Revisiting Linearized Bregman Iterations under Lipschitz-like
Convexity Condition

Hui Zhang∗ Lu Zhang† Hao-Xing Yang‡
March 7, 2022

Abstract
The linearized Bregman iterations (LBrel) and its variants have received considerable attention in signal/image processing and compressed sensing. Recently, LBrel has been extended to a larger class of nonconvex functions, along with several theoretical issues left for further investigation. In particular, the gradient Lipschitz continuity assumption precludes its use in many practical applications. In this study, we propose a generalized algorithmic framework to unify LBrel-type methods. Our main discovery is that the gradient Lipschitz continuity assumption can be replaced by a Lipschitz-like convexity condition in both convex and nonconvex cases. The proposed framework and theory are then applied to linear/quadratic inverse problems.

Keywords. Bregman distance, linearized Bregman iterations, Lipschitz-like convexity condition, linear inverse problem, quadratic inverse problem, regularization, nonconvex minimization, Kurdyka-Lojasiewicz property
AMS subject classifications. 49M37,65K05,90C25,90C26,90C30

1 Introduction
Let \( x_0 \in \mathbb{R}^n \) be an unknown vector which stands for a signal or an image. Assume that we obtain some linear measurements of \( x_0 \) but the number of measurements is much less than the dimension of \( x_0 \). Suppose that the measure procedure and data are modeled by a matrix \( A \in \mathbb{R}^{m \times n} \) and a vector \( b \in \mathbb{R}^m \). Can we recover \( x_0 \) from the measure data \( b \) such that \( Ax_0 \simeq b \)?

The question above lies in the research field of linear inverse problems that subsume a great number of important applications such as signal denoising/deblurring and compressed sensing. In the case of \( m \ll n \), the reconstruction of a general vector \( x_0 \) from \( b \) is impossible even though the data \( b \) is not contaminated since the linear system \( Ax = b \) is underdetermined. Fortunately, the vector \( x_0 \) in many assignments has certain structure such as sparsity and low-rankness in compressed sensing. If we know the prior structure of \( x_0 \) beforehand, then regularization techniques can help us find the desired \( x_0 \) from the solution set of \( Ax = b \). For example, if we previously know that many

∗Corresponding author. Department of Mathematics, National University of Defense Technology, Changsha, Hunan 410073, China. Email: h.zhang1984@163.com
†Department of Mathematics, National University of Defense Technology, Changsha, Hunan 410073, China.
‡Department of Mathematics, National University of Defense Technology, Changsha, Hunan 410073, China.
entries of \( x_0 \) are zero (in other words, \( x_0 \) is sparse), then one can use the \( \ell_1 \) norm as a regularizer to approach the sparse \( x_0 \) by solving

\[
\min_x \|x\|_1 := \sum_{i=1}^n |x_i|, \quad \text{subject to } Ax = b, \tag{1.1}
\]

which is the well-known basis pursuit pursued in the field of signal processing [15]. More generally, one may need to consider the optimization problem in the following form

\[
\min_x \{E(x) + \mu R(x)\}, \tag{1.2}
\]

where \( E(x) := E(Ax, b) \) is chosen to quantify the “error” between \( Ax \) and \( b \), \( R(x) \) is an appropriate regularizer used to reflect the desired feature of \( x_0 \), and \( \mu > 0 \) is a penalty parameter playing the role of controlling the trade-off between the data fidelity and the degree of regularization. In order to obtain further improvement in variational image restoration, the Bregman iteration method related to (1.2), introduced in [24], consists of iteratively solving the following problems

\[
x^{k+1} := \arg \min_x \{E(x) + \mu D_R^p (x, x^k)\}, \tag{1.3}
\]

where \( p^k \) is a subgradient of \( R \) at \( x^k \), i.e., \( p^k \in \partial R(x^k) \), and \( D_R^p (x, x^k) = R(x) - R(x^k) - \langle p^k, x - x^k \rangle \) is the Bregman distance of \( R \) between \( x \) and \( x^k \). Compared with (1.2), the generalized Bregman distance of \( R \) between \( x \) and \( x^k \) replaces the term \( R(x) \) to play the role of regularization. If we start with an initial value \( x^0 \) such that \( p^0 = 0 \in \partial R(x^0) \), then the first step of the Bregman iteration method (1.3) is exactly (1.2) since \( D_R^0 (x, x^0) = R(x) - R(x^0) \). The second and consecutive steps of the Bregman iteration method aim to generate better and better reconstructions. There are a huge amount of research results, both experimentally and analytically, confirming the superior properties of the Bregman iterations; see for example [5, 10, 22, 24, 29, 35].

### 1.1 Linearized Bregman iterations

The linearized Bregman iterations (LBreI) method, suggested by Darbon and Osher (2007) and formally introduced in the influential work [31], replaces the data fidelity term \( E(x) \) by the sum of its linearized approximation at \( x^k \) and a Euclidean norm proximity term \( \frac{1}{2\delta} \|x - x^k\|^2 \) in (1.3). Mathematically, it iteratively solves the following subproblems

\[
x^{k+1} := \arg \min_x \{\langle \nabla E(x^k), x - x^k \rangle + \frac{1}{2\delta} \|x - x^k\|^2 + \mu D_R^p (x, x^k)\}, \tag{1.4}
\]

After a rearrangement of the terms, the LBreI can be equivalently written in the following form

\[
x^{k+1} := \arg \min_x \{R(x) + \frac{1}{2\delta\mu} \|x - (x^k - (\delta \nabla E(x^k) - \delta p^k))\|^2\}, \tag{1.5}
\]

which can be solved easily in many cases or even in a closed-form such as in the case of \( R(\cdot) = \|\cdot\|_1 \). Preliminary convergence results of the LBreI was given in [31] by assuming that \( R \in C^2 \) is strictly convex over a compact set \( \Omega \supset \{x^k\} \). Further convergence properties on the LBreI was deduced in [31] under the assumption that the convex function \( R(\cdot) \) is continuously differentiable and satisfies some Lipschitz property. However, the simple case of \( R(\cdot) = \|\cdot\|_1 \), which is a central tool in compressed sensing, fails to satisfy the required assumption. Fortunately, this assumption can be dropped, as shown in the following elegant convergence result.
Theorem 1.1 (13). Assume that the matrix $A$ is full-rank and $0 < \delta < \frac{1}{\|AA^T\|}$. Suppose that $R(x)$ is convex and $\partial R(x)$ is bounded. Then for any fixed $\mu > 0$, the sequence $\{x^k\}$ generated by (1.5) with $E(x) = \frac{1}{2}\|Ax - b\|^2$ converges to the unique solution of the following problem:

$$\text{minimize} \quad \{\mu R(x) + \frac{1}{2\delta}\|x\|^2 : Ax = b\}. \quad (1.6)$$

As the parameter $\mu$ tends to infinity, the term $\frac{1}{2\delta}\|x\|^2$ becomes more and more trivial and finally loses its effect. Actually, this point was clarified in [13] by showing that the unique solution of (1.6), denoted by $x_{\mu}$, tends to a solution of minimizing $R(x)$ subject to $Ax = b$ as $\mu \to +\infty$; a refiner analysis was done in [30] where the parameter $\mu$ only requires to be greater than a certain value. By phrasing (1.6) as a split feasibility problem and using the concept of Bregman projections, a new deduction of the LBreI and its several new variants were provided in [19]. In order to show global linear convergence of the LBreI, the restricted strongly convex inequality was introduced in [18], which was further developed in [34] for accelerated LBreI variants. The first unified framework that unifies the LBreI and its matrix variant—the singular value thresholding algorithm [10] was proposed in [32] by introducing augmented convex signal recovery models and gauge regularized functions. In order to deal with image deblurring in tight frame domains, the LBreI was modified in [12] to solve the following problem

$$\text{minimize} \quad \{\mu R(x) + \frac{1}{2\delta}\|x\|^2 : x \in \arg \min \frac{1}{2}\|Az - b\|^2\}. \quad (1.7)$$

It should be pointed out that all these mentioned works only consider the special fidelity term $E(x) = \frac{1}{2}\|Ax - b\|^2$. Very recently, some nonconvex extension of the LBreI, allowing $E(x) = E(Ax, b)$ to be in a general form which has a Lipschitz continuous gradient, was made in [4]. Although a group of numerical tests were reported in [4] to demonstrate that the LBreI in nonconvex optimization still leads to superior performance than that of the regularized problems (1.2), the current theory is far from satisfying. On one hand, as partially mentioned in section 4.2 in [4], the required gradient Lipschitz continuity assumption precludes the application of LBreI to many practical problems such as blind deconvolution problems, Poisson inverse problems, and quadratic inverse problems. On the other hand, it is unclear whether similar results to Theorem 1.1 can be established for general convex energy function $E$. These two aspects contribute the main motivation of this study.

1.2 Beyond gradient Lipschitz continuity

In order to relax the gradient Lipschitz continuity, we follow a new notion which was recently introduced under the name of Lipschitz-like convexity condition in [1,28] (also independently rediscovered with the name of relative smoothness in [20]). If the fidelity term $E(x)$ is two-differentiable, then the gradient Lipschitz continuity is essentially equivalent to that the norm of the Hessian of $E(x)$ can be bounded by a constant. Let us take a quartic function $E(x) = \frac{1}{4}x^4 + \frac{1}{2}x^2$ as an example; as the Hessian $x^2 + 1$ “grows” to infinity as $|x|$ tends to infinity, it fails to be gradient Lipschitz continuous. The main idea of the Lipschitz-like convexity condition is to choose a suitable relative function $h(x)$ such that the Hessian of $E(x)$ could be bounded by that of $h(x)$ in the sense of $L\nabla^2 h(x) - \nabla^2 E(x) \succeq 0$ for some constant $L > 0$; please refer to [1] for other equivalent conditions. As the first contribution, we replace the norm proximity term $\frac{1}{2\delta}\|x - x^k\|^2$ in (1.4) with the
Bregman distance term $D_h(x, x^k)$ by choosing a suitable relative function $h(x)$ so that the gradient Lipschitz continuity can be relaxed. Consequently, the generalized method consists in solving the following optimization problems

$$x^{k+1} := \arg\min_x \{\langle \nabla E(x^k), x - x^k \rangle + \frac{1}{\delta}D_h(x, x^k) + \mu D_{R_k}^p(x, x^k)\}. \quad (1.8)$$

If $\mu = 0$, then the Bregman regularization term $\mu D_{R_k}^p(x, x^k)$ disappears, in which case the generalized method (1.8) overlaps with the NoLips algorithm and the Bregman proximal gradient method in [17]. In other words, our proposed method can also be obtained by adding the term $\mu D_{R_k}^p(x, x^k)$ to the NoLips algorithm or the Bregman proximal gradient method. This point will be highlighted in Section 3 where a formal unified framework will be constructed.

1.3 Contribution and organization

We summarize the contributions as follows:

- For a convex energy function $E$, if it satisfies the Lipschitz-like convexity condition, then we are able to extend Theorem 1.1. This discovery will correspond to Theorem 4.1.
- For a nonconvex energy function $E$, if it satisfies the Lipschitz-like convexity condition and has some “nice” local properties, then we are able to show the global convergence of generalized LBreI methods. This discovery will correspond to Theorem 4.2.

The remainder of the paper is organized as follows. The Bregman distances and the Lipschitz-like convexity condition, as two basic tools, are recalled in Section 2. The unified framework is introduced in Section 3. The convergence analysis is presented in Section 4. Two application examples are discussed in Section 5. Concluding remarks and research directions for future work are given in Section 6.

