Bregman forward-backward splitting for nonconvex composite optimization: superlinear convergence to nonisolated critical points

MASOUD AHOOKHOSH, ANDREAS THEMELIS AND PANAGIOTIS PATRINOS

Abstract. We introduce BELLA, a locally superlinearly convergent Bregman forward-backward splitting method for minimizing the sum of two nonconvex functions, one of which satisfying a relative smoothness condition [9, 46] and the other one possibly nonsmooth. A key tool of our methodology is the Bregman forward-backward envelope (BFBE), an exact and continuous penalty function with favorable first- and second-order properties, and enjoying a nonlinear error bound when the objective function satisfies a Łojasiewicz-type property. The proposed algorithm is of linesearch type over the BFBE along candidate update directions, and converges subsequentially to stationary points, globally under a KL condition, and owing to the given nonlinear error bound can attain superlinear convergence rates even when the limit point is a nonisolated minimum, provided the directions are suitably selected.

1. Introduction

In this paper, we address the composite minimization problem

$$\operatorname*{\mathbf{minimize}}_{x\in\mathbb{R}^n} \varphi(x) \equiv f(x) + g(x) \tag{1.1}$$

under the following hypotheses (see Section 2.2):

Assumption I (requirements for composite minimization (1.1))**.** *The following hold:*

A1 $h : \mathbb{R}^n \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ *is strictly convex,* 1*-coercive*[1] *and essentially smooth;*[2]

A2 $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ *is* L_f-smooth relative to h*: namely, functions* $L_f h \pm f$ *are convex on* $\mathbf{dom}\, h$*;*

A3 $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ *is proper and lower semicontinuous (lsc);*

A4 $\mathbf{dom}\,\varphi \subseteq \mathbf{int}\,\mathbf{dom}\, h$*, and* $\operatorname{\mathbf{arg\,min}} \varphi \neq \emptyset$*.*

Despite its simple structure, (1.1) encompasses a variety of optimization problems appearing frequently in scientific areas such as signal and image processing, machine learning, control and system identification; see, e.g., [33, 46]. The notion of relative smoothness has been recently discovered in seminal works [9, 46] as a generalization of smooth functions with Lipschitz-continuous gradients. Studying optimization problems involving relatively smooth functions has received much attention during the last few years [9, 19, 32, 33, 46, 49, 62]. In the composite setting (1.1), since f is relatively smooth and g is nonsmooth nonconvex, we can cover a wide spectrum of applications.

1991 *Mathematics Subject Classification.* 90C06, 90C25, 90C26, 49J52, 49J53.

Key words and phrases. Nonsmooth nonconvex optimization, Bregman forward-backward envelope, relative smoothness, prox-regularity, KL inequality, nonlinear error bound, nonisolated critical points, superlinear convergence.

Department of Electrical Engineering (ESAT-STADIUS) – KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium

This work was supported by the *Research Foundation Flanders (FWO)* research projects G086518N and G086318N; *Research Council KU Leuven* C1 project No. C14/18/068; *Fonds de la Recherche Scientifique — FNRS and the Fonds Wetenschappelijk Onderzoek — Vlaanderen* under EOS project no 30468160 (SeLMA)..

[1] h is 1-coercive if $h(x)/\|x\|$ is coercive.

[2] h is essentially smooth if $\mathbf{int}\,\mathbf{dom}\, h \neq \emptyset$, $h \in C^1(\mathbf{int}\,\mathbf{dom}\, h)$, and $\|\nabla h(x_k)\| \to \infty$ for any sequence $(x_k)_{k\in\mathbb{N}}$ converging to a point in the boundary of $\mathbf{dom}\, h$.

There are plenty of optimization algorithms that can handle composite minimization of the form (1.1), such as [1, 2, 12, 17, 48, 64] for convex problems and [5, 18, 21, 20, 22, 30, 57, 66] for nonconvex problems. Recently, in the relatively smooth setting for convex f and g, [9] proposed a Bregman proximal gradient method, [46] developed primal and dual algorithms, [49] proposed an accelerated tensor method, and [32, 33] suggested a Nesterov-type accelerated method and a stochastic mirror descent method. Moreover, [19, 62] extended the results of [9] in the nonconvex setting. More recently, [8] showed linear convergence of the gradient method for relatively smooth functions. To our knowledge, apart from the latter three papers, there have not been many attempts to deal with (1.1) in the relative smooth setting for nonsmooth nonconvex problems.

One of the most significant discussions in the field of numerical optimization has been related to designing iterative schemes guaranteeing a superlinear convergence rate; see, e.g., [50] for many algorithms attaining a superlinear convergence rate for smooth problems and [30, 31, 57, 64, 66] for other related works in the nonconvex setting. In most of these attempts, the key element is the so-called Dennis-Moré condition [25, 26] which guarantees superlinear convergence to an isolated critical point of the objective function. However, there are many applications that have nonisolated critical points such as low-rank matrix completion [60], low-rank matrix recovery [13], phase retrieval [59], and deep learning [37]. Up to now, besides some attempts for minimizing smooth nonlinear least-squares problems (see, e.g., [3, 4, 35] and references therein) far too little attention has been paid to the superlinear convergence to nonisolated critical points for nonconvex nonsmooth problems. Our main motivation is thus to design an algorithmic framework that requires only a first-order black-box oracle with guarantees of superlinear convergence to nonisolated critical points in a fully nonsmooth nonconvex setting.

1.1. **Contribution.** We propose the Bregman EnveLope Linesearc Algorithm (BELLA) to address problem (1.1), a method that generalizes Bregman forward-backward splitting (BFBS). Our contribution can be summarized as follows.

1) *Bregman forward-backward envelope: a new key tool.* We introduce an envelope function for forward-backward splitting using Bregman distance, the Bregman forward-backward envelope (BFBE), which is a generalization of its Euclidean counterpart introduced in [52] and later further analyzed in [42, 57, 64, 66, 71]. A local equivalence of the BFBE and its Euclidean version allows to provide first- and second-order differential properties of the BFBE based on the known Euclidean properties of prox-regularity and epi-differentiability (Theorems 4.11 and 4.13). As a byproduct of our results, we also present the first- and second-order differential properties of the Bregman-Moreau envelope as a special case. Moreover, the existence of first derivatives of the BFBE in a neighbourhood of critical points under such assumptions allows to provide a local nonlinear error bound for the BFBE around local (not necessarily isolated) minima of the original function, whenever the latter satisfies the Kurdyka-Łojasiewicz (KL) property.

2) *Superlinear convergence to nonisolated critical points.* Using the aforementioned favorable properties of the BFBE around critical points, an accelerated Bregman forward-backward splitting (BELLA) is developed, and the global and linear convergence of the sequence generated by this algorithm under the KL property are given (Theorem 5.5). Remarkably, under mild assumptions and thanks to the nonlinear error bound for the BFBE, the superlinear convergence to nonisolated critical points is shown (Theorem 5.8). To the best of our knowledge, this is the first analysis that exploits this nonlinear error bound to guarantee superlinear convergence to a nonisolated critical point.

1.2. **Paper organization.** The rest of the paper is organized as follows. In Section 2 we introduce the notation and some preliminary results. In Section 3 we review some basic properties of the Bregman forward-backward mapping, needed in Section 4 to construct

and analyze the BFBE, key tool of our analysis. In Section 5, we introduce the proposed BELLA BFBE-based linesearch algorithm, and finally Section 6 concludes the paper.

2. PRELIMINARIES

2.1. **Notation.** The extended-real line is denoted by $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$. The open and closed balls of radius $r \geq 0$ centered at $x \in \mathbb{R}^n$ are denoted as $\mathbf{B}(x; r)$ and $\overline{\mathbf{B}}(x; r)$, respectively. We say that $(x^k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ is *summable* if $\sum_{k\in\mathbb{N}} \|x^k\|$ is finite, and *square-summable* if $(\|x^k\|^2)_{k\in\mathbb{N}}$ is summable.

A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *proper* if $f \not\equiv \infty$, in which case its *domain* is defined as the set $\mathbf{dom}\, f := \{x \in \mathbb{R}^n \mid f(x) < \infty\}$. For $\alpha \in \mathbb{R}$, $[f \leq \alpha] := \{x \in \mathbb{R}^n \mid f(x) \leq \alpha\}$ is the *α-(sub)level set* of f; $[f \geq \alpha]$, $[f = \alpha]$, etc. are defined accordingly. We say that f is *level bounded* if $[f \leq \alpha]$ is bounded for all $\alpha \in \mathbb{R}$. The *convex conjugate* of a function h is denoted as $h^* := \mathbf{sup}_z \{\langle \cdot\,, z\rangle - h(z)\}$.

A vector $v \in \partial f(x)$ is a *subgradient* of f at x, where $\partial f(x)$ is the *subdifferential*

$$\partial f(x) := \left\{v \in \mathbb{R}^n \mid \exists (x^k, v^k)_{k\in\mathbb{N}} \text{ s.t. } x^k \to x,\ f(x^k) \to h(x),\ \hat{\partial} f(x^k) \ni v^k \to v\right\}$$

and $\hat{\partial} f(x)$ is the set of *regular subgradients* of f at x, namely

$$v \in \hat{\partial} f(x) \quad \text{iff} \quad \liminf_{\substack{z\to x\\ z\neq x}} \frac{f(z) - f(x) - \langle v, z - x\rangle}{\|z - x\|} \geq 0.$$

Following the terminology of [55], we say that $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *strictly continuous at* $\bar{x}$ if

$$\mathbf{lip}\, f(\bar{x}) := \limsup_{\substack{y,z\to\bar{x}\\ y\neq z}} \frac{|f(y) - f(z)|}{\|y - z\|} < \infty,$$

and *strictly differentiable at* $\bar{x}$ if $\nabla f(\bar{x})$ exists and

$$\lim_{\substack{y,z\to\bar{x}\\ y\neq z}} \frac{f(y) - f(z) - \langle \nabla f(\bar{x}), y - z\rangle}{\|y - z\|} = 0.$$

With $C^{1+}(\mathbb{R}^n)$ and $C^{1,1}(\mathbb{R}^n)$, we indicate the set of functions from $\mathbb{R}^n$ to $\mathbb{R}$ with locally and globally Lipschitz continuous gradient, respectively. If f is strictly continuous in an open set O, then its gradient exists almost everywhere on O, and as such its *Bouligand subdifferential*

$$\partial_B f(x) := \left\{v \mid \exists x^k \to x \text{ with } \nabla f(x^k) \to v\right\}$$

is nonempty and compact for all $x \in O$ [55, Thm. 9.61].

For a point-to-set mapping $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, the set of its *fixed points* and *zeros* are denoted as $\mathbf{fix}\, F := \{x \in \mathbb{R}^n \mid x \in F(x)\}$ and $\mathbf{zer}\, F := \{x \in \mathbb{R}^n \mid 0 \in F(x)\}$, respectively.

2.2. **Relative smoothness and hypoconvexity.** Here, after giving some definitions, we establish necessary facts regarding relative smoothness and hypoconvexity.

Definition 2.1. *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper, lsc, convex function with $\mathbf{int}\,\mathbf{dom}\, h \neq \emptyset$ and such that $h \in C^1(\mathbf{int}\,\mathbf{dom}\, h)$. Then, h is said to be*

(i) *a* kernel function *if it is* 1-coercive*, i.e., $\mathbf{lim}_{\|x\|\to\infty} {}^{h(x)}/_{\|x\|} = \infty$;*

(ii) essentially smooth*, if $\|\nabla h(x_k)\| \to \infty$ for every sequence $(x_k)_{k\in\mathbb{N}} \subseteq \mathbf{int}\,\mathbf{dom}\, h$ converging to a boundary point of $\mathbf{dom}\, h$;*

(iii) *of* Legendre *type if it is essentially smooth and strictly convex.*

Definition 2.2 (Bregman distance [23])**.** *For a kernel function h, the* Bregman distance *$\mathbf{D}_h : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ is given by*

$$\mathbf{D}_h(x, y) := \begin{cases} h(x) - h(y) - \langle \nabla h(y), x - y\rangle & \text{if } y \in \mathbf{int}\,\mathbf{dom}\, h, \\ \infty & \text{otherwise.} \end{cases} \tag{2.1}$$

If h is a strictly convex kernel function, then $\mathbf{D}_h$ serves as a pseudo-distance, having $\mathbf{D}_h \geq 0$ and $\mathbf{D}_h(x, y) = 0$ iff $x = y \in \mathbf{int}\,\mathbf{dom}\, h$. In general, however, $\mathbf{D}_h$ is nonsymmetric and fails to satisfy the triangular inequality. There are many popular kernel functions such as energy, Boltzmann-Shannon entropy, Fermi-Dirac entropy and so on leading to variant Bregman distances that appear in many applications; see, e.g., [11, Ex. 2.3].

Remark 2.3. The following assertions hold:

(i) If $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is of Legendre type and 1-coercive, then $h^* \in C^1(\mathbb{R}^n)$ is strictly convex. In fact, $\nabla h : \mathbf{int}\,\mathbf{dom}\, h \to \mathbb{R}^n$ is a (continuous) bijection with $\nabla h^{-1} = \nabla h^*$ [56, Thm. 26.5, Cor. 13.3.1].

(ii) If $h \in C^2$ is of Legendre type and $\nabla^2 h \succ 0$ on $\mathbf{int}\,\mathbf{dom}\, h$, then $h^* \in C^2(\mathbb{R}^n)$ [54].

(iii) $\mathbf{D}_h$ is continuous on $\mathbf{int}\,\mathbf{dom}\, h \times \mathbf{int}\,\mathbf{dom}\, h$.

(iv) $\mathbf{D}_h(\cdot\,, x)$ and $\mathbf{D}_h(x, \cdot\,)$ are level bounded locally uniformly in x [10, Lem. 7.3(v)-(viii)].[3]

(v) If ∇h is $\tilde{\sigma}_h$-strongly monotone on an open convex set $\mathcal{U} \subseteq \mathbf{int}\,\mathbf{dom}\, h$, then $\mathbf{D}_h(y, x) \geq \frac{\tilde{\sigma}_h}{2}\|y - x\|^2$ for all $x, y \in \mathcal{U}$.

(vi) If ∇h is $\tilde{L}_h$-Lipschitz on an open convex set $\mathcal{U} \subseteq \mathbf{int}\,\mathbf{dom}\, h$, then $\mathbf{D}_h(y, x) \leq \frac{\tilde{L}_h}{2}\|y-x\|^2$ for all $x, y \in \mathcal{U}$. □

We will sometimes require properties such as Lipschitz differentiability or strong convexity to hold locally, where locality amounts to the existence for any point of a neighborhood in which such property holds.

Definition 2.4 (relative smoothness [9, 46])**.** *We say that a proper and lsc function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is* smooth relative to *a kernel $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ if there exists $L_f \geq 0$ such that $L_f h - f$ and $L_f h + f$ are convex functions. Whenever h is clear from context, we will simply say that f is relatively smooth, or L_f-relatively smooth to make the modulus L_f explicit.*

Whenever there exists $\sigma_f \in \mathbb{R}$ such that function $f - \sigma_f h$ is convex, we will say that f is *σ_f-hypoconvex relative to h*. In particular, any L_f-relatively smooth function f is also σ_f-relatively hypoconvex with $\sigma_f = -L_f$. There are however cases in which a tighter (*i.e.*, larger) σ_f can be considered, as it is the case of convex functions f for which $\sigma_f \geq 0$. As we will formally elaborate in Proposition 2.6, the modulus $L_f \geq 0$ provides some upper bounding inequalities on f, whereas the modulus $\sigma_f \in [-L_f, L_f]$ provides lower bounds.

Proposition 2.5. *Let f be smooth relative to a kernel h. Then, the following hold:*

(i) $f \in C^1(\mathbf{int}\,\mathbf{dom}\, h)$*;*

(ii) if h is Lipschitz differentiable on an open set $\mathcal{U}$, then so is f.

Proof.

♠ 2.5*(i)* Convexity of $L_f h \pm f$ and continuous differentiability of h on $\mathbf{int}\,\mathbf{dom}\, h$ ensure that $\mathbf{dom}\, f \supseteq \mathbf{int}\,\mathbf{dom}\, h$, and through [55, Ex. 8.20(b) and Cor. 9.21] that both f and $-f$ are subdifferentially regular on $\mathbf{int}\,\mathbf{dom}\, h$, in the sense of [55, Def. 7.25], with $\hat{\partial} f$ and $\hat{\partial}(-f)$ both nonempty. The proof now follows from [55, Thm. 9.18(d)].

♠ 2.5*(ii)* Let $\tilde{L}_h$ be a Lipschitz modulus for ∇h on $\mathcal{U}$. Convexity of $f - \sigma_f h$ yields

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \sigma_f \langle \nabla h(x) - \nabla h(y), x - y\rangle \geq \mathbf{min}\{\sigma_f, 0\}\tilde{L}_h\|x - y\|^2$$

for $x, y \in \mathcal{U}$, while due to concavity of $f - L_f h$ it holds that

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \leq L_f \langle \nabla h(x) - \nabla h(y), x - y\rangle \leq L_f \tilde{L}_h\|x - y\|^2.$$

The two inequality together prove that f is $\tilde{L}_f$-smooth and $\tilde{\sigma}_f$-hypoconvex (in the classical Euclidean sense) with $\tilde{L}_f = L_f \tilde{L}_h$ and $\tilde{\sigma}_f = \tilde{L}_h\, \mathbf{min}\{\sigma_f, 0\}$. □

[3] Although [10] only states level boundedness, a trivial modification of the proof shows local uniformity too.

