Math. Program., Ser. B (2018) 170:67–96
https://doi.org/10.1007/s10107-018-1284-2

FULL LENGTH PAPER

# A simplified view of first order methods for optimization

**Marc Teboulle**[1]

Received: 30 January 2018 / Accepted: 26 April 2018 / Published online: 12 May 2018


**Abstract** We discuss the foundational role of the proximal framework in the development and analysis of some iconic first order optimization algorithms, with a focus on non-Euclidean proximal distances of Bregman type, which are central to the analysis of many other fundamental first order minimization relatives. We stress simplification and unification by highlighting self-contained elementary proof-patterns to obtain convergence rate and global convergence both in the convex and the nonconvex settings, which in turn also allows to present some novel results.

**Keywords** Proximal framework · Non-Euclidean Bregman distance · First order algorithms · Convex and nonconvex minimization · Descent Lemma · Kurdyka–Łosiajewicz property

**Mathematics Subject Classification** 90C25 · 65K05

## 1 Introduction

The notion of proximal map of a convex function, invented about half a century ago by Moreau in his seminal work [48], is kicking and alive! This fundamental regularization process gave birth 5 years later to the so-called *proximal minimization* algorithm by Martinet [47], followed by its extension in Rockafellar [59] for solv-

This research was partially supported by the Israel Science Foundation, under ISF Grant 1844-16, and the German Israel Foundation, under GIF Grant 1253304.6-2014.

✉ Marc Teboulle
teboulle@post.tau.ac.il

[1] School of Mathematical Sciences, Tel-Aviv University, 69978 Ramat-Aviv, Israel

ing monotone inclusions. Merely abandoned in favor of the sophisticated interior point methods, and unpopular among many optimization researchers for over three decades, proximal based methods and their relatives are nowadays "starring" in modern optimization algorithms based on first order information, e.g., function values and gradient/subgradients. This resurrection and renewed interest in first order methods (FOM) is motivated by the current high demand in solving large and huge scale problems arising in a wide spectrum of disparate fundamental applications (e.g., signal processing, image sciences, machine learning, communication systems, computerized tomography, astronomy), which admit structured convex and nonconvex optimization formulations. The computational simplicity of FOM makes them appealing and ideal candidates for solving such problems to medium accuracy. Current intensive research activities can be seen from the very large volume of literature which is constantly growing at a rapid pace. Clearly, space limitation prevents us to discuss and review all the past and recent achievements. Instead, to get an initial sense of the intensity of this body of research which started about a decade ago, we refer the reader to, e.g., [54,62] and references therein. For a more recent comprehensive treatment of convex optimization algorithms, we refer the reader to the book of Bertsekas [23], and to the just released book of Beck [15], which focuses on first order methods, and provides a unique self-contained and rigorous study underlying the theoretical foundations of FOM. Both monographs include many relevant up-to-date and annotated sources, but note that new works in this area are basically appearing on a daily basis!

In this work we discuss in a concise way the foundational role of the proximal approach in the development and analysis of first order optimization algorithms, with a focus on non-Euclidean proximal schemes based on Bregman distances. The aim of this paper is to stress simplification and unification. Starting with the proximal framework (Sect. 2), and building on two fundamental pillars, one which is quite old and one very recent (Sect. 3), elementary proof-patterns emerge. This allows to simplify the derivation of the rate of convergence and global convergence of some of the most fundamental first order algorithmic icons: gradient/subgradient, proximal gradient, and mirror descent for composite convex optimization models. Note that these schemes are also central to the analysis of many other first order minimization relatives that will not be reviewed here, but see Sect. 2 for further discussion and references. In addition to some of these classical results, we highlight the benefits of the simplicity of the proof techniques (Sect. 4), which also allow to derive these results under weaker assumptions than the usual ones, as well as novel results, see respectively Sect. 4.2, where we introduce a single condition in terms of the problem's data allowing to refine the classical analysis of the Mirror Descent (Theorem 4.3), and Sect. 4.3 where we derive a descent like property for the more general convexly composite model (Lemma 4.2), and the corresponding complexity result in Theorem 4.4. The nonconvex setting is far more difficult than its convex counterpart, and remains highly challenging. In Sect. 5, we give a brief tour on some very recent advances in this area, describing a flexible framework to derive global convergence of the Bregman based proximal gradient scheme to a critical point of a nonconvex composite minimization model. We end the paper in Sect. 6, with some discussion on the benefits and limitations of the Bregman proximal framework, and concluding remarks leading to an open challenging question.

**Notation** We use standard notation and concepts from convex and variational analysis, which unless otherwise specified, can all be found in the classical monographs [57,58]. Throughout, $\mathbb{E}$ stands for a finite dimensional vector space with inner product $\langle \cdot, \cdot \rangle$, norm $\| \cdot \|$ and its usual dual norm, denoted by $\| \cdot \|_*$, residing in the dual space $\mathbb{E}^*$.

## 2 The proximal framework

We start with the seminal work of Moreau [48], which provides the foundation and the central ideas underlying the proximal methodology.

### 2.1 The master chef: Moreau's proximal map

Given a proper, lower semicontinuous (lsc) and convex function $\varphi : \mathbb{E} \to (-\infty, \infty]$ and $\lambda > 0$, the proximal map of Moreau [48], is defined as the unique minimizer:

$$\text{prox}_{\lambda\varphi}(x) := \text{argmin} \left\{ \varphi(u) + \frac{1}{2\lambda}\|u - x\|^2 : \ u \in \mathbb{E} \right\},$$

and the resulting optimal value, that is, the Moreau envelope of $\varphi$ is the function

$$x \to \varphi_\lambda(x) := \min \left\{ \varphi(u) + \frac{1}{2\lambda}\|u - x\|^2 : \ u \in \mathbb{E} \right\}.$$

This *regularization* process for $\varphi$ produces a very neat function $\varphi_\lambda$ which is finite everywhere, convex and with $\lambda^{-1}$-Lipschitz continuous gradient on $\mathbb{E}$ given by:

$$\nabla\varphi_\lambda(x) = \lambda^{-1}(x - \text{prox}_{\lambda\varphi}(x)).$$

Moreover, minimization of $\varphi$ and $\varphi_\lambda$ are equivalent in the sense that:

$$\inf_x \varphi_\lambda(x) = \inf_x \varphi(x) \quad \text{and} \quad \text{argmin}_x \ \varphi(x) = \text{argmin}_x \ \varphi_\lambda(x),$$

with minimizers satisfying the equation $x = \text{prox}_{\lambda\varphi}(x)$.

This naturally provides the rationale for the so-called *proximal minimization* algorithm of Martinet [47], which consists of iterating the above fixed point equation:

$$x^0 \in \mathbb{E}, \quad x^{k+1} = \text{prox}_{\lambda\varphi}(x^k), \quad k = 0, 1, \ldots.$$

Alternatively, this is equivalent to the basic gradient iteration with step size $\lambda$ applied on the smooth function $\varphi_\lambda$.

Despite the obvious difficulties that can often emerge from the practical computational side of a proximal map (which by itself requires to solve a nonsmooth optimization problem), the proximal framework is fundamental, and nowadays has regained a *starring* position in the development and analysis of modern optimization algorithms based on first order information, both in their primal and dual forms.

In fact, the proximal methodology is not only underlying the essence of the central first order schemes, e.g., gradient, proximal gradient, and mirror descent which can be recovered through a specific choice of $\varphi$, but also at the root of many other fundamental first order minimization methods. For instance, these include incremental and stochastic versions, coordinate descent and smoothing methods, as well as fundamental primal-dual methods, within augmented Lagrangian and their many relatives, e.g., alternating direction method of multipliers, and related decomposition schemes. Because of space limitations, we obviously barely scratch the surface, and will not discuss the many theoretical, computational and applications aspects of past and recent achieved results in this area. For a taste of these and additional modern coverage on these topics, we refer the reader to some recent works, e.g., [13,15,18,19,23,30,61] which also provide ample relevant references.

### 2.2 Extending the scope of Moreau's proximal map: motivation

Mimicking Moreau, one can consider replacing the squared Euclidean distance with a non-Euclidean proximity measure, also called *distance-like* function. Formally, a general form of the algorithm for minimizing a function $\psi : \mathbb{E} \to (-\infty, \infty]$ then reads as

$$x^{k+1} \in \operatorname{argmin} \left\{ \psi(u) + \frac{1}{\lambda} D(u, x^k) : \ u \in \mathbb{E} \right\},$$

where $D(\cdot, \cdot)$ stands for a distance-like function which replaces the squared Euclidean distance. There exist various ways to define adequate proximity measures, see for instance the general framework of Auslender and Teboulle [8] and references therein, which analyzes first order methods based on a variety of distance-like functions, including the so-called Csiszar $\phi$-divergence [37] and the Bregman distance [31].

The rationale which underlines the relevance and usefulness of considering more general proximal measures stems from algorithmic contexts, i.e., toward improving and extending convergence properties of classical algorithms, as well as from specific applications areas. Let us briefly describe some contexts where the use of non-Euclidean proximal schemes have been particularly useful, and which include for example:

(a) *The possibility to better adapt to the geometry of a given problem*, which allows to derive simpler proximal computational steps, i.e., by naturally eliminating the constraint through a suitable choice of $D$ which best match the problem's data information and structures, see e.g. [7,9,14]. Moreover, this allows for deriving a complexity constant in terms of $D(x^*, x^0)$ between the starting point $x^0$, and an optimal solution $x^*$, which can be significantly smaller than the usual term $\|x^* - x^0\|^2$, and is often nearly optimal for very large scale problems, see e.g., [16,20,50].

(b) *Augmented Lagrangian Methods*. It is well known that the classical augmented Lagrangian method for the standard inequality constrained convex programming problem is nothing else but the proximal minimization algorithm applied on its

dual [21]. Likewise, applying a non-Euclidean proximal scheme produces an augmented Lagrangian function which will involve the conjugate of $D$. However, in contrast to its classical quadratic counterpart for inequality constraints, (which lacks twice continuous differentiability, even if the problem's data is $C^2$), here one can obtain *smooth* augmented Lagrangians, and this provides computational advantages for their minimization, see Bertsekas [21] for early work in this direction, and [10,40,56,63] for more examples and results within this approach.

(c) *Links with Dynamical Systems* In that context, non-Euclidean proximal based methods are particularly useful to study some continuous dynamical systems associated with constrained optimization problems, and have resulted in producing new *interior* descent methods and elliptic barrier operators [26], and also allows to make connections with physical and other applied science phenomena, such as, the heavy ball with friction in mechanics [3], and the Lotka–Volterra systems in biology [2].

The above description is far from being exhaustive, but rather just an appetizer on the prospect of studying non-Euclidean proximal methods. For further details, elaborations and enhancements with general non-Euclidean proximal schemes, including extensions to variational inequalities, see, for instance, [6,7,9,49] and references therein, as well as the concluding Sect. 6 for further discussion.

This paper will focus only on schemes based on the Bregman distance, stressing its natural appearance and usefulness in most of the fundamental basic FOM. We refer the reader to [12,26,33–35,40,63] and references therein for early foundational papers, more motivation, and key results on proximal-based methods associated to Bregman distances. Some very recent works on Bregman based FOM in various contexts can be found in [14,24,41,52].

The rest of this section provides the basic definitions and preliminaries for working in the proximal Bregman setting.

### 2.3 The Bregman distance

To define non-Euclidean proximity measures in general, and here specifically the Bregman distance, it is convenient to work with *Legendre* functions. Basically, as we shall see below, this allows to handle the nonlinear geometry of a problem, e.g., acting as a natural barrier for a constraint set. This is very flexible and more general than the classical squared Euclidean distance, allowing adaptation to the structure and data information of the problem at hand.

