A Bregman Forward-Backward Linesearch Algorithm for Nonconvex Composite Optimization: Superlinear Convergence to Nonisolated Local Minima

Ahookhosh, Masoud
Department of Mathematics and Computer Science, University of Antwerp

Themelis, Andreas
Faculty of Information Science and Electrical Engineering (ISEE), Kyushu University

Patrinos, Panagiotis
Department of Electrical Engineering (ESAT-STADIUS) - KU Leuven

出版情報：SIAM Journal on Optimization. 31 (1), pp.653-685, 2021-02-22. SIAM
バージョン：

SIAM J. OPTIM.
Vol. 31, No. 1, pp. 653–685

A BREGMAN FORWARD-BACKWARD LINESEARCH ALGORITHM FOR NONCONVEX COMPOSITE OPTIMIZATION: SUPERLINEAR CONVERGENCE TO NONISOLATED LOCAL MINIMA*

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 satisfies a relative smoothness condition and the other one is 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, which enjoys a nonlinear error bound when the objective function satisfies a Łojasiewicz-type property. The proposed algorithm is of linesearch type over the BFBE along user-defined update directions and converges subsequentially to stationary points and globally under the Kurdyka–Łojasiewicz condition. Moreover, when the update directions are superlinear in the sense of Facchinei and Pang [Finite-Dimensional Variational Inequalities and Complementarity Problems, Volume I, Springer, New York, 2003], owing to the given nonlinear error bound unit stepsize is eventually always accepted and the algorithm attains superlinear convergence rates even when the limit point is a nonisolated minimum.

Key words. nonsmooth nonconvex optimization, Bregman–Moreau and Bregman forward-backward envelopes, relative smoothness, KL inequality, nonlinear error bound, nonisolated local minima, superlinear convergence

AMS subject classifications. 90C06, 90C25, 90C26, 49J52, 49J53

DOI. 10.1137/19M1264783

1. Introduction.

In this paper, we address the composite minimization problem

$$\textbf{minimize}\, \varphi(x) \equiv f(x) + g(x) \quad \textbf{subject to}\, x \in \overline{C}. \tag{1.1}$$

Here, C is an open convex set ($\overline{C}$ denotes the closure of C), g is proper and lower semicontinuous (lsc), and f is relatively smooth with respect to a Legendre kernel h (see subsection 2.2) with $\mathbf{dom}\, \nabla h = C$ (for detailed assumptions we refer to section 4). Despite its simple structure, (1.1) encompasses a variety of optimization problems frequently encountered in scientific areas such as signal and image processing, machine learning, and inverse problems [11, 24, 40, 55, 60]. The notion of *Lipschitz-like convexity* was recently discovered in the seminal work [11] as a generalization of the Lipschitz smoothness condition and was later renamed *relative smoothness* in [55]. Studying optimization problems involving relatively smooth functions has received much attention in the last few years [11, 24, 34, 39, 40, 55, 63, 79]. In our setting (1.1), since f is relatively smooth and g is nonsmooth nonconvex, we can cover a

*Received by the editors May 28, 2019; accepted for publication (in revised form) November 23, 2020; published electronically February 22, 2021.
https://doi.org/10.1137/19M1264783
Funding: This work was supported by Research Foundation Flanders (FWO) research projects G086518N, G086318N, and G0A0920N; Research Council KU Leuven C1 project C14/18/068; Fonds de la Recherche Scientifique–FNRS; and the Fonds Wetenschappelijk Onderzoek–Vlaanderen under EOS project 30468160 (SeLMA).

†Department of Mathematics and Computer Science, University of Antwerp, Middelheimlaan 1, B-2020 Antwerp, Belgium (masoud.ahookhosh@uantwerp.be).
‡Faculty of Information Science and Electrical Engineering (ISEE), Kyushu University, 819-0395 Fukuoka, Japan (andreas.themelis@ees.kyushu-u.ac.jp).
§Department of Electrical Engineering (ESAT-STADIUS), KU Leuven, 3001 Leuven, Belgium (panos.patrinos@esat.kuleuven.be).

wide spectrum of applications. In the Euclidean setting, there are plenty of optimization algorithms that can handle composite minimization of the form (1.1), such as [2, 1, 16, 21, 62, 81] for convex problems and [7, 23, 27, 26, 28, 37, 74, 83] for 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., [64] for many algorithms attaining a superlinear convergence rate for smooth problems and [37, 38, 83] for other related works in the nonconvex nonsmooth setting. In most of these attempts, the key element is the so-called Dennis–Moré condition [31, 32] 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 [77], low-rank matrix recovery [17], phase retrieval [76], and deep learning [43]. Up to now, besides some attempts for minimizing smooth nonlinear least-squares problems (see, e.g., [3, 4, 41] and references therein) far too little attention has been paid to the superlinear convergence to nonisolated critical points for nonconvex nonsmooth problems.

1.1. Related work. In order to guarantee convergence, most first-order methods for problem (1.1) in the nonconvex setting require Lipschitz differentiability of f; however, there are plenty of examples that fail to comply with this assumption (see, e.g., [6, 5, 11, 34, 55, 60]). Recently, from the seminal work conducted in [11] it emerged that the Lipschitz smoothness assumption of f can be relaxed by introducing the notion of *relative smoothness* (see Definition 2.4), which was further developed in [55, 88]. Assuming the convexity of f and g and the relative smoothness of f, a Bregman proximal-gradient method was proposed in [11], while primal and dual algorithms were developed in [55]. More recently, in the convex setting, [63] proposed an accelerated tensor method, and [39, 40] suggested a Nesterov-type accelerated method and a stochastic mirror descent method.

The developments of relative smoothness also led to a renewed interest in theory and algorithms of nonconvex optimization. Recently, [24] extended the results of [11] for the Bregman proximal-gradient method, and [79] discussed several first-order algorithms. More recently, the linear convergence of the gradient method for relatively smooth functions was studied in [10]. In [66], a generic Bregman linesearch was proposed for not necessarily Lipschitz smooth problems, which covers the Bregman forward-backward splitting as a special case for nonconvex smooth f and convex g. In [61], some Bregman proximal-gradient algorithms with inertial effects were presented in the nonconvex setting for smooth f and hypoconvex g. Furthermore, a stochastic convex model-based minimization algorithm was proposed in [30] for hypoconvex functions under relative smoothness and high-order growth conditions. The notion of relative smoothness was further extended to its block version [6, 5], opening the possibility of alternating minimization algorithms in the fully nonconvex setting. To the best of our knowledge, apart from the latter papers, there have not been many attempts to deal with (1.1) in the relatively smooth and fully nonconvex setting.

1.2. Contribution. Our goal is to design an efficient framework for addressing the structured nonconvex problem (P) with superlinear guarantees of convergence, even when the limit point is a nonisolated local minimum. We aim at devising a linesearch strategy that globalizes the convergence of fast local methods, stemming, for instance, from Newton-type schemes. The lack of differentiability in problem (1.1)

makes classical smooth optimization methodologies, such as Armijo backtracking, not applicable. Nevertheless, favorable properties of the Bregman forward-backward envelope (BFBE) introduced here lead to the *Bregman envelope linesearch algorithm* (BELLA), which overcomes said limitations exclusively by means of Bregman forward-backward operations. Our contribution can be summarized as follows:

(i) *Bregman–Moreau envelope analysis.* We provide new insights on the Bregman–Moreau envelope complementing the ones in [42, 15, 45]. Among these, we highlight properties of fixed points, a local equivalence with the forward-backward envelope (FBE) [68], and ultimately a second-order differentiability result with classical generalized differentiability tools [69, 73].

(ii) *Bregman forward-backward splitting and linesearch extension.* We highlight a connection between Bregman forward-backward mapping and Bregman proximal mapping (Theorem 4.1), revealing that the proximal point algorithm is as general as the forward-backward splitting in the Bregman setting. Correspondingly, we introduce a BFBE function which is at the core of BELLA, a linesearch algorithm that can globalize the convergence of fast local methods for problem (1.1).

(iii) *Superlinear convergence to nonisolated critical points.* Differentiability properties of the BFBE (Theorem 4.7) allow us to link the Kurdyka–Łojasiewicz (KL) property to a nonlinear error bound involving the distance to sublevel sets (Lemma 5.11). This observation is the key for showing that BELLA can converge superlinearly to local minima even if nonisolated (Theorem 5.12), when "fast" directions (in a sense that will be made precise in Definition 5.9) are selected. More generally, global and linear convergence are shown utilizing the KL inequality (Theorems 5.7 and 5.8).

1.3. Paper organization. Section 2 introduces the notation used and some known facts. In section 3 we investigate properties of the Bregman proximal mapping and the Bregman–Moreau envelope, which we then use in section 4 to derive similar results for the Bregman forward-backward mapping and the corresponding *envelope* function, the BFBE, which is a key tool of our analysis. In section 5 we introduce BELLA, a linesearch algorithm on the BFBE, and show its convergence properties. Section 6 concludes the paper.

2. Preliminaries.

2.1. Notation. The extended-real line is denoted by $\overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{\infty\}$. The open and closed balls of radius $r > 0$ centered at $x \in \mathbb{R}^n$ are denoted as $\mathbf{B}(x; r)$ and $\overline{\mathbf{B}}(x; r)$, respectively. We say that $(x^k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ converges at R-linear rate (to a point $x_\star$) if there exist $c > 0$ and $\rho \in (0, 1)$ such that $\|x^k - x_\star\| \leq c\rho^k$ holds for every k. The distance of a point $x \in \mathbb{R}^n$ to a nonempty set $S \subseteq \mathbb{R}^n$ is given by $\mathbf{dist}(x, S) = \mathbf{inf}_{z\in S} \|z - x\|$. The interior, closure, and boundary of S are respectively denoted as $\mathbf{int}\, S$, $\overline{S}$, and $\mathbf{bdry}\, S = \overline{S} \setminus \mathbf{int}\, S$. The *indicator function* of S is $\delta_S : \mathbb{R}^n \to \overline{\mathbb{R}}$ defined as $\delta_S(x) = 0$ if $x \in S$ and ∞ otherwise.