The proof of the following result is a a simple adaptation of that of [46, Prop. 1.1].

Proposition 2.6 (characterization of relative smoothness and hypoconvexity)**.** *The following assertions are equivalent for a proper lsc function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$:*

(a) f is L_f-smooth and σ_f-hypoconvex relative to h;

(b) $\sigma_f \mathbf{D}_h(y, x) \le f(y) - f(x) - \langle \nabla f(x), y - x\rangle \le L_f \mathbf{D}_h(y, x)$ for all $x, y \in \mathbf{int}\,\mathbf{dom}\, h$;

(c) $\sigma_f\langle \nabla h(x) - \nabla h(y), x - y\rangle \le \langle \nabla f(x) - \nabla f(y), x - y\rangle \le L_f\langle \nabla h(x) - \nabla h(y), x - y\rangle$ for all $x, y \in \mathbf{int}\,\mathbf{dom}\, h$;

(d) $\sigma_f \nabla^2 h \preceq \nabla^2 f \preceq L_f \nabla^2 h$ on $\mathbf{int}\,\mathbf{dom}\, h$, provided that $f, h \in C^2(\mathbf{int}\,\mathbf{dom}\, h)$.

2.3. **Bregman proximal mapping.** Let us now recall the definition of Bregman proximal mapping and Bregman-Moreau envelope and some of their fundamental properties.

Definition 2.7 (Bregman proximal mapping and Moreau envelope)**.** *Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper and lsc and let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a kernel function. The* Bregman proximal mapping *is the set-valued mapping $\mathbf{prox}^h_{\gamma g} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ defined as*

$$\mathbf{prox}^h_{\gamma g}(x) := \operatorname*{\mathbf{arg\,min}}_{z\in\mathbb{R}^n} \Big\{ g(z) + \tfrac{1}{\gamma}\mathbf{D}_h(z, x)\Big\}, \tag{2.2}$$

while the Bregman-Moreau envelope *is the single-valued function $g^{h/\gamma} : \mathbb{R}^n \to \overline{\mathbb{R}}$ defined as*

$$g^{h/\gamma}(x) := \inf_{z\in\mathbb{R}^n} \Big\{ g(z) + \tfrac{1}{\gamma}\mathbf{D}_h(z, x)\Big\}. \tag{2.3}$$

Some results of the paper will make use of an important connection relating the proximal mapping and Moreau envelope as defined here with a Bregman kernel with similar objects in the Euclidean case. To ease the notation, the kernel h will be omitted in the Euclidean case $h = \frac{1}{2}\|\cdot\|^2$, and thus write $g^{1/\gamma}$ and $\mathbf{prox}_{\gamma g}$ to indicate the γ-Moreau envelope function and γ-proximal point mapping for $h = \frac{1}{2}\|\cdot\|^2$.

Definition 2.8 (h-prox-boundedness)**.** *Given a kernel h, a function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is said to be* h-prox-bounded *if there exists $\gamma > 0$ such that $g^{h/\gamma}(x) > -\infty$ for some $x \in \mathbb{R}^n$. The supremum of the set of all such γ is the threshold γ^h_g of the h-prox-boundedness, i.e.,*

$$\gamma^h_g := \sup\Big\{\gamma > 0 \mid \exists x \in \mathbb{R}^n \text{ s.t. } g^{h/\gamma}(x) > -\infty\Big\}.$$

3. Bregman forward-backward mapping

We now introduce the forward-backward operator with a Bregman kernel, and analyze some properties of its fixed points. We first make a technical observation that allows to drop prox-boundedness requirements in the sequel.

Remark 3.1. Notice that Assumption I ensures that g is h-prox-bounded with threshold $\gamma^h_g \ge 1/L_f$. In fact, for any $x \in \mathbf{int}\,\mathbf{dom}\, h$ concavity of $f - L_f h$ yields

$$\begin{aligned}\mathbf{min}\,\varphi \le f(z) + g(z) &= (f - L_f h)(z) + L_f h(z) + g(z)\\ &\le (f - L_f h)(x) + \langle \nabla(f - L_f h)(x), z - x\rangle + L_f h(z) + g(z)\\ &= f(x) + \langle \nabla f(x), z - x\rangle + g(z) + L_f\,\mathbf{D}_h(z, x).\end{aligned}$$

Thus, for any $\gamma < 1/L_f$ the function $g + \frac{1}{\gamma}\mathbf{D}_h(\,\cdot\,, x)$ is lower bounded, owing to 1-coercivity of h (which entails that of $\mathbf{D}_h(\,\cdot\,, x)$). □

The L_f-relative smoothness of f implies through Prop. 2.6*(b)* that for any $\gamma \in (0, 1/L_f)$ the function $\mathcal{M}^{f,g}_{h/\gamma}(z, x) : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ given by

$$\mathcal{M}^{f,g}_{h/\gamma}(z, x) := f(x) + \langle \nabla f(x), z - x\rangle + \tfrac{1}{\gamma}\mathbf{D}_h(z, x) + g(z) \tag{3.1}$$

provides a majorization model for φ, *i.e.*, $\varphi(z) \le \mathcal{M}^{f,g}_{h/\gamma}(z, x)$. More specifically,

$$\varphi(z) + \tfrac{1-\gamma L_f}{\gamma}\mathbf{D}_h(z, x) \le \mathcal{M}^{f,g}_{h/\gamma}(z, x) \le \varphi(z) + \tfrac{1-\gamma\sigma_f}{\gamma}\mathbf{D}_h(z, x) \quad \forall x, z. \tag{3.2}$$

The *Bregman forward-backward splitting* mapping is the (set-valued) *majorization-minimization* operator $\mathbf{T}^{f,g}_{h/\gamma} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ defined as

$$\mathbf{T}^{f,g}_{h/\gamma}(x) := \operatorname*{\mathbf{arg\,min}}_{z\in\mathbb{R}^n} \mathcal{M}^{f,g}_{h/\gamma}(z, x), \tag{3.3}$$

and the *Bregman forward-backward residual* mapping is $\mathbf{R}^{f,g}_{h/\gamma} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ given by

$$\mathbf{R}^{f,g}_{h/\gamma}(x) := x - \mathbf{T}^{f,g}_{h/\gamma}(x). \tag{3.4}$$

3.1. **Stationarity, criticality, and optimality.** We investigate different aspects of suboptimality and show their connection to $\varphi = f + g$, similarly to what has been done in [66] for the Euclidean case.

Definition 3.2. *Relative to problem* (1.1)*, we say that a point* $x_\star \in \mathbf{dom}\,\varphi$ *is*

(i) stationary *if* $0 \in \hat{\partial}\varphi(x_\star)$*;*

(ii) critical *if it is* γ*-critical for some* $\gamma > 0$*, namely if* $x_\star \in \mathbf{T}^{f,g}_{h/\gamma}(x_\star)$*;*

(iii) optimal *if* $x_\star \in \operatorname{\mathbf{arg\,min}} \varphi$*, i.e.,* $x_\star$ *is a solution of* (1.1)*.*

As we will see in Proposition 3.5, criticality is a halfway property between stationarity and optimality. In fact, the higher the value of γ the more restrictive the property of γ-criticality. As a measure of this suboptimality, we thus introduce the *criticality threshold*.

Definition 3.3. *Relative to problem* (1.1)*, the* h-criticality threshold *is the function* $\mathbf{\Gamma}^{f,g}_h : \mathbb{R}^n \to [0, \gamma^h_g]$ *defined by*

$$\mathbf{\Gamma}^{f,g}_h(x) := \mathbf{sup}\left(\left\{\gamma > 0 \mid x \in \mathbf{T}^{f,g}_{h/\gamma}(x)\right\} \cup \{0\}\right).$$

Note that for $\gamma > \gamma^h_g$ the mapping $\mathbf{T}^{f,g}_{h/\gamma}$ is empty-valued, hence the inclusion $\mathbf{\Gamma}^{f,g}_h \in [0, \gamma^h_g]$. In the next two results, we show how the three notions introduced in Definition 3.2 are interrelated and identify some useful properties of critical points that will be used to derive regularity of $\mathbf{T}^{f,g}_{h/\gamma}$ and $\mathbf{R}^{f,g}_{h/\gamma}$. Although simple adaptations of [66, Thm. 3.4 and Prop. 3.5] where the Euclidean case $h = \frac{1}{2}\|\cdot\|^2$ is considered, for self containedness we detail the proofs.

Proposition 3.4 (critical point characterization)**.** *The following hold for a point* $x_\star \in \mathbb{R}^n$*:*

(i) $x_\star$ *is* γ*-critical iff* $g(x) \geq g(x_\star) - \langle \nabla f(x_\star), x - x_\star\rangle - \frac{1}{\gamma}\mathbf{D}_h(x, x_\star)$ *for all* $x \in \mathbf{int}\,\mathbf{dom}\,h$*;*

(ii) if $x_\star$ *is critical, then it is* γ*-critical for all* $\gamma \in (0, \mathbf{\Gamma}^{f,g}_h(x_\star))$*;*

(iii) $\mathbf{T}^{f,g}_{h/\gamma}(x_\star) = \{x_\star\}$ *and* $\mathbf{R}^{f,g}_{h/\gamma}(x_\star) = \{0\}$ *for any* $\gamma \in (0, \mathbf{\Gamma}^{f,g}_h(x_\star))$*.*

Proof.

♠ 3.4*(i)* By definition, $x_\star$ is γ-critical iff $\mathcal{M}^{f,g}_{h/\gamma}(x_\star, x_\star) \leq \mathcal{M}^{f,g}_{h/\gamma}(x, x_\star)$ for all x, *i.e.*, iff

$$f(x_\star) + g(x_\star) \leq f(x_\star) + \langle \nabla f(x_\star), x - x_\star\rangle + g(x) + \tfrac{1}{\gamma}\mathbf{D}_h(x, x_\star) \quad \forall x \in \mathbb{R}^n.$$

By suitably rearranging, the claim readily follows.

♠ 3.4*(ii)* Directly follows from assert 3.4*(i)*.

♠ 3.4*(iii)* Let $x_\star$ be a critical point, and let $x \in \mathbf{T}^{f,g}_{h/\gamma}(x_\star)$ for some $\gamma < \Gamma(x_\star)$. Fix $\gamma' \in (\gamma, \Gamma(x_\star))$. From 3.4*(i)* and 3.4*(ii)*, it then follows that

$$g(x) \geq g(x_\star) + \langle -\nabla f(x_\star), x - x_\star\rangle - \tfrac{1}{\gamma'}\mathbf{D}_h(x, x_\star). \tag{3.5}$$

Since $x, x_\star \in \mathbf{T}^{f,g}_{h/\gamma}(x_\star)$, it holds that $\mathcal{M}^{f,g}_{h/\gamma}(x_\star, x_\star) = \mathcal{M}^{f,g}_{h/\gamma}(x, x_\star)$, *i.e.*,

$$g(x_\star) = \langle \nabla f(x_\star), x - x_\star\rangle + \tfrac{1}{\gamma}\mathbf{D}_h(x, x_\star) + g(x) \overset{(3.5)}{\geq} g(x_\star) + (\tfrac{1}{\gamma} - \tfrac{1}{\gamma'})\mathbf{D}_h(x, x_\star).$$

Since $\frac{1}{\gamma} - \frac{1}{\gamma'} > 0$, necessarily $x = x_\star$. □

Proposition 3.5 (optimality, criticality, and stationarity)**.** *The following statements hold:*

(i) (criticality $\Rightarrow$ stationarity) $\mathbf{fix}\, \mathbf{T}^{f,g}_{h/\gamma} \subseteq \mathbf{zer}\, \hat{\partial}\varphi$ *for all* $\gamma \in (0, \gamma_g^h)$;

(ii) (optimality $\Rightarrow$ criticality) $\mathbf{\Gamma}^{f,g}_h(x_\star) \geq 1/L_f$ *for all* $x_\star \in \mathbf{arg\,min}\, \varphi$.

Proof.

♠ 3.5*(i)* Let $\gamma \in (0, \gamma_g^h)$ and $x \in \mathbf{fix}\, \mathbf{T}^{f,g}_{h/\gamma}$. Since x minimizes $\mathcal{M}^{f,g}_{h/\gamma}(\,\cdot\,, x)$, we have

$$0 \in \hat{\partial}[\mathcal{M}^{f,g}_{h/\gamma}(\,\cdot\,, x)](x) = \nabla f(x) + \tfrac{1}{\gamma}[\nabla h(x) - \nabla h(x)] + \hat{\partial} g(x) = \hat{\partial} g(x) + \nabla f(x) = \hat{\partial}\varphi(x),$$

where the inclusion follows from [55, Thm. 10.1] and the equalities from [55, Thm. 8.8(c)].

♠ 3.5*(ii)* Fix $\gamma \in (0, 1/L_f)$, $x_\star \in \mathbf{arg\,min}\, \varphi$ and $z \in \mathbf{T}^{f,g}_{h/\gamma}(x_\star)$. Then, (3.2) with $x = x_\star$ yields

$$\begin{aligned}\varphi(z) &\leq \mathcal{M}^{f,g}_{h/\gamma}(z, x_\star) - \tfrac{1-\gamma L_f}{\gamma} \mathbf{D}_h(z, x_\star) = \min_w \mathcal{M}^{f,g}_{h/\gamma}(w, x_\star) - \tfrac{1-\gamma L_f}{\gamma} \mathbf{D}_h(z, x_\star) \\ &\leq \mathcal{M}^{f,g}_{h/\gamma}(x_\star, x_\star) - \tfrac{1-\gamma L_f}{\gamma} \mathbf{D}_h(z, x_\star) = \varphi(x_\star) - \tfrac{1-\gamma L_f}{\gamma} \mathbf{D}_h(z, x_\star),\end{aligned}$$

where the first equality is due to the inclusion $z \in \mathbf{T}^{f,g}_{h/\gamma}(x_\star)$. We then conclude that $z = x_\star$, for otherwise $\varphi(z) < \varphi(x_\star)$ would contradict global minimality of $x_\star$. □

4. Properties of Bregman forward-backward envelope

In this section, we first introduce the *Bregman forward-backward envelope* (BFBE) and study its fundamental properties that we need in subsequent sections. Let us define the function BFBE $\varphi^{f,g}_{h/\gamma} : \mathbb{R}^n \to \overline{\mathbb{R}}$ as

$$\varphi^{f,g}_{h/\gamma}(x) := \inf_{z \in \mathbb{R}^n} \left\{ f(x) + \langle \nabla f(x), z - x \rangle + g(z) + \tfrac{1}{\gamma} \mathbf{D}_h(z, x) \right\}, \tag{4.1}$$

namely the value function of the minimization subproblem defining the Bregman forward-backward mapping (3.3). We next give alternative expressions for $\mathbf{T}^{f,g}_{h/\gamma}$ and $\varphi^{f,g}_{h/\gamma}$.

Proposition 4.1 (alternative expressions of $\mathbf{T}^{f,g}_{h/\gamma}$ and $\varphi^{f,g}_{h/\gamma}$)**.** *Suppose that Assumption I is satisfied. Then, the following hold:*

(i) $\mathbf{T}^{f,g}_{h/\gamma}(x) = \mathbf{prox}^h_{\gamma g}(\nabla h^*(\nabla h(x) - \gamma \nabla f(x)))$ *for every* $x \in \mathbf{int\,dom}\, h$;

(ii) $\frac{h}{\gamma} - f$ *is a Legendre kernel,* $\varphi^{f,g}_{h/\gamma} = \varphi^{\frac{h}{\gamma} - f}$ *and* $\mathbf{T}^{f,g}_{h/\gamma} = \mathbf{prox}^{\frac{h}{\gamma} - f}_{\varphi}$.

Proof.

♠ 4.1*(i)* By expanding the Bregman distance and discarding constant terms in (3.3), one has

$$\mathbf{T}^{f,g}_{h/\gamma}(x) = \operatorname*{arg\,min}_z \left\{ g(z) + \tfrac{1}{\gamma}[h(z) - \langle \nabla h(x) - \gamma \nabla f(x), z - x \rangle] \right\} = \operatorname*{arg\,min}_z \left\{ g(z) + \tfrac{1}{\gamma} \mathbf{D}_h(z, \bar{z}) \right\}$$

for $\bar{z} = \nabla h^*(\nabla h(x) - \gamma \nabla f(x))$, owing to the identity $\nabla h(x) - \gamma \nabla f(x) = \nabla h(\bar{z})$ (cf. Rem. 2.3*(i)*).