To introduce Legendre functions, we first need to recall some basic facts from convex analysis which can be found in [58, Section 26].

A proper and convex function $h : \mathbb{E} \to (-\infty, \infty]$ is called essentially smooth if $\operatorname{int}\operatorname{dom} h \neq \emptyset$, $h$ is differentiable on $\operatorname{int}\operatorname{dom} h$, and $\|\nabla h(x^k)\| \to \infty$ whenever $\{x^k\}_{k\in\mathbb{N}} \subset \operatorname{int}\operatorname{dom} h$ converges to a boundary point $x \in \operatorname{dom} h$ as $k \to +\infty$.

Recall that an extended valued function on $\mathbb{E}$ is *smooth*, only if it is actually finite and differentiable throughout $\mathbb{E}$. Clearly, a smooth convex function on $\mathbb{E}$ is in particular essentially smooth. The next result records important facts characterizing essential smoothness.

**Lemma 2.1** (Essential smoothness) [58, Theorem 26.1] *Let $h : \mathbb{E} \to (-\infty, \infty]$ be a proper, lsc and convex function. The following statements are equivalent:*

(a) *$h$ is essentially smooth.*
(b) *$\partial h(x) = \{\nabla h(x)\}$ when $x \in \operatorname{int} \operatorname{dom} h$, while $\partial h(x) = \emptyset$ for $x \notin \operatorname{int} \operatorname{dom} h$.*
(c) *$\operatorname{dom} \partial h = \operatorname{int} \operatorname{dom} h \neq \emptyset$.*

**Definition 2.1** (*Legendre function*) [58]. A function $h : \mathbb{E} \to (-\infty, \infty]$ which is proper, lsc, strictly convex and essentially smooth will be called a Legendre function.

In addition to the properties (b) and (c) given in Lemma 2.1, a Legendre function also enjoys the following useful properties [58, Thm. 26.5]:

- $h$ is Legendre if and only if its conjugate $h^*$ is Legendre.
- The gradient of a Legendre function $h$ is a bijection from $\operatorname{int} \operatorname{dom} h$ to $\operatorname{int} \operatorname{dom} h^*$, and its inverse is the gradient of the conjugate, that is, we have $(\nabla h)^{-1} = \nabla h^*$.

In this work we adopt the following definition of a Bregman distance, which origin goes back to the work of Bregman [31].

**Definition 2.2** (*Bregman distance*) Let $h : \mathbb{E} \to (-\infty, \infty]$ be a Legendre function. The Bregman distance associated to $h$, denoted by $D_h : \operatorname{dom} h \times \operatorname{int} \operatorname{dom} h \to \mathbb{R}_+$ is defined by:

$$D_h(x, y) := h(x) - [h(y) + \langle \nabla h(y), x - y \rangle].$$

It naturally measures the proximity between the two points $(x, y)$. Indeed, thanks to the gradient inequality we have,

$$h \text{ is convex if and only if } D_h(x, y) \geq 0, \quad \forall\, x \in \operatorname{dom} h, \quad y \in \operatorname{int} \operatorname{dom} h.$$

In addition, thanks to the strict convexity of $h$ we have that $D_h(x, y) = 0$ if and only if $x = y$, which implies a basic distance-like property. However, note that $D_h$ is not symmetric in general, unless $h$ is the energy function, that is, $h(\cdot) = (1/2)\|\cdot\|^2$, and in this case $D_h$ recovers the classical squared Euclidean distance.

We list below some of the most popular and useful choices for $h$ to generate relevant associated Bregman distances $D_h$, which are well documented in the literature, see e.g., [8,12,40,63] where more examples are given, including *nonseparable* Bregman distances, as well as, Bregman distances on the space of symmetric matrices.

*Example 2.1* To generate a Bregman distance in the $n$ dimensional vector space $\mathbb{E} = \mathbb{R}^n$, it is often enough to consider a one dimensional function $h_j : \mathbb{R} \to (-\infty, \infty]$ which is convex on $\mathbb{R}$. Then, with $h(x) = \sum_{j=1}^n h_j(x_j)$, we simply get that $D_h(x, y) = \sum_{j=1}^n D_{h_j}(x_j, y_j)$. Typical and useful examples include the following $h$ (with $n = 1$):

- **(Energy)** $h(x) = \frac{1}{2}x^2$ with $\operatorname{dom} h = \operatorname{dom} h^* = \mathbb{R}$, and likewise, $h(x) = \frac{1}{p}|x|^p$, $p \geq 2$, with $h^*(y) = \frac{1}{q}|y|^q$, $p + q = pq$.
- **(Shannon Entropy)** $h(x) = x \log x$ with $\operatorname{dom} h = [0, \infty]$, $(0 \log 0 = 0)$, and $h^*(y) = e^{y-1}$.

- **(Fermi–Dirac)** $h(x) = x\log x + (1-x)\log(1-x)$ with $\operatorname{dom} h = [0,1]$, and $h^*(y) = \log(1+e^y)$.
- **(Hellinger)** $h(x) = -\sqrt{1-x^2}$ with $\operatorname{dom} h = [-1,1]$ and $h^*(y) = \sqrt{1+y^2}$.
- **(Burg)** $h(x) = -\log x$ with $\operatorname{dom} h = (0,\infty)$, and $h^*(y) = -\log(-y) - 1$.

Note that, in all the above examples, $h$ is Legendre. Moreover, except for the last example, all their conjugates function $h^*$ share the nice and useful property of $\operatorname{dom} h^* = \mathbb{R}$; see Sect. 6 for further discussion on the relevance of these examples in applications.

The definition of a Bregman distance varies in the literature. In particular, most often the function $h$ generating $D_h$ is assumed to be $\sigma$-*strongly convex*, a stringent assumption that might also prevent the use of classical $D_h$, e.g., all the last four functions $h$ given in Example 2.1 *are not* strongly convex on $\operatorname{dom} h$. In fact, as we shall see later on, strong convexity of $h$ is often not needed.

The *structural form* of $D_h$ is also useful when $h$ *is not* convex. The function $D_h$ measures the *gap* between the value of $h$ at a given point $x \in \operatorname{dom} h$ from its first order Taylor linear approximation around $y \in \operatorname{int}\operatorname{dom} h$, and could be naturally dubbed as the *First Order Taylor Gap* associated to $h$. In that case, obviously, the distance-like property of $D_h$ is lost, yet this interpretation can be beneficially used, see Sect. 4.

We end this section with the *three points identity*, whose naturally extends the Euclidean Pythagoras type identity. This identity, first observed in Chen and Teboulle [35] is very simple, yet crucial in the analysis of any optimization method based on Bregman distances.

**Lemma 2.2** (The Three Points Identity) [35, Lemma 3.1] *Suppose* $h : \mathbb{E} \to (-\infty,\infty]$ *is Legendre. Then, for any three points* $x, y \in \operatorname{int}\operatorname{dom} h$ *and* $z \in \operatorname{dom} h$*, the following three points identity holds true:*

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

Note that the Legendre property of $h$ *plays no role* in the derivation of this identity. Indeed, for any differentiable function $h$, the identity follows by simple algebra relying on the structural definition of $D_h$. With $h(u) := \|u\|^2/2, \mathbb{E} = \mathbb{R}^n$, we recover as expected, the fundamental Pythagoras identity

$$\|z-x\|^2 - \|z-y\|^2 - \|x-y\|^2 = 2\langle x-y, y-z\rangle,$$

where here $\|\cdot\|$ stands for the Euclidean norm in $\mathbb{R}^n$.

### 2.4 The Bregman proximal map

Consider the following convex optimization problem

$$\varphi_* = \inf\{\varphi(u) : u \in C\}.$$

We make the following standing assumptions.

**Assumption A** (i) $\varphi : \mathbb{E} \to (-\infty, \infty]$ is proper, lsc and convex,
(ii) $C \subseteq \mathbb{E}$ is nonempty, closed and convex,
(iii) $h : \mathbb{E} \to (-\infty, \infty]$ is Legendre,
(iv) $(\operatorname{dom}\varphi \cap C) \cap \operatorname{dom} h \neq \emptyset$.

Mimicking Moreau one more time, we replace the squared Euclidean distance with $D_h(\cdot,\cdot)$ to define the Bregman proximal map (with the same notation as the one of Moreau, except for the presence of $h$).

**Definition 2.3** (*Bregman proximal map*) Suppose that Assumption **A** holds. For any $x \in \operatorname{int}\operatorname{dom} h$ and $\lambda > 0$, the Bregman proximal map is defined by

$$\operatorname{prox}^h_{\lambda\varphi}(x) := \operatorname{argmin}\left\{\varphi(u) + \frac{1}{\lambda}D_h(u,x) : \ u \in C\right\}.$$

Various types of assumptions can be made to warrant well-posedness of the Bregman proximal map, and which has been developed in the cited literature just alluded above. We state here one result which warrants that the Bregman proximal map associated with a convex function is well-defined. To make this part self-contained, we first recall some elementary results which can be found in [58].

**Definition 2.4** (*Level boundedness*) A function $\psi : \mathbb{E} \to (-\infty, \infty]$ is called (lower) level bounded if for every $\alpha \in \mathbb{R}$, the level set $\operatorname{Lev}(\psi, \alpha) := \{x \in \mathbb{E} : \ \psi(x) \leq \alpha\}$ is bounded (possibly empty).

We then have the following properties, (cf. [58, Corollary 8.7.1]). Let $\psi : \mathbb{E} \to (-\infty, \infty]$ be a proper, lsc and convex function.

- If the level set $\operatorname{Lev}(\psi, \alpha)$ is nonempty and bounded for some $\alpha \in \mathbb{R}$, it is bounded for every $\alpha \in \mathbb{R}$.
- If $\psi$ is level bounded, then argmin $\psi$ is nonempty and compact, and inf $\psi$ is finite.

The proof of Lemma 2.3 is patterned after similar approaches used, e.g., in [8,40, 64].

**Lemma 2.3** *Suppose that Assumption* **A** *holds. For any* $x \in \operatorname{int}\operatorname{dom} h$ *and* $\lambda > 0$ *let*

$$x^+ := \operatorname{prox}^h_{\lambda\varphi}(x) = \operatorname{argmin}\left\{\varphi(u) + \frac{1}{\lambda}D_h(u,x) : \ u \in C\right\}.$$

*The following statements hold.*

(a) *If* $\inf\{\varphi(u) : u \in C\} = \varphi_* > -\infty$, *then the minimizer* $x^+$ *exists and is unique.*
(b) *If in addition,* $\operatorname{ri}(\operatorname{dom}\varphi \cap C) \subset \operatorname{int}\operatorname{dom} h$, *then* $x^+ \in \operatorname{dom}\varphi \cap \operatorname{int}\operatorname{dom} h$, *which is characterized via its first order optimality condition:*

$$0 \in \lambda\partial\varphi(x^+) + \nabla h(x^+) - \nabla h(x). \tag{2.1}$$

*Proof* Fix $x \in \operatorname{int}\operatorname{dom} h$ and $\lambda > 0$.