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 $\mathbf{dom}\, f \coloneqq \{x \in \mathbb{R}^n \mid f(x) < \infty\}$. For $\alpha \in \mathbb{R}$, $[f \leq \alpha] \coloneqq \{x \in \mathbb{R}^n \mid f(x) \leq \alpha\}$ is the *α-(sub)level set* of f; the *α-level set* $[f = \alpha]$ is defined accordingly. We say that f is *level bounded* if $[f \leq \alpha]$ is bounded for all $\alpha \in \mathbb{R}$. A point $x_\star \in \mathbf{dom}\, f$ is a *local minimum* for f if $f(x) \geq f(x_\star)$ holds for all x in a neighborhood of $x_\star$. If the inequality can be strengthened to $f(x) \geq f(x_\star) + \frac{\mu}{2}\|x - x_\star\|^2$ for some $\mu > 0$, then $x_\star$ is a *strong local minimum*. The *convex conjugate* of f is denoted as $f^* \coloneqq \mathbf{sup}_z \{\langle \cdot\, , z\rangle - f(z)\}$.

Given $x \in \mathbf{dom}\, f$, a vector $v \in \partial f(x)$ is a *subgradient* of f at x, where $\partial f(x)$ is the *(limiting) subdifferential*

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

and $\hat{\partial} f(x)$ is the set of *regular subgradients* of f at x, namely vectors $v \in \mathbb{R}^n$ such that $\mathbf{liminf}_{\substack{z\to x \\ z\neq x}} \frac{f(z)-f(x)-\langle v, z-x\rangle}{\|z-x\|} \geq 0$. Following the terminology of [73], we say that $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *strictly continuous at* $\bar{x}$ if $\mathbf{lip}\, f(\bar{x}) \coloneqq \mathbf{limsup}_{\substack{y,z\to\bar{x} \\ y\neq z}} \frac{|f(y)-f(z)|}{\|y-z\|} < \infty$ and *strictly differentiable at* $\bar{x}$ if $\nabla f(\bar{x})$ exists and satisfies $\mathbf{lim}_{\substack{y,z\to\bar{x} \\ y\neq z}} \frac{f(y)-f(z)-\langle \nabla f(\bar{x}), y-z\rangle}{\|y-z\|} = 0$. If f is everywhere strictly continuous on an open set $\mathcal{U}$, then its gradient exists almost everywhere on $\mathcal{U}$, and as such its *Bouligand subdifferential*

$$\partial_B f(x) \coloneqq \big\{ v \mid \exists x^k \to x \text{ with } \nabla f(x^k) \to v \big\}$$

is nonempty and compact for all $x \in \mathcal{U}$ [73, Thm. 9.61]. $\mathcal{C}^k(\mathcal{U})$ is the set of functions $\mathcal{U} \to \mathbb{R}$ which are k times continuously differentiable. We write $\mathcal{C}^k$ if $\mathcal{U}$ is clear from context. For a function $F : \mathbb{R}^n \to \mathbb{R}^m$, we denote by $JF : \mathbb{R}^n \to \mathbb{R}^{m\times n}$ its Jacobian, defined whenever it makes sense. The notation $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ indicates a set-valued mapping, whose domain and range are respectively defined as $\mathbf{dom}\, T = \{x \in \mathbb{R}^n \mid T(x) \neq \emptyset\}$ and $\mathbf{range}\, T = \bigcup_{x\in\mathbb{R}^n} T(x)$.

2.2. Relative smoothness. Here, after giving some definitions, we provide a list of results regarding relative smoothness that will be useful in what follows.

DEFINITION 2.1 (Bregman distance [29]). *Relative to a convex function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ that is continuously differentiable on $\mathbf{int}\,\mathbf{dom}\, h \neq \emptyset$, the* Bregman distance $\mathbf{D}_h : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ *is*

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

Function h will be referred to as a distance-generating function.

Throughout the paper we will consider distance-generating functions that are Legendre kernels, as defined next. We refer the reader to [15] for a list of popular such functions.

DEFINITION 2.2. *A proper, lsc, and strictly convex function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ with $\mathbf{int}\,\mathbf{dom}\, h \neq \emptyset$ and such that $h \in \mathcal{C}^1(\mathbf{int}\,\mathbf{dom}\, h)$ is said to be a Legendre kernel if it is* (i) 1-coercive*, i.e., such that $\mathbf{lim}_{\|x\|\to\infty} h(x)/\|x\| = \infty$, and* (ii) essentially smooth*, i.e., 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 $\mathbf{dom}\, h$.*

FACT 2.3. *The following assertions hold for a Legendre kernel $h : \mathbb{R}^n \to \overline{\mathbb{R}}$:*

(i) *$h^* \in \mathcal{C}^1(\mathbb{R}^n)$ is strictly convex and $\nabla h^{-1} = \nabla h^*$* [72, Thm. 26.5 and Cor. 13.3.1]*;*

(ii) *$\mathbf{D}_h(\cdot, x)$ and $\mathbf{D}_h(x, \cdot)$ are level bounded locally uniformly in x on $\mathbf{int}\,\mathbf{dom}\, h \times \mathbf{int}\,\mathbf{dom}\, h$* [12, Lem. 7.3(v)–(viii)].[1]

Moreover, for any open convex set $\mathcal{U} \subseteq \mathbf{int}\,\mathbf{dom}\, h$ the following hold:

[1] Although [12] only states level boundedness, a trivial modification of the proof shows local uniformity too.

(iii) *If h is $\tilde{\sigma}_h$-strongly convex on $\mathcal{U}$, then $\mathbf{D}_h(y,x) \geq \frac{\tilde{\sigma}_h}{2}\|y-x\|^2$ for all $x, y \in \mathcal{U}$.*

(iv) *If ∇h is $\tilde{L}_h$-Lipschitz on $\mathcal{U}$, 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 convex neighborhood in which such property holds.

DEFINITION 2.4 (relative smoothness [11]). *We say that a proper, lsc function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is* smooth relative to *a Legendre kernel $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ if $\mathbf{dom}\, f \supseteq \mathbf{dom}\, h$, and there exists $L_f \geq 0$ such that $L_f h \pm f$ are convex functions on $\mathbf{int}\,\mathbf{dom}\, h$. We will simply say that f is relatively smooth when h is clear from context, or L_f-relatively smooth to make the modulus L_f explicit.*

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

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

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

Proof.

Proposition 2.5(i). Convexity of $L_f h \pm f$ and continuous differentiability of h on $\mathbf{int}\,\mathbf{dom}\, h$ ensure through [73, 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 [73, Def. 7.25], with $\hat{\partial} f$ and $\hat{\partial}(-f)$ both nonempty. The proof now follows from [73, Thm. 9.18(d)].

Proposition 2.5(ii). Let $\tilde{L}_h$ be a Lipschitz modulus for ∇h on $\mathcal{U}$. Convexity of $L_f h + f$ yields

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

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 inequalities together prove that ∇f is $\tilde{L}_f$-Lipschitz on $\mathcal{U}$ with $\tilde{L}_f = L_f \tilde{L}_h$. ∎

The proof of the following result is a simple adaptation of that of [55, Prop. 1.1].

PROPOSITION 2.6 (characterization of relative smoothness). *Let a proper lsc function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a Legendre kernel $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be fixed. The following are equivalent:*

(a) *f is L_f-smooth relative to h;*

(b) *$\big|f(y) - f(x) - \langle \nabla f(x), y - x\rangle\big| \leq L_f\, \mathbf{D}_h(y,x)$ for all $x, y \in \mathbf{int}\,\mathbf{dom}\, h$.*

3. Bregman proximal mapping and Moreau envelope. This section is devoted to the analysis of the Bregman proximal mapping and the Bregman–Moreau envelope, which we will need in the coming sections. Our main results include their local equivalence with Euclidean objects, namely the forward-backward mapping and the FBE (Theorem 3.8), respectively, which in turn will be used to derive novel second-order properties (Theorem 3.11) as simple byproducts of similar results available in the literature.

Relative to a Legendre kernel $\hat{h} : \mathbb{R}^n \to \overline{\mathbb{R}}$, the *Bregman proximal mapping of* $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ with stepsize $\gamma > 0$ is the set-valued map $\mathbf{prox}^{\hat{h}}_{\gamma\varphi} : \mathbf{int}\,\mathbf{dom}\, \hat{h} \rightrightarrows \mathbb{R}^n$ given by

$$\mathbf{prox}^{\hat{h}}_{\gamma\varphi}(x) = \mathbf{prox}^{\hat{h}/\gamma}_{\varphi}(x) \coloneqq \operatorname*{\mathbf{arg\,min}}_{z \in \mathbb{R}^n} \Big\{ \varphi(z) + \tfrac{1}{\gamma}\, \mathbf{D}_{\hat{h}}(z,x) \Big\}, \tag{3.1}$$

and the corresponding *Bregman–Moreau envelope* is $\varphi^{\hat{h}/\gamma} : \mathbb{R}^n \to [-\infty, \infty]$ defined as

$$\varphi^{\hat{h}/\gamma}(x) := \inf_{z\in\mathbb{R}^n} \left\{\varphi(z) + \tfrac{1}{\gamma}\mathbf{D}_{\hat{h}}(z, x)\right\}. \tag{3.2}$$

The first equality in (3.1) owes to the invariance of the **arg min** operator under positive scalings, whereas the notation in (3.2) is justified from the identity $\frac{1}{\gamma}\mathbf{D}_{\hat{h}} = \mathbf{D}_{\hat{h}/\gamma}$. Although writing $\mathbf{prox}_{\varphi}^{\hat{h}/\gamma}$ better reflects the kinship with the envelope $\varphi^{\hat{h}/\gamma}$, the (equally well-posed) notation $\mathbf{prox}_{\gamma\varphi}^{\hat{h}}$ is more consistent with the classical $\mathbf{prox}_{\gamma\varphi}$ adopted in the Euclidean setting, i.e., when $\hat{h} = \frac{1}{2}\|\cdot\|^2$. To maintain this consistency and for simplicity of exposition, the distance-generating function $\hat{h}$ will be omitted in the Euclidean case; we will thus write $\mathbf{prox}_{\gamma\varphi}$ and $\varphi^{1/\gamma}$ to indicate the Euclidean proximal map and Moreau envelope of φ with stepsize γ.