2 Preliminaries

Throughout the paper, we assume that $\mathbb{E}$ is a finite dimensional vector space with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$. For a nonempty subset $\Omega \subseteq \mathbb{E}$ and a point $x \in \mathbb{E}$, we define by $\text{dist}(x, \Omega) := \inf_{y \in \Omega} \|x - y\|$ the distance function from $x$ to $\Omega$. The notation and almost all the facts about convex analysis we employ are standard and can be found in Rockafellar’s classic book [26]. Let $f$ be a convex function on $\mathbb{E}$, the domain (conjugate function of $f$, gradient of $f$, and subgradient of $f$) is denoted by $\text{dom} f$ ($f^*$, $\nabla f$, $\partial f$, respectively).

2.1 The Bregman distances

The most important technical ingredient in the linearized Bregman iterations (also in our proposed algorithmic framework) is the concept of Bregman distance, which was originally introduced by Bregman in the 1967’s paper [8] for extending the classical method of cyclic orthogonal projections to the case of non-orthogonal projections. Given a differentiable convex function $h$, the Bregman distance associated with $h$ between two points $x$ and $y$ is defined as

$$D_h(x, y) := h(x) - h(y) - \langle \nabla h(y), x - y \rangle.$$
It is worth noting that $D_h$ is not a metric since the symmetry and the triangle inequality fail to hold. In order to guarantee that projection methods equipped with the Bregman distance still behave well, a “nice” convex function $h$ has to be chosen. The class of Legendre functions, which was investigated in [2], have become a popular choice to define the Bregman distance [1]. Now, we recall its definition below.

**Definition 2.1 (Legendre functions, [26]).** Let $h : E \to (-\infty, +\infty]$ be a proper lower semicontinuous (lsc) convex function. We say that

- it is essentially smooth if int dom $h \neq \emptyset$, $h$ is differentiable on int dom $h$, and $\|\nabla h(x^k)\| \to \infty$ for every sequence $\{x^k\}_{k \geq 0} \subseteq$ int dom $h$ converging to a boundary point of dom $h$ as $k \to \infty$,
- and it is of Legendre type if $h$ is essentially smooth and strictly convex on int dom $h$.

Note that the Legendre functions are essentially smooth and hence differentiable on int dom $h$. However, the associated function $R$ in the linearized Bregman iterations may be non-differentiable on int dom $R$; for example $R(\cdot) = \|\cdot\|_1$ is not of Legendre type. Therefore, we need the concept of generalized Bregman distances, introduced by Kiwiel in [17]. It is worth emphasizing that we do not completely follow the definition of Kiwiel in [17] where the associated function is needed to be strictly convex on its domain, excluding the case of $R(\cdot) = \|\cdot\|_1$ again. Given a proper lsc convex function $h$, the generalized Bregman distance associated with $h$ between $x, y \in E$ with respect to a subgradient $y^* \in \partial h(y)$ is defined by

$$D^g_h(x, y) := h(x) - h(y) - \langle y^*, x - y \rangle, \forall x \in \text{dom } h, y \in \text{dom } \partial h.$$  

(2.1)

Using this generalized definition, we can obtain important lemma which generalizes the three points identity in [14].

**Lemma 2.1 ([16, 17]).** Let $h : E \to (-\infty, +\infty]$ be a proper lsc convex function with dom $\partial h \neq \emptyset$. For any $z \in \text{dom } h$ and $x, y \in \text{dom } \partial h$ and $x^* \in \partial h(x), y^* \in \partial h(y)$, we have that

$$D^g_h(z, x) - D^g_h(z, y) - D^g_h(y, x) = \langle x^* - y^*, y - z \rangle.$$  

(2.2)

Recall the fact that for a Legendre function $h$, it holds that dom $\partial h = \text{int dom } h$ with $\partial h(x) = \{\nabla h(x)\}$ for any $x \in \text{int dom } h$. Therefore, applying Lemma 2.1 we recover the well-known three points identity

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

(2.3)

which holds for any Legendre function $h$ with the three points $z \in \text{dom } h$ and $x, y \in \text{int dom } h$.

At last, we introduce a measure for the lack of symmetry in $D_h$.

**Definition 2.2 ([11]).** Given a Legendre function $h : E \to (-\infty, +\infty]$, its symmetry coefficient is defined by

$$\alpha(h) := \inf \left\{ \frac{D_h(x, y)}{D_h(y, x)} : x, y \in \text{int dom } h, x \neq y \right\} \in [0, 1].$$
2.2 The Lipschitz-like convexity condition

The applied scope of the well-known proximal gradient method and its variants is limited by the gradient Lipschitz continuity assumption. Recently, the authors of [1] proposed the Lipschitz-like/convexity condition as an alternative to go beyond the gradient Lipschitz continuity. Below, we recall its definition.

**Definition 2.3** (Lipschitz-like/convexity condition, [1] [28]). Let \( h : E \to (-\infty, +\infty] \) be a Legendre function and let \( E : E \to (-\infty, +\infty] \) be a proper lsc function with \( \text{dom} E \supset \text{dom} h \), and \( E \) is differentiable on \( \text{int dom} h \). Given such a pair of functions \((E, h)\), the Lipschitz-like/convexity condition denoted by \((\text{LC})\) is:

\[
(\text{LC}) \quad \exists L > 0 \text{ with } Lh - E \text{ convex on } \text{int dom} h.
\]

The following three points extended descent lemma will play a very important role in the forthcoming analysis.

**Lemma 2.2.** Consider the pair of functions \((E, h)\) as above and take \( L > 0 \). Then, the function \( Lh - E \) is convex on \( \text{int dom} h \) if and only if for any \((x, y, z) \in (\text{int dom} h)^3\):

\[
E(x) \leq E(y) + \langle \nabla E(z), x - y \rangle + LD_h(x, z) - DE(y, z).
\]  \((2.4)\)

In particular, assume that \((\text{LC})\) holds for the pair of convex functions \((E, h)\). Then, for any \((x, y, z) \in \text{int dom} h \times \text{dom} h \times \text{int dom} h\), we have

\[
E(x) \leq E(y) + \langle \nabla E(z), x - y \rangle + LD_h(x, z).
\]  \((2.5)\)

It is not hard to see that \((2.5)\) follows from \((2.4)\) by noting that \( D_E(y, z) \geq 0 \) when \( E \) is convex. For a possibly nonconvex function \( E \), the term \( D_E(y, z) \) is difficult to bound. Therefore, we usually consider the case of \( y = z \), in which the term \( D_E(y, z) \) disappears and the corresponding descent inequality becomes:

\[
E(x) \leq E(y) + \langle \nabla E(y), x - y \rangle + LD_h(x, y).
\]  \((2.6)\)

When \( h(\cdot) = \frac{1}{2} \| \cdot \|^2 \), it further reduces to

\[
E(x) \leq E(y) + \langle \nabla E(y), x - y \rangle + \frac{L}{2} \| x - y \|^2,
\]  \((2.7)\)

which is exactly the classical descent lemma; see for example Lemma 5.7 in [3].

3 The proposed algorithmic framework

Let \( x_0 \in E, p^0 \in \partial R(x_0) \) be given. We propose the following linearized Bregman iteration framework, abbreviated as LBreIF, to extend the LBreI method.

\[
\begin{cases}
  x^{k+1} = \arg \min_x \{ \langle \nabla E(x^k), x - x^k \rangle + \beta R(x) + \frac{1}{\delta^k} D_h(x, x^k) + \mu^k D_R^k(x, x^k) \}, \\
  p^{k+1} = \frac{\mu^k}{\mu^k + \beta} p^k - \frac{1}{\delta^k (\mu^k + \beta)} \left[ \nabla h(x^{k+1}) - \nabla h(x^k) + \delta^k \nabla E(x^k) \right].
\end{cases}
\]  \((3.1a, 3.1b)\)

To help the reader understand LBreIF better, we add a few remarks:
• The relative function $h$ is chosen to be strongly convex on its domain so that the subproblem in (3.1a) has the unique solution $x^{k+1}$. The term $D_h(x, x^k)$ is used to replace/generalize the proximity term $\frac{1}{2}\|x - x^k\|^2$ in the LBreI (1.4) so that the traditional gradient Lipschitz continuity assumption can be relaxed. Meanwhile, a potential difficulty is how to determine such a desired function $h$ so that the pair of $(E, h)$ satisfies the Lipschitz-like convexity condition and at the same time the subproblem (3.1a) can be solved efficiently.

• The term $\beta R(x)$ is taken into account in (3.1a) so the NoLips algorithm in [1] can be included as a special case of (3.1a) by setting $\mu^k \equiv 0$. In this way, we can see that the main difference between NoLips and LBreIF is the term $D^p_k R(x, x^k)$, which helps us generate regularized solutions via the nonsmooth Bregman distance $D_R(\cdot, \cdot)$; whereas NoLips as well as the traditional regularization algorithms produce solutions via optimizing the sum of a fidelity function and a regularized function. Theoretically, the sequence $\{x^k\}$ generated by LBreIF enjoys a refiner convergence phenomenon, as shown in Theorem 4.1.

• The subgradient $p^k \in \partial R(x^k)$ in (3.1b) is updated iteratively according to the optimality condition of (3.1a). The reader will find that (3.1b) plays a key role in the forthcoming convergence analysis.

Now, we present an elegant expression for LBreIF in the special case of $\beta \equiv 0$, $\mu_k \equiv \mu$, and $\delta^k \equiv \delta$. To do this, we first reformulate (3.1b) as follows

$$\nabla h(x^{k+1}) + \delta \mu p^{k+1} = \nabla h(x^k) + \delta \mu p^k - \delta \nabla E(x^k).$$

(3.2)

Denote $y^k := \nabla h(x^k) + \delta \mu p^k$; then we have

$$\begin{cases}
y^{k+1} = y^k - \delta \nabla E(x^k), \quad (3.3a) \\
x^{k+1} = (\nabla h + \delta \mu \partial R)^{-1}(y^{k+1}). \quad (3.3b)
\end{cases}$$

In particular, the iterative scheme above with $R(x) \equiv 0$ returns to the well-known mirror descent

$$x^{k+1} = (\nabla h)^{-1}(\nabla h(x^k) - \delta \nabla E(x^k)),$$

where $\nabla h$ is the mirror mapping and $(\nabla h)^{-1}$ is the inverse mirror mapping [23]. Thereby, the iterative scheme (3.3a)-(3.3b) may be viewed as a generalized mirror descent method if we view $\nabla h + \delta \mu \partial R$ as the mirror mapping and $(\nabla h + \delta \mu \partial R)^{-1}$ as the inverse mirror mapping [33].

At last, we introduce a natural assumption on the regularized function $R$ and the relative function $h$ so that the iterate (3.1a) is well defined if $\mu^k > 0$, $\delta^k > 0$ and $x^k \in \text{int dom } h$.

**Lemma 3.1.** If $h + \lambda R$ is supercoercive for all $\lambda > 0$ in the sense that

$$\liminf_{\|x\| \to \infty} \frac{h(x) + \lambda R(x)}{\|x\|} = \infty,$$

(3.4)

then for any $x^k \in \text{int dom } h$, the subproblem in (3.1a) has the unique solution $x^{k+1}$ and it must belong to $\text{int dom } h$. 

7
Proof. Using the definitions of the Bregman distances, the objective function in (3.1a) can be rewritten as

\[ g(x) := h(x) + \lambda R(x) + \langle \bar{u}, x \rangle + C, \]

where \( \bar{u} \in E \) and \( C \in \mathbb{R} \) are constant quantities. Note that \( |\langle \bar{u}, x \rangle| \leq \|\bar{u}\| \|x\| \) by the Cauchy-Schwarz inequality. We conclude that the objective function \( g \) is still supercoercive. Thus, the set of minimizers of \( g \) must be nonempty and compact by Weierstrass’ theorem [3]. The uniqueness of solution follows from the strict convexity of \( h \) on its domain. Finally, by the optimality condition we have

\[ 0 \in \partial h(x^{k+1}) + \lambda \partial R(x^{k+1}) + \bar{u}, \]

which implies that \( x^{k+1} \in \text{dom} \partial h = \text{int dom} h \). This completes the proof.

