On Linear Convergence of non-Euclidean Gradient Methods without Strong Convexity and Lipschitz Gradient Continuity

Heinz H. Bauschke[*], Jérôme Bolte[†], Jiawei Chen[‡], Marc Teboulle[§] and Xianfu Wang[¶]

October 17, 2018

Abstract

The gradient method is well known to globally converge linearly when the objective function is strongly convex and admits a Lipschitz continuous gradient. In many applications, both assumptions are often too stringent, precluding the use of gradient methods. In the early 60, after the amazing breakthrough of Łojasiewicz on gradient inequalities, it was observed that uniform convexity assumptions could be relaxed and replaced by these inequalities. On the other hand, very recently, it has been shown that the Lipschitz gradient continuity can be lifted and replaced by a class of functions satisfying a Non-Euclidean descent property expressed in terms of a Bregman distance. In this note, we combine these two ideas to introduce a class of non-Euclidean gradient-like inequalities, allowing to prove linear convergence of a Bregman gradient method for nonconvex minimization, even when neither strong convexity nor Lipschitz gradient continuity holds.

Keywords: Non-Euclidean gradient methods, nonconvex minimization, Bregman distance, Lipschitz-like convexity condition, descent lemma without Lipschitz gradient, gradient dominated inequality, Łojasiewicz gradient inequality, linear rate of convergence.

AMS 2010 Mathematics Subject Classification: Primary 65K05, 49M10, 90C26; Secondary 90C30, 65K10

1 Introduction

The gradient descent method is one of the oldest and most fundamental first order iterative algorithm in optimization. Its low computational complexity makes it an ideal algorithm for solving very large scale problems where medium accuracy is sufficient. Modern applications which are often very large or even huge scale have thus provoked a resurgence of gradient based schemes. This can be seen through the intensive recent research activities in many disparate fields, e.g., machine learning, signal processing, image sciences, communication systems, see for instance [32, 36] and references therein, as well as the more recent books [13, 9].

[*] Mathematics, Irving K. Barber School, University of British Columbia Okanagan, Kelowna, British Columbia V1V 1V7, Canada. E-mail: heinz.bauschke@ubc.ca.

[†] Toulouse School of Economics, Université Toulouse 1 Capitole, 31015 Toulouse, France. E-mail: jerome.bolte@ut-capitole.fr.

[‡] School of Mathematics and Statistics, Southwest University, Chongqing 400715, China. E-mail: J.W.Chen713@163.com.

[§] School of Mathematical Sciences, Tel Aviv University, Ramat Aviv 69978, Israel. E-mail: teboulle@post.tau.ac.il.

[¶] Mathematics, Irving K. Barber School, University of British Columbia Okanagan, Kelowna, British Columbia V1V 1V7, Canada. E-mail: shawn.wang@ubc.ca.

A central assumption in most first order minimization methods is to require Lipschitz continuity of the gradient of the objective function. This property implies the fundamental descent lemma (see e.g., [13]). In turn, it allows to establish the following well-known inequality for the unconstrained minimization of a function $f \in C_L^{1,1}$, namely a differentiable function with an L-smooth gradient:

$$\frac{1}{2L}\|\nabla f(x)\|^2 \leq f(x) - f_*, \ \forall x \in \mathbb{R}^n, \tag{1.1}$$

where $f_* := \min\{f(x) : x \in \mathbb{R}^n\} > -\infty$. The latter inequality combined with the assumption that f is also assumed σ-strongly convex, implies the linear rate of the gradient method for such a class of functions [33]. In the same work, it is observed that a *reverse* inequality (1.1),

$$\exists c > 0 : \ (\forall x \in \mathbb{R}^n) \ \|\nabla f(x)\|^2 \geq c(f(x) - f_*), \tag{1.2}$$

allows to relax the strong convexity assumption on the function f, (which is in fact implied by (1.2) with $c := 2\sigma$), and yet preserve the linear rate of convergence for the class of nonconvex functions $f \in C_L^{1,1}$. This inequality (1.2) was actually a special case of the much more general Łojasiewicz gradient inequality [27] which holds for a wide class of functions, including all real-analytic functions, and which was a turning point for real semi-algebraic/semi-analytic geometry [28]. Today's strong renewed interest for these inequalities in optimization is due to a nonsmooth Łojasiewicz inequality [15], which allows to cope with many genuine optimization problems featuring constraints and nonsmoothness, see [18]. In a recent paper [17] complexity results for first order methods in this broad setting are provided under a Łojasiewicz type assumption. In the present work, we make an attempt to understand how these inequalities could be generalized in a non-Euclidean setting, but we shall avoid the terminology Łojasiewicz inequality (which could be misleading) and we will pertain to the more neutral vocabulary of gradient dominated functions. Before considering these questions let us present our enlarged framework by recalling how the lack of Lipschitz continuity of the gradient can be dealt with (see e.g., Bertero et al. [12] for examples of problems without Lipschitz gradient).

Recently, Bauschke, Bolte, and Teboulle [7] recast the Lipschitz gradient condition into a simple and more general convexity assumption whose generalization is immediate and far reaching. Their approach captures all at once the geometry of a given minimization problem, the key point being the existence of a powerful descent lemma involving the so-called Bregman distance [19]; see Section 2 for details. For a related approach, see also the recent work [31]. This opened a new path to a broad spectrum of optimization models arising in many applications, see e.g., the very recent work [18] and references therein, which further investigated the Bregman proximal gradient method for nonconvex nonsmooth minimization problem. Sublinear efficiency estimates in terms of value functions were established in [7] for Bregman based proximal gradient methods in the convex setting. Moreover, as shown in [38, Proposition 4.1], linear convergence of the Bregman proximal gradient method follows as an easy consequence of the framework given in [7], by assuming a Bregman-like strong convexity assumption [3], see Section 4. In this work, we focus on a Bregman gradient scheme for nonconvex objective functions, see Section 2. Motivated by the above old [27, 33] and new [7] ideas, this paper attempts to answer the following challenging natural question:

Can we develop a linear convergence "theory" for the gradient descent in the framework of Bregman distances freed from Lipschitz gradient continuity and any type of strong convexity?

More specifically, our main objective is to discover what should be a gradient-like dominated inequality with respect to a kernel function defining a Bregman distance, that would: (i) naturally replace its Euclidean version given through the norm squared of a gradient in (1.2); (ii) allow to derive linear convergence of Bregman gradient methods for minimizing nonconvex functions lacking strong convexity and Lipschitz continuity. Our approach towards these goals is developed in Section 3 where we introduce the key tools. Global linear convergence of the Bregman gradient scheme is derived in Section 4.

Notation. The notation we employ is standard and follows, e.g., [8, 34, 35]. Let $\mathbb{R}^n$ be the n-dimensional Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\|x\| := \sqrt{\langle x, x \rangle}$ for $x \in \mathbb{R}^n$. We set $\mathcal{Q}(x) := \|x\|^2/2$ the energy function. For a subset C in $\mathbb{R}^n$, its interior and closure are respectively denoted by $\operatorname{int} C$ and $\overline{C}$. For a function $f : \mathbb{R}^n \to]-\infty, +\infty]$, we use $\operatorname{dom} f$ for its domain, and f^* for its Legendre-Fenchel conjugate.

We let $\mathbb{R}_+ := [0, +\infty[$ and $\mathbb{R}_{++} :=]0, +\infty[$.

2 Preliminaries, the Problem and the Algorithm

In this section, we describe the optimization problem setting, and basic algorithm. We start with some auxiliary results on Bregman distances and symmetric coefficients of Legendre functions. The gradient descent method described below in terms of Bregman distances (termed Bregman Gradient (BG) algorithm) we study is not new, and it has been investigated by many researchers under various conditions on the problem's data. The literature on the subject is wide and it is not our intent to further elaborate on these. For more details and applications on first order methods based on Bregman distances, including earlier works as well as recent ones, see the very recent survey [38] and references therein.

