MATHEMATICS OF OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0364-765X | EISSN 1526-5471 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000
© 0000 INFORMS

A descent Lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications

Heinz H. Bauschke
Mathematics, University of British Columbia, Kelowna, B.C. V1V 1V7, Canada, heinz.bauschke@ubc.ca

Jérôme Bolte
Toulouse School of Economics, Université Toulouse Capitole, Manufacture des Tabacs, 21 allée de Brienne, 31015 Toulouse, France, jerome.bolte@ut-capitole.fr

Marc Teboulle
School of Mathematical Sciences, Tel Aviv University, Ramat Aviv 69978, Israel, teboulle@post.tau.ac.il

The proximal gradient and its variants is one of the most attractive first-order algorithm for minimizing the sum of two convex functions, with one being nonsmooth. However, it requires the differentiable part of the objective to have a Lipschitz continuous gradient, thus precluding its use in many applications. In this paper we introduce a framework which allows to circumvent the intricate question of Lipschitz continuity of gradients by using an elegant and easy to check convexity condition which captures the geometry of the constraints. This condition translates into a new descent Lemma which in turn leads to a natural derivation of the proximal-gradient scheme with Bregman distances. We then identify a new notion of asymmetry measure for Bregman distances, which is central in determining the relevant step-size. These novelties allow to prove a global sublinear rate of convergence, and as a by-product, global pointwise convergence is obtained. This provides a new path to a broad spectrum of problems arising in key applications which were, until now, considered as out of reach via proximal gradient methods. We illustrate this potential by showing how our results can be applied to build new and simple schemes for Poisson inverse problems.

Key words: first-order methods, composite nonsmooth convex minimization, descent lemma, proximal-gradient algorithms, complexity, Bregman distance, multiplicative Poisson linear inverse problems
MSC2000 subject classification: 90C25, 65K05
OR/MS subject classification: Convex Programming/Algorithms

1. Introduction First-order methods have occupied the forefront of research in continuous optimization for more than a decade. This is due to their wide applicability in a huge spectrum of fundamental and disparate applications such as signal processing, image sciences, machine learning, communication systems, and astronomy to mention just a few, but also to their computational simplicity which makes them ideal methods for solving big data problems within medium accuracy levels. Recent research activities in this field are still conducted at a furious path in all the aforementioned applications (and much more), as testified by the large volume of literature; see e.g., [29, 34] and references therein for an appetizer.

A fundamental generic optimization model that encompasses various classes of smooth/nonsmooth convex models arising in the alluded applications is the well known composite minimization problem which consists in minimizing the sum of a possibly nonsmooth extended valued function with a differentiable one over a real Euclidean space X (see more precise description in §2):

$$(\mathcal{P}) \qquad \inf\{f(x) + g(x) : x \in X\}.$$

Despite its striking simplicity, this model is very rich and has led to the development of fundamental and well known algorithms. A *mother* scheme is the so-called forward-backward splitting method, which goes back at least to Passty [30] and Bruck [12] and which was developed in the more general setting of maximal monotone operators. When specialized to the convex problem $(\mathcal{P})$, this method is often called the *proximal gradient method* (PGM), a terminology we adopt in this article. One of the earliest work describing and analyzing the PGM includes for example the work of Fukushima and Milne [21]. The more recent work by Combettes and Wajs [17] provides important foundational insights and has popularized the method for a wide audience. More recently, the introduction of fast versions of the PGM such as FISTA by Beck-Teboulle [5] – which extends the seminal and fundamental work on the optimal gradient methods of Nesterov [27]– has resulted in a burst of research activities.

A central property required in the analysis of gradient methods, like the PGM, is that of the Lipschitz continuity of the gradient of the smooth part. Such a property implies (for a convex function is equivalent to) the so-called descent Lemma, e.g., [9], which provides a quadratic upper approximation to the smooth part. This simple process is at the root of the proximal gradient method, as well as many other methods. However, in many applications the differentiable function does not have such a property, e.g., in the broad class of Poisson inverse problems, (see e.g. the recent review paper [8] which also includes over 130 references), thus precluding therefore the use of the PGM methodology. When both f and g have an easily computable proximal operator, one could also consider tackling the composite model $(\mathcal{P})$ by applying the alternating direction of multipliers ADM scheme [23]. For many problems, these schemes are known to be quite efficient. However, note that even in simple cases, one faces several serious difficulties that we now briefly recall. First, being a primal-dual splitting method, the ADM scheme may considerably increase the dimension of the problem (by the introduction of auxiliary splitting variables). Secondly, the method depends on one (or more) unknown penalty parameter that needs to be heuristically chosen. Finally, to our knowledge, the convergence rate results of ADM based schemes are weaker, holding only for primal-dual gap in terms of ergodic sequences, see [14, 25, 33] and references therein. Moreover, the complexity bound constant not only depends on the unknown penalty parameter, but also on the norm of the matrix defining the splitting, which in many applications can be huge.

The main goal of this paper is to rectify this situation. We introduce a framework which allows to derive a class of proximal gradient based algorithms which are proven to share most of the convergence properties and complexity of the classical proximal-gradient, yet where the usual restrictive condition of Lipschitz continuity of the gradient of the differentiable part of problem $(\mathcal{P})$ is not required. It is instead traded with a more general and flexible convexity condition which involves the problem's data and can be specified by the user for each given problem. This is a new path to a broad spectrum of optimization models arising in key applications which were not accessible before. Surprisingly, the derivation and the development of our results starts from a very simple fact (which appears to have been overlooked) which underlines that the main ingredient in the success of PGM is to have an appropriate descent Lemma, i.e., an adequate upper approximation of the objective function.

Contribution and Outline The methodology underlying our approach and leading to a proximal-based algorithm freed from Lipschitz gradient continuity is developed in Section 2. A key player is a new simple, yet useful descent Lemma which allows to trade Lipschitz continuity of the gradient with an elementary convexity property. We further clarify these results by deriving several properties and examples and highlighting the key differences with the traditional proximal gradient method. In particular, an important notion of asymmetry coefficient is introduced and shown to play a central role in determining the relevant step size of the proposed scheme. The method is presented in Section 3 and its analysis is developed in Section 4, where a sublinear

$O(1/k)$ rate of convergence is established without the traditional Lipschitz gradient continuity of the smooth function. As a by-product, pointwise convergence of the method is also established. To demonstrate the benefits and potential of our new approach, we illustrate in Section 5 how it can be successfully applied to a broad class of Poisson linear inverse problems, leading to new proximal-based algorithms for these problems.

Notation Throughout the paper, the notation we employ is standard and as in [32] or [4]. We recall that for any set C, $i_C(\cdot)$ stands for the usual indicator function, which is equal to 0 if $x\in C$ and ∞ otherwise, and $\overline{C}$ denotes the closure of C. We set $\mathbb{R}_{++}=(0,+\infty)$.

2. A New Look at The Proximal Gradient Method

We start by recalling the basic elements underlying the proximal gradient method and its analysis which motivates the forthcoming developments.

Let $X=\mathbb{R}^d$ be a real Euclidean space with inner product $\langle\cdot,\cdot\rangle$ and induced norm $\|\cdot\|$. Given a closed convex set C with nonempty interior consider the convex problem

$$\inf\{\Phi(x):=f(x)+g(x): x\in C\}$$

where f,g are proper, convex and lower semicontinuous (lsc), with g continuously differentiable on $\operatorname{int}\operatorname{dom} g\neq\emptyset$, (see later on below for a precise description).

First consider the case when $C=\mathbb{R}^d$. For any fixed given point $x\in X$ and any $\lambda>0$, the main step of the proximal gradient method consists in minimizing an *upper approximation* of the objective obtained by summing a quadratic majorant of the differentiable part g and f, leaving thus untouched the nonsmooth part f of Φ:

$$x^+=\arg\min\{g(x)+\langle\nabla g(x),u-x\rangle+\frac{1}{2\lambda}\|u-x\|^2+f(u):\ u\in\mathbb{R}^d\}.$$

This is the proximal gradient algorithm, see e.g. [6]. Clearly, the minimizer x^+ exists and is unique, and ignoring the constant terms in x reduces to

$$x^+=\arg\min_u\{f(u)+\frac{1}{2\lambda}\|u-(x-\lambda\nabla g(x))\|^2\}\equiv\operatorname{prox}_{\lambda f}\left(x-\lambda\nabla g(x)\right), \tag{1}$$

where $\operatorname{prox}_\varphi(\cdot)$ stands for the so called Moreau's proximal map [26] of a proper lsc convex function φ. Thus, the PG scheme consists of the composition of a proximal (implicit/backward) step on f with a gradient (explicit/forward) step of g.[1]

A key assumption needed in the very construction and in the analysis of PG scheme is that g admits a Lipschitz continuous gradient L_g. As a simple consequence of this assumption (for a convex function g, this is an equivalence), we obtain the so-called descent Lemma, see e.g., [9] , namely for any $L\geq L_g$,

$$g(x)\leq g(y)+\langle x-y,\nabla g(y)\rangle+\frac{L}{2}\|x-y\|^2,\ \forall x,y\in\mathbb{R}^d. \tag{2}$$

This inequality not only naturally provides a upper quadratic approximation of g, but is also a crucial pillar in the analysis of *any* PG based method.

This leads us to the following simple observation:

[1] This also can be seen by convex calculus which gives $0\in\lambda(\partial f(x^+)+\nabla g(x)+x^+-x)$, which is equivalent to $x^+=(\mathrm{Id}+\lambda\partial f)^{-1}\circ(\mathrm{Id}-\lambda\nabla g)(x)\equiv\operatorname{prox}_{\lambda f}\left(x-\lambda\nabla g(x)\right)$.

Main Observation Developing the squared norm in (2), simple algebra shows that it can be equivalently written as:

$$\left(\frac{L}{2}\|x\|^2 - g(x)\right) - \left(\frac{L}{2}\|y\|^2 - g(y)\right) \geq \langle Ly - \nabla g(y), x - y\rangle \quad \forall x, y \in \mathbb{R}^d,$$

which in turn is nothing else but the gradient inequality for the convex function $\frac{L}{2}\|x\|^2 - g(x)$. Thus, for a given smooth convex function g on $\mathbb{R}^d$, the descent Lemma is equivalent to say that $\frac{L}{2}\|x\|^2 - g(x)$ is convex on $\mathbb{R}^d$.

This elementary and known fact, (see, e.g., [4, Theorem 18.15(vi)]) seems to have been overlooked. It naturally suggests to consider, instead of the squared norm used for the unconstrained case $C = \mathbb{R}^d$, a more general convex function that captures the geometry of the constraint C. This provides the motivation for the forthcoming proximal gradient based algorithm and its analysis for the constrained composite problem $(\mathcal{P})$.