It can be easily verified that for any strongly convex function \( h \) and nonnegative function \( R \) (in the sense of \( R(x) \geq 0 \)), the composition function \( h + \lambda R \) is always supercoercive for all \( \lambda > 0 \).

4 Convergence analysis

In this section, we present a detailed convergence analysis for the proposed algorithmic framework.

4.1 Assumptions and the key lemma

The following assumption is a basic setting about the involved functions.

**Assumption 4.1.** We assume that the relative function \( h \), the regularized function \( R \), and the loss function \( E \) satisfy the following conditions:

(i) \( h : E \to (-\infty, +\infty] \) is of Legendre type.

(ii) \( R : E \to [0, +\infty] \) is proper lsc convex with \( \text{dom} \partial R \supset \text{int dom} h \) and \( \text{dom} R \cap \text{int dom} h \neq \emptyset \).

(iii) \( E : E \to (-\infty, +\infty] \) is proper lsc with \( \text{dom} E \supset \text{dom} h \) and is differentiable on \( \text{int dom} h \) and continuous on \( \text{dom} h \). Moreover, (LC) holds for the pair of functions \((E, h)\).

(iv) \(-\infty < \inf_{x \in \text{dom} h} f_{\beta}(x) \) with \( f_{\beta}(x) := \beta R(x) + E(x) \) for some \( \beta \geq 0 \).

The conditions \( \text{dom} \partial R \supset \text{int dom} h \) and \( \text{dom} R \cap \text{int dom} h \neq \emptyset \) in item (ii) and the differentiability of \( E \) on \( \text{int dom} h \) in item (iii) are posed to guarantee the objective function in (3.1a) is well-defined if \( x^k \in \text{int dom} h \). Item (ii) can be satisfied when \( R \) is real-valued since \( \text{dom} \partial R = \text{dom} R = E \). Hence, many exiting regularized functions like the \( \ell_1 \) norm obey these restrictions on \( R \). Items (i) and (iii) essentially requires the (LC) condition. The last item holds trivially for nonnegative regularized function \( R \) and energy function \( E \).

In order to derive point convergence results, we will rely on the following assumption:

**Assumption 4.2.** The Bregman distance associated with the relative function \( h \) and the regularized function \( R \) satisfy the following conditions:

(i) For every \( x \in \text{dom} h \) and \( \eta \in \mathbb{R} \), one of the level sets \( \{y \in \text{int dom} h : D_h(x, y) \leq \eta\} \) and \( \{y \in \text{dom} R : D_R(x, y) \leq \eta\} \) is bounded.
(ii) If \( \{x^k\} \subset \text{int dom } h \) converges to some \( x \) in dom \( h \), then \( D_h(x, x^k) \to 0 \) and \( D_R(x, x^k) \to 0 \).

(iii) If \( x \in \text{int dom } h \) and if \( \{x^k\} \) is such that \( D_h(x, x^k) \to 0 \) or \( D_R(x, x^k) \to 0 \), then \( x^k \to x \).

If we drop the requirement \( D_R(x, x^k) \to 0 \) in item (ii) above, then Assumption 4.2 can be implied by Assumption H in [1] and hence it holds for many entropies such as the Boltzmann-Shannon function \( h(x) = x \log x \) which will appear in the section of application. If \( R \) itself is real-valued convex, then we always have that \( D_R(x, x^k) \to 0 \) as \( x^k \to x \) due to the continuity of \( R \) and the boundedness of subgradients over compact sets; see Theorem 3.16 in [3].

For nonconvex convergence analysis, we require the following assumptions. The first one was used in [4] and the second one in [7]. Again, if \( R \) itself is real-valued convex, then the first assumption below holds trivially.

**Assumption 4.3.** The regularized function \( R \) has locally bounded subgradients in the sense that if for every bounded set \( U \subset \mathbb{R} \) there exists a constant \( C > 0 \) such that for every \( x \in U \) and for all \( p \in \partial R(x) \) we have \( \|p\| \leq C \).

**Assumption 4.4.** A function \( f \) is said to be locally gradient-Lipschitz-continuous if for every bounded set \( U \subset \text{int dom } f \) there exists a constant \( L_f > 0 \) such that for any \( x, y \in U \) we have \( \|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\| \).

The next lemma provides key descent estimations for the objective function in terms of the Bregman distances.

**Lemma 4.1** (Descent inequalities). Under Assumption 4.1, we have

\[
\beta R(x^{k+1}) + E(x^{k+1}) - (\beta R(x^k) + E(x^k)) \leq LD_h(x^{k+1}, x^k) - \frac{1}{\delta_k} D_h(x^k, x^{k+1}) - \mu^k D_R(x^{k+1}, x^k).
\]

If \( E \) is also assumed convex, then for any \( x \in \text{dom } h \), it holds that

\[
\beta R(x^{k+1}) + E(x^{k+1}) - (\beta R(x) + E(x)) \leq \mu^k D_R(x, x^k) - \mu^k D_R(x, x^{k+1}) - \mu^k D_R(x^{k+1}, x^k) + \frac{1}{\delta_k} D_h(x, x^{k+1}) - \frac{1}{\delta_k} D_h(x^{k+1}, x^k) + LD_h(x^{k+1}, x^k).
\]

**Proof.** We only show (4.2); the other relationship can be shown in a similar way. To this end, we first rephrase (3.1b) as the following equality:

\[
\mu^k (p^{k+1} - p^k) + \beta p^{k+1} + \nabla E(x^k) + \frac{1}{\delta_k} (\nabla h(x^{k+1}) - \nabla h(x^k)) = 0.
\]

Now, let us do the inner product between the left-hand side of (4.3) and the term \( x - x^{k+1} \). According to the generalized three points identity in Lemma 2.1, we have

\[
\mu^k (p^{k+1} - p^k, x - x^{k+1}) = \mu^k D_R(x, x^k) - \mu^k D_R(x, x^{k+1}) - \mu^k D_R(x^{k+1}, x^k).
\]

Similarly, applying the three points identity to the relative function \( h \), we obtain

\[
\frac{1}{\delta_k} (\nabla h(x^{k+1}) - \nabla h(x^k), x - x^{k+1}) = \frac{1}{\delta_k} D_h(x^k, x^k) - \frac{1}{\delta_k} D_h(x, x^{k+1}) - \frac{1}{\delta_k} D_h(x^{k+1}, x^k).
\]
Using the three points extended descent property in Lemma 2.2, we have

\[
\langle \nabla E(x^k), x - x^{k+1} \rangle \leq E(x) - E(x^{k+1}) + LD_h(x^{k+1}, x^k).
\] (4.6)

Using the subgradient inequality for the convex function \( R \), we have

\[
\beta(p^{k+1}, x - x^{k+1}) \leq \beta R(x) - \beta R(x^{k+1}).
\] (4.7)

Thereby, the descent inequality (4.2) follows from (4.3)-(4.7). \( \square \)

4.2 Convergence for convex optimization

The basic convergence properties of LBreIF in the convex case are presented in the following lemma.

**Lemma 4.2** (Basic convergence). Let \( \sigma_k := \sum_{i=0}^k \delta^k \), \( \nu := \inf_{x \in \text{dom} \ h} f_\beta(x) \), and let \( \{x^k\} \) be the sequence generated by (3.1a)-(3.1b) with the parameters satisfying

\[
\begin{align*}
0 < \delta^k &< \frac{1 + \alpha(h) - \omega}{L}, \exists \omega \in (0, 1 + \alpha(h)), \quad \text{(4.8a)} \\
0 < \tau < \delta^{k+1} &\leq \delta^k, \forall k \geq 0. \quad \text{(4.8b)}
\end{align*}
\]

If Assumption 4.1 holds and \( E \) is also assumed convex, then we have that

(a) (Monotonicity). \( \{f_\beta(x^k)\} \) is nonincreasing.

(b) (Summability). \( \sum_{k=0}^\infty D_h(x^{k+1}, x^k) < \infty. \)

(c) (Convergence of the function values). If \( \sigma_k \to \infty \), then \( \lim_{k \to \infty} f_\beta(x^k) = \nu. \)

**Proof.** First of all, denote \( H(x, x^k) := \delta^k \mu^k D_R(x, x^k) + D_h(x, x^k) \). Using (1.2) of Lemma 4.1 and the condition (4.8b) on \( \delta^k \) and \( \mu^k \), we obtain that for all \( x \in \text{dom} \ h \),

\[
\delta^k (f_\beta(x^{k+1}) - f_\beta(x)) \leq H(x, x^k) - H(x, x^{k+1}) - (1 - \delta^k L)D_h(x^{k+1}, x^k). \] (4.9)

Setting \( x = x^k \) in (4.9), recalling the definition of \( H(x, x^k) \), and using the condition (4.8a) on \( \delta^k \), we derive that

\[
\begin{align*}
\delta^k (f_\beta(x^{k+1}) - f_\beta(x^k)) &\leq -H(x^k, x^{k+1}) - (1 - \delta^k L)D_h(x^{k+1}, x^k) \\
&\leq -D_h(x^k, x^{k+1}) - (1 - \delta^k L)D_h(x^{k+1}, x^k) \\
&\leq -D_h(x^k, x^{k+1}) + \alpha(h)D_h(x^{k+1}, x^k) - \omega D_h(x^{k+1}, x^k) \\
&\leq -\omega D_h(x^{k+1}, x^k),
\end{align*}
\] (4.10)

where the last relationship follows from the definition of \( \alpha(h) \). Therefore, the sequence \( \{f_\beta(x^k)\} \) is nonincreasing. Note that \( x^k \in \text{dom} \ h \) and the assumption that \( \nu = \inf_{x \in \text{dom} \ h} f_\beta(x) > -\infty \), we conclude that

\[
\lim_{k \to \infty} f_\beta(x^k) \geq \nu > -\infty.
\] (4.11)

Using (4.10), we get

\[
D_h(x^{k+1}, x^k) \leq \frac{\delta^k}{\omega} (f_\beta(x^k) - f_\beta(x^{k+1})) \leq \frac{1 + \alpha(h) - \omega}{\omega L} (f_\beta(x^k) - f_\beta(x^{k+1})).
\]
Thus, it follows from (4.11) and the inequality above that for any $n \in \mathbb{N}$,
\[
\sum_{k=0}^{n} D_h(x^{k+1}, x^k) \leq \frac{1 + \alpha(h) - \omega}{\omega L} (f_\beta(x^0) - f_\beta(x^{n+1})) \leq \frac{1 + \alpha(h) - \omega}{\omega L} (f_\beta(x^0) - v).
\]

This implies the following summability
\[
\sum_{k=0}^{\infty} D_h(x^{k+1}, x^k) \leq \frac{1 + \alpha(h) - \omega}{\omega L} (f_\beta(x^0) - v) < \infty. \tag{4.12}
\]

Now, let us show the convergence of the function values. Summing (4.9) over $k = 0, \cdots, n$, using the fact $\delta^k L - 1 \leq \alpha(h)$, and noting the nonnegativity of $H(x, x^k)$, we derive that
\[
\sum_{k=0}^{n} \delta^k(f_\beta(x^{k+1}) - f_\beta(x)) \leq H(x, x^0) - H(x, x^{n+1}) + \alpha(h) \sum_{k=0}^{n} D_h(x^{k+1}, x^k) \\
\leq H(x, x^0) + \alpha(h) \sum_{k=0}^{\infty} D_h(x^{k+1}, x^k). \tag{4.13}
\]

Since the sequence $\{f_\beta(x^k)\}$ is nonincreasing, we get
\[
\sum_{k=0}^{n} \delta^k(f_\beta(x^{k+1}) - f_\beta(x)) \geq \sum_{k=0}^{n} \delta^k((f_\beta(x^{n+1}) - f_\beta(x)).
\]

Therefore, invoking (4.13) and the notation $\sum_{k=0}^{n} \delta^k = \sigma_n$ we further get
\[
f_\beta(x^{n+1}) - f_\beta(x) \leq \frac{H(x, x^0) + \alpha(h) \sum_{k=0}^{\infty} D_h(x^{k+1}, x^k)}{\sigma_n}.
\]

Taking the limit with $\sigma_n \to \infty$ above and recalling that $\sum_{k=0}^{\infty} D_h(x^{k+1}, x^k) < \infty$, we get that for every $x \in \text{dom } h$, $\lim_{n \to \infty} f_\beta(x^{n+1}) \leq f_\beta(x)$ and hence $\lim_{k \to \infty} f_\beta(x^k) \leq v$. Together with (4.11), it follows that $\lim_{k \to \infty} f_\beta(x^k) = v$. This completes the proof. \(\square\)

In order to obtain point convergence results, we need the following basic result about sequence convergence.