(a) Under the premises of the lemma, a standard argument shows that

$$\Phi_\lambda(u) := \varphi(u) + \lambda^{-1} D_h(u, x) + \delta_C(u),$$

is proper, lsc and strictly convex (since $h$ is strictly convex). Thus, if a minimizer of $\inf_u \Phi_\lambda(u)$ exists, it must be unique. To show existence of a minimizer, we will prove that $\Phi_\lambda$ is level bounded. For convenience, let $u \to d(u) := D_h(u, x)$. Then $d$ is proper, lsc and strictly convex, with $d(u) \geq 0$, and equal to zero if and only if $x = u$. Therefore, $\mathrm{Lev}(d, 0) = \{x\}$, and hence (cf. above recalled properties) $\mathrm{Lev}(d, \alpha)$ is bounded for any $\alpha \geq 0$. Now, let $m := \inf_u \Phi_\lambda(u)$. Since $d(u) \geq 0$, then $m \geq \varphi_* > -\infty$, and the following holds:

$$L(m) := \mathrm{Lev}(\Phi_\lambda, m) \subset \mathrm{Lev}(d, \lambda(m - \varphi_*)).$$

Since we have just seen that $\mathrm{Lev}(d, \alpha)$ is bounded for any $\alpha \geq 0$, then $\mathrm{Lev}(d, \lambda(m - \varphi_*))$ is bounded, and hence by the above inclusion, so is $L(m)$, and the claimed existence result is proved.

For (b), under the premises of the lemma, we can apply standard subdifferential calculus rules [58], which thanks to the additional fact that $h$ is given Legendre, characterizes a well-defined unique optimal $x^+ \in \operatorname{int} \operatorname{dom} h$ via the inclusion $0 \in \partial\varphi(x^+) + \lambda(\nabla h(x^+) - \nabla h(x))$, as desired. □

Throughout the rest of this paper, we take as our blanket assumption the validity of Lemma 2.3, that is, the proximal map $x^+ = \mathrm{prox}^h_{\lambda\varphi}(x)$ is uniquely defined and satisfies $x^+ \in \operatorname{int} \operatorname{dom} h$.

## 3 Two fundamental pillars

The analysis of basic proximal schemes and their variants, essentially relies on the following two main pillars:

- One well-known and quite old: the Bregman proximal inequality derived by Chen and Teboulle [35], which provides a fundamental estimate in the objective function value gap for Bregman based proximal schemes.
- One very recent: A Lipschitz-Like Convexity Condition, recently introduced by Bauschke et al. [14], which naturally captures through one convexity condition, the nonlinear geometry information of the problem's at hand.

### 3.1 The Bregman proximal inequality

The first pillar of the analysis is the following result which naturally extends a similar fundamental estimate for the Euclidean squared proximal map as first derived by Güler [43].

**Lemma 3.1** (Bregman proximal inequality) [35, Lemma 3.2] *Let $\varphi : \mathbb{E} \to (-\infty, \infty]$ be a proper, lsc and convex function. Given $\lambda > 0$ and $x \in \operatorname{int}\operatorname{dom} h$, define:*

$$x^+ := \operatorname{prox}_{\lambda\varphi}(x) = \operatorname{argmin}_u \left\{ \varphi(u) + \frac{1}{\lambda} D_h(u, x) \right\}.$$

*Then, $x^+ \in \operatorname{dom}\varphi \cap \operatorname{int}\operatorname{dom} h$, and*

$$\lambda(\varphi(x^+) - \varphi(u)) \leq D_h(u, x) - D_h(u, x^+) - D_h(x^+, x), \quad \forall u \in \operatorname{dom} h.$$

*Proof* Let $x \in \operatorname{int}\operatorname{dom} h$ and $\lambda > 0$. Lemma 2.3 warrants the existence of a unique $x^+ \in \operatorname{int}\operatorname{dom} h$, such that

$$\lambda\,\xi + \nabla h(x^+) - \nabla h(x) = 0, \quad \text{with} \quad \xi \in \partial\varphi(x^+).$$

Therefore, using the subgradient inequality for the convex function $\varphi$, it follows that

$$\lambda(\varphi(x^+) - \varphi(u)) \leq \lambda\langle \xi, \quad x^+ - u\rangle = \langle \nabla h(x) - \nabla h(x^+), \quad x^+ - u\rangle, \quad \forall u \in \operatorname{dom} h.$$

Invoking the three points identity of Lemma 2.2 with $z := u$ and $y := x^+$ gives the desired result. □

### 3.2 A Lipschitz-like convexity condition

The second main pillar of the analysis relies on the recent work of Bauschke et al. [14] which introduces a simple framework capturing the non-Euclidean geometry of a given problem through one *Lipschitz-like/Convexity condition*. It elegantly translates into a general Descent Lemma precisely given in terms of Bregman distances, and in particular allows to handle problems lacking a global Lipschitz gradient, a common assumption used in almost all FOM.

We adopt this framework, with a slight extension, which allows to consider non-convex functions $g$, as recently done in Bolte et al. [24], stressing its usefulness and flexibility in various contexts.

*A Lipschitz-like/Convexity Condition*. [14] Let $h : \mathbb{E} \to (-\infty, \infty]$ be a Legendre function and let $g : \mathbb{E} \to (-\infty, \infty]$ be a proper and lsc function with $\operatorname{dom} g \supset \operatorname{dom} h$, and $g$ is differentiable on $\operatorname{int}\operatorname{dom} h$. Given such a pair of functions $(g, h)$, the *Lipschitz-like/Convexity Condition* denoted by **(LC)** is:

$$\textbf{(LC)} \qquad \exists L > 0 \quad \text{with} \quad Lh - g \text{ convex on } \operatorname{int}\operatorname{dom} h.$$

Note that the above definition holds for any convex function $h$ which is differentiable on any open subset of $\operatorname{dom} h$. Clearly, this condition does not need the Legendre property on $h$. Only the convexity of $Lh - g$ plays a central role. Moreover, the convexity condition **(LC)** nicely translates in terms of Bregman distances to produce a new Descent Lemma (called NoLips in [14]).

**Lemma 3.2** (NoLips Descent Lemma) [14, Lemma 1] *Consider the pair of functions* $(g, h)$ *as above. Take* $L > 0$. *The following statements are equivalent:*

(i) $Lh - g$ *is convex on* int dom $h$, *i.e., condition* **(LC)** *holds.*
(ii) $D_g(x, y) \leq LD_h(x, y) \iff D_{Lh-g}(x, y) \geq 0$ *for all* $x, y \in$ int dom $h$.

*Proof* Simply follows from the gradient inequality for the convex function $Lh - g$, and the fact that $0 \leq D_{Lh-g}(x, y) = LD_h(x, y) - D_g(x, y)$. □

Note that if $L' \geq L$ property **LC** holds with $L'$. When both functions $g$ and $h$ are assumed to be $C^2$ on int dom $h$, the above statement is equivalent to

$$\exists L > 0, \ L\nabla^2 h(x) - \nabla^2 g(x) \succeq 0, \quad \text{for all} \quad x \in \text{int dom}\, h,$$

which can be useful to the check condition **(LC)**, see [14] for examples.

Clearly, in the Euclidean setting, with $h(x) = \frac{1}{2}\|x\|_2^2$, item (ii) above recovers the classical and fundamental Descent Lemma [22] which under the assumption that $g$ is convex, is equivalent to the standard smoothness property: $\nabla g$ is Lipschitz continuous with constant $L$ on $\mathbb{E}$. Observe however, that contrary to the usual Descent Lemma of a differentiable function, the first inequality (ii) of Lemma 3.2 is one-sided. As recently shown in [24], it can be complemented by asking both $Lh - g$ and $Lh + g$ to be convex on int dom $h$, which immediately yields a full Descent Lemma, which reads compactly:

$$\left|D_g(x, y)\right| \leq LD_h(x, y), \quad \forall\, x, y \in \text{int dom}\, h.$$

Clearly, when $g$ is convex, the convexity condition on $Lh + g$ trivially holds.

The structural form of $D_h$ allows to derive another useful and more general descent-like result than Lemma 3.2.

**Lemma 3.3** (The Three Points Descent Lemma) *Consider the pair of functions* $(g, h)$ *as above. Take* $L > 0$. *Then, the function* $Lh - g$ *is convex on* int dom $h$ *if and only if for any* $(x, y, z) \in (\text{int dom}\, h)^3$*:*

$$g(x) \leq g(y) + \langle \nabla g(z), x - y\rangle + LD_h(x, z) - D_g(y, z). \tag{3.1}$$

*Proof* Thanks to the structural form of $D_g$, it is easy to see that for any $(x, y, z) \in (\text{int dom}\, h)^3$, the claimed inequality can be equivalently written as:

$$g(x) - g(z) - \langle \nabla g(z), x - z\rangle \leq LD_h(x, z) \iff D_g(x, z) \leq LD_h(x, z),$$

and the later inequality is equivalent to $D_{Lh-g}(x, z) \geq 0$, namely that $Lh - g$ convex. □

When $g$ is also assumed convex, we have that $D_g(y, z) \geq 0$, and hence one can drop the last term in the inequality (3.1), thus recovering [14, Lemma 4].

We end this section by stressing again, the relevance of the condition **(LC)**. To that end, first recall that $h$ is $\sigma$-strongly convex, (cf. [57, Section 12H]) if there exists $\sigma > 0$ such that

$$\langle u - v, x - y\rangle \geq \sigma \|x - y\|^2, u \in \partial h(x), v \in \partial h(y).$$

For simplicity, here we set $\sigma = 1$, as $h$ can always be re-scaled if necessary.

First order methods based on Bregman distances usually requires *both* the smoothness of $g$ and the 1-strong convexity of $h$. This is stronger than what is actually necessary. Indeed, the following very simple result shows that these two assumptions *imply* condition **(LC)**, thus showing that condition **(LC)** is very natural, and has the nice feature and generality of capturing the essential information in the non-Euclidean setting.

**Proposition 3.1** *Let $\|\cdot\|$ be an arbitrary norm in $\mathbb{E}$. If $\nabla g$ is $L$-Lipschitz continuous on* int dom $h$ *and $h$ is 1-strongly convex, then condition **(LC)** holds true, that is, $Lh - g$ is convex on* int dom $h$.

*Proof* The result follows immediately from the premises made on the pair $(g, h)$. Indeed, for all $x, y \in$ int dom $h$ we have

$$\begin{aligned}\langle \nabla g(x) - \nabla g(y), x - y\rangle &\leq \|\nabla g(x) - \nabla g(y)\|_* \|x - y\|, \text{ [Cauchy}-\text{Schwartz]}\\ &\leq L\|x - y\|^2 \quad \text{[L-Lipschitz gradient of } g\text{]},\\ &\leq L\langle \nabla h(x) - \nabla h(y), x - y\rangle \text{ [}h \text{ is 1-strongly convex]}.\end{aligned}$$

This proves that the gradient of $(Lh - g)$ is monotone on int dom $h$, and hence the convexity of $Lh - g$. □

## 4 Analysis of non-Euclidean proximal based schemes

Thanks to the two main pillars described above in Lemmas 3.1 and 3.3, we now revisit the analysis of some of the most fundamental FOM, stressing simplification in the proofs, which in turn, also allows to present less well-known results, as well as to derive new ones.

We start with the popular additive convex composite model, which despite its simplicity, underscores the basic elements leading to most fundamental FOM, and also covers a broad class of problems arising in a wide spectrum of applications.

### 4.1 The additive composite optimization model

*The Problem and Assumptions.* Consider the following additive convex composite model:

$$\text{(CM)} \qquad v(\mathcal{P}) = \inf\{\Phi(u) := f(u) + g(u) : \ u \in C\},$$

where $C \subseteq \mathbb{E}$ is a closed convex set with a nonempty interior. The following assumptions are made throughout this section.

