



**HAL**  
open science

# Non-smooth Non-convex Bregman Minimization: Unification and new Algorithms

Peter Ochs, Jalal M. Fadili, Thomas Brox

► **To cite this version:**

Peter Ochs, Jalal M. Fadili, Thomas Brox. Non-smooth Non-convex Bregman Minimization: Unification and new Algorithms. Journal of Optimization Theory and Applications, In press. hal-01921804

**HAL Id: hal-01921804**

**<https://hal.science/hal-01921804>**

Submitted on 14 Nov 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Non-smooth Non-convex Bregman Minimization: Unification and new Algorithms

Peter Ochs\*, Jalal Fadili<sup>†</sup>, and Thomas Brox<sup>‡</sup>

\* Saarland University, Saarbrücken, Germany

<sup>†</sup> Normandie Univ, ENSICAEN, CNRS, GREYC, France

<sup>‡</sup> University of Freiburg, Freiburg, Germany

## Abstract

We propose a unifying algorithm for non-smooth non-convex optimization. The algorithm approximates the objective function by a convex model function and finds an approximate (Bregman) proximal point of the convex model. This approximate minimizer of the model function yields a descent direction, along which the next iterate is found. Complemented with an Armijo-like line search strategy, we obtain a flexible algorithm for which we prove (subsequential) convergence to a stationary point under weak assumptions on the growth of the model function error. Special instances of the algorithm with a Euclidean distance function are, for example, Gradient Descent, Forward–Backward Splitting, ProxDescent, without the common requirement of a “Lipschitz continuous gradient”. In addition, we consider a broad class of Bregman distance functions (generated by Legendre functions), replacing the Euclidean distance. The algorithm has a wide range of applications including many linear and non-linear inverse problems in signal/image processing and machine learning.

## 1 Introduction

When minimizing a non-linear function on the Euclidean vector space, a fundamental strategy is to successively minimize approximations to the actual objective function. We refer to such an approximation as model (function). A common model example in smooth optimization is linearization (first order Taylor approximation) around the current iterate. However, in general, the minimization of a linear function does not provide a finite solution, unless, for instance, the domain is compact. Therefore, the model is usually complemented by a proximity measure, which favors a solution (the next iterate) close to the current iterate. For the Euclidean norm as proximity measure, computing the next iterate (minimizer of the sum of the model function and the Euclidean proximity measure) is equivalent to a Gradient Descent step, i.e. the next iterate is obtained by moving along the direction of the negative gradient at the current point for a certain step size.

Since sequential minimization of model functions does not require smoothness of the objective or the model function, non-smoothness is handled naturally. The crucial aspect is the “approximation quality” of the model function, which is controlled by a growth function, that describes the approximation error around the current iterate. Drusvyatskiy et al. [19] refer to such model functions as Taylor-like models. The difference among algorithms lies in the properties of such a growth function, rather than the specific choice of a model function.

For the example of the Gradient Descent model function (linearization around the current iterate) for a continuously differentiable function, the value and the derivative of the growth function (approximation error) vanish at the current iterate. In this case, a line search strategy is required to determine a suitable step size that reduces the objective value. If the gradient of the objective function is additionally  $L$ -Lipschitz continuous, then the growth function satisfies a quadratic growth globally, and step sizes can be controlled analytically.

A large class of algorithms, which are widely popular in machine learning, statistics, computer vision, signal and image processing can be cast in the same framework. This includes algorithms such as Forward–Backward Splitting [26] (Proximal Gradient Descent), ProxDescent [25, 20] (or proximal Gauss–Newton method), and many others. They all obey the same growth function as Gradient Descent. This allows for a unified analysis of all these algorithms, which is a key contribution of this paper. Moreover, we allow for a broad class of (iteration dependent) Bregman proximity functions (e.g., generated by common entropies such as Boltzmann–Shannon, Fermi–Dirac, and Burg’s entropy), which leads to new algorithms. To be generic in the choice of the objective, the model, and the Bregman functions, the algorithm is complemented with an Armijo-like line search strategy. Subsequential convergence to a stationary point is established for different types of growth functions.

The above mentioned algorithms are ubiquitous in applications of machine learning, computer vision, image/signal processing, and statistics as is illustrated in Section 5 and in our numerical experiments in Section 6. Due to the unifying framework, the flexibility of these methods is considerably increased further.

## 2 Contributions and Related Work

For smooth functions, Taylor’s approximation is unique. However, for non-smooth functions, there are only “Taylor-like” model functions [32, 31, 19]. Each model function yields another algorithm. Some model functions [32, 31] could also be referred to as lower-Taylor-like models, as there is only a lower bound on the approximation quality of the model. Noll et al. [31] addressed the problem by bundle methods based on cutting planes, which differs from our setup.

The goal of Drusvyatskiy et al. [19] is to measure the proximity of an approximate solution of the model function to a stationary point of the original objective, i.e., a suitable

stopping criterion for non-smooth objectives is sought. On the one hand, their model functions may be non-convex, unlike ours. On the other hand, their growth functions are more restrictive. Considering their abstract level, the convergence results may seem satisfactory. However, several assumptions that do not allow for a concrete implementation are required, such as a vanishing distance between successive iterates and convergence of the objective values along a generated convergent subsequence to the objective value of the limit point. This is in contrast to our framework.

We assume more structure of the subproblems: They are given as the sum of a model function and a Bregman proximity function. With this mild assumption on the structure and a suitable line-search procedure, the algorithm can be implemented and the convergence results apply. **We present the first implementable algorithm in the abstract model function framework and prove subsequential convergence to a stationary point.**

Our algorithm **generalizes ProxDescent [20, 25] with convex subproblems**, which is known for its broad applicability. We provide more flexibility by considering Bregman proximity functions, and our **backtracking line-search need not solve the subproblems for each trial step.**

The algorithm and convergence analysis is a **far-reaching generalization** of Bonettini et al. [11], which is similar to the instantiation of our framework where the model function leads to Forward–Backward Splitting. The proximity measure of Bonettini et al. [11] is assumed to satisfy a strong convexity assumption. Our **proximity functions can be generated by a broad class of Legendre functions**, which includes, for example, the non-strongly convex Burg’s entropy [13, 3] for the generation of the Bregman proximity function.

### 3 Preliminaries and Notations

Throughout the whole paper, we work in a Euclidean vector space  $\mathbb{R}^N$  of dimension  $N \in \mathbb{N}$  equipped with the standard *inner product*  $\langle \cdot, \cdot \rangle$  and associated *norm*  $|\cdot|$ .

**Variational analysis.** We work with extended-valued functions  $f: \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$ ,  $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ . The *domain* of  $f$  is  $\text{dom } f := \{x \in \mathbb{R}^N \mid f(x) < +\infty\}$  and a function  $f$  is *proper*, if it is nowhere  $-\infty$  and  $\text{dom } f \neq \emptyset$ . It is *lower semi-continuous* (or *closed*), if  $\liminf_{x \rightarrow \bar{x}} f(x) \geq f(\bar{x})$  for any  $\bar{x} \in \mathbb{R}^N$ . Let  $\text{int } \Omega$  denote the *interior* of  $\Omega \subset \mathbb{R}^N$ . We use the notation of *f-attentive convergence*  $x \xrightarrow{f} \bar{x} \Leftrightarrow (x, f(x)) \rightarrow (\bar{x}, f(\bar{x}))$ , and the notation  $k \xrightarrow{K} \infty$  for some  $K \subset \mathbb{N}$  to represent  $k \rightarrow \infty$  where  $k \in K$ .

As in [19], we introduce the following concepts. For a closed function  $f: \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$  and a

point  $\bar{x} \in \text{dom } f$ , we define the *slope* of  $f$  at  $\bar{x}$  by

$$|\nabla f|(\bar{x}) := \limsup_{x \rightarrow \bar{x}, x \neq \bar{x}} \frac{[f(\bar{x}) - f(x)]_+}{|x - \bar{x}|},$$

where  $[s]_+ := \max\{s, 0\}$ . It is the maximal instantaneous rate of decrease of  $f$  at  $\bar{x}$ . For a differentiable function, it coincides with the norm of the gradient  $|\nabla f(\bar{x})|$ . Moreover, the *limiting slope*

$$\overline{|\nabla f|}(\bar{x}) := \liminf_{x \xrightarrow{f} \bar{x}} |\nabla f|(x)$$

is key. For a convex function  $f$ , we have  $\overline{|\nabla f|}(\bar{x}) = \inf_{v \in \partial f(\bar{x})} |v|$ , where  $\partial f(\bar{x})$  is the (*convex*) *subdifferential*  $\partial f(\bar{x}) := \{v \in \mathbb{R}^N \mid \forall x: f(x) \geq f(\bar{x}) + \langle x - \bar{x}, v \rangle\}$ , whose *domain* is given by  $\text{dom } \partial f := \{x \in \mathbb{R}^N \mid \partial f(x) \neq \emptyset\}$ . A point  $\bar{x}$  is a *stationary point* of the function  $f$ , if  $\overline{|\nabla f|}(\bar{x}) = 0$  holds. Obviously, if  $|\nabla f|(\bar{x}) = 0$ , then  $\overline{|\nabla f|}(\bar{x}) = 0$ . We define the *set of (global) minimizers* of a function  $f$  by

$$\text{Argmin}_{x \in \mathbb{R}^N} f(x) := \left\{ x \in \mathbb{R}^N \mid f(x) = \inf_{\bar{x} \in \mathbb{R}^N} f(\bar{x}) \right\},$$

and the (*unique*) *minimizer* of  $f$  by  $\text{argmin}_{x \in \mathbb{R}^N} f(x)$ , if  $\text{Argmin}_{x \in \mathbb{R}^N} f(x)$  consists of a single element. As shorthand, we also use  $\text{Argmin } f$  and  $\text{argmin } f$ .

**Definition 1 (Growth function [19]).** A differentiable univariate function  $\omega: \mathbb{R}_+ \rightarrow \mathbb{R}_+$  is called *growth function* if it satisfies  $\omega(0) = \omega'_+(0) = 0$ . If, in addition,  $\omega'_+(t) > 0$  for  $t > 0$  and equalities  $\lim_{t \searrow 0} \omega'_+(t) = \lim_{t \searrow 0} \omega(t)/\omega'_+(t) = 0$  hold, we say that  $\omega$  is a *proper growth function*.

Concrete instances of growth functions will be generated for example by the concept of  $\psi$ -uniform continuity, which is a generalization of Lipschitz and Hölder continuity.

**Definition 2.** A mapping  $F: \mathbb{R}^N \rightarrow \mathbb{R}^M$  is called  $\psi$ -uniform continuous with respect to a continuous function  $\psi: \mathbb{R}_+ \rightarrow \mathbb{R}_+$  with  $\psi(0) = 0$ , if the following holds:

$$|F(x) - F(\bar{x})| \leq \psi(|x - \bar{x}|) \quad \text{for all } x, \bar{x} \in \mathbb{R}^N.$$

**Example 3.** Let  $F$  be  $\psi$ -uniform continuous. If, for some  $c > 0$ , we have  $\psi(s) = cs^\alpha$  with  $\alpha \in ]0, 1]$ , then  $F$  is Hölder continuous, which for  $\alpha = 1$  is the same as Lipschitz continuity.

In analogy to the case of Lipschitz continuity, we can state a generalized Descent Lemma:

**Lemma 4 (Generalized Descent Lemma).** Let  $f: \mathbb{R}^N \rightarrow \mathbb{R}$  be continuously differentiable and let  $\nabla f: \mathbb{R}^N \rightarrow \mathbb{R}^N$  be  $\psi$ -uniform continuous. Then, the following holds

$$|f(x) - f(\bar{x}) - \langle \nabla f(\bar{x}), x - \bar{x} \rangle| \leq \int_0^1 \frac{\varphi(s|x - \bar{x}|)}{s} ds \quad \text{for all } x, \bar{x} \in \mathbb{R}^N,$$

where  $\varphi: \mathbb{R}_+ \rightarrow \mathbb{R}_+$  is given by  $\varphi(s) := s\psi(s)$ .

*Proof.* We follow the proof of the Descent Lemma for functions with Lipschitz gradient:

$$\begin{aligned}
 |f(x) - f(\bar{x}) - \langle \nabla f(\bar{x}), x - \bar{x} \rangle| &= \left| \int_0^1 \langle \nabla f(\bar{x} + s(x - \bar{x})) - \nabla f(\bar{x}), x - \bar{x} \rangle ds \right| \\
 &\leq \int_0^1 |\nabla f(\bar{x} + s(x - \bar{x})) - \nabla f(\bar{x})| |x - \bar{x}| ds \\
 &\leq \int_0^1 \psi(s|x - \bar{x}|) |x - \bar{x}| ds = \int_0^1 \frac{\varphi(s|x - \bar{x}|)}{s} ds. \quad \square
 \end{aligned}$$

**Example 5.** The function  $\omega(t) = \int_0^1 \frac{\varphi(st)}{s} ds$  is an example for a growth function. Obviously, we have  $\omega(0) = 0$  and, using the Dominated Convergence Theorem (with majorizer  $\sup_{s \in [0,1]} \psi(s) < +\infty$  for small  $t \geq 0$ ), we conclude

$$\omega'_+(0) = \lim_{t \searrow 0} \int_0^1 \frac{\varphi(st)}{st} ds = \lim_{t \searrow 0} \int_0^1 \psi(st) ds = \int_0^1 \lim_{t \searrow 0} \psi(st) ds = 0.$$

It becomes a proper growth function, for example, if  $\psi(s) = 0 \Leftrightarrow s = 0$  and we impose the additional condition  $\lim_{t \searrow 0} \omega(t)/\psi(t) = 0$ . The function  $\psi(s) = cs^\alpha$  with  $\alpha > 0$ , i.e.  $\varphi(s) = cs^{1+\alpha}$ , is an example for a proper growth function.

**Bregman distance.** In order to introduce the notion of a Bregman function [12], we first define a set of properties for functions to generate nicely behaving Bregman functions.

**Definition 6 (Legendre function [4, Def. 5.2]).** The proper, closed, convex function  $h: \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$  is

- (i) *essentially smooth*, if  $\partial h$  is both locally bounded and single-valued on its domain,
- (ii) *essentially strictly convex*, if  $(\partial h)^{-1}$  is locally bounded on its domain and  $h$  is strictly convex on every convex subset of  $\text{dom } \partial h$ , and
- (iii) *Legendre*, if  $h$  is both essentially smooth and essentially strictly convex.

Note that we have the duality  $(\partial h)^{-1} = \partial h^*$  where  $h^*$  denotes the conjugate of  $h$ .