**Lemma 4.3** (\cite{1,25}). Let $\{v_k\}$ and $\{\epsilon_k\}$ be nonnegative sequences. Assume that $\sum_{k=0}^{\infty} \epsilon_k < \infty$ and that
\[
v_{k+1} \leq v_k + \epsilon_k.
\]

Then, $\{v_k\}$ converges.

We now present the point convergence results.

**Lemma 4.4** (Point convergence). Let Assumption 4.2 hold, $E$ be convex, and $\beta \geq 0$. Denote
\[
X_\beta := \text{arg min}\{f_\beta(x) : x \in \text{dom } h\}.
\]

(1) If $X_\beta$ is nonempty, then any limit point of $\{x^k\}$ (if it exists) belongs to $X_\beta$.

(2) If $\text{dom } h = \text{dom } h$, Assumption 4.2 holds, and $X_\beta$ is nonempty, then the sequence of $\{x^k\}$ converges to some solution $x^* \in X_\beta$. 

11
Proof. Let \( x^* \) be a limit point of \( \{x^k\} \). In other words, there exists a subsequence \( \{x^{k_i}\} \) such that \( \lim x^{k_i} = x^* \). Then, \( x^* \in \text{dom } h \) due to the closedness of \( \text{dom } h \) and \( \{x^k\} \subset \text{dom } h \). By using (a) and (c) of Lemma 4.2 and the lower semicontinuity of \( E \) and \( R \), we derive that

\[
\min\{f_\beta(x) : x \in \text{dom } h\} = \min\{f_\beta(x) : x \in \text{dom } h\} = \lim_{k \to \infty} f_\beta(x^k) = \liminf_{k \to \infty} f_\beta(x^k) = \liminf_{k \to \infty} \{E(x^{k_i}) + \beta R(x^{k_i})\} \geq \liminf_{k \to \infty} E(x^{k_i}) + \liminf_{k \to \infty} \beta R(x^{k_i}) \geq E(x^*) + \beta R(x^*) = f_\beta(x^*),
\]

which implies that \( x^* \in X_\beta \). This completes the proof of (1).

Recalling (4.9), we have for all \( x \in \text{dom } h = \text{dom } h \),

\[
d^k(f_\beta(x^{k+1}) - f_\beta(x)) \leq H(x, x^k) - H(x, x^{k+1}) - (1 - d^k L) D_h(x^{k+1}, x^k).
\]

Fix \( \bar{x} \in X_\beta \). Due to \( d^k(f_\beta(x^{k+1}) - f_\beta(\bar{x})) \geq 0 \) and the condition (4.8a) on \( d^k \), the inequality above yields

\[
H(\bar{x}, x^{k+1}) \leq H(\bar{x}, x^k) + (\alpha h - \omega) D_h(x^{k+1}, x^k).
\]

If \( \alpha h - \omega \leq 0 \), then the nonnegative sequence \( \{H(\bar{x}, x^k)\} \) is nonincreasing and hence it must converge. If \( \alpha h - \omega > 0 \), then we can use Lemma 4.3 with \( \epsilon_k = (\alpha h - \omega) D_h(x^{k+1}, x^k) \) to conclude that \( \{H(\bar{x}, x^k)\} \) converge since \( \sum_{k=0}^{\infty} D_h(x^{k+1}, x^k) < \infty \) from (b) of Lemma 4.2. The convergence of \( \{H(\bar{x}, x^k)\} \) and condition (i) of Assumption 4.2 imply that the sequence \( \{x^k\} \) is bounded. Hence, we can take one of its cluster points, denoted by \( x^* \), which must lies in \( \text{dom } h = \text{dom } h \). Further, we deduce that \( x^* \in X_\beta \) from part (1).

Let \( \{x^{n_k}\} \) be the subsequence of \( \{x^k\} \) such that \( x^{n_k} \to x^* \). By condition (ii) of Assumption 4.2, we have that \( D_h(x^*, x^{n_k}) \to 0 \) and \( D_R(x^*, x^{n_k}) \to 0 \). Thus, \( H(x^*, x^{n_k}) = \delta^{n_k} \mu^{n_k} D_R(x^*, x^{n_k}) \) converges to zero. Note that \( \{H(x^*, x^{n_k})\} \) is a convergence sequence. Hence it must converge to zero. Recall that \( \delta^k \mu^k \) are bounded by \( \tau \). We can deduce that both \( \{D_R(x^*, x^{k})\} \) and \( \{D_h(x^*, x^{k})\} \) converge to zero, which implies that \( \{x^k\} \) converges to \( x^* \) by condition (iii) of Assumption 4.2.

Built on the lemmas above, we are now to present the first main convergence theorem in this study.

Theorem 4.1. Suppose that Assumptions 4.1, 4.3 hold, the minimizer set \( X_0 \) is nonempty, \( \text{dom } h = \text{dom } h \), and the function \( E(x) \) has the form of \( E(Ax, b) \), satisfying that \( A \) is surjective and

\[
\psi(E(u, b) - E(\hat{u}, b)) \geq \|u - \hat{u}\|, \quad \forall u \in \text{dom } E(\cdot, b), \tag{4.17}
\]

where \( \psi(t) \) is some real function with \( \lim_{t \to 0} \psi(t) = 0 \) and \( \hat{u} \in \arg \min_w \{E(u, b)\} \). Suppose also that \( h \) is continuously differentiable on \( \text{int } \text{dom } h \). Denote \( \hat{x} := \arg \min \{\mu R(x) + \frac{1}{\delta} h(x), x \in X_0\} \).

Then, the sequence \( \{x^k\} \), generated by (3.1a) - (3.1b) with \( \mu^k \equiv \mu, \ \delta^k \equiv \delta, \ \beta = 0, \) and \( x^0 = p^0 = 0 \), converges either to the unique minimizer \( \hat{x} \) or to a boundary point of \( \text{dom } h \).

Proof. From Lemma 4.3 we know that the sequence \( \{x^k\} \) must converge to some point \( x^* \). If this point is not a boundary point of \( \text{dom } h \), let us show \( x^* = \hat{x} \). We begin with the relationship (3.1b) in the current setting, which has the following form

\[
\mu p^{k+1} + \frac{1}{\delta} \nabla h(x^{k+1}) = \mu p^k + \frac{1}{\delta} \nabla h(x^k) - \nabla E(x^k). \tag{4.18}
\]
Denote $q^k := \mu p^k + \frac{1}{q} \nabla h(x^k)$ and let $g(x) = \mu R(x) + \frac{1}{q} h(x)$, which must be a proper lsc convex function; then $q^k \in \partial g(x^k)$. Using the subgradient inequality, we derive that

$$
g(x^k) \leq g(\hat{x}) - \langle q^k, \hat{x} - x^k \rangle = g(\hat{x}) + \langle \sum_{i=0}^{k-1} \nabla E(Ax^i, b), Ax^k - Ax^k \rangle \\leq g(\hat{x}) + \| \sum_{i=0}^{k-1} \nabla E(Ax^i, b) \| \| Ax^k - Ax^k \|. 
$$

Denote $z^{k-1} := \sum_{i=0}^{k-1} \nabla E(Ax^i, b)$. Now, we show two results: (a). $\{z^k\}$ is bounded and (b). $\| Ax^k - Ax^k \| \to 0$ as $k \to \infty$. Actually, from (4.18) we know

$$q^{k+1} = q^k - \nabla E(x^k) = \cdots = q^0 - A^T \sum_{i=0}^{k} \nabla E(Ax^i, b) = q^0 - A^T z^k.$$

Since $A$ is a surjective mapping, the verification of (a) can be reduced to showing the boundedness of $\{q^k\}$. Due to the convergence of $\{x^k\}$ and Assumption 4.3, $\{p^k\}$ must be bounded. On the other hand, since $\{x^k\}$ converges to $x^* \in \text{int dom } h$ and $\nabla h(x)$ is continuously differentiable at $x^*$, $\{\nabla h(x^k)\}$ must converge and hence it is a bounded sequence. Recalling $q^k = \mu p^k + \frac{1}{q} \nabla h(x^k)$, we can conclude that $\{q^k\}$ is bounded. It remains to show (b). Since $\hat{x}$ is an interior point of $h$ and that $\hat{x} \in \mathcal{X}_0 = \arg \min \{E(x) : x \in \text{dom } h\}$, we have $\hat{x} \in \arg \min \{E(x)\}$ and hence $\nabla E(\hat{x}) = A^T \nabla E(A\hat{x}, b) = 0$, which further implies that $\nabla E(A\hat{x}, b) = 0$ due to the fact of $A$ being surjective. Therefore, we get $\hat{x} \in \arg \min_{w} \{E(w, b)\}$ and hence the condition (4.17) now can be used to deduce that as $k \to \infty$,

$$\| Ax - Ax \| \leq \psi(E(A\hat{x}, b) - E(Ax^k, b) = \psi(E(\hat{x}) - E(x^k)) \to 0,$$

where the last relationship follows from $\lim_{t \to 0} \psi(t) = 0$ and the fact $\lim_{k \to \infty} E(x^k) = \lim_{k \to \infty} E(\hat{x})$, implied by (c) of Lemma 4.2.

Now, using (4.19) and the verified results (a)-(b), we have $g(x^*) \leq \liminf_{k \to \infty} g(x^k) \leq g(\hat{x})$. Therefore, $x^* \in \arg \min \{\mu R(x) + \frac{1}{q} h(x) : x \in \mathcal{X}_0\}$, which implies that $x^* = \hat{x}$ by the uniqueness of solutions. Note that the uniqueness follows from the strict convexity of $h$. This completes the proof.

In particular, if $h(x) = \frac{1}{2} \| x \|^2$ whose boundary set is empty, we have the following result which generalizes Theorem 1 about LBrel.

**Corollary 4.1.** Let $h(x) = \frac{1}{2} \| x \|^2$ and suppose that Assumptions 4.1-4.3 hold for $E$ and $R$, the minimizer set $\mathcal{X}_0$ is nonempty, and the function $E(x)$ has the form of $E(Ax, b)$, satisfying that $A$ is surjective and the condition (4.17) holds. Denote $\hat{x} := \arg \min \{\mu R(x) + \frac{1}{q} h(x) : x \in \mathcal{X}_0\}$. Then, the sequence $\{x^k\}$, generated by (3.1a)-(3.1b) with $\mu^k \equiv \mu$, $\delta_k \equiv \delta$, $\beta = 0$, and $x^0 = p^0 = 0$, converges to the unique minimizer $\hat{x}$.

