbol{x}^k) \right\}, \tag{1.2}$$

where $\lambda > 0$ and $D_\phi(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) - \langle \nabla\phi(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle$ is the Bregman distance [21] with a convex and continuously differentiable function $\phi : \mathbb{R}^n \to (-\infty, +\infty]$. The method has been refined and extended from various perspectives, as described below. Ding et al. [22] proposed the stochastic BPG and applied it to deep learning problems. Gao et al. [23] proposed an alternating version of BPG and its acceleration version. Hanzely et al. [24] proposed an accelerated BPG for convex optimization problems. Takahashi et al. [6] proposed a Bregman proximal gradient

[1] The Bregman gradient method was first introduced for the case of $g \equiv 0$ as the mirror descent [20]. The main motivation of developing the method is perhaps to simplify the computation of the projection operator rather than to weaken the assumption of global Lipschitz continuity. For example, the mirror descent is used for minimizing the Kullback–Leibler divergence over the unit simplex [10, Example 9.10].

algorithm exploiting the difference of convex functions (DC) structure and its acceleration version for nonconvex optimization problems.

BPG globally converges to a stationary point (also called a critical point) under the smooth adaptable property (so-called L-smad) [5] for (f, ϕ) (see Definition 5), also called relative smoothness [25]. The smooth adaptable property is more permissive than the global Lipschitz continuity. From the viewpoint of the smooth adaptable property, (1.2) is majorized by the upper approximation of f with ϕ. While Subproblem (1.2) is expressed in a closed form when g and ϕ have simple structures, such a combination of g and ϕ is limited to some examples, e.g., [5, 9, 26, 27].

One of the weaknesses of BPG is that Subproblem (1.2) is not easily solved for a general ϕ. Even if $g \equiv 0$, it is not clear whether the subproblem of BPG has a closed-form solution. In fact, Subproblem (1.2) reduces to $\min_{\boldsymbol{x}} \phi(\boldsymbol{x}) + \langle \lambda \nabla f(\boldsymbol{x}^k) - \nabla \phi(\boldsymbol{x}^k), \boldsymbol{x} \rangle$, which often have no closed-form solutions (for example, $\phi(\boldsymbol{x}) = \frac{1}{p}\|\boldsymbol{x}\|_p^p$ with $1 < p \in \mathbb{R}$). To resolve this issue, assuming that ϕ is $\mathcal{C}^2$, we define the approximate Bregman distance $\tilde{D}_\phi$, given by

$$\tilde{D}_\phi(\boldsymbol{x}, \boldsymbol{y}) := \frac{1}{2}\langle \nabla^2 \phi(\boldsymbol{y})(\boldsymbol{x} - \boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \simeq D_\phi(\boldsymbol{x}, \boldsymbol{y}),$$

which is the second-order approximation of $\phi(\boldsymbol{x})$ around $\boldsymbol{y}$ in $D_\phi(\boldsymbol{x}, \boldsymbol{y})$. Using $\tilde{D}_\phi$ instead of D_ϕ, we propose a new algorithm, named the approximate Bregman proximal gradient algorithm (ABPG). Because the subproblem of ABPG can be formulated as a quadratic formula, it is easier to solve than (1.2). When ϕ is strongly convex and $g \equiv 0$, the subproblem of ABPG always has a closed-form solution given by $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \frac{1}{\lambda}\nabla^2 \phi(\boldsymbol{x}^k)^{-1} \nabla f(\boldsymbol{x}^k)$. When ϕ is separable, i.e., $\nabla^2 \phi(\boldsymbol{x}^k)$ is diagonal, and g is prox-friendly, the subproblem of ABPG reduces to the proximal calculus (see also Sect. 5 and Remark 25 for specific examples). Table 1 summarizes these facts.

Similarly to BPG, we can analyze the convergence results of ABPG under the smooth adaptable property. ABPG uses the line search procedure to ensure the accuracy of the approximate Bregman distance $\tilde{D}_\phi(\boldsymbol{x}, \boldsymbol{y})$. We have proven that the line search procedure is well-defined under the smooth adaptable property. We have established the global subsequential convergence under standard assumptions. We have also established the global convergence for the special case $g \equiv 0$ under the Kurdyka–Łojasiewicz (KL) property.

ABPG includes some existing algorithms depending on the choice of ϕ. We summarize this in Table 2 (also includes the case $g \equiv 0$). When $\phi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2$, i.e. $\nabla^2 \phi(\boldsymbol{x}) = \boldsymbol{I}$, ABPG is equivalent to the proximal gradient method. Assuming that f is convex and $\mathcal{C}^2$, when $\phi(\boldsymbol{x}) = f(\boldsymbol{x})$, ABPG is equivalent to the proximal Newton method, and when $\phi(\boldsymbol{x}) = f(\boldsymbol{x}) + \frac{\kappa_k}{2}\|\boldsymbol{x}\|^2$, i.e. $\nabla^2 \phi(\boldsymbol{x}^k) = \nabla f(\boldsymbol{x}^k) + \kappa_k \boldsymbol{I}$ for $\kappa_k > 0$, ABPG is equivalent to the regularized proximal Newton method [29] with a parameter $\kappa_k > 0$. When $\nabla^2 \phi(\boldsymbol{x}^k) = \boldsymbol{Q}_k$ with symmetric positive definite matrices $\boldsymbol{Q}_k \in \mathbb{R}^{n \times n}$ at the kth iteration, ABPG is equivalent to the proximal quasi-Newton method. ABPG corresponds to these existing algorithms with the above $\nabla^2 \phi$. The convergence properties of these algorithms are often investigated under the assumption of

Table 1 ABPG has a closed-form solution in several settings when $t_k = 1$ (see Algorithm 1). Let f_0 be a continuously differentiable function with Lipschitz continuous gradients, ϕ_0 be a continuously differentiable convex function, and prox_g be a proximal mapping of g. Let v_i be the ith element of $\nabla f(\boldsymbol{x}^k)$ and s_i be the (i,i) element of $\nabla^2\phi(\boldsymbol{x}^k)^{-1}$. The Kullback–Leibler divergence $D_{\mathrm{KL}}(\cdot,\cdot)$ will be defined by (5.7). Let $\boldsymbol{A} \in \mathbb{R}_+^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}_+^m$, $\boldsymbol{W} \in \mathbb{R}_+^{m\times r}$, $\boldsymbol{H} \in \mathbb{R}_+^{r\times n}$, and $\boldsymbol{B} \in \mathbb{R}_+^{m\times n}$. See also Sect. 5

Reference	(f,ϕ)	Calculus of the subproblem
Sects. 5.1 & 5.2	$f = f_0 + \theta\phi_0$, $\phi = \phi_0 + \frac{1}{2}\|\cdot\|^2$ with a separable ϕ_0	
Sect. 5.4, [26]	$f(\boldsymbol{x}) = D_{\mathrm{KL}}(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{b})$, $\phi(\boldsymbol{x}) = \sum_i x_i \log x_i + \frac{1}{2}\|\boldsymbol{x}\|^2$	For separable $g(\boldsymbol{x}) = \sum_i g_i(x_i)$, $\boldsymbol{x}^{k+1} = \left(\mathrm{prox}_{\lambda s_i g_i}(x_i^k - s_i v_i)\right)_{i=1}^n$
[26]	$f(\boldsymbol{x}) = D_{\mathrm{KL}}(\boldsymbol{b}, \boldsymbol{A}\boldsymbol{x})$, $\phi(\boldsymbol{x}) = -\sum_i \log x_i$	
[28][2]	$f(\boldsymbol{W},\boldsymbol{H}) = D_{\mathrm{KL}}(\boldsymbol{B}, \boldsymbol{W}\boldsymbol{H})$, $\phi(\boldsymbol{x}) = -\sum_i \log x_i + \frac{1}{2}\|\boldsymbol{x}\|^2$, $x = (\boldsymbol{W},\boldsymbol{H})$	
Sect. 5.3	Any (f,ϕ) is L-smad, $\phi(\boldsymbol{x})$ is strongly convex.	$g \equiv 0 \implies \boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \frac{1}{\lambda}\nabla^2\phi(\boldsymbol{x}^k)^{-1}\nabla f(\boldsymbol{x}^k)$

[2] ABPG can be applied to the auxiliary function $\hat{f}_k$ defined in [28]

Table 2 ABPG corresponds to existing algorithms in setting some $\nabla^2\phi(\boldsymbol{x}^k)$. In this table, κ_k is a positive scalar, and $\boldsymbol{Q}_k$ is a symmetric positive definite matrix at the kth iteration. Note that f is required to be convex for some algorithms

$\nabla^2\phi(\boldsymbol{x}^k)$	Algorithm	Algorithm ($g \equiv 0$)
$\nabla^2\phi(\boldsymbol{x}^k) = \boldsymbol{I}$	Proximal gradient method [13–15]	Gradient descent
$\nabla^2\phi(\boldsymbol{x}^k) = \nabla^2 f(\boldsymbol{x}^k)$	Proximal Newton method [17]	Newton method
$\nabla^2\phi(\boldsymbol{x}^k) = \nabla^2 f(\boldsymbol{x}^k) + \frac{\kappa_k}{2}\boldsymbol{I}$	Regularized proximal Newton method [29]	Regularized Newton method [30]
$\nabla^2\phi(\boldsymbol{x}^k) = \boldsymbol{Q}_k$	Proximal quasi-Newton method [18, 19]	Quasi-Newton method

the global Lipschitz continuity for ∇f, but this paper gives convergence guarantees under the weaker assumption, L-smad property.

In numerical experiments, we have applied ABPG to the convex ℓ_p regularized least squares problem [31, 32], the ℓ_p loss problem [33], and the nonnegative linear system [26]. In the ℓ_p regularized least squares problem, f includes the pth power of the ℓ_p norm term with $p > 1$. Its gradient for a general p is not globally Lipschitz continuous, and it is also not locally Lipschitz continuous at $\boldsymbol{x} \in (-1,1)^n$ for $p \in (1,2)$. Moreover, when ϕ includes the ℓ_p norm in this case, the subproblem of BPG is the minimization of a pth degree polynomial function. While the subproblem of BPG has no closed-form solutions, that of ABPG has a closed-form solution even if $g \not\equiv 0$. We also have compared ABPG with BPG on the application to the nonnegative linear system.

This paper is organized as follows. Section 2 summarizes key notations such as subdifferentials, the Bregman distance, the smooth adaptable property, and the KL property. These mathematical tools are necessary to tackle nonconvex optimization problems. Section 3 proposes the approximate Bregman proximal gradient algorithm, which does not require the global Lipschitz continuity for ∇f. Section 4 establishes the properties and global convergence of our proposed algorithm. We proved the global subsequential convergence under standard assumptions and the global convergence when $g \equiv 0$. Section 5 shows applications to the ℓ_p regularized least squares problem, the ℓ_p loss problem, and the nonnegative linear system.

Notation 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 $\mathbb{R}^n$, $\mathbb{R}^n_+$, and $\mathbb{R}^n_{++}$ be the real space of n dimension, the nonnegative orthant of $\mathbb{R}^n$, and the positive orthant of $\mathbb{R}^n$, respectively. Let $\mathbb{R}^{n\times m}$ be the set of $n \times m$ real matrices. Vectors and matrices are shown in boldface. The all one vector is $\mathbf{1} \in \mathbb{R}^n$, the all zero vector is $\mathbf{0} \in \mathbb{R}^n$, and the identity matrix is $\boldsymbol{I} \in \mathbb{R}^{n\times n}$. Let $|\boldsymbol{x}|$ and $\boldsymbol{x}^p$ be the elementwise absolute and pth power vectors of $\boldsymbol{x} \in \mathbb{R}^n$, respectively. Given a real number $p \geq 1$, the ℓ_p norm is defined by $\|\boldsymbol{x}\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$. Let $\lambda_{\max}(\boldsymbol{M})$ be the largest eigenvalue of a symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$.

Let $\operatorname{int} C$ and $\operatorname{cl} C$ be the interior and the closure of a set $C \subset \mathbb{R}^n$, respectively. We also define the distance from a point $\boldsymbol{x} \in \mathbb{R}^n$ to C by $\operatorname{dist}(\boldsymbol{x}, C) := \inf_{\boldsymbol{y}\in C} \|\boldsymbol{x} - \boldsymbol{y}\|$. The indicator function δ_C is defined by $\delta_C(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in C$ and $\delta_C(\boldsymbol{x}) = +\infty$ otherwise. The sign function $\operatorname{sgn}(x)$ is defined by $\operatorname{sgn}(x) = -1$ for $x < 0$, $\operatorname{sgn}(x) = 0$ for $x = 0$, and $\operatorname{sgn}(x) = 1$ for $x > 0$.

2 Preliminaries

2.1 Subdifferentials

We first introduce the definitions of subdifferentials. For an extended-real-valued function $f : \mathbb{R}^n \to [-\infty, +\infty]$, the effective domain of f is defined by $\operatorname{dom} f := \{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) < +\infty\}$. The function f is said to be proper if $f(\boldsymbol{x}) > -\infty$ for all $\boldsymbol{x} \in \mathbb{R}^n$ and $\operatorname{dom} f \neq \emptyset$.