**Assumption CM** (i) $f : \mathbb{E} \to (-\infty, \infty]$ is proper, lsc and convex,
(ii) $h : \mathbb{E} \to (-\infty, \infty]$ is Legendre, with $\overline{\text{dom}\, h} = C$,
(iii) $g : \mathbb{E} \to (-\infty, \infty]$ is proper, lsc and convex with $\text{dom}\, g \supset \text{dom}\, h$, which is differentiable on $\text{int dom}\, h$, and condition **(LC)** holds: there exists $L > 0$ with $Lh - g$ is convex on $\text{int dom}\, h$,
(iv) $\text{dom}\, f \cap \text{int dom}\, h \neq \emptyset$,
(v) $-\infty < v(\mathcal{P}) = \inf\{\Phi(x) : x \in C\}$.

Given a Legendre function $h$, for all $x \in \text{int dom}\, h$ and any $\lambda > 0$, we define

$$T_\lambda(x) := \text{argmin}\left\{ f(u) + \langle \nabla g(x), u - x\rangle + \frac{1}{\lambda} D_h(u, x) : u \in C \right\}. \tag{4.1}$$

This map emerges from the usual approach which consists of linearizing the differentiable part $g$ around $x$, and regularize it with a proximal distance from that point. Again, when $h$ is the energy function, this is nothing else but the classical proximal gradient map, also called/known as proximal forward–backward splitting map [32,36,55].

Throughout, we suppose that assumption **CM** holds, and recall that we systematically assume that $T_\lambda$ is well-defined and resides in $\text{int dom}\, h$ (cf. Sect. 2 and [14]).

We consider solving problem (CM) with the Bregman Proximal Gradient Method, that is we generate a sequence $\{x^k\}_{k\in\mathbb{N}}$ via the following fixed point iteration for the map $T_\lambda(\cdot)$.

**Bregman Proximal Gradient – BPG**

$$x^0 \in \text{int dom}\, h, \quad x^{k+1} = T_\lambda(x^k),\ k = 0, 1, 2 \ldots \quad (\lambda > 0).$$

The Bregman proximal gradient scheme and its convergence rate have been investigated in other contexts in past studies, e.g., [8,65] under various assumptions. We will come back to this point at the end of this subsection.

*A Mixer in Action: An "All-in-One" Proximal Inequality*. Combining the two main inequalities established in Lemmas 3.1 and 3.3, the mixer in action, gives an "all in one" proximal inequality. Recall that for the result below, we *do not* assume that $g$ is convex.

**Lemma 4.1** (All in One Proximal Inequality) *Let $x \in \text{int dom}\, h$, $\lambda > 0$, and*

$$x^+ = \text{argmin}\left\{ f(u) + \langle \nabla g(x), u - x\rangle + \frac{1}{\lambda} D_h(u, x) :\ u \in C \right\}.$$

*Then, for any $u \in \text{dom}\, h$, we have*

$$\lambda(\Phi(x^+) - \Phi(u)) \leq D_h(u, x) - D_h(u, x^+) - (1 - \lambda L) D_h(x^+, x) - \lambda D_g(u, x). \tag{4.2}$$

*Proof* Writing successively the Bregman proximal inequality (Lemma 3.1) for the convex function $u \to \varphi(u) := f(u) + \langle \nabla g(x), u - x \rangle$, and the Three Points Descent Lemma 3.3 for $g$, yields the following inequalities:

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

Adding these two inequalities, proves the desired result (4.2). □

As a special case, when $g$ is assumed convex, the term $\lambda D_g(u, x)$ is nonnegative and thus can be dropped. In this case we recover the key descent inequality established in Bauschke et al. [14, Lemma 4], that is,

$$\lambda(\Phi(x^+) - \Phi(u)) \le D_h(u, x) - D_h(u, x^+) - (1 - \lambda L) D_h(x^+, x), \ \forall u \in \operatorname{dom} h. \tag{4.3}$$

This inequality at hand is all what is needed to establish the $O(1/n)$ rate of convergence for the BPG method, as established in Bauschke et al. [14, Theorem 1], as well as its relatives: the Bregman proximal minimization and the Bregman projected gradient scheme (hence including as special cases the classical results with the squared Euclidean norm).

**Theorem 4.1** (Basic complexity estimate) [14] *Let* $\{x^k\}_{k \in \mathbb{N}}$ *be the sequence generated by* BPG *with* $\lambda \in (0, L^{-1}]$. *Then,*

(a) *the sequence* $\{\Phi(x^k)\}_{k \in \mathbb{N}}$ *is nonincreasing.*
(b) *Let* $\lambda = L^{-1}$, *then for any* $n \ge 1$,

$$\Phi(x^n) - \Phi(u) \le \frac{L}{n} D_h(u, x^0), \quad \forall\, u \in \operatorname{dom} h.$$

*Proof* Let $k \ge 1$. Invoking Lemma 4.1 (e.g., (4.3)) with $x = x^k, x^+ = x^{k+1} = T_\lambda(x^k)$, with $\lambda \le L^{-1}$ we obtain for any $u \in \operatorname{dom} h$:

$$\lambda(\Phi(x^{k+1}) - \Phi(u)) \le D_h(u, x^k) - D_h(u, x^{k+1}).$$

For $u = x^k$ we obviously have that $D_h(x^k, x^k) = 0$, and since $D_h(\cdot, \cdot) \ge 0$, the above inequality implies that $\Phi(x^{k+1}) \le \Phi(x^k)$, proving (a). Summing the above inequality over $k = 0, \ldots, n-1$, we thus obtain, with $\lambda = L^{-1}$.

$$\sum_{k=0}^{n-1} (\Phi(x^{k+1}) - \Phi(u)) \le L(D_h(u, x^0) - D_h(u, x^n)) \le L D_h(u, x^0).$$

Since by (a) the sequence $\{\Phi(x^k)\}_{k \in \mathbb{N}}$ is nonincreasing, with $s^k := \Phi(x^k) - \Phi(u)$, we have $s^{k+1} \le s^k$, and hence it follows that $s^n \le n^{-1} \sum_{k=0}^{n-1} s^{k+1}$, which together with the above inequality proves the desired result (b). □

*Remark 4.1.* A close inspection of Lemma 4.1 shows again the key role of condition **(LC)**. Indeed, thanks to the structural definition of Bregman distance, it is easily seen by a simple algebraic manipulation that with $\lambda L = 1$ in (4.2), the complexity constant derived in Theorem 4.1(b) given by $\Gamma := L D_h(u, x^0)$ can be replaced by the complexity constant $\Gamma' := D_{Lh-g}(u, x^0)$, which assuming that $g$ is convex, clearly implies $D_{Lh-g}(u, x^0) \leq L D_h(u, x^0)$, thus providing the improved constant $\Gamma' \leq \Gamma$.

As a straightforward consequence of the above analysis, much like in the classical Euclidean squared norm setting, assuming here that $g - \sigma h$ is convex on int dom $h$, for some $\sigma > 0$ (e.g., that $g$ is $\sigma$-strongly convex with respect to $h$, see [11, Definition 4.1]), we can immediately obtain the following linear rate of convergence of BPG for the value gap function $\Phi(x^n) - \Phi_*$. Here we denote by $x^*$ the unique minimizer of $\Phi$, and by $\Phi_*$ its minimal value $\Phi(x^*)$.

**Proposition 4.1.** (Linear Rate of BPG). *Let $\{x^k\}_{k\in\mathbb{N}}$ be the sequence generated by* BPG *with $\lambda = L^{-1}$, and assume that $g - \sigma h$ is convex on* int dom $h$ *for some $\sigma > 0$. Then, for any $n \geq 0$,*

$$\Phi(x^{n+1}) - \Phi_* \leq \left(1 - \frac{\sigma}{L}\right)^{n+1} L D_h(x^*, x^0).$$

*Proof* Simply invoke again Lemma 4.1 with $u := x^*, x = x^k$, and $x^+ = x^{k+1}$, followed by the convexity of $g - \sigma h$, i.e., $D_g(\cdot, \cdot) \geq \sigma D_h(\cdot, \cdot)$, and the nonnegativity of $D_h(\cdot, \cdot)$ to obtain for all $k \geq 0$,:

$$\begin{aligned}
\Phi(x^{k+1}) - \Phi_* &\leq L(D_h(x^*, x^k) - D_h(x^*, x^{k+1})) - D_g(x^*, x^k) \\
&\leq (L - \sigma) D_h(x^*, x^k) - L D_h(x^*, x^{k+1}) \\
&\leq L\left(1 - \frac{\sigma}{L}\right) D_h(x^*, x^k)
\end{aligned}$$

Since $\Phi_* \leq \Phi(x^{k+1})$, applying recursively the resulting second inequality above implies

$$D_h(x^*, x^k) \leq \left(1 - \frac{\sigma}{L}\right)^k D_h(x^*, x^0),$$

and hence the claimed result immediately follows from the third inequality. □

*Global Pointwise Convergence of BPG.* As usual, deriving the global convergence of the sequence generated by proximal based schemes require more subtle arguments, in particular, to determine the "largest possible step size" in a given algorithm, see for instance the analysis in [36] for the classical proximal gradient. As recently shown in [14], it turns out that the notion of a *symmetry coefficient measure* for $D_h$, plays a fundamental role in determining the most aggressive step size for global pointwise convergence of the BPG.

Indeed, Bregman distances are in general not symmetric, except when $h$ is the energy function, for which it simply reduces to the squared Euclidean distance. It is thus natural to introduce a measure for the lack of symmetry in $D_h$, as done in [14].

**Definition 4.1** (*Symmetry coefficient*) Given a Legendre function $h : \mathbb{E} \to (-\infty, \infty]$, its symmetry coefficient is defined by

$$\alpha(h) := \inf \left\{ D(x, y)/D(y, x) \;\middle|\; (x, y) \in \operatorname{int} \operatorname{dom} h \times \operatorname{int} \operatorname{dom} h,\ x \neq y \right\} \in [0, 1]. \tag{4.4}$$

Clearly, perfect symmetry occurs when $\alpha(h) = 1$, i.e., when $h(\cdot) = 2^{-1}\|\cdot\|^2$. Total lack of symmetry, namely $\alpha(h) = 0$, occurs for the key examples $h(x) = x \log x$ and $h(x) = -\log x$, which often arise in applications, while with $h(x) = x^4$ one obtains $\alpha(h) = 2 - \sqrt{3}$, [14].

Equipped with the symmetry coefficient, it was shown in [14] that the most aggressive step size that can be used in BPG satisfies

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

This allows to nicely recover the standard step size $\lambda \in (0, 2/L)$ which warrants global convergence in the classical proximal gradient [36], since in that case $\alpha(h) = 1$.

To establish the global convergence to an optimal solution, additional assumptions on the Bregman proximal distance (satisfied by all examples in Example 2.1) is needed to ensure separation properties of the Bregman distance at the boundary, so that we can use arguments "à la Opial" [53] as done with classical norms.

**Assumption H** (i) For every $x \in \operatorname{dom} h$ and $\beta \in \mathbb{R}$, the level set $\left\{ y \in \operatorname{int} \operatorname{dom} h : D_h(x, y) \leq \beta \right\}$ is bounded.
(ii) If $\{x^k\}_{k\in\mathbb{N}}$ converges to some $x$ in $\operatorname{dom} h$ then $D_h(x, x^k) \to 0$.
(iii) Reciprocally, if $x$ is in $\operatorname{dom} h$ and if $\{x^k\}_{k\in\mathbb{N}}$ is such that $D_h(x, x^k) \to 0$, then $x^k \to x$.

We then have the following global convergence result.