### 4.3 Convergence for nonconvex optimization

This part is about the convergence analysis of LBrelIF for minimizing a nonconvex objective function $E(x)$. We start with a sufficient descent lemma, which generalizes the central result–Lemma 4.2 in [3].
Lemma 4.5 (Sufficient descent). Let \( \{x^k\} \) be the sequence generated by (3.1) with inf \( \mu^k \geq \mu > 0 \) and the stepsize \( \delta^k \) satisfying (4.8a). Denote

\[
\rho := \frac{L\omega}{1 + \alpha(h) - \omega}, \quad \omega \in (0, 1 + \alpha(h)).
\]

If Assumption 4.1 holds, then we have the following sufficient decrease property:

\[
f_\beta(x^{k+1}) + \rho D_h(x^{k+1}, x^k) + \mu D_R^{symm}(x^{k+1}, x^k) \leq f_\beta(x^k).
\]

(4.20)

In particular, we observe that

\[
\lim_{k \to \infty} D_h(x^{k+1}, x^k) = \lim_{k \to \infty} D_R^{symm}(x^{k+1}, x^k) = 0.
\]

(4.21)

Here, \( D_R^{symm} \) is the symmetric generalized Bregman distance, defined as

\[
D_R^{symm}(u, v) := D_R^q(u, v) + D_R^p(v, u) = \langle p - q, u - v \rangle
\]

for \( u, v \in \text{dom } R \) with \( p \in \partial R(u) \) and \( q \in \partial R(v) \).

Proof. Using the inequality (4.1) in Lemma 4.1, the definition of \( D_R^{symm} \), and the condition (4.8a) on \( \delta^k \), we derive that

\[
\delta^k (f_\beta(x^{k+1}) - f_\beta(x^k)) \leq -\mu^k \delta^k D_R^{symm}(x^{k+1}, x^k) - D_h(x^k, x^{k+1}) - (1 - \delta^k L)D_h(x^{k+1}, x^k)
\]

\[
\leq -\mu^k \delta^k D_R^{symm}(x^{k+1}, x^k) - D_h(x^k, x^{k+1}) + (\alpha - \omega)D_h(x^{k+1}, x^k)
\]

\[
\leq -\mu^k \delta^k D_R^{symm}(x^{k+1}, x^k) - \omega D_h(x^{k+1}, x^k),
\]

(4.22)

where the last relationship follows from the definition of \( \alpha(h) \). Thus, (4.20) follows from (4.22) by rearranging the terms and using the conditions on \( \delta^k \) and \( \mu^k \) and also the notation \( \rho \).

To obtain (4.21), one can sum (4.20) over \( k = 0, \cdots, n \) to get that

\[
\sum_{k=0}^{n} (\rho D_h(x^{k+1}, x^k) + \mu D_R^{symm}(x^{k+1}, x^k)) \leq f_\beta(x^0) - f_\beta(x^n) \leq f_\beta(x^0) - \inf_{x \in \text{dom } h} f_\beta(x).
\]

Therefore, \( \sum_{k=0}^{\infty} D_h(x^{k+1}, x^k) < \infty \) and \( \sum_{k=0}^{\infty} D_R^{symm}(x^{k+1}, x^k) < \infty \). Hence, (4.21) follows immediately. This completes the proof. \( \square \)

The set of all limit points of \( \{x^k\} \) is denoted by \( \Omega \). In other words,

\[
\Omega := \{x^* \in E : \text{there exists an increasing sequence of integers } \{k_i\} \text{such that } \lim_{i \to \infty} x^{k_i} = x^*\}.
\]

Lemma 4.6 (Point and function value convergence). In addition to Assumption 4.1 and (ii) of Assumption 4.2, we assume that \( h \) is strongly convex on \( \text{dom } h \) with \( \text{dom } h = \text{dom } h \) and that the level set \( \{x : f_\beta(x) \leq f_\beta(x^0)\} \) is bounded. Then, we have \( \Omega \neq \emptyset \) and for any limit point \( x^* \in \Omega \),

\[
\lim_{k \to \infty} f_\beta(x^k) = f_\beta(x^*).
\]
Proof. The boundedness of \( \{ x : f_\beta(x) \leq f_\beta(x^0) \} \) and the nonincreasing property of \( \{ f_\beta(x^k) \} \) from (4.20) ensure the boundedness of \( \{ x^k \} \). Hence, \( 0 \not\in \Omega \). Take \( x^* \in \Omega \). This means there exists a subsequence \( \{ x^{k_i} \} \subset \{ x^k \} \subset \text{int dom } h \) such that \( \lim_{i \to \infty} x^{k_i} = x^* \in \text{dom } h = \text{dom } h \). Together with (4.21) in Lemma 4.5 and using the strong convexity of \( h \), we can conclude that as \( i \to \infty \),

\[
D_h(x^{k_i+1}, x^{k_i}) \to 0, \quad \| x^{k_i+1} - x^{k_i} \| \to 0, \quad D_R(x^{k_i+1}, x^{k_i}) \to 0.
\]

(4.23)

Note that \( \{ x^k \} \subset \text{int dom } h \subset \text{int dom } E \) due to (iii) of Assumption 4.1. Using boundedness \( \{ x^k \} \) and Theorem 3.16 in [3], we know that \( \{ \nabla E(x^k) \} \) is bounded. Hence, as \( i \to \infty \),

\[
\langle \nabla E(x^{k_i}), x^* - x^{k_i} \rangle \leq \| \nabla E(x^{k_i}) \| \| x^* - x^{k_i} \| \to 0,
\]

\[
\langle \nabla E(x^{k_i}), x^{k_i+1} - x^{k_i} \rangle \leq \| \nabla E(x^{k_i}) \| \| x^{k_i+1} - x^{k_i} \| \to 0.
\]

(4.24)

In light of (3.1a), we have

\[
\langle \nabla E(x^{k_i}), x^{k_i+1} - x^{k_i} \rangle + \beta R(x^{k_i+1}) + \frac{1}{\beta} D_h(x^{k_i+1}, x^{k_i}) + \mu_k D_R(x^{k_i+1}, x^{k_i}) \leq \langle \nabla E(x^{k_i}), x^* - x^{k_i} \rangle + \beta R(x^*) + \frac{1}{\beta} D_h(x^*, x^{k_i}) + \mu_k D_R(x^*, x^{k_i}).
\]

(4.25)

Letting \( i \to \infty \) in the above inequality, using the results (4.23)-(4.24) and the assumption that \( D_h(y, y^k) \to 0 \) and \( D_R(y, y^k) \to 0 \) if \( \{ y^k \} \subset \text{int dom } h \) converges to some \( y \in \text{dom } h \), we obtain

\[
\limsup_{i \to \infty} R(x^{k_i+1}) \leq R(x^*).
\]

(4.26)

Combining the continuity of \( E \) over \( \text{dom } h = \overline{\text{dom } h} \) and noting that \( x^{k_i+1} \to x^* \) as \( i \to \infty \) as well since \( \| x^{k_i+1} - x^{k_i} \| \to 0 \), we further have

\[
\limsup_{i \to \infty} (E(x^{k_i+1}) + \beta R(x^{k_i+1})) \leq E(x^*) + \beta R(x^*).
\]

(4.27)

On the other hand, using the lower semicontinuity of \( R \) and \( E \), we derive that

\[
\liminf_{i \to \infty} (E(x^{k_i+1}) + \beta R(x^{k_i+1})) \geq \liminf_{i \to \infty} E(x^{k_i+1}) + \beta R(x^{k_i+1}) \geq E(x^*) + \beta R(x^*).
\]

(4.28)

Therefore, we get

\[
\lim_{i \to \infty} f_\beta(x^{k_i+1}) = \lim_{i \to \infty} (E(x^{k_i+1}) + \beta R(x^{k_i+1})) = E(x^*) + \beta R(x^*) = f_\beta(x^*).
\]

(4.29)

Note that \( \{ f_\beta(x^k) \} \) is a nonincreasing sequence from Lemma 4.3 and is lower bounded by the finite value \( \inf_{x \in \text{dom } h} f_\beta(x) \) and hence it is convergent. Therefore, we have \( \lim_{k \to \infty} f_\beta(x^k) = f_\beta(x^*) \).

In order to derive the global convergence of \( \{ x^k \} \) without the gradient Lipschitz continuity and convexity of \( E \), we combine the method in [4] and that in [7], both of which were originally inspired by [6]. First, we introduce a modified surrogate function \( F : \mathbb{E} \times \mathbb{E} \to \mathbb{R} \cup \{ +\infty \} \) in the same spirit of that in [4]:

\[
F_{\beta, \mu}(x, y) := E(x) + (\beta + \mu) R(x) + \mu R^*(y) - \mu \langle x, y \rangle.
\]

(4.30)

Note that the case of \( \beta = 0 \) and \( \mu = 1 \) reduces to the surrogate function in [4], which is only defined for functions with a Lipschitz continuous gradient. In what follows, we will deduce a sufficient decrease property of the surrogate function \( F_{\beta, \mu} \) and its subgradient bound. To this
end, we first link this function with the known function \( f_\beta \) and some Bregman distance. For any fixed \( y \in \text{dom} \partial R \), we take \( z \in \partial R^*(y) \). Then by the Fenchel theorem, we have \( y \in \partial R(z) \) and \( R^*(y) + R(z) = \langle y, z \rangle \), with which the surrogate function \( F_{\beta,\mu} \) now can be reformulated as

\[
F_{\beta,\mu}(x, y) = f_\beta(x) + \mu D_R(x, z). \tag{4.31}
\]

In particular, we have \( F_{\beta,\mu}(x^{k+1}, p^k) = f_\beta(x^{k+1}) + \mu D_R(x^{k+1}, x^k) \) since \( p^k \in \partial R(x^k) \) and hence \( x^k \in \partial R^*(p^k) \). Note that

\[
\partial F_{\beta,\mu}(x, y) = \left\{ \left( \nabla E(x) + (\beta + \mu)s_1 - \mu y, \mu s_2 - \mu x \right) : s_1 \in \partial R(x), s_2 \in \partial R^*(y) \right\}. \tag{4.32}
\]

Using again the fact that \( p^{k+1} \in \partial R(x^{k+1}) \) and \( x^k \in \partial R^*(p^k) \), we can get

\[
x^{k+1} := \left( \nabla E(x^{k+1}) + (\beta + \mu)p^{k+1} - \mu p^k \right) / \mu x^k - \mu x^{k+1} \in \partial F_{\beta,\mu}(x^{k+1}, p^k). \tag{4.33}
\]

Denote \( (s^k) := (x^k, p^{k-1}) \). The set of all limit points of \( \{s^k\} \) is denoted by \( \Omega_0 \). In other words,

\[
\Omega_0 := \{ s^* := (x^*, p^*) \in \mathbb{E} \times \mathbb{E} : \text{there exists an increasing sequence of integers} \{k_i\} \text{ such that } \lim_{i \to \infty} x^{k_i} = x^*, \lim_{i \to \infty} p^{k_i-1} = p^* \}.
\]

We point out that the set \( \Omega_0 \) above is slightly different from the set of limit points in [3], where the following definition was used.

\[
\omega(s^0) := \{ s^* := (x^*, p^*) \in \mathbb{E} \times \mathbb{E} : \text{there exists an increasing sequence of integers} \{k_i\} \text{ such that } \lim_{i \to \infty} x^{k_i} = x^*, \lim_{i \to \infty} p^{k_i} = p^* \}.
\]

**Lemma 4.7.** Under the same setting as Lemma 4.5, we have the following sufficient decrease property

\[
F_{\beta,\mu}(x^{k+1}, p^k) + \rho_D h(x^{k+1}, x^k) + \mu D_R(x^k, x^{k+1}) + \mu D_R(x^k, x^{k-1}) \leq F_{\beta,\mu}(x^k, p^{k-1}). \tag{4.34}
\]