**Definition 7 (Bregman distance [12]).** Let  $h: \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$  be proper, closed, convex and Gâteaux differentiable on  $\text{int dom } h \neq \emptyset$ . The *Bregman distance* associated with  $h$  is the function

$$D_h: \mathbb{R}^N \times \mathbb{R}^N \rightarrow [0, +\infty], \quad (x, \bar{x}) \mapsto \begin{cases} h(x) - h(\bar{x}) - \langle x - \bar{x}, \nabla h(\bar{x}) \rangle, & \text{if } \bar{x} \in \text{int dom } h; \\ +\infty, & \text{otherwise.} \end{cases}$$

In contrast to the Euclidean distance, the Bregman distance is lacking symmetry.

We focus on Bregman distances that are generated by Legendre functions from the following class:

$$\mathcal{L} := \left\{ h: \mathbb{R}^N \rightarrow \overline{\mathbb{R}} \left| \begin{array}{l} h \text{ is a proper, closed, convex} \\ \text{Legendre function that is} \\ \text{Fréchet differentiable on } \text{int dom } h \neq \emptyset \end{array} \right. \right\}.$$

To control the variable choice of Bregman distances throughout the algorithm's iterations, we introduce the following ordering relation for  $h_1, h \in \mathcal{L}$ :

$$h_1 \succeq h \quad \Leftrightarrow \quad \forall x \in \text{dom } h: \forall \bar{x} \in \text{int dom } h: D_{h_1}(x, \bar{x}) \geq D_h(x, \bar{x}).$$

As a consequence of  $h_1 \succeq h$ , we have  $\text{dom } D_{h_1} \subset \text{dom } D_h$ .

In order to conveniently work with Bregman distances, we collect a few properties.

**Proposition 8.** Let  $h \in \mathcal{L}$  and  $D_h$  be the associate Bregman distance.

- (i)  $D_h$  is strictly convex on every convex subset of  $\text{dom } \partial h$  with respect the first argument.
- (ii) For  $\bar{x} \in \text{int dom } h$ , it holds that  $D_h(x, \bar{x}) = 0$  if and only if  $x = \bar{x}$ .
- (iii) For  $x \in \mathbb{R}^N$  and  $\bar{x}, \hat{x} \in \text{int dom } h$  the following *three point identity* holds:

$$D_h(x, \bar{x}) = D_h(x, \hat{x}) + D_h(\hat{x}, \bar{x}) + \langle x - \hat{x}, \nabla h(\hat{x}) - \nabla h(\bar{x}) \rangle.$$

*Proof.* (i) and (ii) follow directly from the definition of  $h$  being essentially strictly convex. (iii) is stated in [16]. It follows from the definition of a Bregman distance.  $\square$

Associated with such a distance function is the following proximal mapping.

**Definition 9 (Bregman proximal mapping [5, Def. 3.16]).** Let  $f: \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$  and  $D_h$  be a Bregman distance associated with  $h \in \mathcal{L}$ . The  $D_h$ -prox (or Bregman proximal mapping) associated with  $f$  is defined by

$$P_f^h(\bar{x}) := \underset{x}{\text{argmin}} f(x) + D_h(x, \bar{x}). \tag{1}$$

In general, the proximal mapping is set-valued, however for a convex function, the following lemma simplifies the situation.

**Lemma 10.** Let  $f: \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$  be a proper, closed, convex function that is bounded from below, and  $h \in \mathcal{L}$  such that  $\text{int dom } h \cap \text{dom } f \neq \emptyset$ . Then the associated Bregman proximal mapping  $P_f^h$  is single-valued on its domain and maps to  $\text{int dom } h \cap \text{dom } f$ .

*Proof.* Single-valuedness follows from [5, Corollary 3.25(i)]. The second claim is from [5, Prop. 3.23(v)(b)].  $\square$

**Proposition 11.** Let  $f: \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$  be a proper, closed, convex function that is bounded from below, and  $h \in \mathcal{L}$  such that  $\text{int dom } h \cap \text{dom } f \neq \emptyset$ . For  $\bar{x} \in \text{int dom } h$ ,  $\hat{x} = P_f^h(\bar{x})$ , and any  $x \in \text{dom } f$  the following inequality holds:

$$f(x) + D_h(x, \bar{x}) \geq f(\hat{x}) + D_h(\hat{x}, \bar{x}) + D_h(x, \hat{x}).$$

*Proof.* See [16, Lemma 3.2]. □

For examples and more useful properties of Bregman functions, we refer the reader to [3, 5, 6, 30].

**Miscellaneous.** We make use of little-o notation  $f \in o(g)$  (or  $f = o(g)$ ), which indicates that the asymptotic behavior of a function  $f$  is dominated by that of the function  $g$ . Formally, it is defined by

$$f \in o(g) \quad \Leftrightarrow \quad \forall \varepsilon > 0: |f(x)| \leq \varepsilon |g(x)| \text{ for } |x| \text{ sufficiently small.}$$

Note that a function  $\omega$  is in  $o(t)$  if, and only if  $\omega$  is a growth function.

## 4 Line Search Based Bregman Minimization Algorithms

In this paper, we solve optimization problems of the form

$$\min_{x \in \mathbb{R}^N} f(x) \tag{2}$$

where  $f: \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$  is a proper, closed function on  $\mathbb{R}^N$ . We assume that  $\text{Argmin } f \neq \emptyset$  and  $\underline{f} := \min f > -\infty$ . The main goal is to develop a provably (subsequentially) convergent algorithm that finds a stationary point  $x$  of (2) in the sense of the limiting slope  $|\overline{\nabla} f|(x) = 0$ .

We analyze abstract algorithms that sequentially minimize convex models of the objective function.

### 4.1 The Abstract Algorithm

For each point  $\bar{x}$ , we consider a proper, closed, convex *model function*  $f_{\bar{x}}: \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$  with

$$|f(x) - f_{\bar{x}}(x)| \leq \omega(|x - \bar{x}|), \tag{3}$$

where  $\omega$  is a growth function as defined in Definition 1. The *model assumption* (3) is an abstract description of a (local) first order oracle. For examples, we refer to Section 5.

Before delving further into the details, we need a bit of notation. Let

$$f_{\bar{x}, \bar{z}}^h(x) := f_{\bar{x}}(x) + D_h(x, \bar{z}) \quad \text{and} \quad f_{\bar{x}}^h := f_{\bar{x}, \bar{x}}^h,$$



where  $h \in \mathcal{L}$ . Note that  $f_{\bar{x}}^h(\bar{x}) = f(\bar{x})$ . Moreover, the following quantity defined for generic points  $\bar{x}$ ,  $x$  and  $\tilde{x}$  will be important:

$$\Delta_{\bar{x}}^h(x, \tilde{x}) := f_{\bar{x}}^h(x) - f_{\bar{x}}^h(\tilde{x}). \quad (4)$$

For  $\tilde{x} = \bar{x}$ , it measures the decrease of the surrogate function  $f_{\bar{x}}^h$  from the current iterate  $\bar{x}$  to any other point  $x$ . Obviously, the definition implies that  $\Delta_{\bar{x}}^h(x, x) = 0$  for all  $x$ .

**Algorithm.** We consider the following Algorithm 1.

**Algorithm 1 (Inexact Bregman Proximal Minimization Line Search).**

- **Basic prerequisites:** Fix  $\gamma \in ]0, 1[$  and  $h \in \mathcal{L}$ . Let
  - $(x_k)_{k \in \mathbb{N}}$  and  $(\tilde{y}_k)_{k \in \mathbb{N}}$  be sequences in  $\mathbb{R}^N$ ;
  - $(f_{x_k})_{k \in \mathbb{N}}$  be a sequence of model functions with  $\inf_{k \in \mathbb{N}} \inf_x f_{x_k}(x) > -\infty$ ;
  - $(h_k)_{k \in \mathbb{N}}$  be a sequence in  $\mathcal{L}$  with  $h_k \succeq h$ ;
  - $(\eta_k)_{k \in \mathbb{N}}$  be a sequence of positive real numbers.
- **Initialization:** Select  $x_0 \in \text{dom } f \cap \text{int dom } h$ .
- **For each  $k \geq 0$ :** Generate the sequences such that the following relations hold:

$$\Delta_{x_k}^{h_k}(\tilde{y}_k, x_k) < 0 \text{ with } \tilde{y}_k \in \text{int dom } h \quad (5)$$

$$x_{k+1} = x_k + \eta_k(\tilde{y}_k - x_k) \in \text{int dom } h \quad (6)$$

$$f(x_{k+1}) \leq f(x_k) + \gamma \eta_k \Delta_{x_k}^{h_k}(\tilde{y}_k, x_k) \quad (7)$$

If (5) cannot be satisfied, then the algorithm terminates.

The algorithm starts with a feasible point<sup>1</sup>  $x_0$ . At each iteration, it computes a point  $\tilde{y}_k$  that satisfies (5), which is an inexact solution of the *Bregman proximal mapping*

$$\tilde{x}_k = P_{f_{x_k}}^{h_k}(x_k) := \underset{x \in \mathbb{R}^N}{\operatorname{argmin}} f_{x_k}(x) + D_{h_k}(x, x_k) \quad (8)$$

that, at least, improves the (model) value compared to  $x_k$ . Thanks to the class of Legendre functions  $\mathcal{L}$ , this proximal mapping is well-defined and single-valued on its domain. The exact version of the algorithm solves the proximal mapping exactly for the global optimal solution. The optimal solution of the proximal mapping will always be denoted by  $\tilde{x}_k$  instead of  $\tilde{y}_k$ , which refers to an approximate solution. The direction  $\tilde{y}_k - x_k$  can be considered as a descent direction for the function  $f$ . Given this direction, the goal of (6) and (7) is the

---

<sup>1</sup>It is often easy to find a feasible point. Of course, there are cases, where finding an initialization is a problem itself. We assume that the user provides a feasible initial point.

estimation of a step size  $\eta_k$  (by line search, cf. Algorithm 2) that reduces the value of the objective function. In case that the proximal mapping has a solution but the first relation (5) can only be satisfied with equality, we will see that  $x_k = \tilde{y}_k$  must be a stationary point of the objective, hence, the algorithm terminates.

**Remark 12.** Instead of performing backtracking on the objective values as in Algorithm 1, backtracking on the scaling of the Bregman distance in (5) is also possible. For a special model function, this leads to ProxDescent [25, 20] (with Euclidean proximity function). If a scaled version of (5) yields a descent on  $f$ , we can set  $\eta_k = 1$ , and accept this point. However, this can be expensive when the proximal subproblem in (5) is hard to solve, since each trial step requires to solve the subproblem. In order to break the backtracking, the new objective value must be computed anyway. Therefore, a computational advantage of the line search (6) and (7) is to be expected (cf. Section 6.1).

**Algorithm 2 (Line Search for Algorithm 1).**

- **Basic prerequisites:** Fix  $\delta, \gamma \in ]0, 1[$ ,  $\tilde{\eta} > 0$ , and  $k \in \mathbb{N}$ .
- **Input:** Current iterates  $x_k \in \text{int dom } h$  and  $\tilde{y}_k$  satisfy (5).
- **Solve:** Find the smallest  $j \in \mathbb{N}$  such that  $\tilde{\eta}_j := \tilde{\eta}\delta^j$  satisfies (6) and (7).
- **Return:** Set the feasible step size  $\eta_k$  for iteration  $k$  to  $\tilde{\eta}_j$ .

Algorithm 1–2 is well defined as the following lemmas show.

**Lemma 13 (Well-definedness).** Let  $\omega$  in (3) be a growth function. Algorithm 1 is well-defined, i.e., for all  $k \in \mathbb{N}$ , the following holds:

- (i) there exists  $\tilde{y}_k$  that satisfies (5) or  $x_k = \tilde{x}_k$  and the algorithm terminates;
- (ii)  $x_k \in \text{dom } f \cap \text{int dom } h$ ; and
- (iii) there exists  $\eta_k$  that satisfies (6) and (7).

*Proof.* (i) For  $x_k \in \text{int dom } h$ , Lemma 10 shows that  $P_{f_{x_k}}^{h_k}$  maps to  $\text{int dom } h_k \cap \text{dom } f_{x_k} \subset \text{int dom } h \cap \text{dom } f$  and is single-valued. Thus, for example,  $\tilde{y}_k = \tilde{x}_k$  satisfies (5). Otherwise,  $x_k = \tilde{x}_k$ , which shows (i). (ii) Since  $x_0 \in \text{dom } f \cap \text{int dom } h$  and  $f(x_{k+1}) \leq f(x_k)$  by (7) it holds that  $x_k \in \text{dom } f$  for all  $k$ . Since  $x_k \in \text{int dom } h$  and  $\tilde{y}_k \in \text{dom } h$ , for small  $\eta_k$  also  $x_{k+1} \in \text{int dom } h$ , hence  $x_{k+1} \in \text{dom } f \cap \text{int dom } h$ . Inductively, we conclude the statement. (iii) This will be shown in Lemma 14. □

**Lemma 14 (Finite termination of Algorithm 2).** Consider Algorithm 1 and fix  $k \in \mathbb{N}$ . Let  $\omega$  in (3) be a growth function. Let  $\delta, \gamma \in ]0, 1[$ ,  $\tilde{\eta} > 0$ ,  $\bar{h} := h_k$ , and  $\bar{x} := x_k$ ,  $\tilde{y} := \tilde{y}_k$  be such that  $\Delta_{\bar{x}}^{\bar{h}}(\tilde{y}, \bar{x}) < 0$ . Then, there exists  $j \in \mathbb{N}$  such that  $\tilde{\eta}_j := \tilde{\eta}\delta^j$  satisfies

$$f(\bar{x} + \tilde{\eta}_j(\tilde{y} - \bar{x})) \leq f(\bar{x}) + \gamma\tilde{\eta}_j\Delta_{\bar{x}}^{\bar{h}}(\tilde{y}, \bar{x}).$$

*Proof.* This result is proved by contradiction. Define  $v := \tilde{y} - \bar{x}$ . By our assumption in (3), we observe that

$$f(\bar{x} + \tilde{\eta}_j v) - f(\bar{x}) \leq f_{\bar{x}}(\bar{x} + \tilde{\eta}_j v) - f(\bar{x}) + o(\tilde{\eta}_j). \quad (9)$$

Using Jensen's inequality for the convex function  $f_{\bar{x}}$  provides:

$$f_{\bar{x}}(\bar{x} + \tilde{\eta}_j v) - f_{\bar{x}}(\bar{x}) \leq \tilde{\eta}_j f_{\bar{x}}(\bar{x} + v) + (1 - \tilde{\eta}_j) f_{\bar{x}}(\bar{x}) - f_{\bar{x}}(\bar{x}) = \tilde{\eta}_j \cdot \left( f_{\bar{x}}(\bar{x} + v) - f_{\bar{x}}(\bar{x}) \right). \quad (10)$$

Now, suppose  $\gamma \Delta_{\bar{x}}^{\bar{h}}(\tilde{y}, \bar{x}) < \frac{1}{\tilde{\eta}_j} (f(\bar{x} + \tilde{\eta}_j v) - f(\bar{x}))$  holds for any  $j \in \mathbb{N}$ . Then, using (9) and (10), we conclude the following:

$$\begin{aligned} \gamma \Delta_{\bar{x}}^{\bar{h}}(\tilde{y}, \bar{x}) &< f_{\bar{x}}(\bar{x} + v) - f_{\bar{x}}(\bar{x}) + o(\tilde{\eta}_j)/\tilde{\eta}_j \\ &\leq f_{\bar{x}}(\bar{x} + v) - f_{\bar{x}}(\bar{x}) + D_{\bar{h}}(\tilde{y}, \bar{x}) + o(\tilde{\eta}_j)/\tilde{\eta}_j \\ &= (f_{\bar{x}}^{\bar{h}}(\tilde{y}) - f_{\bar{x}}^{\bar{h}}(\bar{x})) + o(\tilde{\eta}_j)/\tilde{\eta}_j = \Delta_{\bar{x}}^{\bar{h}}(\tilde{y}, \bar{x}) + o(\tilde{\eta}_j)/\tilde{\eta}_j, \end{aligned}$$

which for  $j \rightarrow \infty$  yields the desired contradiction, since  $\gamma \in ]0, 1[$  and  $\Delta_{\bar{x}}^{\bar{h}}(\tilde{y}, \bar{x}) < 0$ .  $\square$

## 4.2 Finite Time Convergence Analysis

First, we study the case when the algorithm terminates after a finite number of iterations, i.e., there exists  $k_0 \in \mathbb{N}$  such that (5) cannot be satisfied. Then, the point  $\tilde{y}_{k_0}$  is a global minimizer of  $f_{x_{k_0}}^{h_{k_0}}$  and  $\Delta_{x_{k_0}}^{h_{k_0}}(\tilde{y}_{k_0}, x_{k_0}) = 0$ . Moreover, the point  $x_{k_0}$  turns out to be a stationary point of  $f$ .