Definition 1 (Subdifferentials [34, Definition 8.3]) Let $f : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper and lower semicontinuous function.

(i) The regular subdifferential of f at $\boldsymbol{x} \in \operatorname{dom} f$ is defined by

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

When $\boldsymbol{x} \notin \operatorname{dom} \phi$, we set $\hat{\partial} f(\boldsymbol{x}) = \emptyset$.

(ii) The limiting subdifferential of f at $\boldsymbol{x} \in \mathbb{R}^n$ is defined by

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

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 any $\boldsymbol{x} \in \mathbb{R}^n$ [34, Theorem 8.6]. See also [35, Example 1.31] and [34] for examples of subdifferential calculus. We also define $\operatorname{dom} \partial f := \{\boldsymbol{x} \in \mathbb{R}^n \mid \partial f(\boldsymbol{x}) \neq \emptyset\}$. Note that when f is convex, the limiting subdifferential coincides with the (classical) subdifferential [34, Proposition 8.12].

2.2 Bregman distances

Let C be a nonempty open convex subset of $\mathbb{R}^n$. We introduce the kernel generating distance [5] and the Bregman distance.

Definition 2 (Kernel generating distance [5, Definition 2.1]) Associated with C, a function $\phi : \mathbb{R}^n \to (-\infty, +\infty]$ is called a kernel generating distance if it satisfies 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 $\operatorname{int} \operatorname{dom} \phi \equiv C$.

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

Definition 3 (Bregman distance [21]) Given a kernel generating distance $\phi \in \mathcal{G}(C)$, a Bregman distance $D_\phi : \operatorname{dom} \phi \times \operatorname{int} \operatorname{dom} \phi \to \mathbb{R}_+$ associated with ϕ is defined by

$$D_\phi(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) - \langle \nabla \phi(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle. \tag{2.1}$$

Note that the Bregman distance D_ϕ is not a distance because it does not satisfy the symmetry and the triangle inequality. Due to the convexity of ϕ, it holds that $D_\phi(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ for any $\boldsymbol{x} \in \operatorname{dom} \phi$ and $\boldsymbol{y} \in \operatorname{int} \operatorname{dom} \phi$. In addition, when ϕ is strictly convex, $D_\phi(\boldsymbol{x}, \boldsymbol{y}) = 0$ holds if and only if $\boldsymbol{x} = \boldsymbol{y}$.

Example 4 We show some famous examples of Bregman distances:

- Squared Euclidean distance: Let $\phi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2$ and $\operatorname{dom} \phi = \mathbb{R}^n$. Then, we have the squared Euclidean distance $D_\phi(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$.
- Kullback–Leibler divergence [36]: The Boltzmann–Shannon entropy $\phi(\boldsymbol{x}) = \sum_{i=1}^n x_i \log x_i$ with $0 \log 0 = 0$ and $\operatorname{dom} \phi = \mathbb{R}^n_+$. Then, we obtain the Kullback–Leibler divergence $D_\phi(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^n x_i \log \frac{x_i}{y_i}$ on the unit simplex $\{\boldsymbol{x} \in \mathbb{R}^n_+ \mid \sum_{i=1}^n x_i = 1\}$.
- Itakura–Saito divergence [37]: The Burg entropy $\phi(\boldsymbol{x}) = -\sum_{i=1}^n \log x_i$ and $\operatorname{dom} \phi = \mathbb{R}^n_{++}$. Then, we obtain the Itakura–Saito divergence $D_\phi(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^n \left(\frac{x_i}{y_i} - \log \frac{x_i}{y_i} - 1 \right)$.

See [25, 26, 38] and [39, Table 2.1] for more examples.

2.3 Smooth adaptable property

Now let us define the smooth adaptable property [5], also called relative smoothness [25].

Definition 5 (L-smooth adaptable property [5, Definition 2.2]) Consider a pair of functions (f, ϕ) satisfying the following conditions:

(i) $\phi \in \mathcal{G}(C)$,
(ii) $f : \mathbb{R}^n \to (-\infty, +\infty]$ 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.

The smooth adaptable property for the pair (f, ϕ) implies that f is globally majorized by ϕ with some constant $L > 0$. The smooth adaptable property induces a useful lemma.

Lemma 6 *(Extended descent lemma* [5, Lemma 2.1]*) The pair of functions (f, ϕ) is L-smad on C if and only if*

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

Recalling $\phi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2$ on $C = \mathbb{R}^n$, Lemma 6 coincides with the classical descent lemma, i.e., ∇f is Lipschitz continuous.

2.4 Kurdyka–ojasiewicz property

The Kurdyka–Łojasiewicz (KL) property plays a central role in global convergence. Attouch et al. [40] extended the Łojasiewicz gradient inequality [41, 42] to nonsmooth functions.

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 define the Kurdyka–Łojasiewicz property.

Definition 7 (Kurdyka–Łojasiewicz property [40, Definition 7]) Let $f : \mathbb{R}^n \to (-\infty, +\infty]$ 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

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

the following inequality holds:

$$\psi'(f(\boldsymbol{x}) - f(\hat{\boldsymbol{x}}))\operatorname{dist}(\boldsymbol{0}, \partial f(\boldsymbol{x})) \geq 1. \tag{2.2}$$

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

Let $\psi(s) = cs^{1-\theta}$ for some $\theta \in [0, 1)$ and $c > 0$. The parameter θ is called the KL exponent, which affects the rate of convergence. When f is $\mathcal{C}^1$, (2.2) coincides with the Łojasiewicz gradient inequality [41, 42], defined by

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

When f is $\mathcal{C}^1$, the KL exponent θ is also called the Łojasiewicz exponent. Li and Pong [43] established calculus rules of the KL exponent.

The uniformized KL property is induced from the KL property.

Lemma 8 *(Uniformized KL property* [44, Lemma 6]*) Suppose that $f : \mathbb{R}^n \to (-\infty, +\infty]$ 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*

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

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

3 Proposed algorithm

Throughout this paper, we make the following assumptions.

Assumption 9 (i) $\phi \in \mathcal{G}(C)$ with $\operatorname{cl} C = \operatorname{cl} \operatorname{dom} \phi$ is $\mathcal{C}^2$ on $C = \operatorname{int} \operatorname{dom} \phi$.

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

(iii) $g : \mathbb{R}^n \to (-\infty, +\infty]$ is a proper and convex function with $\operatorname{dom} g \cap C \neq \emptyset$.

(iv) $\Psi^* := \inf_{\boldsymbol{x} \in \operatorname{cl} C} \Psi(\boldsymbol{x}) > -\infty$.

(v) For any $\boldsymbol{x} \in \operatorname{int} \operatorname{dom} \phi$ and $\lambda > 0$, $\lambda g(\boldsymbol{u}) + \frac{1}{2}\langle \nabla^2 \phi(\boldsymbol{x})(\boldsymbol{u} - \boldsymbol{x}), \boldsymbol{u} - \boldsymbol{x}\rangle$ is supercoercive, that is,

$$\lim_{\|\boldsymbol{u}\| \to \infty} \frac{\lambda g(\boldsymbol{u}) + \frac{1}{2}\langle \nabla^2 \phi(\boldsymbol{x})(\boldsymbol{u} - \boldsymbol{x}), \boldsymbol{u} - \boldsymbol{x}\rangle}{\|\boldsymbol{u}\|} = \infty.$$

Assumption 9(i-iv) are standard for Bregman-type algorithms [5, 6, 22] and satisfied in practice. Assumption 9(v) is a coercive condition, which is used for the well-posedness of the ABPG mapping (Lemma 12). If ϕ is strongly convex, then Assumption 9(v) holds [45, Corollary 11.17].

Remark 10 In Problem (1.1), $\operatorname{cl} C$ is more of the effective domain of ϕ than a constraint. We use $C = \operatorname{int} \operatorname{dom} \phi$ to guarantee that points are in the effective domain of ϕ. Therefore, C depends on the choice of ϕ. In Sects. 4.2 and 4.3, we use $C \equiv \mathbb{R}^n$ following [5].

3.1 Approximate bregman proximal gradient algorithm

The Bregman proximal gradient mapping [5] at $\boldsymbol{x} \in C$ is defined by

$$\mathcal{T}_\lambda(\boldsymbol{x}) := \underset{\boldsymbol{u}\in \mathrm{cl}\, C}{\operatorname{argmin}} \left\{ \langle \nabla f(\boldsymbol{x}), \boldsymbol{u}-\boldsymbol{x}\rangle + g(\boldsymbol{u}) + \frac{1}{\lambda} D_\phi(\boldsymbol{u},\boldsymbol{x}) \right\}, \tag{3.1}$$

where $\lambda > 0$ is called a stepsize. For general ϕ and g, it is difficult to calculate $\mathcal{T}_\lambda(\boldsymbol{x})$ in a closed form. Using Assumption 9(i), we obtain the second-order approximation to $\phi(\boldsymbol{u})$ around $\boldsymbol{x}$, given by

$$\begin{aligned} D_\phi(\boldsymbol{u},\boldsymbol{x}) &= \phi(\boldsymbol{u}) - \phi(\boldsymbol{x}) - \langle \nabla\phi(\boldsymbol{x}), \boldsymbol{u}-\boldsymbol{x}\rangle \\ &\simeq \phi(\boldsymbol{x}) + \langle \nabla\phi(\boldsymbol{x}), \boldsymbol{u}-\boldsymbol{x}\rangle + \frac{1}{2}\langle \nabla^2\phi(\boldsymbol{x})(\boldsymbol{u}-\boldsymbol{x}), \boldsymbol{u}-\boldsymbol{x}\rangle \\ &\quad - \phi(\boldsymbol{x}) - \langle \nabla\phi(\boldsymbol{x}), \boldsymbol{u}-\boldsymbol{x}\rangle \\ &= \frac{1}{2}\langle \nabla^2\phi(\boldsymbol{x})(\boldsymbol{u}-\boldsymbol{x}), \boldsymbol{u}-\boldsymbol{x}\rangle =: \tilde{D}_\phi(\boldsymbol{u},\boldsymbol{x}). \end{aligned}$$

We call $\tilde{D}_\phi(\boldsymbol{u},\boldsymbol{x}) \geq 0$ the (second-order) approximate Bregman distance. Note that it does not always hold that $D_\phi(\boldsymbol{u},\boldsymbol{x}) \leq \tilde{D}_\phi(\boldsymbol{u},\boldsymbol{x})$ or $D_\phi(\boldsymbol{u},\boldsymbol{x}) \geq \tilde{D}_\phi(\boldsymbol{u},\boldsymbol{x})$ for any $\boldsymbol{u},\boldsymbol{x} \in \mathbb{R}^n$. Because of this, our algorithm shown later needs a line search procedure. We consider, instead of (3.1), the approximate Bregman proximal gradient mapping at $\boldsymbol{x} \in C$, defined by

$$\tilde{\mathcal{T}}_\lambda(\boldsymbol{x}) := \underset{\boldsymbol{u}\in \mathrm{cl}\, C}{\operatorname{argmin}} \left\{ \langle \nabla f(\boldsymbol{x}), \boldsymbol{u}-\boldsymbol{x}\rangle + g(\boldsymbol{u}) + \frac{1}{\lambda} \tilde{D}_\phi(\boldsymbol{u},\boldsymbol{x}) \right\}. \tag{3.2}$$

Using Assumption 9(iii) and the positive semidefiniteness of $\nabla^2\phi(\boldsymbol{x})$, (3.2) is a convex optimization problem. In addition, $\tilde{\mathcal{T}}_\lambda(\boldsymbol{x})$ is a singleton when ϕ is strongly convex. If ϕ is not strongly convex, it is enough to use $\phi_0 := \phi + \frac{\sigma}{2}\|\cdot\|^2$ with $\sigma > 0$ as a new kernel generating distance. Setting $\phi = \frac{1}{2}\|\cdot\|^2$, i.e., $\nabla^2\phi(\boldsymbol{x}) = \boldsymbol{I}$, $\tilde{\mathcal{T}}_\lambda(\boldsymbol{x})$ is equivalent to a proximal mapping. Additionally, we make the following assumption.

Assumption 11 For any $\boldsymbol{x} \in C$ and $\lambda > 0$, we have $\tilde{\mathcal{T}}_\lambda(\boldsymbol{x}) \subset C$.

Assumption 11 guarantees that a sequence generated by ABPG is in $C \equiv \operatorname{int}\operatorname{dom}\phi$ (see also Remark 10). Assumption 11 holds when $C \equiv \mathbb{R}^n$. In the same discussion as [5, p.2136], another approach to warrant Assumption 11 is to consider extending the prox-boundedness [34, Definition 1.23] to the approximate Bregman proximal gradient mapping and its envelope. Moreover, if ϕ is strongly convex, an envelope function $\inf_{\boldsymbol{u}\in \mathrm{cl}\, C} \left\{ \langle \nabla f(\boldsymbol{x}), \boldsymbol{u}-\boldsymbol{x}\rangle + g(\boldsymbol{u}) + \frac{1}{\lambda}\tilde{D}_\phi(\boldsymbol{u},\boldsymbol{x}) \right\}$ is bounded from below, which implies the prox-boundedness from [34, Exercise 1.24]. More discussions for Assumption 11 have been in [5]. We can prove that $\tilde{\mathcal{T}}_\lambda(\boldsymbol{x})$ is well-posed, as shown in the following lemma, by using Assumptions 9 and 11.