The proximal mapping is a fundamental building block of splitting algorithms, and the corresponding envelope function offers an extremely valuable tool for the convergence analysis of such schemes. For this reason, starting from the pioneering work [59] a lot of research has been devoted to the study of their properties in the Euclidean case [69, 71, 70, 47, 58]. It is well known, for instance, that when φ is a (proper, lsc, and) convex function, its (Euclidean) proximal mapping is (single-valued and) nonexpansive, and its Moreau envelope is convex and Lipschitz differentiable [14, sect. 12]. More generally, even in the nonconvex setting, the Moreau envelope is a continuous function sharing infimum and minimizers with φ, and local smoothness properties have been established with variational analysis tools such as prox-regularity and epigraphical differentiation [69, 71, 70].

As detailed in the introduction, the extension to non-Euclidean $\hat{h}$ offers a significant additional degree of flexibility. Furthermore, the Bregman proximal mapping can encapsulate an entire splitting algorithm, as is the case of the (Euclidean) proximal-gradient operator $\mathbf{prox}_{\gamma g}(\mathrm{id} - \gamma\nabla f)$ that can be expressed as $\mathbf{prox}_{f+g}^{\hat{h}}$ for $\hat{h} := \frac{1}{2\gamma}\|\cdot\|^2 - f$. In fact, it will be shown in Theorem 4.1 that this is still true even for the proximal-gradient operator with arbitrary Bregman metrics. In other words, from a theoretical standpoint *the proximal-gradient scheme offers no advantage in generality over the proximal point algorithm in the Bregman setting.*

This awareness emphasizes the importance of studying the Bregman proximal mapping in full generality. Nonetheless, in the nonconvex setting, there is a big discrepancy between the well-studied Euclidean setting and the less mature Bregman generalization. In an attempt to partially fill this gap, this section complements the analysis of [42, 45] for the proximal mapping and the Moreau envelope in the Bregman setting. The extension—or better, the "translation"—of the results for the proximal gradient will then be derived as simple byproducts in the following section. As the *translation* entails a change of Bregman metric, in order to avoid confusion we use the hat version $\hat{h}$ in this section and reserve the notation h for the distance-generating function involved in the proximal-gradient mapping. We begin by introducing the notion of Bregman-type prox-boundedness, which is a technical requirement ensuring the well definedness of the proximal map and the properness of the Moreau envelope.

DEFINITION 3.1 ($\hat{h}$-prox-boundedness). *Given a Legendre kernel $\hat{h}$, a function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is said to be* $\hat{h}$-prox-bounded *if there exists $\gamma > 0$ such that $\varphi^{\hat{h}/\gamma}(x) > -\infty$ for some $x \in \mathbb{R}^n$. The supremum of the set of all such γ is the threshold $\gamma_{\varphi}^{\hat{h}}$ of $\hat{h}$-prox-boundedness.*

Note that whenever a proper and lsc function φ is lower bounded by an affine function on $\mathbf{dom}\,\hat{h}$ (as is the case when φ is convex or lower bounded on $\mathbf{dom}\,\hat{h}$) then it is $\hat{h}$-prox-bounded with threshold $\gamma_\varphi^{\hat{h}} = \infty$. For more general functions, instead, the threshold plays a central role in dictating the range of feasible stepsizes γ.

FACT 3.2 (regularity properties of $\mathbf{prox}_{\gamma\varphi}^{\hat{h}}$ and $\varphi^{\hat{h}/\gamma}$ [42, Thm. 2.2, 2.3, and 2.4]). *Let $\hat{h}$ be a Legendre kernel, $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, lsc, and $\hat{h}$-prox-bounded, and let $\gamma \in (0, \gamma_\varphi^{\hat{h}})$. Then,*

(i) $\mathbf{dom}\,\varphi^{\hat{h}/\gamma} = \mathbf{dom}\,\mathbf{prox}_{\gamma\varphi}^{\hat{h}} = \mathbf{int}\,\mathbf{dom}\,\hat{h}$;
(ii) $\mathbf{range}\,\mathbf{prox}_{\gamma\varphi}^{\hat{h}} \subseteq \mathbf{dom}\,\varphi \cap \mathbf{dom}\,\hat{h}$;
(iii) $\mathbf{prox}_{\gamma\varphi}^{\hat{h}}$ *is locally bounded, compact-valued, and outer semicontinuous (osc; cf.* [73, Def. 5.4]*) on* $\mathbf{int}\,\mathbf{dom}\,\hat{h}$;
(iv) $\varphi^{\hat{h}/\gamma}$ *is real-valued and continuous on* $\mathbf{int}\,\mathbf{dom}\,\hat{h}$*; in fact, it is locally Lipschitz if so is* $\nabla\hat{h}$.

Next, we furnish some elementary connections between φ and its envelope $\varphi^{\hat{h}/\gamma}$.

PROPOSITION 3.3 (relation between φ and $\varphi^{\hat{h}/\gamma}$). *Let $\hat{h}$ be a Legendre kernel and $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, lsc, and $\hat{h}$-prox-bounded. Then, for every $\gamma \in (0, \gamma_\varphi^{\hat{h}})$*

(i) $\varphi(\bar{x}) + \frac{1}{\gamma}\mathbf{D}_{\hat{h}}(\bar{x}, x) = \varphi^{\hat{h}/\gamma}(x) \le \varphi(x)$ *for $x \in \mathbf{int}\,\mathbf{dom}\,\hat{h}$ and $\bar{x} \in \mathbf{prox}_{\gamma\varphi}^{\hat{h}}(x)$, with $\varphi^{\hat{h}/\gamma}(x) = \varphi(x)$ iff $x \in \mathbf{prox}_{\gamma\varphi}^{\hat{h}}(x)$.*

Moreover, if $\mathbf{range}\,\mathbf{prox}_{\gamma\varphi}^{\hat{h}} \subseteq \mathbf{int}\,\mathbf{dom}\,\hat{h}$*, then the following also hold:*

(ii) $\mathbf{inf}\,\varphi^{\hat{h}/\gamma} = \mathbf{inf}_{\mathbf{int}\,\mathbf{dom}\,\hat{h}}\,\varphi$ *and* $\mathbf{arg\,min}\,\varphi^{\hat{h}/\gamma} = \mathbf{arg\,min}_{\mathbf{int}\,\mathbf{dom}\,\hat{h}}\,\varphi$;
(iii) $\varphi^{\hat{h}/\gamma}$ *is level bounded iff φ is level bounded on* $\mathbf{int}\,\mathbf{dom}\,\hat{h}$.

Proof.

Proposition 3.3(i). The first equality is the definition of the Bregman proximal map and Moreau envelope; the inequality follows by considering $z = x$ in the subproblem (3.1) defining $\varphi^{\hat{h}/\gamma}$. In turn, the "iff" condition owes to the fact that $\mathbf{D}_{\hat{h}}(z, x) = 0$ iff $z = x$ for all $x, z \in \mathbf{int}\,\mathbf{dom}\,\hat{h}$.

Proposition 3.3(ii). It follows from assertion 3.3(i) that $\mathbf{inf}\,\varphi^{\hat{h}/\gamma} \le \mathbf{inf}_{\mathbf{int}\,\mathbf{dom}\,\hat{h}}\,\varphi$. Let a sequence $(x^k)_{k\in\mathbb{N}}$ be such that $\varphi^{\hat{h}/\gamma}(x^k) \to \mathbf{inf}\,\varphi^{\hat{h}/\gamma}$ as $k \to \infty$. Then, taking $\bar{x}^k \in \mathbf{prox}_{\gamma\varphi}^{\hat{h}}(x^k) \subseteq \mathbf{int}\,\mathbf{dom}\,\hat{h}$, assertion 3.3(i) ensures that $\mathbf{lim\,inf}_{k\to\infty}\,\varphi(\bar{x}^k) \le \mathbf{inf}_{\mathbf{int}\,\mathbf{dom}\,\hat{h}}\,\varphi$, hence the claimed equivalence of infima. If $x \in \mathbf{arg\,min}\,\varphi^{\hat{h}/\gamma}$, necessarily $x \in \mathbf{dom}\,\varphi^{\hat{h}/\gamma} = \mathbf{int}\,\mathbf{dom}\,\hat{h}$ and there thus exists $\bar{x} \in \mathbf{prox}_{\gamma\varphi}^{\hat{h}}(x)$ (cf. Fact 3.2), which satisfies $\bar{x} \in \mathbf{int}\,\mathbf{dom}\,\hat{h}$ by assumption. Then,

$$\varphi(\bar{x}) \le \varphi^{\hat{h}/\gamma}(x) - \tfrac{1}{\gamma}\mathbf{D}_{\hat{h}}(\bar{x}, x) = \mathbf{inf}\,\varphi^{\hat{h}/\gamma} - \tfrac{1}{\gamma}\mathbf{D}_{\hat{h}}(\bar{x}, x) = \inf_{\mathbf{int}\,\mathbf{dom}\,\hat{h}}\varphi - \tfrac{1}{\gamma}\mathbf{D}_{\hat{h}}(\bar{x}, x),$$

where the inequality follows from assertion 3.3(i). Therefore, $\mathbf{D}_{\hat{h}}(\bar{x}, x) = 0$ or, equivalently, $x = \bar{x} \in \mathbf{arg\,min}_{\mathbf{int}\,\mathbf{dom}\,\hat{h}}\,\varphi$. Similarly, for $x \in \mathbf{arg\,min}_{\mathbf{int}\,\mathbf{dom}\,\hat{h}}\,\varphi$ it follows from assertion 3.3(i) that $\varphi^{\hat{h}/\gamma}(x) \le \varphi(x) = \mathbf{inf}_{\mathbf{int}\,\mathbf{dom}\,\hat{h}}\,\varphi = \mathbf{inf}\,\varphi^{\hat{h}/\gamma}$, proving that $x \in \mathbf{arg\,min}\,\varphi^{\hat{h}/\gamma}$.