**Lemma 15.** For  $\bar{x} \in \text{dom } f$  and a model  $f_{\bar{x}}$  that satisfies (3), where  $\omega$  is a growth function, the following holds:

$$|\nabla f_{\bar{x}}|(\bar{x}) = |\nabla f|(\bar{x}).$$

*Proof.* Since  $\omega(0) = 0$ , we have from (3) that  $f_{\bar{x}}(\bar{x}) = f(\bar{x})$ . This, together with sub-additivity of  $[\cdot]_+$ , entails

$$\begin{aligned} \frac{[f_{\bar{x}}(\bar{x}) - f_{\bar{x}}(x)]_+}{|x - \bar{x}|} &\leq \frac{[f(\bar{x}) - f(x)]_+ + [f(x) - f_{\bar{x}}(x)]_+}{|x - \bar{x}|} \leq \frac{[f(\bar{x}) - f(x)]_+}{|x - \bar{x}|} + \frac{|f(x) - f_{\bar{x}}(x)|}{|x - \bar{x}|} \\ &\leq \frac{[f(\bar{x}) - f(x)]_+}{|x - \bar{x}|} + \frac{\omega(|x - \bar{x}|)}{|x - \bar{x}|} \end{aligned}$$

Passing to the lim sup on both sides and using that  $\omega \in o(t)$ , we get

$$|\nabla f_{\bar{x}}|(\bar{x}) \leq |\nabla f|(\bar{x}).$$

Arguing similarly but now starting with  $|\nabla f|(\bar{x})$ , we get the reverse inequality, which in turn shows the claimed equality.  $\square$

**Proposition 16 (Stationarity for finite time termination).** Consider the setting of Algorithm 1. Let  $\omega$  in (3) be a growth function. Let  $k_0 \in \mathbb{N}$  be fixed, and set  $\tilde{x} = \tilde{y}_{k_0}$ ,  $\bar{x} = x_{k_0}$ ,  $\bar{h} = h_{k_0}$ , and  $\bar{x}, \tilde{x} \in \text{dom } f \cap \text{int dom } h$ . If  $\Delta_{\bar{x}}^{\bar{h}}(\tilde{x}, \bar{x}) \geq 0$ , then  $\tilde{x} = \bar{x}$ ,  $\Delta_{\bar{x}}^{\bar{h}}(\tilde{x}, \bar{x}) = 0$ , and  $|\nabla f|(\bar{x}) = 0$ , i.e.  $\bar{x}$  is a stationary point of  $f$ .

*Proof.* Since  $\tilde{x}$  is the unique solution of the proximal mapping, obviously  $\Delta_{\tilde{x}}^{\bar{h}}(\tilde{x}, \bar{x}) = 0$  and  $\tilde{x} = \bar{x}$ . Moreover,  $\tilde{x}$  is the minimizer of  $f_{\tilde{x}}^{\bar{h}}$ , i.e. we have

$$0 = |\nabla f_{\tilde{x}}^{\bar{h}}|(\tilde{x}) = |\nabla f_{\bar{x}}^{\bar{h}}|(\bar{x}) = \limsup_{\substack{x \rightarrow \bar{x} \\ x \neq \bar{x}}} \frac{[f_{\bar{x}}(\bar{x}) - f_{\bar{x}}(x) - D_{\bar{h}}(x, \bar{x})]_+}{|x - \bar{x}|} = |\nabla f_{\bar{x}}|(\bar{x}) = |\nabla f|(\bar{x}),$$

where we used that  $\bar{h}$  is Fréchet differentiable at  $\bar{x}$  and Lemma 15. □

### 4.3 Asymptotic Convergence Analysis

We have established stationarity of the algorithm's output, when it terminates after a finite number of iterations. Therefore, without loss of generality, we now focus on the case where (5) can be satisfied for all  $k \in \mathbb{N}$ . We need to make the following assumptions.

**Assumption 1.** The sequence  $(\tilde{y}_k)_{k \in \mathbb{N}}$  satisfies  $f_{x_k}^{h_k}(\tilde{y}_k) \leq \inf f_{x_k}^{h_k} + \varepsilon_k$  for some  $\varepsilon_k \rightarrow 0$ .

**Remark 17.** Assumption 1 states that asymptotically (for  $k \rightarrow \infty$ ) the Bregman proximal mapping (8) must be solved accurately. In order to obtain stationarity of a limit point, Assumption 1 is necessary, as shown by Bonettini et al. [11, after Theorem 4.1] for a special setting of model functions.

**Assumption 2.** Let  $h \in \mathcal{L}$ . For every bounded sequences  $(x_k)_{k \in \mathbb{N}}$  and  $(\bar{x}_k)_{k \in \mathbb{N}}$  in  $\text{int dom } h$ , and  $(h_k)_{k \in \mathbb{N}}$  such that  $h_k \succeq h$ , it is assumed that:

$$x_k - \bar{x}_k \rightarrow 0 \quad \Leftrightarrow \quad D_{h_k}(x_k, \bar{x}_k) \rightarrow 0.$$

**Remark 18.** (i) Assumption 2 states that (asymptotically) a vanishing Bregman distance reflects a vanishing Euclidean distance. This is a natural assumption and satisfied, e.g., by most entropies such as Boltzmann–Shannon, Fermi–Dirac, and Burg entropy.

(ii) The equivalence in Assumption 2 is satisfied, for example, when there exists  $c \in \mathbb{R}$  such that  $ch \succeq h_k$  holds for all  $k \in \mathbb{N}$  and the following holds:

$$x_k - \bar{x}_k \rightarrow 0 \quad \Leftrightarrow \quad D_h(x_k, \bar{x}_k) \rightarrow 0.$$

**Proposition 19 (Convergence of objective values).** Consider the setting of Algorithm 1. Let  $\omega$  in (3) be a growth function. The sequence of objective values  $(f(x_k))_{k \in \mathbb{N}}$  is non-increasing and converging to some  $f^* \geq \underline{f} > -\infty$ .

*Proof.* This statement is a consequence of (7) and (5), and the lower-boundedness of  $f$ . □

Asymptotically, under some condition on the step size, the improvement of the model objective value between  $\tilde{y}_k$  and  $x_k$  must tend to zero. Since we do not assume that the step sizes  $\eta_k$  are bounded away from zero, this is a non-trivial result.

**Proposition 20 (Vanishing model improvement).** Consider the setting of Algorithm 1. Let  $\omega$  in (3) be a growth function. Suppose, either<sup>2</sup>  $\inf_k \eta_k > 0$  or  $\eta_k$  is selected by the Line Search Algorithm 2. Then,

$$\sum_{k=0}^{\infty} \eta_k (-\Delta_{x_k}^{h_k}(\tilde{y}_k, x_k)) < +\infty \quad \text{and} \quad \Delta_{x_k}^{h_k}(\tilde{y}_k, x_k) \rightarrow 0 \quad \text{as } k \rightarrow \infty.$$

*Proof.* The first part follows from rearranging (7), and summing both sides for  $k = 0, \dots, n$ :

$$\gamma \sum_{k=0}^n \eta_k (-\Delta_{x_k}^{h_k}(\tilde{y}_k, x_k)) \leq \sum_{k=0}^n (f(x_k) - f(x_{k+1})) = f(x_0) - f(x_{n+1}) \leq f(x_0) - f^*.$$

In the remainder of the proof, we show that  $\Delta_{x_k}^{h_k}(\tilde{y}_k, x_k) \rightarrow 0$ , which is not obvious unless  $\inf_k \eta_k > 0$ . The model improvement is bounded. Boundedness from above is satisfied by construction of the sequence  $(\tilde{y}^k)_{k \in \mathbb{N}}$ . Boundedness from below follows from the following observation and the uniform boundedness of the model functions from below:

$$\Delta_{x_k}^{h_k}(\tilde{y}_k, x_k) = f_{x_k}^{h_k}(\tilde{y}_k) - f_{x_k}^{h_k}(x_k) \geq f_{x_k}^{h_k}(\tilde{x}_k) - f_{x_k}^{h_k}(x_k) \geq f_{x_k}(\tilde{x}_k) - f(x_0).$$

Therefore, there exists  $K \subset \mathbb{N}$  such that the subsequence  $\Delta_{x_k}^{h_k}(\tilde{y}_k, x_k)$  converges to some  $\Delta^*$  as  $k \xrightarrow{K} \infty$ . Suppose  $\Delta^* < 0$ . Then, the first part of the statement implies that the step size sequence must tend to zero, i.e.,  $\eta_k \rightarrow 0$  for  $k \xrightarrow{K} \infty$ . For  $k \in K$  sufficiently large, the line search procedure in Algorithm 2 reduces the step length from  $\eta_k/\delta$  to  $\eta_k$ . (Note that  $\eta_k$  can be assumed to be the “first” step length that achieves a reduction in (7)). Before multiplying with  $\delta$ , no descent of (7) was observed, i.e.,

$$(\eta_k/\delta)\gamma\Delta_{x_k}^{h_k}(\tilde{y}_k, x_k) < f(x_k + (\eta_k/\delta)v_k) - f(x_k),$$

where  $v_k = \tilde{y}_k - x_k$ . Using (9) and (10), we can make the same observation as in the proof of Lemma 14:

$$\begin{aligned} \gamma\Delta_{x_k}^{h_k}(\tilde{y}_k, x_k) &< f_{x_k}(x_k + v) - f_{x_k}(x_k) + o(\eta_k/\delta)/(\eta_k/\delta) \\ &\leq f_{x_k}(x_k + v) - f_{x_k}(x_k) + D_{h_k}(\tilde{y}_k, x_k) + o(\eta_k)/\eta_k \\ &= (f_{x_k}^{h_k}(\tilde{y}_k) - f_{x_k}^{h_k}(x_k)) + o(\eta_k)/\eta_k \\ &= \Delta_{x_k}^{h_k}(\tilde{y}_k, x_k) + o(\eta_k)/\eta_k, \end{aligned}$$

which for  $\eta_k \rightarrow 0$  yields a contradiction, since  $\gamma \in ]0, 1[$  and  $\Delta_{x_k}^{h_k}(\tilde{y}_k, x_k) < 0$ . Therefore, any cluster point  $\Delta^*$  of  $(\Delta_{x_k}^{h_k}(\tilde{y}_k, x_k))_{k \in K}$  must be 0, which concludes the proof.  $\square$

### 4.3.1 Asymptotic Stationarity with a Growth Function

In order to establish stationarity of limit points generated by Algorithm 1 additional assumptions are required. We consider three different settings for the model assumption (3):  $\omega$  in the model assumption (3) is a growth function (this section),  $\omega$  is a proper growth function (Section 4.3.2), and  $\omega$  is global growth function of the form  $\omega = D_h$  (Section 4.3.3).

---

<sup>2</sup>Note that  $\inf_k \eta_k > 0$  is equivalent to  $\liminf_k \eta_k > 0$  as we assume  $\eta_k > 0$  for all  $k \in \mathbb{N}$ .

**Assumption 3.** Let  $x^*$  be a limit point of  $(x_k)_{k \in \mathbb{N}}$  and  $x_k \xrightarrow{f} x^*$  as  $k \xrightarrow{K} \infty$  with  $K \subset \mathbb{N}$ . Then

$$|\nabla f_{x_k}|(x_k) = |\nabla f|(x_k) \rightarrow 0 \quad \text{as } k \xrightarrow{K} \infty.$$

**Remark 21.** Assumption 3 is common for abstract algorithms. Attouch et al. [2], for example, use a relative error condition of the form  $|\nabla f|(x_{k+1}) \leq b|x_{k+1} - x_k|$ ,  $b \in \mathbb{R}$ . A weaker sufficient condition for Assumption 3 is  $|\nabla f|(x_{k+1}) \leq \psi(|x_{k+1} - x_k|)$  for some continuous function  $\psi: \mathbb{R}_+ \rightarrow \mathbb{R}_+$  with  $\psi(0) = 0$ ; See Corollary 25. See also Remark 24. For explicit examples, we refer to Section 5.1.

Using this assumption, we can state one of our main theorems, which shows convergence to a stationary point under various condition. The conditions are easily verified in many applications (see Section 5).