2.1 Bregman Distances

Definition 2.1 (Legendre function) *[34, Section 26]. Let $h : \mathbb{R}^n \to]-\infty, +\infty]$ be a lsc proper convex function. We say that h is Legendre if h is essentially smooth and strictly convex on $\operatorname{int} \operatorname{dom} h \neq \emptyset$.*

Essentially smooth means that h is differentiable on $\operatorname{int} \operatorname{dom} h \neq \emptyset$ with $\|\nabla h(x^k)\| \to \infty$ for each sequence $(x^k)_{k \in \mathbb{N}} \subset \operatorname{int} \operatorname{dom} h$ converging to a boundary point of $\operatorname{dom} h$ as $k \to \infty$.

Note that we have adopted here the simple terminology for a *Legendre Function*. Functions of *Legendre type* are defined in [34] under *essential* strict convexity (i.e., strict convexity on every convex subset of $\operatorname{dom} \partial h$). When h is essentially smooth, essential strict convexity reduces to strict convexity on $\operatorname{int} \operatorname{dom} h$; for more details and more general facts on functions of Legendre type, see [34, Section 26].

Recall that with h Legendre, so is its conjugate h^*, and the following useful properties hold:

- $\operatorname{dom} \nabla h = \operatorname{int} \operatorname{dom} h$, $\operatorname{dom} \nabla h^* = \operatorname{int} \operatorname{dom} h^*$ and $\operatorname{ran} \nabla h = \operatorname{dom} \nabla h^*$.
- $(\nabla h)^{-1} = \nabla h^*$ and $h^*(\nabla h(x)) = \langle x, \nabla h(x) \rangle - h(x)$ $\forall x \in \operatorname{int} \operatorname{dom} h$.

Using the Legendre function h, the Bregman distance [19] associated with h is defined by

$$D_h(x, y) := h(x) - h(y) - \langle \nabla h(y), x - y \rangle, \forall x \in \operatorname{dom} h, y \in \operatorname{int} \operatorname{dom} h.$$

Bregman distance is a proximity measure in the sense that $D_h(x, y) \geq 0$ and from the strict convexity of $D_h(\cdot, y)$, we have $D_h(x, y) = 0 \;\Leftrightarrow\; x = y,\; \forall\, (x, y) \in \operatorname{dom} h \times \operatorname{int} \operatorname{dom} h$.

When no confusion arises, we also write $D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle$ even when f is not convex.

The proximity measure defined through the Bregman distance D_h has been extensively studied. For early developments, examples of Bregman distances, as well and many other useful properties, see e.g., [26, 21, 37, 22, 24, 20, 16, 4] and references therein. Below, we recall some of these basic properties that will be used repeatedly in sequel, and give some classical examples.

(P1) $D_h(x,y) = D_h(x,z) + D_h(z,y) + \langle \nabla h(z) - \nabla h(y), x - z\rangle,\ \forall\, x \in \mathbb{R}^n,\ y,z \in \operatorname{int}\operatorname{dom} h.$

(P2) $D_h(x,y) = D_{h^*}(\nabla h(y), \nabla h(x)),\ \forall\, x,y \in \operatorname{int}\operatorname{dom} h.$

(P3) On the set of differentiable functions, $h \mapsto D_h$ is a linear operator, namely,

$$(\forall h_1, h_2 \in \mathcal{C}^1)(\forall \beta_1, \beta_2 \in \mathbb{R})\ D_{\beta_1 h_1 + \beta_2 h_2} = \beta_1 D_{h_1} + \beta_2 D_{h_2}.$$

In $\mathbb{R}^n$, often one uses Legendre functions with separable structure $h(x) = \sum_{i=1}^n h_i(x_i)$ where $h_i : \mathbb{R} \to\ ]-\infty, +\infty]$ is Legendre. Then $h^*(y) = \sum_{i=1}^n h_i^*(y_i)$ and

$$\nabla h(x) = (h_1'(x_1), \dots, h_n'(x_n)), \text{ and } \nabla h^*(y) = ((h_1^*)'(y_1), \dots, (h_n^*)'(y_n)).$$

The corresponding Bregman distance D_h then preserves this separable structure. Below we list well-known concrete examples.

Example 2.2 Let $h : \mathbb{R} \to\ ]-\infty, +\infty]$.

(i) The function $h(x) = |x|^p/p$, $p > 1$. It has $h^*(y) = |y|^q/q$, $h'(x) = |x|^{p-1}\operatorname{sign} x$, and $(h^*)'(y) = |y|^{q-1}\operatorname{sign} y$, where $1/p + 1/q = 1$. In particular, the energy function $h(x) = x^2/2$ has $h^*(y) = y^2/2$, $h'(x) = x$, and $(h^*)'(y) = y$.

(ii) The Boltzmann-Shannon entropy $h(x) = x \ln x - x$, $\operatorname{dom} h = [0, +\infty)$. It has $h^*(y) = e^y$, $h'(x) = \ln x$, and $(h^*)'(y) = e^y$.

(iii) The Burg entropy $h(x) = -\ln x$, $\operatorname{dom} h = \mathbb{R}_{++}$. It has $h^*(y) = -\ln(-y) - 1$ with $\operatorname{dom} h^* = \mathbb{R}_{--}$, $h'(x) = -1/x$, and $(h^*)'(y) = -1/y$.

(iv) The strongly convex Fermi-Dirac entropy $h(x) = x \ln x + (1-x)\ln(1-x)$, $\operatorname{dom} h = [0,1]$. It has $h^*(y) = \ln(e^y + 1)$, $h'(x) = \ln \frac{x}{1-x}$, and $(h^*)'(y) = e^y/(1+e^y)$.

(v) The strongly convex Hellinger entropy $h(x) = -\sqrt{1-x^2}$, $\operatorname{dom} h = [-1,1]$. It has $h^*(y) = \sqrt{1+y^2}$, $h'(x) = -x/\sqrt{1-x^2}$, and $(h^*)'(y) = y/\sqrt{1+y^2}$.

(vi) The fractional power $h(x) = px - x^p/(1-p)$, $\operatorname{dom} h = \mathbb{R}_+$. It has $h^*(y) = (1 + y/q)^q$, $h'(x) = p - px^{p-1}/(1-p)$, and $(h^*)'(y) = (1+y/q)^{q-1}$. Here $\operatorname{dom} h^* = (-\infty, -q], 0 < p < 1, 1/p + 1/q = 1$.

Symmetry coefficient of a Legendre function

In general, D_h is not symmetric. Thus, in order to relate $D_h(x,y)$ to $D_h(y,x)$, the following useful measure for the lack of symmetry in D_h was introduced in [7].

Definition 2.3 (symmetry coefficient) (see [7]) Given a Legendre function $h : \mathbb{R}^n \to]-\infty, \infty]$, its symmetry coefficient is defined by

$$\alpha(h) := \inf\left\{ \frac{D_h(x,y)}{D_h(y,x)} : (x,y) \in \operatorname{int}\operatorname{dom} h \times \operatorname{int}\operatorname{dom} h,\ x \neq y \right\} \in [0,1]. \tag{2.1}$$

Clearly, $\alpha(h) = 1$ if and only if D_h is symmetric. Moreover, as noted in [7], the latter happens if and only if h is a strictly convex linear-quadratic function; this can be deduced from [6, Lemma 3.16].

The following result collects some useful properties of the symmetry coefficient $\alpha(h)$, which readily follows from the definition, see [7].