2.1. The Constrained Composite Problem

Our strategy to handle the constraint set C is standard: a Legendre function on C is chosen and its associated Bregman distance is used as a proximity measure. Let us first recall the definition of a Legendre function.

Definition 1 (Legendre functions). [32, Chapter 26] Let $h : X \to (-\infty, \infty]$ be a lsc proper convex function. It is called:
(i) essentially smooth, if h is differentiable on $\operatorname{int}\operatorname{dom} h$, with moreover $\|\nabla h(x^k)\| \to \infty$ for every sequence $\{x^k\}_{k\in\mathbb{N}} \subset \operatorname{int}\operatorname{dom} h$ converging to a boundary point of $\operatorname{dom} h$ as $k \to +\infty$;
(ii) of Legendre type if h is essentially smooth and strictly convex on $\operatorname{int}\operatorname{dom} h$.

Also, let us recall the useful fact that h is of Legendre type if and only if its conjugate h^* is of Legendre type. Moreover, 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 ([32, Thm 26.5], that is we have,

$$(\nabla h)^{-1} = \nabla h^* \quad \text{and} \quad h^*(\nabla h(x)) = \langle x, \nabla h(x)\rangle - h(x). \tag{3}$$

Recall also that

$$\operatorname{dom}\partial h = \operatorname{int}\operatorname{dom}\, h \text{ with } \partial h(x) = \{\nabla h(x)\},\ \forall x \in \operatorname{int}\operatorname{dom}\, h. \tag{4}$$

The Problem and Blanket Assumptions Our aim is thus to solve

$$v(\mathcal{P}) = \inf\{\Phi(x) := f(x) + g(x) \mid x \in \overline{\operatorname{dom}}\, h\},$$

where $\overline{\operatorname{dom}}\, h = C$ denotes the closure of $\operatorname{dom} h$.

The following assumptions on the problem's data are made throughout the paper (and referred to as the blanket assumptions).

Assumption A
(i) $f : X \to (-\infty, \infty]$ is proper lower semicontinuous (lsc) convex,
(ii) $h : X \to (-\infty, \infty]$ is of Legendre type,
(iii) $g : X \to (-\infty, \infty]$ is proper lsc convex with $\operatorname{dom} g \supset \operatorname{dom} h$, which is differentiable on $\operatorname{int}\operatorname{dom} h$,
(iv) $\operatorname{dom} f \cap \operatorname{int}\operatorname{dom} h \neq \emptyset$,
(v) $-\infty < v(\mathcal{P}) = \inf\{\Phi(x) : x \in \overline{\operatorname{dom}}\, h\} = \inf\{\Phi(x) : x \in \operatorname{dom} h\}$.

Note that the second equality in (v) follows e.g. from [4, Proposition 11.1(iv)] and (iv) because $\operatorname{dom}(f + g) \cap \operatorname{int}\operatorname{dom}\, h = \operatorname{dom} f \cap \operatorname{int}\operatorname{dom}\, h \neq \varnothing$.

2.2. A New Descent Lemma Beyond Lipschitz Continuity Following our basic observation on the classical proximal gradient, we introduce a condition on the couple (g,h) which replaces the usual Lipschitz continuity property required on the gradient of g, by a convexity assumption capturing in a very simple manner the geometry of the constraints.

A Lipschitz-like/Convexity Condition

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

where we recall that $h: X\rightarrow(-\infty,\infty]$ is a Legendre function, and $g: X\rightarrow(-\infty,\infty]$ is a convex function with $\operatorname{dom} g\supset\operatorname{dom} h$, and g is continuously differentiable on $\operatorname{int}\operatorname{dom} h$.

Note that if $L'\geq L$ the same property holds with L'.

Clearly, as seen above, when $h(x)=\frac{1}{2}\|x\|^2$ one recovers the property that ∇g is Lipschitz continuous with constant L.

We shall see soon that the mere translation of condition (LC) into its first-order characterization immediately yields the new descent Lemma we seek for (see Lemma 1 below).

Beforehand, we introduce the fundamental proximity measure associated to any given Legendre function h; it is called the *Bregman distance* [11]

$$D_h(x,y)=h(x)-h(y)-\langle\nabla h(y),x-y\rangle,\ \forall x\in\operatorname{dom} h, y\in\operatorname{int}\operatorname{dom} h. \tag{5}$$

The use of Bregman distances in optimization within various contexts is well spread and cannot be reviewed here. For initial works on proximal Bregman based methods we refer the reader to [13, 35, 15, 19]. Many interesting results connecting for example Bregman proximal distance with dynamical systems can be found in [10] and references therein, and much more properties and applications can be found in the fundamental and comprehensive work [2].

Clearly, D_h is strictly convex with respect to its first argument. Moreover, $D_h(x,y)\geq 0$ for all $(x,y)\in\operatorname{dom} h\times\operatorname{int}\operatorname{dom} h$, and it is equal to zero if and only if $x=y$. Hence D_h provides, as announced, a natural proximity measure between points in the domain of h. Observe however that D_h is in general asymmetric, (see more below in §2.3).

We are ready to establish, the simple but key extended descent lemma.

Lemma 1 **(Descent lemma without Lipschitz Gradient Continuity)**. *Let $h: X\rightarrow(-\infty,\infty]$ be a Legendre function, and let $g: X\rightarrow(-\infty,\infty]$ be a convex function with* $\operatorname{dom} g\supset\operatorname{dom} h$ *which is continuously differentiable on* $\operatorname{int}\operatorname{dom} h$. *Then, the condition* (LC) *for the pair of functions* (h,g) *is equivalent to*

$$\big(\forall (x,y)\in\operatorname{int}\operatorname{dom}\, h\times\operatorname{int}\operatorname{dom}\, h\big)\quad g(x)\leq g(y)+\langle\nabla g(y),x-y\rangle+LD_h(x,y). \tag{6}$$

Proof. For any $y\in\operatorname{int}\operatorname{dom} h$, the function $Lh-g$ is convex on $\operatorname{int}\operatorname{dom}\, h$ if and only if the gradient inequality holds, i.e.,

$$\big(\forall x\in\operatorname{int}\operatorname{dom}\, h\big)\quad (Lh(x)-g(x))-(Lh(y)-g(y))\geq\langle L\nabla h(y)-\nabla g(y),x-y\rangle.$$

Rearranging the above inequality gives

$$\begin{aligned} g(x) &\leq g(y)+\langle\nabla g(y),x-y\rangle+L\left(h(x)-h(y)-\langle\nabla h(y),x-y\rangle\right)\\ &= g(y)+\langle\nabla g(y),x-y\rangle+LD_h(x,y), \end{aligned}$$

where the equality follows from the definition of D_h given in (5). ■

It is easy to see that the condition (LC) admits various alternative reformulations which can facilitate its checking, and which we conveniently collect in the following.

Proposition 1. *Consider the pair of functions (g,h) and assume that the above regularity conditions on h and g holds. Take $L>0$. The following statements are equivalent*
(i) *$Lh-g$ is convex on* $\operatorname{int}\operatorname{dom} h$, *i.e.* (LC) *holds,*
(ii) *$D_g(x,y)\le LD_h(x,y)$ for all $x,y\in\operatorname{int}\operatorname{dom} h$,*
(iii) *$D_{Lh-g}\ge 0$ on* $\operatorname{int}\operatorname{dom} h$,
(iv) *$\langle\nabla g(x)-\nabla g(y),x-y\rangle\le L\left(D_h(x,y)+D_h(y,x)\right)$, for all $x,y\in\operatorname{int}\operatorname{dom} h$.*
Moreover, when both g,h are assumed C^2 on the interior of their domain, then the above are equivalent to

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

Proof. The proof easily follows from the definition of the Bregman distance and the usual convexity properties. Indeed, first note that elementary algebra yields $LD_h-D_g=D_{Lh-g}$, which, in view of Lemma 1, immediately proves the equivalence between items (i), (ii) and (iii). Likewise, the convexity of $Lh-g$ is equivalent to the monotonicity of its gradient, which together with the definition of D_h yields

$$\langle\nabla g(x)-\nabla g(y),x-y\rangle\le L\langle\nabla h(x)-\nabla h(y),x-y\rangle=L\left(D_h(x,y)+D_h(y,x)\right),$$

proving the equivalence of (LC)-(iv). Finally, with g and h being C^2, we have the equivalence (LC)-(iv). ■

Remark 1. Note that if we assume,

$$\text{(D-Lip)}\quad \|\nabla g(x)-\nabla g(y)\|\le L\,\frac{D_h(x,y)+D_h(y,x)}{\|x-y\|},\ \text{for all}\ x\ne y\in\operatorname{int}\operatorname{dom} h,$$

then by Cauchy-Schwarz inequality we immediately obtain

$$\langle\nabla g(x)-\nabla g(y),x-y\rangle\le L\left(D_h(x,y)+D_h(y,x)\right),\ \text{for all}\ x\ne y\in\operatorname{int}\operatorname{dom} h,$$

which shows that (iv) holds, and hence so does (i), i.e., (LC). Thus, (D-Lip) provides a sort of Lipschitz-like gradient property of g with respect to D_h. Clearly, when $h=\frac{1}{2}\|\cdot\|^2$, (D-Lip) reduces to the Lipschitz gradient continuity of g with constant L.

One may construct various examples where (LC) holds. For simplicity, we focus on the case when $X=\mathbb{R}$, which can be utilized to obtain higher-dimensional Bregman distances with separable structure, i.e., with $h(x)=\sum_{j=1}^{d}h_j(x_j)$ and h_j defined on $\mathbb{R}$ (with possibly real-extended values), and which are the most fundamental class (in most cases, all the h_i are also identical). A more involved and interesting example is given in Section 5.

Example 1. We first list some of the most popular choices for h which are well documented in the literature, see e.g., [11, 35, 19, 2], and where more examples are given. Each example is a one dimensional h which is Legendre. To obtain the corresponding Legendre function $\tilde{h}$ and Bregman distance in $\mathbb{R}^d$ simply use the formulae $\tilde{h}(x)=\sum_{j=1}^{n}h(x_j)$ and $D_{\tilde{h}}(x,y)=\sum_{j=1}^{n}D_h(x_j,y_j)$.