♠ 4.1*(ii)* Let us first show that $\frac{1}{\gamma} h - f$ is a Legendre kernel. Let $\hat{h} := \frac{1}{\gamma} h - f$ and $C := \mathbf{int\,dom}\, h$. Since $\hat{h} = \frac{1-\gamma L_f}{\gamma} h + L_f h - f$ and $L_f h - f$ is convex, strict convexity and 1-coercivity of $\hat{h}$ follow from the similar properties of h. We now show essential smoothness; to arrive to a contradiction, suppose that there exists a sequence $(x^k)_{k \in \mathbb{N}} \subset C$ converging to a boundary point $x_\star$ of C and such that $\sup_{k \in \mathbb{N}} \|\nabla \hat{h}(x^k)\| < \infty$. By possibly extracting, we may assume that $\nabla h(x^k)/\|\nabla h(x^k)\| \to v$ for some vector v with unitary norm. For every $y \in C$, since $\nabla(L_f h - f)(x^k) = \nabla \hat{h}(x^k) - \frac{1-\gamma L_f}{\gamma} \nabla h(x^k)$, it holds that

$$\langle \nabla(L_f h - f)(x^k) - \nabla(L_f h - f)(y), x^k - y \rangle \leq c - \tfrac{1-\gamma L_f}{\gamma} \langle \nabla h(x^k) - \nabla h(y), x^k - y \rangle, \tag{4.2}$$

where $c := \sup \langle \nabla \hat{h}(x^k) - \nabla \hat{h}(y), x^k - y \rangle$ is a finite quantity. Moreover, $0 \leq \frac{1}{\|\nabla h(x^k)\|} \langle \nabla h(x^k) - \nabla h(y), x^k - y \rangle \to \langle v, x_\star - y \rangle$ as $k \to \infty$, and from the arbitrarity of $y \in C$ we conclude that

$v \in \{u \mid \langle u, x_\star - y\rangle \geq 0 \ \forall y \in C\}$. Since C is open, $\mathbf{B}(x_\star;\varepsilon) \cap C \neq \emptyset$ for any $\varepsilon > 0$, and in particular there exists $y \in C$ such that $\langle v, x_\star - y\rangle \gneqq 0$. Pluggin this y in (4.2) yields

$$\langle \nabla(L_f h - f)(x^k) - \nabla(L_f h - f)(y), x^k - y\rangle \leq c - \tfrac{1-\gamma L_f}{\gamma}\|\nabla h(x^k)\|\langle \tfrac{\nabla h(x^k)-\nabla h(y)}{\|\nabla h(x^k)\|}, x^k - y\rangle \to -\infty,$$

contradicting convexity of $L_f h - f$. Therefore, $\frac{1}{\gamma}h - f$ is a Legendre kernel. Now, observe that

$$\begin{aligned}
\mathbf{T}^{f,g}_{h/\gamma}(x) &= \operatorname*{\mathbf{arg\,min}}_z \left\{ f(x) + \langle \nabla f(x), z - x\rangle + g(z) + \tfrac{1}{\gamma}\mathbf{D}_h(z,x)\right\} \\
&= \operatorname*{\mathbf{arg\,min}}_z \left\{ f(z) + \tfrac{1}{\gamma}[(h-\gamma f)(z) - (h-\gamma f)(x) - \langle \nabla(h-\gamma f)(x), z - x\rangle] + g(z)\right\} \\
&= \operatorname*{\mathbf{arg\,min}}_z \left\{ \varphi(z) + \tfrac{1}{\gamma}\mathbf{D}_{h-\gamma f}(z,x)\right\},
\end{aligned}$$

hence the identity $\mathbf{T}^{f,g}_{h/\gamma} = \mathbf{prox}^{\frac{h}{\gamma}-f}_{\varphi}$. The same reasoning with the infima proves also the identity $\varphi^{f,g}_{h/\gamma} = \varphi^{\frac{h}{\gamma}-f}$, completing the proof. □

The next two results characterize the fundamental relationship between the Bregman forward-backward envelope $\varphi^{f,g}_{h/\gamma}$ and the original function φ that are essential to analyze the convergence of the Bregman forward-backward scheme that will be given in Section 5.

Theorem 4.2 (relation between φ and $\varphi^{f,g}_{h/\gamma}$)**.** *Under conditions given in Assumption I and $\gamma \in (0, 1/L_f)$, the following statements are true:*

(i) $\varphi^{f,g}_{h/\gamma} \leq \varphi$;

(ii) $\frac{1-\gamma L_f}{\gamma}\mathbf{D}_h(\bar{x},x) \leq \varphi^{f,g}_{h/\gamma}(x) - \varphi(\bar{x}) \leq \frac{1-\gamma\sigma_f}{\gamma}\mathbf{D}_h(\bar{x},x)$ *for all* $x \in \mathbb{R}^n$ *and* $\bar{x} \in \mathbf{T}^{f,g}_{h/\gamma}(x)$;

(iii) $\varphi^{f,g}_{h/\gamma}(x) = \varphi(x)$ *iff* $x \in \mathbf{T}^{f,g}_{h/\gamma}(x)$;

(iv) $\mathbf{inf}\,\varphi^{f,g}_{h/\gamma} = \mathbf{inf}\,\varphi$ *and* $\operatorname*{\mathbf{arg\,min}}\varphi^{f,g}_{h/\gamma} = \operatorname*{\mathbf{arg\,min}}\varphi$;

Proof.

♠ 4.2*(i)* Follows from [36, Prop. 2.1(i)] in light of the identification of Prop. 4.1*(ii)*.

♠ 4.2*(ii)* & 4.2*(iii)* Follow from (3.2) with $z = \bar{x}$ (since $\varphi^{f,g}_{h/\gamma}(x) = \mathcal{M}^{f,g}_{h/\gamma}(\bar{x},x)$) and from the fact that $\mathbf{D}_h(\bar{x},x) \geq 0$ for every $x, \bar{x}$ with equality holding iff $x = \bar{x}$.

♠ 4.2*(iv)* It follows from 4.2*(i)* that $\mathbf{inf}\,\varphi^{f,g}_{h/\gamma} \leq \mathbf{inf}\,\varphi$. Let $(x^k)_{k\in\mathbb{N}}$ be such that $\varphi^{f,g}_{h/\gamma}(x^k) \to \mathbf{inf}\,\varphi^{f,g}_{h/\gamma}$ as $k \to \infty$. Then, taking $\bar{x}^k \in \mathbf{T}^{f,g}_{h/\gamma}(x^k)$ assert 4.2*(ii)* ensures that $\varphi(\bar{x}^k) \to \mathbf{inf}\,\varphi$, hence the claimed equivalence of infima.

If $x \in \operatorname*{\mathbf{arg\,min}}\varphi^{f,g}_{h/\gamma}$ and $\bar{x} \in \mathbf{T}^{f,g}_{h/\gamma}(x)$ then

$$\varphi(\bar{x}) \overset{4.2(ii)}{\leq} \varphi^{f,g}_{h/\gamma}(x) - \tfrac{1-\gamma L_f}{\gamma}\mathbf{D}_h(\bar{x},x) = \mathbf{inf}\,\varphi^{f,g}_{h/\gamma} - \tfrac{1-\gamma L_f}{\gamma}\mathbf{D}_h(\bar{x},x) = \mathbf{inf}\,\varphi - \tfrac{1-\gamma L_f}{\gamma}\mathbf{D}_h(\bar{x},x),$$

hence necessarily $x = \bar{x} \in \operatorname*{\mathbf{arg\,min}}\varphi$. Similarly, for $x \in \operatorname*{\mathbf{arg\,min}}\varphi$ one has

$$\varphi^{f,g}_{h/\gamma}(x) \overset{4.2(i)}{\leq} \varphi(x) = \mathbf{inf}\,\varphi = \mathbf{inf}\,\varphi^{f,g}_{h/\gamma},$$

hence $x \in \operatorname*{\mathbf{arg\,min}}\varphi^{f,g}_{h/\gamma}$. □

Proposition 4.3 (regularity of $\mathbf{T}^{f,g}_{h/\gamma}$ and $\varphi^{f,g}_{h/\gamma}$)**.** *Suppose that Assumption I holds. Then, for every $\gamma \in (0, \gamma^h_g)$ the following statements are true:*

(i) $\mathbf{dom}\,\mathbf{T}^{f,g}_{h/\gamma} = \mathbf{dom}\,\varphi^{f,g}_{h/\gamma} = \mathbf{int}\,\mathbf{dom}\,h$;

(ii) $\mathbf{range}\,\mathbf{T}^{f,g}_{h/\gamma} \subseteq \mathbf{dom}\,\varphi \subseteq \mathbf{int}\,\mathbf{dom}\,h$;

(iii) $\mathbf{T}^{f,g}_{h/\gamma}$ *and* $\mathbf{R}^{f,g}_{h/\gamma}$ *are osc, locally bounded and compact-valued;*

(iv) $\varphi^{f,g}_{h/\gamma}$ *is real valued and continuous on* $\mathbf{int}\,\mathbf{dom}\,h$*; if additionally* h *is* C^{1+} *on* $\mathbf{int}\,\mathbf{dom}\,h$*, then* $\varphi^{f,g}_{h/\gamma}$ *is locally Lipschitz continuous on* $\mathbf{int}\,\mathbf{dom}\,h$*.*

Proof. The first inclusion of 4.3*(ii)* follows from Thm.s 4.2*(i)* and 4.2*(ii)*, and the second one from Assumption I$_{A4}$. The other claims can invoke similar ones for the Bregman-Moreau envelope and Bregman proximal mapping, owing to the identification of Prop. 4.1*(ii)*:

♠ 4.3*(i)* See [36, Thm. 2.2(ii)].

♠ 4.3*(iv)* Follow from [36, Cor. 2.2 and Thm. 2.3] in light of Prop. 2.5*(ii)*.

♠ 4.3*(iii)* Compact-valuedness and osc are shown in [36, Thm. 2.2(i)]. Further, the continuity of $\varphi^{f,g}_{h/\gamma}$ entails its local upper boundedness; as such, local boundedness of $\mathbf{T}^{f,g}_{h/\gamma}$ follows from [55, Ex. 5.22], which applies as ensured by [36, Thm. 2.1]. □

We now show two other important properties relating the BFBE with the original cost function, namely equivalence of level boundedness and of (strong/local) minimality.

Theorem 4.4 (equivalence of level boundedness)**.** *Suppose that Assumption I is satisfied, and let* $\gamma \in (0, 1/L_f)$*. Then,* $\varphi^{f,g}_{h/\gamma}$ *is level bounded iff so is* φ*.*

Proof. It follows from 4.2*(i)* that if $\varphi^{f,g}_{h/\gamma}$ is level bounded then so is φ. Conversely, suppose that there exists $\alpha \in \mathbb{R}$ together with an unbounded sequence $(x^k)_{k\in\mathbb{N}} \subseteq [\varphi^{f,g}_{h/\gamma} \leq \alpha]$. Then, it follows from Prop. 4.3*(i)* that $x^k \in \mathbf{dom}\,\varphi^{f,g}_{h/\gamma} = \mathbf{int}\,\mathbf{dom}\,h$ for all k, and in turn that for any k there exists $\bar{x}^k \in \mathbf{T}^{f,g}_{h/\gamma}(x^k)$ which satisfies $\varphi(\bar{x}^k) \leq \alpha - \frac{1-\gamma L_f}{\gamma}\mathbf{D}_h(\bar{x}^k, x^k)$, as it follows from Thm. 4.2*(ii)*. Local boundedness of $\mathbf{D}_h$ with respect to the second variable (Rem. 2.3*(iv)*) then ensures that $(\bar{x}^k)_{k\in\mathbb{N}}$ is not bounded, hence that φ is not level bounded. □

Theorem 4.5 (equivalence of local minimality)**.** *Additionally to Assumption I, suppose that the following requirements are satisfied:*

A1 $x_\star$ *is (a critical point) such that* $\mathbf{T}^{f,g}_{h/\gamma}(x_\star) = \{x_\star\}$ *for some* $\gamma < 1/L_f$ *(as it is the case when* $\gamma < \mathbf{\Gamma}^{f,g}_h(x_\star)$*, cf. Prop. 3.4(iii));*

A2 h *is* $\tilde{\sigma}_h$*-strongly convex around* $x_\star$ *for some* $\tilde{\sigma}_h > 0$*.*

Then, $x_\star$ *is a (strong) local minimum for* $\varphi^{f,g}_{h/\gamma}$ *iff it is a (strong) local minimum for* φ*.*

Proof. That (strong) local minimality for $\varphi^{f,g}_{h/\gamma}$ implies that for φ follows from the fact that $\varphi^{f,g}_{h/\gamma}$ "supports" φ at $x_\star$, namely that $\varphi^{f,g}_{h/\gamma} \leq \varphi$ and $\varphi^{f,g}_{h/\gamma}(x_\star) = \varphi(x_\star)$ (Thm. 4.2*(i)*). Conversely, suppose that there exists $\mu \geq 0$ such that $\varphi(x) \geq \varphi(x_\star) + \frac{\mu}{2}\|x - x_\star\|^2$ for all x sufficiently close to $x_\star$. Let $\delta := \frac{1}{2}\,\mathbf{min}\left\{\mu, \frac{\tilde{\sigma}_h}{2}\frac{1-\gamma L_f}{\gamma}\right\} \geq 0$, and note that $\delta = 0$ iff $\mu = 0$. Thus, contrary to the claim suppose that for all $k \in \mathbb{N}_{\geq 1}$ there exists $x^k \in \mathbf{B}(x_\star; 1/k)$ such that $\varphi^{f,g}_{h/\gamma}(x^k) < \varphi^{f,g}_{h/\gamma}(x_\star) + \frac{\delta}{2}\|x^k - x_\star\|^2$. Let $\bar{x}^k \in \mathbf{T}^{f,g}_{h/\gamma}(x^k)$; since $\mathbf{T}^{f,g}_{h/\gamma}$ is osc and $\mathbf{T}^{f,g}_{h/\gamma}(x_\star) = \{x_\star\}$, necessarily $\bar{x}^k \to x_\star$ as $k \to \infty$. We have

$$\begin{aligned}\varphi(\bar{x}^k) &\overset{4.2(i)}{\leq} \varphi^{f,g}_{h/\gamma}(x^k) - \tfrac{1-\gamma L_f}{\gamma}\mathbf{D}_h(\bar{x}^k, x^k) \overset{2.3(v)}{\leq} \varphi^{f,g}_{h/\gamma}(x^k) - \tfrac{\tilde{\sigma}_h}{2}\tfrac{1-\gamma L_f}{\gamma}\|x^k - \bar{x}^k\|^2 \\ &< \varphi^{f,g}_{h/\gamma}(x_\star) + \tfrac{\delta}{2}\|x^k - x_\star\|^2 - \tfrac{\tilde{\sigma}_h}{2}\tfrac{1-\gamma L_f}{\gamma}\|x^k - \bar{x}^k\|^2 \\ &= \varphi(x_\star) + \tfrac{\delta}{2}\|x^k - x_\star\|^2 - \tfrac{\tilde{\sigma}_h}{2}\tfrac{1-\gamma L_f}{\gamma}\|x^k - \bar{x}^k\|^2.\end{aligned}$$

By using the inequality $\frac{1}{2}\|a - c\|^2 \leq \|a - b\|^2 + \|b - c\|^2$ holding for any $a, b, c \in \mathbb{R}^n$, we have

$$\varphi(\bar{x}^k) < \varphi(x_\star) + \delta\|\bar{x}^k - x_\star\|^2 + (\delta - \tfrac{\tilde{\sigma}_h}{2}\tfrac{1-\gamma L_f}{\gamma})\|x^k - \bar{x}^k\|^2 \leq \varphi(x_\star) + \tfrac{\mu}{2}\|\bar{x}^k - x_\star\|^2,$$

where the last inequality follows from the definition of δ. Thus, $\varphi(\bar{x}^k) < \varphi(x_\star) + \frac{\mu}{2}\|\bar{x}^k - x_\star\|^2$ for all $k \in \mathbb{N}$, which contradicts μ-strong local minimality of $x_\star$ for φ (since $\bar{x}^k \to x_\star$). □

It was first observed in [42] that the Euclidean forward-backward envelope can be interpreted as a Bregman-Moreau envelope. The following theorem furnishes a local converse relation, namely that when h is locally strongly convex and locally Lipschitz differentiable the Bregman FBE, hence in particular the Bregman-Moreau envelope, can locally be identified with a Euclidean FBE. This equivalence is a key certificate for analyzing the local properties of BFBE (in particular Bregman-Moreau envelope) close to critical points under the prox-regularity condition using the existing analysis of FBE; cf. [53, 64]. To do so, we first state a technical lemma, whose proof is an immediate consequence of the osc and local boundedness of $\mathbf{T}^{f,g}_{h/\gamma}$ together with the inclusion $\mathbf{range}\,\mathbf{T}^{f,g}_{h/\gamma} \subseteq \mathbf{int}\,\mathbf{dom}\,h$ shown in Prop. 4.3.