**Theorem 22 (Asymptotic stationarity with a growth function).** Consider the setting of Algorithm 1. Let  $\omega$  in (3) be a growth function. Moreover, let either  $\inf_k \eta_k > 0$  or  $\eta_k$  be selected by the Line Search Algorithm 2. Let  $(x_k)_{k \in \mathbb{N}}$  and  $(\tilde{y}_k)_{k \in \mathbb{N}}$  be bounded sequences such that Assumptions 1 and 2 hold and let  $f_{x_k}$  obey (3) with growth function  $\omega$ . Then,  $x_k - \tilde{y}_k \rightarrow 0$  and for  $\tilde{x}_k = P_{f_{x_k}}^{h_k}(x_k)$ , it holds that  $x_k - \tilde{x}_k \rightarrow 0$  and  $\tilde{x}_k - \tilde{y}_k \rightarrow 0$ . Moreover,  $f(x_k) - f(\tilde{y}_k) \rightarrow 0$  and  $f(\tilde{x}_k) - f(x_k) \rightarrow 0$  as  $k \rightarrow \infty$ . Suppose Assumption 3 is satisfied. If  $x^*$  is a limit point of the sequence  $(x_k)_{k \in \mathbb{N}}$ , and one of the following conditions is satisfied:

- (i)  $f$  is continuous on the closure of  $\text{dom } h$ ,
- (ii)  $x^* \in \text{int dom } h$ ,
- (iii)  $x^* \in \text{dom } h$  and  $D_{h_k}(x^*, \tilde{y}_k) \rightarrow 0$  as  $k \xrightarrow{K} \infty$ ,
- (iv)  $x^* \in \text{cl dom } h$  and
  - for all  $x \in \text{int dom } h \cap \text{dom } f$  holds that  $D_{h_k}(x, \tilde{x}_k) - D_{h_k}(x, x_k) \rightarrow 0$  as  $k \xrightarrow{K} \infty$ ,
  - and for all  $x \in \text{dom } f$  the model functions obey  $f_{x_k}(x) \rightarrow f_{x^*}(x)$  as  $k \xrightarrow{K} \infty$ ,

then  $x^*$  is a stationary point of  $f$ .

*Proof.* First, we show that for  $k \rightarrow \infty$  the pairwise distances between the sequences  $(x_k)_{k \in \mathbb{N}}$ ,  $(\tilde{y}_k)_{k \in \mathbb{N}}$ , and  $(\tilde{x}_k)_{k \in \mathbb{N}}$  vanishes. Proposition 11, reformulated in our notation, can be stated as

$$\Delta_{x_k}^{h_k}(x, \tilde{x}_k) = f_{x_k}^{h_k}(x) - f_{x_k}^{h_k}(\tilde{x}_k) \geq D_{h_k}(x, \tilde{x}_k), \quad \forall x \in \text{dom } f. \quad (11)$$

As a direct consequence, using  $x = \tilde{y}_k$  together with Assumptions 1 and 2, we obtain

$$D_{h_k}(\tilde{y}_k, \tilde{x}_k) \rightarrow 0 \quad \text{thus} \quad \tilde{x}_k - \tilde{y}_k \rightarrow 0.$$

Moreover, from Proposition 20, we have  $\Delta_{x_k}^{h_k}(\tilde{y}_k, x_k) \rightarrow 0$ , and from

$$\Delta_{x_k}^{h_k}(\tilde{y}_k, x_k) = \Delta_{x_k}^{h_k}(\tilde{y}_k, \tilde{x}_k) - \Delta_{x_k}^{h_k}(x_k, \tilde{x}_k) \leq \Delta_{x_k}^{h_k}(\tilde{y}_k, \tilde{x}_k) - D_{h_k}(x_k, \tilde{x}_k), \quad (12)$$

and Assumptions 1 and 2, we conclude that  $x_k - \tilde{x}_k \rightarrow 0$ , hence also  $x_k - \tilde{y}_k \rightarrow 0$ .

The next step is to show that  $f(x_k) - f(\tilde{y}_k) \rightarrow 0$  as  $k \rightarrow \infty$ . This follows from the following estimation:

$$\begin{aligned} |f(x_k) - f(\tilde{y}_k)| &\leq |f_{x_k}(x_k) - f_{x_k}(\tilde{y}_k)| + \omega(|\tilde{y}_k - x_k|) \\ &\leq |f_{x_k}^{h_k}(x_k) - f_{x_k}^{h_k}(\tilde{y}_k)| + D_{h_k}(\tilde{y}_k, x_k) + \omega(|\tilde{y}_k - x_k|) \\ &= |\Delta_{x_k}^{h_k}(x_k, \tilde{y}_k)| + D_{h_k}(\tilde{y}_k, x_k) + \omega(|\tilde{y}_k - x_k|), \end{aligned} \quad (13)$$

where the right hand side vanishes for  $k \rightarrow \infty$ . Analogously, we can show that  $f(\tilde{x}_k) - f(x_k) \rightarrow 0$  as  $k \rightarrow \infty$ .

Let  $x^*$  be the limit point of the subsequence  $(x_k)_{k \in K}$  for some  $K \subset \mathbb{N}$ . The remainder of the proof shows that  $f(\tilde{y}_k) \rightarrow f(x^*)$  as  $k \rightarrow \infty$ . Then  $f(x_k) - f(\tilde{y}_k) \rightarrow 0$  implies that  $x_k \xrightarrow{f} x^*$  as  $k \xrightarrow{K} \infty$ , and by Assumption 3, the slope vanishes, hence the limiting slope  $|\nabla f|(x^*)$  at  $x^*$  also vanishes, which concludes the proof.

(i) implies  $f(\tilde{y}_k) \rightarrow f(x^*)$  as  $k \rightarrow \infty$  by definition. For (ii) and (iii), we make the following observation:

$$f(\tilde{y}_k) - \omega(|\tilde{y}_k - x_k|) \leq f_{x_k}^{h_k}(\tilde{y}_k) = f_{x_k}^{h_k}(\tilde{x}_k) + (f_{x_k}^{h_k}(\tilde{y}_k) - f_{x_k}^{h_k}(\tilde{x}_k)) \leq f_{x_k}^{h_k}(x^*) + \Delta_{x_k}^{h_k}(\tilde{y}_k, \tilde{x}_k), \quad (14)$$

where  $\tilde{x}_k = P_{f_{x_k}^{h_k}}(x_k)$ . Taking “ $\limsup_{k \xrightarrow{K} \infty}$ ” on both sides,  $D_{h_k}(x^*, x_k) \rightarrow 0$  (Assumption 2 for (ii) or the assumption in (iii)), and  $\Delta_{x_k}^{h_k}(\tilde{y}_k, \tilde{x}_k) \rightarrow 0$  (Assumption 1) shows that  $\limsup_{k \xrightarrow{K} \infty} f(\tilde{y}_k) \leq f(x^*)$ . Since  $f$  is closed,  $f(\tilde{y}_k) \rightarrow f(x^*)$  holds.

We consider (iv). For all  $x \in \text{int dom } h \cap \text{dom } f$ , we have (11) or, reformulated,  $f_{x_k}^{h_k}(x) - D_{h_k}(x, \tilde{x}_k) \geq f_{x_k}^{h_k}(\tilde{x}_k)$ , which implies the following:

$$f_{x_k}(x) + D_{h_k}(x, x_k) - D_{h_k}(x, \tilde{x}_k) - D_{h_k}(\tilde{x}_k, x_k) \geq f(\tilde{x}_k) - \omega(|\tilde{x}_k - x_k|).$$

Note that for any  $x$  the limits for  $k \xrightarrow{K} \infty$  on the left hand side exist. In particular, we have

$$D_{h_k}(x, x_k) - D_{h_k}(x, \tilde{x}_k) - D_{h_k}(\tilde{x}_k, x_k) \rightarrow 0 \quad \text{as } k \xrightarrow{K} \infty,$$

by the assumption in (iv), and Assumption 2 together with  $\tilde{x}_k - x_k \rightarrow 0$ . The limit of  $f_{x_k}(x)$  exists by assumption and coincides with  $f_{x^*}(x)$ . Choosing a sequence  $(z_k)_{k \in \mathbb{N}}$  in  $\text{int dom } h \cap \text{dom } f$  with  $z_k \rightarrow x^*$  as  $k \xrightarrow{K} \infty$ , in the limit, we obtain

$$f(x^*) \geq \lim_{k \xrightarrow{K} \infty} f(\tilde{x}_k) =: f^*,$$

since  $f_{x^*}(z_k) \rightarrow f_{x^*}(x^*) = f(x^*)$  for  $z_k \rightarrow x^*$  as  $k \xrightarrow{K} \infty$ . Invoking that  $f$  is closed, we conclude the  $f$ -attentive convergence  $f^* = \liminf_{k \rightarrow \infty} f(x_k) \geq f(x^*) \geq f^*$ .  $\square$

**Remark 23.** Existence of a limit point  $x^*$  is guaranteed by assuming that  $(x_k)_{k \in \mathbb{N}}$  is bounded. Alternatively, we could require that  $f$  is coercive (i.e.  $f(x_k) \rightarrow \infty$  for  $|x_k| \rightarrow \infty$ ), which implies boundedness of the lower level sets of  $f$ , hence by Proposition 19 the boundedness of  $(x_k)_{k \in \mathbb{N}}$ .

**Remark 24.** From Theorem 22, clearly, also  $\tilde{y}_k \xrightarrow{f} x^*$  and  $\tilde{x}_k \xrightarrow{f} x^*$  as  $k \xrightarrow{K} \infty$  holds. Therefore, Assumption 3 could also be stated as the requirement

$$|\nabla f|(\tilde{x}_k) \rightarrow 0 \quad \text{or} \quad |\nabla f|(\tilde{y}_k) \rightarrow 0 \quad \text{as } k \xrightarrow{K} \infty,$$

in order to conclude that limit points of  $(x_k)_{k \in \mathbb{N}}$  are stationary points.

As a simple corollary of this theorem, we replace Assumption 3 with the relative error condition mentioned in Remark 21.

**Corollary 25 (Asymptotic stationarity with a growth function).** Consider the setting of Algorithm 1. Let  $\omega$  in (3) be a growth function. Moreover, let either  $\inf_k \eta_k > 0$  or  $\eta_k$  be selected by the Line Search Algorithm 2. Let  $(x_k)_{k \in \mathbb{N}}$  and  $(\tilde{y}_k)_{k \in \mathbb{N}}$  be bounded sequences such that Assumptions 1 and 2 hold and let  $f_{x_k}$  obey (3) with growth function  $\omega$ . Suppose there exists a continuous function  $\psi: \mathbb{R}_+ \rightarrow \mathbb{R}_+$  with  $\psi(0) = 0$  such that  $|\nabla f|(x_{k+1}) \leq \psi(|x_{k+1} - x_k|)$  is satisfied. If  $x^*$  is a limit point of the sequence  $(x_k)_{k \in \mathbb{N}}$  and one of the conditions (i)–(iv) in Theorem 22 is satisfied, then  $x^*$  is a stationary point of  $f$ .

*Proof.* Theorem 22 shows that  $\tilde{y}_k - x_k \rightarrow 0$ , thus,  $\inf_k \eta_k > 0$  implies  $x_{k+1} - x_k \rightarrow 0$  by (6). Therefore, the relation  $|\nabla f|(x_{k+1}) \leq \psi(|x_{k+1} - x_k|)$  shows that Assumption 3 is automatically satisfied and we can apply Theorem 22 to deduce the statement.  $\square$

**Some more results on the limit point set.** In Theorem 22 we have shown that limit points of the sequence  $(x_k)_{k \in \mathbb{N}}$  generated by Algorithm 1 are stationary, and in fact the sequence  $f$ -converges to its limit points. The following proposition shows some more properties of the set of limit points of  $(x_k)_{k \in \mathbb{N}}$ . This is a well-known result [10, Lem. 5] that follows from  $x_{k+1} - x_k \rightarrow 0$  as  $k \rightarrow \infty$ .

**Proposition 26.** Consider the setting of Algorithm 1. Let  $\omega$  in (3) be a growth function and  $\inf_k \eta_k > 0$ . Let  $(x_k)_{k \in \mathbb{N}}$  and  $(\tilde{y}_k)_{k \in \mathbb{N}}$  be bounded sequences such that Assumptions 1, 2 and 3 hold. Suppose one of the conditions (i)–(iv) in Theorem 22 is satisfied for each limit point of  $(x_k)_{k \in \mathbb{N}}$ . Then, the set  $\mathfrak{S} := \left\{ x^* \in \mathbb{R}^N \mid \exists K \subset \mathbb{N}: x_k \rightarrow x^* \text{ as } k \xrightarrow{K} \infty \right\}$  of limit points of  $(x_k)_{k \in \mathbb{N}}$  is connected, each point  $x^* \in \mathfrak{S}$  is stationary for  $f$ , and  $f$  is constant on  $\mathfrak{S}$ .

*Proof.* Theorem 22 shows that  $\tilde{y}_k - x_k \rightarrow 0$ . Thus, boundedness of  $\eta_k$  away from 0 implies  $x_{k+1} - x_k \rightarrow 0$  by (6). Now, the statement follows from [10, Lem. 5] and Theorem 22.  $\square$

### 4.3.2 Asymptotic Stationarity with a Proper Growth Function

Our proof of stationarity of limit points generated by Algorithm 1 under the assumption of a proper growth function  $\omega$  in (3) relies on an adaptation of a recently proved result by



Drusvyatskiy et al. [19, Corollary 5.3], which is stated in Lemma 27 before the main theorem of this subsection. The credits for this lemma should go to [19].

**Lemma 27 (Perturbation result under approximate optimality).** Let  $f: \mathbb{R}^N \rightarrow \overline{\mathbb{R}}$  be a proper closed function. Consider bounded sequences  $(x_k)_{k \in \mathbb{N}}$  and  $(\tilde{y}_k)_{k \in \mathbb{N}}$  with  $x_k - \tilde{y}_k \rightarrow 0$  for  $k \rightarrow \infty$ , and model functions  $f_{x_k}^{h_k}$  according to (3) with proper growth functions. Suppose Assumption 1 and 2 hold. If  $(x^*, f(x^*))$  is a limit point of  $(x_k, f(x_k))_{k \in \mathbb{N}}$ , then  $x^*$  is stationary for  $f$ .