Fact 2.4 (i) *$\alpha(ph) = \alpha(h)$ for $p > 0$.*

(ii) *$\alpha(h) = \alpha(h^*)$.*

(iii) *If $\operatorname{dom} h$ is not open, then $\alpha(h) = 0$.*

(iv) *$(\forall x \in \operatorname{int}\operatorname{dom} h)(\forall y \in \operatorname{int}\operatorname{dom} h) \quad \alpha(h)D_h(x,y) \le D_h(y,x) \le \alpha(h)^{-1}D_h(x,y)$, where we have adopted the convention that $0^{-1} = +\infty$ and $+\infty \times r = +\infty$ for all $r \ge 0$.*

As shown in [7], total lack of symmetry, i.e., $\alpha(h) = 0$ occurs both for the Boltzmann–Shannon (this follows from (iii)), and the Burg entropies respectively given by $h(x) = \sum_{i=1}^{n} x_i \ln x_i$, and $h(x) = -\sum_{i=1}^{m} \log x_i$, while $\alpha(h) = 2 - \sqrt{3} > 0$ for $h(x) = x^4$. More generally, it can be verified that $\alpha(h) > 0$ for the Legendre function on $\mathbb{R}$ given by $h(x) = x^{2p}$ for every $p \ge 1$. Likewise, if h is σ-strongly convex and L-smooth on $\operatorname{int}\operatorname{dom} h$, with $\sigma, L > 0$, (i.e., $h - \sigma\mathcal{Q}$, and $L\mathcal{Q} - h$ are convex, which equivalently reads as $\sigma D_{\mathcal{Q}} \le D_h \le LD_{\mathcal{Q}}$), then it immediately follows that $\alpha(h) \ge \sigma/L > 0$.

2.2 The Problem, Algorithm and Blanket Assumption

We consider the following nonconvex minimization problem:

$$(\mathcal{P}) \quad \nu(\mathcal{P}) := \inf_{x \in C} f(x), \text{ where } C := \overline{\operatorname{dom} h}, \tag{2.2}$$

under the following standing assumption:

Assumption A

(i) $h : \mathbb{R}^n \to \left]-\infty, +\infty\right]$ is a Legendre function.

(ii) $f : \mathbb{R}^n \to \left]-\infty, +\infty\right]$ is a lower semicontinuous (lsc) function with $\operatorname{dom} f \supset \operatorname{dom} h$ which is differentiable on $\operatorname{int}\operatorname{dom} h$.

(iii) $\inf_{x \in \operatorname{dom} h} f(x) = \inf_{x \in \overline{\operatorname{dom} h}} f(x) = \nu(\mathcal{P}) > -\infty$.

To solve $(\mathcal{P})$, we consider the following *Bregman-Gradient*, BG for short, algorithm, which replaces the squared Euclidean distance in the usual gradient method with a Bregman distance:

BG – Bregman Gradient Algorithm

For any initial point $x_0 \in \operatorname{int}\operatorname{dom} h$, and chosen stepsize $\lambda > 0$, generate the sequence (x_k) via the iteration

$$x_{k+1} := \nabla h^* \left(\nabla h(x_k) - \lambda \nabla f(x_k) \right), \; k = 0, 1, \ldots$$

The iterative step of BG is modeled on the gradient method whereby the usual squared Euclidean norm regularization of the linearization of f at x_k is replaced by a Bregman distance, namely:

$$x_{k+1} = \operatorname*{argmin}_{y \in \mathbb{R}^n} \left(f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{1}{\lambda} D_h(y, x_k) \right), \tag{2.3}$$

As already mentioned above, this algorithm is not new, and has appeared in the literature under various settings and names, see e.g., [1, 2, 11, 16, 30] and references therein.

It is clear that once we know h^*, the computation in the iterative step of BG is straightforward. The common choices of Legendre functions exhibited in Example 2.2 all admit explicit Legendre conjugates h^*.

For simplicity, to warrant the well-posedness of BG, throughout the paper we make the following assumption.

Assumption B There exists $c > 0$ such that

$$(\forall x \in \operatorname{int}\operatorname{dom} h)\ \nabla h(x) - c\nabla f(x) \in \operatorname{int}\operatorname{dom} h^*.$$

Note in particular, that Assumption B holds if h is cofinite, i.e., $\operatorname{dom} h^* = \mathbb{R}^n$. For other sufficient conditions implying the well-posedness of x^{k+1}, see e.g. [7, 18, 38].

Lemma 2.5 *The iterative step is well-defined under* Assumption B.

Proof. In (2.3), the minimizer is unique thanks to the strict convexity of $y \mapsto D_h(y, x_k)$. The existence of minimizer is guaranteed by Assumption B, and the fact that $x_{k+1} \in \operatorname{int}\operatorname{dom} h$ because h is Legendre. ■

Throughout the rest of this paper, Assumptions A and B are our blanket assumptions.

3 The Toolbox

In this section, we introduce the key new objects/conditions needed to study the linear convergence of the Bregman gradient descent method.

3.1 Lipschitz-like Convexity – (LC) Condition

We adopt the recent framework of [7], where it was shown that the usual Lipschitz gradient continuity assumption can be replaced by a more flexible Lipschitz-like convexity (LC) condition, which captures the geometry of the problem's data, and defined as follows.

Definition 3.1 (LC-condition) A pair of functions (f, h) satisfies the *L-convexity condition* (LC-condition), if there exists $L > 0$ such that $Lh - f$ is convex on $\operatorname{int}\operatorname{dom} h$.

The LC-condition is equivalent to a descent lemma [7, Lemma 1], but there it was assumed that f is convex. However, as already observed in [18, 38], the convexity assumption of f plays no role in the LC-condition, and f can be nonconvex.[1]

Lemma 3.2 (descent lemma) *The LC-condition for the pair of functions (f, h) is equivalent to*

$$(\forall (x, y) \in \operatorname{int}\operatorname{dom} h \times \operatorname{int}\operatorname{dom} h) \quad f(x) \le f(y) + \langle \nabla f(x), x - y\rangle + LD_h(x, y), \tag{3.1}$$

i.e., $(\forall (x, y) \in \operatorname{int}\operatorname{dom} h \times \operatorname{int}\operatorname{dom} h)\ D_f(x, y) \le LD_h(x, y)$.

Proof. The proof is as in [7, Lemma 1]. In fact, it is immediate since $Lh - f$ convex is equivalent to $D_{Lh-f} = LD_h - D_f \ge 0$, where the equality uses the linearity property (P3) of D_h. ■

Clearly, when $h = \mathcal{Q}$, Lemma 3.1 yields the standard descent lemma [13]. Moreover, if f is assumed convex, and $h = \mathcal{Q}$, then the LC-condition is equivalent to the usual L-Lipschitz continuity of ∇f. In general the LC-condition for (f, h) does not imply that the gradient operator ∇f is L-Lipschitz continuous on $\operatorname{int}\operatorname{dom} h$ as the following examples show.

[1] We can also handle a nonsmooth function f assumed locally Lipschitz on $\operatorname{int}\operatorname{dom} h$. It that case, f admits a Clarke-Rockafellar subdiffferential ∂f, [23]. Then $\nabla f(x)$ can be replaced by a subgradient element in ∂f, and our results hold in this larger setting. For simplicity and clarity of exposition, we work with differentiable f.

Example 3.3 Let $h(x) := -\ln x$ on $\operatorname{dom} h = \mathbb{R}_{++}$ and $f(x) := -x^{-2}$ on $\operatorname{dom} f = \mathbb{R} \setminus \{0\}$. Then $\operatorname{dom} h \subset \operatorname{dom} f$, $\nabla f(x) = 2x^{-3}$ for all $x \in \operatorname{int} \operatorname{dom} h$, and the second order derivative $(Lh(x) - f(x))'' = \frac{L}{x^2} + \frac{6}{x^4} > 0$ for all $x \in \operatorname{int} \operatorname{dom} h$, where $L > 0$. So, $Lh(x) - f(x)$ is convex on $\operatorname{int} \operatorname{dom} h$, i.e., the LC-condition for the pair of functions (f, h) holds. But the derivative f' is not Lipschitz continuous on $\operatorname{int} \operatorname{dom} h$ since $f''(x) = -\frac{6}{x^4}$ is not bounded on $\operatorname{int} \operatorname{dom} h$.