Lemma 4.6. *Suppose that Assumption I is satisfied, and let $\mathcal{U} \subset \mathbf{int}\,\mathbf{dom}\,h$ be a compact set. Then $\mathbf{T}^{f,g}_{h/\gamma}(\mathcal{U}) \subset \mathbf{int}\,\mathbf{dom}\,h$ is also compact.*

Theorem 4.7 (local equivalence of Bregman and Euclidean FBE)**.** *Suppose that Assumption I holds and let $\gamma < 1/L_f$ be fixed. Suppose further that h is locally Lipschitz differentiable and locally strongly convex (as it is the case when $h \in \mathcal{C}^2$ with $\nabla^2 h \succ 0$ on $\mathbf{int}\,\mathbf{dom}\,h$). Then, for all $\tilde{\gamma} > 0$ and all bounded sets $\mathcal{U}$ such that $\mathbf{cl}(\mathcal{U}) \subset \mathbf{int}\,\mathbf{dom}\,h$ one has*

$$\varphi^{f,g}_{h/\gamma} = \varphi^{\tilde{f},\tilde{g}}_{1/\tilde{\gamma}} \quad \text{and} \quad \mathbf{T}^{f,g}_{h/\gamma} = \mathbf{T}^{\tilde{f},\tilde{g}}_{1/\tilde{\gamma}} \ \text{on } \mathcal{U}, \tag{4.3}$$

where

$$\tilde{f} \coloneqq f - \tfrac{1}{\gamma}h + \tfrac{1}{2\tilde{\gamma}}\|\cdot\|^2 \quad \text{and} \quad \tilde{g} \coloneqq g + \tfrac{1}{\gamma}h - \tfrac{1}{2\tilde{\gamma}}\|\cdot\|^2 + \delta_B \tag{4.4}$$

for some nonempty and compact set $B \subseteq \mathbf{dom}\,h$. Moreover, $\tilde{g}$ is proper, lsc, and prox-bounded (in the Euclidean sense) with $\gamma_{\tilde{g}} = \infty$, and for $\tilde{\gamma}$ small enough $\tilde{f}$ is $L_{\tilde{f}}$-Lipschitz-differentiable on $\mathbf{dom}\,\tilde{g}$ with $\tilde{\gamma} < 1/L_{\tilde{f}}$.

Proof. It follows from Lem. 4.6 that $B \coloneqq \mathbf{T}^{f,g}_{h/\gamma}(\mathcal{U})$ is compact. We have

$$\begin{aligned}
\varphi^{f,g}_{h/\gamma}(x) &= \min_{z\in B}\left\{f(x) + \langle \nabla f(x), z - x\rangle + \tfrac{1}{\gamma}\,\mathbf{D}_h(z,x) + g(z)\right\} \\
&= \min_{z\in B}\left\{f(x) - \tfrac{1}{\gamma}h(x) + \langle \nabla f(x) - \tfrac{1}{\gamma}\nabla h(x), z - x\rangle + g(z) + \tfrac{1}{\gamma}h(z)\right\} \\
&= \min_{z\in \mathbb{R}^n}\left\{\tilde{f}(x) + \langle \nabla \tilde{f}(x), z - x\rangle + \tilde{g}(z) + \tfrac{1}{2\tilde{\gamma}}\|z - x\|^2\right\} = \varphi^{\tilde{f},\tilde{g}}_{1/\tilde{\gamma}}(x).
\end{aligned}$$

Notice that $\tilde{g}$ is proper, lsc and with bounded domain, hence its claimed prox-boundedness. Let now $\Omega \coloneqq \mathbf{conv}(\mathcal{U} \cup \mathbf{T}^{f,g}_{h/\gamma}(\mathcal{U}))$, and observe that $\mathbf{cl}\,\Omega$ is a bounded subset of $\mathbf{int}\,\mathbf{dom}\,h$. In fact, boundedness follows from that of $\mathcal{U}$ and its image under $\mathbf{T}^{f,g}_{h/\gamma}$, and the inclusion from convexity of $\mathbf{int}\,\mathbf{dom}\,h$. Thus, h is $L_{h,\Omega}$-smooth and $\sigma_{h,\Omega}$-strongly convex on Ω for some constants $L_{h,\Omega}, \sigma_{h,\Omega} > 0$. Then, from the equalities

$$\tilde{f} = f - L_f h - \tfrac{1-\gamma L_f}{\gamma}h + \tfrac{1}{2\tilde{\gamma}}\|\cdot\|^2 = f + L_f h - \tfrac{1+\gamma L_f}{\gamma}h + \tfrac{1}{2\tilde{\gamma}}\|\cdot\|^2,$$

the convexity of $f + L_f h$, and the concavity of $f - L_f h$, it follows that

$$\left(\tfrac{1}{\tilde{\gamma}} - \tfrac{1+\gamma L_f}{\gamma}\sigma_{h,\Omega}\right)\|x - y\|^2 \le \langle \nabla \tilde{f}(x) - \nabla \tilde{f}(y), x - y\rangle \le \left(\tfrac{1}{\tilde{\gamma}} - \tfrac{1-\gamma L_f}{\gamma}L_{h,\Omega}\right)\|x - y\|^2$$

for every $x, y \in \Omega$. Therefore, $\tilde{f}$ is $L_{\tilde{f}}$-smooth on B, with

$$L_{\tilde{f}} = \max\left\{\left|\tfrac{1}{\tilde{\gamma}} - \tfrac{1+\gamma L_f}{\gamma}\sigma_{h,\Omega}\right|,\ \left|\tfrac{1}{\tilde{\gamma}} - \tfrac{1-\gamma L_f}{\gamma}L_{h,\Omega}\right|\right\}.$$

Imposing $\tilde{\gamma} < 1/L_{\tilde{f}}$ yields $\tilde{\gamma} < \frac{1}{2\gamma}\min\left\{(1 - \gamma L_f)L_{h,\Omega},\ (1 + \gamma L_f)\sigma_{h,\Omega}\right\}$, hence the claim. □

4.1. **First-order properties.** We here discuss first-order properties of the BFBE. Let us begin with a subdifferential inclusion that extends known facts about the Bregman-Moreau envelope [36].

Proposition 4.8. *Additionally to Assumption I, suppose that $f, h \in C^2(\mathbf{int\,dom}\, h)$. Then, $\varphi^{f,g}_{h/\gamma}$ is strictly continuous on $\mathbf{int\,dom}\, h$ and is strictly differentiable wherever it is differentiable. Moreover,*

(i) $\mathbf{lip}\, \varphi^{f,g}_{h/\gamma}(x) = \mathbf{max}_{r \in \mathbf{R}^{f,g}_{h/\gamma}(x)} \left\| \left(\frac{1}{\gamma}\nabla^2 h(x) - \nabla^2 f(x) \right) r \right\|$;

(ii) $\partial \varphi^{f,g}_{h/\gamma}(x) = \partial_B \varphi^{f,g}_{h/\gamma}(x) \subseteq \left(\frac{1}{\gamma}\nabla^2 h(x) - \nabla^2 f(x) \right) \mathbf{R}^{f,g}_{h/\gamma}(x)$.

Proof. As shown in Prop. 4.1*(ii)*, $\varphi^{f,g}_{h/\gamma} = \varphi^{\hat{h}}$ with $\hat{h} := \frac{1}{\gamma}h - f$. We will pattern the proof of [55, Ex. 10.32], and thus observe that in light of Lem. 4.6 for every open set $O \subseteq \mathbf{cl}\, O \subseteq \mathbf{int\,dom}\, h$ there exists a compact set $Y \subseteq \mathbf{int\,dom}\, h$ such that $-\varphi^{\hat{h}}(x) = \mathbf{max}_{y \in Y} \Phi(x, y)$, where $\Phi(x, y) := -\varphi(y) - \frac{1}{\gamma} \mathbf{D}_{\hat{h}}(y, x)$ is continuously differentiable in x, its derivatives depending continuously on (y, x) with $\nabla_x \Phi(y, x) = \nabla^2 \hat{h}(x)(y - x)$. In fact, the maxima are attained for $y \in \mathbf{T}^{f,g}_{h/\gamma}(x)$. Function $-\varphi^{\hat{h}}$ is thus *lower-C^1* in the sense of [55, Def. 10.29], so that [55, Def. 10.31] ensures its strict continuity, the equivalence of differentiability and strict differentiability, and the relations

$$\mathbf{lip}\, \varphi^{\hat{h}}(x) = \max_{\bar{x} \in \mathbf{T}^{f,g}_{h/\gamma}(x)} \left\| \nabla^2 \hat{h}(x)(x - \bar{x}) \right\| \quad \text{and} \quad \partial \varphi^{\hat{h}}(x) = \partial_B \varphi^{\hat{h}}(x) \subseteq \nabla^2 \hat{h}(x) \left(x - \mathbf{T}^{f,g}_{h/\gamma}(x) \right).$$

The proof now follows from the identities $\varphi^{f,g}_{h/\gamma} = \varphi^{\hat{h}}$ and $\hat{h} = \frac{1}{\gamma}h - f$. □

Although strict continuity ensures almost everywhere differentiability, with mild additional assumptions the BFBE can be shown to be (Lipschitz-continuously) differentiable around critical points. Thanks to the local equivalence shown in Theorem 4.7, these requirements are the same as those ensuring similar properties in the Euclidean case. These amount to prox-regularity, a condition which was first proposed in [53] that we state next.

Definition 4.9 (prox-regularity)**.** *A function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is prox-regular at $\bar{x} \in \mathbf{int\,dom}\, h$ for $\bar{v} \in \partial g(\bar{x})$ if it is locally lsc at $\bar{x}$ and there exists $r, \varepsilon > 0$ such that*

$$g(x') \geq g(x) + \langle v, x' - x \rangle - \tfrac{r}{2}\|x' - x\|^2 \tag{4.5}$$

holds for all $x, x' \in \mathbf{B}(\bar{x}; \varepsilon)$ and $(x, v) \in \mathbf{gph}\, \partial g$ with $v \in \mathbf{B}(\bar{v}; \varepsilon)$ and $g(x) \leq g(\bar{x}) + \varepsilon$.

In order to ease the terminology, since prox-regularity will only be needed at critical points $x_\star$ for $v = -\nabla f(x_\star)$, we will introduce a slight abuse of notation and define prox-regularity *of critical points* as follows.

Definition 4.10 (prox-regularity of critical points)**.** *Relative to problem (1.1), we say that a critical point $x_\star$ is prox-regular if g is prox-regular at $x_\star$ for $-\nabla f(x_\star)$.*

The subsequent result connects prox-regularity of g in (1.1) with the first-order properties of BFBE, owing to the relation of BFBE and the Euclidean forward-backward envelope given in Theorem 4.7. To shorten the notation, we introduce the matrix-valued mapping

$$Q^f_{h/\gamma}(x) := \tfrac{1}{\gamma}\nabla^2 h(x) - \nabla^2 f(x), \tag{4.6}$$

defined wherever it makes sense.

Theorem 4.11 (continuous differentiability of BFBE)**.** *Suppose that Assumption I holds and that $h \in C^2$ with $\nabla^2 h \succ 0$ on $\mathbf{int\,dom}\, h$. Suppose further that f is of class C^2 around a prox-regular critical point $x_\star$. Then, for all $\gamma \in (0, \mathbf{\Gamma}^{f,g}_h(x_\star))$ there exists a neighborhood $\mathcal{U}$ of $x_\star$ on which the following statements are true:*

(i) *$\mathbf{T}^{f,g}_{h/\gamma}$ and $\mathbf{R}^{f,g}_{h/\gamma}$ are Lipschitz continuous (hence single valued);*

(ii) *$\varphi^{f,g}_{h/\gamma} \in C^{1+}$ with $\nabla \varphi^{f,g}_{h/\gamma} = Q_\gamma \mathbf{R}^{f,g}_{h/\gamma}$, where $Q^f_{h/\gamma}$ is as in (4.6).*

Proof. For any compact neighborhood $\mathcal{U} \subset \mathbf{int\,dom}\, h$ of $x_\star$ we may invoke Thm. 4.7 and identify $\varphi^{f,g}_{h/\gamma}$ with the Euclidean FBE $\varphi^{\tilde f,\tilde g}_{1/\tilde\gamma}$ on $\mathcal{U}$, for some $\tilde\gamma$ small enough and with $\tilde f$ and $\tilde g$ as in (4.4). It follows from [55, Ex. 13.35] and the continuous differentiability of f and h that $\tilde g$ is prox-regular at $x_\star$ for $-\nabla\tilde f(x_\star)$. Up to possibly restricting $\mathcal{U}$ so that f is twice continuously differentiable there, we may thus invoke the similar result in the Euclidean [66, Thm. 4.7] to infer that $\mathbf{T}^{\tilde f,\tilde g}_{1/\tilde\gamma}$ is Lipschitz continuous around $x_\star$ and the Euclidean FBE $\varphi^{\tilde f,\tilde g}_{1/\tilde\gamma}$ is Lipschitz differentiable around $x_\star$ with

$$\nabla\varphi^{f,g}_{h/\gamma} = \nabla\varphi^{\tilde f,\tilde g}_{1/\tilde\gamma} = \tilde\gamma^{-1}[\mathrm{I} - \tilde\gamma\nabla^2\tilde f\,](\mathrm{id} - \mathbf{T}^{\tilde f,\tilde g}_{1/\tilde\gamma}) = Q^f_{h/\gamma}(\mathrm{id} - \mathbf{T}^{f,g}_{h/\gamma}), \tag{4.7}$$

where the last equality follows from the fact that

$$\tilde\gamma^{-1}[\mathrm{I} - \tilde\gamma\nabla^2\tilde f] = \tilde\gamma^{-1}\mathrm{I} - \nabla^2(f - \tfrac{1}{\gamma}h + \tfrac{1}{2\tilde\gamma}\|\cdot\|^2) = \tfrac{1}{\gamma}\nabla^2 h - \nabla^2 f = Q^f_{h/\gamma}, \tag{4.8}$$

together with the identity $\mathbf{T}^{\tilde f,\tilde g}_{1/\tilde\gamma} = \mathbf{T}^{f,g}_{h/\gamma}$, cf. (4.3) and (4.4). This last identity in particular proves local Lipschitz continuity of $\mathbf{T}^{f,g}_{h/\gamma}$ and $\mathbf{R}^{f,g}_{h/\gamma}$, as claimed. □

We note that the results presented in Theorem 4.11 cover those of Bregman-Moreau envelope by setting $f = 0$, which has been studied in [61, Thm. 4.1] and [7, Prop. 3.12] for jointly convex Bregman distances in the convex setting, as discussed also in [11, Prop. 2.19]. Differently from the global continuous differentiability result in [36, Cor. 3.1] which requires global convexity of $h + \gamma g$, ours is a local result that requires local properties of g around critical points.

Corollary 4.12 (continuous differentiability of the Bregman-Moreau envelope)**.** *Suppose that h is a Legendre kernel twice continuously differentiable with $\nabla^2 h \succ 0$ on* **int dom** *h. Let g be a proper, lsc function and let $\gamma > 0$ and $x_\star$ be such that $\mathbf{prox}_{\gamma g}(x_\star) = \{x_\star\}$. If g is prox-regular at $x_\star$ for 0, then there exists a neighborhood $\mathcal{U}$ of $x_\star$ on which the following hold:*

(i) $\mathbf{prox}^h_{\gamma g}$ *is Lipschitz continuous (hence single valued);*

(ii) $g^{h/\gamma} \in C^{1+}$ *with* $\nabla g^{h/\gamma}(x) = \gamma^{-1}\nabla^2 h(x)(x - \mathbf{prox}^h_{\gamma g}(x))$.

4.2. **Second-order properties.** We now investigate sufficient conditions ensuring twice differentiability of $\varphi^{f,g}_{h/\gamma}$ at critical points, which will be needed in Section 5.3 to show superlinear convergence of the proposed algorithm under a Dennis-Moré condition. To do so, the following extra assumption is required.

Assumption II (second-order properties)**.** *Function h is twice continuously differentiable with $\nabla^2 h \succ 0$ on* **int dom** *h, and relative to a given critical point $x_\star$, we have that*

A1 *$\nabla^2 f$ exists and is (strictly) continuous around $x_\star$;*

A2 *g is prox-regular and (strictly) twice epi-differentiable at $x_\star$ for $-\nabla f(x_\star)$, with its second-order epi-derivative being generalized quadratic:*

$$\mathrm{d}^2 g(x_\star|-\nabla f(x_\star))(w) = \langle w, Mw\rangle + \delta_S(w) \quad \forall w \in \mathbb{R}^n, \tag{4.9}$$

where $M \in \mathbb{R}^{n\times n}$ and $S \subseteq \mathbb{R}^n$ is a linear subspace. Without loss of generality, we take M symmetric, and such that **range** *$M \subseteq S$ and* **ker** *$M \supseteq S^\perp$.*[4]

The assumptions are "strictly" satisfied if the stronger conditions in parentheses hold.

Theorem 4.13 (twice differentiability of $\varphi^{f,g}_{h/\gamma}$)**.** *Suppose that Assumption I holds, that Assumption II is (strictly) satisfied with respect to a critical point $x_\star$, and let $Q^f_{h/\gamma}$ be as in (4.6). Then, for $\gamma \in (0, \mathbf{\Gamma}^{f,g}_h(x_\star))$, the following statements are true:*

[4]This can indeed be done without loss of generality: if M and S satisfy (4.9), then it suffices to replace M with $M' = \frac{1}{2}\,\mathbf{Proj}_S(M + M^\top)\mathbf{Proj}_S$ to ensure the desired properties.