*Proof.* Recall  $\varepsilon_k$  from Assumption 1. Theorem 5.1 from [19] guarantees, for each  $k$  and any  $\rho_k > 0$ , the existence of points  $\hat{y}_k$  and  $z_k$  such that the following hold:

(i) (point proximity)

$$|\tilde{y}_k - z_k| \leq \frac{\varepsilon_k}{\rho_k} \quad \text{and} \quad |z_k - \hat{y}_k| \leq 2 \cdot \frac{\omega(|z_k - x_k|)}{\omega'_+(|z_k - x_k|)},$$

under the convention  $\frac{0}{0} = 0$ ,

(ii) (value proximity)  $f(\hat{y}_k) \leq f(\tilde{y}_k) + 2\omega(|z_k - x_k|) + \omega(|\tilde{y}_k - x_k|)$ , and

(iii) (near-stationarity)  $|\nabla f|(\hat{y}) \leq \rho_k + \omega'_+(|z_k - x_k|) + \omega'_+(\hat{y}_k - x_k)$ ,

Setting  $\rho_k = \sqrt{\varepsilon_k}$ , using  $\varepsilon_k \rightarrow 0$  and the point proximity, shows that  $|\tilde{y}_k - z_k| \rightarrow 0$ . Moreover  $|z_k - x_k| \leq |z_k - \tilde{y}_k| + |\tilde{y}_k - x_k| \rightarrow 0$ , which implies that  $|z_k - \hat{y}_k| \rightarrow 0$ . Now, we fix a convergent subsequence  $(x_k, f(x_k)) \rightarrow (x^*, f(x^*))$  as  $k \xrightarrow{K} \infty$  for some  $K \subset \mathbb{N}$ . Using (11), we observe  $\tilde{x}_k - \tilde{y}_k \rightarrow 0$ , hence  $x_k - \tilde{x}_k \rightarrow 0$ . From Proposition 20 and Assumption 1, we conclude that  $\Delta_{x_k}^{h_k}(\tilde{y}_k, x_k) \rightarrow 0$ , and, therefore  $f(x_k) - f(\tilde{y}_k) \rightarrow 0$  using (13). Consider the value proximity. Combined with the lower semi-continuity of  $f$ , it yields

$$f(x^*) \leq \liminf_{k \xrightarrow{K} \infty} f(\hat{y}_k) \leq \limsup_{k \xrightarrow{K} \infty} f(\hat{y}_k) \leq \limsup_{k \xrightarrow{K} \infty} f(\tilde{y}_k) \leq f(x^*),$$

hence  $(\hat{y}_k, f(\hat{y}_k)) \rightarrow (x^*, f(x^*))$  as  $k \xrightarrow{K} \infty$ . Near-stationarity implies that  $|\nabla f|(\hat{y}_k) \rightarrow 0$ , which proves that  $|\overline{\nabla f}|(x^*) = 0$ , hence  $x^*$  is a stationary point.  $\square$

**Remark 28.** The setting in [19, Corollary 5.3] is recovered when  $(x_k)_{k \in \mathbb{N}}$  is given by  $x_{k+1} = \tilde{y}_k$ .

**Theorem 29 (Asymptotic stationarity with a proper growth function).** Consider the setting of Algorithm 1. Let  $\omega$  in (3) be a proper growth function. Moreover, let either  $\inf_k \eta_k > 0$  or  $\eta_k$  be selected by the Line Search Algorithm 2. Let  $(x_k)_{k \in \mathbb{N}}$  and  $(\tilde{y}_k)_{k \in \mathbb{N}}$  be bounded sequences such that Assumptions 1 and 2 hold. If  $x^*$  is a limit point of the sequence  $x_k$  and one of the conditions (i)–(iv) in Theorem 22 is satisfied, then  $x^*$  is a stationary point of  $f$ .

*Proof.* Propositions 19 and 20, and the proof of  $f$ -attentive convergence from Theorem 22 only rely on a growth function. Instead of assuming that the slope vanishes, here we apply Lemma 27 to conclude stationarity of the limit points.  $\square$

Of course, Proposition 26 can also be stated in the context here.

### 4.3.3 Asymptotic Analysis with a Global Growth Function

Suppose, for  $\bar{x} \in \text{int dom } h$  for some  $h \in \mathcal{L}$ , the model error can be estimated as follows:

$$|f(x) - f_{\bar{x}}(x)| \leq LD_h(x, \bar{x}) \quad \forall x. \quad (15)$$

Since  $h$  is Fréchet differentiable on  $\text{int dom } h$ , the right hand side is bounded by a growth function. Without loss of generality, we restrict ourselves to a fixed function  $h \in \mathcal{L}$  (this section analyses a single iteration). In order to reveal similarities to well-known step size rules, we scale  $h$  in the definition of  $f_{\bar{x}}^h$  to  $D_{h/\alpha} = \frac{1}{\alpha}D_h$  with  $\alpha > 0$  instead of  $D_h$ . Here, decreasing objective values can be assured without the line search procedure (see Proposition 30), i.e.,  $\eta_k = 1$  is always feasible.

In order to obtain the result of stationarity of limit points (Theorem 22 or 29), we can either verify by hand that Assumption 3 holds or we need to assume that  $D_h(x, \bar{x})$  is bounded by a proper growth function.

**Proposition 30.** Consider the setting of Algorithm 1 and let (15) be satisfied.

- (i) For points  $\tilde{y}$  that satisfy  $\Delta_{\bar{x}}^h(\tilde{y}, \bar{x}) < 0$ ,

$$\frac{1 - \alpha L}{\alpha} D_h(\tilde{y}, \bar{x}) \leq f(\bar{x}) - f(\tilde{y})$$

holds, where the left-hand-side is strictly larger than 0 for  $\alpha \in ]0, 1/L[$ .

- (ii) For points  $\tilde{x} = P_{f_{\bar{x}}^h}(\bar{x})$ , the following descent property holds:

$$\frac{1 + \rho - \alpha L}{\alpha} D_h(\tilde{x}, \bar{x}) \leq f(\bar{x}) - f(\tilde{x}),$$

where the left-hand-side is strictly larger than 0 for  $\alpha \in ]0, (1 + \rho)/L[$ , and  $\rho$  is the Bregman symmetry factor defined by  $\rho := \inf \left\{ \frac{D_h(x, \bar{x})}{D_h(\bar{x}, x)} \mid x, \bar{x} \in \text{int dom } h, x \neq \bar{x} \right\}$ ; (see [6]).

*Proof.* The following relations hold:

$$\Delta_{\bar{x}}^h(\tilde{y}, \bar{x}) \leq 0 \quad \Leftrightarrow \quad f_{\bar{x}}^h(\tilde{y}) \leq f_{\bar{x}}^h(\bar{x}) \quad \Leftrightarrow \quad f_{\bar{x}}(\tilde{y}) + \frac{1}{\alpha} D_h(\tilde{y}, \bar{x}) \leq f_{\bar{x}}(\bar{x}) = f(\bar{x}). \quad (16)$$

Bounding the left hand side of the last expression using (15), we obtain

$$f(\tilde{y}) - LD_h(\tilde{y}, \bar{x}) + \frac{1}{\alpha} D_h(\tilde{y}, \bar{x}) \leq f(\bar{x}), \quad (17)$$

which proves part (i). Part (ii) follows analogously. However, thanks to the three point inequality from Proposition 11 and optimality of  $\tilde{x}$  the rightmost inequality of (16) improves to

$$f_{\bar{x}}(\tilde{x}) + \frac{1}{\alpha}D_h(\tilde{x}, \bar{x}) + \frac{1}{\alpha}D_h(\bar{x}, \tilde{x}) \leq f_{\bar{x}}(\bar{x}) = f(\bar{x}),$$

and the statement follows. □

## 4.4 A Remark on Convex Optimization

In this section, let  $f$  be convex, and consider the following global model assumption

$$0 \leq f(x) - f_{\bar{x}}(x) \leq LD_h(x, \bar{x}). \quad (18)$$

We establish a convergence rate of  $\mathcal{O}(1/k)$  for Algorithm 1 with  $\eta_k \equiv 1$ . For Forward–Backward Splitting, this has been shown by Bauschke et al. [6]. We only require  $f_{\bar{x}}$  to be a model w.r.t. (18).

**Proposition 31.** Consider Algorithm 1 with  $\eta_k \equiv 1$  and model functions that obey (18). For  $x_{k+1} = P_{f_{x_k}}^{h/\alpha}(x_k)$  and  $\alpha = \frac{1}{L}$ , the following rate of convergence on the objective values holds:

$$f(x_{k+1}) - f(x) \leq \frac{LD_h(x^*, x_0)}{2k} \quad (= \mathcal{O}(1/k)).$$

*Proof.* The three point inequality in Proposition 11 combined with the model assumption (18) yields the following inequality:

$$f(\tilde{x}) + \frac{1 - \alpha L}{\alpha}D_h(\tilde{x}, \bar{x}) + \frac{1}{\alpha}D_h(x, \tilde{x}) \leq f(x) + \frac{1}{\alpha}D_h(x, \bar{x})$$

for all  $x$ . Restricting to  $0 < \alpha \leq \frac{1}{L}$ , we obtain

$$f(\tilde{x}) - f(x) \leq \frac{1}{\alpha} (D_h(x, \bar{x}) - D_h(x, \tilde{x})). \quad (19)$$

Let  $x^*$  be a minimizer of  $f$ . We make the following choices:

$$x = x^*, \quad \tilde{x} = x_{k+1}, \quad \text{and} \quad \bar{x} = x_k.$$

Summing both sides up to iteration  $k$  and the descent property yield the convergence rate:

$$f(x_{k+1}) - f(x) \leq \frac{D_h(x^*, x_0)}{2\alpha k} \stackrel{\alpha=\frac{1}{L}}{=} \frac{LD_h(x^*, x_0)}{2k}. \quad (20)$$

□

**Example 32.** This section goes beyond the Forward–Backward Splitting setting, e.g., we may specialize Example 35 to convex problems of the form  $\min_x g(F(x))$ , for instance, using  $F(x) = (f_1(x), f_2(x))$  and  $g(z_1, z_2) = \max\{z_1, z_2\}$ .

## 5 Examples

We discuss several classes of problems that can be solved using our framework. To apply Algorithm 1, in Section 5.1, we define a suitable model and mention the associated algorithmic step that arises from exactly minimizing the sum of the model and an Euclidean proximity measure. However our algorithm allows for inexact solutions and very flexible (also iteration dependent) Bregman proximity functions. Examples are provided in Section 5.2. For brevity, we define the symbols  $\Gamma_0$  for the set of proper, closed, convex functions and  $C^1$  for the set of continuously differentiable functions.

### 5.1 Examples of Model Functions

**Example 33 (Forward–Backward Splitting).** Problems of the form

$$f = f_0 + f_1 \quad \text{with } f_0 \in \Gamma_0 \text{ and } f_1 \in C^1$$

can be modeled by

$$f_{\bar{x}}(x) = f_0(x) + f_1(\bar{x}) + \langle x - \bar{x}, \nabla f_1(\bar{x}) \rangle .$$

This model is associated with Forward–Backward Splitting (FBS). We assume that one of the following error models is satisfied:

$$|f(x) - f_{\bar{x}}(x)| = |f_1(x) - f_1(\bar{x}) - \langle x - \bar{x}, \nabla f_1(\bar{x}) \rangle| \leq \begin{cases} \frac{L}{2}|x - \bar{x}|^2, & \text{if } \nabla f_1 \text{ is L-Lipschitz;} \\ \int_0^1 \frac{\varphi(t|x-\bar{x}|)}{t} dt, & \text{if } \nabla f_1 \text{ is } \psi\text{-uniformly} \\ & \text{continuous;} \\ \omega(|x - \bar{x}|), & \text{otherwise,} \end{cases}$$

which is the linearization error of the smooth part  $f_1$ . The first case obeys a global (proper) growth function, derived from the common Descent Lemma. The second case is the generalization to a  $\psi$ -uniformly continuous gradient with  $\varphi(s) = s\psi(s)$  as in Lemma 4. The bound  $\int_0^1 \frac{\varphi(t|x-\bar{x}|)}{t} dt$  is a growth function but not necessarily a proper growth function. The third case is the most general and assumes that the error obeys a growth function. In any case, the model satisfies the model consistency required in Theorem 22(iv). For any  $x \in \text{dom } f$  and  $\bar{x} \rightarrow x^*$ ,

$$|f_0(x) + f_1(\bar{x}) + \langle x - \bar{x}, \nabla f_1(\bar{x}) \rangle - (f_0(x) + f_1(x^*) + \langle x - x^*, \nabla f_1(x^*) \rangle)| \rightarrow 0$$

holds, thanks to the continuous differentiability of  $f_1$  and continuity of the inner product.

In order to verify Assumption 3, we make use of Remark 24 and show that  $|\nabla f|(\tilde{x}_k) \rightarrow 0$  as  $k \xrightarrow{K} \infty$  where  $K \subset \mathbb{N}$  is such that  $x_k \xrightarrow{K} x^*$ . Note that  $\tilde{x}_k$  satisfies the following relation:

$$\begin{aligned} 0 &\in \partial f_0(\tilde{x}_k) + \nabla f_1(x_k) + \nabla h_k(\tilde{x}_k) - \nabla h_k(x_k) \\ &\Rightarrow \nabla f_1(\tilde{x}_k) - \nabla f_1(x_k) + \nabla h_k(x_k) - \nabla h_k(\tilde{x}_k) \in \partial f_0(\tilde{x}_k) + \nabla f_1(\tilde{x}_k) = \partial f(\tilde{x}_k) \end{aligned}$$

Moreover, we know that  $\tilde{x}_k - x_k \rightarrow 0$  as  $k \xrightarrow{K} \infty$ . Since  $\nabla f_1$  is continuous, if  $|\nabla h_k(x_k) - \nabla h_k(\tilde{x}_k)| \rightarrow 0$  for  $k \xrightarrow{K} \infty$ , then Assumption 3/Remark 24 is satisfied. The condition  $|\nabla h_k(x_k) - \nabla h_k(\tilde{x}_k)| \rightarrow 0$  is naturally fulfilled by many Legendre functions, e.g., if  $\nabla h_k$  is  $\psi$ -uniformly continuous (uniformly in  $k$ ) with  $\alpha > 0$  or uniformly continuous (independent of  $k$ ) on bounded sets or continuous at  $x^*$  (uniformly w.r.t.  $k$ ), and will be discussed in more detail in Section 5.2.

**Example 34 (Variable metric FBS).** We consider an extension of Examples 33. An alternative feasible model for a twice continuously differentiable function  $f_1$  is the following:

$$f_{\bar{x}}(x) = f_0(x) + f_1(\bar{x}) + \langle x - \bar{x}, \nabla f_1(\bar{x}) \rangle + \frac{1}{2} \langle x - \bar{x}, B(x - \bar{x}) \rangle ,$$

where  $B := [\nabla^2 f_1(\bar{x})]_+$  is a positive definite approximation to  $\nabla^2 f_1(\bar{x})$ , which leads to a Hessian driven variable metric FBS. It is easy to see that the model error satisfies the growth function  $\omega(s)$ . Again, Theorem 22(iv) obviously holds and the same conclusions about Assumption 3 can be made as in Example 33.