**Theorem 4.2** (BPG: Global Convergence) [14, Theorem 2] *Let $\{x^k\}_{k\in\mathbb{N}}$ be the sequence generated by BPG with $\lambda \in (0, (1 + \alpha(h))/L)$. Assume that $C = \overline{\operatorname{dom} h} = \operatorname{dom} h$, the solution set of $(\mathcal{P})$ $\operatorname{argmin}_C \Phi$ is nonempty and compact, and that Assumption **H** is satisfied. Then, the sequence $\{x^k\}_{k\in\mathbb{N}}$ converges to some solution $x^*$ of problem $(\mathcal{P})$.*

To conclude this part it is interesting to re-emphasize the key role played by the Lipschitz-like convexity condition **(LC)** which is *sufficient* and *weaker* than past studies. For instance, when $f = 0$ the algorithm BPG is the gradient method with Bregman distance studied in [8], and later extended in [65] to handle the composite model (CM). An important difference with both works, is the fact that the usual assumptions: $\nabla g$ globally $L$-Lipschitz continuous and the strong convexity of $h$ are not needed anymore, see also the concluding remarks for further discussion.

### 4.2 Non-Euclidean projected subgradient schemes: mirror descent

Consider the special case of problem (CM) with $g = 0$, namely the *nonsmooth* convex problem

$$(NS) \qquad \min\{f(u) : u \in C\}.$$

We use the notation $f'(u)$ for a subgradient element of the subdifferential set $\partial f(u)$.

To solve problem (NS), we consider the so-called *Mirror Descent* method introduced by Nemirovsky and Yudin [50, Chapter 4]. As shown in Beck and Teboulle [16], the Mirror Descent can be derived and analyzed through the lenses of the proximal framework as a non-Euclidean Bregman projected subgradient scheme and reads as follows:

**Mirror Descent - MD**

$$x^0 \in C \cap \operatorname{int} \operatorname{dom} h,\ x^{k+1} = \operatorname{argmin}\{t_k\langle f'(x^k), u - x^k\rangle + D_h(u, x^k) : u \in C\},\ t_k > 0,\ k = 0, 1, 2, \ldots.$$

The usual and standard assumptions to establish the rate of convergence and iteration complexity of MD require:

(a) $\sigma$-strong convexity of $h$,
(b) Lipschitz continuity of the convex function $f$, namely there exists $L_f > 0$ such that

$$|f(x) - f(y)| \leq L_f \|x - y\|, \quad \text{for all} \quad x, y \in \operatorname{dom} f,$$

which implies that for all $x \in \operatorname{int} \operatorname{dom} f$, and $f'(x) \in \partial f(x)$, we have $\|f'(x)\|_* \leq L_f$.

As we shall see, we can *relax* these requirements with a *weaker* and more general assumption, which is in the spirit of condition **(LC)**, namely with a condition capturing the data information on both $f$ and $h$ within a *single condition*, and which naturally emerges from a close inspection of the proof derived in [16].

To see this, applying Lemma 3.1 on $\varphi(u) = \langle f'(x^k), u - x^k\rangle + \delta_C(u)$ with $\lambda = t_k$, $x = x^k$ and $x^+ = x^{k+1}$, we get the iterate produced by (MD), which then yields the first inequality below, while the second inequality follows from the subgradient inequality for the convex function $f$:

$$\begin{aligned} t_k\langle f'(x^k), x^{k+1} - u\rangle &\leq D_h(u, x^k) - D_h(u, x^{k+1}) - D_h(x^{k+1}, x^k), \\ t_k(f(x^k) - f(u)) &\leq t_k\langle f'(x^k), x^k - u\rangle \\ &= t_k\langle f'(x^k), x^{k+1} - u\rangle + t_k\langle f'(x^k), x^k - x^{k+1}\rangle. \end{aligned}$$

Adding these two inequalities, we then obtain for all $u \in \operatorname{dom} h$:

$$\begin{aligned} t_k(f(x^k) - f(u)) \leq &\left[D_h\left(u, x^k\right) - D_h\left(u, x^{k+1}\right)\right] \\ &+ t_k\left\langle f'(x^k), x^k - x^{k+1}\right\rangle - D_h\left(x^{k+1}, x^k\right). \end{aligned} \tag{4.5}$$

To continue the standard proof of MD and obtain a meaningful estimate in terms of some objective function gap, we observe that using standard telescoping arguments over $k = 0, \ldots, n$, clearly the first term in the squared brackets will behave nicely and collapse to $D_h(x^0, x^n)$. Therefore, to complete the proof, all what we need is to find an adequate bound for the expression given by the two last terms in (4.5). In the classical proof of MD, this is obtained through the standard Assumptions (a) and (b) alluded above, (cf. [16, Theorem 4.1, inequality (4.21)]). This naturally leads us to consider instead, the following weaker condition on the pair $[f, h]$.

**Condition W[f,h]**. For any $t > 0$, there exists $G > 0$, such that

$$t\langle f'(x), x - u\rangle - D_h(u, x) \leq \frac{t^2}{2}G^2, \quad \text{for all} \quad u \in \operatorname{dom} h, \quad x \in \operatorname{int} \operatorname{dom} h.$$

The very simple result below shows that this condition is weaker (hence the **W**!) than the usual *separate* Assumptions (a) and (b) on $[f, h]$ alluded above which are made in the classical analysis of the MD method.

**Proposition 4.2** *Assumptions (a) and (b) imply the validity of **W[f,h]** with* $G = L_f/\sqrt{\sigma}$.

*Proof* Using the the Fenchel–Young inequality $2\langle a, b\rangle \leq s^{-1}\|a\|_*^2 + s\|b\|^2$ for all $a, b \in \mathbb{E}$ and $s > 0$, with $a := tf'(x), b := x - u$ and $s := \sigma$ followed by the $\sigma$-strong convexity of $h$, we obtain for any $t > 0$, $u \in \operatorname{dom} h$, and $x \in \operatorname{int} \operatorname{dom} f$:

$$\begin{aligned} t\langle f'(x), x - u\rangle - D_h(u, x) &\leq \frac{t^2}{2\sigma}\|f'(x)\|_*^2 + \frac{\sigma}{2}\|x - u\|^2 - D_h(u, x) \\ &\leq \frac{t^2}{2\sigma}\|f'(x)\|_*^2 \leq \frac{t^2}{2\sigma}L_f^2, \end{aligned}$$

proving that **W[f,h]** holds with $G = L_f/\sqrt{\sigma}$. □

Under our standing assumption on problem $(NS)$, summarizing the above developments for the MD algorithm, we obtain the following main result expressed in terms of the best value function $\min_{0\leq k\leq n} f(x^k)$ (or, likewise, thanks to Jensen's inequality, in terms of the average value function, namely $f(\hat{x}^n)$, where $\hat{x}^n = (\sum_{k=0}^n t_k)^{-1}\sum_{k=0}^n t_k f(x^k)$).

**Theorem 4.3** (Best value function gap estimate for MD) *Let $\{x^k\}_{k\in\mathbb{N}}$ be the sequence generated by the MD algorithm, and suppose that condition **W[f,h]** holds for all $u \in \operatorname{dom} h$ and $x^k \in \operatorname{int} \operatorname{dom} h$, $t_k > 0$. Then, for any $k \geq 0$ and $u \in \operatorname{dom} h$,*

$$\min_{0\leq k\leq n} f(x^k) - f(u) \leq \frac{D_h(u, x^0) + \frac{G^2}{2}\sum_{k=0}^n t_k^2}{\sum_{k=0}^n t_k}.$$

*Proof* Thanks to condition **W[f,h]**, with $t := t_k$, and $x := x^k$ we get from (4.5)

$$t_k(f(x^k) - f(u)) \leq \Big[D_h(u, x^k) - D_h(u, x^{k+1})\Big] + \frac{t_k^2}{2}G^2, \quad \text{for all} \quad u \in \operatorname{dom} h.$$

Summing the above inequality over $k = 0, \dots n$ we obtain for all $u \in \operatorname{dom} h$

$$\sum_{k=0}^{n} t_k(f(x^k) - f(u)) \le D_h(u, x^0) - D_h(u, x^{n+1}) + \frac{G^2}{2}\sum_{k=0}^{n} t_k^2,$$

and hence, the claimed result follows at once. □

From the above result, assuming that the optimal set $X_*$ of problem (NS) is nonempty, let $x^* \in X_*$, and suppose that $R(x^0) := \max_{x\in C} D_h(x^*, x^0) < \infty$. Then, the $O(1/\sqrt{n})$ rate of convergence result easily follows as done in [16] by minimizing the right hand side above with respect to $t_k > 0$ (cf. [16, Proposition 4.1]) to obtain $t_k = \frac{1}{G}\sqrt{\frac{2R(x^0)}{n+1}}$, $k = 0, \dots, n$, from which we get

$$\min_{0\le k\le n} f(x^k) - f(x^*) \le G\sqrt{\frac{2R(x^0)}{(n+1)}}.$$

The same line of analysis can be adapted in a straightforward way for the composite model (CM), when both $f$ *and* $g$ *are nonsmooth*, $C = \mathbb{E}$, which results in a Bregman proximal-subgradient scheme:

**Bregman Proximal Subgradient - Composite MD**

$$x^{k+1} = \operatorname{prox}^h_{t_k g}(x^k - t_k f'(x^k)) = \operatorname{argmin}\{t_k g(u) + t_k\langle f'(x^k), u - x^k\rangle + D_h(u, x^k) : u \in \mathbb{E}\}.$$

This scheme was analyzed by Duchi et al. [39], under the usual strong convexity of $h$ and Lipschitz continuity of $f$, and by assuming in addition that $g(\cdot)$ is a nonnegative function. Here, our analysis implies that the proof can be done in analogous way, but under the weaker condition **W[f,h]** on $f$. Alternatively, one can also get away from the additional nonnegativity assumption on $g$, at the price of increasing the constant in the estimation rate, where $G$ which was attributed to $f$, should be replaced with a constant $G'$, so that the sum $f + g$, satisfies the condition **W[f+g,h]**.

### 4.3 Extension to convexly composite model

We outline here an extension that can benefit from the above developments, and consider the convexly composite model

$$(\mathrm{Cvx} - \mathrm{CM}) \quad \inf\{c(x) := f(x) + \theta(G(x)) : x \in \mathbb{R}^d\},$$

where $G(x) = (g_1(x), \dots, g_m(x))$.

Here $\theta : \mathbb{R}^m \to (-\infty, \infty]$ is a proper, lsc and convex function, which is assumed to be nonincreasing with respect to its $i$-th argument for nonlinear $g_i$. The assumptions are as before for $f$ and $h$, where we also assume that condition **(LC)** holds for each pair $(g_i, h)$, i.e., there exists $L_i > 0$ such that $L_i h - g_i$, where for simplicity we assume here the same $h$ for each pair (in this respect, see also Remark 4.2 below.)

Under the above assumptions, clearly (Cvx-CM) is a convex problem, which is far more flexible than the additive (CM) (easily recovered with $\theta$ being the identity map, and with $g(x) = g_1 \equiv g, m = 1$) and covers many important classical convex optimization models, such as, minimax of a finite collection, standard constrained convex programming, Lagrangians-type/penalty objectives, ect.. This model was popular in the mid 80's and merely abandoned. However, it has recently re-appeared on the surface, see for instance the recent works: [45] which also includes additional elaborated coverage on proximal methods and earlier references motivating this model; [38] which provides a comprehensive analysis on the role of error bounds in the linear convergence of proximal methods for this class, and the very recent work [25] in the context of nonconvex Lagrangian methods and their variants.