(i) $\mathbf{prox}^h_{\gamma g} \circ \nabla h^*$ *is (strictly) differentiable at* $\nabla h(x_\star) - \gamma \nabla f(x_\star)$ *with symmetric positive semidefinite Jacobian* $P_\gamma(x_\star) \coloneqq J(\mathbf{prox}^h_{\gamma g} \circ \nabla h^*)(\nabla h(x_\star) - \gamma \nabla f(x_\star))$*;*

(ii) $\mathbf{R}^{f,g}_{h/\gamma}$ *is (strictly) differentiable at* $x_\star$ *with Jacobian* $J\,\mathbf{R}^{f,g}_{h/\gamma}(x_\star) = x_\star - P_\gamma(x_\star)Q_\gamma(x_\star)$*;*

(iii) $\varphi^{f,g}_{h/\gamma}$ *is (strictly) twice differentiable at* $x_\star$ *with symmetric Hessian*

$$\nabla^2 \varphi^{f,g}_{h/\gamma}(x_\star) = Q_\gamma(x_\star) J\,\mathbf{R}^{f,g}_{h/\gamma}(x_\star).$$

Proof. As shown in the proof of Thm. 4.13, for some $\tilde\gamma$ small enough and with $\tilde f$ and $\tilde g$ as in (4.4) we may identify $\varphi^{f,g}_{h/\gamma}$ with the Euclidean FBE $\varphi^{\tilde f,\tilde g}_{1/\tilde\gamma}$ around $x_\star$. It follows from [55, Ex.s 13.18 and 13.25] and the continuous differentiability of f and h that Assumption II remains valid if one replaces γ, f, g and h respectively with $\tilde\gamma, \tilde f, \tilde g$ and $\tilde h = \frac{1}{2}\|\cdot\|^2$. We may thus invoke the similar result in the Euclidean case [66, Thm. 4.10] to infer that

(iv) $\mathbf{prox}_{\tilde\gamma\tilde g}$ is (strictly) differentiable at $x_\star - \tilde\gamma\nabla\tilde f(x_\star)$ with symmetric positive semidefinite Jacobian $\tilde P_\gamma(x_\star) \coloneqq J\,\mathbf{prox}_{\tilde\gamma\tilde g}(x_\star - \tilde\gamma\nabla\tilde f(x_\star))$,

(v) $\mathbf{R}^{\tilde f,\tilde g}_{1/\tilde\gamma}$ is (strictly) differentiable at $x_\star$ with Jacobian $J\,\mathbf{R}^{\tilde f,\tilde g}_{1/\tilde\gamma}(x_\star) = x_\star - \tilde P_{\tilde\gamma}(x_\star)Q^{\tilde f}_{1/\tilde\gamma}(x_\star)$,

(vi) and $\varphi^{\tilde f,\tilde g}_{1/\tilde\gamma}$ is (strictly) twice differentiable at $x_\star$ with symmetric Hessian $\nabla^2\varphi^{\tilde f,\tilde g}_{1/\tilde\gamma}(x_\star) = Q^{\tilde f}_{1/\tilde\gamma} J\,\mathbf{R}^{\tilde f,\tilde g}_{1/\tilde\gamma}(x_\star)$,

where $Q^{\tilde f}_{1/\tilde\gamma} = \tilde\gamma^{-1}\mathrm{I} - \nabla^2\tilde f$. From the identities $J\,\mathbf{R}^{\tilde f,\tilde g}_{1/\tilde\gamma} = J\,\mathbf{R}^{f,g}_{h/\gamma}$ and $Q^{\tilde f}_{1/\tilde\gamma} = Q^f_{h/\gamma}$ (cf. (4.8)), the claimed expression for $\nabla^2\varphi^{f,g}_{h/\gamma}(x_\star)$ follows. Moreover, the chain rule of differentiation yields

$$J\,\mathbf{T}^{f,g}_{h/\gamma}(x_\star) = J(\mathbf{prox}_{\gamma g} \circ \nabla h^*)(\nabla h(x_\star) - \gamma\nabla f(x_\star))Q^f_{h/\gamma}(x_\star) = P_\gamma(x_\star)Q^f_{h/\gamma}(x_\star)$$

while from 4.13*(v)*

$$J\,\mathbf{T}^{\tilde f,\tilde g}_{1/\tilde\gamma}(x_\star) = \tilde P_{\tilde\gamma}(x_\star)Q^{\tilde f}_{1/\tilde\gamma}(x_\star) \overset{(4.8)}{=} \tilde P_{\tilde\gamma}(x_\star)Q^f_{h/\gamma}(x_\star),$$

and from the identity $J\,\mathbf{T}^{f,g}_{h/\gamma}(x_\star) = J\,\mathbf{T}^{\tilde f,\tilde g}_{1/\tilde\gamma}(x_\star)$, the invertibility of Q_γ and 4.13*(iv)*, we conclude that indeed $\mathbf{prox}^h_{\gamma g} \circ \nabla h^*$ is (strictly) differentiable at $\nabla h(x_\star) - \gamma\nabla f(x_\star)$ with symmetric and positive semidefinite Jacobian $P_\gamma(x_\star) = \tilde P_{\tilde\gamma}(x_\star)$. The expression in 4.13*(ii)* then follows from the identity $J\,\mathbf{R}^{f,g}_{h/\gamma} = \mathrm{id} - \mathbf{T}^{f,g}_{h/\gamma}$. □

Setting $f = 0$, one immediately obtains a similar result for the Bregman-Moreau envelope. Clearly, twice differentiability of h^* ensured by Remark 2.3*(ii)* allows to infer (strict) differentiability of $\mathbf{prox}^h_{\gamma g}$ from that of $\mathbf{prox}^h_{\gamma g} \circ \nabla h^*$.

Corollary 4.14 (twice differentiability of the Bregman-Moreau envelope)**.** *Suppose that* h *is a Legendre kernel twice continuously differentiable with* $\nabla^2 h \succ 0$ *on* $\mathbf{int}\,\mathbf{dom}\,h$*. Let* g *be a proper, lsc function and let* $\gamma > 0$ *and* $x_\star$ *be such that* $\mathbf{prox}_{\gamma g}(x_\star) = \{x_\star\}$*. If* g *is prox-regular and (strictly) twice epi-differentiable at* $x_\star$ *for* $-\nabla f(x_\star)$*, with its second-order epi-derivative being generalized quadratic (see* (4.9)*), then*

(i) $\mathbf{prox}^h_{\gamma g}$ *is (strictly) differentiable at* $x_\star$*, and* $\mathbf{prox}^h_{\gamma g} \circ \nabla h^*$ *is (strictly) differentiable at* $\nabla h(x_\star)$ *and has symmetric and positive semidefinite Jacobian there;*

(ii) $g^{h/\gamma}$ *is (strictly) twice differentiable at* $x_\star$ *with symmetric Hessian*

$$\nabla^2 g^{h/\gamma}(x_\star) = \tfrac{1}{\gamma}\nabla^2 h(x_\star)[x_\star - J\,\mathbf{prox}^h_{\gamma g}(x_\star)].$$

4.3. **Kurdyka-Łojasiewicz property and local nonlinear error bound.** We conclude this section by giving a discussion on KL property and a nonlinear error bound for the BFBE which are essential tools for our algorithm in the next section.

Definition 4.15 (KL property)**.** *A proper lsc function* $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ *is said to have the* Kurdyka-Łojasiewicz *property (KL property) at* $x_\star \in \mathbf{dom}\,F$ *if there exist a concave* desingularizing function $\psi : [0, \eta] \to [0, \infty)$ *(for some* $\eta > 0$*) and an* $\varepsilon > 0$ *such that*

P1 $\psi(0) = 0$;

P2 ψ *is of class* C^1 *on* $(0, \eta)$;

P3 *for all* $x \in \mathbf{B}(x_\star; \varepsilon)$ *such that* $F(x_\star) < F(x) < F(x_\star) + \eta$ *it holds that*

$$\psi'(F(x) - F(x_\star))\,\mathbf{dist}(0, \partial F(x)) \geq 1. \tag{4.10}$$

The first inequality of this type is given in the seminal work of Łojasiewicz [44, 45] for analytic functions, which we nowadays call Łojasiewicz's gradient inequality. Later, Kurdyka [38] showed that this inequality is valid for C^1 functions whose graph belongs to an *o-minimal structure* [69, 68]. The first extensions of the KL property to nonsmooth tame functions was given in [14, 15, 16].

In the subsequent proposition, we show that the functions $g^{h/\gamma}$ and $\varphi^{f,g}_{h/\gamma}$ and the mappings $\mathbf{prox}^h_{\gamma g}$ and $\mathbf{T}^{f,g}_{h/\gamma}$ are semialgebraic provided that f, g and h are, thus ensuring $g^{h/\gamma}$ and $\varphi^{f,g}_{h/\gamma}$ to satisfy the KL inequality. Although the proof can be generalized to *tame* functions [24], for simplicity we restrict the analysis to the semialgebraic case.

Proposition 4.16. *Additionally to Assumption I, suppose that* f, g *and* h *are semialgebraic. Then, the following statements are true:*

(i) $g^{h/\gamma}$ *and* $\varphi^{f,g}_{h/\gamma}$ *are semialgebraic functions and in particular have the KL property with desingularizing function* $\psi(s) = \varrho s^\vartheta$, *for some* $\varrho > 0$ *and* $\vartheta \in (0, 1)$;

(ii) $\mathbf{prox}^h_{\gamma g}$ *and* $\mathbf{T}^{f,g}_{h/\gamma}$ *are semialgebraic mappings.*

Proof. Our arguments follow known properties of semialgebraic mappings, see *e.g.*, [34, §8.3.1]. Since f, g and h are semialgebraic, $g(z) + \frac{1}{2\gamma}\mathbf{D}_h(z, x)$ and $\mathcal{M}^{f,g}_{h/\gamma}(z, x)$ are semialgebraic. Moreover, since the parametric minimization of a semialgebraic function is still semialgebraic, it follows that $g^{h/\gamma}$ and $\varphi^{f,g}_{h/\gamma}$ are semialgebraic and consequently satisfy the KL property with desingularizing function of the form $\psi(s) = \varrho s^\vartheta$ [14, 15]. Moreover, notice that

$$\begin{aligned}
\mathbf{T}^{f,g}_{h/\gamma}(x) &= \left\{z \in \mathbb{R}^n \mid g(z) + f(x) + \langle \nabla f(x), z - x\rangle + \tfrac{1}{\gamma}\mathbf{D}_h(z, x) \leq \varphi^{f,g}_{h/\gamma}(x)\right\} \\
&= \left\{z \in \mathbb{R}^n \mid g(z) + f(x) + \langle \nabla f(x), z - x\rangle + \tfrac{1}{\gamma}\mathbf{D}_h(z, x) - \varphi^{f,g}_{h/\gamma}(x) \leq 0\right\} \\
&= \{z \in \mathbb{R}^n \mid q(x, z) \leq 0\},
\end{aligned}$$

where $q(x, z) \coloneqq g(z) + f(x) + \langle \nabla f(x), z - x\rangle + \frac{1}{2\gamma}\mathbf{D}_h(z, x) - \varphi^{f,g}_{h/\gamma}(x)$ is a semialgebraic function. Then, the graph of $\mathbf{T}^{f,g}_{h/\gamma}(\cdot)$ is given by

$$\mathbf{gph}\,\mathbf{T}^{f,g}_{h/\gamma} = \left\{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n \mid y \in \mathbf{T}^{f,g}_{h/\gamma}(x)\right\} = \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n \mid q(x, y) \leq 0\} = q^{-1}((-\infty, 0]),$$

proving that $\mathbf{T}^{f,g}_{h/\gamma}$, and thus $\mathbf{prox}^h_{\gamma g}$ for $f = 0$, are semialgebraic mappings. □

In [6] a desingularizing property stronger than the KL inequality is investigated, namely with $\mathbf{dist}(0, \partial F(x))$ being replaced by the *strong slope* $|\nabla F|(x) \coloneqq \mathbf{lim\,sup}_{x \neq z \to x} \frac{F(x) - F(z)}{\|x - z\|} \leq \mathbf{dist}(0, \partial F(x))$, and it is then related to a nonlinear growth condition on function F of the form $\psi(F(x) - F(x_\star)) \geq \mathbf{dist}(x, [F \leq F(x_\star)])$ [6, Thm. 4.1]. Whenever F is continuously differentiable around $x_\star$, both the strong slope and minimum norm subgradient reduce to the norm of the gradient, so that the KL property and the one given in [6] coincide. Using the differentiability properties of the BFBE under the assumptions of Theorem 4.11*(ii)*, we can thus specialize this result in the next two lemmas that will be used in Section 5.3 as a key tool for guaranteeing a superlinear convergence to nonisolated critical points of φ given in (1.1).

Lemma 4.17 (nonlinear error bound [6, Thm. 4.1])**.** *Additionally to Assumption I, suppose that the following requirements are satisfied:*

A1 h *is of class* C^2 *with* $\nabla^2 h \succ 0$ *on* $\mathbf{int}\,\mathbf{dom}\,h$;

A2 *f is of class C^2 around a prox-regular critical point $x_\star$;*

A3 *$\gamma \in (0, \mathbf{\Gamma}_h^{f,g}(x_\star))$;*

A4 *$\varphi_{h/\gamma}^{f,g}$ has the KL property at $x_\star$ with desingularizing function ψ.*

Then, denoting $\varphi_\star := \varphi_{h/\gamma}^{f,g}(x_\star)$ there exist $\varepsilon, \eta > 0$ such that

$$\psi(\varphi_{h/\gamma}^{f,g}(x) - \varphi_\star) \geq \mathbf{dist}(x, [\varphi_{h/\gamma}^{f,g} \leq \varphi_\star]) \quad \forall x \in \mathbf{B}(x_\star; \varepsilon) \text{ such that } \varphi_\star < \varphi_{h/\gamma}^{f,g}(x) < \varphi_\star + \eta.$$

Whenever the desingularizing function can be taken of the form $\psi(s) = \varrho s^\vartheta$ with $\varrho > 0$ an $\vartheta \in (0, 1)$, it is usually referred to as Łojasiewicz function (with exponent $1 - \vartheta$). It has been shown in [71, Thm. 5.2] that whenever the kernel function h is twice continuously differentiable and (locally) strongly convex, the function φ admits a Łojasiewicz desingularizing function with exponent $\vartheta \geq 1/2$ iff so does the Bregman envelope φ^h, in which case the exponent is preserved. Clearly, the lower the ϑ the stronger the property, in the sense that whenever φ admits a desingularizing function with exponent $\vartheta \in (0, 1)$, then it also admits a desingularizing function with exponent ϑ' for any $\vartheta' \in [\vartheta, 1)$. Combined with the relation existing among the BFBE and the Bregman-Moreau envelope as in Proposition 4.1*(ii)*, we can specialize the result as follows.

Lemma 4.18 (equivalence of Łojasiewicz property [71, Thm. 5.2])**.** *Suppose that the assumptions of Lemma 4.17 are satisfied, and that the desingularizing function can be taken of the form $\psi(s) = \varrho s^\vartheta$ for some $\varrho > 0$ and $\vartheta \in (0, 1)$. Then, denoting $\bar{\vartheta} := \mathbf{min}\{\vartheta, 1/2\}$ there exists $\bar{\varrho} > 0$ such that $\bar{\psi}(s) = \bar{\varrho} s^{\bar{\vartheta}}$ is a desingularizing function for $\varphi_{h/\gamma}^{f,g}$ at $x_\star$.*

We conclude the section with a simple result showing that the BFBE enjoys a "mild growth" property around critical points.

Lemma 4.19 (mild growth at critical points)**.** *Suppose that Assumption I is satisfied. Then,*

$$\varphi_{h/\gamma}^{f,g}(x) \leq \varphi(z) + \tfrac{1-\gamma\sigma_f}{\gamma} \mathbf{D}_h(z, x) \quad \forall x, z \in \mathbb{R}^n. \tag{4.11}$$

In particular, whenever $x_\star \in \mathbf{T}_{h/\gamma}^{f,g}(x_\star)$ one has

$$\varphi_{h/\gamma}^{f,g}(x) \leq \varphi_\star + \tfrac{1-\gamma\sigma_f}{\gamma} \mathbf{D}_h(x_\star, x) \quad \forall x \in \mathbb{R}^n, \tag{4.12}$$

where $\varphi_\star := \varphi_{h/\gamma}^{f,g}(x_\star) = \varphi(x_\star)$.

Proof. The first inequality directly follows from the definition (4.1) of the BFBE combined with the upper bound in (3.2). In turn, the second inequality follows from the identity $\varphi(x_\star) = \varphi_{h/\gamma}^{f,g}(x_\star)$ holding for any γ-critical point $x_\star$ (Thm. 4.2*(iii)*). □

5. Bregman forward-backward splitting algorithm

In this section, we discuss a Bregman forward-backward algorithm that is accelerated by a linesearch along some suitable directions. The subsequential, global, and local superlinear convergence of a sequence generated by this algorithm are investigated.