- **Energy** $h(x)=\frac{1}{2}x^2, \operatorname{dom} h=\mathbb{R}$ and coincides with h^*.
- **Boltzmann-Shannon entropy** $h(x)=x\log x$, $\operatorname{dom} h=[0,\infty],(0\log 0=0)$. Then, $h^*(y)=\exp y-1$ with $\operatorname{dom} h^*=\mathbb{R}$.
- **Burg's entropy** $h(x)=-\log x$, $\operatorname{dom} h=(0,\infty)$. Then, $h^*(y)=-\log(-y)-1$, with $\operatorname{dom} h^*=(-\infty,0)$.
- **Fermi-Dirac entropy** $h(x)=x\log x+(1-x)\log(1-x)$, $\operatorname{dom} h=[0,1]$. Then, $h^*(y)=\log(1+\exp y)$, with $\operatorname{dom} h^*=\mathbb{R}$.
- **Hellinger** $h(x)=-\sqrt{1-x^2}$, $\operatorname{dom} h=[-1,1]$. Then, $h^*(y)=\sqrt{1+y^2}$, with $\operatorname{dom} h^*=\mathbb{R}$.
- **Fractional Power** $h(x)=(px-x^p)/(1-p), p\in(0,1)$, $\operatorname{dom} h=[0,\infty)$, [35]. Then, $h^*(y)=(1+y/q)^q$, with $\operatorname{dom} h^*=(-\infty,-q]$, and where $p+q=pq$.

For all these examples, h is twice differentiable with $h'' > 0$ on $\operatorname{int}\operatorname{dom} h$. Thus, (LC) is equivalent to

$$\sup_{x\in\operatorname{int}\operatorname{dom}\, h} \frac{g''(x)}{h''(x)} < +\infty. \tag{7}$$

Let us give two examples where g is C^2 and *does not* have a classical Lipschitz continuous gradient, yet where (7) holds.

- Let h be the Fermi-Dirac entropy. Then, (7) turns into $\sup_{0<x<1} x(1-x)g''(x) < +\infty$, which clearly holds when $[0,1] \subseteq \operatorname{int}\operatorname{dom} g$. For instance, this holds with $g(x) = x\log x$ which *does not* have a Lipschitz gradient.
- Let h be the Burg's entropy, and $g(x) = -\log x$ which *does not* have a Lipschitz gradient. Then, (7) trivially holds.

2.3. A Symmetry Measure for D_h It is well known that Bregman distances are in general not symmetric, except when h is the energy function because then $D_h(x,y) = \frac{1}{2}\|x-y\|^2$. In fact as pointed out by Iusem, (see [3]), D_h is symmetric if and only if h is a nondegenerate convex quadratic form (with $\operatorname{dom} h = \mathbb{R}^d$).[2] It is thus natural to introduce a measure for the lack of symmetry in D_h.

Definition 2 (Symmetry coefficient). Given a Legendre function $h : X \to (-\infty,\infty]$, its symmetry coefficient is defined by

$$\alpha(h) := \inf\left\{ D_h(x,y)/D_h(y,x) \;\middle|\; (x,y) \in \operatorname{int}\operatorname{dom}\, h \times \operatorname{int}\operatorname{dom}\, h,\ x \neq y \right\} \in [0,1]. \tag{8}$$

Remark 2. (a) The fact that $\alpha(h) \le 1$ follows from this simple observation: given any distinct points x, y in $\operatorname{int}\operatorname{dom} h$, either $D(x,y)/D(y,x)$ or $D(y,x)/D(x,y)$ is less than or equal to 1.
(b) Note that by definition of D_h, and with h Legendre, using (3) we immediately obtain $D_h(x,y) = D_{h^*}(\nabla h(y), \nabla h(x))$ for all $(x,y) \in \left(\operatorname{int}\operatorname{dom} h\right)^2$ and hence it follows that

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

(c) As we shall see in the next section the symmetry coefficient happens to play a fundamental role as a safeguard for descent properties of the proposed proximal gradient based method. The biggest stepsize that can be chosen in our method is indeed strictly upper bounded by the quantity

$$\frac{1+\alpha(h)}{L} \tag{9}$$

where $L > 0$ stands in place of the usual Lipschitz constant of ∇g from (LC).

By definition

$$(\forall x \in \operatorname{int}\operatorname{dom}\, h)(\forall y \in \operatorname{int}\operatorname{dom}\, h) \quad \alpha(h) D_h(x,y) \le D_h(y,x) \le \alpha(h)^{-1} D_h(x,y), \tag{10}$$

where we have adopted the convention that $0^{-1} = +\infty$ and $+\infty \times r = +\infty$ for all $r \ge 0$.

Clearly, the closer is $\alpha(h)$ to 1 the more symmetric D_h is, with perfect symmetry when $\alpha(h) = 1$ i.e., when h is strictly convex and quadratic.

A total lack of symmetry may occur for functions that do not have full domain. For the two key examples $h(x) = x\log x$ and $h(x) = -\log x$ namely the Boltzmann-Shannon and Burg entropy kernels which often arise in applications one can indeed verify that $\alpha(h) = 0$.

In fact, for the first example this is a consequence of the following result.

[2] Explicitly: $\frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle + c$, $\forall x \in X$, where A is positive definite, $b \in X$, and $c \in \mathbb{R}$.

PROPOSITION 2 **(Absence of symmetry)**. *Suppose that $h : X \to (-\infty, \infty]$ is a Legendre function and that $\operatorname{dom} h$ is not open. Then $\alpha(h) = 0$.*

Proof. Fix $x \in \operatorname{int}\operatorname{dom} h$ and let $z \in \operatorname{dom} h \setminus \operatorname{int}\operatorname{dom} h$. Let $\{\varepsilon_k\}_{k\in\mathbb{N}}$ be in $(0,1)$ such that $\varepsilon_k \to 0$ and set for all integer k, $x^k = (1-\varepsilon_k)z + \varepsilon_k x$. Then $x^k \to z$ and $\{x^k\}_{k\in\mathbb{N}}$ lies in $\operatorname{int}\operatorname{dom} h$. On the one hand, [2, Theorem 3.7] implies that $D(x^k, x) \to D(z, x) \in \mathbb{R}_{++}$. On the other hand, [2, Theorem 3.8.(i)] yields $D(x, x^k) \to +\infty$. Altogether,

$$\frac{D(x^k, x)}{D(x, x^k)} \to 0,$$

which implies $\alpha(h) = 0$. ■

Hence if $\alpha(h) > 0$, then $\operatorname{dom} h$ is open, (and the same holds true for $\operatorname{dom} h^*$). This necessary condition is however not sufficient. Indeed, for the second example with h being the Burg's entropy, with open domain $(0, \infty)$, one computes

$$\frac{D(1,x)}{D(x,1)} = \frac{x\log x - x + 1}{x^2 - x - x\log x} \to 0 \quad \text{as } x \to +\infty,$$

and we deduce that $\alpha(h) = 0$. On the positive side, with $h(x) = x^4$, using calculus, one can verify by a direct computation that $\alpha(h) = 2 - \sqrt{3} > 0$.

3. The Proximal Gradient Algorithm without a Lipschitz Continuous Gradient

Equipped with the generalized descent Lemma we can now develop the necessary tools to build and analyze the proximal gradient method without the usual Lipschitz gradient continuity assumption. For ease of reference the algorithm is called NoLips.

It should be noted that the proximal gradient method which uses Bregman distances is not by itself a new algorithm. Indeed, it was already investigated through various works/contexts, see for instance, [1, 7, 36], for more details and references therein.

The novel aspect of our approach resides in several facts:

– we circumvent the intricate question of Lipschitz continuity of gradients by using a simple, elegant and easy to check condition on the geometry of g. Besides, this condition can immediately be translated into a descent Lemma, a type of inequality which is at the heart of most works on the complexity of first-order methods;

– our notion of asymmetry for D_h is identified as a sharp measure of the step-sizes allowed in the proximal gradient method;

– when put together the above novelties allow for a transparent derivation and analysis of the (Bregman) proximal gradient method and they open many possibilities for problems which were, until now, considered as out of reach via proximal gradient methods.

Given a Legendre function h, for all $x \in \operatorname{int}\operatorname{dom} h$ and any step-size $\lambda > 0$, we define formally

$$T_\lambda(x) := \arg\min\left\{ f(u) + g(x) + \langle \nabla g(x), u - x\rangle + \frac{1}{\lambda} D_h(u,x) : u \in X \right\}. \tag{11}$$

Again, note that when h is the energy, the above boils down to the classical proximal gradient operator, the principle behind being called forward-backward splitting, [12, 30].

In the remainder, we shall systematically assume that the proximal gradient operator T_λ is well defined on $\operatorname{int}\operatorname{dom} h$, meaning:

$$\boxed{T_\lambda \text{ is nonempty, single-valued and maps } \operatorname{int}\operatorname{dom} h \text{ in } \operatorname{int}\operatorname{dom} h.}$$

Giving a universally relevant set of conditions for this well-posedness aspect is quite involved, technical and somehow useless. Indeed this does not impact any other compartment of the subsequent analysis and practical examples show that the method is generally well defined.

We provide however with two natural assumptions very often met in practice. We recall beforehand that for any given $F:\mathbb{R}^d \to (-\infty,\infty]$, we say that F is supercoercive if $\liminf_{\|x\|\to\infty} F(x)/\|x\| = \infty$, see e.g., [4, p.158] for further details.

Lemma 2 **(Well-posedness of the method).** *Under the assumption* **A**, *if one of the following assumptions holds:*
(i) $\arg\min\left\{\Phi(x) : x \in \overline{\operatorname{dom} h}\right\}$ *is compact*
(ii) $(\forall\lambda > 0)\quad h+\lambda f$ *is supercoercive,*
then the map T_λ *defined in* (11) *is nonempty and single-valued from* $\operatorname{int}\operatorname{dom} h$ *to* $\operatorname{int}\operatorname{dom} h$.

Proof. Fix $x \in \operatorname{int}\operatorname{dom} h$. Since h is strictly convex the objective in (11) may have at most one minimizer. Assuming (i), we obtain that $\Phi + i_{\overline{\operatorname{dom} h}}$ is coercive, since Bregman distances are nonnegative, the objective within T_λ is also coercive and thus T_λ is nonempty. When assuming (ii), the argument follows by the supercoercivity properties of the same objective, see [4]. The property $T_\lambda(x)$ is contained in $\operatorname{int}\operatorname{dom} h$ is a general fact. It can be seen through the optimality condition for $T_\lambda(x)$ which implies that $\partial h(T_\lambda(x))$ must be nonempty. This forces $T_\lambda(x)$ to belong to $\operatorname{int}\operatorname{dom} h$ by the Legendre property (4). ■

3.1. The NoLips Algorithm

We are now ready to describe our algorithm for solving

$$(\mathcal{P})\quad \inf\{\Phi(x) := f(x) + g(x) : x \in C\},$$

where C is closed convex set with nonempty interior.