Proposition 3.3(iii). It follows from Proposition 3.3(i) that if $\varphi^{\hat{h}/\gamma}$ is level bounded, then so is φ on $\mathbf{int}\,\mathbf{dom}\,\hat{h}$. Conversely, suppose that there exists $\alpha \in \mathbb{R}$ together with an unbounded sequence $(x^k)_{k\in\mathbb{N}} \subseteq [\varphi^{\hat{h}/\gamma} \le \alpha]$. Then, it follows from Fact 3.2 that $x^k \in \mathbf{int}\,\mathbf{dom}\,\hat{h}$ for all k and in turn that for any k there exists $\bar{x}^k \in \mathbf{prox}_{\gamma\varphi}^{\hat{h}}(x^k)$ which satisfies $\varphi(\bar{x}^k) \le \alpha - \mathbf{D}_{\hat{h}}(\bar{x}^k, x^k)$ by Proposition 3.3(i). Local boundedness of $\mathbf{D}_h$ with respect to the second variable (Fact 2.3) then ensures that $(\bar{x}^k)_{k\in\mathbb{N}} \subset \mathbf{dom}\,\hat{h}$ is not bounded, hence that $\varphi^{\hat{h}/\gamma}$ is not level bounded. ∎

It is apparent from Fact 3.2 that the assertions of Proposition 3.3 cannot be extended outside of $\mathbf{int}\,\mathbf{dom}\,\hat{h}$. This hindrance is also at the base of the requirement $\mathbf{range}\,\mathbf{prox}^{\hat{h}}_{\gamma\varphi} \subseteq \mathbf{int}\,\mathbf{dom}\,\hat{h}$; while guaranteed in case φ is either a convex function or if it is strictly continuous on $\mathbf{dom}\,\varphi \cap \mathbf{dom}\,\hat{h} \cap \mathbf{bdry}\,\mathbf{dom}\,\hat{h}$,[2] to the best of our knowledge it always stands as a blanket assumption in nonconvex Bregman optimization. It is also common in practice that $\hat{h}$ is twice continuously differentiable, in which case we can easily derive subdifferential properties of the Bregman–Moreau envelope similarly to what was done in [42, sect. 3].

PROPOSITION 3.4 (subdifferential properties of the Bregman–Moreau envelope). *Let $\hat{h}$ be a Legendre kernel with $\hat{h} \in \mathcal{C}^2(\mathbf{int}\,\mathbf{dom}\,\hat{h})$ and let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, lsc, and $\hat{h}$-prox-bounded. Then, for every $\gamma \in (0, \gamma^{\hat{h}}_{\varphi})$ the envelope $\varphi^{\hat{h}/\gamma}$ is strictly differentiable wherever it is differentiable. Moreover, for every $x \in \mathbf{int}\,\mathbf{dom}\,\hat{h}$*

(i) $\mathbf{lip}\,\varphi^{\hat{h}/\gamma}(x) = \mathbf{max}_{\bar{x}\in\mathbf{prox}^{\hat{h}}_{\gamma\varphi}(x)} \big\|\frac{1}{\gamma}\nabla^2\hat{h}(x)(x-\bar{x})\big\|$;

(ii) $\partial\varphi^{\hat{h}/\gamma}(x) = \partial_B\varphi^{\hat{h}/\gamma}(x) \subseteq \frac{1}{\gamma}\nabla^2\hat{h}(x)\big(x - \mathbf{prox}^{\hat{h}}_{\gamma\varphi}(x)\big)$.

Proof. We pattern the proof after [73, Ex. 10.32]. Let $\mathcal{U} \subset \overline{\mathcal{U}} \subset \mathbf{int}\,\mathbf{dom}\,\hat{h}$ be a bounded open set containing x, and observe that by osc of $\mathbf{prox}^{\hat{h}}_{\gamma\varphi}$ (Fact 3.2) there exists a compact set $\mathcal{V} \subseteq \mathbf{dom}\,\hat{h}$ such that $-\varphi^{\hat{h}}(u) = \mathbf{max}_{v\in\mathcal{V}}\,\Phi(u,v)$ for all $u \in \mathcal{U}$, where $\Phi(u,v) \coloneqq -\varphi(v) - \frac{1}{\gamma}\mathbf{D}_{\hat{h}}(v,u)$ is $\mathcal{C}^1$ in u, its derivatives depending continuously on (u,v) with $\nabla_u\Phi(u,v) = \frac{1}{\gamma}\nabla^2\hat{h}(u)(v-u)$. In fact, the maxima are attained for $v \in \mathbf{prox}^{\hat{h}}_{\gamma\varphi}(u)$. Function $-\varphi^{\hat{h}}$ is thus *lower-$\mathcal{C}^1$* in the sense of [73, Def. 10.29], and all claims then follow from [73, Thm. 10.31]. ∎

3.1. Fixed points. Similarly to the Euclidean setting, being a fixed point of the Bregman proximal mapping is an intermediate property between stationarity and global minimality, the three conditions being equivalent under convexity. As we will see in later subsections, single-valuedness of the proximal mapping at fixed points is of paramount importance for local regularity properties and thus deserves a dedicated definition.

DEFINITION 3.5 (nondegenerate fixed point). *We say that $x \in \mathbb{R}^n$ is a* fixed point *of the set-valued mapping $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ if $x \in T(x)$. We say that x is* nondegenerate *if $T(x) = \{x\}$.*

Fortunately, degenerate fixed points are rarely encountered; in fact, any fixed point x of $\mathbf{prox}^{\hat{h}}_{\gamma\varphi}$ is also a nondegenerate fixed point of $\mathbf{prox}^{\hat{h}}_{\gamma'\varphi}$ for every $\gamma' \in (0,\gamma)$. As such, out of a continuous set of favorable stepsizes γ' there exists at most one such that $\{x\} \subsetneq \mathbf{prox}^{\hat{h}}_{\gamma\varphi}(x)$. In this sense, we may consider nondegeneracy as a negligible condition. These claims are validated in the following result.

LEMMA 3.6. *Let $\hat{h} : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a Legendre kernel and $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, lsc, and $\hat{h}$-prox-bounded. If $x \in \mathbf{int}\,\mathbf{dom}\,\hat{h}$ is a fixed point of $\mathbf{prox}^{\hat{h}}_{\gamma\varphi}$, then it is a nondegenerate fixed point of $\mathbf{prox}^{\hat{h}+\hat{h}'}_{\gamma\varphi}$ for every strictly convex function $\hat{h}' : \mathbb{R}^n \to \overline{\mathbb{R}}$ differentiable on $\mathbf{int}\,\mathbf{dom}\,\hat{h}' \supseteq \mathbf{int}\,\mathbf{dom}\,\hat{h}$. In particular, x is a nondegenerate fixed point of $\mathbf{prox}^{\hat{h}}_{\gamma'\varphi}$ for every $\gamma' \in (0,\gamma)$.*

Proof. By definition, $x \in \mathbf{prox}^{\hat{h}}_{\gamma\varphi}$ iff $\varphi(z) + \frac{1}{\gamma}\mathbf{D}_{\hat{h}}(z,x) \geq \varphi(x)$ holds for any $z \in \mathbb{R}^n$. Since $\hat{h}+\hat{h}'$ is a Legendre kernel, $\mathbf{D}_{\hat{h}+\hat{h}'} = \mathbf{D}_{\hat{h}} + \mathbf{D}_{\hat{h}'}$ on $\mathbf{dom}\,\hat{h} \times \mathbf{int}\,\mathbf{dom}\,\hat{h}$,

[2] The convex case is discussed in the proof of [11, Lem. 2]; the other case can use the same arguments therein owing to the subdifferential calculus rule of [73, Ex. 10.10]. An easy counterexample is given by $\hat{h}(x) = x\,\mathbf{ln}\,x - x$ and $\varphi(x) = x^2 + x - \hat{h}(x)$, having $\mathbf{prox}^{\hat{h}}_{\varphi}(x) = \{0\} \not\subseteq \mathbf{int}\,\mathbf{dom}\,h$ for $x \leq \mathbf{exp}(1)$.

and $\mathbf{D}_{\hat{h}'}(z,x) = 0$ iff $z = x$, apparently x is also a fixed point of $\mathbf{prox}_{\gamma\varphi}^{\hat{h}+\hat{h}'}$, with $\varphi(z) + \frac{1}{\gamma}\mathbf{D}_{\hat{h}+\hat{h}'}(z,x) > \varphi(x)$ for all $z \neq x$. Thus, if contrary to the claim there exists $x' \in \mathbf{prox}_{\gamma\varphi}^{\hat{h}+\hat{h}'}(x) \setminus \{x\}$, then

$$\varphi(x') + \tfrac{1}{\gamma}\mathbf{D}_{\hat{h}+\hat{h}'}(x',x) > \varphi(x) = \varphi^{(\hat{h}+\hat{h}')/\gamma}(x) = \varphi(x') + \tfrac{1}{\gamma}\mathbf{D}_{\hat{h}+\hat{h}'}(x',x),$$

which is a contradiction. ∎

The following result shows that being a fixed point for the proximal map (for some stepsize γ) is actually a necessary condition for *local* minimality in the interior of the domain of the Legendre kernel $\hat{h}$. An equivalence of local minimality for function φ and its envelope $\varphi^{\hat{h}/\gamma}$ at fixed points is also shown, in which nondegeneracy plays a key role. We remark that this condition is necessary even in the Euclidean setting, as can be verified from the counterexample $\varphi = \frac{1}{2}\|\cdot\|^2 + \delta_{\{0,1\}}$ (see [80, Fig. 3.1] and discussion therein).

THEOREM 3.7 (equivalence of local minimality). *The following hold for a Legendre kernel $\hat{h} : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a proper, lsc, $\hat{h}$-prox-bounded function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$:*

(i) *If $x_\star \in \mathbf{int}\,\mathbf{dom}\,\hat{h}$ is a local minimum for φ, then it is a nondegenerate fixed point of $\mathbf{prox}_{\gamma\varphi}^{\hat{h}}$ for any γ small enough.*

(ii) *Conversely, any nondegenerate fixed point $x_\star$ of $\mathbf{prox}_{\gamma\varphi}^{\hat{h}}$ is a local minimum for φ iff it is a local minimum for $\varphi^{\hat{h}/\gamma}$. Moreover, equivalence of strong local minimality also holds provided that $\hat{h}$ is strongly convex in a neighborhood of $x_\star$.*

Proof.