**Example 35 (ProxDescent).** Problems of the form

$$f_0 + g \circ F \quad \text{with } f_0 \in \Gamma_0, F \in C^1, \text{ and } g \in \Gamma_0 \text{ finite-valued,}$$

which often arise from non-linear inverse problems, can be approached by the model function

$$f_{\bar{x}}(x) = f_0(x) + g(F(\bar{x}) + DF(\bar{x})(x - \bar{x})),$$

where  $DF(\bar{x})$  is the Jacobian matrix of  $F$  at  $\bar{x}$ . The associated algorithm is connected to ProxDescent [25, 20]. If  $g$  is a quadratic function, the algorithm reduces to the Levenberg–Marquardt algorithm [27]. The error model can be computed as follows:

$$\begin{aligned} |f(x) - f_{\bar{x}}(x)| &= |g(F(x)) - g(F(\bar{x}) + DF(\bar{x})(x - \bar{x}))| \\ &\leq \ell |F(x) - F(\bar{x}) - DF(\bar{x})(x - \bar{x})| \\ &\leq \begin{cases} \frac{\ell L}{2} |x - \bar{x}|^2, & \text{if } DF \text{ is } L\text{-Lipschitz and } g \text{ is } \ell\text{-Lipschitz;} \\ \ell \int_0^1 \frac{\varphi(t|x-\bar{x}|)}{t} dt, & \text{if } DF \text{ is } \psi\text{-uniform continuous and } g \text{ is } \ell\text{-Lipschitz;} \\ \omega(|x - \bar{x}|), & \text{otherwise,} \end{cases} \end{aligned} \tag{21}$$

where  $\ell$  is the (possibly local) Lipschitz constant of  $g$  around  $F(\bar{x})$ . Since  $g$  is convex and finite-valued, it is always locally Lipschitz continuous. Since  $F$  is continuously differentiable, for  $x$  sufficiently close to  $\bar{x}$ , both  $F(x)$  and  $F(\bar{x}) + DF(\bar{x})(x - \bar{x})$  lie in a neighborhood of  $F(\bar{x})$  where the local Lipschitz constant  $\ell$  of  $g$  is valid, which shows the first inequality in (21). The second case uses the concept of  $\psi$ -uniform continuity from Definition 2 and Lemma 4. The third case is the assumption that the error obeys a growth function. With a similar

reasoning, we can show that Theorem 22(iv) is satisfied.

We consider Assumption 3 (see also Remark 24). Let  $x_k \rightarrow x^*$  as  $k \xrightarrow{K} \infty$  for  $K \subset \mathbb{N}$  and  $\tilde{x}_k - x_k \rightarrow 0$ . Since  $g$  is finite-valued, using [7, Corollary 16.38] (sum-rule for the subdifferential), and [36, Theorem 10.6], we observe that

$$0 \in \partial f_0(\tilde{x}_k) + DF(x_k)^* \partial g(F(x_k) + DF(x_k)(\tilde{x}_k - x_k)) + \nabla h_k(\tilde{x}_k) - \nabla h_k(x_k), \quad (22)$$

where  $DF(x_k)^*$  denotes the adjoint of  $DF(x_k)$ . We can assume that, for  $k$  large enough,  $F(x_k) + DF(x_k)(\tilde{x}_k - x_k)$  and  $F(\tilde{x}_k)$  lie in a neighborhood of  $F(x^*)$  on which  $g$  has the Lipschitz constant  $\ell > 0$ . By [36, Theorem 9.13],  $\partial g$  is locally bounded around  $F(x^*)$ , i.e. there exists a compact set  $G$  such that  $\partial g(z) \subset G$  for all  $z$  in a neighborhood of  $F(x^*)$ . We conclude that

$$\sup_{\substack{v \in \partial g(F(x_k) + DF(x_k)(\tilde{x}_k - x_k)) \\ w \in \partial g(F(\tilde{x}_k))}} |DF(x_k)^* v - DF(\tilde{x}_k)^* w| \leq \sup_{v, w \in G} |DF(x_k)^* v - DF(\tilde{x}_k)^* w| \rightarrow 0$$

for  $k \xrightarrow{K} \infty$  since  $DF(x_k) \rightarrow DF(x^*)$  and  $DF(\tilde{x}_k) \rightarrow DF(x^*)$ . Again assuming that  $\nabla h_k(\tilde{x}_k) - \nabla h_k(x_k) \rightarrow 0$  we conclude that the outer set-limit of the right hand side of (22) is included in  $\partial f(\tilde{x}_k)$  and, therefore, the slope  $|\nabla f|(\tilde{x}_k)$  vanishes for  $k \xrightarrow{K} \infty$ .

**Example 36.** Problems of the form

$f_0 + g \circ F$  with  $f_0 \in \Gamma_0$ ,  $g \in C^1$ ,  $\nabla g_i \geq 0$ , and  $F = (F_1, \dots, F_M)$  is Lipschitz with  $F_i \in \Gamma_0$  can be modeled by

$$f_{\bar{x}}(x) = f_0(x) + g(F(\bar{x})) + \langle F(x) - F(\bar{x}), \nabla g(F(\bar{x})) \rangle.$$

Such problems appear for example in non-convex regularized imaging problems in the context of iteratively reweighted algorithms [33]. For the error of this model function, we observe the following:

$$\begin{aligned} |f(x) - f_{\bar{x}}(x)| &= |g(F(x)) - (g(F(\bar{x})) + \langle F(x) - F(\bar{x}), \nabla g(F(\bar{x})) \rangle)| \\ &\leq \begin{cases} \frac{\ell}{2} |F(x) - F(\bar{x})|^2, & \text{if } \nabla g \text{ is } \ell\text{-Lipschitz}; \\ \int_0^1 \frac{\varphi(t|F(x) - F(\bar{x})|)}{t} dt, & \text{if } \nabla g \text{ is } \psi\text{-uniform continuous}; \\ \omega(|F(x) - F(\bar{x})|), & \text{otherwise}; \end{cases} \\ &\leq \begin{cases} \frac{\ell L^2}{2} |x - \bar{x}|^2, & \text{if } \nabla g \text{ is } \ell\text{-Lipschitz and } F \text{ is } L\text{-Lipschitz}; \\ \int_0^1 \frac{\varphi(tL|x - \bar{x}|)}{t} dt, & \text{if } \nabla g \text{ is } \psi\text{-uniformly continuous and} \\ & F \text{ is } L\text{-Lipschitz}; \\ \omega(|x - \bar{x}|), & \text{otherwise,} \end{cases} \end{aligned}$$

which shows the same growth functions are obeyed as in Example 33 and 35. The explanation for the validity of the reformulations are analogue to those of Example 35. It is easy to

see that Theorem 22(iv) holds.

We consider Assumption 3/Remark 24. Let  $x_k \rightarrow x^*$  as  $k \xrightarrow{K} \infty$  for  $K \subset \mathbb{N}$  and  $\tilde{x}_k - x_k \rightarrow 0$ . Since  $g$  is continuously differentiable, the sum-formula for the subdifferential holds. Moreover, we can apply [36, Corollary 10.09] (addition of functions) to see that  $\tilde{x}_k$  satisfies the following relation:

$$0 \in \partial f_0(\tilde{x}_k) + \sum_{i=1}^M \partial F_i(\tilde{x}_k)(\nabla g(F(x_k)))_i + \nabla h_k(\tilde{x}_k) - \nabla h_k(x_k),$$

Note that by [36, Theorem 10.49] the subdifferential of  $g \circ F$  at  $\tilde{x}_k$  is  $\sum_{i=1}^M \partial F_i(\tilde{x}_k)(\nabla g(F(\tilde{x}_k)))_i$ . As in Example 35, using the Lipschitz continuity of  $F$ , hence local boundedness of  $\partial F$ , and using the continuous differentiability of  $g$ , the sequence of sets  $\sum_{i=1}^M \partial F_i(\tilde{x}_k)(\nabla g(F(x_k)))_i - \partial F_i(\tilde{x}_k)(\nabla g(F(\tilde{x}_k)))_i$  vanishes for  $k \xrightarrow{K} \infty$ , which implies that the slope  $|\nabla f|(\tilde{x}_k)$  vanishes for  $k \xrightarrow{K} \infty$ .

**Example 37 (Problem adaptive model function).** Our framework allows for a problem specific adaptation using a combination of Examples 33, 34, 35, and 36. Consider the following objective function<sup>3</sup>  $f: \mathbb{R}^N \times \mathbb{R}^2 \rightarrow \overline{\mathbb{R}}$  with  $a \in \mathbb{R}^N$  and  $b \in \mathbb{R}$ :

$$f(x, z) = f_1(z) + \delta_{[-1,1]^2}(z) + f_2(x) \text{ with } \begin{cases} f_1 \in C^2(\mathbb{R}^2) \text{ is strongly convex;} \\ f_2(x) := \max\{(\langle a, x \rangle - b)^2, 1 - \exp(-|x|)\}. \end{cases}$$

We define our model function as:  $f_{(\bar{x}, \bar{z})}(x, z) = \bar{f}_1(z; \bar{z}) + \delta_{[-1,1]}(z) + \bar{f}_2(x; \bar{x})$  with

$$\begin{aligned} \bar{f}_1(z; \bar{z}) &:= f_1(\bar{z}) + \langle \nabla f_1(\bar{z}), z - \bar{z} \rangle + \frac{1}{2} \langle z - \bar{z}, \nabla^2 f_1(\bar{z})(z - \bar{z}) \rangle; \\ \bar{f}_2(x; \bar{x}) &:= \max\{(\langle a_i, x \rangle - b_i)^2, 1 + \exp(-|\bar{x}|)(|x| - |\bar{x}| - 1)\}. \end{aligned}$$

The strong convexity of  $f_1$  allows for a convex second order approximation with positive definite Hessian. We linearize only the second component of the “max” (w.r.t.  $|x|$ ) in  $f_2$  to obtain a convex approximation that is as close as possible to original function  $f_2$ .

As  $\nabla f_1$  is Lipschitz continuous on the compact set  $[0, 1]^2$ , the growth function w.r.t.  $z$  is of the form  $L|z - \bar{z}|^2$ . Moreover, using 1-Lipschitz continuity of  $\exp$  on  $\mathbb{R}_-$ , we can bound  $|\exp(-|x|) - \exp(-|\bar{x}|)(1 + |\bar{x}| - |x|)|$  by  $||x| - |\bar{x}||^2$  and, using Lipschitz continuity of  $|x|$ , by  $|x - \bar{x}|^2$ . Therefore, the model error is given by a growth function  $w(t) = \max\{1, L\}t^2$ .

## 5.2 Examples of Bregman functions

Let us explore some of the Bregman functions, that are most important to our applications and show that our assumptions are satisfied.

---

<sup>3</sup>The example is not meant to be meaningful and the model function to be algorithmically the best choice. This example shall demonstrate the flexibility and problem adaptivity of our framework.

**Example 38 (Euclidean Distance).** The most natural Bregman proximity function is the Euclidean distance

$$D_h(x, \bar{x}) = \frac{1}{2}|x - \bar{x}|^2,$$

which is generated by the Legendre function  $h(x) = \frac{1}{2}|x|^2$ . The domain of  $h$  is the whole space  $\mathbb{R}^N$ , which implies that Condition (ii) in Theorem 22 is satisfied for any limit point. Assumption 2 is trivial, and for the model functions in Section 5.1, Assumption 3 is satisfied, if  $x_k - \bar{x}_k \rightarrow 0$  implies  $\nabla h(x_k) - \nabla h(\bar{x}_k) \rightarrow 0$ , which is clearly true. Therefore, for the models in Section 5.1 combined with the Euclidean proximity measure, we conclude subsequential convergence to a stationary point.

**Example 39 (Variable Euclidean Distance).** A simple but far-reaching extension of Example 38 is the following. Let  $(A_k)_{k \in \mathbb{N}}$  be a sequence of symmetric positive definite matrices such that the smallest and largest eigenvalues are in  $[c_1, c_2]$  for some  $0 < c_1 < c_2 < +\infty$ , i.e.  $0 < \inf_k \langle x, A_k x \rangle < \sup_k \langle x, A_k x \rangle < +\infty$  for all  $x \in \mathbb{R}^N$ . Each matrix  $A_k$  induces a metric on  $\mathbb{R}^N$  via the inner product  $\langle x, A \bar{x} \rangle$  for  $x, \bar{x} \in \mathbb{R}^N$ . The induced norm is a Bregman proximity function

$$D_{h_k}(x, \bar{x}) = \frac{1}{2}|x - \bar{x}|_{A_k}^2 := \frac{1}{2} \langle x - \bar{x}, A_k(x - \bar{x}) \rangle,$$

generated analogously to Example 38. Except the boundedness of the eigenvalues of  $(A_k)_{k \in \mathbb{N}}$  there are no other restrictions. All the conditions mentioned in Example 38 are easily verified.

From now on, we restrict to iteration-independent Bregman distance functions, knowing that we can flexibly adapt the Bregman distance in each iteration.

**Example 40 (Boltzmann–Shannon entropy).** The Boltzmann–Shannon entropy is

$$D_h(x, \bar{x}) = \sum_{i=1}^N (x^{(i)}(\log(x^{(i)}) - \log(\bar{x}^{(i)})) - (x^{(i)} - \bar{x}^{(i)}))$$

where  $x^{(i)}$  denotes the  $i$ -th coordinate of  $x \in \mathbb{R}^N$ .  $D_h$  is generated by the Legendre function  $h(x) = \sum_{i=1}^N x^{(i)} \log(x^{(i)})$ , which has the domain  $[0, +\infty[^N$ . Since  $h$  is additively separable, w.l.o.g., we restrict the discussion to  $N = 1$  in the following.

We verify Assumption 2. Let  $(x_k)_{k \in \mathbb{N}}$  and  $(\bar{x}_k)_{k \in \mathbb{N}}$  be bounded sequences in  $\text{int dom } h = ]0, +\infty[$  with  $x_k - \bar{x}_k \rightarrow 0$  for  $k \rightarrow \infty$ . For any convergent subsequence  $x_k \rightarrow x^*$  as  $k \xrightarrow{K} \infty$  for some  $K \subset \mathbb{N}$  also  $\bar{x}_k \rightarrow x^*$  as  $k \xrightarrow{K} \infty$  and  $x^* \in ]0, +\infty[$ . Since  $h$  is continuous on  $\text{cl dom } h = [0, +\infty[$  (define  $h(0) = 0 \log(0) = 0$ ),  $D_h(x_k, \bar{x}_k) \rightarrow 0$  for any convergent subsequence, hence for the full sequence. The same argument shows that the converse implication is also true, hence the Boltzmann–Shannon entropy satisfies Assumption 2.