NoLips Algorithm

0. **Input.** Choose a Legendre function h with $C = \overline{\operatorname{dom} h}$ such that there exists $L > 0$ with $Lh - g$ convex on $\operatorname{int}\operatorname{dom} h$.
1. **Initialization.** Start with any $x^0 \in \operatorname{int}\operatorname{dom} h$.
2. **Recursion.** For each $k = 1,\ldots$ with $\lambda_k > 0$, generate a sequence $\left\{x^k\right\}_{k\in\mathbb{N}}$ in $\operatorname{int}\operatorname{dom} h$ via

$$x^k = T_{\lambda_k}(x^{k-1}) = \operatorname*{arg\,min}_{x\in X}\left\{f(x) + \left\langle \nabla g(x^{k-1}), x - x^{k-1}\right\rangle + \frac{1}{\lambda_k} D_h(x, x^{k-1})\right\}. \tag{12}$$

Recalling our standing assumption on the non vacuity of T_λ the algorithm is well defined.

Splitting mechanism It is interesting to observe that T_λ shares the same structural decomposition principle as the usual proximal gradient. Under very mild assumptions the above recursion can indeed actually be split for computational purpose into “elementary” steps. Writing the optimality condition for $x^+ = T_\lambda(x)$, we obtain

$$0 \in \lambda\left(\partial f(x^+) + \nabla g(x)\right) + \nabla h(x^+) - \nabla h(x).$$

When $\nabla h(x) - \lambda\nabla g(x) \in \operatorname{dom}\nabla h^*$ one can define

$$p_\lambda(x) = \nabla h^*(\nabla h(x) - \lambda\nabla g(x)), \tag{13}$$

then recalling that $\nabla h \circ \nabla h^* = I$ (cf. (3)), the optimality condition reduces to

$$0 \in \lambda\partial f(x^+) + \nabla h(x^+) - \nabla h(p_\lambda(x)),$$

which is nothing else but the optimality condition which characterizes x^+ via the Bregman proximal step

$$x^+ = \arg\min\left\{ f(u) + \frac{1}{\lambda} D_h(u, p_\lambda(x)) : u \in X \right\}. \tag{14}$$

Note that the formula (13) which defines $p_\lambda(\cdot)$ is nothing but an *interior Bregman gradient step* (see [1] and references therein), i.e.,

$$p_\lambda(x) = \arg\min\left\{ \langle \nabla g(x), u \rangle + \frac{1}{\lambda} D_h(u, x) : u \in X \right\}, \tag{15}$$

which clearly reduces to the usual explicit gradient step when $h = \frac{1}{2}\|\cdot\|^2$. Finally observe that a formal introduction of the proximal Bregman operator, see e.g., [35]

$$\mathrm{Prox}^h_{\lambda f}(y) = \arg\min\left\{ \lambda f(u) + D_h(u, y) : u \in X \right\},\ y \in \mathrm{int}\,\mathrm{dom}\, h$$

allows to rewrite the NoLips algorithm simply as the composition of a Bregman proximal step with an interior Bregman gradient step:

$$x^k = \mathrm{Prox}^h_{\lambda f}(p_\lambda(x^{k-1})). \tag{16}$$

Remark 3. It should be carefully noted that the above decomposition necessitates that the subproblems (14) (the prox/implicit step) and (15) (the gradient/explicit step) are well defined. As previously pointed out, well-definedness is not in general an issue but formal guarantees depend strongly on the geometry of the problem (see, e.g., Lemma 2).

Computing the mappings $p_\lambda(\cdot)$, and $\mathrm{Prox}^h_{\lambda f}(\cdot)$.

As explained above, the algorithm NoLips requires to compute two specific objects:

- An (interior) Bregman projected-gradient step: $p_\lambda(\cdot)$ as defined in (15).
- A prox-like step based on a Bregman distance: $\mathrm{Prox}^h_{\lambda f}(p_\lambda(\cdot))$, as defined in (14).

Bregman-like gradients: computing $p_\lambda(\cdot)$. In the classical Euclidean proximal gradient method, the first computation above reduces to a standard gradient step, i.e., $p_\lambda(x) = x - \lambda \nabla g(x)$. We now give some examples for NoLips.

For convenience, for any given $\lambda > 0$ and for any $x \in \mathrm{int}\,\mathrm{dom}\, h$, define $v(x) := \nabla h(x) - \lambda \nabla g(x)$. Then, as shown in (13), we have

$$p_\lambda(x) = \nabla h^*(v(x)),\ \ v(x) \in \mathrm{dom}\, \nabla h^*. \tag{17}$$

Therefore, once we know h^*, the computation of p_λ is straightforward. This is the case for the six examples listed in Example 1. Below, we further illustrate with some relevant examples where it is easy to evaluate p_λ.

Example 2. (i) *Regularized Burg's Entropy.* Let $h(x) = \frac{\sigma}{2}x^2 - \mu \log x$ with $\mathrm{dom}\, h = (0, \infty)$, where $(\sigma, \mu > 0)$. Then one computes $h^*(s) = \frac{\sigma}{2}t^2(s) + \mu \log t(s) - \mu$, where

$$t(s) := \frac{s + \sqrt{s^2 + 4\mu\sigma}}{2\sigma} > 0, \text{ and } \mathrm{dom}\, h^* = \mathbb{R}.$$

Some algebra shows that $\nabla h^*(s) = (\sigma t^2(s) + \mu)(s^2 + 4\mu\sigma)^{-1/2}$, which yields the desired $p_\lambda(x)$ via (17).

(ii) *"Hellinger-Like function".* Let $h(x) = -\sqrt{1 - \|x\|^2}$; $\mathrm{dom}\, h = \{x \in \mathbb{R}^d : \|x\| \le 1\}$. Note that this yields a nonseparable Bregman distance which is relevant for ball constraints. We then obtain, $h^*(y) = \sqrt{1 + \|y\|^2}$, $\mathrm{dom}\, h^* = \mathbb{R}^n$, and hence $p_\lambda(x) = (1 + v^2(x))^{-1/2} v(x)$.

(iii) *Composition with a norm.* Let $\phi : \mathbb{R} \to (-\infty, \infty]$ be Legendre and monotone increasing. Suppose that ϕ is even, i.e., $\phi(x) = \phi(-x)$ for all $x \in \mathbb{R}$. Note that ϕ even implies that $\operatorname{dom}\phi \ni 0$, and by convexity, that 0 is a minimizer of ϕ on $\mathbb{R}$. Define $h(x) := \phi(\|x\|)$. Then,

$$h^*(y) = \sup_x\{\langle x, y\rangle - \phi(\|x\|)\} = \sup\{t\|y\| - \phi(t) : t \geq 0\} = \phi^*(\|y\|),$$

from which p_λ can be evaluated for a given ϕ. The above example (ii) corresponds to the choice $\phi(t) = -\sqrt{1-t^2}$, with $\operatorname{dom}\phi = [-1, 1]$. Another example is with the choice $\phi(t) = -\log(1-t^2)$ on $(-1, 1)$, which we leave to the reader.

(iv) *Semidefinite constraints.* Bregman distances can be defined on the space of $d \times d$ symmetric matrices S^d, and are useful to handle semidefinite constraints. Denote by S^d_{++} the cone of positive definite matrices of S^d. Let $h : S^d_{++} \to \mathbb{R}$ defined by $h(x) = -\log\det(x)$. Then, $\nabla h(x) = x^{-1}$ the inverse of x, and we obtain $p_\lambda(x) = v(x)^{-1}$, $v(x), x \in S^d_{++}$. For more examples on handling conic constraints with other Bregman distances on matrices, and evaluating p_λ, see e.g., [1, Examples B, C, p.718].

Proximal calculus with Bregman kernels: computing $\operatorname{Prox}^h_{\lambda f}(\cdot)$. The classical Moreau proximal map of f is in general explicitly computable when f is norm-like, or when f is the indicator of sets whose geometry is favorable to Euclidean projections. Although quite frequent in applications (orthant, second-order cone, ℓ^1 norm), these prox-friendly functions are very rare, see e.g., [17, Section 2.6]. Concerning NoLips the situation is exactly the same: for a given kernel h, sets and functions which are prox-friendly are scarce and are modeled on h. However a major advantage in our approach is that one can choose the kernel h to adapt to the geometry of the given function/set. This situation will be illustrated in Section 5, for a broad class of inverse problems involving Poisson noise.

Below, we give examples for which $\operatorname{Prox}^h_{\lambda f}(y) = \arg\min\{\lambda f(u) + D_h(u, y) : u \in X\}$, $y \in \operatorname{int}\operatorname{dom} h$ is easy to compute in closed form, whereas the standard Moreau proximal map is not explicitly known (and would thus require a numerical procedure, implying a nested scheme if used in an algorithm).

Example 3. (i) (Entropic thresholding) Let $f(u) = |u - a|$ where $a > 0$ and take $h(x) = x\log x$, $\operatorname{dom} h = [0, \infty)$. Then,

$$\operatorname{Prox}^h_{\lambda f}(y) = \begin{cases} \exp(\lambda) y & \text{if } y < \exp(-\lambda)a, \\ a & \text{if } y \in [\exp(-\lambda)a, \exp(\lambda)a], \\ \exp(-\lambda) y & \text{if } y > \exp(\lambda)a. \end{cases}$$

(ii) (Log thresholding) Let $f(u) = |u - a|$ where $a > 0$ and take $h(x) = -\log x$, $\operatorname{dom} h = (0, \infty)$. Assume $\lambda a < 1$. Then,

$$\operatorname{Prox}^h_{\lambda f}(y) = \begin{cases} \frac{y}{1+\lambda y} & \text{if } y < \frac{a}{1-\lambda a}, \\ a & \text{if } y \in \left[\frac{a}{1-\lambda a}, \frac{a}{1+\lambda a}\right], \\ \frac{y}{1-\lambda y} & \text{if } y > \frac{a}{1+\lambda a}. \end{cases}$$

Similar formulas may be derived when $\lambda a > 1$.

(iii) Let $f(u) = ce^u$, $c > 0$, and take $h(x) = e^x$, $\operatorname{dom} h = \mathbb{R}$. Then $\operatorname{Prox}^h_{\lambda f}(y) = y - \log(1 + \lambda c)$.

(iv) (Squared norms and Burg entropy) Let $f(u) = \frac{c}{2}u^2, c > 0$ and take $h(x) = -\log x$, $\operatorname{dom} h = (0, \infty)$. Then $\operatorname{Prox}^h_{\lambda f}(y) = (2c\lambda y)^{-1}\left(\sqrt{1 + 4c\lambda y^2} - 1\right)$.

Step-size choices. Back to the algorithm, it remains to determine the choice of the step-size in terms of the problem's data. Here the symmetry coefficient $\alpha(h)$ and the relative convexity constant

L play central roles. In the rest of the paper, and when no confusion occurs, we use the simpler notation α instead of $\alpha(h)$. We let $(\lambda_k)_{k\in\mathbb{N}}$ to be a sequence in $\mathbb{R}_{++}$ with

$$0<\lambda_k\le\frac{(1+\alpha)-\delta}{L}\quad\text{for some}\quad\delta\in(0,1+\alpha),\tag{18}$$

where $\alpha\in[0,1]$ is the symmetry coefficient of h as defined in (8). Observe that, when h is the energy function, then $\alpha(h)=1$ and the above boils down to $\lambda_k\le\frac{2-\delta}{L}$, the usual step size of the Euclidean proximal gradient method, see e.g., [22].