Suppose further that Assumption 4.4 holds for \( h \) and \( E \) and that the level set \( \{x : f_\beta(x) \leq f_\beta(x^0)\} \) is bounded. Then, we have the subgradient bound by the iterates gap

\[
\|r^{k+1}\| \leq \rho_2 \|x^{k+1} - x^k\| + (\mu - \mu)\|p^{k+1} - p^k\|, \tag{4.35}
\]

where \( \rho_2 := L_E + L_f + \mu \) and \( \delta := \sup \{\delta^k\} \). Moreover, if (ii) of Assumption 4.3 and Assumption 4.4 also hold, \( \lim_k \mu^k = \mu \), and we also assume that \( h \) is strongly convex on \( \text{dom} h \) with \( \text{dom} h = \text{dom} h \), then \( \Omega_0 \) must be a nonempty and compact set, and for every \( s^* = (x^*, p^*) \in \Omega_0 \) we have \( \lim_{k \to \infty} \text{dist}(s^k, \Omega_0) = 0 \) and

\[
\lim_{k \to \infty} F_{\beta,\mu}(s^k) = f_\beta(x^*). \tag{4.36}
\]

**Proof.** Adding \( \mu D_R(x^k, x^{k-1}) \) to both sides of (4.20) and using the formulation (4.31), we deduce (4.34). The boundedness of the level set \( \{x : f_\beta(x) \leq f_\beta(x^0)\} \) and the monotonicity of \( \{f_\beta(x^k)\} \)
ensure the boundedness of \( \{x^k\} \) and hence Assumption 4.4 can be employed to bound the difference of gradient below. Actually, we can derive that

\[
\begin{align*}
||r^{k+1}|| &\leq ||\nabla E(x^{k+1}) + (\beta + \mu)p^{k+1} - \mu p_k|| + \mu||x^k - x^{k+1}|| \\
&\leq ||\nabla E(x^{k+1}) + (\beta + \mu)p^{k+1} - \mu p_k|| + (\mu - \mu)||p^{k+1} - p_k|| + \mu||x^k - x^{k+1}|| \\
&= ||\nabla E(x^{k+1}) - \nabla E(x^k) - \frac{1}{\delta k} (\nabla h(x^{k+1}) - \nabla h(x^k))|| \\
&\quad + (\mu - \mu)||p^{k+1} - p_k|| + \mu||x^k - x^{k+1}|| \\
&\leq ||\nabla E(x^{k+1}) - \nabla E(x^k)|| + \frac{1}{\delta k} ||\nabla h(x^{k+1}) - \nabla h(x^k)|| \\
&\quad + (\mu - \mu)||p^{k+1} - p_k|| + \mu||x^k - x^{k+1}|| \\
&\leq (L_E + \frac{L_E}{\mu} + \mu)||x^{k+1} - x^k|| + (\mu - \mu)||p^{k+1} - p_k||,
\end{align*}
\]

where the equality follows from (3.11).

Now, we show the nonemptiness of \( \Omega_0 \). By the boundedness of \( \{x^k\} \), there exists an increasing of integers \( \{i_j\} \in \mathbb{N} \) such that \( \lim_{j \to \infty} x^{i_j} = x^* \). Recall that \( p^{i_j} \in \partial R(x^{i_j}) \). Using the locally bounded subgradient Assumption 4.3, we know that \( \{p^{i_j}\} \) must be bounded (actually \( \{p^{i_j}\} \) is bounded due to the same argument) and hence there exists a subsequence \( \{k_{i_j}\} \subset \{i_j\} \) such that \( \lim_{i \to \infty} p^{k_{i_j}} = \bar{p} \). From (3.11), it holds that

\[
(\mu + \beta)p^{k_{i_j}} - \mu p^{k_{i_j}} - \nabla E(x^{k_{i_j}}) - \frac{1}{\delta k} (\nabla h(x^{k_{i_j}}) - \nabla h(x^{k_{i_j}})).
\]

Note that \( \lim_{i \to \infty} x^{k_{i_j}} = \lim_{i \to \infty} x^{k_{i_{j-1}}} = x^* \), \( \lim_{i \to \infty} \mu p^{k_{i_j}} = \mu \) and \( \{\delta k_{i_j}\} \) is bounded. Together with Assumption 4.4, we conclude that there exists a point \( p^* \) such that \( \lim_{i \to \infty} p^{k_{i_j}} = p^* \) (such point may be different from \( \bar{p} \)). Therefore, \( s^* = (x^*, p^*) \) indeed belongs to \( \Omega_0 \) and hence it is nonempty. In particular, \( x^* \in \Omega \) for each \( s^* = (x^*, p^*) \in \Omega_0 \). Thus, in light of Lemma 4.5 and Lemma 4.6, we derive that

\[
\lim_{k \to \infty} F_{\beta, \mu}(s^k) = \lim_{k \to \infty} F_{\beta, \mu}(x^k, p^{k-1}) = \lim_{k \to \infty} (f_{\beta}(x^k) + \mu D_R(x^k, x^{k-1})) = f_{\beta}(x^*).
\]

From Theorem 3.7 in [27], we know that the set \( \Omega_0 \) must be closed since it is the set of cluster points of \( \{s^k\} \). The boundedness of \( \Omega_0 \) is due to the boundedness of \( \{x^k, p^k\} \). Therefore, the set \( \Omega_0 \) is compact and hence \( \lim_{k \to \infty} \text{dist}(s^k, \Omega_0) = 0 \) by the definition of limit points. This completes the proof.

\[\square\]

**Lemma 4.8.** Let \( \{a^k\} \) and \( \{b^k\} \) be given sequences of \( E \). If \( a^k \to 0 \) and \( b^k \to b \neq 0 \) as \( k \to \infty \), then as \( n \to \infty \) we have

\[
\| \sum_{k=0}^{n} (a^k + b^k) \| \to \infty.
\]

**Proof.** Using the condition that \( a^k \to 0 \) and \( b^k \to b \neq 0 \) as \( k \to \infty \), we can find an index \( k_0 \) such that for any \( k \geq k_0 \), it holds that

\[
\|a^k\| < \frac{1}{4}\|b\|, \quad \|b^k - b\| < \frac{1}{4}\|b\|.
\]

Letting \( n > k_0 \) and using the reverse triangle inequality \( \|u + \sum_i u_i\| \geq \|u\| - \sum_i \|u_i\| \), we drive that

\[
\begin{align*}
\| \sum_{k=0}^{n} (a^k + b^k) \| &= \| \sum_{k=0}^{n} b + \sum_{k=0}^{n} a^k + (b^k - b) \| \\
&\geq \| \sum_{k=0}^{n} b \| - \sum_{k=0}^{n} \|a^k\| - \sum_{k=0}^{n} \|b^k - b\| \\
&\geq (n + 1)\|b\| - \sum_{k=0}^{n} \|a^k\| - \|b^k - b\| - \frac{n-k_0}{2}\|b\| \\
&= \frac{n+2-k_0}{2}\|b\| - \sum_{k=0}^{n} \|a^k\| + \|b^k - b\|.
\end{align*}
\]
from which the conclusion follows. This completes the proof.

Now, we are ready to present the second main convergence theorem in this study.

**Theorem 4.2.** Suppose that $F_{\beta, \mu}$ is a KL function, Assumptions 4.1, 4.3 and (ii) of Assumption 4.2 hold, Assumption 4.4 holds for $h$ and $E$, the level set $\{x : f_\beta(x) \leq f_\beta(x^0)\}$ is bounded, and $h$ is strongly convex with $\text{dom } h = \text{dom } h$. Let the parameters $\delta^k$ and $\mu^k$ satisfy (4.8a)-(4.8b) and $\sum_{k=0}^{\infty} (\mu^k - \mu) < \infty$. Let $\{(x^k, p^k)\}$ be the sequence generated by (3.1a)-(3.1b). Then, the sequence $\{x^k\}$ has finite length in the sense that

$$
\sum_{k=0}^{\infty} \|x^{k+1} - x^k\| < \infty.
$$

Moreover, the sequence $\{x^k\}$ converges to a critical point $x^*$ of $E$ in the sense that $\nabla E(x^*) = 0$ if $\beta = 0$. Furthermore, if $R^*$ is assumed to be locally strongly convex, then the dual sequence $\{p^k\}$ also converges and the limit point $x^*$ of $\{x^k\}$ is a critical point of $f_\beta$ in the sense that $0 \in \nabla E(x^*) + \beta \partial R(x^*)$.

**Proof.** We divide the proof into two parts. The first part is to show (4.43) by modifying the methodology in [6]. Let us begin with any point $s^* = (x^*, p^*) \in \Omega_0$. Then, there exists an increasing sequence of integers $\{k_i\}_{i \in \mathbb{N}}$ such that $x^{k_i} \to x^*$ as $i \to \infty$. From Lemma 4.7 and recalling that $s^k = (x^k, p^{k-1})$, we know

$$
\lim_{k \to \infty} F_{\beta, \mu}(s^k) = f_\beta(x^*),
$$

(4.44)

Note that the convergent sequence $\{F_{\beta, \mu}(s^k)\}$ is nonincreasing. If there exists an integer $\tilde{k}$ such that $F_{\beta, \mu}(s^k) = f_\beta(x^*)$, then $F_{\beta, \mu}(s^k) = f_\beta(x^*)$ for $k \geq \tilde{k}$ and hence $D_h(x^{k+1}, x^k) = 0$ for $k \geq \tilde{k}$ from (4.43), which implies that $x^k \equiv x^\tilde{k}$ for $k \geq \tilde{k}$ due to the strong convexity of $h$. Hence, the result (4.43) follows trivially. If there does not exist such an index, then it must hold that $F_{\beta, \mu}(s^k) > f_\beta(x^*)$ holds for all $k > 0$. Since $\lim_{k \to \infty} F_{\beta, \mu}(s^k) = f_\beta(x^*)$, for any $\eta > 0$ there must exist an integer $\hat{k} > 0$ such that $F_{\beta, \mu}(s^k) < f_\beta(x^*) + \eta$ for all $k > \hat{k}$. Similarly, $\lim_{k \to \infty} \text{dist}(s^k, \Omega_0) = 0$ implies for any $\zeta > 0$ there must exist an integer $\hat{k} > 0$ such that $\text{dist}(s^k, \Omega_0) < \zeta$ for all $k > \hat{k}$. Therefore, for all $k > l := \max\{\tilde{k}, \hat{k}\}$ we have

$$
s^k \in \{s : \text{dist}(s, \Omega_0) < \zeta\} \cap \{s : f_\beta(x^*) < F_{\beta, \mu}(s) < f_\beta(x^*) + \eta\}.
$$

(4.45)

Thus, we apply Lemma 6.1 to get,

$$
\varphi'(F_{\beta, \mu}(s^k) - f_\beta(x^*)) \text{dist}(0, \partial F_{\beta, \mu}(s^k)) \geq 1.
$$

(4.46)

Recall that $r^k \in \partial F_{\beta, \mu}(s^k)$. Using (4.35) in Lemma 4.7 we get that

$$
\text{dist}(0, \partial F_{\beta, \mu}(s^k)) \leq \|r^k\| \leq p_2\|x^k - x^{k-1}\| + (\mu^{k-1} - \mu)\|p^k - p^{k-1}\|.
$$

(4.47)

On the other hand, from the concavity of $\varphi$ we know that

$$
\varphi'(x) \leq \frac{\varphi(x) - \varphi(y)}{x - y}
$$
holds for all \( x, y \in [0, \eta], x > y \). Hence, by taking \( x = F_{\beta, \mu}(s^k) - f_{\beta}(x^*) \) and \( y = F_{\beta, \mu}(s^{k+1}) - f_{\beta}(x^*) \) in the inequality above, we get

\[
\varphi'(F_{\beta, \mu}(s^k) - f_{\beta}(x^*)) \leq \frac{\varphi^k - \varphi^{k+1}}{F_{\beta, \mu}(s^k) - F_{\beta, \mu}(s^{k+1})} \leq \frac{\varphi^k - \varphi^{k+1}}{\rho \nu \|x^{k+1} - x^k\|^2},
\]

where \( \varphi^k := \varphi(F_{\beta, \mu}(s^k) - f_{\beta}(x^*)) \) and the last inequality follows from (4.34) and the strong convexity of \( D_h(x^{k+1}, x^k) \geq \nu \|x^{k+1} - x^k\|^2 \) for some constant \( \nu > 0 \). Therefore, from (4.46)-(4.48) we get

\[
\|x^{k+1} - x^k\|^2 \leq \frac{\rho_2}{\rho \nu} \left( \varphi^k - \varphi^{k+1} \right) \left( \|x^k - x^{k-1}\| + \frac{\mu^{k-1} - \mu}{\rho_2} \|p^k - p^{k-1}\| \right).
\]