Theorem 3.7(i). For $\gamma \in (0, \gamma_\varphi^{\hat{h}})$, let $x_\gamma \in \mathbf{prox}_{\gamma\varphi}^{\hat{h}}(x_\star)$. It is easy to see that $x_\gamma \to x_\star$ as $\gamma \searrow 0$ (cf. proof of [42, Thm. 2.5]). Local minimality of $x_\star$ thus implies that $\varphi(x_\gamma) \geq \varphi(x_\star)$ holds for γ small enough. Combined with Proposition 3.3(i) we obtain that $\varphi(x_\gamma) + \frac{1}{\gamma}\mathbf{D}_{\hat{h}}(x_\gamma, x_\star) \leq \varphi(x_\star) \leq \varphi(x_\gamma)$ holds for γ small enough, hence that $\mathbf{D}_{\hat{h}}(x_\gamma, x_\star) = 0$ or, equivalently, $x_\gamma = x_\star$.

Theorem 3.7(ii). That (strong) local minimality for $\varphi^{\hat{h}/\gamma}$ implies that for φ follows from the fact that $\varphi^{\hat{h}/\gamma}$ "supports" φ at $x_\star$, namely that $\varphi^{\hat{h}/\gamma} \leq \varphi$ and $\varphi^{\hat{h}/\gamma}(x_\star) = \varphi(x_\star)$ (Proposition 3.3(i)). Conversely, suppose that $\hat{h}$ is $\sigma_{\hat{h},\mathcal{U}}$-strongly convex in a neighborhood $\mathcal{U}$ of $x_\star$ for some $\sigma_{\hat{h},\mathcal{U}} \geq 0$ and that there exists $\mu \geq 0$ such that $\varphi(x) \geq \varphi(x_\star) + \frac{\mu}{2}\|x - x_\star\|^2$ for $x \in \mathcal{U}$. Notice that in allowing $\sigma_{\hat{h},\mathcal{U}} = 0$ and $\mu = 0$ we also cover nonstrong minimality and nonstrong convexity. Let $\delta := \frac{1}{2}\mathbf{min}\{\mu, \frac{\sigma_{\hat{h},\mathcal{U}}}{2\gamma}\} \geq 0$, and note that $\delta = 0$ iff either $\sigma_{\hat{h},\mathcal{U}}$ or μ is zero. To arrive to a contradiction, suppose that for all $k \geq 1$ there exists $x^k \in \mathbf{B}(x_\star; 1/k)$ such that $\varphi^{\hat{h}/\gamma}(x^k) < \varphi^{\hat{h}/\gamma}(x_\star) + \frac{\delta}{2}\|x^k - x_\star\|^2 = \varphi(x_\star) + \frac{\delta}{2}\|x^k - x_\star\|^2$. Let $\bar{x}^k \in \mathbf{prox}_{\gamma\varphi}^{\hat{h}}(x^k)$; since $\mathbf{prox}_{\gamma\varphi}^{\hat{h}}$ is osc on $\mathbf{int}\,\mathbf{dom}\,\hat{h} \ni x_\star$ (cf. Fact 3.2) and $\mathbf{prox}_{\gamma\varphi}^{\hat{h}}(x_\star) = \{x_\star\}$, necessarily $\bar{x}^k \to x_\star$ as $k \to \infty$. We have

$$\varphi(\bar{x}^k) \overset{3.3(i)}{=} \varphi^{\hat{h}/\gamma}(x^k) - \tfrac{1}{\gamma}\mathbf{D}_{\hat{h}}(\bar{x}^k, x^k) \overset{2.3}{\leq} \varphi^{\hat{h}/\gamma}(x^k) - \tfrac{\sigma_{\hat{h},\mathcal{U}}}{2\gamma}\|x^k - \bar{x}^k\|^2 < \varphi(x_\star) + \tfrac{\delta}{2}\|x^k - x_\star\|^2 - \tfrac{\sigma_{\hat{h},\mathcal{U}}}{2\gamma}\|x^k - \bar{x}^k\|^2.$$

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

$$\varphi(\bar{x}^k) < \varphi(x_\star) + \delta\|\bar{x}^k - x_\star\|^2 + \big(\delta - \tfrac{\sigma_{\hat{h},\mathcal{U}}}{2\gamma}\big)\|x^k - \bar{x}^k\|^2 \leq \varphi(x_\star) + \tfrac{\mu}{2}\|\bar{x}^k - x_\star\|^2,$$

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 $\bar{x}^k \to x_\star$). ∎

3.2. Local Euclidean reparametrization. While the Euclidean proximal mapping and Moreau envelope are special instances of the more general Bregman variants, the converse is not true even locally and in the convex case. To see this, observe that while the Euclidean Moreau envelope preserves convexity, this is not at all the case for more general Legendre kernels $\hat{h}$ even if strongly convex and Lipschitz smooth on the entire space. Nevertheless, under a local strong convexity and Lipschitz differentiability assumption on $\hat{h}$, yet with no requirement on its domain, it is possible to locally identify the Bregman proximal mapping and its Moreau envelope with Euclidean objects, namely the forward-backward mapping and the corresponding FBE function [68, 83]. This result complements what was first observed in [51], namely that the Euclidean FBE is, in fact, a Bregman–Moreau envelope. The advantage of this identification will be revealed in the next subsections, where local differentiability results will be deduced with virtually no effort based on already established results in the Euclidean setting.

We remind the reader that given a decomposition $\varphi = \tilde{f} + \tilde{g}$ with $\tilde{f}$ continuously differentiable, the (Euclidean) forward-backward operator with stepsize $\tilde{\gamma} > 0$ is

$$\mathbf{prox}_{\tilde{\gamma}\tilde{g}}\left(x - \tilde{\gamma}\nabla\tilde{f}(x)\right) = \operatorname*{\mathbf{arg\,min}}_{w\in\mathbb{R}^n}\left\{\tilde{f}(x) + \langle\nabla\tilde{f}(x), w - x\rangle + \tilde{g}(w) + \tfrac{1}{2\tilde{\gamma}}\|w - x\|^2\right\},$$

while the FBE is the associated value function, namely

$$\varphi^{\tilde{f},\tilde{g}}_{1/\tilde{\gamma}}(x) = \operatorname*{\mathbf{inf}}_{w\in\mathbb{R}^n}\left\{\tilde{f}(x) + \langle\nabla\tilde{f}(x), w - x\rangle + \tilde{g}(w) + \tfrac{1}{2\tilde{\gamma}}\|w - x\|^2\right\}. \tag{3.3}$$

THEOREM 3.8 (local equivalence of Bregman–Moreau envelope and FBE). *Let $\hat{h} : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a Legendre kernel, let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, lsc, and $\hat{h}$-prox-bounded, and let $\gamma \in (0, \gamma^{\hat{h}}_{\varphi})$ be fixed. Suppose further that $\hat{h}$ is locally Lipschitz differentiable and locally strongly convex on $\operatorname{\mathbf{int}}\operatorname{\mathbf{dom}}\hat{h}$ (as is the case when $\hat{h} \in \mathcal{C}^2$ with $\nabla^2\hat{h} \succ 0$ on $\operatorname{\mathbf{int}}\operatorname{\mathbf{dom}}\hat{h}$) and that $\operatorname{\mathbf{range}}\mathbf{prox}^{\hat{h}}_{\gamma\varphi} \subseteq \operatorname{\mathbf{int}}\operatorname{\mathbf{dom}}\hat{h}$. Then, for every compact set $\mathcal{U} \subset \operatorname{\mathbf{int}}\operatorname{\mathbf{dom}}\hat{h}$ there exist $\bar{\gamma} > 0$ and a convex compact set $\mathcal{V}$ with $\mathcal{U} \subseteq \mathcal{V} \subset \operatorname{\mathbf{int}}\operatorname{\mathbf{dom}}\hat{h}$ such that for all $\tilde{\gamma} \in (0, \bar{\gamma})$ it holds that*

$$\varphi^{\hat{h}/\gamma} = \varphi^{\tilde{f},\tilde{g}}_{1/\tilde{\gamma}} \quad \text{and} \quad \mathbf{prox}^{\hat{h}}_{\gamma\varphi} = \mathbf{prox}_{\tilde{\gamma}\tilde{g}}\left(\mathrm{id} - \tilde{\gamma}\nabla\tilde{f}\right) \text{ on } \mathcal{U}, \tag{3.4}$$

where

$$\tilde{f} := -\tfrac{1}{\gamma}\hat{h} + \tfrac{1}{2\tilde{\gamma}}\|\cdot\|^2 \quad \text{and} \quad \tilde{g} := \varphi + \tfrac{1}{\gamma}\hat{h} - \tfrac{1}{2\tilde{\gamma}}\|\cdot\|^2 + \delta_{\mathcal{V}}\,. \tag{3.5}$$

Moreover, $\tilde{g}$ is proper, lsc, and prox-bounded (in the Euclidean sense) with $\gamma_{\tilde{g}} = \infty$, and $\tilde{f}$ is $L_{\tilde{f}}$-Lipschitz-differentiable on $\mathcal{V}$ with $\tilde{\gamma} < 1/L_{\tilde{f}}$.