The next section addresses the two fundamental issues regarding NoLips

– *What is the complexity of the method?*

– *Can we assert that the sequence converges to a single minimizer?*

These issues are strongly related to the geometric features of h, but also to their adequation with the couple $(f+g,\overline{\operatorname{dom} h})$.

4. Analysis of the NoLips **algorithm: complexity and convergence** In this section, we establish the main convergence properties of the proposed algorithm. In particular, we prove its global rate of convergence, showing that it shares the claimed sublinear rate $O(1/k)$ of basic first-order methods such as the classical PG. We also derive a global convergence of the sequence generated by NoLips to a minimizer of $(\mathcal{P})$ under an additional mild assumption on h which holds for most practical choices. Our analysis builds on [24] initially developed for the classical quadratic proximal minimization method, and follows its extensions in [15] and [1].

Throughout this section, we remind the reader that, we work under the blanket assumption, the (LC) condition and that we assume the algorithm to be well defined.

We start with some elementary preliminaries. As usual with the analysis of Bregman based schemes, the following very simple three points identity for D_h is very useful.

Lemma 3 **(Three points identity).** [15] *Let $h:\mathbb{R}^d\to(-\infty,\infty]$ be a proper lsc convex function. For any $x\in\operatorname{dom} h$, and $y,z\in\operatorname{int}\operatorname{dom} h$ the following identity holds:*

$$D_h(x,z)-D_h(x,y)-D_h(y,z)=\langle\nabla h(y)-\nabla h(z),x-y\rangle\,.\tag{19}$$

It is also useful to record the following inequality which is a consequence of the new extended descent Lemma 1.

Lemma 4 **(Three points extended descent Lemma).** *Assume that* (LC) *holds for the pair of convex functions (h,g). Then, for any $(x,y,z)\in\operatorname{int}\operatorname{dom}\,h\times\operatorname{dom} h\times\operatorname{int}\operatorname{dom}\,h$, we have*

$$g\,(x)\le g\,(y)+\langle\nabla g\,(z)\,,x-y\rangle+LD_h(x,z).$$

Proof. Take x,y,z as specified. By Lemma 1 and since $x,z\in\operatorname{int}\operatorname{dom} h$ we have

$$g\,(x)\le g\,(z)+\langle\nabla g\,(z)\,,x-z\rangle+LD_h(x,z).$$

Since g is convex, differentiable and $\operatorname{dom} g\supset\operatorname{dom} h$, the gradient inequality yields

$$0\le g\,(y)-g\,(z)-\langle\nabla g\,(z)\,,y-z\rangle\,.$$

Adding these two inequalities gives the desired result. ■

The next lemma provides a key estimation inequality for the forthcoming analysis.

LEMMA 5 **(Descent inequality for** NoLips**).** *Let* $\lambda > 0$*. For all* x *in* $\operatorname{int}\operatorname{dom} h$*, let* $x^+ := T_\lambda(x)$*. Then,*

$$\lambda\left(\Phi(x^+) - \Phi(u)\right) \le D_h(u,x) - D_h(u,x^+) - (1-\lambda L)D_h(x^+,x),\ \forall u \in \operatorname{dom} h. \tag{20}$$

Proof. Fix any $x \in \operatorname{int}\operatorname{dom} h$. By assumption, $x^+ = T_\lambda(x) \in \operatorname{int}\operatorname{dom} h$ is unique and characterized by the optimality condition:

$$0 \in \lambda\left(\partial f(x^+) + \nabla g(x)\right) + \nabla h(x^+) - \nabla h(x).$$

Substituting the latter in the subgradient inequality for the convex function f, we then obtain for any $u \in \operatorname{dom} h$ (with $\lambda > 0$),

$$\begin{aligned}\lambda(f(x^+) - f(u)) &\le \lambda\langle \nabla g(x), u - x^+\rangle + \langle \nabla h(x^+) - \nabla h(x), u - x^+\rangle \\ &= \lambda\langle \nabla g(x), u - x^+\rangle + D_h(u,x) - D_h(u,x^+) - D_h(x^+,x),\end{aligned} \tag{21}$$

where the equality follows from the three points identity (cf. Lemma 3). On the other hand, since $(x^+, u, x) \in \operatorname{int}\operatorname{dom}\, h \times \operatorname{dom} h \times \operatorname{int}\operatorname{dom}\, h$, Lemma 4 applies:

$$\lambda(g(x^+) - g(u)) \le \lambda\langle \nabla g(x), x^+ - u\rangle + \lambda L D_h(x^+,x). \tag{22}$$

Adding the last inequality to (21), recalling that $\Phi(x) = f(x) + g(x)$, we thus obtain

$$\lambda\left(\Phi(x^+) - \Phi(u)\right) \le D_h(u,x) - D_h(u,x^+) - (1-\lambda L)D_h(x^+,x),$$

as announced. ■

We are now ready for our main result. Recall that $\lambda_k > 0$, $(k \in \mathbb{N})$, must satisfy

$$0 < \lambda_k \le \frac{(1+\alpha) - \delta}{L} \quad \text{for some} \quad \delta \in (0, 1+\alpha). \tag{23}$$

THEOREM 1. *Let* $\{x^k\}_{k\in\mathbb{N}}$ *be the sequence generated by* NoLips *and let* $\sigma_k = \sum_{l=1}^k \lambda_l$*. Then the following hold:*

(i) **(Monotonicity)** $\{\Phi(x^k)\}_{k\in\mathbb{N}}$ *is nonincreasing.*
(ii) **(Summability)** $\sum_{k=1}^\infty D_h(x^k, x^{k-1}) < \infty$.
(iii) **(Convergence of the function values)** *If* $\sigma_k \to \infty$*, then* $\lim_{n\to\infty} \Phi(x^k) = v(\mathcal{P})$.
(iv) **(Global estimate in function values)** *Now, let* $\lambda_k := \frac{1+\alpha}{2L}$ *for all positive integer* k*. Then,*

$$\Phi(x^k) - \Phi(u) \le \frac{2L}{(1+\alpha)k} D_h(u,x^0),\ \forall u \in \operatorname{dom} h.$$

Proof. Fix $k \ge 1$. Using Lemma 5 with $x^k = T_{\lambda_k}(x^{k-1})$, we obtain, for all $u \in \operatorname{dom} h$,

$$\lambda_k\left(\Phi(x^k) - \Phi(u)\right) \le D_h(u,x^{k-1}) - D_h(u,x^k) - (1-\lambda_k L)D_h(x^k,x^{k-1}). \tag{24}$$

Basically, all the claims of the Theorem easily follow from this inequality.

Items (i) and (ii): By definition of λ_k, we have $1 - \lambda_k L \ge \delta - \alpha$, and hence (24) reduces to

$$\lambda_k\left(\Phi(x^k) - \Phi(u)\right) \le D_h(u,x^{k-1}) - D_h(u,x^k) + (\alpha - \delta)D_h(x^k,x^{k-1}),\ \forall u \in \operatorname{dom} h. \tag{25}$$

Set $u = x^{k-1}$ in (25), using $D_h(x^{k-1}, x^{k-1}) = 0$, and recalling that by definition of α, (recall (10))

$$-D_h(x^{k-1}, x^k) + \alpha D_h(x^k, x^{k-1}) \le 0,$$

we thus deduce from (25) that

$$\begin{aligned}\lambda_k\left(\Phi(x^k)-\Phi(x^{k-1})\right) &\le -D_h(x^{k-1},x^k)+\alpha D_h(x^k,x^{k-1})-\delta D_h(x^k,x^{k-1})\\ &\le -\delta D_h(x^k,x^{k-1})\\ &\le 0.\end{aligned} \tag{26}$$

Therefore the sequence $\{\Phi(x^k)\}_{k\in\mathbb{N}}$ is nonincreasing, and since by Assumption **A**, $v(P)>-\infty$, we get

$$\lim_{k\to\infty}\Phi(x^k)\ge v(P)>-\infty. \tag{27}$$

Now, again using the condition on λ_k given in (23), and summing (26) over $k=1,\ldots,n$, we obtain

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

Hence with (27) this implies that $\sum_{k=1}^{\infty}D_h(x^k,x^{k-1})<\infty$ which is (ii).

(iii) Using $\sigma_k=\lambda_k+\sigma_{k-1}$ (with $\sigma_0=0$), multiplying (26) by σ_{k-1}, we obtain

$$\sigma_k\Phi(x^k)-\sigma_{k-1}\Phi(x^{k-1})-\lambda_k\Phi(x^k)\le-\delta\frac{\sigma_{k-1}}{\lambda_k}D_h(x^k,x^{k-1}).$$

Summing over $k=1,\ldots,n$ the above inequality, as well as inequality (25), we then get respectively

$$\begin{aligned}\sigma_n\Phi(x^n)-\sum_{k=1}^{n}\lambda_k\Phi(x^k)\;&\le\;-\delta\sum_{k=1}^{n}\frac{\sigma_{k-1}}{\lambda_k}D_h(x^k,x^{k-1}),\\ \sum_{k=1}^{n}\lambda_k\Phi(x^k)-\sigma_n\Phi(u)\;&\le\;D_h(u,x^0)-D_h(u,x^n)+\alpha\sum_{k=1}^{n}D_h(x^k,x^{k-1})-\delta\sum_{k=1}^{n}D_h(x^k,x^{k-1}).\end{aligned}$$

Adding these two inequalities, and recalling that $\delta>0$ and $D_h(\cdot,\cdot)\ge0$, it follows that

$$\Phi(x^n)-\Phi(u)\le\frac{D_h(u,x^0)}{\sigma_n}+\frac{\alpha}{\sigma_n}\sum_{k=1}^{n}D_h(x^k,x^{k-1}). \tag{28}$$

Therefore, passing to the limit with $\sigma_n\to\infty$, and recalling that $(\Phi(x^k))_{k\in\mathbb{N}}$ is decreasing, and $sum_{k=1}^{\infty}D_h(x^k,x^{k-1})<\infty$, we obtain $\lim_{n\to\infty}\Phi(x^n)\le\Phi(u)$ for every $u\in\operatorname{dom}h$, and hence together with (27) it follows that $\lim_{n\to\infty}\Phi(x^n)=v(\mathcal{P})$.