Algorithm 1 Bella (**B**regman **E**nve**L**ope **L**inesearch **A**lgorithm)

Require stepsize $\gamma \in (0, 1/L_f)$, initial point $x^0 \in \mathbf{dom}\,\varphi$, tolerance $\varepsilon > 0$, $\sigma \in (0, \frac{1-\gamma L_f}{\gamma})$

Initialize $k = 0$

1: choose $\bar{x}^k \in \mathbf{T}_{h/\gamma}^{f,g}(x^k)$

2: **if** $\mathbf{D}_h(\bar{x}^k, x^k) \leq \varepsilon$ **then return** $\hat{x} := \bar{x}^k$; **end if**

3: choose a direction $d^k \in \mathbb{R}^n$;

4: let $\tau_k \in \left\{2^{-i} \mid i \in \mathbb{N}\right\}$ be the largest such that $x^{k+1} = (1 - \tau_k)\bar{x}^k + \tau_k(x^k + d^k)$ satisfies

$$\varphi_{h/\gamma}^{f,g}(x^{k+1}) \leq \varphi_{h/\gamma}^{f,g}(x^k) - \sigma \mathbf{D}_h(\bar{x}^k, x^k) \tag{5.1}$$

5: $k \leftarrow k + 1$ and go to step 1.

Notice that by setting $d^k = \bar{x}^k - x^k$ BELLA reduces to the Bregman proximal gradient algorithm given in [19] (the linesearch condition (5.1) is satsified regardless of the stepsize τ_k owing to Theorems 4.2*(i)* and 4.2*(ii)*), while for the Euclidean kernel $h(\cdot) = \frac{1}{2}\|\cdot\|$, one obtains the PANOC algorithm given in [58]. Let us begin by showing that step 4 in BELLA is well defined.

Lemma 5.1 (well definedness of the algorithm)**.** *Suppose that Assumption I holds and let* $\gamma \in (0, 1/L_f)$ *and* $\sigma \in (0, \frac{1-\gamma L_f}{\gamma})$ *be fixed. Then, for any* $x \in \mathbf{int}\,\mathbf{dom}\, h$, $\bar{x} \in \mathbf{T}^{f,g}_{h/\gamma}(x) \setminus \{x\}$ *and* $d \in \mathbb{R}^n$ *there exists* $\bar{\tau} \in (0, 1]$ *such that for any* $\tau \in [0, \bar{\tau}]$ *the point* $x^+_\tau := (1-\tau)\bar{x} + \tau(x+d)$ *satisfies*

$$\varphi^{f,g}_{h/\gamma}(x^+_\tau) \le \varphi^{f,g}_{h/\gamma}(x) - \sigma\, \mathbf{D}_h(\bar{x}, x).$$

In particular, since necessarily any such x^+_τ *again belongs to* $\mathbf{int}\,\mathbf{dom}\, h$*, the iterates of* BELLA *are well defined with linesearch at step 4 terminating in a finite number of backtrackings regardless of the choice of directions* $(d^k)_{k\in\mathbb{N}}$.

Proof. It follows from Thm.s 4.2*(i)* and 4.2*(ii)* that the strict inequality

$$\varphi^{f,g}_{h/\gamma}(x') < \varphi^{f,g}_{h/\gamma}(x) - \sigma\, \mathbf{D}_h(\bar{x}, x)$$

holds for $x' = \bar{x}$. Continuity of the envelope $\varphi^{f,g}_{h/\gamma}$ as asserted in Prop. 4.3*(iv)* then ensures that there exists a neighborhood $\mathcal{U}$ of $\bar{x}$ such that the inequality remains valid for any $x' \in \mathcal{U}$. The proof now follows by observing that $x^+_\tau \to \bar{x}$ as $\tau \searrow 0$ for any $d \in \mathbb{R}^n$. That x^+_τ as in the claim belongs to $\mathbf{int}\,\mathbf{dom}\, h$ follows from the inclusion $\mathbf{dom}\,\varphi^{f,g}_{h/\gamma} \subseteq \mathbf{int}\,\mathbf{dom}\, h$ (Prop. 4.3*(i)*). □

Our first main result is about the iteration complexity of BELLA, which is the number of iterations needed to find a point x^k satisfying $\mathbf{D}_h(\bar{x}^k, x^k) \le \varepsilon$.

Theorem 5.2 (iteration complexity of BELLA)**.** *Suppose that Assumption I holds. Then,*

(i) BELLA *terminates within* $k \le \frac{\varphi(x^0) - \mathbf{inf}\,\varphi}{\sigma\varepsilon}$ *iterations;*

(ii) if h *is* $\tilde{\sigma}_h$*-strongly convex and* $\tilde{L}_h$*-Lipschitz differentiable on an open convex set containing all the iterates* x^k *and* $\bar{x}^k$*, then the point* $\hat{x}$ *returned by the algorithm satisfies* $\mathbf{dist}(0, \hat{\partial}\varphi(\hat{x})) \le \frac{1-\gamma\sigma_f}{\gamma}\sqrt{\frac{2\tilde{L}_h}{\tilde{\sigma}_h}\varepsilon}$.

Proof.

♠ 5.2*(i)* By telescoping the linesearch condition (5.1) over the first $K > 0$ iterations we have

$$\sigma \sum_{k=0}^{K-1} \mathbf{D}_h(\bar{x}^k, x^k) \le \sum_{k=0}^{K-1} (\varphi^{f,g}_{h/\gamma}(x^k) - \varphi^{f,g}_{h/\gamma}(x^{k+1})) = \varphi^{f,g}_{h/\gamma}(x^0) - \varphi^{f,g}_{h/\gamma}(x^K) \le \varphi(x^0) - \mathbf{inf}\,\varphi, \quad (5.2)$$

where the last equality follows from Thm.s 4.2*(i)* and 4.2*(iv)*. Since all the iterates up to the $(K-1)$-th satisfy $\mathbf{D}_h(\bar{x}^k, x^k) > \varepsilon$, if $\varepsilon > 0$ necessarily $K \le \frac{\varphi(x^0) - \mathbf{inf}\,\varphi}{\sigma\varepsilon}$ as claimed.

♠ 5.2*(ii)* Let $\mathcal{U}$ be an open convex set containing the sequences and where h is $\tilde{L}_h$-Lipschitz differentiable. For the majorization model $\mathcal{M}^{f,g}_{h/\gamma}$ as in (3.1) and any $x \in \mathcal{U}$ it holds that the majorization gap $\delta_x(w) := \mathcal{M}^{f,g}_{h/\gamma}(w, x) - \varphi(w)$ satisfies

$$\nabla\delta_x(w) = \nabla(\tfrac{1}{\gamma}h - f)(w) - \nabla(\tfrac{1}{\gamma}h - f)(x).$$

By using convexity of $L_f h - f$ and concavity of $\sigma_f h - f$ as in the proof of Prop. 2.5*(ii)*, it is easy to verify that the gradient of the (convex) function $\frac{1}{\gamma}h - f$ is $\frac{1-\gamma\sigma_f}{\gamma}\tilde{L}_h$-Lipschitz continuous on $\mathcal{U}$, hence so is $\nabla\delta_x$ independently of x. The proof can now trace that of [63, Lem. 2.15] in the Euclidean case. Since $\nabla\delta_x(x) = 0$, for any $\bar{x} \in \mathcal{U}$ one has $\|\nabla\delta_x(\bar{x})\| \le \frac{1-\gamma\sigma_f}{\gamma}\tilde{L}_h\|x - \bar{x}\|$. In particular, for $\bar{x} \in \mathbf{T}^{f,g}_{h/\gamma}(x) \cap \mathcal{U}$ one has

$$0 \in \hat{\partial}[\mathcal{M}^{f,g}_{h/\gamma}(\,\cdot\,, x)](\bar{x}) = \hat{\partial}\varphi(\bar{x}) + \nabla\delta_x(\bar{x}),$$

that is, $-\nabla\delta_x(\bar{x}) \in \hat{\partial}\varphi(\bar{x})$. Thus, for $\hat{x} = \bar{x}^k$ as in the last iteration of BELLA one has

$$\mathbf{dist}(0, \hat{\partial}\varphi(\bar{x}^k)) \leq \|-\nabla\delta_{x^k}(\bar{x}^k)\| \leq \tfrac{1-\gamma\sigma_f}{\gamma}\tilde{L}_h\|x^k - \bar{x}^k\| \leq \tfrac{1-\gamma\sigma_f}{\gamma}\sqrt{\tfrac{2\tilde{L}_h}{\tilde{\sigma}_h}\mathbf{D}_h(\bar{x}^k, x^k)} \leq \tfrac{1-\gamma\sigma_f}{\gamma}\sqrt{\tfrac{2\tilde{L}_h}{\tilde{\sigma}_h}}\varepsilon,$$

as claimed. □

5.1. **Subsequential convergence.** We here show the subsequential convergence of the sequence generated by BELLA.

Theorem 5.3 (subsequential convergence)**.** *Suppose that Assumption I holds and consider the iterates generated by* BELLA *with tolerance $\varepsilon = 0$. Then, the following hold:*

(i) $\sum_k \mathbf{D}_h(\bar{x}^k, x^k)$ *is finite.*

(ii) The sequences $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$ have same cluster points, all of which are γ-critical and on which φ and $\varphi^{f,g}_{h/\gamma}$ attain the same finite value $\varphi_\star$, this being the limit of the real-valued sequences $(\varphi^{f,g}_{h/\gamma}(x^k))_{k\in\mathbb{N}}$ and $(\varphi(\bar{x}^k))_{k\in\mathbb{N}}$.

(iii) Both $\mathbf{cl}\left\{x^k \mid k \in \mathbb{N}\right\}$ *and* $\mathbf{cl}\left\{\bar{x}^k \mid k \in \mathbb{N}\right\}$ *are contained in* $\mathbf{int}\,\mathbf{dom}\, h$.

(iv) If φ is level bounded, then $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$ are bounded. If in addition h is locally strongly convex and $\|d^k\| \to 0$ as $k \to \infty$, then their set of accumulation points ω is compact, connected and such that both $\mathbf{dist}(x^k, \omega)$ *and* $\mathbf{dist}(\bar{x}^k, \omega)$ *vanish as $k \to \infty$.*

Proof. To rule out trivialities, let us assume that $\bar{x}^k \neq x^k$ for every $k \in \mathbb{N}$ so that the algorithm runs infinite many iterations.

♠ 5.3*(i)* Readily follows from the fact that the partial sums in (5.2) are bounded by the same finite constant for any $K \in \mathbb{N}$.

♠ 5.3*(ii)* The linesearch condition (5.1) ensures that $(\varphi^{f,g}_{h/\gamma}(x^k))_{k\in\mathbb{N}}$ is decreasing, hence it admits a limit, be it $\varphi_\star$, which is lower bounded by $\mathbf{inf}\,\varphi^{f,g}_{h/\gamma} = \mathbf{inf}\,\varphi$ and is thus finite. Then, necessarily also $\varphi(\bar{x}^k) \to \varphi_\star$ (although not necessarily monotonically), as it follows from Thm. 4.2*(ii)* and the fact that $\mathbf{D}_h(\bar{x}^k, x^k) \to 0$. This last limit together with Rem. 2.3*(iii)* also ensures that $(x^k)_{k\in K} \to x_\star$ iff $(\bar{x}^k)_{k\in K} \to x_\star$, and in particular x^k and $\bar{x}^k$ have same limit points. Since $\bar{x}^k \in \mathbf{T}^{f,g}_{h/\gamma}(x^k)$, if $(x^k)_{k\in K}$ converges to $x_\star$ and thus so does $(\bar{x}^k)_{k\in K}$, from outer semicontinuity of $\mathbf{T}^{f,g}_{h/\gamma}$ (Prop. 4.3*(iii)*) it follows that $x_\star \in \mathbf{T}^{f,g}_{h/\gamma}(x_\star)$. Continuity of $\varphi^{f,g}_{h/\gamma}$ (Prop. 4.3*(iv)*) then implies that $\varphi^{f,g}_{h/\gamma}(x_\star) = \varphi_\star$, and from Thm. 4.2*(ii)* and the fact that $x_\star \in \mathbf{T}^{f,g}_{h/\gamma}(x_\star)$ we conclude that $\varphi(x_\star) = \varphi_\star$ as well.

♠ 5.3*(iii)* Easily follows from the fact that since any limit point of either sequence is critical as claimed in 5.3*(ii)*, it necessarily belongs to $\mathbf{dom}\,\mathbf{T}^{f,g}_{h/\gamma} = \mathbf{int}\,\mathbf{dom}\, h$ (Prop. 4.3*(iii)*).

♠ 5.3*(iv)* The monotonic behavior of $(\varphi^{f,g}_{h/\gamma}(x^k))_{k\in\mathbb{N}}$ ensures that the sequence $(x^k)_{k\in\mathbb{N}}$ is contained in the level set $[\varphi^{f,g}_{h/\gamma} \leq \varphi^{f,g}_{h/\gamma}(x^0)]$, which is bounded as ensured by Thm. 4.4. In fact, it follows from the assertion 5.3*(iii)* that both sequences are contained in a compact set $\Omega \subset \mathbf{int}\,\mathbf{dom}\, h$ that in light of convexity of $\mathbf{dom}\, h$ we may take to be convex. As such, if h is locally strongly convex then it is σ_Ω-strongly convex on Ω for some $\sigma_\Omega > 0$. Then, $\frac{\sigma_\Omega}{2}\|x^k - \bar{x}^k\|^2 \leq \mathbf{D}_h(\bar{x}^k, x^k)$ for all k, and 5.3*(i)* then ensures that $\bar{x}^k - x^k \to 0$. We then have

$$x^{k+1} - x^k = (1-\tau_k)(\bar{x}^k - x^k) + \tau_k d^k \to 0 \quad \text{as } k \to \infty,$$

and since $x^k - \bar{x}^k \to 0$, similarly $\bar{x}^{k+1} - \bar{x}^k \to 0$ as $k \to \infty$. The claimed properties of the sequences now follow from [18, Rem. 5]. □

Remark 5.4 (adaptive variant of BELLA for unknown L_f)**.** If the constant L_f is not available, then it can be retrieved adaptively by initializing it with an estimate $L > 0$ and adding the following instruction after step 1:

1bis: **if** $f(\bar{x}^k) > f(x^k) + \langle\nabla f(x^k), \bar{x}^k - x^k\rangle + L\,\mathbf{D}_h(\bar{x}^k, x^k)$ **then**
$\gamma \leftarrow \gamma/2$, $L \leftarrow 2L$, $\sigma \leftarrow 2\sigma$, and go to step 1.

Whenever L exceeds the actual value L_f, this procedure will terminate and L will be constant starting from that iteration; consequently, L be increased only a finite number of times. Whether or not the final constant L exceeds the actual value L_f, all the claims of Theorem 5.3 remain valid. In order to replicate the proof of Theorem 5.3, it suffices to show that $(\varphi^{f,g}_{h/\gamma}(x^k))_{k\in\mathbb{N}}$ converges to a finite value $\varphi_\star$, which here cannot be inferred from the lower boundedness of $\varphi^{f,g}_{h/\gamma}$ being it ensured only for $\gamma < 1/L_f$ (Theorem 4.2*(iv)*). Nevertheless,

$$\begin{aligned}\inf\varphi &\le f(\bar{x}^k) + g(\bar{x}^k) \le f(x^k) + \langle \nabla f(x^k), \bar{x}^k - x^k\rangle + L\,\mathbf{D}_h(\bar{x}^k, x^k) + g(\bar{x}^k)\\ &= \varphi^{f,g}_{h/\gamma}(x^k) - \tfrac{1-\gamma L}{\gamma}\,\mathbf{D}_h(\bar{x}^k, x^k),\end{aligned}$$

proving that $\varphi_\star \ge \inf\varphi$. □

5.2. **Global and linear convergence.** In this subsection, we provide sufficient conditions ensuring global and linear convergence of the sequence generated by BELLA.

Theorem 5.5 (global convergence)**.** *Suppose that Assumption I is satisfied and consider the iterates generated by* BELLA *with tolerance $\varepsilon = 0$. Suppose further that the following assumptions are satisfied:*

- A1 *φ is level bounded;*
- A2 *$f, h \in C^2$ with $\nabla^2 h \succ 0$ on $\operatorname{int}\operatorname{dom} h$;*
- A3 *$\mathbf{T}^{f,g}_{h/\gamma}(x^k)$ is single valued for k large enough (as it is the case when the limit points of the sequence are prox-regular);*
- A4 *there exists a constant $c \ge 0$ such that $\|d^k\| \le c\|x^k - \bar{x}^k\|$ for all k;*
- A5 *$\varphi^{f,g}_{h/\gamma}$ satisfies the KL property.*

Then, $(\|x^k - \bar{x}^k\|)_{k\in\mathbb{N}}$ is summable and both $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$ converge to a (γ-critical point).