Lemma 12 *(Well-posedness of $\tilde{\mathcal{T}}_\lambda(\boldsymbol{x})$) Suppose that Assumptions 9 and 11 hold. Let $\boldsymbol{x} \in \operatorname{int}\operatorname{dom}\phi$. Then, the mapping $\tilde{\mathcal{T}}_\lambda(\boldsymbol{x})$ is a nonempty and compact subset of C for any $\lambda > 0$.*

Proof Take any $\boldsymbol{x} \in \operatorname{int}\operatorname{dom}\phi$ and $\lambda > 0$. Let

$$\Phi(\boldsymbol{u}) := \langle \lambda \nabla f(\boldsymbol{x}), \boldsymbol{u}\rangle + \lambda g(\boldsymbol{u}) + \frac{1}{2}\langle \nabla^2 \phi(\boldsymbol{x})(\boldsymbol{u}-\boldsymbol{x}), \boldsymbol{u}-\boldsymbol{x}\rangle,$$

which implies $\tilde{\mathcal{T}}_\lambda(\boldsymbol{x}) = \operatorname{argmin}_{\boldsymbol{u}\in \operatorname{cl} C} \Phi(\boldsymbol{u})$. We have

$$\begin{aligned}\Phi(\boldsymbol{u}) &\geq -|\langle \lambda \nabla f(\boldsymbol{x}), \boldsymbol{u}\rangle| + \lambda g(\boldsymbol{u}) + \frac{1}{2}\langle \nabla^2 \phi(\boldsymbol{x})(\boldsymbol{u}-\boldsymbol{x}), \boldsymbol{u}-\boldsymbol{x}\rangle \\ &= \|\boldsymbol{u}\| \left(\frac{\lambda g(\boldsymbol{u}) + \frac{1}{2}\langle \nabla^2 \phi(\boldsymbol{x})(\boldsymbol{u}-\boldsymbol{x}), \boldsymbol{u}-\boldsymbol{x}\rangle}{\|\boldsymbol{u}\|} - \frac{|\langle \lambda \nabla f(\boldsymbol{x}), \boldsymbol{u}\rangle|}{\|\boldsymbol{u}\|} \right).\end{aligned}$$

Because $\lambda g(\boldsymbol{u}) + \frac{1}{2}\langle \nabla^2 \phi(\boldsymbol{x})(\boldsymbol{u}-\boldsymbol{x}), \boldsymbol{u}-\boldsymbol{x}\rangle$ is supercoercive, we have $\lim_{\|\boldsymbol{u}\|\to\infty} \Phi(\boldsymbol{u}) = \infty$. Since Φ is also proper and continuous due to the convexity of g, we can use Weierstrass's theorem (see, for example, [34, Theorem 1.9]). It follows that $\tilde{\mathcal{T}}_\lambda(\boldsymbol{u})$ is a nonempty and compact set.

Inspired by existing Newton-type methods, we take the search direction procedure and the line search procedure. The search direction of our proposed method solves the subproblem at $\boldsymbol{x}^k \in C$,

$$\boldsymbol{d}^k = \operatorname*{argmin}_{\boldsymbol{x}^k+\boldsymbol{d}\in \operatorname{cl} C} \left\{ \langle \nabla f(\boldsymbol{x}^k), \boldsymbol{d}\rangle + g(\boldsymbol{x}^k + \boldsymbol{d}) + \frac{1}{2\lambda}\langle \nabla^2 \phi(\boldsymbol{x}^k)\boldsymbol{d}, \boldsymbol{d}\rangle \right\}, \tag{3.3}$$

where $\boldsymbol{d}$ is a variable and $\boldsymbol{x}^k$ is fixed. Subproblem (3.3) is essentially the same as (3.2) of $\tilde{\mathcal{T}}_\lambda(\boldsymbol{x}^k)$ due to $\boldsymbol{d}^k = \tilde{\mathcal{T}}_\lambda(\boldsymbol{x}^k) - \boldsymbol{x}^k$. When ϕ and g are separable, (3.3) may be solved in a closed-form expression. Because $\nabla^2\phi(\boldsymbol{x})$ is a diagonal matrix, (3.3) reduces to the proximal calculus (see, for the proximal calculus, [10, Chapter 6 and Appendix B]). For example, when $\phi(\boldsymbol{x}) = \frac{1}{p}\|\boldsymbol{x}\|_p^p$ and $g(\boldsymbol{x}) = \theta_1\|\boldsymbol{x}\|_1$ with $\theta_1 > 0$, we have a closed-form solution of (3.3) (see Remark 25).

We use the line search procedure to select $t_k > 0$ at the kth iteration so that the sufficient decreasing condition holds for $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + t_k\boldsymbol{d}^k$,

$$\Psi(\boldsymbol{x}^{k+1}) \leq \Psi(\boldsymbol{x}^k) + \alpha t_k(\langle \nabla f(\boldsymbol{x}^k), \boldsymbol{d}^k\rangle + g(\boldsymbol{x}^k + \boldsymbol{d}^k) - g(\boldsymbol{x}^k)),$$

where $\boldsymbol{x}^k \in C$, $\alpha \in (0, 1)$, and $\boldsymbol{d}^k$ is given by (3.3). Since C is convex, $\boldsymbol{x}^k + t_k\boldsymbol{d}^k = t_k(\boldsymbol{x}^k + \boldsymbol{d}^k) + (1 - t_k)\boldsymbol{x}^k \in C$ with $\boldsymbol{x}^k \in C$ and $\boldsymbol{x}^k + \boldsymbol{d}^k \in \tilde{\mathcal{T}}_\lambda(\boldsymbol{x}^k) \subset C$ (Lemma 12). The simple way to calculate t_k would be the backtracking line search (see, for example, [46]), and we also use the backtracking approach.

Now we are ready to describe our algorithm for solving the nonconvex composite optimization problem (1.1).

Algorithm 1 Approximate Bregman proximal gradient algorithm (ABPG)

Input: A $\mathcal{C}^2$ function $\phi \in \mathcal{G}(C)$ such that the pair (f, ϕ) is L-smad on C.
Initialization: $\boldsymbol{x}^0 \in \operatorname{int}\operatorname{dom}\phi$, $\eta \in (0,1)$, $\alpha \in (0,1)$, and $\lambda > 0$.
for $k = 0, 1, 2, \ldots,$ **do**
Compute a search direction with a fixed $\boldsymbol{x}^k$:

$$\boldsymbol{d}^k = \operatorname*{argmin}_{\boldsymbol{x}^k + \boldsymbol{d} \in \operatorname{cl} C} \left\{ \langle \nabla f(\boldsymbol{x}^k), \boldsymbol{d} \rangle + g(\boldsymbol{x}^k + \boldsymbol{d}) + \frac{1}{2\lambda} \langle \nabla^2 \phi(\boldsymbol{x}^k)\boldsymbol{d}, \boldsymbol{d} \rangle \right\}. \tag{3.4}$$

Set $t_k = 1$.
while $\Psi(\boldsymbol{x}^k + t_k \boldsymbol{d}^k) > \Psi(\boldsymbol{x}^k) + \alpha t_k (\langle \nabla f(\boldsymbol{x}^k), \boldsymbol{d}^k \rangle + g(\boldsymbol{x}^k + \boldsymbol{d}^k) - g(\boldsymbol{x}^k))$ **do**
$t_k \leftarrow \eta t_k$
Update $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + t_k \boldsymbol{d}^k$.

Remark 13 Assuming that f is $\mathcal{C}^2$, we see the useful property [26, Proposition 1]: There exists $L > 0$ such that $L\phi - f$ is convex if and only if

$$\exists L > 0,\ L\nabla^2 \phi(\boldsymbol{x}) \succeq \nabla^2 f(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \operatorname{int}\operatorname{dom}\phi. \tag{3.5}$$

Condition (3.5) is useful for estimating L in applications such that the Poisson linear inverse problem [26, Lemma 7], phase retrieval [5, Lemma 5.1][6, Propositions 5 and 6], and blind deconvolution [9, Theorem 1]. Regarding a role of $\frac{1}{\lambda}\nabla^2\phi(\boldsymbol{x})$ in (3.2), Condition (3.5) implies that $\frac{1}{\lambda}\nabla^2\phi(\boldsymbol{x})$ is an upper approximation to $\nabla^2 f(\boldsymbol{x})$ when $\lambda \in (0, 1/L)$.

In the next section, we show the well-definedness of the search direction and the line search, and then establish a global convergence result to a stationary point.

4 Convergence analysis

Throughout this section, we make the following assumption.

Assumption 14 The pair (f, ϕ) is L-smad on C.

4.1 Properties of the proposed algorithm

We first show the search direction property. It induces the decreasing property.

Proposition 15 *(Search direction properties) Suppose that Assumptions 9, 11 and 14 hold. For any $\lambda > 0$, $\boldsymbol{x} \in \operatorname{int}\operatorname{dom}\phi$, and $\boldsymbol{d} \in \mathbb{R}^n$ defined by*

$$\boldsymbol{d} = \operatorname*{argmin}_{\boldsymbol{u}} \left\{ \langle \nabla f(\boldsymbol{x}), \boldsymbol{u} \rangle + g(\boldsymbol{x} + \boldsymbol{u}) + \frac{1}{2\lambda} \langle \nabla^2 \phi(\boldsymbol{x})\boldsymbol{u}, \boldsymbol{u} \rangle \right\} \ \textit{s.t.}\ \boldsymbol{x} + \boldsymbol{u} \in \operatorname{cl} C, \tag{4.1}$$

we have

$$\Psi(\boldsymbol{x}^+) - \Psi(\boldsymbol{x}) \leq t(\langle \nabla f(\boldsymbol{x}), \boldsymbol{d} \rangle + g(\boldsymbol{x} + \boldsymbol{d}) - g(\boldsymbol{x})) + o(t\|\boldsymbol{d}\|), \tag{4.2}$$

where $\boldsymbol{x}^+ = \boldsymbol{x} + t\boldsymbol{d}$ *and* $t \in (0, 1]$. *Furthermore, we have*

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{d} \rangle + g(\boldsymbol{x} + \boldsymbol{d}) - g(\boldsymbol{x}) \leq -\frac{1}{2\lambda} \langle \nabla^2 \phi(\boldsymbol{x})\boldsymbol{d}, \boldsymbol{d} \rangle. \tag{4.3}$$

Proof By using Taylor's theorem and $g(\boldsymbol{x} + t\boldsymbol{d}) \leq tg(\boldsymbol{x} + \boldsymbol{d}) + (1 - t)g(\boldsymbol{x})$ due to the convexity of g, we obtain

$$\begin{aligned} \Psi(\boldsymbol{x}^+) - \Psi(\boldsymbol{x}) &= f(\boldsymbol{x}^+) - f(\boldsymbol{x}) + g(\boldsymbol{x}^+) - g(\boldsymbol{x}) \\ &\leq t(\langle \nabla f(\boldsymbol{x}), \boldsymbol{d} \rangle + g(\boldsymbol{x} + \boldsymbol{d}) - g(\boldsymbol{x})) + o(t\|\boldsymbol{d}\|), \end{aligned}$$

which proves (4.2).

Using the optimality condition of (4.1), we have

$$\begin{aligned} \langle \nabla f(\boldsymbol{x}), \boldsymbol{d} \rangle + g(\boldsymbol{x} + \boldsymbol{d}) + \frac{1}{2\lambda} \langle \nabla^2 \phi(\boldsymbol{x})\boldsymbol{d}, \boldsymbol{d} \rangle &\leq t\langle \nabla f(\boldsymbol{x}), \boldsymbol{d} \rangle + g(\boldsymbol{x} + t\boldsymbol{d}) \\ &\quad + \frac{t^2}{2\lambda} \langle \nabla^2 \phi(\boldsymbol{x})\boldsymbol{d}, \boldsymbol{d} \rangle \\ &\leq t\langle \nabla f(\boldsymbol{x}), \boldsymbol{d} \rangle + tg(\boldsymbol{x} + \boldsymbol{d}) \\ &\quad + (1 - t)g(\boldsymbol{x}) + \frac{t^2}{2\lambda} \langle \nabla^2 \phi(\boldsymbol{x})\boldsymbol{d}, \boldsymbol{d} \rangle, \end{aligned}$$

where in the first inequality, function values are compared between the optimal solution $\boldsymbol{d}$ and a feasible solution $t\boldsymbol{d}$ with $t \in (0, 1]$, and the second inequality holds due to the convexity of g. Moving the right-hand side to the left-hand side and simplifying the above inequality, we obtain

$$(1 - t)\langle \nabla f(\boldsymbol{x}), \boldsymbol{d} \rangle + \frac{1 - t^2}{2\lambda} \langle \nabla^2 \phi(\boldsymbol{x})\boldsymbol{d}, \boldsymbol{d} \rangle + (1 - t)(g(\boldsymbol{x} + \boldsymbol{d}) - g(\boldsymbol{x})) \leq 0,$$

and moreover, by using $1 - t \geq 0$,

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{d} \rangle + g(\boldsymbol{x} + \boldsymbol{d}) - g(\boldsymbol{x}) \leq -\frac{1 + t}{2\lambda} \langle \nabla^2 \phi(\boldsymbol{x})\boldsymbol{d}, \boldsymbol{d} \rangle \leq -\frac{1}{2\lambda} \langle \nabla^2 \phi(\boldsymbol{x})\boldsymbol{d}, \boldsymbol{d} \rangle$$

holds for any $t \in (0, 1]$. The desired inequality is proved.