3.2 Gradient Domination Inequalities through Bregman Lenses

We propose to extend the quadratic[2] case in Łojasiewicz inequality [27] given in (1.2) to the framework of Bregman distances.

Definition 3.4 (gradient dominated conditions/inequalities) The pair of functions (f, h) satisfies a gradient dominated condition if one of the two following conditions hold:

$$\text{(GD1)} \qquad \exists \tau > 0 : \; D_h(\nabla h^*(\nabla h(x) - \nabla f(x)), x) \geq \tau(f(x) - \nu(\mathcal{P})), \; \forall\, x \in \operatorname{int} \operatorname{dom} h. \tag{3.2}$$

$$\text{(GD2)} \qquad \exists \mu > 0 : \; D_h(x, \nabla h^*(\nabla h(x) - \nabla f(x))) \geq \mu(f(x) - \nu(\mathcal{P})), \; \forall\, x \in \operatorname{int} \operatorname{dom} h. \tag{3.3}$$

Inequality (3.2) or (3.3) implies that:

- when $\nabla f(x) = 0$ with $x \in \operatorname{int} \operatorname{dom} h$, then x is a *global* minimizer of f.
- when $h = \mathcal{Q}$, they reduce to a classical gradient dominated condition (cf. (1.2)) with $\tau \equiv \mu$:

$$(\forall x \in \mathbb{R}^n) \; \frac{1}{2}\|\nabla f(x)\|^2 \geq \mu \left(f(x) - \nu(\mathcal{P})\right). \tag{3.4}$$

Some remarks are in order with respect to the proposed gradient dominated conditions.

(i) Since Bregman distances are, in general, not symmetric, each of the above inequalities (GD1) and (GD2), whereby the arguments of D_h are reversed, simply marked the fact that one can consider both as a natural candidate for an extension of the classical gradient dominated just alluded above.

(ii) Both inequalities admit dual equivalent formulations. Indeed, recalling property P2 of a Bregman distance, they translate respectively into:

$$\text{(GD1)} \qquad \exists \tau > 0 : \; D_{h^*}(\nabla h(x), \nabla h(x) - \nabla f(x)) \geq \tau(f(x) - \nu(\mathcal{P})), \; \forall x \in \operatorname{int} \operatorname{dom} h.$$

$$\text{(GD2)} \qquad \exists \mu > 0 : \; D_{h^*}(\nabla h(x) - \nabla f(x), \nabla h(x)) \geq \mu(f(x) - \nu(\mathcal{P})), \; \forall x \in \operatorname{int} \operatorname{dom} h.$$

(iii) From the definition of $\alpha(h)$, assuming that $\alpha(h) > 0$, it follows from Fact 2.4(iv) that (GD1) and (GD2) are equivalent (up to the positive constant $\alpha(h)$).

However, note that it is also possible that (f, h) satisfies both gradient dominated conditions, but $\alpha(h) = 0$, this is illustrated in Example 3.8 below.

(iv) As mentioned, although both gradient dominated conditions can be considered, it turns out that when used in the context of the linear convergence analysis of the BG scheme, the condition (GD2) seems more restrictive, compare Theorem 4.3 and Remark 4.4(a).

[2] We pertain to this case for simplicity of exposition.

The following simple lemma (and its notation) is repeatedly used in the forthcoming analysis.

Lemma 3.5 (gradient envelope) *For every $x \in \operatorname{int}\operatorname{dom} h$, and any $t > 0$, define*

$$\begin{aligned} x_t^+ &:= \operatorname*{argmin}_u \{\langle \nabla f(x), u - x\rangle + t^{-1} D_h(u,x)\} \equiv \nabla h^*(\nabla h(x) - t\nabla f(x)), && (3.5)\\ \mathcal{G}_t(x) &:= \min_u \{\langle \nabla f(x), u - x\rangle + t^{-1} D_h(u,x)\}. && (3.6)\end{aligned}$$

Then for every $x \in \operatorname{dom} h$ one has $\mathcal{G}_t(x) = -t^{-1} D_h(x, x_t^+)$.

Proof. Writing the optimality condition for $x_t^+ \in \operatorname{int}\operatorname{dom} h$, uniquely defined by (3.5), we get

$$t\nabla f(x) + \nabla h(x_t^+) - \nabla h(x) = 0.$$

Therefore, using the definition of $\mathcal{G}_t(\cdot)$, combined with the above, we obtain for every $x \in \operatorname{int}\operatorname{dom} h$,

$$\begin{aligned} t\mathcal{G}_t(x) &= t\langle \nabla f(x), x_t^+ - x\rangle + D_h(x_t^+, x)\\ &= -\langle \nabla h(x_t^+) - \nabla h(x), x_t^+ - x\rangle + D_h(x_t^+, x),\\ &= -\{D_h(x_t^+, x) + D_h(x, x_t^+)\} + D_h(x_t^+, x) = -D_h(x, x_t^+),\end{aligned}$$

which proves the result. ■

3.3 Sufficient Conditions Implying gradient dominated Conditions

Below, we give two sufficient conditions for (f, h) to satisfy the gradient dominated condition (GD2). We recall that for $\alpha(h) > 0$, (GD1)-(GD2) are equivalent up to a positive constant, so these sufficient conditions also hold for the gradient dominated condition (GD1).

The first sufficient condition below naturally extends the classical case [33], namely for $h = \mathcal{Q}$, which states that if f is σ-strongly convex, then the gradient dominated (1.2) holds with $c = 2\sigma$. To establish the corresponding extension, we first need the relevant notion of strong convexity with respect to a Legendre function h, see e.g., [3].

Definition 3.6 (Legendre-strongly convex) *A function f is Legendre-strongly convex with respect to a Legendre function h if there exists $\sigma > 0$ such that*

$$(\forall\, x, y \in \operatorname{int}\operatorname{dom} h) \quad f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \sigma D_h(y, x), \tag{3.7}$$

i.e., $D_f(y, x) \geq \sigma D_h(y, x)$; consequently, $f - \sigma h$ is convex on $\operatorname{int}\operatorname{dom} h$*.*

Clearly, with $h = \mathcal{Q}$, this recovers the standard σ-strong convexity of f.

Lemma 3.7 (Legendre strongly convex implies gradient dominated (GD2)) *Assume that there exists $\sigma \in [1, +\infty[$ such that f is σ-strongly convex with respect to the Legendre function h. Then,*

$$(\forall x \in \operatorname{int}\operatorname{dom} h)\ D_{h^*}(\nabla h(x) - \nabla f(x), \nabla h(x)) \geq f(x) - \nu(\mathcal{P}),$$

i.e., (f, h) satisfies the gradient dominated condition (GD2) *with $\mu = 1$, as well as* (GD1) *when $\alpha(h) > 0$.*

Proof. Since $f - \sigma h$ and h are convex, and $\sigma - 1 \geq 0$, we have $f - h = f - \sigma h + (\sigma - 1)h$ convex on $\operatorname{int}\operatorname{dom} h$, and hence,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + D_h(y, x),\ \forall\, x, y \in \operatorname{int}\operatorname{dom} h.$$

Taking the infimum with respect to $y \in \operatorname{int}\operatorname{dom} h$ on both sides, we obtain

$$\inf_{y \in \operatorname{int}\operatorname{dom} h} f(y) \geq f(x) + \inf_{y \in \operatorname{dom} h} \{\langle \nabla f(x), y - x\rangle + D_h(y,x)\} =: f(x) + \inf_{y \in \operatorname{dom} h} R(y,x)$$

for all $x, y \in \operatorname{int}\operatorname{dom} h$.