Suppose we want to solve (Cvx-CM) via BPG. Before hand, it will be convenient to introduce the following notations. For $i = 1, \ldots, m$, let $l_{g_i}(x, y) := g_i(y) + \langle \nabla g_i(y), x - y \rangle$, and define

$$
\begin{aligned}
G(x, y) &:= (l_{g_1}(x, y), \ldots, l_{g_m}(x, y)), \\
\mathbf{L} &:= (L_1, \ldots, L_m), \\
I_a &:= \{1 \leq i \leq m : \ g_i \text{ is affine}\} \quad \text{and} \quad V := \left\{v \in \mathbb{R}^m_+ : \ v_i = 0, \ \text{for } i \in I_a\right\}.
\end{aligned}
$$

Let $x^0 \in \operatorname{dom} f \cap G^{-1}(\operatorname{dom} \theta) \cap \operatorname{int} \operatorname{dom} h$. The Bregman Proximal Gradient iteration for (Cvx-CM) reads

**BPG for (Cvx-CM)**

$$x^{k+1} = \operatorname{argmin}\{f(u) + \theta(l_G(u, x^k)) + \lambda^{-1} D_h(u, x^k) : u \in \mathbb{R}^d\}, k = 0, 1, 2, \ldots \quad (\lambda > 0).$$

Throughout, we assume that $x^{k+1}$ is well defined, and in $\operatorname{int} \operatorname{dom} h$ (this can be ensured through additional technical assumptions along the lines of Lemma 2.3; for simplicity we omit the details).

From the proof-mechanism revealed in the analysis of the previous schemes, it is clear that all what is needed to analyze this case is an appropriate Descent-like Lemma for the composite term $\theta(G(x))$. From there, the rate of convergence analysis of BPG will then follow in an analogous way. It turns out, that we can express the adequate choice of $\lambda$ in term of the asymptotic (recession) function of $\mathbf{L}$, a notion that we now recall with its basic properties, see [58, p. 87] and [5, p. 53].

**Proposition 4.3** (Asymptotic function) *Let $\theta : \mathbb{R}^m \to (-\infty, \infty]$ be a proper, lsc and convex function. Then, the following properties hold for the asymptotic function of $\theta$ which is denoted by $\theta_\infty$:*

(a) $\theta_\infty(d) = \sup\{\theta(x + d) - \theta(x) : \ x \in \operatorname{dom} \theta\}$ *for all* $d \in \mathbb{R}^m$.
(b) $\theta_\infty$ *is proper, lsc, convex and positively homogeneous, i.e.,* $\theta_\infty(sd) = s\theta_\infty(d)$ *for every* $s > 0$.
(c) *Let* $L_\theta \geq 0$. *Then,* $\operatorname{dom} \theta = \mathbb{R}^m$ *and* $\theta$ *is* $L_\theta$*-continuous if and only if*

$$\theta_\infty(d) \leq L_\theta \|d\|, \quad \textit{for all} \quad d \in \mathbb{R}^m.$$

We then have the following descent-like property for the convexly composite function $\theta(G(x))$.

**Lemma 4.2** (Descent-like property for a convexly composite function) *Let $\theta : \mathbb{R}^m \to (-\infty, \infty]$ be proper, lsc, convex and increasing for $i \notin I_a$, and let $x \in \operatorname{int} \operatorname{dom} h$. Then, for any $u \in \operatorname{dom} h$, the following statements hold:*

(a) $\theta(G(u)) - \theta(l_G(u, x)) \leq \theta_\infty(\mathbf{L}) D_h(u, x)$, *with* $\mathbf{L} \in V$ *and* $\theta_\infty(\mathbf{L}) \geq 0$.
(b) *Suppose that $\theta : \mathbb{R}^m \to \mathbb{R}$ is Lipschitz continuous, with constant $L_\theta > 0$. Then, we have*

$$\theta(G(u)) - \theta(l_G(u, x)) \leq L_\theta \|\mathbf{L}\| D_h(u, x).$$

*Proof* For any $i = 1, \ldots, m$, the convexity of $g_i$ together with the condition **(LC)** of each pair[1] $(g_i, h)$ imply the validity of the following inequality for each coordinate:

$$0 \leq G(u) - l_G(u, x) \leq \mathbf{L} D_h(u, x),$$

that is, $G(u) - l_G(u, v) \in V$. Since $\theta$ is nondecreasing in each of its arguments $i \notin I_a$, it follows from Proposition 4.3(a) that $\theta_\infty(\mathbf{L}) \geq 0$. To establish the claimed inequality, using $G(u) - l_G(u, v) \in V$, and the monotonicity of $\theta$, we get:

$$\begin{aligned}
\theta(G(u)) - \theta(l_G(u, x)) &\leq \theta(l_G(u, x) + G(u) - l_G(u, x)) - \theta(l_G(u, x)) \\
&\leq \theta(l_G(u, x) + \mathbf{L} D_h(u, x)) - \theta(l_G(u, x)) \\
&\leq \theta_\infty(\mathbf{L} D_h(u, x)) \\
&= D_h(u, x) \theta_\infty(\mathbf{L}),
\end{aligned}$$

where the last inequality and the equality follow from Proposition 4.3(a) and (b), respectively.

(b) Since we assumed that $\theta$ if finite-valued and $L_\theta$ continuous on $\mathbb{R}^m$, the desired statement follows from item (a) and Proposition 4.3(c). □

Thanks to this lemma, the convergence rate result of BPG for the convexly composite model (Cvx-CM) follows by the same arguments as done in Theorem 4.1 for the BPG for the additive composite model (CM), where the only change here is in the choice of $\lambda$ to produce the same $O(1/n)$ rate for problem (Cvx-CM) with an adjusted complexity constant; this is recorded in the following theorem.

**Theorem 4.4** (Complexity estimate of BPG for (Cvx-CM)) *Let $\{x^k\}_{k \in \mathbb{N}}$ be the sequence generated by BPG for (Cvx-CM) with $\lambda^{-1} = L_\theta \|\mathbf{L}\|$. Then, for all $n \geq 1$,*

$$c(x^n) - c(u) \leq \frac{L_\theta \|\mathbf{L}\| D_h(u, x^0)}{n}, \quad \forall\, u \in \operatorname{dom} h.$$

[1] Note that when $i \in I_a$, we can take $L_i = 0$ and hence the condition **(LC)** still holds, since $L_i h - g_i = -g_i$ with $g_i$ affine.

Note that the same rate with complexity constant $\theta_\infty(\mathbf{L})$ would remain valid when $\theta$ is extended valued, provided one can warrant in this case that $0 < \theta_\infty(\mathbf{L}) < \infty$.

*Remark 4.2* An extension presuming a different $h_i$ for each $g_i$ with the condition $L_i h_i - g_i$ convex for each $i$, can also be established. For brevity we omit the additional technical details. It can be shown that in that case it would result in a simple adjustment in the right hand side of Lemma 4.2 and reads for any $u \in \operatorname{dom} h$:

$$\theta(G(u)) - \theta(l_G(u,x)) \leq L_\theta \|\mathbf{L}\| \max_{1\leq i\leq m} D_{h_i}(u,x).$$

## 5 From convex to nonconvex models

The nonconvex setting is far more difficult than its convex counterpart, and remains highly challenging. Convergence to a global optimum is usually out of reach and rate of convergence results are obviously limited and weaker. For the *classical* proximal gradient equipped with the squared Euclidean norm, the analysis goes back to Fukushima and Milne [42], and more recent works include e.g., [4,29]. Moreover, these works also imposed the usual restrictive global Lipschitz continuity of the gradient of $g$. Very recently, these results have been significantly improved in Bolte et al. [24] within the non-Euclidean proximal framework for the fully nonconvex composite model, namely both $f$ and $g$ being nonconvex, and with $g$ lacking a global Lipschitz continuous gradient.

For brevity, here we describe some of these recent results derived in [24] by focusing on the following simpler nonconvex model, which consists of minimizing the sum of a *convex* nonsmooth function with a *nonconvex* differentiable function satisfying the condition **(LC)**. This simplified model is already of sufficient importance in many applications, and we refer the reader to [24] for details on analyzing the more general nonconvex model.

### 5.1 Nonconvex BPG: rate of convergence

Throughout this section we consider the composite problem $(\mathcal{P})$ defined on $C \equiv \mathbb{E} = \mathbb{R}^d$, namely

$$(NC) \qquad v_* = \inf\left\{\Psi(u) := f(u) + g(u) : u \in \mathbb{R}^d\right\},$$

under the following standing assumptions.

**Assumption NC** (i) $f : \mathbb{R}^d \to (-\infty, \infty]$ is proper, lsc and convex,
(ii) $\operatorname{dom} h = \mathbb{R}^d$ and $h$ is $\sigma$-strongly convex on $\mathbb{R}^d$,
(iii) $g : \mathbb{R}^d \to \mathbb{R}$ is *nonconvex* differentiable, and satisfies condition **(LC)**,
(iv) $v_* := \inf\left\{\Psi(u) : u \in \mathbb{R}^d\right\} > -\infty$.

As in the convex case, we consider the BPG iteration, which here we start with $x^0 \in \mathbb{R}^d$, and generate the sequence $\{x^k\}_{k\in\mathbb{N}}$ via its Bregman proximal gradient map:

$$x^{k+1} = T_\lambda(x^k) := \operatorname{argmin}\left\{ f(u) + \langle \nabla g(x^k),\ u - x^k\rangle + \frac{1}{\lambda} D_h(u, x^k) :\ u \in \mathbb{R}^d \right\}.$$

Clearly, here since $\operatorname{dom} h = \mathbb{R}^d$ and $h$ is strongly convex, the mapping $T_\lambda$ is well-defined and single-valued. We have the following rate of convergence properties recently derived in [24].

**Theorem 5.1** *Suppose that Assumption NC holds. Let $\{x^k\}_{k\in\mathbb{N}}$ be the sequence generated by BPG with $0 < \lambda L < 1$. Then, the sequence $\left\{\Psi\left(x^k\right)\right\}_{k\in\mathbb{N}}$ is nonincreasing, and for every $n \geq 1$ we have*

$$\min_{1\leq k\leq n} \|x^{k+1} - x^k\| \leq \frac{1}{\sqrt{n}} \left( \frac{\lambda(\Psi\left(x^0\right) - \Psi_*)}{\sigma(1-\lambda L)} \right)^{1/2}.$$

*Proof* Under the condition **(LC)**, we can invoke Lemma 4.1 with $u = v = x^k$ and $u^+ = x^{k+1} = T_\lambda(x^k)$ to obtain

$$\begin{aligned} \lambda(\Psi(x^{k+1}) - \Psi(x^k)) &\leq -(1-\lambda L) D_h(x^{k+1}, x^k) - D_h(x^k, x^{k+1}) \\ &\leq -(1-\lambda L) D_h(x^k, x^{k+1}). \end{aligned} \tag{5.1}$$

Since $0 < \lambda L < 1$, this proves that $\left\{\Psi\left(x^k\right)\right\}_{k\in\mathbb{N}}$ is nonincreasing. Summing the inequality (5.1) over $k = 0, \ldots, n$, and using the $\sigma$-strong convexity of $h$, it follows that

$$\begin{aligned} n\frac{\sigma(1-\lambda L)}{2} \min_{1\leq k\leq n} \|x^k - x^{k+1}\|^2 &\leq (1-\lambda L) \sum_{k=0}^{n} D_h(x^k, x^{k+1}) \\ &\leq \lambda(\Psi(x^0) - \Psi(x^n)), \end{aligned}$$

and with $\Psi(x^n) \geq \Psi_* > -\infty$, the claimed desired result follows. □

Note that this result recovers the classical rate of convergence result of the proximal gradient algorithm for the nonconvex composite model [18, Theorem 2.3], with $h$ being the energy function, $\sigma = 1$ and $\lambda L = 1/2$. However, a major difference here is that we *do not* require the gradient of the nonconvex function $g$ to be globally Lipschitz continuous. Instead, once again, condition **(LC)** plays the central role to adapt to the geometry of the problem at hand. This allows to treat important models; see [24] for an interesting application with simple globally convergent algorithms for the so-called class of *quadratic inverse problems* with sparsity constraints, which having a quartic objective, fails to have a global Lipschitz gradient.