Proof. Let $\mathcal{V} := \mathbf{conv}(\mathcal{U} \cup \mathbf{prox}^{\hat{h}}_{\gamma\varphi}(\mathcal{U}))$ and observe that $\mathcal{V}$ is a convex compact subset of $\operatorname{\mathbf{int}}\operatorname{\mathbf{dom}}\hat{h}$, as it follows from osc and local boundedness of $\mathbf{prox}^{\hat{h}}_{\gamma\varphi}$ (Fact 3.2) and the fact that compactness is preserved by the convex hull. It follows from the assumptions that there exist $L_{\hat{h},\mathcal{V}} \geq \sigma_{\hat{h},\mathcal{V}} > 0$ such that $\hat{h}$ is $L_{\hat{h},\mathcal{V}}$-smooth and $\sigma_{\hat{h},\mathcal{V}}$-strongly convex on $\mathcal{V}$. Define $\bar{\gamma} := \frac{2\gamma}{L_{\hat{h},\mathcal{V}}}$, let $\tilde{\gamma} \in (0, \bar{\gamma})$ be fixed, and let $\tilde{f}$ and $\tilde{g}$ be as in (3.5). Clearly, $\varphi = \tilde{f} + \tilde{g}$ on $\mathcal{V}$. Moreover,

$$\varphi(z) + \tfrac{1}{\gamma}\mathbf{D}_{\hat{h}}(z, x) + \delta_{\mathcal{V}}(z) = \tilde{f}(x) + \langle\nabla\tilde{f}(x), z - x\rangle + \tilde{g}(z) + \tfrac{1}{2\tilde{\gamma}}\|z - x\|^2 \quad \forall x \in \mathcal{U}, z \in \mathbb{R}^n.$$

Since $\mathbf{prox}^{\hat{h}}_{\gamma\varphi}(\mathcal{U}) \subseteq \mathcal{V}$,

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

Similarly, considering the minimizers it is apparent $\mathbf{prox}^{\hat{h}}_{\gamma\varphi}(x) = \mathbf{prox}_{\tilde{\gamma}\tilde{g}}\,(x - \tilde{\gamma}\nabla\tilde{f}(x))$ for any $x \in \mathcal{U}$. Notice that $\tilde{g}$ is proper, lsc, and with bounded domain, hence its claimed prox-boundedness follows. Moreover,

$$\Big(\tfrac{1}{\tilde{\gamma}} - \tfrac{L_{\hat{h},\mathcal{V}}}{\gamma}\Big)\|x-y\|^2 \le \langle \nabla\tilde{f}(x) - \nabla\tilde{f}(y), x-y\rangle \le \Big(\tfrac{1}{\tilde{\gamma}} - \tfrac{\sigma_{\hat{h},\mathcal{V}}}{\gamma}\Big)\|x-y\|^2$$

for every $x, y \in \mathcal{V}$. Therefore $\tilde{f}$ is $L_{\tilde{f}}$-smooth on $\mathcal{V}$ with $L_{\tilde{f}} = \mathbf{max}\{|\frac{1}{\tilde{\gamma}} - \frac{L_{\hat{h},\mathcal{V}}}{\gamma}|,\ |\frac{1}{\tilde{\gamma}} - \frac{\sigma_{\hat{h},\mathcal{V}}}{\gamma}|\}$. Since $\gamma < \bar{\gamma} = \frac{2\gamma}{L_{\hat{h},\mathcal{V}}}$, it follows that $-\frac{1}{\tilde{\gamma}} < \frac{1}{\tilde{\gamma}} - \frac{L_{\hat{h},\mathcal{V}}}{\gamma} \le \frac{1}{\tilde{\gamma}} - \frac{\sigma_{\hat{h},\mathcal{V}}}{\gamma} < \frac{1}{\tilde{\gamma}}$, proving that $\tilde{\gamma} < 1/L_{\tilde{f}}$. ∎

3.3. First- and second-order properties. Although strict continuity ensures almost everywhere differentiability, with mild additional assumptions the Bregman–Moreau envelope can be shown to be continuously differentiable around fixed points. Thanks to the local equivalence shown in Theorem 3.8, 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 [71] and which has been recently extended to its $\hat{h}$-relative version in [45].

DEFINITION 3.9 (prox-regularity). *A function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is prox-regular at $\bar{x}$ for $\bar{v} \in \partial\varphi(\bar{x})$ if it is locally lsc at $\bar{x}$ and there exist $r, \varepsilon > 0$ such that*

$$\varphi(x') \ge \varphi(x) + \langle v, x' - x\rangle - \tfrac{r}{2}\|x' - x\|^2 \tag{3.6}$$

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

Differentiability properties of the Bregman–Moreau envelope have been studied in [78, 13] for jointly convex Bregman distances in the convex setting, and a similar analysis for the "right" envelope is provided in [15]. For a nonconvex function φ, [42] shows global continuous differentiability of $\varphi^{\hat{h}/\gamma}$ under a global convexity assumption on $\hat{h} + \gamma\varphi$. What we provide next is instead a local result that requires local properties of φ around nondegenerate fixed points of the proximal map. We remark that after the first submission of our paper a similar result appeared in [45] in a more general setting. We, however, offer our alternative proof as a means of emphasizing the favorable theoretical implications of the Euclidean equivalence stated in Theorem 3.8, which, ultimately, will lead us to the second-order result of Theorem 3.11 which is instead novel.

THEOREM 3.10 (continuous differentiability of the Bregman–Moreau envelope). *Suppose that the assumptions of Theorem 3.8 hold, and let $x_\star$ be a nondegenerate fixed point of $\mathbf{prox}^{\hat{h}}_{\gamma\varphi}$. If φ is prox-regular at $x_\star$ for $v = 0$, then there exists a neighborhood $\mathcal{U}$ of $x_\star$ on which the following statements are true:*

(i) *$\mathbf{prox}^{\hat{h}}_{\gamma\varphi}$ is Lipschitz continuous (hence single-valued);*

(ii) *$\varphi^{\hat{h}/\gamma} \in \mathcal{C}^1$ with $\nabla\varphi^{\hat{h}/\gamma}(x) = \frac{1}{\gamma}\nabla^2\hat{h}(x)\big(x - \mathbf{prox}^{\hat{h}}_{\gamma\varphi}(x)\big)$.*

Proof. For any compact neighborhood $\mathcal{U} \subset \mathbf{int}\,\mathbf{dom}\,\hat{h}$ of $x_\star$ we may invoke Theorem 3.8 and identify $\varphi^{\hat{h}/\gamma}$ with the Euclidean FBE $\varphi^{\tilde{f},\tilde{g}}_{1/\tilde{\gamma}}$ on $\mathcal{U}$ and $\mathbf{prox}^{\hat{h}}_{\gamma\varphi}$ with $T \coloneqq \mathbf{prox}_{\tilde{\gamma}\tilde{g}}\,(\mathrm{id} - \tilde{\gamma}\nabla\tilde{f})$ for some $\tilde{\gamma} > 0$ and $\tilde{f}$ and $\tilde{g}$ as in (3.5). It follows from [73, Ex. 13.35] and the continuous differentiability of $\hat{h}$ that $\tilde{g}$ is prox-regular at $x_\star$ for $-\nabla\tilde{f}(x_\star)$. Since $\tilde{f}$ is $\mathcal{C}^2$ around $x_\star$ and $T(x_\star) = \mathbf{prox}^{\hat{h}}_{\gamma\varphi}(x_\star) = \{x_\star\}$ by assumption,

the setting of [83, Thm. 4.7] is satisfied[3] and thus T is Lipschitz continuous around $x_\star$ and the Euclidean FBE $\varphi^{\tilde{f},\tilde{g}}_{1/\tilde{\gamma}}$ is $\mathcal{C}^1$ around $x_\star$ with

(3.7)
$$\nabla\varphi^{\hat{h}/\gamma}(x) = \nabla\varphi^{\tilde{f},\tilde{g}}_{1/\tilde{\gamma}}(x) = \tilde{\gamma}^{-1}\big[\mathrm{I} - \tilde{\gamma}\nabla^2\tilde{f}(x)\big]\big(x - T(x)\big) = \tfrac{1}{\gamma}\nabla^2\hat{h}(x)\big(x - \mathbf{prox}^{\hat{h}}_{\gamma\varphi}(x)\big),$$

which completes the proof. ∎

We conclude the section with a second-order analysis *at* (as opposed to *around*) fixed points. In the spirit of Theorem 3.10, thanks to the local identities assessed in Theorem 4.1 we will simply invoke known generalized differentiability properties; we refer the interested reader to [69, 73] for an extensive discussion.

THEOREM 3.11 (twice differentiability of the Bregman–Moreau envelope). *Additionally to the assumptions of Theorem* 3.10*, suppose that φ is (strictly) twice epi-differentiable at $x_\star$ for $v = 0$, with generalized quadratic second-order epi-derivative. Then,*

(i) $\mathbf{prox}^{\hat{h}}_{\gamma\varphi}$ *is (strictly) differentiable at $x_\star$;*
(ii) $\mathbf{prox}^{\hat{h}}_{\gamma\varphi} \circ \nabla\hat{h}^*$ *is (strictly) differentiable at $\nabla\hat{h}(x_\star)$ with symmetric and positive semidefinite Jacobian;*
(iii) $\varphi^{\hat{h}/\gamma}$ *is (strictly) twice differentiable at $x_\star$ with symmetric Hessian*

$$\nabla^2\varphi^{\hat{h}/\gamma}(x_\star) = \tfrac{1}{\gamma}\nabla^2\hat{h}(x_\star)\big(\mathrm{I} - J\,\mathbf{prox}^{\hat{h}}_{\gamma\varphi}(x_\star)\big).$$

Proof. As shown in the proof of Theorem 3.10, for some $\tilde{\gamma} > 0$ and with $\tilde{f}$ and $\tilde{g}$ as in (3.5) we may identify $\varphi^{\hat{h}/\gamma}$ with the Euclidean FBE $\varphi^{\tilde{f},\tilde{g}}_{1/\tilde{\gamma}}$ around $x_\star$ and $\mathbf{prox}^{\hat{h}}_{\gamma\varphi}$ with the Euclidean proximal-gradient operator $\mathbf{prox}_{\tilde{\gamma}\tilde{g}}(\mathrm{id} - \tilde{\gamma}\nabla\tilde{f})$. It follows from [73, Ex. 13.18 and 13.25] and the continuous differentiability of $\hat{h}$ that $\tilde{g}$ is prox-regular and (strictly) twice epi-differentiable at $x_\star$ for $-\nabla\tilde{f}(x_\star)$, with generalized quadratic second-order epi-derivative. The setting of [83, Thm. 4.10] is thus satisfied (cf. footnote 3) and therefore