Since here f and $R(\cdot, x)$ are lsc proper convex we have[3] $\inf_{y \in \operatorname{int}\operatorname{dom} h} f(y) = \inf_{y \in \overline{\operatorname{dom} h}} f(y) = \nu(\mathcal{P})$, and likewise, the infimum on the right-hand side can be taken over $\operatorname{dom} h$. Invoking Lemma 3.5, it follows that forall $x \in \operatorname{int}\operatorname{dom} h$,

$$\nu(\mathcal{P}) \geq f(x) + \mathcal{G}_1(x) = f(x) - D_h(x, x_1^+) = f(x) - D_h(x, \nabla h^*(\nabla h(x) - \nabla f(x))).$$

Rearranging the above is exactly (GD2) with $\mu = 1$. The claim for (GD1) with $\alpha(h) > 0$ follows from Fact 2.4(iv). ■

Recall that if (f, h) satisfies the gradient dominated condition (GD2) for some $\mu > 0$, this *does not* imply that $f - \mu h$ is convex on $\operatorname{int}\operatorname{dom} h$ for some $\mu > 0$. It was already one of the interest of the Łojasiewicz inequality. It is also possible that f satisfies both (GD1) and (GD2) but $\alpha(h) = 0$. This is illustrated in the example below.

Example 3.8 Let $f(x) := x$ and $h(x) := x \ln x - x$ for all $x \in \operatorname{dom} h = \mathbb{R}_+$. Then $h^*(y) = e^y, h'(x) = \ln x$, and $h'(x) - f'(x) = \ln x - 1$. Also, we have $\nu(\mathcal{P}) = 0$. Therefore, using the above, for all $x > 0$, we obtain

$$D_{h^*}(h'(x) - f'(x), h'(x)) - \mu(f(x) - \nu(\mathcal{P})) = e^{-1}x - \mu x = (e^{-1} - \mu)x,$$

and hence (f, h) satisfies the gradient dominated condition (3.3) for each $\mu \in \left]0, e^{-1}\right]$

Likewise, for all $x > 0$, we have $D_{h^*}(h'(x), h'(x) - f'(x)) = (1 - 2e^{-1})x \geq \tau x = \tau(f(x) - \nu(\mathcal{P}))$, where $\tau \in \left]0, \frac{e-2}{e}\right]$. So f also satisfies (GD1). However, note that $\alpha(h) = 0$ by Fact 2.4(iii); and for each $\mu > 0$, the function $f - \mu h$ is not convex on $\operatorname{int}\operatorname{dom} h$.

Our second sufficient condition says that when f is convex, and a *Bregman distance growth* condition of the objective function f holds, then a gradient dominated condition holds for the pair (f, h). The Bregman distance growth condition naturally extends the growth condition used in the quadratic case (see [17]). At the time this paper was under preparation, we discovered that this growth condition à la Bregman has also been proposed in the recent preprint [39] to analyze rate of convergence of proximal-like schemes for convex minimization problems.

Lemma 3.9 (Bregman growth condition implies gradient dominated (GD2)) *Let f be convex on* $\operatorname{dom} h$*, and let $x^* \in C$ with $f(x^*) = \nu(P)$. Suppose that f satisfies a Bregman Growth Condition:*

$$(GC) \quad \exists\, \gamma > 0: \ (\forall x \in \operatorname{int}\operatorname{dom} h) \ \ f(x) - f(x^*) \geq \gamma D_h(x^*, x).$$

Then, for every $\lambda > \gamma^{-1}$, λf satisfies the gradient dominated condition (GD2) *with* $\tau = (1 - (\gamma\lambda)^{-1})$*, i.e., one has*

$$(\forall x \in \operatorname{int}\operatorname{dom} h) \ D_h(x, \nabla h^*(\nabla h(x) - \lambda \nabla f(x))) \geq \Big(1 - (\lambda\gamma)^{-1}\Big)(\lambda f(x) - \lambda \nu(P)).$$

[3] Recall that for ψ proper lsc convex on $\mathbb{R}^n$ and P a convex subset of $\mathbb{R}^n$ with $\operatorname{int} P \neq \emptyset$ and $P \subset \operatorname{dom} \psi$, we have $\inf_{\operatorname{int} P} \psi = \inf_{\overline{P}} \psi$, see e.g., [8, Proposition 11.1].

Proof. Let $x \in \operatorname{int}\operatorname{dom} h$ and $\lambda > 0$. Invoking Lemma 3.5 (we use here the same notation as in that lemma), we have

$$\lambda^{-1} D_h(x, x_\lambda^+) = -\mathcal{G}_\lambda(x) = -\min_u \{\langle \nabla f(x), u - x\rangle + \lambda^{-1} D_h(u, x)\}.$$

Therefore, for every $u \in \operatorname{dom} h$, we obtain

$$\begin{aligned} D_h(x, x_\lambda^+) &\geq -\lambda\langle \nabla f(x), u - x\rangle - D_h(u, x) \\ &\geq \lambda(f(x) - f(u)) - D_h(u, x), \end{aligned}$$

where the second inequality uses the gradient inequality for the convex function f. Thus, with $u := x^*$ in the latter inequality, and using the condition (GC), it follows that

$$D_h(x, x_\lambda^+) \geq \left(\lambda - \gamma^{-1}\right)(f(x) - \nu(P)) = \left(1 - (\lambda\gamma)^{-1}\right)(\lambda(f(x) - \nu(P))).$$

Hence with $\lambda\gamma > 1$, (GD2) holds for λf with $\tau := \left(1 - (\lambda\gamma)^{-1}\right) > 0$. ■

When $\alpha(h) > 0$, a similar growth condition (just reverse the order in $D_h(x^*, x)$ in (GC) above) implies that (GD2) (3.3) holds for λf, the details are omitted.

The following result provides the link between the iteration of BG defined by x_λ^+ for a positive step size, and a *pure* gradient step x_1^+ (i.e., with $\lambda = 1$) in terms of Bregman distance. This connection plays a role in the convergence rate analysis developed in Section 4.

Lemma 3.10 *Let $\lambda > 0$ and $x \in \operatorname{int}\operatorname{dom} h$. Then for any $\lambda \geq 1$, one has:*

$$D_h(\nabla h^*(\nabla h(x) - \lambda \nabla f(x)), x) \quad \geq \quad D_h(\nabla h^*(\nabla h(x) - \nabla f(x)), x). \tag{3.8}$$

Proof. Let $x \in \operatorname{int}\operatorname{dom} h$ and $\lambda > 0$. For convenience, define $a := \nabla h(x), b := \nabla f(x)$. Then, using the definition of D_h, after rearranging, proving (3.8) is equivalent to prove

$$h(\nabla h^*(a - \lambda b)) - h(\nabla h^*(a - b)) \geq \langle \nabla h^*(a - \lambda b) - \nabla h^*(a - b), a\rangle =: R. \tag{3.9}$$

The gradient inequality for the convex function h implies

$$\begin{aligned} h(\nabla h^*(a - \lambda b)) - h(\nabla h^*(a - b)) &\geq \langle \nabla h^*(a - \lambda b) - \nabla h^*(a - b), a - b\rangle \\ &= \langle \nabla h^*(a - \lambda b) - \nabla h^*(a - b), a\rangle + \langle \nabla h^*(a - b) - \nabla h^*(a - \lambda b), b\rangle \\ &= R + \langle \nabla h^*(a - b) - \nabla h^*(a - \lambda b), b\rangle. \end{aligned} \tag{3.10}$$

Now, thanks to the monotonicity of ∇h^* we get,

$$(\lambda - 1)\langle \nabla h^*(a - b) - \nabla h^*(a - \lambda b), b\rangle \geq 0.$$

Therefore, with $\lambda \geq 1$, the latter inequality combined with (3.10) proves (3.9), i.e., the desired (3.8). ■

Note that when $0 < \lambda \leq 1$, we were not able to derive Lemma 3.10 (and it is not clear that such inequalities hold). Therefore, to overcome this difficulty we introduce a uniform control assumption on the behavior of x_λ^+ versus x_1^+ with respect to D_h. This plays an important role in our main results in Section 4.