### 5.2 Global convergence to a critical point

To prove the global convergence of the sequence $\{x^k\}_{k\in\mathbb{N}}$ generated by BPG to a critical point of $\Psi$, we first outline a general abstract convergence mechanism as recently developed in [29], which is very flexible and can be applied to any given algorithm satisfying the premises described below. To this end, let $F : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function which is bounded from below and consider the problem

$$\inf \left\{ F(x) : \ x \in \mathbb{R}^d \right\}.$$

Consider a generic algorithm $\mathcal{A}$ which generates a sequence $\{x^k\}_{k\in\mathbb{N}}$ via the following:

$$\text{start with any } x^0 \in \mathbb{R}^d \quad \text{and set} \quad x^{k+1} \in \mathcal{A}\left(x^k\right), \quad k = 0, 1, \ldots.$$

The main goal is to prove that the *whole sequence* $\{x^k\}_{k\in\mathbb{N}}$, generated by the algorithm $\mathcal{A}$, converges to a critical point $x^*$ of $F$, namely $0 \in \partial F(x^*)$. Note that in this general setting, $\partial F$ stands for the limiting subdifferential of $F$, cf. [57].

**Definition 5.1** (*Gradient-like Descent Sequence*) A sequence $\{x^k\}_{k\in\mathbb{N}}$ is called *a gradient-like descent sequence* for $F$ if the following three conditions hold:

(C1) *Sufficient decrease property.* There exists a positive scalar $\rho_1$ such that

$$\rho_1 \left\| x^{k+1} - x^k \right\|^2 \leq F(x^k) - F(x^{k+1}), \quad \forall\, k \in \mathbb{N}.$$

(C2) *A subgradient lower bound for the iterates gap.* There exists a positive scalar $\rho_2$ such that

$$\left\| w^{k+1} \right\| \leq \rho_2 \left\| x^{k+1} - x^k \right\|, \quad \text{for some } w^{k+1} \in \partial F(x^{k+1}), \quad \forall\, k \in \mathbb{N}.$$

(C3) Let $\overline{x}$ be a limit point of any subsequence $\left\{x^k\right\}_{k\in\mathcal{K}}$, then $\limsup_{k\in\mathcal{K}\subset\mathbb{N}} F\left(x^k\right) \leq F(\overline{x})$.

These three conditions are typical for any *descent* type algorithm, and basic to prove subsequential convergence, see e.g., [1,29,66].

To establish global convergence of the *whole* sequence, we need an additional assumption on the class of functions $F$: it must satisfy the so-called nonsmooth Kurdyka–Łojasiewicz (KL) property [27] (see [44,46] for smooth cases). We refer the reader to [28] for an in depth study of the class of KL functions, as well as references therein.

Verifying the KL property of a given function might often be a difficult task. However, thanks to a fundamental result established by Bolte et al. [27], it holds for the broad class of *semi-algebraic* functions, which abound in applications, see [29,66], and references therein.

The last ingredient needed to activate this generic framework, is a key uniformization of the KL property proven in [29, Lemma 6, p. 478]. Then, we can prove the following general theorem (see [29] for details leading to this result).

**Theorem 5.2** (Global Convergence) *Let $\{x^k\}_{k\in\mathbb{N}}$ be a bounded gradient-like descent sequence for $F$. If $F$ satisfies the KL property, then the sequence $\{x^k\}_{k\in\mathbb{N}}$ has finite length, i.e., $\sum_{k=1}^{\infty} \left\|x^{k+1} - x^k\right\| < \infty$, and it converges to a critical point $x^*$ of $F$.*

Equipped with this generic result, we can establish the global convergence result of BPG to a critical point of $\Psi$. Thanks to Fermat's rule [57, Theorem 10.1, p. 422], the set of critical points of $\Psi$ here is given by

$$\operatorname{crit}\Psi = \left\{x \in \mathbb{R}^d : \ 0 \in \partial\Psi(x) \equiv \partial f(x) + \nabla g(x)\right\}.$$

In brief, let us mention that the sufficient descent property established in Theorem 5.1 (cf. (5.1)) allows to prove conditions C1 and C3, while the following mild additional assumption on the pair $(g, h)$ is needed to establish condition C2.

**Assumption NC[+]** $\nabla h$ and $\nabla g$ are Lipschitz continuous on any *bounded subset* of $\mathbb{R}^d$.

We then have the following simplified version of the more general convergence result proved in [24].

**Theorem 5.3** (Convergence of BPG) *Suppose that assumptions **NC** and **NC[+]** hold. Let $\{x^k\}_{k\in\mathbb{N}}$ be a sequence generated by BPG which is assumed to be bounded, and let $0 < \lambda L < 1$. The following assertions hold:*

(i) Subsequential convergence. *Any limit point of the sequence $\{x^k\}_{k\in\mathbb{N}}$ is a critical point of $\Psi$.*

(ii) Global convergence. *Suppose that $\Psi$ satisfies the KL property on* dom $\Psi$*, which is in particular true when $f$ and $g$ are semialgebraic. Then, the sequence $\{x^k\}_{k\in\mathbb{N}}$ has finite length, and converges to a critical point $x^*$ of $\Psi$.*

Convergence rate results of the generated sequence described in Theorem 5.3 can also be derived by applying the generic rate of convergence result of Attouch and Bolte [1].

## 6 Concluding remarks

We have provided a concise synthesis outlining the key theoretical elements playing a central role in the convergence analysis of Non Euclidean Bregman based proximal methods, and their most fundamental first order algorithm relatives, stressing clarifications and simplifications through elementary proof-patterns. The volume of literature on first order methods has exploded over the last decade, and we refer the reader to some of the recent pointers given in the bibliography for further elaborations, extensions and applications.

Below, we briefly further discuss the benefits and limitations of the non-Euclidean proximal framework, and end with an open challenging question.

Much like any proximal minimization scheme, the underlying framework depends on how efficiently we can compute the iteration formula of BPG (cf. (4.1)):

$$x^+ = T_\lambda(x) := \operatorname{argmin}\left\{ f(u) + \langle \nabla g(x), u - x\rangle + \frac{1}{\lambda} D_h(u, x) : u \in C \right\}.$$

Although not visible at first sight, it is interesting to observe that $T_\lambda(\cdot)$ shares the same structural *splitting* principle as the classical Euclidean proximal-gradient map, which can be useful for computational purpose allowing to decompose it through two specific operators. Indeed, as shown in [14, Section 3.1], writing the optimality condition for $x^+$, formal computations show that by defining the following operators:

- *A Bregman gradient step*

$$p_\lambda(x) := \operatorname{argmin}\left\{ \langle \nabla g(x), u\rangle + \frac{1}{\lambda} D_h(u, x) : u \in \mathbb{E} \right\}, \quad x \in \operatorname{int}\operatorname{dom} h, \tag{6.1}$$

- *A Bregman proximal map*

$$\operatorname{prox}^h_{\lambda f}(y) := \operatorname{argmin}\{\lambda f(u) + D_h(u, y) : u \in C\}, \quad y \in \operatorname{int}\operatorname{dom} h, \tag{6.2}$$

one can rewrite the map $T_\lambda(\cdot)$ simply as the composition of a Bregman proximal step with a Bregman gradient step:

$$x^+ = \operatorname{prox}^h_{\lambda f}(p_\lambda(x)). \tag{6.3}$$

Thus, computing $T_\lambda$ depends on whether the above two operators can output closed-form solutions or can be numerically computed in an efficient way. In the classical Euclidean proximal gradient method, i.e., when $h$ is the energy function, the first computation above reduces to a standard gradient step, $p_\lambda(x) = x - \lambda\nabla g(x)$, and the second one is the usual Moreau proximal map of $\lambda f$. In the general case, for any $x \in \operatorname{int}\operatorname{dom} h$, define $v(x) := \nabla h(x) - \lambda\nabla g(x)$. Then, the Bregman gradient step (6.1) reduces to

$$p_\lambda(x) = \nabla h^*(v(x)), \quad \text{with} \quad v(x) \in \operatorname{dom}\nabla h^*.$$

Therefore, once we know the conjugate function $h^*$ of $h$, the computation of $p_\lambda$ is straightforward. This is the case for the five examples listed in Example 2.1, as well as for the $n$-dimensional "Hellinger-Like function" defined by $h(x) = -\sqrt{1 - \|x\|^2}$ with $\operatorname{dom} h = \{x \in \mathbb{R}^n : \|x\| \leq 1\}$, which is relevant for ball constraints. In the latter case, one obtains $h^*(y) = \sqrt{1 + \|y\|^2}$, with $\operatorname{dom} h^* = \mathbb{R}^n$, and hence $p_\lambda(x) = (1 + v^2(x))^{-1/2} v(x)$. For further details and interesting examples evaluating $p_\lambda$ for nonseparable Bregman distances, and for handling conic constraints with Bregman distances see [14, Section 3.1] and [8, Examples B, C, p. 718].

We now discuss the computation of the second operator, namely the Bregman proximal map given in (6.2). In the classical case, the Moreau proximal map of $f$ is in

general computable in closed form whence $f$ is "norm-like", or when $f$ is the indicator of sets whose geometry is favorable to Euclidean projections, which we refer here as prox-friendly, see e.g., [36, Section 2.6] and [15, pp. 156 and 177] for many such examples. However, beyond such prox-friendly functions/sets, in general computing the classical proximal map can be very difficult. Concerning (6.2) the situation is similar. Sets and functions which will be prox-friendly with respect to a Bregman proximal term $D_h$, will share a mathematical structure similar to $h$, which is chosen to adapt to the geometry of the given function/set. Moreover, to activate BPG, the choice for $h$ is also intimately related to the step-size expressed in terms of $L$, which is determined by the condition (LC). Such computations are illustrated in details for the convex setting in [14, Section 5], leading to explicit simple algorithms to tackle the important class of Poisson linear inverse problems which is prevalent in the statistical and image sciences application areas. In the nonconvex setting, computations of $T_\lambda$ and $L$, can be found in the very recent work [24, Section 5] for the broad class of *quadratic inverse* problems described by an unconstrained problem with an $l_1$-norm regularizer and a nonconvex sparsity constrained problem, resulting in two simple closed-form Bregman proximal gradient algorithms. It would be interesting to expand proximal calculus with Bregman distances that can be of benefit to further relevant applications.