- $\mathbf{prox}_{\tilde{\gamma}\tilde{g}}(\mathrm{id} - \tilde{\gamma}\nabla\tilde{f}) = \mathbf{prox}^{\hat{h}}_{\gamma\varphi}$ is (strictly) differentiable at $x_\star$, which is assertion 3.11(i);
- $\varphi^{\tilde{f},\tilde{g}}_{1/\tilde{\gamma}} = \varphi^{\hat{h}/\gamma}$ is (strictly) twice differentiable at $x_\star$ with symmetric Hessian

$$\tfrac{1}{\tilde{\gamma}}\big[\mathrm{I}-\tilde{\gamma}\nabla^2\tilde{f}(x_\star)\big]\big(\mathrm{I}-J\big(\mathbf{prox}_{\tilde{\gamma}\tilde{g}}(x_\star - \tilde{\gamma}\nabla\tilde{f}(x_\star))\big)\big) = \tfrac{1}{\gamma}\nabla^2\hat{h}(x_\star)\big(\mathrm{I}-J\,\mathbf{prox}^{\hat{h}}_{\gamma\varphi}(x_\star)\big),$$

  which is assertion 3.11(iii);
- $\mathbf{prox}_{\tilde{\gamma}\tilde{g}}$ is (strictly) differentiable at $x_\star - \tilde{\gamma}\nabla\tilde{f}(x_\star)$ with symmetric Jacobian. We have

$$\begin{aligned}
\mathbf{prox}_{\tilde{\gamma}\tilde{g}}(s) &= \operatorname*{\mathbf{arg\,min}}_{w\in\mathcal{V}} \Big\{\varphi(w) + \tfrac{1}{\gamma}\hat{h}(w) - \tfrac{1}{2\tilde{\gamma}}\|w\|^2 + \tfrac{1}{2\tilde{\gamma}}\|w - s\|^2\Big\} \\
&= \operatorname*{\mathbf{arg\,min}}_{w\in\mathcal{V}} \Big\{\varphi(w) + \tfrac{1}{\gamma}\hat{h}(w) - \tfrac{1}{\gamma}\langle\tfrac{\gamma}{\tilde{\gamma}}s, w\rangle\Big\} \\
&= \operatorname*{\mathbf{arg\,min}}_{w\in\mathcal{V}} \Big\{\varphi(w) + \tfrac{1}{\gamma}\,\mathbf{D}_{\hat{h}}\big(w, \nabla\hat{h}^*(\tfrac{\gamma}{\tilde{\gamma}}s)\big)\Big\}.
\end{aligned}$$

[3] The requirement $\gamma \in (0, \Gamma(x_\star))$ in [83, Thm. 4.7] is needed only for ensuring that $T(x_\star) = \{x_\star\}$ is a singleton (which here follows by assumption), and consequently that strict inequality in (4.4) therein holds for $x \neq x_\star$. In fact, the notion of *criticality* in [83] corresponds to that of being a fixed point (cf. [83, Def. 3.1]), and the bound $\gamma \in (0, \Gamma(x_\star))$ therein guarantees the nondegeneracy as defined in Definition 3.5 (cf. [83, Thm. 3.4(iii)]).

Since $\mathbf{prox}^{\hat{h}}_{\gamma\varphi}(x)$ is contained in $\mathcal{V}$ for all points close to $x_\star$, as apparent from the definition of $\mathcal{V}$ in the proof of Theorem 3.8, for all points s such that $\frac{\gamma}{\tilde{\gamma}}s$ is close to $\nabla\hat{h}(x_\star)$ the above formula coincides with $\mathbf{prox}_{\gamma\varphi}\circ\nabla\hat{h}^*(\frac{\gamma}{\tilde{\gamma}}s)$. The proof of assertion 3.11(ii) now follows by observing that $x_\star - \tilde{\gamma}\nabla\tilde{f}(x_\star) = \frac{\tilde{\gamma}}{\gamma}\nabla\hat{h}(x_\star)$. ∎

4. Bregman forward-backward mapping and forward-backward envelope. As a last step toward the algorithm presented in the next section, we here analyze its main building block, the Bregman forward-backward operator. We thus go back to the composite minimization setting of the investigated problem, stated again here for the reader's convenience,

$$\text{(P)} \qquad \textbf{minimize}\, \varphi(x) \equiv f(x) + g(x) \quad \textbf{subject to}\, x \in \overline{C},$$

and which will be addressed under the following assumptions.

Assumption 1 (requirements for the composite minimization (P)). The following hold:

- A1. $h : \mathbb{R}^n \to \overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{\infty\}$ is a Legendre kernel with $\mathbf{int}\,\mathbf{dom}\, h = C$;
- A2. $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is L_f-smooth relative to h;
- A3. $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper and lower semicontinuous (lsc);
- A4. $\mathbf{arg\,min}\left\{\varphi(x) \mid x \in \overline{C}\right\} \neq \emptyset$;
- A5. $\mathbf{range}\, \mathbf{T}^{f,g}_{h/\gamma} \subseteq C$ for $\gamma \in (0, 1/L_f)$, where

$$\mathbf{T}^{f,g}_{h/\gamma}(x) \coloneqq \operatorname*{\mathbf{arg\,min}}_{z\in\mathbb{R}^n} \left\{ f(x) + \langle \nabla f(x), z - x\rangle + g(z) + \tfrac{1}{\gamma}\mathbf{D}_h(z,x) \right\} \tag{4.1}$$

is the Bregman forward-backward mapping.

Similarly to what was noted in the previous section, special care is needed when dealing with boundary points. Although some of the results would hold in a more general setting, including requirement 1.A5 among the blanket assumptions considerably simplifies the analysis. Moreover, the range inclusion ensures that any output of $\mathbf{T}^{f,g}_{h/\gamma}$ can again be fed to $\mathbf{T}^{f,g}_{h/\gamma}$—having $\mathbf{dom}\, \mathbf{T}^{f,g}_{h/\gamma} = C$; cf. Proposition 4.2—and is essential for the well definedness of the BELLA algorithm that will be presented in the next section.

Our approach hinges on two analogies, one based on the local equivalence of the Bregman proximal map and the Euclidean forward-backward mapping given in Theorem 3.8 and particularly useful for asymptotic analyses, and the other one based on the equivalence of forward-backward and proximal mappings in the Bregman setting. The latter identity, which we show next in Theorem 4.1, leads to a simpler analysis of the Bregman forward-backward mapping in allowing us to disregard the decomposition $f + g$ to solely focus on the cost function φ. Most importantly, it enables the possibility of making use of the Bregman–Moreau envelope in the algorithmic analysis, whence the thorough study carried out in the previous section will be heavily exploited. In this perspective, in the spirit of the Bregman–Moreau envelope and its relation with the Bregman proximal mapping, we construct an "envelope" for the Bregman forward-backward operator by considering the value function associated to the minimization problem (4.1) defining $\mathbf{T}^{f,g}_{h/\gamma}$: we define the BFBE as the function $\varphi^{f,g}_{h/\gamma} : \mathbb{R}^n \to \overline{\mathbb{R}}$ given by

$$\varphi^{f,g}_{h/\gamma}(x) \coloneqq \inf_{z\in\mathbb{R}^n} \left\{ f(x) + \langle \nabla f(x), z - x\rangle + g(z) + \tfrac{1}{\gamma}\mathbf{D}_h(z,x) \right\}. \tag{4.2}$$

The BFBE is the value function of the majorization-minimization (MM) problem (4.1) defining the Bregman proximal mapping and serves as the Lyapunov function for the iterates generated by the linesearch algorithm discussed in the next section. Differently from other works that also employ value functions in the convergence analysis of MM-type algorithms (see, e.g., [56, 22]), our analysis makes extensive use of the *continuity* of the BFBE, which is shown in Proposition 4.2. This property will fall as a simple corollary of Fact 3.2 once the equivalence between the Bregman–Moreau envelope and BFBE is established in the next result.

THEOREM 4.1 (equivalence of forward-backward and proximal point mappings). *Suppose that f is L_f-relatively smooth with respect to a Legendre kernel h. Then, for every $\gamma \in (0, 1/L_f)$ the function $\hat{h} \coloneqq \frac{h}{\gamma} - f$ (with the convention $\infty - \infty = \infty$) is a Legendre kernel. Moreover,*

$$\varphi(z) + \mathbf{D}_{\hat{h}}(z, x) = f(x) + \langle \nabla f(x), z - x \rangle + g(z) + \tfrac{1}{\gamma} \mathbf{D}_h(z, x) \tag{4.3}$$

holds for any $(z, x) \in \mathbb{R}^n \times \mathbf{int}\,\mathbf{dom}\, h$, and in particular

$$\mathbf{T}_{h/\gamma}^{f,g}(x) = \mathbf{prox}_{\varphi}^{h/\gamma - f}(x) = \mathbf{prox}_{\gamma\varphi}^{h - \gamma f}(x) \tag{4.4}$$

and

$$\varphi_{h/\gamma}^{f,g}(x) = \varphi^{h/\gamma - f}(x). \tag{4.5}$$

Proof. Let $C = \mathbf{int}\,\mathbf{dom}\, h$. Observe that $\hat{h} = \frac{1-\gamma L_f}{\gamma} h + L_f h - f$ on $\mathbf{dom}\,\hat{h} = \mathbf{dom}\, h$; since $L_f h - f$ is convex on C by assumption, 1-coercivity and strict convexity on C of $\hat{h}$ follow from the similar properties of h. We now show essential smoothness; clearly, $\hat{h}$ is differentiable on C with $\nabla \hat{h} = \frac{1}{\gamma} \nabla h - \nabla f$. 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 $\mathbf{sup}_{k \in \mathbb{N}} \|\nabla \hat{h}(x^k)\| < \infty$. By possibly extracting a subsequence, we may assume that $\nabla h(x^k)/\|\nabla h(x^k)\| \to v$ for some unitary vector v. 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

$$\langle \nabla(L_f h - f)(x^k) - \nabla(L_f h - f)(y), x^k - y \rangle \leq c_y - \tfrac{1-\gamma L_f}{\gamma} \langle \nabla h(x^k) - \nabla h(y), x^k - y \rangle, \tag{4.6}$$

where $c_y \coloneqq \mathbf{sup}_{k \in \mathbb{N}} \langle \nabla \hat{h}(x^k) - \nabla \hat{h}(y), x^k - y \rangle$ is a finite quantity. Moreover, since $\|\nabla h(x^k)\| \to \infty$,

$$0 \leq \tfrac{1}{\|\nabla h(x^k)\|} \langle \nabla h(x^k) - \nabla h(y), x^k - y \rangle \to \langle v, x_\star - y \rangle \quad \text{as } k \to \infty,$$

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 $\langle v, x_\star - y \rangle \gneqq 0$. Plugging this y into (4.6) yields

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

contradicting convexity of $L_f h - f$ on C. Therefore, $\frac{1}{\gamma} h - f$ is a Legendre kernel, and in particular the right-hand side in (4.3) is well defined, with equality therein and consequent validity of (4.4) and (4.5) of immediate verification. ∎

In light of the equivalence of Theorem 4.1, properties of the Bregman proximal mapping and Bregman–Moreau envelope can directly be imported.