For the model functions from Section 5.1, we show that Assumption 3 holds for  $x^* \in \text{int dom } h$ , i.e.  $\nabla h(x_k) - \nabla h(\bar{x}_k) \rightarrow 0$  for sequence  $(x_k)_{k \in \mathbb{N}}$  and  $(\bar{x}_k)_{k \in \mathbb{N}}$  with  $x_k \rightarrow x^*$  and



$x_k - \bar{x}_k \rightarrow 0$  for  $k \xrightarrow{K} \infty$  for some  $K \subset \mathbb{N}$ . This condition is satisfied, because  $\nabla h$  is continuous on  $\text{int dom } h$ , hence  $\lim_{k \xrightarrow{K} \infty} \nabla h(x_k) = \lim_{k \xrightarrow{K} \infty} \nabla h(\bar{x}_k) = \nabla h(x^*)$ .

Since  $\text{dom } h = \text{cl dom } h$ , it suffices to verify Condition (iii) of Theorem 22 to guarantee subsequential convergence to a stationary point. For  $x^* \in ]0, +\infty[$  and a bounded sequence  $(\tilde{y}_k)_{k \in \mathbb{N}}$  in  $\text{int dom } h$  as in Theorem 22, we need to show that  $D_h(x^*, \tilde{y}_k) \rightarrow 0$  as  $k \xrightarrow{K} \infty$  for  $K \subset \mathbb{N}$  such that  $\tilde{y}_k \rightarrow x^*$  as  $k \xrightarrow{K} \infty$ . This result is clearly true for  $x^* > 0$ , thanks to the continuity of  $\log$ . For  $x^* = 0$ , we observe  $x^* \log(\tilde{y}_k) \rightarrow 0$  for  $k \xrightarrow{K} \infty$ , hence Condition (iii) of Theorem 22 holds, and subsequential convergence to a stationary point is guaranteed.

**Example 41 (Burg's entropy).** For optimization problems with non-negativity constraint, Burg's entropy is a powerful distance measure. The associated Bregman distance

$$D_h(x, \bar{x}) = \sum_{i=1}^N \left( \frac{x^{(i)}}{\bar{x}^{(i)}} - \log \left( \frac{x^{(i)}}{\bar{x}^{(i)}} \right) - 1 \right)$$

is generated by the Legendre function  $h(x) = -\sum_{i=1}^N \log(x^{(i)})$  which is defined on the domain  $]0, +\infty[^N$ . Approaching 0, the function  $h$  grows towards  $+\infty$ . In contrast to the Bregman functions in the examples above, Burg's entropy does not have a Lipschitz continuous gradient, and is therefore interesting for objective functions with the same deficiency.

W.l.o.g. we consider  $N = 1$ . Assumption 2 for two bounded sequences  $(x_k)_{k \in \mathbb{N}}$  and  $(\bar{x}_k)_{k \in \mathbb{N}}$  in  $]0, +\infty[$  reads

$$x_k - \bar{x}_k \rightarrow 0 \quad \Leftrightarrow \quad \frac{x_k}{\bar{x}_k} - \log \left( \frac{x_k}{\bar{x}_k} \right) \rightarrow 1,$$

which is satisfied if the limit points lie in  $]0, +\infty[$  since  $x_k - \bar{x}_k \rightarrow 0 \Leftrightarrow x_k/\bar{x}_k \rightarrow 1$  for  $k \xrightarrow{K} \infty$  and  $\log$  is continuous at 1.

For the model functions in Section 5.1, Assumption 3 requires  $\nabla h(x_k) - \nabla h(\bar{x}_k) \rightarrow 0$  for sequence  $(x_k)_{k \in \mathbb{N}}$  and  $(\bar{x}_k)_{k \in \mathbb{N}}$  in  $\text{int dom } h$  with  $x_k \rightarrow x^*$  and  $x_k - \bar{x}_k \rightarrow 0$  for  $k \xrightarrow{K} \infty$  for some  $K \subset \mathbb{N}$ . By continuity, this statement is true for any  $x^* > 0$ . For  $x^* = 0$ , the statement is in general not true. Also Condition (iv) in Theorem 22 can, in general, not be verified. Therefore, if a model functions is complemented with Burg's entropy, then the objective should be continuous on the  $\text{cl dom } h$ . Stationarity of limit points is obtained, if they lie in  $\text{int dom } h$ .

## 6 Applications

We discuss in this section some numerical experiments whose goal is to illustrate the wide applicability of our algorithmic framework. The applicability of our results follows from the considerations in Section 5. Actually, the considered objective functions are all continuous, i.e. Theorem 22(i) is satisfied.

## 6.1 Robust Non-linear Regression

We consider a simple non-smooth and non-convex robust regression problem [22] of the form

$$\min_{u:=(a,b) \in \mathbb{R}^P \times \mathbb{R}^P} \sum_{i=1}^M \|F_i(u) - y_i\|_1, \quad F_i(u) := \sum_{j=1}^P b_j \exp(-a_j x_i), \quad (23)$$

where  $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$ ,  $i = 1, \dots, M$  is a sequence of covariate-observation pairs. We assume that  $(x_i, y_i)$  are related by  $y_i = F_i(u) + n_i$ , where  $n_i$  is the error term and  $u = (a, b)$  are the unknown parameters. We assume that the errors are iid with Laplacian distribution, in which case the data fidelity devised by a maximum likelihood argument is  $\ell_1$ -norm as used in (23).

We define model functions by linearizing the inner functions  $F_i$  as suggested by the model function in Example 35. Complemented by an Euclidean proximity measure (with  $\tau > 0$ ) the convex subproblem (5) to be solved inexactly is the following:

$$\tilde{u} = \operatorname{argmin}_{u \in \mathbb{R}^P \times \mathbb{R}^P} \sum_{i=1}^M \|\mathcal{K}_i u - y_i^\diamond\|_1 + \frac{1}{2\tau} |u - \bar{u}|^2, \quad y_i^\diamond := y_i - F(\bar{u}) + \mathcal{K}_i \bar{u},$$

where  $\mathcal{K}_i := DF_i(\bar{u}): \mathbb{R}^P \times \mathbb{R}^P \rightarrow \mathbb{R}$  is the Jacobian of  $F_i$  at the current parameters  $\bar{u}$ . We solve the (convex) dual problem (cf. [14, 18]) with warm starting up to absolute step difference  $10^{-3}$ .

As mentioned in Remark 12, backtracking on  $\tau$  could be used (cf. ProxDescent [25]); denoted `prox-linear` and `prox-linear2` in the following. This requires to solve the subproblem for each trial step. This is the bottleneck compared to evaluating the objective. The line search in Algorithm 2 only has to evaluate the objective value. This variant is denoted `prox-linear-LS` in the following. A representative convergence result in terms of the number of accumulated iterations of the subproblems is shown in Figure 1. For this random example, the maximal noise amplitude is 12.18, and the maximal absolute deviation of the solution from the ground truth is 0.53, which is reasonable for this noise level. Algorithm `prox-linear-LS` requires significantly fewer subproblem iterations than `prox-linear` and `prox-linear2`. For `prox-linear2` the initial  $\tau$  is chosen such that initially no backtracking is required.

For large scale problems, frequently solving the subproblems can be prohibitively expensive. Hence, ProxDescent cannot be applied, whereas our algorithm is still practical.

## 6.2 Image Deblurring under Poisson Noise

Let  $b \in \mathbb{R}^{n_x \times n_y}$  represent a blurry image of size  $n_x \times n_y$  corrupted by Poisson noise. Recovering a clean image from  $b$  is an ill-posed inverse problem. It is a common problem, for example, in fluorescence microscopy and optical/infrared astronomy; see [8] and references therein. A popular way to solve it is to formulate an optimization problem [40] of the form

$$\min_{u \in \mathbb{R}^{n_x \times n_y}} f(u) := D_{KL}(b, \mathcal{A}u) + \frac{\lambda}{2} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \phi(|(\mathcal{D}u)_{i,j}|^2), \quad s.t. \ u_{i,j} \geq 0, \quad (24)$$

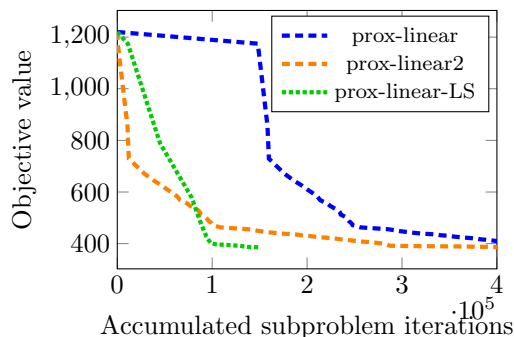


Figure 1: Objective value vs. accumulated number of subproblem iterations for (23).



Figure 2: Deblurring and Poisson noise removal by solving (24). From left to right: clean, noisy, and reconstructed image (PSNR: 25.86).

where  $\mathcal{A}$  is a circular convolution (blur) operator. The first term (coined data term) in the objective  $f$  is the Kullback–Leibler divergence (Bregman distance generated by the Boltzmann–Shannon entropy  $x \log(x)$ ), which, neglecting additive constant terms, is given by

$$f_1(u) := D_{KL}(b, \mathcal{A}u) := \sum_{i,j} (\mathcal{A}u)_{i,j} - b_{i,j} \log((\mathcal{A}u)_{i,j}),$$

$f_1$  is well-suited for Poisson noise removal [38]. The second term (coined regularization term) involves a penalty  $\phi: \mathbb{R}^2 \rightarrow \mathbb{R}$  applied to spatial finite differences  $(\mathcal{D}u)_{i,j} := ((\mathcal{D}u)_{i,j}^1, (\mathcal{D}u)_{i,j}^2)^\top$  in horizontal direction  $(\mathcal{D}u)_{i,j}^1 := u_{i+1,j} - u_{i,j}$  for all  $(i, j)$  with  $i < n_x$ , and 0 otherwise; and vertical direction  $(\mathcal{D}u)_{i,j}^2$  (defined analogously). The function  $\phi$  in the regularization is usually chosen to favor “smooth” images with sharp edges. The relative importance of both the data and regularization terms is weighted by  $\lambda > 0$ .

For convex penalties  $\phi$ , algorithms for solving problem (24) are available (e.g. primal-dual proximal splitting) provided that  $\phi$  is simple (in the sense that its Euclidean proximal mapping can be computed easily). But if one would like to exploit the gradient of  $f_1$  explicitly, things become more intricate. The difficulty comes from the lack of global Lipschitz continuity of  $\nabla f_1(u)$ . A remedy is provided by Bauschke et al. [6]. They have shown that, instead of the global Lipschitz continuity, the key property is the convexity of  $Lh - f_1$  for a Legendre function  $h$  and sufficiently large  $L$ , which can be achieved using Burg’s entropy  $h(u) = -\sum_{i,j} \log(u_{i,j})$  ([6, Lemma 7]).

However, non-convex penalties  $\phi$  are known to yield a better solution [21, 9, 29]. In this case, the algorithmic framework of Bauschke et al. [6] is not applicable anymore, whereas our framework is applicable. Due to the lack of strong convexity of Burg’s entropy also the algorithm of Bonettini et al. [11] cannot be used. Note that Burg’s entropy is strongly convex on bounded subsets of  $]0, +\infty[$ , however, the subset cannot be determined a priori.

The abstract framework proposed in this paper appears to be the first algorithm with convergence guarantees for solving (24) with a smooth non-convex regularizer.

In our framework, we choose  $\phi: t \in \mathbb{R}^2 \mapsto \log(1 + \rho|t|^2)$ , which is smooth but non-convex. The model functions are defined as in Example 33. We also use  $h$  as the Burg’s entropy

to generate the Bergman proximity function (see Example 41). Thus, the subproblems (5) which emerge from linearizing the objective  $f$  in (24) around the current iterate  $\bar{u}$

$$\tilde{u} = \underset{u \in \mathbb{R}^{n_x \times n_y}}{\operatorname{argmin}} \langle u - \bar{u}, \nabla f(\bar{u}) \rangle + \frac{1}{\tau} \sum_{i,j} \left( \frac{u_{i,j}}{\bar{u}_{i,j}} - \log \left( \frac{u_{i,j}}{\bar{u}_{i,j}} \right) \right)$$

can be solved exactly in closed-form  $\tilde{u}_{i,j} = \bar{u}_{i,j} / (1 + \tau (\nabla f(\bar{u}))_{i,j} \bar{u}_{i,j})$  for all  $i, j$ . A result for the successful Poisson noise removal and deblurring is shown in Figure 2.

### 6.3 Structured Matrix Factorization

Structured matrix factorization problems are crucial in data analysis. It has many applications in various areas including blind deconvolution in signal processing, clustering, source separation, dictionary learning, etc.. There is a large body of literature on the subject and we refer to e.g. [17, 15, 37, 39] and references therein for a comprehensive account.

**The problem.** Given a data matrix  $A \in \mathbb{R}^{M \times N}$  whose  $N$   $M$ -dimensional columns are the data vectors. The goal is to find two matrices  $U \in \mathbb{R}^{M \times K}$  and  $Z \in \mathbb{R}^{K \times N}$  such that

$$A = UZ + Q,$$

where  $Q \in \mathbb{R}^{M \times N}$  accounts for an unknown error. The matrices  $U$  and  $Z$  (called also factors) enjoy features arising in a specific application at hand (see more below).

To solve the matrix factorization problem, we adopt the optimization approach and we consider the non-convex and non-smooth minimization problem

$$\min_{U \in \mathcal{U}, Z \in \mathcal{Z}} f(A, UZ) + \lambda g(Z), \quad f(A, UZ) := \frac{1}{2} \|A - UZ\|_F^2. \quad (25)$$

The term  $f(A, UZ)$  stands for proximity function that measures fidelity of the approximation of  $A$  by the product  $UZ$  of the two factors. We here focus on the classical case where the fidelity is measured via the Frobenius norm  $\|\cdot\|_F$ , but other data fidelity measures can also be used just as well in our framework, such as divergences (see [17] and references therein). The sets  $\mathcal{U}$ ,  $\mathcal{Z}$ , which are non-empty closed and convex, and the function  $g \in \Gamma_0$  are used to capture specific features of the matrices  $U$  and  $Z$  arising in a specific application as we will exemplify shortly. The influence of  $g$  is weighted by the parameter  $\lambda > 0$ .

Many (if not most) algorithms to solve the matrix factorization problem (25) are based on Gauss-Seidel alternating minimization with limited convergence guarantees [17, 15, 37]<sup>4</sup>. The PALM algorithm proposed recently by Bolte et al. [10], was designed specifically for the structure of the optimization problem (25). It can then be successfully applied to solve

---

<sup>4</sup>For very specific instances, a recent line of research proposes to lift the problem to the space of low-rank matrices, and then use convex relaxation and computationally intensive conic programming that are only applicable to small-dimensional problems; see, e.g., [1] for blind deconvolution.

instances of such a problem with provably guaranteed convergence under some assumptions including the Kurdyka-Lojasiewicz property. However, though it can handle non-convex constraint sets and functions  $g$ , it does not allow to incorporate Bregman proximity functions.