Note that Proposition 15 does not require the L-smad property. Proposition 15 implies the search direction is the descent direction because, substituting (4.3) into (4.2), it holds that

$$\Psi(\boldsymbol{x}^+) - \Psi(\boldsymbol{x}) \leq -\frac{t}{2\lambda} \langle \nabla^2 \phi(\boldsymbol{x})\boldsymbol{d}, \boldsymbol{d} \rangle + o(t\|\boldsymbol{d}\|).$$

Let $\alpha \in (0, 1)$. We next show that our line search procedure is well-defined, i.e., the sufficient decreasing condition

$$\Psi(\boldsymbol{x}^+) \leq \Psi(\boldsymbol{x}) + \alpha t(\langle \nabla f(\boldsymbol{x}), \boldsymbol{d}\rangle + g(\boldsymbol{x}+\boldsymbol{d}) - g(\boldsymbol{x})) \tag{4.4}$$

is satisfied after finite steps of the line search procedures under the following assumption.

Assumption 16 ϕ is σ-strongly convex on C with $\sigma > 0$.

We discussed the strong convexity of ϕ in Sect. 3.1.

Lemma 17 *Suppose that Assumptions 9, 11, 14, and 16 hold. For any $\alpha \in (0, 1)$, $\lambda > 0$, and some $\beta > 0$, if*

$$t \leq \min\left\{1, \frac{1-\alpha}{\beta\lambda L}\right\}, \tag{4.5}$$

then the sufficient decreasing condition (4.4) is satisfied after finite steps of the line search procedures.

Proof When $\boldsymbol{d} = \boldsymbol{0}$, (4.4) obviously holds. We will prove (4.4) when $\boldsymbol{d} \neq \boldsymbol{0}$. Because the pair (f, ϕ) is L-smad, using Lemma 6, we have

$$\begin{aligned}\Psi(\boldsymbol{x}^+) - \Psi(\boldsymbol{x}) &= f(\boldsymbol{x}^+) - f(\boldsymbol{x}) + g(\boldsymbol{x}^+) - g(\boldsymbol{x}) \\ &\leq t\langle \nabla f(\boldsymbol{x}), \boldsymbol{d}\rangle + LD_\phi(\boldsymbol{x}^+, \boldsymbol{x}) + t(g(\boldsymbol{x}+\boldsymbol{d}) - g(\boldsymbol{x})) \\ &= t(\langle \nabla f(\boldsymbol{x}), \boldsymbol{d}\rangle + g(\boldsymbol{x}+\boldsymbol{d}) - g(\boldsymbol{x})) + LD_\phi(\boldsymbol{x}^+, \boldsymbol{x}),\end{aligned} \tag{4.6}$$

where the inequality holds due to the convexity of g. Applying Taylor's Theorem to $\phi(\boldsymbol{x}^+)$ in $D_\phi(\boldsymbol{x}^+, \boldsymbol{x})$ of (2.1), there exists $T \in (0, 1)$ such that

$$\begin{aligned}D_\phi(\boldsymbol{x}^+, \boldsymbol{x}) &= \phi(\boldsymbol{x}) + t\langle \nabla\phi(\boldsymbol{x}), \boldsymbol{d}\rangle + \frac{t^2}{2}\langle \nabla^2\phi(\boldsymbol{x} + \tau t\boldsymbol{d})\boldsymbol{d}, \boldsymbol{d}\rangle - \phi(\boldsymbol{x}) - t\langle \nabla\phi(\boldsymbol{x}), \boldsymbol{d}\rangle \\ &= \frac{t^2}{2}\langle \nabla^2\phi(\boldsymbol{x} + \tau t\boldsymbol{d})\boldsymbol{d}, \boldsymbol{d}\rangle.\end{aligned}$$

We consider $h(s) = \langle \nabla^2\phi(\boldsymbol{x}+s\boldsymbol{d})\boldsymbol{d}, \boldsymbol{d}\rangle > 0$ for $\boldsymbol{d} \neq \boldsymbol{0}$ and $s \in [0, 1]$, which is positive and finite-valued because of the strong convexity of ϕ and $\boldsymbol{x} + s\boldsymbol{d} \in \operatorname{int}\operatorname{dom}\phi$ for any $s \in [0, 1]$. We see $h(0) = \langle \nabla^2\phi(\boldsymbol{x})\boldsymbol{d}, \boldsymbol{d}\rangle > 0$ for $\boldsymbol{d} \neq \boldsymbol{0}$. Since ϕ is $\mathcal{C}^2$, h is continuous on $[0, 1]$. Using the extreme value theorem, h has the maximum value on $[0, 1]$. Because $\nabla^2\phi(\boldsymbol{x})$ is positive definite, there exists $\beta > 0$ such that

$$h(s) \leq \beta\langle \nabla^2\phi(\boldsymbol{x})\boldsymbol{d}, \boldsymbol{d}\rangle, \quad \forall s \in [0, 1],$$

which implies

$$D_\phi(\boldsymbol{x}^+, \boldsymbol{x}) \leq \frac{\beta t^2}{2}\langle \nabla^2\phi(\boldsymbol{x})\boldsymbol{d}, \boldsymbol{d}\rangle. \tag{4.7}$$

Note that β is finite because $h(s)$ is also finite-valued. Substituting (4.7) into (4.6), we obtain

$$\begin{aligned}\Psi(\boldsymbol{x}^+) - \Psi(\boldsymbol{x}) &\leq t\left(\langle \nabla f(\boldsymbol{x}), \boldsymbol{d}\rangle + g(\boldsymbol{x}+\boldsymbol{d}) - g(\boldsymbol{x}) + \frac{\beta t L}{2}\langle \nabla^2\phi(\boldsymbol{x})\boldsymbol{d}, \boldsymbol{d}\rangle\right)\\ &\leq t\left(\langle \nabla f(\boldsymbol{x}), \boldsymbol{d}\rangle + g(\boldsymbol{x}+\boldsymbol{d}) - g(\boldsymbol{x}) + \frac{1-\alpha}{2\lambda}\langle \nabla^2\phi(\boldsymbol{x})\boldsymbol{d}, \boldsymbol{d}\rangle\right)\\ &\leq t\,(\rho - (1-\alpha)\rho) = \alpha t\rho,\end{aligned}$$

where the second inequality holds from $t \leq \frac{1-\alpha}{\beta\lambda L}$, and the third inequality follows from the notation $\rho := \langle \nabla f(\boldsymbol{x}), \boldsymbol{d}\rangle + g(\boldsymbol{x}+\boldsymbol{d}) - g(\boldsymbol{x})$ together with using (4.3) to the second term. The desired inequality is proved.

Lemma 17 induces that (4.4) holds after finitely many numbers of backtracking steps. Substituting (4.3) into (4.4), we obtain

$$\Psi(\boldsymbol{x}^+) \leq \Psi(\boldsymbol{x}) - \frac{\alpha t}{2\lambda}\langle \nabla^2\phi(\boldsymbol{x})\boldsymbol{d}, \boldsymbol{d}\rangle, \tag{4.8}$$

which also implies a sufficient decrease in the objective function value Ψ. Inequality (4.8) does not require $0 < \lambda L < 1$ while BPG requires it to prove the sufficient decreasing [5, Lemma 4.1]. We can choose λ larger than $1/L$, that is, $\lambda L \geq 1$, while t might be small in practice as (4.5) implies.

4.2 Global subsequential convergence

Throughout Sects. 4.2 and 4.3, we consider (1.1) with $C \equiv \mathbb{R}^n$. In this subsection, we establish the global subsequential convergence, i.e., any accumulation point of the sequence $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ generated by ABPG is a stationary point. Inspired by Fermat's rule [34, Theorem 10.1], we use the limiting subdifferential and define the stationary point, also called the critical point [5].

Definition 18 We say that $\tilde{\boldsymbol{x}} \in \mathbb{R}^n$ is a stationary point of (1.1) if

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

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

Using [47, Theorem 1.107(ii)], for any $\boldsymbol{x} \in \mathbb{R}^n$, we obtain

$$\partial\Psi(\boldsymbol{x}) \equiv \nabla f(\boldsymbol{x}) + \partial g(\boldsymbol{x}),$$

where $\partial g(\boldsymbol{x})$ is a (classical) subdifferential since g is convex [34, Proposition 8.12].

To prove the global convergence, we additionally make the following assumptions.

Assumption 19 (i) ∇f and $\nabla\phi$ are Lipschitz continuous on any bounded subset of $\mathbb{R}^n$.

(ii) The objective function Ψ is level-bounded, i.e., for any $r \in \mathbb{R}$, lower level sets $\{\boldsymbol{x} \in \mathbb{R}^n \mid \Psi(\boldsymbol{x}) \leq r\}$ are bounded.

Assumption 19(i) is weaker than the global Lipschitz continuity for ∇f. Some researchers [5, 23] make Assumption 19(i). Note that $\nabla\phi$ is locally Lipschitz continuous because ϕ is $\mathcal{C}^2$.

Now, we establish the global subsequential convergence.

Theorem 20 *(Global subsequential convergence of ABPG) Suppose that Assumptions 9, 11, 14, 16, and 19 hold. Let $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ be a sequence generated by ABPG for solving (1.1). Then, the following statements hold:*

(i) *The sequence $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ is bounded.*
(ii) $\lim_{k\to\infty} \|\boldsymbol{d}^k\| = 0$.
(iii) *Any accumulation point of $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ is a stationary point of (1.1).*

Proof (i) Using (4.8) and the positive definiteness of $\nabla^2\phi(\boldsymbol{x}^k)$ due to Assumption 16, we know $\Psi(\boldsymbol{x}^{k+1}) \leq \Psi(\boldsymbol{x}^k)$, which induces $\Psi(\boldsymbol{x}^k) \leq \Psi(\boldsymbol{x}^0)$ for all $k \in \mathbb{N}$. Because of this and Assumption 19(ii), we have the desired result.

(ii) Again using (4.8), we obtain

$$\Psi(\boldsymbol{x}^{k-1}) - \Psi(\boldsymbol{x}^k) \geq \frac{\alpha t_{k-1}}{\lambda}\langle \nabla^2\phi(\boldsymbol{x}^{k-1})\boldsymbol{d}^{k-1}, \boldsymbol{d}^{k-1}\rangle \geq \frac{\alpha\sigma t_{k-1}}{2\lambda}\|\boldsymbol{d}^{k-1}\|^2, \tag{4.9}$$

where the last inequality holds since ϕ is σ-strongly convex. Summing the above inequality from $k = 1$ to ∞, we have

$$\sum_{k=1}^{\infty} \frac{\alpha\sigma t_{k-1}}{2\lambda}\|\boldsymbol{d}^{k-1}\|^2 \leq \Psi(\boldsymbol{x}^0) - \liminf_{N\to\infty} \Psi(\boldsymbol{x}^N) \leq \Psi(\boldsymbol{x}^0) - \Psi^* < \infty,$$

which shows that $\lim_{k\to\infty} \|\boldsymbol{d}^k\| = 0$ because t_k is bounded away from 0 (if necessary, assume that there exists $t_{\min} > 0$ such that $t_{\min} \leq t_k$ for all k).

(iii) Let $\tilde{\boldsymbol{x}} \in \mathbb{R}^n$ be an accumulation point of $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ and let $\{\boldsymbol{x}^{k_j}\}_{j=0}^{\infty}$ be a subsequence such that $\lim_{j\to\infty} \boldsymbol{x}^{k_j} = \tilde{\boldsymbol{x}}$ by Bolzano–Weierstrass theorem. We have $\boldsymbol{d}^k \to \boldsymbol{0}$ because of (ii) and let $\{\boldsymbol{d}^{k_j}\}_{j=0}^{\infty}$ be a subsequence such that $\lim_{j\to\infty} \boldsymbol{d}^{k_j} = \boldsymbol{0}$. We substitute $\boldsymbol{x}^{k_j}$ into the first-order optimality condition of (3.4) to have

$$\boldsymbol{0} \in \nabla f(\boldsymbol{x}^{k_j}) + \partial g(\boldsymbol{x}^{k_j} + \boldsymbol{d}^{k_j}) + \frac{1}{\lambda}\nabla^2\phi(\boldsymbol{x}^{k_j})\boldsymbol{d}^{k_j},$$

which imply

$$-\frac{1}{\lambda}\nabla^2\phi(\boldsymbol{x}^{k_j})\boldsymbol{d}^{k_j} \in \nabla f(\boldsymbol{x}^{k_j}) + \partial g(\boldsymbol{x}^{k_j} + \boldsymbol{d}^{k_j}). \tag{4.10}$$

Since $\nabla\phi$ is Lipschitz continuous on bounded subsets of $\mathbb{R}^n$ and $\{\boldsymbol{x}^{k_j}\}_{j=0}^{\infty}$ is bounded, there exists $A > 0$ such that

$$\left\| \frac{1}{\lambda}\nabla^2\phi(\boldsymbol{x}^{k_j})\boldsymbol{d}^{k_j} \right\| \leq \frac{A}{\lambda}\|\boldsymbol{d}^{k_j}\| \to 0.$$

Taking the limit of (4.10) and using the lower semicontinuity of f and g, we obtain

$$\mathbf{0} \in \nabla f(\tilde{\boldsymbol{x}}) + \partial g(\tilde{\boldsymbol{x}}),$$

which shows that $\tilde{\boldsymbol{x}}$ is a stationary point of (1.1).