(iv) Now, let $\lambda_k=\frac{(1+\alpha)}{2L}$ for all $k\ge0$. Clearly it satisfies (23) with $\delta=(1+\alpha)/2$. Then, from (24) it follows that for all $u\in\operatorname{dom}h$,

$$\Phi(x^k)-\Phi(u)\le\frac{2L}{1+\alpha}\left\{D_h(u,x^{k-1})-D_h(u,x^k)\right\}-\frac{(1-\alpha)}{2}D_h(x^k,x^{k-1}).$$

Therefore, since $\alpha\in[0,1]$, $D_h(\cdot,\cdot)\ge0$, this reduces to

$$\Phi(x^k)-\Phi(u)\le\frac{2L}{1+\alpha}\left\{D_h(u,x^{k-1})-D_h(u,x^k)\right\},\ \forall k\ge1. \tag{29}$$

Define $v_k:=\Phi(x^k)-\Phi(u)$. Since the sequence $\{\Phi(x^k)\}_{k\in\mathbb{N}}$ is decreasing, we have $v_{k+1}\le v_k$, and hence it follows that $v_n\le\frac{1}{n}\sum_{k=1}^{n}v_k$. Therefore, for all $u\in\operatorname{dom}h$

$$\begin{aligned}\Phi(x^n)-\Phi(u)\;&\le\;\frac{1}{n}\sum_{k=1}^{n}[\Phi(x^k)-\Phi(u)]\\ &\le\;\frac{2L}{n(1+\alpha)}D_h(u,x^0),\end{aligned}$$

where the last inequality follows from (29) and the nonnegativity of the Bregman distance. ■

COROLLARY 1 **(Complexity for h with closed domain).** *We make the same assumption as in (iv) above but we assume in addition that* $\operatorname{dom} h = \overline{\operatorname{dom}}\, h$ *and that* $(\mathcal{P})$ *has at least a solution. Then for any solution* $\bar{x}$ *of* $(\mathcal{P})$,

$$\Phi(x^k) - \min_C \Phi \leq \frac{2LD_h(\bar{x}, x^0)}{(1+\alpha(h))\,k}.$$

Proof. It follows directly from (iv) above. ■

When $h(x) = \frac{1}{2}\|x\|^2$, the associated distance is symmetric (i.e. $\alpha = 1$), the number L is a Lipschitz continuity constant of the gradient of g, and we thus recover the classical global rate of the usual proximal gradient method [6]. In view of Corollary 1 above, note also that the entropies of Boltzmann-Shannon, Fermi-Dirac and Hellinger are non trivial examples for which the finiteness assumption ($\overline{\operatorname{dom}}\, h = \operatorname{dom} h$) is obviously satisfied.

We now deduce from all the above results, the global convergence to an optimal solution under the usual finiteness assumptions on h. For that purpose, what is needed are additional assumptions on the Bregman proximal distance ensuring separation properties of this distance at the boundary, so that we can use arguments “à la Opial” [28].

Recall that the blanket assumptions and in particular that $h : X \to (-\infty, \infty]$ is a Legendre function hold.

Assumptions **H**:

(i) For every $x \in \operatorname{dom} h$ and $\beta \in \mathbb{R}$, the level set $\{y \in \operatorname{int}\operatorname{dom} h :\ D_h(x,y) \leq \beta\}$ 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 $x^k \to x$.

REMARK 4. (a) All examples given previously, Boltzmann-Shannon, Fermi-Dirac, Hellinger Burg entropies satisfy the above set of assumptions.
(b) For much more general and accurate results on the interplay between Legendre functions and Bregman separation properties on the boundary we refer the reader to [2].

Before giving our convergence result, we recall the following well known result on nonnegative sequences which will be useful to us, see [31, Lemma 2, p.44].

LEMMA 6. *Let* $\{v_k\}_{k\in\mathbb{N}}$ *and* $\{\varepsilon_k\}_{k\in\mathbb{N}}$ *be nonnegative sequences. Assume that* $\sum_{k=1}^{\infty} \varepsilon_k < \infty$ *and that*

$$(\forall k \in \mathbb{N}),\ v_{k+1} \leq v_k + \varepsilon_k,\ \forall k \geq 0. \tag{30}$$

Then $\{v_k\}_{k\in\mathbb{N}}$ *converges.*

Proof. We include a brief proof here to make the paper self-contained. Set $\beta_k = v_k + \sum_{m=k}^{+\infty} \varepsilon_m$ for all k, this makes $\{\beta_k\}_{k\in\mathbb{N}}$ a nonnegative sequence while (30) makes it nonincreasing. Hence it converges to some l in $\mathbb{R}$. Since $\sum_{m=k}^{+\infty} \varepsilon_m$ tends to zero as k goes to infinity this proves that v_k also converges to l. ■

THEOREM 2 (NoLips**: Point convergence).** *The assumptions are those of Theorem 1.*
(i) **(Subsequential convergence)** *Assume that the solution set of* $(\mathcal{P})$,

$$\arg\min\left\{\Phi(x) : x \in C = \overline{\operatorname{dom}}\, h\right\}$$

is nonempty and compact. Then any limit point of $\{x^k\}_{k\in\mathbb{N}}$ *is a solution to* $(\mathcal{P})$.
(ii) **(Global convergence)** *Assume that* $\overline{\operatorname{dom}}\, h = \operatorname{dom} h$ *and that* **H** *is satisfied. We assume in addition that* $(\mathcal{P})$ *has at least a solution.*

Then the sequence $\{x^k\}_{k\in\mathbb{N}}$ *converges to some solution* x^* *of* $(\mathcal{P})$.

Proof. (i). Let x^* be a limit point of $\{x^k\}_{k\in\mathbb{N}}$ (which exists by compactness), i.e., $x^{k_p} \to x^*$ as $p \to \infty$. By using successively Theorem 1 (iii), the lower-semicontinuity of f and the continuity of g, we obtain

$$\begin{aligned}\min_C f+g &= \lim_{p\to\infty}\left(f(x^{k_p})+g(x^{k_p})\right)\\ &= \liminf_{p\to\infty}\left(f(x^{k_p})+g(x^{k_p})\right)\\ &\geq \liminf_{p\to\infty} f(x^{k_p})+\lim_{p\to\infty} g(x^{k_p})\\ &\geq f(x^*)+g(x^*).\end{aligned}$$

Since $x^* \in C$, the proof is complete.

(ii). From (25), we have for all $x \in \operatorname{dom} h = \overline{\operatorname{dom}}\, h$,

$$D_h(x,x^k) \leq D_h(x,x^{k-1}) + (\alpha-\delta)D_h(x^k,x^{k-1}) - \lambda_k\left(\Phi(x^k)-\Phi(x)\right).$$

Take $\bar{x} \in \arg\min\{\Phi(u) : u \in \operatorname{dom} h\}$. Since $\Phi(x^k) \geq \Phi(\bar{x})$ for all nonnegative integer k, the above inequality yields

$$D_h(\bar{x},x^k) \leq D_h(\bar{x},x^{k-1}) + (\alpha-\delta)D_h(x^k,x^{k-1}). \tag{31}$$

Using Lemma 6 and the fact that $D_h(x^k,x^{k-1})$ is summable (see Theorem 1-(ii)), we deduce that $D_h(\bar{x},x^k)_{k\in\mathbb{N}}$ is convergent for all minimizer $\bar{x}$. This ensures by **H** (i) that the sequence x^k is bounded. If one denotes by x^* a cluster point of x^k in $\overline{\operatorname{dom}}\, h = \operatorname{dom} h$, one deduces from part (i) that x^* is a minimizer of Φ on $\operatorname{dom} h$. As a consequence of the previous result and **H** (ii), $D_h(x^*,x^k)_{k\in\mathbb{N}}$ converges and its limit must be zero. The latter implies that x^k converges to x^* by **H** (iii). ■

Remark 5 (Bregman proximal minimization). When $g=0$, The algorithm NoLips reduces to the well known Bregman proximal minimization scheme ([13]). In this case, Theorem 1 recovers and extends the complexity/convergence results of [15, Theorem 3.4].

Remark 6 (Interior Bregman projected gradient). When $f=0$ our algorithm is the interior method with Bregman distance studied in [1]. An important difference with that work is the fact that we do not require ∇g to be L Lipschitz continuous in general. Instead, we identify in a sharp manner the geometrical assumption needed to ensure the descent property of the algorithm in a Bregman setting, i.e., $Lh-g$ is convex.

As we shall see, this is not a simple formal improvement: in the next section we show that a highly non-Lipschitz problem is amenable to the "interior proximal gradient methodology" through an adequate choice of the underlying geometry, i.e., through a judicious choice of a Bregman function h together with a constant L so that (LC) is satisfied.

5. Application to Linear Inverse Problems

5.1. Poisson Linear Inverse Problems: Some Background A large class of problems in astronomy, nuclear medicine (e.g., Positron Emission Tomography), electronic microscopy, and many other within the broad field of image sciences can be described as inverse problems where data measurements are collected by counting discrete events (e.g., photons, electrons) contaminated by noise and described by a Poisson process. One then needs to recover a nonnegative signal/image for the given problem. There is a huge amount of literature on this class of problems in the statistical and image sciences areas. For some classical works see e.g., [18, 16, 20] and [8] for a more recent review which includes over 130 references.

Formally, we are dealing with linear inverse problems that can be conveniently described as follows. Given a matrix $A \in \mathbb{R}_+^{m\times n}$ modeling the experimental protocol, and $b \in \mathbb{R}_{++}^m$ the vector of measurements, the goal is to reconstruct the signal or image $x \in \mathbb{R}_+^n$ from the noisy measurements b such that

$$Ax \simeq b.$$

Moreover, since the dimension of x is often much larger than the number of observations, there is a need to regularize the problem through an appropriate choice of a regularizer f reflecting desired features of the solution. Thus, given some adequate convex proximity measure $d(\cdot,\cdot)$ that quantifies the "error" between b and Ax, the task of recovering x can be formulated as the optimization problem:

$$(\mathcal{E}) \qquad \text{minimize } \; \{d(b,Ax) + \mu f(x) : \, x \in \mathbb{R}_+^n\}$$

where $\mu > 0$ plays the role of a penalty/regularizing parameter controlling the tradeoff between matching the data fidelity criteria and the weight given to its regularizer.

A natural and very well-known measure of proximity of two nonnegative vectors is based on the so-called Kullback-Liebler divergence (relative entropy functional) see [16],

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

which (up to some constants) corresponds to noise of the Poisson type, more precisely to the negative Poisson log-likelihood function. When $\mu = 0$, i.e., when solving the inverse problem without regularization, problem $(\mathcal{E})$ is the standard (modulo change of sign due to minimization) Poisson type maximum likelihood estimation problem. The latter is typically solved by the Expectation Maximization EM algorithm [37].