In the following, we show how our algorithmic framework can be applied to a broad class of matrix factorization instances. In particular, a distinctive feature of our algorithm is that it can naturally and readily accommodate for different Bregman proximity functions and it has no restrictions on the choice of the step size parameters (except positivity). A descent is enforced in the line search step, which follows the proximal step.

**A generic algorithm.** We apply Algorithm 1 to solve this problem, where the model functions are chosen to linearize the data fidelity function  $f(A, UZ)$ , according to Example 33. The convex subproblems to be solved in the algorithm have the following form:

$$\begin{aligned}
 (\tilde{U}, \tilde{Z}) = \operatorname{argmin}_{U \in \mathcal{U}, Z \in \mathcal{Z}} & \lambda g(Z) + \langle Z - \bar{Z}, \bar{U}^\top (\bar{U} \bar{Z} - A) \rangle_F + D_{h_Z}(Z, \bar{Z}) \\
 & + \langle U - \bar{U}, (\bar{U} \bar{Z} - A) \bar{Z}^\top \rangle_F + D_{h_U}(U, \bar{U})
 \end{aligned}$$

where  $\langle \cdot, \cdot \rangle_F$  stands for the Frobenius inner product. The Bregman proximity functions  $D_{h_Z}(\cdot, \cdot)$  and  $D_{h_U}(\cdot, \cdot)$  provide the flexibility to handle a variety of constraint sets  $\mathcal{U}$  and  $\mathcal{Z}$ . In the following, we list different choices for the constraint sets and explain how to incorporate them into the optimization procedure. Due to the structure of the optimization problem, the variables  $U$  and  $Z$  can be handled separately. The only coupling is the data fidelity function  $f$ , which is linearized and therefore easy to incorporate.

**Examples of constraints  $\mathcal{U}$ .** There are many possible choices for the set  $\mathcal{U}$  depending on the application at hand.

- **Unconstrained case:**

$$\mathcal{U}_1 = \mathbb{R}^{M \times K}.$$

In the unconstrained case, a suitable Bregman proximity function is given by the Euclidean distance  $D_{h_U}(U, \bar{U}) = \frac{1}{2\tau_U} \|U - \bar{U}\|_F^2$  with step size parameter  $\tau_U$ . The resulting update step with respect to the dictionary  $U$  is a gradient descent step.

- **Zero-mean and normalization:**

$$\mathcal{U}_2 = \left\{ U \in \mathbb{R}^{M \times K} \mid \forall j: \sum_{i=1}^M U_{i,j}^2 \leq 1, \forall j \geq 2: \sum_{i=1}^M U_{i,j} = 0 \right\}.$$

This choice of the constraint set leads to a natural normalization of the columns of  $U$  that removes the scale ambiguity due to bilinearity. This choice is very classical in dictionary learning, see, e.g., [39]. As in dictionary learning, the average of the first column may not be enforced to be zero, in order to allow the first column to absorb the mean value of the data points.

By separability of  $\mathcal{U}_2$ , the Euclidean projection onto it is simple. This projector is column-wise achieved by subtracting the mean, and then projecting the result onto the Euclidean unit ball. Thus we advocate  $D_{h_U}(U, \bar{U}) = \frac{1}{2\tau_U} \|U - \bar{U}\|_F^2$  with step size parameter  $\tau_U$ . In turn, the subproblem with respect to  $U$  amounts to a projected gradient descent step.

- **Non-negativity and normalization:**

$$\mathcal{U}_3 = \left\{ U \in \mathbb{R}^{M \times K} \mid \forall j: \sum_{i=1}^M U_{i,j} = 1, \forall i, j: U_{i,j} \geq 0 \right\}.$$

This choice is adopted in non-negative matrix factorization (NMF) [24]. The constraint set  $\mathcal{U}_3$  is column-wise a unit simplex constraint. This constraint can be conveniently handled by choosing  $D_{h_U}(U, \bar{U}) = \frac{1}{\tau_U} \sum_{i,j} U_{i,j} (\log(U_{i,j}) - \log(\bar{U}_{i,j})) - U_{i,j} + \bar{U}_{i,j}$ , which is the Bregman function generated by the entropy  $h_U(U) = \frac{1}{\tau_U} \sum_{i,j} U_{i,j} \log(U_{i,j})$  with step size parameter  $\tau_U$ . This is a more natural choice than the Euclidean proximity distance. Indeed, the update step with respect to  $U$  results in

$$\tilde{U}_{i,j} = \frac{\bar{U}_{i,j} \exp(-\tau_U (C_U)_{i,j})}{\sum_{p=1}^M \bar{U}_{p,j} \exp(-\tau_U (C_U)_{p,j})} \quad \forall i = 1, \dots, M; \forall j = 1, \dots, K,$$

where we use the shorthand notation  $C_U := \nabla_U f(A, \bar{U}\bar{Z}) = \bar{U}^\top (\bar{U}\bar{Z} - A)$  for the partial gradient of  $f$  with respect to  $U$ . The exponential function is applied entry-wise, hence naturally preserving positivity. Note that the Euclidean projector onto  $\mathcal{U}_3$  necessitates to compute the projector on the simplex which can be achieved with sorting [28].

**Examples of constraints  $\mathcal{Z}$ .** There are also several possible choices for the set  $\mathcal{Z}$  and regularizing function  $g$  depending on the application at hand.

- **Unconstrained case:**

$$\mathcal{Z}_1 = \mathbb{R}^{K \times N} \quad \text{and} \quad g(Z) = 0.$$

This case can be handled using a gradient descent step, analogously to the related update step with the constraint set  $\mathcal{U}_1$ .

- **Non-negativity:**

$$\mathcal{Z}_2 = \{Z \in \mathbb{R}^{K \times N} \mid \forall i, j: Z_{i,j} \geq 0\} \quad \text{and} \quad g(Z) = 0.$$

This constraint is used in conjunction with  $\mathcal{U}_3$  in NMF. It can be handled either with a Euclidean proximity function (which amounts to projecting on the non-negative orthant), or via a Bregman proximity function  $D_{h_Z}(Z, \bar{Z})$  generated by the Boltzmann–Shannon entropy ( $h_Z(Z) = \frac{1}{\tau_Z} \sum_{i,j} Z_{i,j} \log(Z_{i,j})$ ) or, alternatively, Burg’s entropy

( $h_Z(Z) = -\frac{1}{\tau_Z} \sum_{i,j} \log(Z_{i,j})$ ), with step size parameter  $\tau_Z$ . The update with respect to  $Z$  then reads

$$\tilde{Z}_{i,j} = \bar{Z}_{i,j} \exp(-\tau_Z (C_Z)_{i,j}) \quad \forall i = 1, \dots, K; \forall j = 1, \dots, N,$$

where we use the shorthand notation  $C_Z := \nabla_Z f(A, \bar{U}\bar{Z}) = (\bar{U}\bar{Z} - A)\bar{Z}^\top$  for the partial gradient of  $f$  with respect to  $Z$ .

- **Sparsity constraints:**

$$\mathcal{Z}_3 = \mathbb{R}^{K \times N} \quad \text{and} \quad g(Z) = \|Z\|_1.$$

The introduction of sparsity has been of prominent importance in several matrix factorization problems, including dictionary learning [34], NMF [23]<sup>5</sup> and source separation [37]. The Euclidean proximal mapping of the  $\ell_1$ -norm is the entry-wise soft-thresholding, hence giving the update step with respect to  $Z$  as

$$\tilde{Z}_{i,j} = \max\{0, 1 - \lambda\tau_Z / |\bar{Z}_{i,j} - \tau_Z (C_Z)_{i,j}|\} (\bar{Z}_{i,j} - \tau_Z (C_Z)_{i,j}) \quad \forall i = 1, \dots, K; \forall j = 1, \dots, N.$$

- **Low rank constraint:**

$$\mathcal{Z}_3 = \mathbb{R}^{K \times N} \quad \text{and} \quad g(Z) = \|Z\|_*.$$

The nuclear norm or 1-Schatten norm  $\|Z\|_*$  is the sum of the singular values. It is known to be the tightest convex relaxation to the rank and was shown to promote low rank solutions [35]. Such a regularization would be useful in the situation where columns of  $A$  are (to a good approximation) clustered on a few linear subspaces spanned by the columns of  $U$ , i.e. the columns of  $A$  can be explained by columns of  $U$  from the same subspace (“cluster”).

The Euclidian proximal mapping of the nuclear norm is the soft-thresholding applied to the singular values. In turn, the update step with respect to  $Z$  reads

$$\tilde{Z}_{i,j} = W \text{diag}((\max\{0, 1 - \lambda\tau_Z / \sigma_i\} \sigma_i)_i) V^\top,$$

where  $W, V$  are respectively the matrices of left and right singular vectors of  $\bar{Z} - \tau_Z C_Z$ , and  $\sigma$  is the associated vector of singular values.

---

<sup>5</sup>Strictly speaking,  $\mathcal{Z}_3$  should be the non-negative orthant for sparse NMF. But this does not change anything to our discussion since computing the Euclidean proximal mapping of the  $\ell_1$  norm restricted to the non-negative orthant is easy.

## 7 Conclusions

We have presented an algorithmic framework, that unifies the analysis of several first order optimization algorithms in non-smooth non-convex optimization such as Gradient Descent, Forward–Backward Splitting, ProxDescent, and many more. The algorithm combines sequential Bregman proximal minimization of model functions, which is the key concept for the unification, with an Armijo-like line search strategy. The framework reduces the difference between algorithms to the model approximation error measured by a growth function. For the developed abstract algorithmic framework, we establish subsequential convergence to a stationary point and demonstrate its flexible applicability in several difficult inverse problems from machine learning, signal and image processing.

## References

- [1] A. Ahmed, B. Recht, and J. Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2014.
- [2] H. Attouch, J. Bolte, and B. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.
- [3] H. Bauschke and J. Borwein. Legendre functions and the method of random Bregman projections. *Journal of Convex Analysis*, 4(1):27–67, 1997.
- [4] H. Bauschke, J. Borwein, and P. Combettes. Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces. *Communications in Contemporary Mathematics*, 3(4):615–647, Nov. 2001.
- [5] H. Bauschke, J. Borwein, and P. Combettes. Bregman monotone optimization algorithms. *SIAM Journal on Control and Optimization*, 42(2):596–636, Jan. 2003.
- [6] H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, Nov. 2016.
- [7] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.
- [8] M. Bertero, P. Boccacci, G. Desiderà, and G. Vicidomini. Image deblurring with Poisson data: from cells to galaxies. *Inverse Problems*, 25(12):123006, 2009.
- [9] A. Blake and A. Zisserman. *Visual Reconstruction*. MIT Press, Cambridge, MA, 1987.
- [10] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for non-convex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.
- [11] S. Bonettini, I. Loris, F. Porta, and M. Prato. Variable metric inexact line-search based methods for nonsmooth optimization. *SIAM Journal on Optimization*, 26(2):891–921, Jan. 2016.
- [12] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.



- 
- [13] J. Burg. The relationship between maximum entropy spectra and maximum likelihood spectra. *Geophysics*, 37(2):375–376, Apr. 1972.
- [14] A. Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20:89–97, 2004.
- [15] S. Chaudhuri, R. Velmurugan, and R. Rameshan. *Blind Image Deconvolution*. Springer, 2014.
- [16] G. Chen and M. Teboulle. Convergence analysis of proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3:538–543, 1993.
- [17] A. Cichocki, R. Zdunek, A. Phan, and S. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. Wiley, New York, 2009.
- [18] P. Combettes, D. Dũng, and B. Vũ. Dualization of signal recovery problems. *Set-Valued and Variational Analysis*, 18(3-4):373–404, Dec. 2010.
- [19] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis. Nonsmooth optimization using Taylor-like models: error bounds, convergence, and termination criteria. *ArXiv e-prints*, Oct. 2016. arXiv:1610.03446.
- [20] D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *ArXiv e-prints*, Feb. 2016. arXiv:1602.06661.
- [21] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.
- [22] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. MIT Press, Cambridge, MA, 1986.
- [23] P. Hoyer. Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.*, 5:1457–1469, 2004.
- [24] D. Lee and H. Seung. Learning the part of objects from nonnegative matrix factorization. *Nature*, 401:788–791, 1999.
- [25] A. Lewis and S. Wright. A proximal method for composite minimization. *Mathematical Programming*, 158(1-2):501–546, July 2016.
- [26] P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Applied Mathematics*, 16(6):964–979, 1979.
- [27] D. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Society for Industrial and Applied Mathematics*, 11:431–441, 1963.
- [28] C. Michelot. A finite algorithm for finding the projection of a point onto the canonical simplex of  $\mathbb{R}^n$ . *J. Optim. Theory Appl.*, 50:195–200, 1986.
- [29] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics*, 42:577–685, 1989.
- [30] Q. Nguyen. Forward–Backward Splitting with Bregman Distances. *Vietnam Journal of Mathematics*, pages 1–21, Jan. 2017.
- [31] D. Noll. Convergence of non-smooth descent methods using the Kurdyka–Łojasiewicz inequality. *Journal of Optimization Theory and Applications*, 160(2):553–572, Sept. 2013.

- [32] D. Noll, O. Prot, and P. Apkarian. A proximity control algorithm to minimize nonsmooth and nonconvex functions. *Pacific Journal of Optimization*, 4(3):571–604, 2008.
- [33] P. Ochs, A. Dosovitskiy, T. Brox, and T. Pock. On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision. *SIAM Journal on Imaging Sciences*, 8(1):331–372, 2015.
- [34] B. Olshausen and D. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research.*, 37, 1996. 3311–3325.
- [35] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.
- [36] R. T. Rockafellar and R.-B. Wets. *Variational Analysis*, volume 317. Springer Berlin Heidelberg, Heidelberg, 1998.
- [37] J.-L. Starck, F. Murtagh, and J. Fadili. *Sparse image and signal processing: wavelets, curvelets, morphological diversity*. Cambridge University Press, 2nd edition, 2015.
- [38] Y. Vardi, L. Shepp, and L. Kaufman. A statistical model for positron emission tomography. *Journal of the American Statistical Association*, 80(389):8–20, 1985.
- [39] Y. Xu, Z. Li, J. Yang, and D. Zhang. A survey of dictionary learning algorithms for face recognition. *IEEE Access*, 5:8502–8514, 2017.
- [40] R. Zanella, P. Boccacci, L. Zanni, and M. Bertero. Efficient gradient projection methods for edge-preserving removal of Poisson noise. *Inverse Problems*, 25(4), 2009.