Before discussing the global convergence, we consider the objective function value Ψ at a stationary point. Consequently, we show that Ψ is constant on the set of accumulation points of $\{\boldsymbol{x}^{k_j}\}_{j=0}^{\infty}$.

Proposition 21 *Suppose that Assumptions 9, 11, 14, 16, and 19 hold. Let $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ be a sequence generated by ABPG for solving (1.1). Then, the following statements hold:*

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

Proof (i) The sequence $\{\Psi(\boldsymbol{x}^k)\}_{k=0}^{\infty}$ is bounded and non-increasing due to Assumption 9(iv) and $\Psi(\boldsymbol{x}^{k+1}) \leq \Psi(\boldsymbol{x}^k)$ (see the proof of Theorem 20). Therefore, $\zeta := \lim_{k\to\infty} \Psi(\boldsymbol{x}^k)$ exists.

(ii) Let $\tilde{\boldsymbol{x}} \in \Omega$ be any accumulation point of $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ and let $\{\boldsymbol{x}^{k_j}\}_{j=0}^{\infty}$ be a subsequence such that $\lim_{j\to\infty} \boldsymbol{x}^{k_j} = \tilde{\boldsymbol{x}}$. Note that $\tilde{\boldsymbol{x}}$ is a stationary point from Theorem 20(iii). From Assumption 9(ii) and (iii), f and g are continuous on $\operatorname{int}\operatorname{dom}\phi = \mathbb{R}^n$. Therefore, we immediately obtain $\lim_{j\to\infty} \Psi(\boldsymbol{x}^{k_j}) = \Psi(\tilde{\boldsymbol{x}}) = \zeta$. Since $\tilde{\boldsymbol{x}}$ is arbitrary, $\Psi \equiv \zeta$ on Ω.

4.3 Global convergence: special case $g \equiv 0$

In the previous subsection, we have shown that any accumulation point of a sequence generated by ABPG is a stationary point of (1.1). We next establish that the sequence converges to a stationary point of (1.1). From the first-order optimality condition of (3.4), it follows that

$$\mathbf{0} \in \nabla f(\boldsymbol{x}^{k-1}) + \partial g(\boldsymbol{x}^{k-1} + \boldsymbol{d}^{k-1}) + \frac{1}{\lambda}\nabla^2\phi(\boldsymbol{x}^{k-1})\boldsymbol{d}^{k-1}. \tag{4.11}$$

In this subsection, we focus on the special case $g \equiv 0$ because the current problem setting that g may not be $\mathcal{C}^1$ makes it difficult to establish the global convergence using (4.11). If $\nabla^2\phi(\boldsymbol{x}^k)$ is nonsingular and $g \equiv 0$, Subproblem (3.4) is always solved in a closed-form expression, given by

$$\boldsymbol{d}^k = -\lambda\nabla^2\phi(\boldsymbol{x}^k)^{-1}\nabla f(\boldsymbol{x}^k).$$

In this case, ABPG is an approximation of the mirror descent [20].

Theorem 22 *(Global convergence of ABPG) Suppose that Assumptions 9, 11, 14, 16, and 19 hold and that Ψ is a KL function. Let $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ be a sequence generated by ABPG for solving (1.1). Then, the following statements hold when $g \equiv 0$:*

(i) $\lim_{k\to\infty} \operatorname{dist}(\mathbf{0}, \partial\Psi(\boldsymbol{x}^k)) = 0$.
(ii) The sequence $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ converges to a stationary point of (1.1); moreover, $\sum_{k=1}^{\infty} \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\| < \infty$.

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

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

We also have $\Omega \subset \mathcal{X}$ by Theorem 20(iii) and Definition 18. Thus, for any $\varepsilon > 0$, there exists a positive integer $k_0 > 0$ such that $\operatorname{dist}(\boldsymbol{x}^k, \Omega) < \varepsilon$ for all $k \geq k_0$.

The subdifferential of Ψ at $\boldsymbol{x}^k$ for any $k \geq k_0$ is $\partial\Psi(\boldsymbol{x}^k) = \{\nabla f(\boldsymbol{x}^k)\}$ due to $g \equiv 0$. Using Assumption 19(i) and (4.11) with $g \equiv 0$, there exists $A_0 > 0$ such that

$$\begin{aligned}
\|\nabla f(\boldsymbol{x}^k)\| &= \|\nabla f(\boldsymbol{x}^k) - \nabla f(\boldsymbol{x}^{k-1}) + \nabla f(\boldsymbol{x}^{k-1})\| \\
&\leq \|\nabla f(\boldsymbol{x}^k) - \nabla f(\boldsymbol{x}^{k-1})\| + \|\nabla f(\boldsymbol{x}^{k-1})\| \\
&= \|\nabla f(\boldsymbol{x}^k) - \nabla f(\boldsymbol{x}^{k-1})\| + \frac{1}{\lambda}\|\nabla^2\phi(\boldsymbol{x}^{k-1})\boldsymbol{d}^{k-1}\| \\
&\leq A_0\|\boldsymbol{d}^{k-1}\|,
\end{aligned}$$

which is equivalent to

$$\operatorname{dist}(\mathbf{0}, \partial\Psi(\boldsymbol{x}^k)) \leq A_0\|\boldsymbol{d}^{k-1}\|, \tag{4.12}$$

where $k \geq k_0 + 1$. Using Theorem 20(ii), we prove the desired result.

(ii) If there exists an integer $\tilde{k} \geq 0$ for which $\Psi(\boldsymbol{x}^{\tilde{k}}) = \Psi(\tilde{\boldsymbol{x}}) = \zeta$, then (4.9) implies $\boldsymbol{x}^{\tilde{k}+1} = \boldsymbol{x}^{\tilde{k}}$. This means that $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ finitely converges to a stationary point.

We next consider the case in which $\Psi(\boldsymbol{x}^k) > \zeta$ for all $k \geq 0$. Since Proposition 21(ii) and $\{\Psi(\boldsymbol{x}^k)\}_{k=0}^{\infty}$ is non-increasing, it holds that $\Psi(\tilde{\boldsymbol{x}}) < \Psi(\boldsymbol{x}^k)$. Again using Proposition 21(ii), for any $\eta > 0$, there exists a nonnegative integer $k_1 \geq 0$ such that $\Psi(\boldsymbol{x}^k) < \Psi(\tilde{\boldsymbol{x}}) + \eta$ for all $k > k_1$. Using Lemma 8, for any $k > l := \max\{k_0, k_1\}$, we have

$$\psi'(\Psi(\boldsymbol{x}^k) - \Psi(\tilde{\boldsymbol{x}}))\operatorname{dist}(\mathbf{0}, \partial\Psi(\boldsymbol{x}^k)) \geq 1. \tag{4.13}$$

Since ψ is a concave function, it follows that, for any $k \geq l$,

$$\begin{aligned}(\psi(\Psi_k) - \psi(\Psi_{k+1})\,\mathrm{dist}(\mathbf{0}, \partial\Psi(\boldsymbol{x}^k)) &\geq \psi'(\Psi_k)\,\mathrm{dist}(\mathbf{0}, \partial\Psi(\boldsymbol{x}^k))\left(\Psi(\boldsymbol{x}^k) - \Psi(\boldsymbol{x}^{k+1})\right)\\ &\geq \Psi(\boldsymbol{x}^k) - \Psi(\boldsymbol{x}^{k+1})\\ &\geq A_1\|\boldsymbol{d}^k\|^2,\end{aligned}$$

where $\Psi_k := \Psi(\boldsymbol{x}^k) - \zeta$, the second inequality holds due to (4.13), and the last inequality holds due to (4.9). Substituting (4.12) into the above inequality, we have

$$\|\boldsymbol{d}^k\|^2 \leq \frac{A_0}{A_1}(\psi(\Psi_k) - \psi(\Psi_{k+1}))\,\|\boldsymbol{d}^{k-1}\|. \tag{4.14}$$

Taking the square root of (4.14) and using the inequality of arithmetic and geometric means, it holds that

$$\begin{aligned}\|\boldsymbol{d}^k\| &\leq \sqrt{\frac{A_0}{A_1}(\psi(\Psi_k) - \psi(\Psi_{k+1}))}\sqrt{\|\boldsymbol{d}^{k-1}\|}\\ &\leq \frac{A_0}{2A_1}(\psi(\Psi_k) - \psi(\Psi_{k+1})) + \frac{1}{2}\|\boldsymbol{d}^{k-1}\|,\end{aligned}$$

that is,

$$\frac{1}{2}\|\boldsymbol{d}^k\| \leq \frac{A_0}{2A_1}(\psi(\Psi_k) - \psi(\Psi_{k+1})) + \frac{1}{2}\|\boldsymbol{d}^{k-1}\| - \frac{1}{2}\|\boldsymbol{d}^k\|. \tag{4.15}$$

Summing (4.15) from $k = l$ to ∞, we obtain

$$\sum_{k=l}^{\infty}\|\boldsymbol{d}^k\| \leq \frac{A_0}{2A_1}\psi(\Psi_l) + \frac{1}{2}\|\boldsymbol{d}^{l-1}\| < \infty.$$

We have $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| = t_k\|\boldsymbol{d}^k\| \leq \|\boldsymbol{d}^k\|$ due to $t_k \in (0, 1]$ and then know

$$\sum_{k=l}^{\infty}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| \leq \sum_{k=l}^{\infty}\|\boldsymbol{d}^k\| < \infty,$$

which implies that $\sum_{k=1}^{\infty}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| < \infty$, i.e., the sequence $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ is a Cauchy sequence. Consequently, together with Theorem 20(iii), the sequence $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ converges to a stationary point of (1.1).

Finally, we show the rate of convergence result. Its proof is almost the same as the existing one (see, for example, [48, Theorem 2], [40], and [6, Theorem 4]).

Theorem 23 *(Rate of convergence) Suppose that Assumptions 9, 11, 14, 16, and 19 hold. Let $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ be a sequence generated by ABPG for solving (1.1) and let $\tilde{\boldsymbol{x}} \in \mathcal{X}$ be a stationary point of (1.1). Suppose further that Ψ is a KL function with ψ in the KL inequality (2.2) taking form $\psi(s) = cs^{1-\theta}$ for some $\theta \in [0, 1)$ and $c > 0$. Then, the following statements hold when $g \equiv 0$:*

(i) If $\theta = 0$, then the sequence $\{\boldsymbol{x}^k\}_{k=0}^{\infty}$ converges to $\tilde{\boldsymbol{x}}$ in a finite number of steps;
(ii) If $\theta \in (0, 1/2]$, then there exist $c_1 > 0$ and $\eta \in [0, 1)$ such that

$$\|\boldsymbol{x}^k - \tilde{\boldsymbol{x}}\| < c_1 \eta^k;$$

(iii) If $\theta \in (1/2, 1)$, then there exists $c_2 > 0$ such that

$$\|\boldsymbol{x}^k - \tilde{\boldsymbol{x}}\| < c_2 k^{-\frac{1-\theta}{2\theta-1}}.$$

5 Numerical experiments

In this section, we conducted numerical experiments to examine the performance of our proposed algorithm. Note that KL exponents for problems in numerical experiments are unknown. All numerical experiments were performed in Python 3.9 on a MacBook Air with an Apple M2 and 16GB LPDDR5 memory.

5.1 The ℓ_p Regularized Least Squares Problem

We consider the sparse ℓ_p (with p slightly larger than 1) regularized least squares problem [31, 32]:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 + \frac{\theta_p}{p}\|\boldsymbol{x}\|_p^p, \tag{5.1}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{b} \in \mathbb{R}^m$, and $\theta_p > 0$. Let $g \equiv 0$,

$$f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 + \frac{\theta_p}{p}\|\boldsymbol{x}\|_p^p, \quad \text{and} \quad \phi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2 + \frac{1}{p}\|\boldsymbol{x}\|_p^p. \tag{5.2}$$

Note that ∇f and $\nabla \phi$ are not Lipschitz continuous for a general p and that when $p > 1$, ϕ and f are $\mathcal{C}^1$. We have $\nabla \phi(\boldsymbol{x}) = \boldsymbol{x} + \operatorname{sgn}(\boldsymbol{x})|\boldsymbol{x}|^{p-1}$.