3.4 Lower Control Function

We introduce a lower control function for the pair of functions (f,h) with respect to D_h which plays an important role in our main results in Section 4.

Assumption C $-$ θ uniform condition Let $\lambda > 0$ and $x \in \operatorname{int}\operatorname{dom} h$. For a pair of functions (f,h), there exists $\theta : \mathbb{R}_{++} \to \mathbb{R}_{++}$ such that

$$D_h(\nabla h^*(\nabla h(x) - \lambda \nabla f(x)), x) \geq \theta(\lambda) D_h(\nabla h^*(\nabla h(x) - \nabla f(x)), x),\ \forall x \in \operatorname{int}\operatorname{dom} h, \tag{3.11}$$

i.e.,

$$D_h(x_\lambda^+, x) \geq \theta(\lambda) D_h(x_1^+, x).$$

Remark 3.11 (a) Note that an alternative and sharper definition of θ would depend on x. Assumption C subsumes a *uniform* (i.e., depends only on λ) validity of the inequality (3.11). Moreover, verifying Assumption C seems to be difficult in general. However, some easy cases remain at hand, for instance by Lemma 3.10, it trivially holds with $\theta(\lambda) \equiv 1$ for any $\lambda \geq 1$.

(b) Assumption C also holds when f is σ-strongly convex with respect to h, and has a Lipschitz gradient with constant κ, with $(\sigma, \kappa > 0)$. Indeed, by [35, Proposition 12.60], the dual version of the above assumptions translates into

$$\exists \sigma, \kappa > 0 :\ \kappa^{-1}\|u - v\|^2 \leq 2D_{h^*}(u,v) \leq \sigma^{-1}\|u-v\|^2,\ \forall u \in \operatorname{dom} h, \forall v \in \operatorname{int}\operatorname{dom} h.$$

Now, recall that $D_h(x_\lambda^+, x) = D_{h^*}(\nabla h(x), \nabla h(x) - \lambda \nabla f(x))$. Using the right-hand side of the above inequality, we immediately get $\|\nabla f(x)\|^2 \geq 2\sigma D_h(x_1^+, x)$, and hence, combining the latter with the lefthand side part it follows that

$$D_h(x_\lambda^+, x) \geq \frac{\lambda^2}{2\kappa}\|\nabla f(x)\|^2 \geq \frac{\sigma\lambda^2}{\kappa} D_h(x_1^+, x)$$

i.e., Assumption C holds with $\theta(\lambda) := \frac{\sigma\lambda^2}{\kappa} > 0$ for any $\lambda > 0$.

4 Global Linear Convergence of the Bregman Gradient Method

Equipped with the LC-condition (see Definition 3.1), the gradient dominated condition, and the existence of a positive uniform control function θ (cf. Assumption C), we can establish the global linear convergence of the BG algorithm for the nonconvex problem $(\mathcal{P})$.

We start with the following simple but key result which is a slight extension of the descent inequality proven in [7, Lemma 4]. As shown in [18, Remark 4.1] and [38, Proposition 4.1], (for the more general Bregman proximal gradient), in that case we get:

Lemma 4.1 (fundamental inequality) *Assume that the pair of functions (f,h) satisfies the LC-condition with constant $L > 0$. For every x in $\operatorname{int}\operatorname{dom} h$, define x^+ in $\operatorname{int}\operatorname{dom} h$ by*

$$x^+ := \nabla h^*(\nabla h(x) - \lambda \nabla f(x)).$$

Then for every u in $\operatorname{dom} h$ one has

$$f(x^+) \leq f(x) + \langle \nabla f(x), u - x\rangle + \frac{1}{\lambda} D_h(u,x) - \frac{1}{\lambda} D_h(u, x^+) + \left(L - \frac{1}{\lambda}\right) D_h(x^+, x). \tag{4.1}$$

Proof. For completeness we include the simple proof as given in [38, Proposition 4.1] (specialized to the Bregman gradient iteration). Since

$$x^+ = \operatorname*{argmin}_u \left\{ \langle \nabla f(x), u \rangle + \frac{1}{\lambda} D_h(u,x) \right\},$$

invoking the well known three point Lemma [22, Lemma 3.2] gives for $u \in \operatorname{dom} h$:

$$\langle \nabla f(x), x^+ - u \rangle \le \frac{1}{\lambda} \left(D_h(u,x) - D_h(u,x^+) - D_h(x^+,x) \right).$$

By Lemma 3.2, the LC-condition gives

$$f(x^+) \le f(x) + \langle \nabla f(x), x^+ - x \rangle + L D_h(x^+,x).$$

Adding the above two inequalities, the desired result (4.1) follows. ■

Equipped with Lemma 4.1, we immediately deduce the sufficient descent property, see also [18, Remark 4.1].

Lemma 4.2 (sufficient descent property) *Assume that the LC-condition holds for the pair of functions (f,h) with constant $L > 0$, and let $\lambda > 0$. Then for every $x \in \operatorname{int} \operatorname{dom} h$ and $x^+ \in \operatorname{int} \operatorname{dom} h$ defined by*

$$x^+ := \nabla h^*(\nabla h(x) - \lambda \nabla f(x)),$$

we have

$$f(x^+) \le f(x) - \left(\frac{1 + \alpha(h)}{\lambda} - L \right) D_h(x^+,x). \tag{4.2}$$

Furthermore, the sufficient decrease property of the objective function value f holds when

$$0 < \lambda L < 1 + \alpha(h). \tag{4.3}$$

Proof. Recalling the definition of the symmetry coefficient $\alpha(h)$, we have $D_h(u,x^+) \ge \alpha(h) D_h(x^+,u)$, and hence the inequality (4.1) derived in Lemma 4.1 implies:

$$f(x^+) \le f(x) + \langle \nabla f(x), u - x \rangle + \frac{1}{\lambda} D_h(u,x) - \frac{\alpha(h)}{\lambda} D_h(x^+,u) + \left(L - \frac{1}{\lambda} \right) D_h(x^+,x).$$

Substituting $u = x$ in the above inequality proves (4.2), and with $\lambda L < (1 + \alpha(h))$, the claimed sufficient decrease. ■

We are now ready to establish the linear convergence of BG for nonconvex objective functions.

Theorem 4.3 *Assume that the pair of functions (f,h) satisfies the LC-condition with constant $L > 0$, the gradient dominated condition* (GD1) *with constant $\tau > 0$, and the θ-uniform condition (Assumption C). Define*

$$r := \left(\frac{1 + \alpha(h)}{\lambda} - L \right) \theta(\lambda) \tau. \tag{4.4}$$

Then for any λ such that $0 < \lambda L < 1 + \alpha(h)$, one has $0 < r \le 1$, and the sequence $(x_k)_{k \in \mathbb{N}}$ generated by BG has a global linear rate of convergence,

$$f(x_k) - \nu(\mathcal{P}) \le (1 - r)^k \left(f(x_0) - \nu(\mathcal{P}) \right), \ \forall k \ge 1.$$

Proof. First, note that since $0 < \lambda L < 1 + \alpha(h), \tau > 0$ and $\theta(\lambda) > 0$, we have $r > 0$. Fix an integer k. Thanks to the LC condition, we can apply Lemma 4.2 to the sequence generated by BG to obtain