Proof. It follows from Thm.s 5.3*(ii)* and 5.3*(iv)* that $\varphi_\star := \lim_{k\to\infty} \varphi^{f,g}_{h/\gamma}(x^k)$ exists and is finite, and that the set of accumulation points ω of the sequences is nonempty, compact, connected and such that $\operatorname{dist}(x^k, \omega) \to 0$ as $k \to \infty$. It then follows from [18, Lem. 6] that there exist $\eta, \varepsilon > 0$ and a *uniformized KL function*, namely a function ψ satisfying 4.15.P1, 4.15.P2 and 4.15.P3 with $F = \varphi^{f,g}_{h/\gamma}$ for all $x_\star \in \omega$ and x such that $\operatorname{dist}(x, \omega) < \varepsilon$ and $\varphi_\star < \varphi^{f,g}_{h/\gamma}(x) < \varphi_\star + \eta$. This will thus happen for all points x^k with k large enough, in which case from Prop. 4.8*(ii)* and single valuedness of $\mathbf{R}^{f,g}_{h/\gamma}$, we obtain

$$\operatorname{dist}(0, \partial\varphi^{f,g}_{h/\gamma}(x^k)) \le \|Q^f_{h/\gamma}(x^k)\|\,\|\mathbf{R}^{f,g}_{h/\gamma}(x^k)\| \le L_\Omega\|x^k - \bar{x}^k\|, \tag{5.3}$$

where $Q^f_{h/\gamma}$ is as in (4.6) and $L_\Omega := \sup_\Omega \|Q^f_{h/\gamma}\|$ is finite, $\Omega \subset \operatorname{int}\operatorname{dom} h$ being a convex and compact set that contains both sequences $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$ (Thm.s 5.3*(iii)* and 5.3*(iv)*). Now, defining $\Delta_k := \psi(\varphi^{f,g}_{h/\gamma}(x^k) - \varphi_\star) > 0$, concavity of ψ yields

$$\begin{aligned}\Delta_k - \Delta_{k+1} &\ge \psi'(\varphi^{f,g}_{h/\gamma}(x^k) - \varphi_\star)(\varphi^{f,g}_{h/\gamma}(x^k) - \varphi^{f,g}_{h/\gamma}(x^{k+1}))\\ &\ge \frac{\varphi^{f,g}_{h/\gamma}(x^k) - \varphi^{f,g}_{h/\gamma}(x^{k+1})}{\operatorname{dist}(0, \partial\varphi^{f,g}_{h/\gamma}(x^k))} \overset{(5.1)}{\ge} \frac{\gamma\sigma}{L_\Omega}\frac{\mathbf{D}_h(\bar{x}^k, x^k)}{\|x^k - \bar{x}^k\|} \ge \frac{\gamma\sigma\sigma_\Omega}{2L_\Omega}\|x^k - \bar{x}^k\|,\end{aligned} \tag{5.4}$$

where the last inequality follows from the fact that $\mathbf{D}_h(y, x) \ge \frac{\sigma_\Omega}{2}\|y - x\|^2$ for $x, y \in \Omega$, where $\sigma_\Omega > 0$ is a strong convexity modulus of h on Ω. Telescoping the above inequality yields

$$\sum_{k\in\mathbb{N}} \|x^k - \bar{x}^k\| \le \frac{2L_\Omega}{\gamma\sigma\sigma_\Omega}\sum_{k\in\mathbb{N}}(\Delta_k - \Delta_{k+1}) \le \frac{2L_\Omega}{\gamma\sigma\sigma_\Omega}\Delta_0,$$

where the last inequality uses the fact that $\Delta_k \ge 0$. Combined with the fact that $\|x^{k+1} - x^k\| \le (1 - \tau_k)\|x^k - \bar{x}^k\| + \tau_k\|d^k\| \le (1 + c)\|x^k - \bar{x}^k\|$, we conclude that the sequence $(x^k)_{k\in\mathbb{N}}$ has

finite length, and as such it has a limit $x_\star$, this being also the limit of $(\bar{x}^k)_{k\in\mathbb{N}}$ and satisfying $x_\star \in \mathbf{T}^{f,g}_{h/\gamma}(x_\star)$ as it follows from Thm. 5.3*(ii)*. □

More can be said when the KL function is of Łojasiewicz type, in which case asymptotic linear convergence holds, as we state next. Notice that, as remarked in Lemma 4.18, thanks to [71, Thm. 5.2] it suffices to require such KL property on the original cost function φ, as the result then ensures the same will hold for the BFBE.

Theorem 5.6 (linear convergence)**.** *Suppose that the assumptions of Theorem 5.5 are satisfied, and that the KL function can be taken of the form $\psi(s) = \rho s^\vartheta$ for some $\rho > 0$ and $\vartheta \in [1/2, 1]$. Then, $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$ converge at R-linear rate to a γ-critical point.*

Proof. As shown in Theorem 5.5, the sequences converge to a γ-critical point $x_\star$. Since $x^{k+1}-x^k = (1-\tau_k)(\bar{x}^k-x^k)+\tau_k d^k$, defining $B_k := \sum_{i\geq k}\|\bar{x}^i-x^i\|$ one has $\|x^k-x_\star\| \leq (1+c)B_k$, and similarly $\|\bar{x}^k - x_\star\| \leq (3+c)B_k$ owing to the inequality

$$\|\bar{x}^{k+1} - \bar{x}^k\| \leq \|\bar{x}^{k+1} - x^{k+1}\| + \|\bar{x}^k - x^k\| + \|x^{k+1} - x^k\| \leq \|\bar{x}^{k+1} - x^{k+1}\| + (2+c)\|\bar{x}^k - x^k\|.$$

As such, it suffices to show that the sequence $(B_k)_{k\in\mathbb{N}}$ converges with asymptotic Q-linear rate. The KL inequality (4.10) reads

$$\rho\vartheta(\varphi^{f,g}_{h/\gamma}(x^k) - \varphi_\star)^{-(1-\vartheta)} = \psi'(\varphi^{f,g}_{h/\gamma}(x^k) - \varphi_\star) \geq \mathbf{dist}(0, \partial\varphi^{f,g}_{h/\gamma}(x^k))^{-1},$$

which combined with (5.3) yields $\varphi^{f,g}_{h/\gamma}(x^k) - \varphi_\star \leq \left(\frac{\rho\vartheta L_\Omega}{\gamma}\|x^k - \bar{x}^k\|\right)^{\frac{1}{1-\vartheta}}$. Since $\|x^k - \bar{x}^k\| \to 0$, up to discarding the first iterates we may assume that this quantity is smaller than 1. Therefore,

$$\Delta_k := \psi(\varphi^{f,g}_{h/\gamma}(x^k) - \varphi_\star) = \rho(\varphi^{f,g}_{h/\gamma}(x^k) - \varphi_\star)^\vartheta \leq \rho\Big(\tfrac{\rho\vartheta L_\Omega}{\gamma}\|x^k - \bar{x}^k\|\Big)^{\frac{\vartheta}{1-\vartheta}} \leq \tfrac{\rho^2 L_\Omega}{\gamma}\|x^k - \bar{x}^k\|, \quad (5.5)$$

where the last inequality uses the fact that $\vartheta \leq 1$ and that $\frac{\vartheta}{1-\vartheta} \geq 1$. Hence,

$$B_k = \sum_{i\geq k}\|x^i - \bar{x}^i\| \overset{(5.4)}{\leq} \frac{2L_\Omega}{\gamma\sigma\sigma_\Omega}\sum_{i\geq k}(\Delta_i - \Delta_{i+1}) \overset{\Delta_i\geq 0}{\leq} \frac{2L_\Omega}{\gamma\sigma\sigma_\Omega}\Delta_k \overset{(5.5)}{\leq} \frac{2\rho^2 L_\Omega^2}{\gamma^2\sigma\sigma_\Omega}\|x^k - \bar{x}^k\| = C\|x^k - \bar{x}^k\|,$$

where $C := \frac{2\rho^2 L_\Omega^2}{\gamma^2\sigma\sigma_\Omega}$. Therefore, $B_k \leq C\|x^k - \bar{x}^k\| = C(B_k - B_{k+1})$, leading to the sought Q-linear rate $B_{k+1} \leq (1 - 1/C)B_k$. □

5.3. **Superlinear convergence.** Although Bella is "robust" to any choice of directions, a suitable selection stemming for instance from Newton-type method can cause a remarkable speed-up. A first-order optimality condition for $x \in \mathbb{R}^n$, stronger in fact that mere stationarity $0 \in \hat{\partial}\varphi(x)$, is criticality, namely $0 \in \mathbf{R}^{f,g}_{h/\gamma}(x)$. As already discussed in Theorem 4.13, the mapping $\mathbf{R}^{f,g}_{h/\gamma}$ behaves nicely at around its solutions under some regularity assumptions. This motivates the quest to derive the direction d^k in Bella by a Newton-type scheme for this inclusion, *i.e.*,

$$x^+ = x - H(x)\,\mathbf{R}^{f,g}_{h/\gamma}(x), \quad (5.6)$$

where $H(x)$ is an approximation of $J\,\mathbf{R}^{f,g}_{h/\gamma}(x)$. In particular, starting with an invertible matrix H_0, quasi-Newton schemes emulate higher-order information by performing low-rank updates satisfying a so-called secant equation

$$H^+ y = s, \quad \text{where } s = x^+ - x,\ y \in \mathbf{R}^{f,g}_{h/\gamma}(x^+) - \mathbf{R}^{f,g}_{h/\gamma}(x). \quad (5.7)$$

A well-known result characterizing the superlinear convergence of this type of schemes is based on the Dennis-Moré condition [25, 26], which amounts to differentiability of $\mathbf{R}^{f,g}_{h/\gamma}$ at the limit point together with the limit $\|\mathbf{R}^{f,g}_{h/\gamma}(x^k) + J\,\mathbf{R}^{f,g}_{h/\gamma}(x_\star)d^k\|/\|d^k\| \to 0$; see also [27] for the extension to generalized equations. In Theorem 5.9, we will see that directions satisfying this condition do trigger asymptotic superlinear rates in Bella. To this end, we first characterize the quality of the update directions with the next definition, and prove an

intermediate result showing how they fit into BELLA. In the sequel, we will make use of the notion of *nonisolated superlinear directions* that we define next.

Definition 5.7 (nonisolated superlinear directions)**.** *Relative to the iterates generated by* BELLA*, we say that* $(d^k)_{k\in\mathbb{N}}$ *is a sequence of* superlinear directions with order $q \geq 1$ *if*

$$\lim_{k\to\infty} \frac{\mathbf{dist}(x^k + d^k, \mathcal{X}_\star)}{\mathbf{dist}(x^k, \mathcal{X}_\star)^q} = 0,$$

where $\mathcal{X}_\star := \left\{x \in \mathbb{R}^n \mid \mathbf{\Gamma}_h^{f,g}(x) \geq \gamma\right\}$ *is the set of* γ*-critical points.*

Note that Definition 5.7 extends the one given in [28, §7.5] to cases in which $\mathcal{X}_\star$ is not a singleton. The next main result of this section constitutes a key component of the proposed methodology, as it shows that the proposed algorithm does not suffer from the *Maratos' effect* [47], a well-known obstacle for fast local methods that inhibits the acceptance of the unit stepsize. On the contrary, we will show that whenever the directions $(d^k)_{k\in\mathbb{N}}$ in BELLA are superlinear, then indeed unit stepsize is eventually always accepted, and the algorithm converges superlinearly.

Theorem 5.8 (acceptance of the unit stepsize and superlinear convergence)**.** *Consider the iterates generated by* BELLA*, and additionally to Assumption I suppose that the following requirements hold:*

- A1 *the original cost* φ *is level bounded;*
- A2 $f, h \in C^2$ *with* $\nabla^2 h \succ 0$ *on* $\mathbf{int}\,\mathbf{dom}\, h$*;*
- A3 $(x^k)_{k\in\mathbb{N}}$ *converges to a prox-regular (not necessarily isolated) local minimum* $x_\star$ *of* φ *with* $\gamma \neq \mathbf{\Gamma}_h^{f,g}(x_\star)$*;*
- A4 φ *has the KL property at* $x_\star$ *with desingularizing function* $\psi(s) = \varrho s^\vartheta$ *for some* $\varrho > 0$ *and* $\vartheta \in (0, 1)$*;*
- A5 d^k *are superlinear directions with order* $q \geq \mathbf{max}\{1, 1/2\vartheta\}$ *(cf. Def. 5.7).*

Then, there exists $k_0 \in \mathbb{N}$ *such that*

$$\varphi_{h/\gamma}^{f,g}(x^k + d^k) \leq \varphi_{h/\gamma}^{f,g}(x^k) - \sigma\, \mathbf{D}_h(\bar{x}^k, x^k) \quad \forall k \geq k_0.$$

In particular,

(i) eventually stepsize $\tau = 1$ *is always accepted at step 4 (that is, no backtrackings eventually occur) and the iterates reduce to* $x^{k+1} = x^k + d^k$*;*

(ii) $\mathbf{dist}(x^k, \mathcal{X}_\star) \to 0$ *at superlinear rate, where* $\mathcal{X}_\star := \left\{x \in \mathbb{R}^n \mid \mathbf{\Gamma}_h^{f,g}(x) \geq \gamma\right\}$ *is the set of* γ*-critical points as in Definition 5.7.*

Proof. Firstly, Lem. 4.18 ensures that $\bar{\psi}(s) := \bar{\varrho} s^{\bar{\vartheta}}$ with $\bar{\vartheta} = \mathbf{min}\{\vartheta, 1/2\}$ and some $\bar{\varrho} > 0$ is a desingularizing function for $\varphi_{h/\gamma}^{f,g}$ at $x_\star$. Denoting $\varphi_\star := \varphi(x_\star) = \varphi_{h/\gamma}^{f,g}(x_\star)$, the equivalence of local minimality asserted in Thm. 4.5 ensures that for small enough $\varepsilon > 0$ it holds that

$$\varphi_{h/\gamma}^{f,g}(x) \geq \varphi_\star \;\text{ and }\; \mathbf{dist}(x, [\varphi_{h/\gamma}^{f,g} \leq \varphi_\star]) = \mathbf{dist}(x, [\varphi_{h/\gamma}^{f,g} = \varphi_\star]) \;\; \forall x \in \mathbf{B}(x_\star; \varepsilon). \tag{5.8}$$

Moreover, since $\varphi_\star < \varphi_{h/\gamma}^{f,g}(x^k) \searrow \varphi_\star$, up to possibly discarding the first iterates and restricting ε we may assume that $\varphi_{h/\gamma}^{f,g}(x^k) \leq \varphi_\star + \eta$, with $\eta, \varepsilon > 0$ as in Def. 4.15 of the KL function. Notice further that any point $x \in \mathbf{B}(x_\star; \varepsilon)$ and such that $\varphi_{h/\gamma}^{f,g}(x) = \varphi_\star$ necessarily is γ-critical owing to Thm. 4.2*(iii)* and local minimality of $x_\star$. Conversely, it follows from Thm. 4.11*(ii)* that $\varphi_{h/\gamma}^{f,g}$ is (Lipschitz-continuously) differentiable around $x_\star$ with $\nabla\varphi_{h/\gamma}^{f,g} = Q_{h/\gamma}^f \mathbf{R}_{h/\gamma}^{f,g}$, where $Q_{h/\gamma}^f > 0$ is as in (4.6), and in particular $\nabla\varphi_{h/\gamma}^{f,g}(x) = 0$ for any γ-critical point x close to $x_\star$. Combined with the KL inequality (4.10), we conclude that close to $x_\star$ a point x is γ-critical iff $\varphi(x) = \varphi_{h/\gamma}^{f,g}(x) = \varphi_\star$. Up to possibly further restricting ε, we may thus modify (5.8) to

$$\mathbf{dist}(x, [\varphi_{h/\gamma}^{f,g} = \varphi_\star]) = \mathbf{dist}(x, \mathcal{X}_\star) \quad \forall x \in \overline{\mathbf{B}}(x_\star; \varepsilon). \tag{5.9}$$

Combined with the error bound in Lem. 4.17, we obtain

$$\varphi^{f,g}_{h/\gamma}(x^k) - \varphi_\star \geq \mathbf{dist}(x^k, \mathcal{X}_\star)^{\mathbf{max}\{1,1/2\vartheta\}} \quad \forall k \in \mathbb{N}. \tag{5.10}$$

Since, as discussed above, $\mathcal{X}_\star$ coincides with a (closed) sublevel set of $\varphi^{f,g}_{h/\gamma}$ close to $x_\star$, for every k there exists a projection point $x^k_\star \in \mathbf{Proj}_{\mathcal{X}_\star}(x^k + d^k)$, which will thus be γ-critical and such that $\varphi(x^k_\star) = \varphi_\star$. In particular,

$$\varphi^{f,g}_{h/\gamma}(x^k + d^k) \leq \varphi_\star + \tfrac{1-\gamma\sigma_f}{\gamma} \mathbf{D}_h(x^\star_k, x^k + d^k) \leq \varphi_\star + \tfrac{\tilde{L}_h}{2}\tfrac{1-\gamma\sigma_f}{\gamma} \mathbf{dist}(x^k + d^k, \mathcal{X}_\star)^2,$$

where the first inequality follows from Lem. 4.19 and $\tilde{L}_h$ in the second one is a Lipschitz modulus of ∇h on $\overline{\mathbf{B}}(x_\star; \varepsilon)$ (since $h \in C^2$). By combining this with (5.10), we get

$$\varepsilon_k := \frac{\varphi^{f,g}_{h/\gamma}(x^k + d^k) - \varphi_\star}{\varphi^{f,g}_{h/\gamma}(x^k) - \varphi_\star} \leq \frac{\tilde{L}_h}{2}\frac{1-\gamma\sigma_f}{\gamma}\frac{\mathbf{dist}(x^k + d^k, \mathcal{X}_\star)^2}{\mathbf{dist}(x^k, \mathcal{X}_\star)^{\mathbf{max}\{1,1/2\vartheta\}}} \to 0 \quad \text{as } k \to \infty. \tag{5.11}$$

Thus, for large enough k so that $\varepsilon_k \leq 1$, we have

$$\begin{aligned}
\varphi^{f,g}_{h/\gamma}(x^k + d^k) - \varphi^{f,g}_{h/\gamma}(x^k) &= (\varphi^{f,g}_{h/\gamma}(x^k + d^k) - \varphi_\star) - (\varphi^{f,g}_{h/\gamma}(x^k) - \varphi_\star) \\
&= (\varepsilon_k - 1)(\varphi^{f,g}_{h/\gamma}(x^k) - \varphi_\star) \\
\text{(since } \varepsilon_k \leq 1,\ \varphi(\bar{x}^k) \geq \varphi_\star) \quad &\leq (\varepsilon_k - 1)(\varphi^{f,g}_{h/\gamma}(x^k) - \varphi(\bar{x}^k)) \\
\text{(use Thm. 4.2}\textit{(ii)}) \quad &\leq (\varepsilon_k - 1)\tfrac{1-\gamma L_f}{\gamma} \mathbf{D}_h(\bar{x}^k, x^k) \leq \sigma\, \mathbf{D}_h(\bar{x}^k, x^k),
\end{aligned}$$

for large enough k, where the last inequality holds since $\sigma < \frac{1-\gamma L_f}{\gamma}$ and $\varepsilon_k \to 0$. □

Despite the importance of nonisolated critical points in nonsmooth nonconvex optimization, there has been little about superlinear directions for such problems. In the convex setting, some studies have shown the potential of variants of regularized Newton [40, 64] and semismooth Newton methods [41, 65] under a local error bound. In the smooth nonconvex setting, there are many works relying on Levenberg-Marquardt [3, 4, 29, 70], cubic regularization [72], and regularized Newton [67] methods under variants of local error bounds and Hölder metric subregularity.