Proposition 24 *Let f and ϕ be as defined above. Then, for any $L > 0$ satisfying*

$$L \geq \lambda_{\max}(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}) + \theta_p, \tag{5.3}$$

the functions $L\phi - f$ and $L\phi + f$ are convex on $\mathbb{R}^n$. Therefore, the pair (f, ϕ) is L-smad.

Proof Because f and ϕ are convex, $L\phi + f$ is convex. We have

$$\begin{aligned} L\phi(\boldsymbol{x}) - f(\boldsymbol{x}) &= L\left(\frac{1}{2}\|\boldsymbol{x}\|^2 + \frac{1}{p}\|\boldsymbol{x}\|_p^p\right) - \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 - \frac{\theta_p}{p}\|\boldsymbol{x}\|_p^p \\ &= \frac{L}{2}\|\boldsymbol{x}\|^2 - \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 + \frac{L-\theta_p}{p}\|\boldsymbol{x}\|_p^p \end{aligned}$$

which is convex for any $\boldsymbol{x} \in \mathbb{R}^n$ because $\frac{L}{2}\|\boldsymbol{x}\|^2 - \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2$ is convex due to Lipschitz continuity of the gradient of $\frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2$ with $\lambda_{\max}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A})$ (see, for example, [16, Example 2.2]) and $\frac{L-\theta_p}{p}\|\boldsymbol{x}\|_p^p$ is convex due to $L - \theta_p \geq 0$. Therefore, the pair (f, ϕ) is L-smad.

The optimality condition of (3.1) is equivalent to a $(p-1)$th polynomial equation. The BPG mapping $\mathcal{T}_\lambda(\boldsymbol{x})$ is not computed in a closed-form expression for a general $p > 1$.

Using the above ϕ, we obtain Subproblem (3.4) in a closed form. We have

$$\nabla^2\phi(\boldsymbol{x}) = \boldsymbol{I} + (p-1)\,\mathrm{diag}(|\boldsymbol{x}|^{p-2}).$$

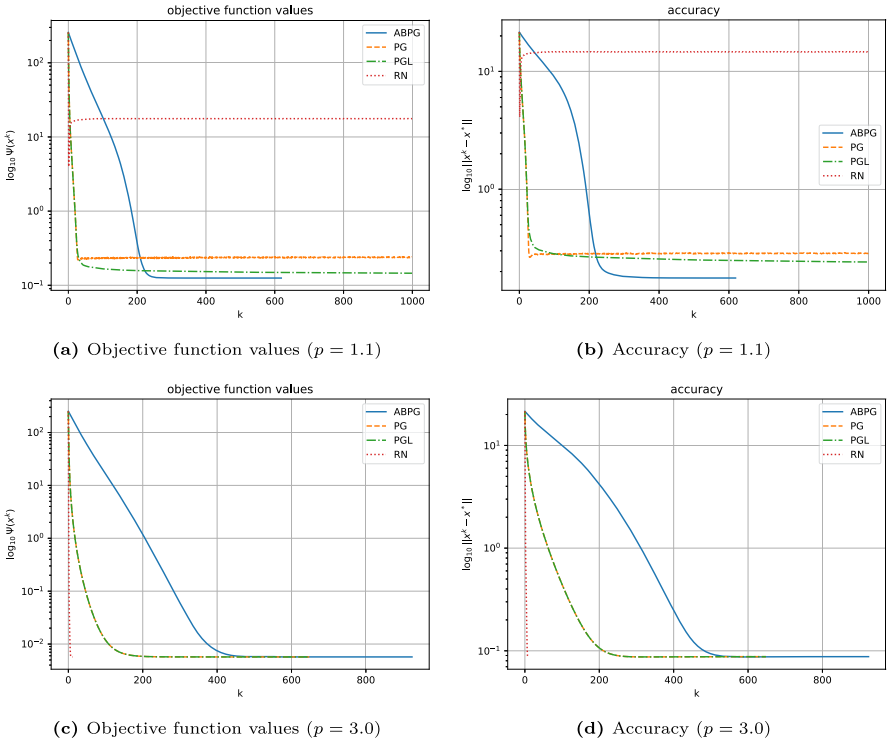
Because the inverse matrix of $\nabla^2\phi(\boldsymbol{x})$ is also a diagonal matrix, Subproblem (3.4) is expressed in a closed form. Note that $\nabla^2\phi(\boldsymbol{x})$ is not continuous on $x_i = 0$ for some $i = 1, \ldots, n$ with $p \in (1, 2)$, while it is continuous on $\mathbb{R}^n$ with $p \geq 2$.

Remark 25 While we handled (5.1) as a case of $g \equiv 0$, we can obtain a closed-form solution of (3.4) even when $g \not\equiv 0$, e.g., $g(\boldsymbol{x}) = \theta_1\|\boldsymbol{x}\|_1$ with $\theta_1 > 0$. For $p > 1$, as a subproblem of (5.1) with the additional term $g(\boldsymbol{x}) = \theta_1\|\boldsymbol{x}\|_1$, we obtain

$$\begin{aligned} \boldsymbol{d}^k &= \left(\operatorname*{argmin}_{x\in\mathbb{R}}\left\{v_i x + \theta_1|x| + \frac{1}{2\lambda}\left(1 + (p-1)|x_i^k|^{p-2}\right)(x - x_i^k)^2\right\}\right)_{i=1}^n - \boldsymbol{x}^k \\ &= \left(\mathrm{soft}_{\theta_1 s_i}\left(x_i^k - s_i v_i\right)\right)_{i=1}^n - \boldsymbol{x}^k, \end{aligned}$$

where v_i is the ith element of $\nabla f(\boldsymbol{x}^k)$, $\mathrm{soft}_s(\cdot)$ is a soft-thresholding operator with $s > 0$, and $s_i = \frac{\lambda}{1+(p-1)|x_i^k|^{p-2}}$. Since $s_i = \frac{\lambda}{1+(p-1)|x_i^k|^{p-2}} = \frac{\lambda|x_i^k|^{2-p}}{|x_i^k|^{2-p}+p-1}$ for $p \in (1, 2)$, $s_i = 0$ when $x_i^k = 0$.

We compared ABPG with the proximal gradient method (PG), the PG with line search (PGL), and the regularized Newton method (RN). We set $\alpha = 0.99$ and $\eta = 0.9$ for ABPG. Although ∇f is not Lipschitz continuous, we used $1/L$ as stepsizes of PG and as an initial stepsize of PGL using the L of the L-smad property given by (5.3). This implies that PG would lose convergence guarantees based on the global-Lipschitz-gradient assumption. We used PGL with the backtracking procedure [10, p.283]. We used RN with the fixed constants of regularization. Note that RN is equivalent to ABPG with $\phi = f + \frac{\kappa}{2}\|\cdot\|^2$. The parameter κ was fixed to $10^{-5} > 0$ in this experiment. We



(a) Objective function values ($p = 1.1$)

(b) Accuracy ($p = 1.1$)

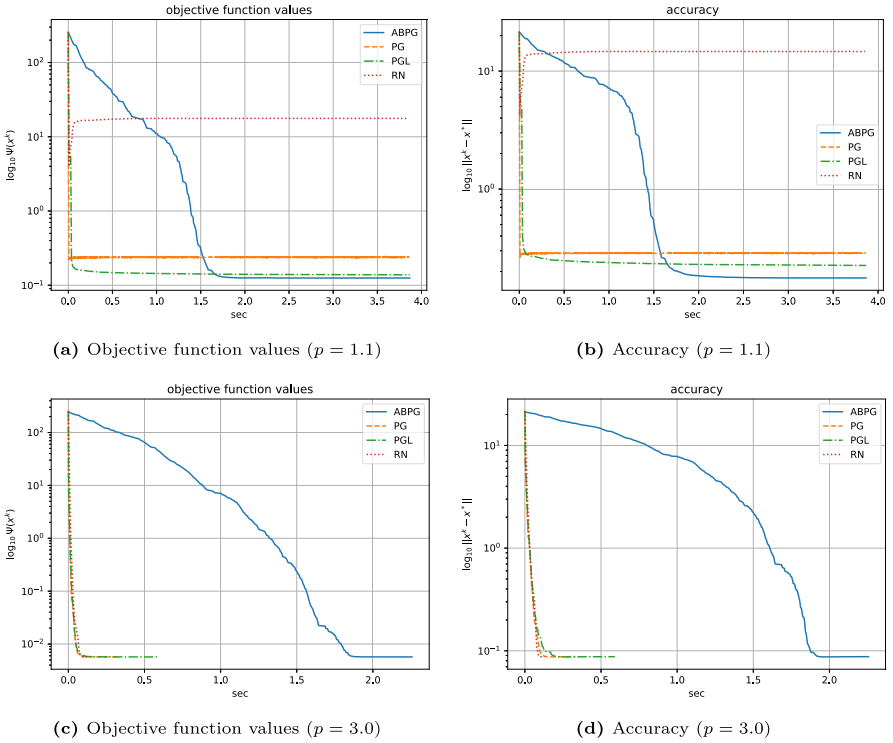
(c) Objective function values ($p = 3.0$)

(d) Accuracy ($p = 3.0$)

Fig. 1 Comparison with ABPG (blue), PG (orange), PGL (green), and RN (red) on the ℓ_p regularized least squares problem (5.1)

generated the initial point $\boldsymbol{x}^0 \in \mathbb{R}^n$ from i.i.d. Gaussian distribution. The maximum iteration is 1000 and the termination condition is $\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\| \leq 10^{-6}$.

We generated the measurement $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and the ground truth $\boldsymbol{x}^*$ from i.i.d. Gaussian distribution, which has 5% nonzero elements. We set $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}^*$. For $(n, m) = (500, 800)$, $p = 1.1, 3.0$, and $\theta_p = 0.05$, Fig. 1 shows the objective function values $\Psi(\boldsymbol{x}^k)$ and the accuracy $\|\boldsymbol{x}^k - \boldsymbol{x}^*\|$ at each iteration on a logarithmic scale. Figure 1a and b show the objective function values and the accuracy at each kth iteration, and Fig. 2a and b show these on time axis. When PG, PGL, and RN did not stop in 1000 iterations, we plotted their performance until the termination time of ABPG. Figures 1a, b, 2a, and b show that ABPG outperforms other algorithms when $p = 1.1$. Only ABPG stopped within 1000 iterations (around 600 iterations). PG and RN did not decrease the objective function value at every iteration; the reason is probably that ∇f is not globally Lipschitz continuous or even locally Lipschitz continuous, more specifically, for $p = 1.1$, $\|\boldsymbol{x}\|_p^p$ does not have any Lipschitz continuous gradients on $(-1, 1)^n$. Stepsizes of PGL were small because PGL requires many line search procedures due to lacking the global Lipschitz continuity of ∇f. For $p = 3.0$, Fig. 1c and d show that ABPG is outperformed by other algorithms. All algorithms stopped within 1000 iterations. The objective function value and the accuracy of all algorithms at the termination were almost the same. In this case, the backtracking procedure of



(a) Objective function values ($p = 1.1$)

(b) Accuracy ($p = 1.1$)

(c) Objective function values ($p = 3.0$)

(d) Accuracy ($p = 3.0$)

Fig. 2 Comparison with ABPG (blue), PG (orange), PGL (green), and RN (red) on the ℓ_p regularized least squares problem (5.1)

PGL was not used. For $p \geq 2$, $\|\boldsymbol{x}\|_p^p$ has Lipschitz continuous gradients on $(-1, 1)^n$ and the backtracking procedure of PGL is not needed on $(-1, 1)^n$. Therefore, PG and PGL outperformed ABPG.

Moreover, Tables 3 and 4 show average performance over 50 random instances for $p = 1.1$, $m = 1000, 2000$, and $n = 100, 200, 500, 1000$. Only ABPG stopped within 1000 iterations and outperformed other algorithms.

5.2 The ℓ_p Regularized least squares problem with equality constraints

We next consider the ℓ_p regularized least squares problem with an equality constraint [49]:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 + \frac{\theta_p}{p}\|\boldsymbol{x}\|_p^p + \delta_S(\boldsymbol{x}), \tag{5.4}$$

where δ_S denotes the indicator function of $S = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{a}^\top \boldsymbol{x} = \gamma\}$ with $\boldsymbol{a} \in \mathbb{R}^n$ and $\gamma \in \mathbb{R}$. Let $g(\boldsymbol{x}) = \delta_S(\boldsymbol{x})$. We use the same definition as (5.2) for f and ϕ.

Let us consider Subproblem (3.4) for solving (5.4). The feasible points $\boldsymbol{x}^k$ need to satisfy $\boldsymbol{a}^\top \boldsymbol{x}^k = \gamma$, which implies that $\boldsymbol{a}^\top \boldsymbol{x}^k = \boldsymbol{a}^\top (\boldsymbol{x}^{k-1} + t_k \boldsymbol{d}^k) = \gamma + t_k \boldsymbol{a}^\top \boldsymbol{d}^{k-1} = \gamma$.