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \left(\frac{1+\alpha(h)}{\lambda} - L\right) D_h(x_{k+1}, x_k) \\ &\leq f(x_k) - \theta(\lambda)\left(\frac{1+\alpha(h)}{\lambda} - L\right) D_h(\nabla h^*(\nabla h(x_k) - \nabla f(x_k)), x_k) \\ &\leq f(x_k) - \tau\theta\left(\frac{1+\alpha(h)}{\lambda} - L\right)(f(x_k) - \nu(\mathcal{P})) \\ &= f(x_k) - r(f(x_k) - \nu(\mathcal{P})), \end{aligned}$$

where the second inequality uses Assumption C, the third uses (GD1), and the last equality follows by definition of r. It ensues

$$f(x_{k+1}) - \nu(\mathcal{P}) \leq (1-r)\,(f(x_k) - \nu(\mathcal{P})),$$

and hence, since the sequence $(f(x_k) - \nu(\mathcal{P}))_{k\in\mathbb{N}}$ is nonincreasing, we must have $r \leq 1$. A simple recursion of the above inequality then proves the desired linear rate of convergence. ■

Remark 4.4 (a) Theorem 4.3 can also be established under the gradient dominated GD2 with parameter $\mu > 0$. However, in that case one must assume that $\alpha(h) > 0$, which is not needed under (GD1). A straightforward adaptation in the proof of Theorem 4.3 shows that in that case the linear rate of convergence holds with $0 < r \leq 1$ given by

$$r = \left(\frac{1+\alpha(h)}{\lambda} - L\right)\theta(\lambda)\alpha(h)\mu.$$

(b) Note that when the pair of functions (f, h) satisfies the LC-condition with constant $L > 0$, and in addition, f is assumed σ strongly convex with respect to a Legendre function h, we then obtain (as a particular case,of the result established for the Bregman proximal gradient in [38, Proposition 4.1]), that the BG scheme with step size $\lambda = 1/L$ linearly converges. More precisely we have for any $u \in \operatorname{dom} h$:

$$f(x_k) - f(u) \leq L\left(1 - \frac{\sigma}{L}\right)^k D_h(u, x_0).$$

(c) The algorithm under consideration could be also ran using varying step-sizes λ_k with the usual requirements

$$0 < \inf_{k\in\mathbb{N}} \lambda_k \text{ and } \sup_{k\in\mathbb{N}} \lambda_k < (1+\alpha(h))/L.$$

The convergence and complexity results are naturally carried out in this setting. This also remark applies to the forthcoming extensions Corollary 4.5 and Proposition 4.6

Theorem 4.3 extends the classical linear rate of convergence of BG [33, Theorem 4] (i.e., with $h \equiv \mathcal{Q}$ and stepsize $\lambda \in \left]0, \frac{2}{L}\right[$), whereby the usual Lipschitz continuity of the gradient of f is replaced by (LC), and the gradient dominated inequality (1.1) is replaced by GD1 (cf. (3.2)). This fact is recorded below:

Corollary 4.5 *Assume that the gradient ∇f is L-Lipschitz continuous with $L > 0$, that $(f, \mathcal{Q})$ satisfies the classical gradient dominated condition with constant $\mu > 0$, and that $\lambda \in \left]0, \frac{2}{L}\right[$. Consider the gradient descent method:*

$$x_{k+1} := x_k - \lambda\nabla f(x_k),\ \forall k \geq 0.$$

Then the following hold:

(i) $0 < r := \mu\lambda(2 - \lambda L) \leq 1$, *and* r *achieves maximum value* μ/L *when* $\lambda = 1/L$;

(ii) *The sequence of function values of iterates has a global linear convergence rate, i.e.,*

$$0 \leq f(x_k) - \nu(\mathcal{P}) \leq (1-r)^k(f(x_0) - \nu(\mathcal{P})),$$

and the optimal linear convergence rate is

$$f(x_k) - \nu(\mathcal{P}) \leq (1 - \mu/L)^k(f(x_0) - \nu(\mathcal{P})),$$

when $\lambda = 1/L$;

(iii) *The iterates* $(x_k)_{k\in\mathbb{N}}$ *converges linearly to an optimal solution of* $(\mathcal{P})$;

(iv) *When* $\lambda = 1/L$, *we then have* $r = \mu/L$ *and the classical linear rate of convergence (optimal according to (i)) of the gradient method is recovered.*

Proof. This follows from Theorem 4.3. Indeed, in this case, since $h := \mathcal{Q}$, we have $\alpha(h) = 1$, and hence $0 < \lambda L < 2$, while LC translates to the Lipschitz continuity of ∇f, and assumption C holds with $\theta(\lambda) = \lambda^2$. Thus, we get $r = \mu\lambda(2-\lambda L) \leq 1$, with maximum value μ/L attains when $\lambda = 1/L$. The linear convergence of the sequence $(x_k)_{k\in\mathbb{N}}$ to an optimal solution of $(\mathcal{P})$ follows as a special case of Proposition 4.6 given below. ■

As a byproduct of Theorem 4.3, other specific convergence results for the sequence $(x_k)_{k\in\mathbb{N}}$ generated by BG can be easily deduced. We illustrate this below with focus on some pointwise convergence results for the sequence $(x_k)_{k\in\mathbb{N}}$ generated by BG.

Proposition 4.6 *Under the assumption of Theorem 4.3, the following assertions hold for the sequence* $(x_k)_{k\in\mathbb{N}}$ *generated by BG:*

(i) (subsequential convergence) *If* f *is coercive, then the sequence* $(x_k)_{k\in\mathbb{N}}$ *is bounded, any cluster point of* $(x_k)_{k\in\mathbb{N}}$ *is an optimal solution of problem* $(\mathcal{P})$ *and* $\mathcal{W}$ *is nonempty and compact, where* $\mathcal{W}$ *is the set of cluster points of the sequence* $(x_k)_{k\in\mathbb{N}}$.

(ii) (summability) $\sum_{k=0}^{\infty} D_h(x_{k+1}, x_k) < \infty$.

(iii) (linear convergence of iterates) *If* h *is strongly convex, then* $(x_k)_{k\in\mathbb{N}}$ *converges linearly to an optimal solution of* $(\mathcal{P})$, *and* $(x_k)_{k\in\mathbb{N}}$ *has a finite length, i.e.,* $\sum_{k=0}^{\infty} \|x_{k+1} - x_k\| < +\infty$.

Proof. (i) Since $(f(x_k))_{k\in\mathbb{N}}$ is decreasing the coercivity of f implies that $(x_k)_{k\in\mathbb{N}}$ is bounded. Thus $\mathcal{W}$ is a nonempty and compact subset of C. Let $x^* \in \mathcal{W}$. Then there exists a subsequence $(x_{k_j})_{j\in\mathbb{N}}$ of $(x_k)_{k\in\mathbb{N}}$ such that $x_{k_j} \to x^* \in C$ as $j \to +\infty$. Since f is lsc, we have

$$f(x^*) \leq \lim_{j\to\infty} f(x_{k_j}) = \lim_{k\to\infty} f(x_k) = \nu(\mathcal{P}), \tag{4.5}$$

i.e., x^* is a solution of problem $(\mathcal{P})$.
(ii) Using Lemma 4.2, we obtain

$$\begin{aligned} D_h(x_{k+1}, x_k) &\leq \left(\frac{1+\alpha(h)}{\lambda} - L\right)^{-1} (f(x_k) - f(x_{k+1}) \\ &\leq \left(\frac{1+\alpha(h)}{\lambda} - L\right)^{-1} (f(x_k) - \nu(\mathcal{P})). \end{aligned} \tag{4.6}$$

Summming the first inequality over $k = 0, \dots, m-1$, and using a simple telescoping argument it follows that that $\sum_{k=0}^{\infty} D_h(x_{k+1}, x_k) < \infty$.
(iii) Using, (4.6) with the given σ strong convexity of h, and Theorem 4.3, we obtain

$$\frac{\sigma}{2}\|x_{k+1} - x_k\|^2 \leq \left(\frac{1+\alpha(h)}{\lambda} - L\right)^{-1} (1-r)^k (f(x_0) - \nu(\mathcal{P})).$$