Quasi-Newton methods constitute an important class of directions widely used in optimization. Superlinear convergence of these type of directions is typically assessed by means of the Dennis-Moré condition. We next show that under regularity assumptions at the limit point the same condition ensures acceptance of unit stepsize in our framework, albeit provided the algorithm converges to an (isolated) strong local minimum.

Theorem 5.9 (superlinear convergence under Dennis-Moré condition)**.** *Consider the iterates generated by* BELLA*. Additionally to Assumption I, suppose that the following requirements are satisfied:*

A1 $(x^k)_{k\in\mathbb{N}}$ *converges to a strong local minimum* $x_\star$ *of* φ*;*

A2 $f, h \in C^2$ *with* $\nabla^2 h \succ 0$ *on* $\mathbf{int}\,\mathbf{dom}\, h$*;*

A3 $\mathbf{R}^{f,g}_{h/\gamma}$ *is strictly differentiable at* $x_\star$ *(see Thm. 4.13 for sufficient conditions);*

A4 $\gamma < \mathbf{\Gamma}^{f,g}_h(x_\star)$*;*

A5 $(d^k)_{k\in\mathbb{N}}$ *satisfy the Dennis-Moré condition*

$$\lim_{k\to\infty} \frac{\mathbf{R}^{f,g}_{h/\gamma}(x^k) + J\,\mathbf{R}^{f,g}_{h/\gamma}(x_\star)d^k}{\|d^k\|} = 0. \tag{5.12}$$

Then, $(d^k)_{k\in\mathbb{N}}$ *are superlinear directions with respect to* $(x^k)_{k\in\mathbb{N}}$*, and in particular all the claims of Theorem 5.8 hold.*

Proof. The Dennis-Moré condition (5.12) and strict differentiability at $x_\star$ imply that

$$\lim_{k\to\infty} \frac{\mathbf{R}^{f,g}_{h/\gamma}(x^k + d^k)}{\|d^k\|} = \lim_{k\to\infty} \left[\frac{\mathbf{R}^{f,g}_{h/\gamma}(x^k) + J\,\mathbf{R}^{f,g}_{h/\gamma}(x_\star)d^k - \mathbf{R}^{f,g}_{h/\gamma}(x^k + d^k)}{\|d^k\|} + \frac{\mathbf{R}^{f,g}_{h/\gamma}(x^k + d^k)}{\|d^k\|} \right] = 0.$$

Further, nonsingularity of $\mathbf{R}^{f,g}_{h/\gamma}(x_\star)$ implies that there exists $\alpha > 0$ such that

$$\|\,\mathbf{R}^{f,g}_{h/\gamma}(x)\| = \|\,\mathbf{R}^{f,g}_{h/\gamma}(x) - \mathbf{R}^{f,g}_{h/\gamma}(x_\star)\| \geq \alpha\|x - x_\star\|$$

holds for all x close enough to $x_\star$. Here, the first equality is due to the fact that $x_\star$ is critical, hence $0 = \mathbf{R}^{f,g}_{h/\gamma}(x_\star)$ (equality, as opposed to inclusion, holds due to the assumption of differentiability). We thus have

$$0 \leftarrow \frac{\|\,\mathbf{R}^{f,g}_{h/\gamma}(x^k + d^k)\|}{\|d^k\|} \geq \alpha \frac{\|x^k + d^k - x_\star\|}{\|d^k\|} \geq \alpha \frac{\|x^k + d^k - x_\star\|}{\|x^k + d^k - x_\star\| + \|x^k - x_\star\|} = \alpha \frac{\frac{\|x^k + d^k - x_\star\|}{\|x^k - x_\star\|}}{1 + \frac{\|x^k + d^k - x_\star\|}{\|x^k - x_\star\|}},$$

as $k \to \infty$, and in particular $\frac{\|x^k + d^k - x_\star\|}{\|x^k - x_\star\|} \to 0$, as claimed. □

6. Final remarks

We proposed Bella, a Bregman-forward-backward-splitting-based algorithm for minimizing the sum of two nonconvex functions, where the first one is relatively smooth and the second one is possibly nonsmooth. Bella is a linesearch algorithm on the Bregman forward-backward envelope (BFBE), a Bregman extension of the forward-backward envelope, and globalizes convergence of fast local methods for finding zeros of the forward-backward residual. Furthermore, thanks to a nonlinear local error bound holding for the BFBE under prox-regularity and the KL property, the algorithm enables acceptance of unit stepsize when the fast local method yields directions that are superlinear with respect to the set of solutions, thus triggering superlinear convergence even when the limit point is not isolated.

In future work we plan to address the following issues: (1) extending existing superlinear direction schemes such as those proposed in [40, 67, 3, 64] for either convex or smooth problems to the more general setting of this paper; (2) assessing the performance of such schemes in the Bella framework with numerical simulations on nonconvex nonsmooth problems such as low-rank matrix completion, sparse nonnegative matrix factorization, phase retrieval, and deep learning; and (3) guaranteeing saddle point avoidance, in the spirit of [51, 39, 43].

References

[1] Masoud Ahookhosh. Accelerated first-order methods for large-scale convex minimization. *arXiv preprint arXiv:1604.08846*, 2016.

[2] Masoud Ahookhosh. Optimal subgradient methods: computational properties for large-scale linear inverse problems. *Optimization and Engineering*, 19(4):815–844, 2018.

[3] Masoud Ahookhosh, Francisco J. Aragón, Ronan M.T. Fleming, and Phan T. Vuong. Local convergence of Levenberg-Marquardt methods under Hölder metric subregularity. *arXiv:1703.07461*, 2017. (To appear in Adv Comp Math).

[4] Masoud Ahookhosh, Ronan M.T. Fleming, and Phan T. Vuong. Finding zeros of Hölder metrically subregular mappings via globally convergent Levenberg-Marquardt methods. *arXiv preprint arXiv:1812.00818*, 2018.

[5] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

[6] Dominique Azé and Jean-Noël Corvellec. Nonlinear error bounds via a change of function. *Journal of Optimization Theory and Applications*, 172(1):9–32, 2017.

[7] Heinz Bauschke, Patrick Combettes, and Dominikus Noll. Joint minimization with alternating Bregman proximity operators. *Pacific Journal of Optimization*, 2005.

[8] Heinz H. Bauschke, Jérôme Bolte, Jiawei Chen, Marc Teboulle, and Xianfu Wang. On linear convergence of non-Euclidean gradient methods without strong convexity and Lipschitz gradient continuity. *Journal of Optimization Theory and Applications*, pages 1–20, 2019.

[9] Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2016.

[10] Heinz H. Bauschke, Jonathan M. Borwein, and Patrick L. Combettes. Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces. *Communications in Contemporary Mathematics*, 03(04):615–647, 2001.

[11] Heinz H. Bauschke, Minh N. Dao, and Scott B. Lindstrom. Regularizing with Bregman-Moreau envelopes. *SIAM Journal on Optimization*, 28(4):3208–3228, 2018.

[12] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[13] Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

[14] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[15] Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007.

[16] Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.

[17] Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, Oct 2017.

[18] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1–2):459–494, 2014.

[19] Jérôme Bolte, Shoham Sabach, Marc Teboulle, and Yakov Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018.

[20] Radu Ioan Boţ and Ernö Robert Csetnek. An inertial Tseng's type proximal algorithm for nonsmooth and nonconvex optimization problems. *Journal of Optimization Theory and Applications*, 171(2):600–616, 2016.

[21] Radu Ioan Boţ, Ernö Robert Csetnek, and Szilárd Csaba László. An inertial forward-backward algorithm for the minimization of the sum of two nonconvex functions. *EURO Journal on Computational Optimization*, 4(1):3–25, 2016.

[22] Radu Ioan Bot, Ernoö Robert Csetnek, and Dang-Khoa Nguyen. A proximal minimization algorithm for structured nonconvex and nonsmooth problems. *SIAM Journal on Optimization*, 29(2):1300–1328, 2019.

[23] Lev 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.

[24] Aris Daniilidis and Jeffrey C. H. Pang. Continuity and differentiability of set-valued maps revisited in the light of tame geometry. *Journal of the London Mathematical Society*, 83(3):637–658, 2011.

[25] John E. Jr. Dennis and Jorge J. Moré. A characterization of superlinear convergence and its application to quasi-Newton methods. *Mathematics of Computation*, 28(126):549–560, 1974.

[26] John E. Jr. Dennis and Jorge J. Moré. Quasi-Newton methods, motivation and theory. *SIAM Review*, 19(1):46–89, 1977.

[27] Asen Dontchev. Generalizations of the Dennis-Moré theorem. *SIAM Journal on Optimization*, 22(3):821–830, 2012.

[28] Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*, volume II. Springer, 2003.

[29] Andreas Fischer. Local behavior of an iterative framework for generalized equations with nonisolated solutions. *Mathematical Programming*, 94(1):91–124, 2002.

[30] Masao Fukushima and Hisashi Mine. A generalized proximal point algorithm for certain non-convex minimization problems. *International Journal of Systems Science*, 12(8):989–1000, 1981.

[31] Masao Fukushima and Liqun Qi. A globally and superlinearly convergent algorithm for nonsmooth convex minimization. *SIAM Journal on Optimization*, 6(4):1106–1120, 1996.

[32] Filip Hanzely and Peter Richtárik. Fastest rates for stochastic mirror descent methods. *arXiv preprint arXiv:1803.07374*, 2018.

[33] Filip Hanzely, Peter Richtarik, and Lin Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. *arXiv preprint arXiv:1808.03045*, 2018.

[34] Alexander D. Ioffe. Variational analysis of regular mappings. *Springer Monographs in Mathematics. Springer, Cham*, 2017.
[35] Alexey F. Izmailov and Mikhail V. Solodov. *Newton-type Methods for Optimization and Variational Problems*. Springer, 2014.
[36] Chao Kan and Wen Song. The Moreau envelope function and proximal mapping in the sense of the Bregman distance. *Nonlinear Analysis: Theory, Methods & Applications*, 75(3):1385 – 1399, 2012.
[37] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in neural information processing systems*, pages 586–594, 2016.
[38] Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'institut Fourier*, 48(3):769–783, 1998.
[39] Jason D. Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, Feb 2019.
[40] Dong-Hui Li, Masao Fukushima, Liqun Qi, and Nobuo Yamashita. Regularized Newton methods for convex minimization problems with singular solutions. *Computational Optimization and Applications*, 28(2):131–147, 2004.
[41] Xudong Li, Defeng Sun, and Kim-Chuan Toh. A highly efficient semismooth Newton augmented Lagrangian method for solving lasso problems. *SIAM Journal on Optimization*, 28(1):433–458, 2018.
[42] Tianxiang Liu and Ting Kei Pong. Further properties of the forward-backward envelope with applications to difference-of-convex programming. *Computational Optimization and Applications*, 67(3):489–520, Jul 2017.
[43] Yanli Liu and Wotao Yin. An envelope for Davis–Yin splitting and strict saddle-point avoidance. *Journal of Optimization Theory and Applications*, 181(2):567–587, May 2019.
[44] Stanislaw Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, pages 87–89, 1963.
[45] Stanislaw Łojasiewicz. Sur la géométrie semi- et sous- analytique. *Annales de l'institut Fourier*, 43(5):1575–1595, 1993.
[46] Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.
[47] Nicholas Maratos. *Exact penalty function algorithms for finite dimensional and control optimization problems*. PhD thesis, Imperial College London (University of London), 1978.
[48] Yurii Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, Aug 2013.
[49] Yurii Nesterov. Implementable tensor methods in unconstrained convex optimization. Technical report, UC Louvain, Center for Operations Research and Econometrics (CORE), Belgium, 2018.
[50] Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer Science & Business Media, 2006.
[51] Michael O'Neill and Stephen J. Wright. Behavior of accelerated gradient methods near critical points of nonconvex functions. *Mathematical Programming*, Oct 2018.
[52] Panagiotis Patrinos and Alberto Bemporad. Proximal Newton methods for convex composite optimization. In *52nd IEEE Conference on Decision and Control*, pages 2358–2363, 2013.
[53] René A. Poliquin and R. Tyrrell Rockafellar. Prox-regular functions in variational analysis. *Transactions of the American Mathematical Society*, 348(5):1805–1838, 1996.
[54] R. Tyrrell Rockafellar. Higher derivatives of conjugate convex functions. *Journal of Applied Analysis*, 1(1):41–43, 1977.
[55] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2011.
[56] Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton university press, 1970.
[57] Lorenzo Stella, Andreas Themelis, and Panagiotis Patrinos. Forward-backward quasi-Newton methods for nonsmooth optimization problems. *Computational Optimization and Applications*, 67(3):443–487, Jul 2017.
[58] Lorenzo Stella, Andreas Themelis, Pantelis Sopasakis, and Panagiotis Patrinos. A simple and efficient algorithm for nonlinear model predictive control. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 1939–1944, Dec 2017.
[59] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018.
[60] Jared Tanner and Ke Wei. Low rank matrix completion by alternating steepest descent methods. *Applied and Computational Harmonic Analysis*, 40(2):417–429, 2016.
[61] Marc Teboulle. Entropic proximal mappings with applications to nonlinear programming. *Mathematics of Operations Research*, 17(3):670–690, 1992.
[62] Marc Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, pages 1–30, 2018.
[63] Andreas Themelis. *Proximal Algorithms for Structured Nonconvex Optimization*. PhD thesis, KU Leuven / IMT Lucca, 2018.

[64] Andreas Themelis, Masoud Ahookhosh, and Panagiotis Patrinos. On the acceleration of forward-backward splitting via an inexact Newton method. In R. Luke, H. Bauschke, and R. Burachik, editors, *Splitting Algorithms, Modern Operator Theory, and Applications*. Springer, 2019. (To appear).

[65] Andreas Themelis and Panagiotis Patrinos. SuperMann: a superlinearly convergent algorithm for finding fixed points of nonexpansive operators. *IEEE Transactions on Automatic Control*, dec 2019.

[66] Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. *SIAM Journal on Optimization*, 28(3):2274–2303, 2018.

[67] Kenji Ueda and Nobuo Yamashita. A regularized Newton method without line search for unconstrained optimization. *Computational Optimization and Applications*, 59(1-2):321–351, 2014.

[68] Lou Van den Dries. *Tame topology and o-minimal structures*, volume 248. Cambridge university press, 1998.

[69] Lou Van den Dries and Chris Miller. Geometric categories and o-minimal structures. *Duke Math. J*, 84(2):497–540, 1996.

[70] Nobuo Yamashita and Masao Fukushima. On the rate of convergence of the Levenberg-Marquardt method. In *Topics in numerical analysis*, pages 239–249. Springer, 2001.

[71] Peiran Yu, Guoyin Li, and Ting Kei Pong. Deducing Kurdyka-Łojasiewicz exponent via inf-projection. *arXiv preprint arXiv:1902.03635*, 2019.

[72] Man-Chung Yue, Zirui Zhou, and Anthony Man-Cho So. On the quadratic convergence of the cubic regularization method under a local error bound condition. *SIAM Journal on Optimization*, 29(1):904–932, 2019.