Clearly, the function $x \to d(b,Ax)$ is convex on $\mathbb{R}_+^n$, however it *does not* admit a globally Lipschitz continuous gradient, and hence the usual proximal gradient cannot be applied. This class of problems is sufficiently broad to illustrate the theory and algorithm we have developed. Many other related problems of this form can benefit from the same treatment via other choices of d and other regularizers. Our purpose here is just to illustrate how the NoLips algorithm can be applied to this class of problems, indicating its potential to the many related problems that can be similarly studied, and are left for future research.

5.2. Two Simple Algorithms for Poisson Linear Inverse Problems

Adopting the above model, we begin with some useful notations that will facilitate the forthcoming development.

Let $b \in \mathbb{R}_{++}^m$, and let $a_i \in \mathbb{R}_+^n$ denotes the rows of the matrix A. We assume that $a_i \neq 0$ for all $i = 1,\ldots,m$, and $\sum_{i=1}^m a_{ij} := r_j > 0$ forall j (which is a standard assumption for this model, [37]), so that for any $x \in \mathbb{R}_{++}^n$, we have $\langle a_i, x\rangle > 0$ for all $i = 1,\ldots,m$.

Let D_ϕ be the Bregman distance in $\mathbb{R}^m$ corresponding to the Boltzmann-Shannon entropy $\phi(u) = u \log u$,

$$D_\phi(u,v) = \sum_{i=1}^m \left[\phi(u_i) - \phi(v_i) - (u_i - v_i)\phi'(v_i)\right].$$

Then, the first component in the objective function of problem $(\mathcal{E})$ reads as

$$g(x) := D_\phi(b,Ax)$$

and the problem of interest can be written (omitting constant terms)

$$(\mathcal{E}) \qquad \text{minimize } \; \left\{ \sum_{i=1}^m \{\langle a_i,x\rangle - b_i \log\langle a_i,x\rangle\} + \mu f(x) : x \in \mathbb{R}_+^n \right\}.$$

To apply the algorithm NoLips, we need to identify an adequate Legendre function h. We use Burg's entropy, let indeed

$$h(x) = -\sum_{j=1}^{n} \log x_j, \ \operatorname{dom} h = \mathbb{R}^n_{++}.$$

Now, we need to find $L > 0$ such that $Lh - g$ is convex on $\mathbb{R}^n_{++}$.

Lemma 7. *Let $g(x) = D_\phi(b, Ax)$ and $h(x)$ as defined above. Then for any L satisfying*

$$L \geq \|b\|_1 = \sum_{i=1}^{m} b_i,$$

the function $Lh - g$ is convex on $\mathbb{R}^n_{++}$.

Proof. Since g and h are C^2 on $\mathbb{R}^n_{++}$, the convexity of $Lh - g$ is warranted for any $L > 0$ such that

$$L\langle \nabla^2 h(x)d, d\rangle \geq \langle \nabla^2 g(x)d, d\rangle, \quad \forall x \in \mathbb{R}^n_{++}, \forall d \in \mathbb{R}^n. \tag{32}$$

A simple computation shows that for any $x \in \mathbb{R}^n_{++}$ and any $d \in \mathbb{R}^n$,

$$L\langle \nabla^2 h(x)d, d\rangle = L\sum_{j=1}^{n} \frac{d_j^2}{x_j^2}. \tag{33}$$

On the other hand, using the definition of g one computes,

$$\nabla g(x) = \sum_{i=1}^{m} \left(1 - \frac{b_i}{\langle a_i, x\rangle}\right) a_i \tag{34}$$

$$\langle \nabla^2 g(x)d, d\rangle = \sum_{i=1}^{m} b_i \frac{\langle a_i, d\rangle^2}{\langle a_i, x\rangle^2}. \tag{35}$$

Now, by a simple application of Jensen's inequality to the nonnegative convex function t^2, it follows that for any $u \in \mathbb{R}^n_+$ (not all zero):

$$\frac{\langle u, d\rangle^2}{\langle u, x\rangle^2} \leq \sum_j \frac{u_j x_j}{\langle u, x\rangle}(d_j/x_j)^2 \leq \sum_{j=1}^{n} \frac{d_j^2}{x_j^2}, \ \forall d \in \mathbb{R}^n, x \in \mathbb{R}^n_{++}.$$

Applying the later with $u := a_i \neq 0$ for each i (by assumption), and recalling that $b > 0$, we obtain from (35),

$$\langle \nabla^2 g(x)d, d\rangle = \sum_{i=1}^{m} b_i \frac{\langle a_i, d\rangle^2}{\langle a_i, x\rangle^2} \leq \left(\sum_{i=1}^{m} b_i\right) \sum_{j=1}^{n} \frac{d_j^2}{x_j^2},$$

and hence with (33) the desired result (32) holds with $L \geq \sum_{i=1}^{m} b_i$. ■

Equipped with Lemma 7, Theorem 1 is applicable and Theorem 2 (i) warrants subsequential convergence to an optimal point. Since here $\alpha(h) = 0$, (cf. §2.3) we can take for example

$$\lambda = \frac{1}{2L} = \frac{1}{2\sum_{i=1}^{m} b_i}.$$

Applying NoLips, given $x \in \mathbb{R}^n_{++}$, the main algorithmic step, namely $x^+ = T_\lambda(x)$ consists of computing the proximal gradient step with a Burg's entropy:

$$x^+ = \arg\min\left\{\mu f(u) + \langle \nabla g(x), u\rangle + \frac{1}{\lambda}\sum_{j=1}^{n}\left(\frac{u_j}{x_j} - \log\frac{u_j}{x_j} - 1\right) : u > 0\right\}. \tag{36}$$

We now show that the above abstract iterative process yields closed form algorithms for Poisson reconstruction problems with two typical regularizers used in applications.

Sparse regularization. We choose the ℓ^1 regularizer $f(x) := \|x\|_1$, which is known to promote sparsity. In that case, since f is separable, the iterate in problem (36) reduces to solve a one dimensional convex problem of the form: (γ below stands for one component of $\nabla g(x)$)

$$x^+ = \arg\min \left\{ \mu u + \gamma u + \frac{1}{\lambda}\left(\frac{u}{x} - \log\frac{u}{x}\right) : u > 0 \right\}.$$

Thus, elementary algebra yields (presuming that $1 + \lambda\gamma x > 0$ to warrant $x^+ > 0$)

$$x^+ = \frac{x}{1 + \lambda\mu x + \lambda\gamma x}.$$

Define,

$$c_j(x) := \sum_{i=1}^m b_i \frac{a_{ij}}{\langle a_i, x\rangle}, \text{ and recall that } r_j := \sum_i a_{ij} > 0,$$

for every $j = 1, \ldots, n$ and $x \in \mathbb{R}^n_{++}$. The j-th component of the gradient of $\nabla g(x)$ defined in (34) can then be written as

$$\gamma_j := (\nabla g(x))_j = r_j - c_j(x), \quad \forall j = 1, \ldots, n, \forall x \in \mathbb{R}^n_{++}.$$

Thus, the algorithm to solve $(\mathcal{E})$ yields the following explicit iteration

$$x_j^+ = \frac{x_j}{1 + \lambda\left(\mu x_j + x_j(r_j - c_j(x))\right)}, \; j = 1, \ldots n; \; \text{ where } \; \lambda \in \left(0, \frac{1}{2L}\right).$$

For $\mu = 0$ problem $(\mathcal{E})$ reduces to solve $\min\{D_\phi(b, Ax) : x \in \mathbb{R}^n_+\}$, and in that particular case the iterates of NoLips simply become

$$x_j^+ = \frac{x_j}{1 + \lambda x_j(r_j - c_j(x))}, \; j = 1, \ldots n.$$

In contrast to the standard EM multiplicative algorithm given by the iteration [37]

$$x_j^+ = \frac{x_j}{r_j} c_j(x), \; j = 1, \ldots, n.$$

Tikhonov regularization. We consider here a regularization à la Tikhonov, i.e., where the regularizer is $f(x) := \frac{1}{2}\|x\|^2$. We recall that this term is used as a penalty in order to promote solutions of $Ax = b$ with small Euclidean norms, see e.g., [20].

As before, we are lead to a simple one dimensional problem (recall that γ is a component of $\nabla g(x)$)

$$x^+ = \arg\min \left\{ \frac{\mu}{2} u^2 + \gamma u + \frac{1}{\lambda}\left(\frac{u}{x} - \log\frac{u}{x}\right) : u > 0 \right\}.$$

Presuming that we are in the case of existence in the above subproblem, one deduces that

$$x^+ = \frac{\sqrt{(1 + \gamma\lambda x)^2 + 4\mu\lambda x^2} - (1 + \gamma\lambda x)}{2\mu\lambda x} > 0.$$

Using the notation introduced above, we obtain a "log-Thikonov method":

$$x_j^+ = \frac{\sqrt{\left(1 + \lambda x_j\left(r_j - c_j(x)\right)\right)^2 + 4\mu\lambda x_j^2} - \left(1 + \lambda x_j\left(r_j - c_j(x)\right)\right)}{2\mu\lambda x_j},$$

where $j = 1, \ldots, n$ and $x \in \mathbb{R}^n_{++}$.

As mentioned above, many other interesting methods can be considered by choosing different kernels for ϕ or by reversing the order of the arguments in the proximity measure. This is illustrated in the next example.

5.3. An algorithm for nonnegative linear systems on $\mathbb{R}^n_+$ An alternative approach to what was developed in the previous section consists in minimizing

$$\min\{\mu f(x) + D_\phi(Ax, b) : x \in \mathbb{R}^n_+\}$$

instead of $\min\{\mu f(x) + D_\phi(b, Ax) : x \in \mathbb{R}^n_+\}$. Since D_ϕ is not symmetric these problems differ and the function $g(x) := D_\phi(Ax, b)$ now reads:

$$g(x) = \sum_{i=1}^{m} \{\langle a_i, x\rangle \log\langle a_i, x\rangle - (\log b_i + 1)\langle a_i, x\rangle + b_i\}.$$

Note that in this case, the Poisson statistical interpretation of the objective g is not anymore valid. Yet, for solving inconsistent linear systems with nonnegative data, we can naturally adopt the distance $D_\phi(Ax, b)$ to measure the residuals between two nonnegative points, instead e.g., the usual least-squares norm, see e.g., [16] and references therein. As we shall see, this leads to a simple *multiplicative* gradient-like iterative scheme.

A judicious choice to implement NoLips in this case is to use the Legendre function $h(x) = \sum_{j=1}^n x_j \log x_j$. As before, we need to find $L > 0$ such that $Lh - g$ is convex on $\mathbb{R}^n_{++}$.

LEMMA 8. *Let $g(x) = D_\phi(Ax, b)$ and $h(x)$ as defined above. Then for any L satisfying*

$$L \geq \max_{1\leq j\leq n} \sum_{i=1} a_{ij}$$

the function $Lh - g$ is convex on $\mathbb{R}^n_{++}$.