Rearranging the above, with $M := \sqrt{\frac{2}{\sigma}\left(\frac{1+\alpha(h)}{\lambda} - L\right)^{-1} (f(x_0) - \nu(\mathcal{P}))}$, we thus have

$$\|x_{k+1} - x_k\| \leq M(\sqrt{1-r})^k. \tag{4.7}$$

The claim (iii) then follows by standard arguments, which we describe for completeness. The finite length of $(x_k)_{k\in\mathbb{N}}$ follows from (4.7):

$$\sum_{k=1}^{\infty} \|x_{k+1} - x_k\| \leq M \sum_{k=1}^{\infty} (\sqrt{(1-r)})^k \leq \frac{M}{1-\sqrt{1-r}}.$$

Let $l > k$. Using (4.7), (and recalling that $\sqrt{1-r} < 1$), we obtain

$$\|x_l - x_k\| \leq \sum_{n=k}^{l-1} \|x_{n+1} - x_n\| \leq M \sum_{n=k}^{l-1} (1-r)^{k/2} \leq \frac{M(\sqrt{1-r})^k}{1-\sqrt{1-r}}.$$

Therefore, $(x_k)_{k\in\mathbb{N}}$ is a Cauchy sequence and converges to some $x^* \in \mathbb{R}^n$. Taking the limit when $l \to \infty$ gives

$$\|x_k - x^*\| \leq \frac{M(\sqrt{1-r})^k}{1-\sqrt{1-r}},$$

which means that $x_k \to x^*$ linearly. Similar arguments as (4.5) show that x^* is an optimal solution to $(\mathcal{P})$. ■

Acknowledgments

Heinz Bauschke was partially supported by the Natural Sciences and Engineering Research Council of Canada. Jérôme Bolte was partially supported by Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant number FA9550-18-1-0226. Jiawei Chen was partially supported by the the Natural Science Foundation of China (No. 11401487), the Basic and Advanced Research Project of Chongqing (cstc2016jcyjA0239) and National Scholarship under the China Scholarship Council. Marc Teboulle was partially supported by the Israel Science Foundation under ISF Grant 1844-16. Xianfu Wang was partially supported by the Natural Sciences and Engineering Research Council of Canada.

References

[1] A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM J. Optim.*, 16 (2006), 697–725.

[2] A. Auslender and M. Teboulle. Projected subgradient methods with non-Euclidean distances for nondifferentiable convex minimization and variational inequalities. *Math. Program.*, Ser. B, 120 (2009), 27–48.

[3] P.L. Bartlett, E. Hazan, and A. Rakhlin, Adaptive online gradient descent, *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis (editors), MIT Press, Cambridge, MA, (2007), pp. 65–72.

[4] H.H. Bauschke and J.M. Borwein, On projection algorithms for solving convex feasibility problems, *SIAM Rev.,* 38 (1996), pp. 367–426.

[5] H.H. Bauschke and J.M. Borwein, Legendre functions and the method of random Bregman projections, *J. Convex Anal.,* 4 (1997), pp. 27–67.

[6] H.H. Bauschke and J.M. Borwein, Joint and separate convexity of the Bregman distance, *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications (Haifa 2000)*, D. Butnariu, Y. Censor, S. Reich (editors), Elsevier, (2001), pp. 23–36.

[7] 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 (2017), pp. 330–348.

[8] H.H. Bauschke and P.L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Second edition, Springer, New York, 2017.

[9] A. Beck, *First-Order Methods in Optimization*, MOS-SIAM Series on Optimization, MO25, SIAM, Philadelphia, PA, 2017.

[10] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM J. Imag. Sci.,* 2 (2009), pp.183–202.

[11] A. Beck and M. Teboulle, Mirror descent and nonlinear projected subgradient methods for convex optimization, *Oper. Res. Lett.* 31 (2003), pp. 167–175.

[12] M. Bertero, P. Boccacci, G. Desider, and G. Vicidomini, Image deblurring with Poisson data: From cells to galaxies, *Inverse Problems,* 25 (2009), 123006.

[13] D.P. Bertsekas, *Nonlinear Programming*, 2nd ed. Athena Scientific, Belmont, MA, 1999.

[14] D.P. Bertsekas, *Convex Optimization Algorithms*, Athena Scientific, Belmont Massachussets, 2015.

[15] J. Bolte, A. Daniilidis, and A.S. Lewis, The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems, *SIAM J. Optim.*, 17 (2007), pp. 1205–1223.

[16] J. Bolte and M. Teboulle, Barrier operators and associated gradient like dynamical systems for constrained minimization problems, *SIAM J. Control Optim.*, 42 (2003), pp. 1266–1292.

[17] J. Bolte, T.P. Nguyen, J. Peypouquet, and B. W. Suter. From error bounds to the complexity of first-order descent methods for convex functions, *Math. Program.*, Ser. A, 165 (2017), 471–507.

[18] 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), 2131-–2151.

[19] L.M. Bregman, The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming, *U.S.S.R. Comput. Math. Math. Phys.,* 7 (1967), pp. 200–217.

[20] R. Burachik, and A. Iusem, A generalized proximal point algorithm for the variational inequality problem in a Hilbert space. SIAM Journal on Optimization, 8(1) (1998), 197-216.

[21] Y. Censor and S. A. Zenios, Proximal minimization algorithm with D-functions, *J. Optim. Theory Appl.*, 73 (1992), pp. 451–464.

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

[23] F.H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983; Republished as Classics in Applied Mathematics, vol. 5, SIAM, Philadelphia, PA, 1990.

[24] J. Eckstein, Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming, *Math. Oper. Res.*, 18 (1993), pp. 202–226.

[25] A.A. Goldstein, On steepest descent, *J. SIAM Control,* 3 (1965), pp. 147–151.

[26] A. Iusem, On dual convergence and the rate of primal convergence of Bregman?s convex programming method, SIAM Journal on Optimization, 1 (1991), pp. 401–423.

[27] S. Łojasiewicz, Une propriété topologique des sous-ensembles analytiques réels, In Les Équations aux Derivées Partielles, Éditions du Centre National de la Recherche Scientifique, Paris, (1963), pp. 87–89.

[28] S. Łojasiewicz, Ensembles semi-analytiques, Cours miméographié de la Faculté des Sciences d'Orsay, I.H.E.S., Bures-sur-Yvette, July 1965. `http://perso.univ-rennes1.fr/michel.coste/Lojasiewicz.pdf`

[29] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.

[30] A. S. Nemirovsky and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, A Wiley-Interscience Publication, John Wiley & Sons Inc., New York, 1983.

[31] Q.V. Nguyen, Forward-backward splitting with Bregman distances, *Vietnam J. Math.* 45 (2017), pp. 519–539.

[32] D.P. Palomar and Y.C. Eldar, *Convex Optimization in Signal Processing and Communications*, Cambridge University Press, New York, 2010.

[33] B.T. Polyak, Gradient methods for minimizing functionals (in Russian), *Zh. Vychisl. Mat. Mat. Fiz.,* 3 (1963), pp. 643–653.

[34] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[35] R.T. Rockafellar and R. J-B Wets, *Variational Analysis*, Springer, corrected 3rd printing, 2009.

[36] S. Sra, S. Nowozin, and S.J. Wright, *Optimization for Machine Learning*, MIT Press, Cambridge, MA, 2011.

[37] M. Teboulle. Entropic proximal mappings with application to nonlinear programming. *Math. Oper. Res.*, 17 (1992), pp. 670–690.

[38] M. Teboulle, A simplified view of first order methods for optimization, *Math. Program.*, Ser. B, 170 (2018), pp. 67–96.

[39] H. Zhang, Y.-H. Dai, L. Guo, Proximal-like incremental aggregated gradient method with linear convergence under Bregman distance growth conditions, arXiv:1711.01136.