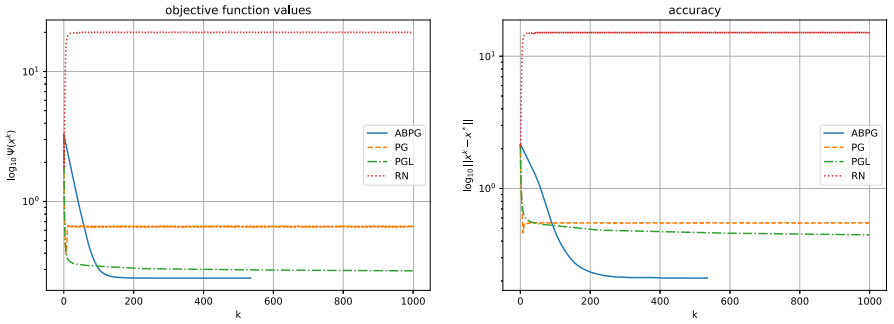
Table 3 Performance results over 50 random instances for $m = 1000$

m	n	Algorithm	Iteration	Obj.	Acc.
1000	100	ABPG	**554**	**0.07502**	**0.09667**
		PG	1000	0.12554	0.17612
		PGL	970	0.07863	0.12067
		RN	1000	0.23388	0.39102
	200	ABPG	**580**	**0.09696**	**0.12952**
		PG	1000	0.16969	0.21781
		PGL	1000	0.10451	0.16348
		RN	1000	0.68147	0.94434
	500	ABPG	**619**	**0.13928**	**0.19568**
		PG	1000	0.26540	0.31184
		PGL	1000	0.15927	0.26231
		RN	1000	7.66003	6.69885
	1000	ABPG	**652**	**0.17662**	**0.26180**
		PG	1000	0.34935	0.38783
		PGL	1000	0.22393	0.40126
		RN	1000	—	—

Numbers in boldface represent the best value among algorithms. This is also used for Table 4. "—" means that its value is too large

Table 4 Performance results over 50 random instances for $m = 2000$

m	n	Algorithm	Iteration	Obj.	Acc.
2000	100	ABPG	**558**	**0.07333**	**0.09635**
		PG	1000	0.13971	0.20044
		PGL	939	0.07660	0.11364
		RN	1000	0.20294	0.32129
	200	ABPG	**575**	**0.09443**	**0.12819**
		PG	1000	0.19450	0.24123
		PGL	1000	0.10177	0.15410
		RN	1000	0.43165	0.57237
	500	ABPG	**602**	**0.13678**	**0.19227**
		PG	1000	0.29906	0.32302
		PGL	1000	0.15525	0.23809
		RN	1000	2.06468	1.90547
	1000	ABPG	**631**	**0.17750**	**0.25676**
		PG	1000	0.42798	0.41943
		PGL	1000	0.21516	0.33322
		RN	1000	15.67614	9.69705



(a) Objective function values ($p = 1.1$) **(b)** Accuracy ($p = 1.1$)

Fig. 3 Comparison with ABPG (blue), PG (orange), PGL (green), and RN (red) on the ℓ_p regularized least squares problem with an equality constraint (5.4)

Therefore, the directions $\boldsymbol{d}^k$ satisfy $\boldsymbol{a}^\mathsf{T}\boldsymbol{d}^k = 0$ for all $k \geq 0$. We have the optimality condition of (3.4),

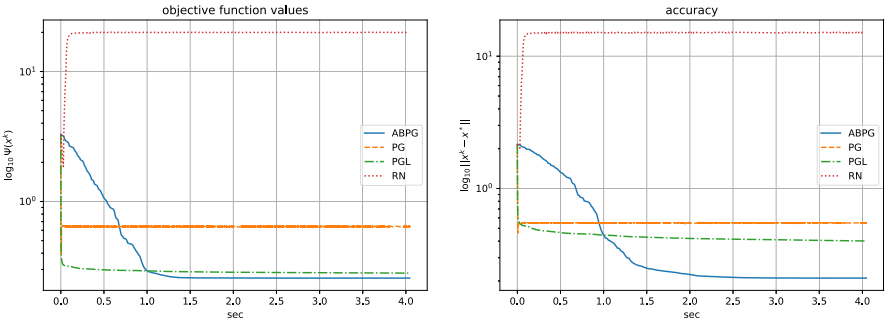
$$\begin{bmatrix} \nabla^2\phi(\boldsymbol{x}^k) & \boldsymbol{a} \\ \boldsymbol{a}^\mathsf{T} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{d}^k \\ \mu \end{bmatrix} = \begin{bmatrix} -\lambda\nabla f(\boldsymbol{x}^k) \\ 0 \end{bmatrix},$$

where $\mu \in \mathbb{R}$. Let $\delta = -\boldsymbol{a}^\mathsf{T}\nabla^2\phi(\boldsymbol{x}^k)^{-1}\boldsymbol{a} \neq 0$. We solve the linear system and have

$$\begin{aligned} \begin{bmatrix} \boldsymbol{d}^k \\ \mu \end{bmatrix} &= \begin{bmatrix} \nabla^2\phi(\boldsymbol{x}^k)^{-1} + \delta^{-1}\boldsymbol{u}\boldsymbol{u}^\mathsf{T} & -\delta^{-1}\boldsymbol{u} \\ -\delta^{-1}\boldsymbol{u}^\mathsf{T} & \delta^{-1} \end{bmatrix} \begin{bmatrix} -\lambda\nabla f(\boldsymbol{x}^k) \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} -\lambda\nabla^2\phi(\boldsymbol{x}^k)^{-1}\nabla f(\boldsymbol{x}^k) - \delta^{-1}\lambda\boldsymbol{u}\boldsymbol{u}^\mathsf{T}\nabla f(\boldsymbol{x}^k) \\ \delta^{-1}\lambda\boldsymbol{u}^\mathsf{T}\nabla f(\boldsymbol{x}^k) \end{bmatrix}, \end{aligned}$$

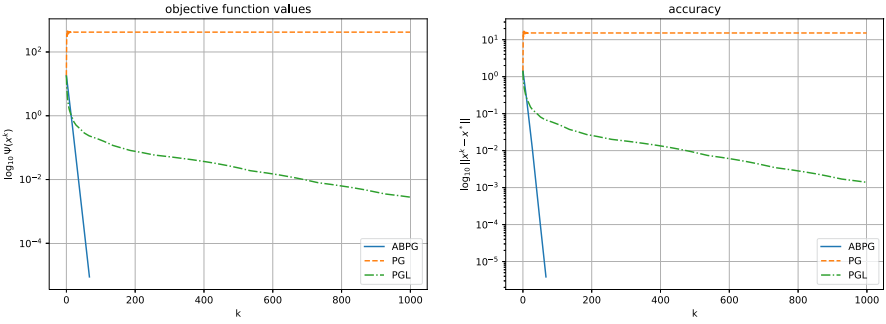
where $\boldsymbol{u} = \nabla^2\phi(\boldsymbol{x}^k)^{-1}\boldsymbol{a}$. While we consider the case of one linear equality constraint, the above result can be extended to the case of multiple linear equality constraints. Set $\nabla^2\phi(\boldsymbol{x}^k) = \boldsymbol{I}$ for the proximal gradient method and $\nabla^2\phi(\boldsymbol{x}^k) = \kappa\boldsymbol{I} + \nabla f(\boldsymbol{x}^k)$ for the regularized Newton method with a fixed parameter $\kappa > 0$.

We generated $\boldsymbol{x}^0 \in S$ and set $\boldsymbol{a} = \boldsymbol{1}$ and $\gamma = 1$. The other setting for parameters and the termination condition is the same as Sect. 5.1. For $(n, m) = (500, 800)$, $p = 1.1$ and $\theta_p = 0.05$, Figs. 3 and 4 show that the objective function values $\Psi(\boldsymbol{x}^k)$ and the accuracy $\|\boldsymbol{x}^k - \boldsymbol{x}^*\|$ for (5.4) on a logarithmic scale. ABPG outperformed other algorithms on the objective function values and the accuracy (Fig. 3a and b for the iteration axis and Figures 4a and b for the time axis). PG, PGL, and RN did not stop within 1000 iterations. The reason for this is the same reason for (5.1). PG did not decrease the objective function value due to lacking the global Lipschitz continuity for ∇f. The line search procedure of PGL did not work well. RN also did not work well because of ill-conditioned inverse matrices.



(a) Objective function values ($p = 1.1$) (b) Accuracy ($p = 1.1$)

Fig. 4 Comparison with ABPG (blue), PG (orange), PGL (green), and RN (red) on the ℓ_p regularized least squares problem with an equality constraint (5.4)



(a) Objective function values (b) Accuracy

Fig. 5 Comparison with ABPG (blue), PG (orange), and PGL (green) on the ℓ_p loss problem (5.5)

5.3 The ℓ_p loss problem

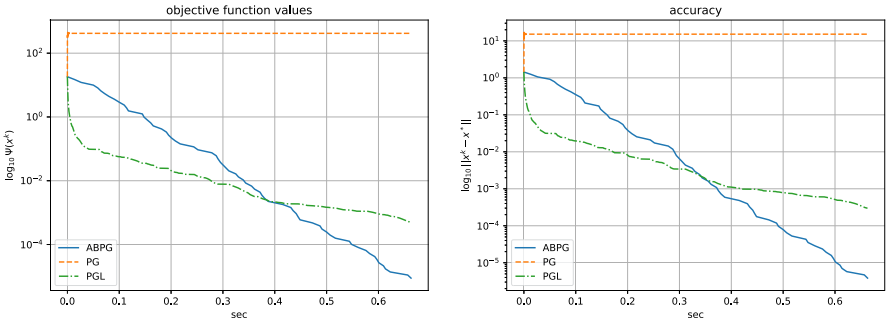
We consider the ℓ_p loss problem [33]:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \quad \frac{1}{p}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_p^p, \tag{5.5}$$

where $\boldsymbol{A}\in\mathbb{R}^{m\times n}$, $\boldsymbol{b}\in\mathbb{R}^m$, and $p>1$. Let $f(\boldsymbol{x})=\frac{1}{p}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_p^p$ and $g\equiv 0$. We have

$$\nabla f(\boldsymbol{x})=\sum_{i=1}^m |\boldsymbol{a}_i^\mathsf{T}\boldsymbol{x}-b_i|^{p-1}\operatorname{sgn}(\boldsymbol{a}_i^\mathsf{T}\boldsymbol{x}-b_i)\boldsymbol{a}_i,$$

which is continuous with $p>1$. We also have $\nabla^2 f(\boldsymbol{x})=(p-1)\sum_{i=1}^m |\boldsymbol{a}_i^\mathsf{T}\boldsymbol{x}-b_i|^{p-2}\boldsymbol{a}_i\boldsymbol{a}_i^\mathsf{T}$. Using $\phi=f+\frac{1}{2}\|\cdot\|^2$, the pair (f,ϕ) is 1-smad and $\nabla^2\phi$ is nonsingular. Moreover, ABPG is equivalent to RN in this case. We compare ABPG (RN) with PG and PGL.



(a) Objective function values (b) Accuracy

Fig. 6 Comparison with ABPG (blue), PG (orange), and PGL (green) on the ℓ_p loss problem (5.5)

We generated the ground truth $\boldsymbol{x}^*$ from i.i.d. Gaussian distribution, which has 10% nonzero elements and the initial point $\boldsymbol{x}^0$ by using the Wirtinger flow initialization [50]. For $(n, m) = (200, 500)$ and $p = 1.1$, Figure 5 shows that the objective function values $\Psi(\boldsymbol{x}^k)$ and the accuracy $\|\boldsymbol{x}^k - \boldsymbol{x}^*\|$ for (5.5) at each iteration on a logarithmic scale. On the other hand, Fig. 6 shows these performance results on the time axis. In this case, PG did not decrease the objective function value (Figs. 5a and 6a). PG and PGL did not stop within 1000 iterations, probably because ∇f is not globally Lipschitz continuous. ABPG outperformed PG and PGL.