Proof. Since g and h are C^2 on $\mathbb{R}^n_{++}$, we proceed as in Lemma 7. We seek a positive constant L such that

$$L\langle\nabla^2 h(x)d, d\rangle \geq \langle\nabla^2 g(x)d, d\rangle, \quad \forall x \in \mathbb{R}^n_{++}, \forall d \in \mathbb{R}^n.$$

A direct computation shows that for any $x \in \mathbb{R}^n_{++}$ and any $d \in \mathbb{R}^n$,

$$L\langle\nabla^2 h(x)d, d\rangle = L\sum_{j=1}^{n} \frac{d_j^2}{x_j} \text{ and } \langle\nabla^2 g(x)d, d\rangle = \sum_{i=1}^{m} \frac{\langle a_i, d\rangle^2}{\langle a_i, x\rangle}. \tag{37}$$

Consider the convex function $\varphi : \mathbb{R} \times (0, \infty)$ defined by $\varphi(s, t) = s^2/t$, and take any $u \in \mathbb{R}^n_+ \setminus \{0\}$, $x \in \mathbb{R}^n_{++}$, and $d \in \mathbb{R}^n$. Apply Jensen's inequality with weights and points

$$\lambda_j := \frac{u_j x_j}{\langle u, x\rangle} > 0,\ (s_j, t_j) := \left(\langle u, x\rangle \frac{d_j}{x_j}, \langle u, x\rangle\right),$$

to obtain

$$\frac{\langle u, d\rangle^2}{\langle u, x\rangle} \leq \sum_{j=1}^{n} u_j \frac{d_j^2}{x_j},\ \forall d \in \mathbb{R}^n,\ \forall x \in \mathbb{R}^n_{++}.$$

Invoking the latter inequality with $u := a_i \in \mathbb{R}^n_+ \setminus \{0\}$ for all i, we thus get:

$$\begin{aligned}\langle \nabla^2 g(x)d, d\rangle &= \sum_{i=1}^m \frac{\langle a_i, d\rangle^2}{\langle a_i, x\rangle}\\ &\le \sum_{i=1}^m \sum_{j=1}^n a_{ij}\frac{d_j^2}{x_j} = \sum_{j=1}^n \frac{d_j^2}{x_j}\left(\sum_{i=1}^m a_{ij}\right)\\ &\le \max_{1\le j\le n}\left(\sum_{i=1}^m a_{ij}\right)\sum_{j=1}^n \frac{d_j^2}{x_j},\end{aligned}$$

and hence the desired result follows with $L \ge \max_j\left(\sum_{i=1}^m a_{ij}\right)$. ■

Remark 7. In some applications, such as positron emission tomography (PET),(see e.g., [37]), it is common to have matrices with unit column sums. In that case $r_j = \sum_{i=1}^m a_{ij} = 1$ for all $j = 1, \ldots, n$, so that $L = 1$.

Let us provide an explicit expression for NoLips in the case of the sparse minimization problem:

$$\min\{\mu\|x\|_1 + g(x) : x \ge 0\} = \min\{\mu\|x\|_1 + D_h(Ax, b) : x \ge 0\}.$$

In this case Theorem 2 (ii) and Corollary 1 are applicable and warrant global convergence to an optimal solution and sublinear rate $O\left(\frac{1}{k}\right)$.

Given $x \in \mathbb{R}^n_{++}$, the iteration $x^+ = T_\lambda(x)$ amounts to solving the one dimensional problem:

$$x^+ = \arg\min\{\lambda\mu u + \lambda\gamma u + u\log\frac{u}{x} + x - u\},$$

where γ is some component of $\nabla g(x)$. Observe that in this setting ∇g is given by

$$(\nabla g(x))_j = \sum_{i=1}^m a_{ij}\log\left(\frac{\langle a_i, x\rangle}{b_i}\right),\ \forall j = 1, \ldots, n.$$

Assuming $\sum_{i=1}^m a_{ij} = 1$, we can take $L = 1$ (e.g., see Remark 7) so that $\lambda = 1/2$. After a few computations, we obtain the following iterative process:

$$x_j^+ = \frac{x_j e^{-\mu/2}}{p_j(x)} \text{ with } p_j(x) := \prod_{i=1}^m \left(\frac{\langle a_i, x\rangle}{b_i}\right)^{a_{ij}/2},\ \forall\, j = 1, \ldots, n;\ x \in \mathbb{R}^n_{++}.$$

6. Concluding Remarks

This work outlines in simple and transparent ways the basic ingredients to apply the proximal gradient methodology when the gradient of the smooth part in the composite model $(\mathcal{P})$ is not Lipschitz continuous. Thanks to a new and natural extension of the descent Lemma and a sharp definition of the step-size through the notion of symmetry, we have shown that NoLips shares convergence and complexity results akin to those of the usual proximal gradient. The last section has illustrated the potential of the new proposed framework when applied to the key research area of linear inverse problems with Poisson noise which arises in image sciences. On the theoretical side, our approach lays the ground for many new and promising perspectives for gradient-based methods that were not conceivable before. Thus, it would be interesting in the future to revisit classical results for first-order methods without Lipschitz gradient continuity and to investigate the impact and extension of our methodology, e.g., on primal-dual schemes, on the derivation and analysis of new first order accelerated NoLips schemes, and on other applications.

Acknowledgments. HHB was partially supported by the Natural Sciences and Engineering Council of Canada (Discovery and Accelerator Supplement Grants). HHB thanks the Toulouse School of Economics for their hospitality during a research visit in Summer 2013 during which this work started. The research of JB benefited from the support of the Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant number FA9550-14-1-0056. The research of MT was partially supported by the Israel Science Foundation under grant ISF-998-12, and by the German-Israel Foundation under grant GIF-G1253304.6-2014. MT thanks the Toulouse School of Economics for their hospitality during a research visit in February 2015 during which this work was completed.

References

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

[2] H.H. Bauschke and J.M. Borwein, Legendre functions and the method of Bregman projections, *Journal of Convex Analysis* 4 (1997), 27–67.

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

[4] H.H. Bauschke and P.L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, 2011.

[5] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM Journal on Imaging Science* 2 (2009), 183–202.

[6] A. Beck and M. Teboulle, Gradient-based algorithms with applications to signal recovery problems. In D. Palomar and Y. C. Eldar, editors, *Convex Optimization in Signal Processing and Communications*, pp. 139–162. Cambridge University Press, 2009.

[7] S.R. Becker, E.J. Candes, and M.C. Grant, Templates for convex cone problems with applications to sparse signal recovery, *Mathematical Programming Computation* 3 (2011), 165–218.

[8] M. Bertero, P. Boccacci, G. Desider, and G. Vicidomini, Image deblurring with Poisson data: from cells to galaxies, *Inverse Problems* 25 (2009), 26 pages.

[9] D.P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont Massachusetts, second edition, 1999.

[10] J. Bolte and M. Teboulle, Barrier operators and associated gradient-like dynamical systems for constrained minimization problems, *SIAM Journal on Control and Optimization* 42 (2003), 1266–1292.

[11] 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. Computational Mathematics and Mathematical Physics* 7 (1967), 200–217.

[12] R. Bruck, On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space, *Journal of Mathematical Analysis and Applications* 61 (1977), 159–164.

[13] Y. Censor and S. A. Zenios, Proximal minimization algorithm with D-functions, *Journal of Optimization Theory and Applications* 73 (1992), 451–464.

[14] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging.G. Chen and M. Teboulle, *Journal of Mathematical Imaging and Vision*, (2010), 1–26.

[15] G. Chen and M. Teboulle, Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization* 3 (1993), 538–543.

[16] I. Csiszár. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *The Annals of Statististics* 19 (1991), 2032–2066.

[17] P.L. Combettes and V.R. Wajs, Signal recovery by proximal forward-backward splitting. *SIAM Multiscale Modeling and Simululation* 4 (2005), 1168–200.

[18] A.P. Dempster, N.M. Laird, and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* 39 (1977), 1–38.

[19] J. Eckstein, Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming, *Mathematics of Operations Research* 18 (1993), 202–226.

[20] H.W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems.* Mathematics and Its Applications, Kluwer Academic Publisher, 2000.

[21] M. Fukushima and H. Milne, A generalized proximal point algorithm for certain nonconvex minimization problems. *International Journal of Systems Science* 12 (1981), 989–1000.

[22] D. Gabay, Applications of the method of multipliers to variational inequalities. In M. Fortin and R. Glowinski, editors, *Applications to the Solution of Boundary-Valued Problems*, pp. 299–331. North-Holland, Amsterdam, 1983.

[23] R. Glowinski and P. Le Tallec. *Augmented Lagrangian and operator splitting methods in nonlinear mechanics*, volume 9. Society for Industrial Mathematics, 1989.

[24] O. Güler, On the convergence of the proximal point algorithm for convex minimization, *SIAM Journal on Control and Optimization* 29 (1991), 403–419.

[25] B. He and X. Yuan. On the $O(1/n)$ convergence rate of the douglas-rachford alternating direction method. *SIAM Journal on Numerical Analysis* 50 (2012), 700–709.

[26] J.J. Moreau, Proximité et dualité dans un espace hilbertien, *Bulletin de la Société Mathématique de France* 90 (1965), 273–299.

[27] Y. Nesterov, A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Doklady Akad Nauk SSSR*, 269 (1983), 543–547.

[28] Z. Opial, Weak convergence of the sequence of successive approximations for nonexpansive mappings, *Bulletin of the AMS* 73 (1967), 591–597.

[29] D.P. Palomar and Y.C. Eldar, *Convex optimization in signal processing and communications*, Cambridge University Press, 2010.

[30] G.B. Passty, Ergodic convergence to a zero of the sum of monotone operators in Hilbert space, *Journal of Mathematical Analysis and Applications* 72 (1979), 383–390.

[31] B.T. Polyak, *Introduction to Optimization*, Optimization Software Inc., New York, 1987.

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

[33] R. Shefi and M. Teboulle. Rate of Convergence Analysis of Decomposition Methods Based on the Proximal Method of Multipliers for Convex Minimization. *SIAM J. Optimization* 24, (2014), 269-297.

[34] S. Sra, S. Nowozin, and S.J. Wright, Read More: http://epubs.siam.org/doi/abs/10.1137/130910774 *Optimization for Machine Learning*, MIT Press, Cambridge, 2011.

[35] M. Teboulle, Entropic proximal mappings with application to nonlinear programming, *Mathematics of Operations Research* 17 (1992), 670–690.

[36] P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization, *Mathematical Programming Series B* 125 (2010), 263–295.

[37] Y. Vardi, L. Shepp, and L. Kaufman, A statistical model for positron emission tomography, *Journal of the American Statistical Association* 80 (1985), 8–37.