Based on the Young’s inequality of the form \( 2\sqrt{ab} \leq a + b \), we further get

\[
2\|x^{k+1} - x^k\| \leq \frac{\rho_2}{\rho \nu} (\varphi^k - \varphi^{k+1}) + \|x^k - x^{k-1}\| + \frac{\mu^{k-1} - \mu}{\rho_2} \|p^k - p^{k-1}\|.
\]

Summing the inequality above from \( k = l, \ldots, N \), we deduce

\[
2 \sum_{k=l}^{N} \|x^{k+1} - x^k\| \leq \sum_{k=l}^{N} \|x^k - x^{k-1}\| + \frac{\rho_2}{\rho \nu} (\varphi^l - \varphi^{N+1}) + \sum_{k=l}^{N} \frac{\mu^{k-1} - \mu}{\rho_2} \|p^k - p^{k-1}\|
\]

\[
\leq \sum_{k=l}^{N} \|x^{k+1} - x^k\| + \|x^l - x^{l-1}\| + \frac{\rho_2}{\rho \nu} \varphi^l + \sum_{k=l}^{N} \frac{\mu^{k-1} - \mu}{\rho_2} \|p^k - p^{k-1}\|.
\]

where the second inequality follows by using the boundedness of \( \{p^k\} \), say \( \|p^k - p^{k-1}\| \leq \frac{1}{C} \) for some constant \( C > 0 \). Thus, we have

\[
\sum_{k=l}^{N} \|x^{k+1} - x^k\| \leq \|x^l - x^{l-1}\| + \frac{\rho_2}{\rho \nu} \varphi^l + \sum_{k=l}^{N} \frac{\mu^{k-1} - \mu}{\rho_2} \|p^k - p^{k-1}\| < \infty,
\]

which immediately implies the result (4.43).

Now, we turn into the second part to analyze the convergence of \( \{x^k\} \) and \( \{p^k\} \). Let us first show that \( \{x^k\} \) is a Cauchy sequence and hence it converges. In fact, the finite length property implies that \( \sum_{k=l}^{\infty} \|x^{k+1} - x^k\| \to 0 \) as \( l \to \infty \). Thus, for any \( m > n \geq l \) we have

\[
\|x^m - x^n\| = \|\sum_{k=n}^{m-1} (x^{k+1} - x^k)\| \leq \sum_{k=n}^{m-1} \|x^{k+1} - x^k\| \leq \sum_{k=l}^{\infty} \|x^{k+1} - x^k\|,
\]

which implies that \( \{x^k\} \) is a Cauchy sequence. Using (3.11a) with \( \beta = 0 \), we get

\[
p^k - p^{k+1} = \frac{1}{\delta^k \mu_k} (\nabla h(x^{k+1}) - \nabla h(x^k)) + \frac{1}{\mu_k} \nabla E(x^k).
\]

Summing (4.50) over \( k = 0, \ldots, n \), we get

\[
p^0 - p^{n+1} = \sum_{k=0}^{n} \left( \frac{1}{\delta^k \mu_k} (\nabla h(x^{k+1}) - \nabla h(x^k)) + \frac{1}{\mu_k} \nabla E(x^k) \right).
\]
Assume that \( \nabla E(x^*) \neq 0 \). Noting that \( \frac{1}{\mu}(\nabla h(x^{k+1}) - \nabla h(x^k)) \to 0 \) and \( \frac{1}{\mu}E(x^k) \to \frac{1}{\mu}\nabla E(x^*) \neq 0 \), we invoke Lemma 4.3 to conclude that \( \|p^0 - p^{n+1}\| \to \infty \) as \( n \to \infty \), which contradicts the boundedness of \( \{p^k\} \). Therefore, we have \( \nabla E(x^*) = 0 \).

If \( R^* \) is locally strongly convex, then for some constant \( \nu > 0 \),

\[
D_R(x^{k}, x^{k-1}) = D_{R^*}(p^{k-1}, p^k) \geq \frac{\nu}{\mu}\|p^k - p^{k-1}\|^2.
\]

Recall that \( D_h(x^{k+1}, x^k) \geq \nu \|x^{k+1} - x^k\|^2 \). Thus, using (4.34) and letting \( \nu_2 = \min\{\frac{\nu\mu}{2}, \rho\nu\} \), we get

\[
F_{\beta, \mu}(s^{k+1}) + \nu_2\|s^{k+1} - s^k\|^2 \leq F_{\beta, \mu}(s^k).
\]

From (4.35), since \( \|x^{k+1} - x^k\| \leq \|s^{k+1} - s^k\| \) we deduce

\[
\|r^{k+1}\| \leq \rho_2\|s^{k+1} - s^k\| + (\mu^k - \mu)\|p^{k+1} - p^k\|,
\]

Now we repeat the argument of the first part to conclude that \( \{s^k\} \) has a finite length and hence it converges. Hence, \( \{p^k\} \) also converges to some \( p^* \). Note that (3.1b) has the following form

\[
\nabla E(x^k) + \beta p^k = (\mu^k + \beta)(p^k - p^{k+1}) - \frac{1}{\delta k}(\nabla h(x^{k+1}) - \nabla h(x^k)).
\]

Letting \( k \to \infty \) above, we immediately get \( 0 = \nabla E(x^*) + \beta p^* \). Finally, using the subgradient inequality and the lsc property of \( R \), we drive that for any \( x \in \text{dom} R \)

\[
R(x) = \liminf_{k \to \infty} R(x) \geq \liminf_{k \to \infty} (R(x^k) + \langle p^k, x - x^k \rangle) \geq R(x^*) + \langle p^*, x - x^* \rangle,
\]

which implies that \( p^* \in \partial R(x^*) \). Thus, we finally get \( 0 \in \nabla E(x^*) + \beta \partial R(x^*) \). This completes the proof.

**5 Application to inverse problems**

In this section, we introduce two examples to show how our algorithmic framework and its convergence theory can be applied.

**5.1 A convex example: Linear inverse problems**

We have described the linear inverse problems in the introduction. Here, we apply the LBrelIF to linear inverse problems with nonnegative data \( b \in \mathbb{R}^m \), considered in [11] in the following form

\[
\min_{x \in \mathbb{R}^n} \{ \Phi(x) := D_\phi(Ax, b) + \lambda R(x) \},
\]

where the distance \( D_\phi(Ax, b) \) with \( \phi(x) = \sum_{i=1}^n x_i \log x_i \) is adopted to measure the “error” between \( b \) and \( Ax \), \( R(\cdot) \) is a regularizer reflecting prior information on the solution, the parameter \( \lambda > 0 \) balances the data fidelity and the solution regularization. In order to apply our method and theory, we let \( E(x) = D_\phi(Ax, b) + \frac{\lambda}{2}\|Ax - b\|^2 \) and \( h(x) = \sum_{i=1}^n x_i \log x_i + \frac{\mu}{2}\|x\|^2 \). It should be noted that we
here add the terms \( \frac{\epsilon}{2} \| Ax - b \|^2 \) and \( \frac{\epsilon}{2} \| x \|^2 \) with \( \epsilon \geq 0 \) to the distance \( D_\phi(Ax, b) \) and the Boltzmann-Shannon entropy \( \sum_{i=1}^n x_i \log x_i \), respectively. For the former, we aim to meet the condition (4.17) in Theorem 4.1. Actually, in this setting \( E(u, b) = D_\phi(u, b) + \frac{\epsilon}{2} \| u - b \|^2 \) is \( \epsilon \)-strongly convex and and its minimizer \( \hat{u} \) is attainable. Hence,

\[
E(u, b) - E(\hat{u}, b) \geq \frac{\epsilon}{2} \| u - \hat{u} \|^2.
\]

For the latter, we aim to meet the (LC) condition. Indeed, due to Lemma 8 in [1], the (LC) condition holds for the pair of \((E, h)\) with

\[
L \geq \max\{\| A \|^2, \max_{1 \leq j \leq n} \sum_{i=1}^n a_{ij}\}.
\]

Since \( \text{dom} (x_i \log x_i) = [0, \infty) \), we can conclude that \( \text{dom} h = \text{dom} h \). From the expression of

\[
\nabla h(x) = e + \log x + \epsilon x \text{ where } e \text{ stands for the vector whose each entry equals to one, we know that } h \text{ is continuously differentiable on } \text{int dom } h. If A is surjective, then \( Ax = b \) is a consistence system and hence \( X_0 = \text{arg min}\{E(x) : x \in \text{dom } h\} \) must be nonempty. Since \( R(x) = \| x \|_1 \) is real-valued convex, Assumptions 4.1.3 about \( R \) can be easily verified to hold. Regarding to the computation, we can reduce (3.1a) to get

\[
x^{k+1} = \text{arg min}_x \{ \delta \mu \| x \|_1 + \langle \overline{p}^k, x \rangle + \sum_{i=1}^n x_i \log x_i + \frac{\epsilon}{2} \| x \|^2 \}, \quad (5.2)
\]

where \( \overline{p}^k = \delta \nabla E(x^k) - \nabla h(x^k) - \delta \mu p^k \). Thus, the entries \( x^{k+1} \) are the roots of the equations \( \log x_i + \epsilon x_i + C_i = 0, i = 1, \ldots, m \) with the constants \( C_i = \delta \mu + \overline{p}^k_i + 1 \). As a consequence of Theorem 4.1, we have the following convergence result for the generated sequence \( \{x^k\} \).

**Corollary 5.1.** Let \( E, h \) and \( R \) be given above with a surjective \( A \). Then, the sequence \( \{x^k\} \), generated by (5.2) with \( x^0 = \rho^0 = 0 \), converges either to the unique minimizer \( \hat{x} := \text{arg min}\{\mu \| x \|_1 + \frac{\epsilon}{2} \| x \|^2 + \frac{1}{\delta} \sum_{i=1}^n x_i \log x_i : x \in X_0 \} \) or to a boundary point of \( \text{dom } h \).