5.4 The nonnegative linear system on $\mathbb{R}^n_+$

Finally, we consider the nonnegative linear system [26] and minimize

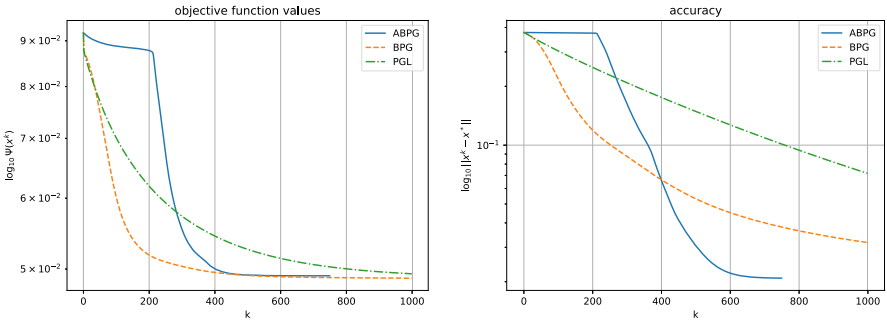
$$\min_{\boldsymbol{x}\in\mathbb{R}^n_+} \quad D_{\mathrm{KL}}(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{b}) + \theta_1\|\boldsymbol{x}\|_1, \tag{5.6}$$

where

$$D_{\mathrm{KL}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{m}\left(x_i \log\frac{x_i}{y_i} + y_i - x_i\right) \tag{5.7}$$

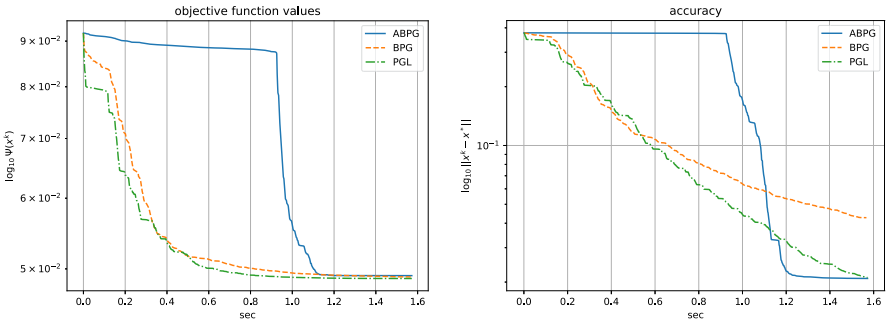
is Kullback–Leibler divergence, $\boldsymbol{A} \in \mathbb{R}^{m\times n}_+$, $\boldsymbol{b} \in \mathbb{R}^m_+$, and $\theta_1 > 0$. Let $f(\boldsymbol{x}) = D_{\mathrm{KL}}(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{b})$ and $g(\boldsymbol{x}) = \theta_1\|\boldsymbol{x}\|_1$. Let $\phi_{\mathrm{BPG}}(\boldsymbol{x}) = \sum_{i=1}^n x_i \log x_i$ be a kernel generating distance for BPG and $\phi_{\mathrm{ABPG}}(\boldsymbol{x}) = \phi_{\mathrm{BPG}}(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x}\|^2$ be that for ABPG. Therefore, we use $C = \operatorname{int}\operatorname{dom}\phi_{\mathrm{BPG}} = \operatorname{int}\operatorname{dom}\phi_{\mathrm{ABPG}} = \mathbb{R}^n_+$. Assuming $\sum_{i=1}^m a_{ij} = 1$ for any j (this assumption is general for applications, for example, positron emission tomography [51]), we find that (f, ϕ_{BPG}) is 1-smad [26] and $(f, \phi_{\mathrm{ABPG}})$ is also 1-smad.

We compare ABPG with BPG to demonstrate the difference in performance by the approximate Bregman distance. The subproblems of BPG and ABPG for (5.6) are expressed in a closed form. We also compare ABPG with PGL.



(a) Objective function values **(b)** Accuracy

Fig. 7 Comparison with ABPG (blue), BPG (orange), and PGL (green) on the nonnegative linear system inverse problem (5.6)



(a) Objective function values **(b)** Accuracy

Fig. 8 Comparison with ABPG (blue), BPG (orange), and PGL (green) on the nonnegative linear system inverse problem (5.6)

We generated the ground truth $\boldsymbol{x}^*$ from i.i.d. Gaussian distribution, which has 5% nonzero elements. For $(n, m) = (200, 500)$ and $\theta_1 = 0.05$, Fig. 7 shows that the objective function values $\Psi(\boldsymbol{x}^k)$ and the accuracy $\|\boldsymbol{x}^k - \boldsymbol{x}^*\|$ for (5.6) at each iteration on a logarithmic scale. Figure 8 shows these performance results on the time axis. In this setting, while PG and BPG did not stop within 1000 iterations, the objective function value of BPG at the 1000th iteration was the best among these algorithms (Fig. 7a). Only ABPG stopped within 1000 iterations and recovered the best point (Fig. 7b). Figure 8b shows PGL (more than the maximum iteration) also recovered a point at which accuracy is almost the same as the point recovered by ABPG.

6 Conclusion

In this paper, we propose the approximate Bregman proximal gradient algorithm (ABPG) for composite nonconvex optimization problems. ABPG does not require the global Lipschitz continuity for ∇f. Its subproblem of ABPG is easy to solve due

to the approximate Bregman distance. In particular, when ϕ is strongly convex and $g \equiv 0$, the subproblem is always expressed in a closed form. Moreover, ABPG coincides with existing algorithms by choosing some particular ϕ. This implies that we can discuss theoretical aspects of existing algorithms by using the ABPG framework. We have established the global subsequential convergence with some standard assumptions. Then, we have guaranteed the global convergence to a stationary point under the KL property when $g \equiv 0$. We also applied ABPG to the ℓ_p regularized least squares problems with some cases, the ℓ_p loss problem, and the nonnegative linear system.

For large-scale optimization problems, ABPG would be very slow due to the line search procedure. It is important for future work to develop the acceleration of ABPG. Recently, some researchers [22, 52, 53] evaluated the iteration complexity of Bregman-type algorithms. Because ABPG uses the approximate Bregman distance, it would be challenging to evaluate the iteration complexity of ABPG. We would like to resolve this matter to evaluate it in the future. Even if $g \not\equiv 0$ is separable, the global convergence would not be established due to the nonsmoothness of g. This might be resolved by improving the line search procedure.

Funding Open Access funding provided by The University of Tokyo. This work was supported by JSPS KAKENHI Grant Number JP23K19953; JSPS KAKENHI Grant Number JP23H03351 and JST ERATO Grant Number JPMJER1903.

Data Availability The datasets generated during and/or analyzed during the current study are available in the Github repository, https://github.com/ShotaTakahashi/ApproximateBPG.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Bouman, C., Sauer, K.: A generalized Gaussian image model for edge-preserving MAP estimation. IEEE Trans. Image Process **2**(3), 296–310 (1993)
2. Elad, M.: Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing. Springer, Berlin (2010)
3. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn., p. 10. Springer, Berlin (2009)
4. Tibshirani, R.: Regression shrinkage and selection via the lasso: a retrospective. J. Stat. Soc. Ser. B Stat. Methodol. **58**(1), 267–288 (1996)
5. Bolte, J., Sabach, S., Teboulle, M., Vaisbourd, Y.: First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. SIAM J. Opt. **28**(3), 2131–2151 (2018)

6. Takahashi, S., Fukuda, M., Tanaka, M.: New Bregman proximal type algorithms for solving DC optimization problems. Comput. Opt. Appl. **83**(3), 893–931 (2022)
7. Chan, T.F., Wong, C.K.: Convergence of the alternating minimization algorithm for blind deconvolution. Linear Algebra Appl. **316**(1–3), 259–285 (2000)
8. Li, X., Ling, S., Strohmer, T., Wei, K.: Rapid, robust, and reliable blind deconvolution via nonconvex optimization. Appl. Comput. Harmon. Anal. **47**(3), 893–934 (2019)
9. Takahashi, S., Tanaka, M., Ikeda, S.: Blind deconvolution with non-smooth regularization via Bregman proximal DCAs. Signal Processing **202**, 108734 (2023)
10. Beck, A.: First-Order Methods in Optimization. MOS-SIAM Series on Optimization, vol. 25. SIAM, Philadelphia (2017)
11. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control Opt. **14**(5), 877–898 (1976)
12. Fukushima, M., Mine, H.: A generalized proximal point algorithm for certain non-convex minimization problems. Int. J. Syst. Sci. **12**(8), 989–1000 (1981)
13. Bruck, R.E.: An iterative solution of a variational inequality for certain monotone operators in Hilbert space. Bull. Am. Math. Soc. **81**(5), 890–892 (1975)
14. Lions, P.L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16**(6), 964–979 (1979)
15. Passty, G.B.: Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. J. Math. Anal. Appl. **72**, 383–390 (1979)
16. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
17. Patriksson, M.: Nonlinear Programming and Variational Inequality Problems: A Unified Approach, p. 102. Springer, Berlin (1999)
18. Becker, S., Fadili, J.: A quasi-Newton proximal splitting method. Adv. Neural Inf. Process. Syst. **25**, 12 (2012)
19. Lee, J.D., Sun, Y., Saunders, M.A.: Proximal Newton-type methods for minimizing composite functions. SIAM J. Opt. **24**(3), 1420–1443 (2014)
20. Nemirovski, A.S., Yudin, D.B.: Problem Complexity and Method Efficiency in Optimization. Wiley Series in Discrete Mathematics. Wiley, New York (1983)
21. Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. Comput. Math. Math. Phys. **7**(3), 200–217 (1967)
22. Ding, K., Li, J., Toh, K.C.: Nonconvex stochastic Bregman proximal gradient method with application to deep learning. arXiv:2306.14522 (2023) [math.OC]
23. Gao, X., Cai, X., Wang, X., Han, D.: An alternating structure-adapted Bregman proximal gradient descent algorithm for constrained nonconvex nonsmooth optimization problems and its inertial variant. J. Global Opt. **87**(1), 277–300 (2023)
24. Hanzely, F., Richtárik, P., Xiao, L.: Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. Comput. Opt. Appl. **79**(2), 405–440 (2021)
25. Lu, H., Freund, R.M., Nesterov, Y.: Relatively smooth convex optimization by first-order methods, and applications. SIAM J. Opt. **28**(1), 333–354 (2018)
26. Bauschke, H.H., Bolte, J., Teboulle, M.: A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. Math. Oper. Res. **42**(2), 330–348 (2017)
27. Dragomir, R.A., d'Aspremont, A., Bolte, J.: Quartic first-order methods for low-rank minimization. J. Opt. Theory Appl. **189**, 341–363 (2021)
28. Takahashi, S., Tanaka, M., Ikeda, S.: Majorization-minimization Bregman proximal gradient algorithms for nonnegative matrix factorization with the Kullback–Leibler divergence. arXiv:2405.11185 (2024) [math.OC]
29. Yue, M.C., Zhou, Z., So, A.M.: A family of inexact SQA methods for non-smooth convex minimization with provable convergence guarantees based on the Luo-Tseng error bound property. Math. Progr. **174**(1), 327–358 (2019)
30. Li, D.H., Fukushima, M., Qi, L., Yamashita, N.: Regularized Newton methods for convex minimization problems with singular solutions. Comput. Opt. Appl. **28**(2), 131–147 (2004)
31. Chung, J., Gazzola, S.: Flexible Krylov methods for ℓ_p regularization. SIAM J. Sci. Comput. **41**(5), 149–171 (2019)

32. Wen, F., Liu, P., Liu, Y., Qiu, R.C., Yu, W.: Robust sparse recovery for compressive sensing in impulsive noise using ℓ_p-norm model fitting. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4643–4647 (2016)
33. Maddison, C.J., Paulin, D., Teh, Y.W., Doucet, A.: Dual space preconditioning for gradient descent. SIAM J. Opt. **31**(1), 991–1016 (2021)
34. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis. Grundlehren der mathematischen Wissenschaften, vol. 317. Springer, Berlin (1997)
35. Mordukhovich, B.S.: Variational Analysis and Applications. Springer Monographs in Mathematics. Springer, Berlin (2018)
36. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**(1), 79–86 (1951)
37. Itakura, F., Saito, S.: Analysis synthesis telephony based on the maximum likelihood method. In: Proceedings of the 6th International Congress on Acoustics (1968)
38. Bauschke, H.H., Borwein, J.M.: Legendre functions and the method of random Bregman projections. J. Convex Anal. **4**(1), 27–67 (1997)
39. Dhillon, I.S., Tropp, J.A.: Matrix nearness problems with Bregman divergences. SIAM J. Matrix Anal. Appl. **29**(4), 1120–1146 (2008)
40. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality. Math. Oper. Res. **35**(2), 438–457 (2010)
41. Kurdyka, K.: On gradients of functions definable in o-minimal structures. Ann. Inst. Fourier **48**(3), 769–783 (1998)
42. Łojasiewicz, S.: Une propriété topologique des sous-ensembles analytiques réels. Les équations aux dérivées partielles (1963)
43. Li, G., Pong, T.K.: Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods. Found. Comut. Math. **18**, 1199–1232 (2018)
44. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Progr. **146**, 459–494 (2014)
45. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. CMS Books in Mathematics, 2nd edn. Springer, Berlin (2017)
46. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer Series in Operations Research and Financial Engineering, 2nd edn. Springer, Berlin (2006)
47. Mordukhovich, B.S.: Variational Analysis and Generalized Differentiation I: Basic Theory. Grundlehren der mathematischen Wissenschaften, vol. 330. Springer, Berlin (2006)
48. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Program. **116**(1), 5–16 (2009)
49. Ni, P., Walker, H.: A linearly constrained least-squares problem in electronic structure computations. Int. Conf. Computat. Exper. Eng. Sci. **7**(1), 43–50 (2008)
50. Candès, E.J., Li, X., Soltanolkotabi, M.: Phase retrieval via Wirtinger flow: theory and algorithms. IEEE Trans. Inf. Theory **61**(4), 1985–2007 (2015)
51. Vardi, Y., Shepp, L.A., Kaufman, L.: A statistical model for positron emission tomography. J. Am. Stat. Assoc. **80**(389), 8–20 (1985)
52. Dragomir, R.A., Taylor, A.B., d'Aspremont, A., Bolte, J.: Optimal complexity and certification of Bregman first-order methods. Math. Program. **194**(1), 41–83 (2022)
53. Yang, L., Toh, K.C.: Bregman proximal point algorithm revisited: a new inexact version and its inertial variant. SIAM J. Opt. **32**(3), 1523–1554 (2022)

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