This survey would be incomplete without discussing the perspectives for fast non-Euclidean proximal based schemes, in the spirit of the classical fast gradient method of Nesterov [51], and its extension to the composite convex model, FISTA of Beck and Teboulle [17]. Actually, both do not seem to be extendible in their basic formulation to the Bregman distance setting. A remedy to this situation was addressed by Auslender and Teboulle [8, Section 5] who proposed a Bregman gradient scheme (called IGA-Improved Interior Gradient Algorithm) and which reads as follows:

**Algorithm AT [8]**
**Initialization.** $L > 0, x^0 \in \operatorname{dom} h, z^0 = x^0, t_0 = 1.$
**Iteration.** For $k \geq 0$, compute

$$\begin{aligned} y^k &= (1 - t_k^{-1})x^k + t_k^{-1} z^k \\ z^{k+1} &= \operatorname{argmin}\{\langle \nabla g(y^k), x - y^k \rangle + t_k^{-1} L D_h(x, z^k) : \ x \in C\} \\ x^{k+1} &= (1 - t_k^{-1})x^k + t_k^{-1} z^{k+1} \\ t_{k+1} &= 2^{-1}(1 + \sqrt{1 + 4t_k^2}). \end{aligned}$$

This algorithm was proven to achieve the faster rate $O(1/n^2)$, [8, Theorem 5.2]. Later on it was extended by Tseng [65] to handle as well, the additive convex composite model (CM), where the Bregman gradient $z$-step should be replaced by the Bregman proximal-gradient iteration $z^{k+1} = T_{\lambda_k}(y_k)$ with $\lambda_k = t_k L^{-1}$. However, for both algorithms, these results were derived only under the assumptions that $g$ as an $L$-globally Lipschitz gradient, and $h$ is strongly convex, two assumptions that we have specifically avoided throughout this paper. As pointed out in [14], one central and interesting topic for further research is to devise fast FOM capable of lifting both restrictions. In a recent study [60], numerical experiments based on the AT scheme (and some other variants) applied to problems satisfying the (**LC**) condition (i.e.,

lacking global Lipschitz gradient), and with $h$ *not* strongly convex exhibit the very same faster rate. The theoretical justification, however, is still lacking. We thus end this paper with the following, which we believe remains as a challenging open question:

*Under the framework of this paper, can we produce first order algorithms with a proven faster convergent rate?*

**Acknowledgements** This synthesis has emerged from my long standing cooperation with Alfred Auslender on this topic. I am deeply indebted to him, not only for generously sharing with me over the past two decades his profound mathematical knowledge, enthusiasm, and vision in optimization, but also for his friendship and the resulting pleasure, fun and adventures beyond our mathematical activities. The material covered here also benefited greatly, and comes from past and recent works written with Amir Beck, Jérôme Bolte and Shoham Sabach, with whom I have had the pleasure to work with over many years. My deepest thanks to all of them for this fructuous past and continuing collaboration, and to Heinz Bauschke, for the pleasure of collaborating with him on this topic over the past 3 years. I am grateful to the editors and the referees of the paper, whose constructive comments and suggestions were very useful to improve the paper presentation.

## References

1. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Program. **116**, 5–16 (2009)
2. Attouch, H., Teboulle, M.: A regularized Lotka–Volterra dynamical system as a continuous proximal-like method in optimization. J. Optim. Theory Appl. **121**, 541–570 (2004)
3. Attouch, H., Bolte, J., Redont, P.: Optimizing properties of an inertial dynamical system with geometric damping: link with proximal methods. Control Cybern. **31**, 643–657 (2002)
4. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. Math. Program. **137**, 91–129 (2013)
5. Auslender, A., Teboulle, M.: Asymptotic Cones and Functions in Optimization and Variational Inequalities. Springer, New York (2003)
6. Auslender, A., Teboulle, M.: Interior gradient and Epsilon-subgradient methods for constrained convex minimization. Math. Oper. Res. **29**, 1–26 (2004)
7. Auslender, A., Teboulle, M.: Interior projection-like methods for monotone variational inequalities. Math. Program. **104**, 39–68 (2005)
8. Auslender, A., Teboulle, M.: Interior gradient and proximal methods for convex and conic optimization. SIAM J. Optim. **16**, 697–725 (2006)
9. Auslender, A., Teboulle, M.: Projected subgradient methods with non-Euclidean distances for non-differentiable convex minimization and variational inequalities. Math. Program. Ser. B **120**, 27–48 (2009)
10. Auslender, A., Teboulle, M., Ben-Tiba, S.: Interior proximal and multiplier methods based on second order homogeneous kernels. Math. Oper. Res. **24**, 645–668 (1999)
11. Bartlett, P.L., Hazan, E., Rakhlin, A.: Adaptive online gradient descent. In: Advances in Neural Information Processing Systems, vol. **20** (2007)
12. Bauschke, H.H., Borwein, J.M.: Legendre functions and the method of Bregman projections. J. Convex Anal. **4**(1), 27–67 (1997)
13. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, New York (2011)
14. Bauschke, H.H., Bolte, J., Teboulle, M.: A descent lemma beyond Lipschitz gradient continuity: first order methods revisited and applications. Math. Oper. Res. **42**(2), 330–348 (2016)
15. Beck, A.: First Order Methods in Optimization. SIAM, Philadelphia (2017)
16. Beck, A., Teboulle, M.: Mirror descent and nonlinear projected subgradient methods for convex optimization. Oper. Res. Lett. **31**, 167–175 (2003)
17. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)

18. Beck, A., Teboulle, M.: Gradient-based algorithms with applications to signal recovery problems. In: Palomar, D., Eldar, Y.C. (eds.) Convex Optimization in Signal Processing and Communications, pp. 139–162. Cambridge University Press, Cambridge (2009)
19. Beck, A., Teboulle, M.: Smoothing and first order methods: a unified framework. SIAM J. Optim. **22**, 557–580 (2012)
20. Ben-Tal, A., Margalit, T., Nemirovsky, A.: The ordered subsets mirror descent optimization method with applications to tomography. SIAM J. Optim. **12**, 79–108 (2001)
21. Bertsekas, D.P.: Constrained Optimization and Lagrange Multiplier Methods. Academic Press, Cambridge (1982)
22. Bertsekas, D.P.: Nonlinear Programming, 2nd edn. Athena Scientific, Belmont (1999)
23. Bertsekas, D.P.: Convex Optimization Algorithms. Athena Scientific, Belmont (2015)
24. Bolte, J., Sabach, S., Teboulle, M., Vaisbourd, Y.: First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. SIAM J. Optim. (2017) (**accepted**)
25. Bolte, J., Sabach, S., Teboulle, M.: Nonconvex Lagrangian-based optimization: monitoring schemes and global convergence. Math. Oper. Res. (2018). https://doi.org/10.1287/moor.2017.0900
26. Bolte, J., Teboulle, M.: Barrier operators and associated gradient like dynamical systems for constrained minimization problems. SIAM J. Control Optim. **42**, 1266–1292 (2003)
27. Bolte, J., Daniilidis, A., Lewis, A.S.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM J. Optim. **17**(4), 1205–1223 (2007)
28. Bolte, J., Daniilidis, A., Ley, O., Mazet, L.: Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. Trans. Am. Math. Soc. **362**, 3319–3363 (2010)
29. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. **146**(1), 459–494 (2014)
30. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
31. Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Math. Phys. **7**, 200–217 (1967)
32. Bruck, R.: On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space. J. Math. Anal. Appl. **61**, 159–164 (1977)
33. Burachik, R.S., Iusem, A.N.: A generalized proximal point algorithm for the variational inequality problem in a Hilbert space. SIAM J. Optim. **8**, 197–216 (1998)
34. Censor, Y., Zenios, S.A.: Proximal minimization algorithm with D-functions. J. Optim. Theory Appl. **73**, 451–464 (1992)
35. Chen, G., Teboulle, M.: Convergence analysis of a proximal-like minimization algorithm using Bregman functions. SIAM J. Optim. **3**, 538–543 (1993)
36. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward–backward splitting. SIAM Multiscale Model. Simul. **4**, 1168–1200 (2005)
37. Csiszár, I.: Information-type measures of difference of probability distributions and indirect observations. Studia Sci. Mat. Hungar. **2**, 299–318 (1967)
38. Drusvyatskiy, D., Lewis A.S.: Error bounds, quadratic growth, and linear convergence of proximal methods. Math. Oper. Res. (2018). https://doi.org/10.1287/moor.2017.0889
39. Duchi, J.C., Shalev-Shwartz, S., Singer, Y., Tewari, A.: Composite objective mirror descent. In: Proceedings of 23rd Annual Conference on Learning Theory, pp. 14–26. (2010)
40. Eckstein, J.: Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming. Math. Oper. Res. **18**, 202–226 (1993)
41. Flammarion, N., Bach, F.: Stochastic composite least-squares regression with convergence rate $O(1/n)$. Proc. Mach. Learn. Res. **65**, 1–44 (2017)
42. Fukushima, M., Mine, H.: A generalized proximal point algorithm for certain nonconvex minimization problems. Int. J. Syst. Sci. **12**, 989–1000 (1981)
43. Güler, O.: On the convergence of the proximal point algorithm for convex minimization. SIAM J. Control Optim. **29**(2), 403–419 (1991)
44. Kurdyka, K.: On gradients of functions definable in o-minimal structures. Ann. Inst. Fourier **48**(3), 769–783 (1998)
45. Lewis, A.S., Wright, S.J.: A proximal method for composite minimization. Math. Program. Ser. A **158**, 501–546 (2016)

46. Łojasiewicz, S.: Une propriété topologique des sous-ensembles analytiques réels. In: Les Équations aux Derivées Partielles, pp. 87–89. Éditions du Centre National de la Recherche Scientifique, Paris (1963)
47. Martinet, B.: Régularisation d'inéquations variationnelles par approximations successives. Rev. Française Informatique et Recherche Opérationnelle **4**, 154–158 (1970)
48. Moreau, J.-J.: Proximité et dualité dans un espace hilbertien. Bull. Soc. Math. Fr. **93**(2), 273–299 (1965)
49. Nemirovsky, A.S.: 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**, 229–251 (2004)
50. Nemirovsky, A.S., Yudin, D.B.: Problem Complexity and Method Efficiency in Optimization. Wiley, New York (1983)
51. Nesterov, Y.: A method for solving the convex programming problem with convergence rate $O(1/k^2)$. Dokl. Akad. Nauk SSSR **269**(3), 543–547 (1983)
52. Nguyen, Q.V.: Forward–backward splitting with Bregman distances. Vietnam J. Math. **45**, 519–539 (2017)
53. Opial, Z.: Weak convergence of the sequence of successive approximations for nonexpansive mappings. Bull. AMS **73**, 591–597 (1967)
54. Palomar, D.P., Eldar, Y.C.: Convex Optimization in Signal Processing and Communications. Cambridge University Press, Cambridge (2010)
55. Passty, G.B.: Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. J. Math. Anal. Appl. **72**, 383–390 (1979)
56. Polyak, R., Teboulle, M.: Nonlinear rescaling and proximal-like methods in convex optimization. Math. Program. **76**, 265–284 (1997)
57. Rockafellar, R.T., Wets, R.: Variational analysis. In: Grundlehren der Mathematischen Wissenschaften, vol. 317. Springer (1998)
58. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
59. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control Optim. **14**(5), 877–898 (1976)
60. Sabach, S., Teboulle, M., Vaisbourd, Y.: Fast non-Euclidean first order algorithms: a numerical study. In: Working Paper. (April 2017)
61. Shefi, R., Teboulle, M.: Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. SIAM J. Optim. **24**, 269–297 (2014)
62. Sra, S., Nowozin, S., Wright, S.J.: Optimization for Machine Learning. The MIT Press, Cambridge (2011)
63. Teboulle, M.: Entropic proximal mappings with application to nonlinear programming. Math. Oper. Res. **17**, 670–690 (1992)
64. Teboulle, M.: Convergence of proximal-like algorithms. SIAM J. Optim. **7**, 1069–1083 (1997)
65. Tseng, P.: Approximation accuracy, gradient methods, and error bound for structured convex optimization. Math. Program. Ser. B **125**, 263–295 (2010)
66. Xu, Y., Yin, W.: A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. SIAM J. Imaging Sci. **6**(3), 1758–1789 (2013)