### 5.2 A nonconvex example: Quadratic inverse problems

We briefly describe the quadratic inverse problems as follows. Given a finite number of symmetric matrices \( A_i \in \mathbb{R}^{d \times d}, i = 1, 2, \ldots, m \), modeling the measure procedure, and a vector \( b \in \mathbb{R}^m \) recording the measure data \( (b_1, b_2, \ldots, b_m) \), the goal is to find \( x \in \mathbb{R}^d \) such that

\[
x^T A_i x \simeq b_i, \quad i = 1, 2, \ldots, m. \quad (5.3)
\]

As a natural extension of the linear inverse problems, the quadratic inverse problems arise in the broad area of signal processing, including for example the phase retrieval problems [21] as special cases. Similar to the linear inverse problems, the system of quadratic equations (5.3) is usually underdetermined. Thus, there also needs the regularization technique to help find the “right” solution via solving the regularized optimization problem

\[
\min_{x \in \mathbb{R}^d} \{ \Psi(x) := \frac{1}{4} \sum_{i=1}^m (x^T A_i x - b_i)^2 + \lambda R(x) \}, \quad (5.4)
\]
where the least-squares models the “error” between $b_i$ and $x^TA_ix$, $R(\cdot)$ is a regularizer reflecting prior information on the solution, the parameter $\lambda > 0$ balances the data fidelity and the solution regularization. Instead of solving \eqref{5.1}, we consider the LBreIF for \eqref{5.3} with sparse prior. To this end, we let $h(x) := \frac{1}{2}\|x\|^4 + \frac{1}{2}\|x\|^2$, $R(x) = \|x\|_1$, and

$$E(x) := \frac{1}{4}\sum_{i=1}^m (x^TA_ix - b_i)^2 + \frac{\epsilon}{2}\|x\|^2.$$  

Here, the term $\frac{\epsilon}{2}\|x\|^2$ with $\epsilon > 0$ is added to ensure the level boundedness of $E(x)$. Due to Lemma 5.1 in \cite{7}, the (LC) condition holds for the pair of $(E, h)$ with

$$L \geq \sum_{i=1}^m (3\|A_i\|^2 + \|A_i \cdot |b_i|\) + \epsilon.$$  

Note that

$$\nabla h(x) = (\|x\|^2 + 1)x, \quad \nabla E(x) = \sum_{i=1}^m (x^TA_ix - b_i)A_ix + \epsilon x.$$  

Now, the iterate sequence $\{x^k\}$, based on LBreIF, is defined via solving

$$x^{k+1} := \arg\min_x \{\delta^k \mu^k \|x\|_1 + \langle \overline{p}^k, x \rangle + \frac{1}{4}\|x\|^4 + \frac{\epsilon}{2}\|x\|^2\},$$  

where $\overline{p}^k = \delta^k \nabla E(x^k) - \nabla h(x^k) - \delta^k \mu^k \eta^k$. Following the same spirit in \cite{7}, we can deduce the closed-form formula $x^{k+1} = -t^* S_{\delta^k \mu^k}(\overline{p}^k)$, where $t^*$ is the unique positive real root of $t^3 \|S_{\delta^k \mu^k}(\overline{p}^k)\|^2 + t - 1 = 0$ and $S_{\delta^k \mu^k}(\overline{p}^k)$ is the soft-thresholding operator $S_{\delta^k \mu^k}(\overline{p}^k) := \max\{|\overline{p}^k| - \delta^k \mu^k, 0\}\text{sign}(|\overline{p}^k|)$.

For any bounded set $\Omega_1 \subset \mathbb{R}^d$, it is easy to verify that both $\nabla h$ and $\nabla E$ are Lipschitz continuous on $\Omega_1$ and hence Assumption 4.4 holds for $E$ and $h$. Assumption 3.1 (ii) of Assumption 4.2 and Assumption 4.3 can also be easily verified for $E$ and $h$. Besides, it is not hard to see that $h$ is strongly convex with dom $h = \mathbb{R}^d$. Therefore, we can apply Theorem 4.2 to get the following result.

**Corollary 5.2.** Let the parameters $\delta^k$ and $\mu^k$ satisfy \eqref{4.8a}--\eqref{4.8b} and $\sum_{k=0}^{\infty} (\mu^k - \mu) < \infty$. Let $\{x^k\}$ be the sequence generated by \eqref{5.5} with $x^0 = p^0 = 0$. Then, the sequence $\{x^k\}$ has finite length. Moreover, the sequence $\{x^k\}$ converges to a critical point $x^*$ of $E$ if $\beta = 0$ and $\lim_{k \to \infty} E(x^k) = E(x^*)$.

### 6 Concluding remarks

The Linearized Bregman iteration was revisited in this paper from the perspective of going beyond the traditional restriction—gradient Lipschitz continuity. In convex case, we found that the important convergence result of LBreIF-Theorem 4.1 still remains true even if the gradient Lipschitz continuity is replaced by the Lipschitz-like convexity condition; In nonconvex case, we were able to show global convergence under some mild assumptions. At last, we presented two examples to demonstrate the widespread application of the proposed algorithmic framework—LBreIF.

In future, we will study the convergence rate of LBreIF and consider randomized variants of LBreIF for large-scale optimization problems.
Appendix

Definition 6.1 (Kurdyka-Lojasiewicz property and function, [6]). (a) The function \( \sigma : \mathbb{E} \to (-\infty, +\infty] \) is said to have the Kurdyka-Lojasiewicz property at \( x^\ast \in \text{dom}(\partial \sigma) \) if there exist \( \eta \in (0, +\infty] \), a neighborhood \( U \) of \( x^\ast \) and a continuous concave function \( \varphi : [0, \eta) \to \mathbb{R}_+ \) such that

1. \( \varphi(0) = 0 \).
2. \( \varphi \) is \( C^1 \) on \((0, \eta)\).
3. for all \( s \in (0, \eta) \), \( \varphi'(s) > 0 \).
4. for all \( x \) in \( U \cap [\sigma(x^\ast) < \sigma < \sigma(x^\ast) + \eta] \), the Kurdyka-Lojasiewicz inequality holds

\[
\varphi'(\sigma(x) - \sigma(x^\ast)) \text{dist}(0, \partial \sigma(x)) \geq 1. \tag{6.1}
\]

(b) Proper lower semicontinuous functions which satisfy the Kurdyka-Lojasiewicz inequality at each point of \( \text{dom}(\partial \sigma) \) are called KL functions.

Lemma 6.1 (Uniformized KL property, [6]). Let \( \Omega \) be a compact set and let \( \sigma : \mathbb{E} \to (-\infty, +\infty] \) be a proper lsc function. Assume that \( \sigma \) is constant on \( \Omega \) and satisfies the KL property at each point of \( \Omega \). Then, there exist \( \zeta > 0, \eta > 0 \) and \( \varphi \) satisfies the same conditions as in Definition 6.1 such that for all \( x^\ast \in \Omega \) and all \( x \) in

\[
\{x : \text{dist}(x, \Omega) < \zeta\} \cap \{x : \sigma(x^\ast) < \sigma(x) < \sigma(x^\ast) + \eta\} \tag{6.2}
\]

the condition (6.1) holds.

Acknowledgements

The first author was supported by the National Science Foundation of China (No.11971480), the Natural Science Fund of Hunan for Excellent Youth (No.2020JJ3038), and the Fund for NUDT Young Innovator Awards (No. 20190105).

References

[1] H. H. Bauschke, J. Bolte, and M. Teboulle, A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications, Math. Oper. Res., 42 (2016), pp. 330–348.
[2] H. H. Bauschke and J. M. Borwein, Legendre functions and the method of random bregman projections, Journal of Convex Analysis, 4 (1997), pp. 27–67.
[3] A. Beck, First-Order Methods in Optimization, SIAM-Society for Industrial and Applied Mathematics, 2017.
[4] M. Benning, M. M. Betcke, M. J. Ehrhardt, and C.-B. Schönlieb, Choose your path wisely: Gradient descent in a bregman distance framework, SIAM Journal on Imaging Sciences, 14 (2021), pp. 814–843.
[5] M. Benning, C. Brune, M. Burger, and J. Miller, Higher-order tv methods - enhancement via bregman iteration, Journal of Scientific Computing, 54 (2013), pp. 269–310.
[6] J. Bolte, S. Sabach, and M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, Mathematical Programming, 146 (2014), pp. 459–494.

[7] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd, First order methods beyond convexity and lipschitz gradient continuity with applications to quadratic inverse problems, SIAM J. Optim., 28 (2018), pp. 2131–2151.

[8] L. M. Bregman, The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming, USSR Computational Mathematics and Mathematical Physics, 7 (1967), pp. 200–217.

[9] M. Burger, G. Gilboa, S. Osher, and J. Xu, Nonlinear inverse scale space methods, Communications in Mathematical Sciences, 4 (2006), pp. 179–212.

[10] J.-F. Cai, E. J. Candès, and Z. Shen, A singular value thresholding algorithm for matrix completion, SIAM Journal on Optimization, 20 (2010), pp. 1956–1982.

[11] J.-F. Cai, S. Osher, and Z. Shen, Linearized bregman iterations for compressed sensing, Mathematics of Computation, 78 (2009), pp. 1515–1536.

[12] J.-F. Cai, S. Osher, and Z. Shen, Linearized bregman iterations for frame-based image deblurring, SIAM Journal on Imaging Sciences, 2 (2009), pp. 226–252.

[13] J.-F. Cai and S. O. Shen, Convergence of the linearized bregman iteration for $\ell_1$-norm minimization, Mathematics of Computation, 78 (2009), pp. 2127–2136.

[14] G. Chen and M. Teboulle, Convergence analysis of a proximal-like minimization algorithm using bregman functions, SIAM J. Optim., 3 (1993), pp. 538–543.

[15] S. S. Chen, D. L. Donoho, and M. A. Saunders, Atomic decomposition by basis pursuit. reprinted from siam j. sci. comp. 20, SIAM Review, (2001), pp. 129–159.

[16] K. C. Kiwiel, Free-steering relaxation methods for problems with strictly convex costs and linear constraints, Math. Oper. Res., 22 (1997), pp. 326–349.

[17] Proximal minimization methods with generalized Bregman functions, SIAM J. Control Optim., 35 (1997), p. 1142C1168.

[18] M.-J. Lai and W. Yin, Augmented $\ell_1$ and nuclear-norm models with a globally linearly convergent algorithm, SIAM Journal on Imaging Sciences, 6 (2013), pp. 1059–1091.

[19] D. A. Lorenz, F. Schöpfer, and S. Wenger, The linearized Bregman method via split feasibility problems: Analysis and generalizations, SIAM Journal on Imaging Sciences, 7 (2014), pp. 1237–1262.

[20] H. Lu, R. M. Freund, and Y. Nesterov, Relatively smooth convex optimization by first-order methods, and applications, SIAM J. Optim., 28 (2018), pp. 333–354.

[21] D. R. Luke, Phase retrieval, what’s new?, SIAG/OPT ViewsNews, 25 (2017), pp. 1–5.

[22] M. Moeller, E. M. Brinkmann, M. Burger, and T. Seybold, Color Bregman tv, Siam Journal on Imaging Sciences, 7 (2014), pp. 2771–2806.

[23] A. S. Nemirovski, Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems, SIAM J. Optim., 15 (2004), pp. 229–251.

[24] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, An iterative regularization method for total variation-based image restoration, Multiscale Model. Simul., 4 (2005), pp. 460–489.

[25] G. Polyak, Introduction to Optimization, Optimization Software, Inc., New York, 1987.

[26] R. T. Rockafellar, Convex analysis, Princeton university press, 2015.
[27] W. Rudin, *Principles of Mathematical Analysis*, MC GRAW HILL INDIA, 2013.

[28] M. Teboulle, *A simplified view of first order methods for optimization*, Math. Program., (2018), pp. 1–30.

[29] J. Xu and S. Osher, *Iterative regularization and nonlinear inverse scale space applied to wavelet-based denoising*, IEEE Trans Image Process, 16 (2007), pp. 534–544.

[30] W. Yin, *Analysis and generalizations of the linearized bregman method*, SIAM Journal on Imaging Sciences, 3 (2010), pp. 856–877.

[31] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, *Bregman iterative algorithms for \( \ell_1 \)-minimization with applications to compressed sensing*, SIAM Journal on Imaging Sciences, 1 (2008), pp. 143–168.

[32] H. Zhang, L. Cheng, and W. Yin, *A dual algorithm for a class of augmented convex models*, Communications in mathematical sciences, 13 (2015), pp. 103–112.

[33] H. Zhang and Y. H. Dai, *Mirror frameworks for relatively lipschitz and monotone-like variational inequalities*, [arXiv:2108.12070](https://arxiv.org/abs/2108.12070) [math.OC], (2021).

[34] H. Zhang and W. Yin, *Gradient methods for convex minimization: better rates under weaker conditions*, CAM Report 13-17, UCLA, (2013).

[35] X. Zhang, M. Burger, X. Bresson, and S. Osher, *Bregmanized nonlocal regularization for deconvolution and sparse reconstruction*, SIAM Journal on Imaging Sciences, 3 (2010), pp. 253–276.