PROPOSITION 4.2. *Suppose that Assumption* 1 *holds and let* $\gamma \in (0, 1/L_f)$. *Then,*

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

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

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

(iv) $\mathbf{T}^{f,g}_{h/\gamma}$ *is locally bounded, compact-valued, and osc on* C;

(v) $\varphi^{f,g}_{h/\gamma}$ *is continuous on* C, *in fact, locally Lipschitz if so is* ∇h;

(vi) *if* $x \in \mathbf{T}^{f,g}_{h/\gamma}(x)$, *then* $\mathbf{T}^{f,g}_{h/\gamma'}(x) = \{x\}$ *for every* $\gamma' \in (0, \gamma)$.

Proof. It suffices to show the second equality in Proposition 4.2(i), while all other claims follow from the similar properties of the Bregman proximal mapping and Bregman–Moreau envelope in light of the equivalence of Theorem 4.1. By expanding the Bregman distance and discarding constant terms in (4.1), one has

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

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. Fact 2.3), hence the claim. ∎

The next two results characterize the fundamental relationship between the BFBE $\varphi^{f,g}_{h/\gamma}$ 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.

PROPOSITION 4.3 (relation between φ and $\varphi^{f,g}_{h/\gamma}$). *Suppose that Assumption* 1 *holds and let* $\gamma \in (0, 1/L_f)$ *be fixed. Then,*

(i) $\varphi^{f,g}_{h/\gamma}(x) \le \varphi(x)$ *for all* $x \in C$, *with equality holding iff* $x \in \mathbf{T}^{f,g}_{h/\gamma}(x)$;

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

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

(iv) $\varphi^{f,g}_{h/\gamma}$ *is level bounded iff* φ *is level bounded on* C.

Proof. All the claims follow from Theorem 4.1 and Proposition 3.3 together with the fact that $\frac{1-\gamma L_f}{\gamma}\, \mathbf{D}_h \le \mathbf{D}_{\hat{h}} \le \frac{1+\gamma L_f}{\gamma}\, \mathbf{D}_h$ for $\hat{h} = \frac{1}{\gamma} h - f$ (owing to convexity of $L_f h \pm f$). ∎

THEOREM 4.4 (equivalence of local minimality). *Suppose that Assumption* 1 *holds. Then,*

(i) *if* $x_\star \in C$ *is a local minimum for* φ, *then it is a nondegenerate fixed point of the proximal-gradient mapping* $\mathbf{T}^{f,g}_{h/\gamma}$ *for any* γ *small enough;*

(ii) *conversely, any nondegenerate fixed point* $x_\star$ *of* $\mathbf{T}^{f,g}_{h/\gamma}$ *is a local minimum for* φ *iff it is a local minimum for* $\varphi^{h/\gamma}$. *Moreover, equivalence of strong local minimality also holds provided that* $h - \gamma f$ *is strongly convex in a neighborhood of* $x_\star$.

Proof. The assertion of Theorem 4.4 directly follows from Theorem 3.7(ii) in light of the equivalence of Theorem 4.1. Suppose now that $x_\star \in C$ is a local minimum for φ. Then, there exists $\bar{\gamma} > 0$ such that $\mathbf{prox}^{h/\gamma'}_{\varphi}(x_\star) = \{x_\star\}$ for all $\gamma' \in (0, \bar{\gamma})$, as shown in Theorem 3.7(i). We now claim that $x_\star$ is a nondegenerate fixed point of

$\mathbf{T}^{f,g}_{h/\gamma}$ for $\gamma \in (0, \frac{\bar\gamma}{1+\bar\gamma L_f})$. To see this, recall that $\mathbf{T}^{f,g}_{h/\gamma} = \mathbf{prox}^{\frac{1}{\gamma}h-f}_{\varphi}$ and observe that $\frac{1}{\gamma}h - f = \frac{1}{\gamma'}h + (L_f h - f)$ for $\gamma' = \frac{\gamma}{1-\gamma L_f}$. By invoking Lemma 3.6 we conclude that $x_\star$ is a nondegenerate fixed point of $\mathbf{T}^{f,g}_{h/\gamma}$ whenever γ is such that $\gamma' \in (0, \bar\gamma)$, that is, for $\gamma \in (0, \frac{\bar\gamma}{1+\bar\gamma L_f})$ as claimed. ∎

Theorem 4.1 also implies through Theorem 3.8 and Proposition 2.5(ii) that the BFBE (3.3) is locally equivalent to its Euclidean version (4.2).

THEOREM 4.5 (local equivalence of Bregman and Euclidean FBE). *Suppose that Assumption 1 holds and let $\gamma < 1/L_f$ be fixed. Suppose further that h is locally Lipschitz differentiable and locally strongly convex on C (as is the case when $h \in \mathcal{C}^2$ with $\nabla^2 h \succ 0$ on C). Then, for every compact set $\mathcal{U} \subset C$ there exist $\bar\gamma > 0$ and a compact convex set $\mathcal{V} \subset C$ such that for all $\tilde\gamma \in (0, \bar\gamma)$ it holds that*

$$\varphi^{f,g}_{h/\gamma} = \varphi^{\tilde f,\tilde g}_{1/\tilde\gamma} \quad \text{and} \quad \mathbf{T}^{f,g}_{h/\gamma} = \mathbf{T}^{\tilde f,\tilde g}_{1/\tilde\gamma} \ \text{ on } \mathcal{U}, \tag{4.7}$$

where

$$\tilde f \coloneqq f - \tfrac{1}{\gamma}h + \tfrac{1}{2\tilde\gamma}\|\cdot\|^2 \quad \text{and} \quad \tilde g \coloneqq g + \tfrac{1}{\gamma}h - \tfrac{1}{2\tilde\gamma}\|\cdot\|^2 + \delta_{\mathcal{V}}\,. \tag{4.8}$$

Moreover, $\tilde g$ is proper, lsc, and prox-bounded (in the Euclidean sense) with $\gamma_{\tilde g} = \infty$, and $\tilde f$ is $L_{\tilde f}$-Lipschitz-differentiable on $\mathcal{V}$ with $\tilde\gamma < 1/L_{\tilde f}$.

We remark that, differently from the case of Theorem 3.8, it is strong convexity of $h - \gamma f$ that is required, while that of h is sufficient but not necessary for the purpose. For instance, local strong convexity of $-f$ would waive the need for a similar requirement on h.

4.1. First- and second-order properties. Here we list some (sub)differential properties of the BFBE and the Bregman forward-backward operator. All the results fall as a direct consequence of similar ones derived in the previous section. We remind the reader that this passage hinges on the key equivalence assessed in Theorem 4.1, namely $\mathbf{T}^{f,g}_{h/\gamma} = \mathbf{prox}^{\hat h}_{\varphi}$ and $\varphi^{f,g}_{h/\gamma} = \varphi^{\hat h}$ for $\hat h = \frac{1}{\gamma}h - f$. In particular, the following result is a direct consequence of Proposition 3.4. For the sake of a lighter notation, it is convenient to introduce the matrix-valued mapping (defined wherever it makes sense)

$$Q^f_{h/\gamma}(x) \coloneqq \tfrac{1}{\gamma}\nabla^2 h(x) - \nabla^2 f(x). \tag{4.9}$$

PROPOSITION 4.6 (subdifferential properties of the BFBE). *Additionally to Assumption 1, suppose that $f, h \in \mathcal{C}^2(C)$. For every $\gamma \in (0, 1/L_f)$ the BFBE $\varphi^{f,g}_{h/\gamma}$ is strictly differentiable wherever it is differentiable. Moreover, for every $x \in C$ and with $Q^f_{h/\gamma}$ as in (4.9)*

(i) $\mathbf{lip}\, \varphi^{f,g}_{h/\gamma}(x) = \mathbf{max}_{\bar x \in \mathbf{T}^{f,g}_{h/\gamma}(x)} \big\|Q^f_{h/\gamma}(x)(x - \bar x)\big\|$;

(ii) $\partial\varphi^{f,g}_{h/\gamma}(x) = \partial_B\varphi^{f,g}_{h/\gamma}(x) \subseteq Q^f_{h/\gamma}(x)\big(x - \mathbf{T}^{f,g}_{h/\gamma}(x)\big)$.

THEOREM 4.7 (continuous differentiability of the BFBE). *Suppose that Assumption 1 holds and that $f, h \in \mathcal{C}^2$ with $\nabla^2 h \succ 0$ on C. Suppose further that g is prox-regular at a nondegenerate fixed point $x_\star$ of $\mathbf{T}^{f,g}_{h/\gamma}$ for $-\nabla f(x_\star)$. Then, there exists a neighborhood $\mathcal{U}$ of $x_\star$ on which the following statements are true:*

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

(ii) *$\varphi^{f,g}_{h/\gamma} \in \mathcal{C}^1(\mathcal{U})$ with $\nabla\varphi^{f,g}_{h/\gamma}(x) = Q^f_{h/\gamma}(x)(x - \mathbf{T}^{f,g}_{h/\gamma}(x))$, where $Q^f_{h/\gamma}$ is as in (4.9).*

Proof. We may invoke [73, Ex. 13.35] to infer that φ is prox-regular at $x_\star$ for $v = 0$. The assumptions of Theorem 3.10 are thus satisfied for the Legendre kernel $\hat{h} = \frac{1}{\gamma}h - f = (L_f h - f) + (\frac{1}{\gamma} - L_f)h$, and the proof then follows from Theorem 3.10 in light of Theorem 4.1. ∎

Twice differentiability of the BFBE will play a key role in the asymptotic analysis of the BELLA algorithm discussed in the next section, when directions of quasi-Newton type are considered (cf. Theorem 5.13). The following result offers sufficient conditions ensuring this property at a fixed point of the Bregman forward-backward mapping $\mathbf{T}^{f,g}_{h/\gamma}$.

THEOREM 4.8 (twice differentiability of $\varphi^{f,g}_{h/\gamma}$). *Additionally to Assumption* 1*, suppose that*

- A1. *$f \in \mathcal{C}^2(C)$ and $\nabla^2 f$ is (strictly) continuous around a nondegenerate fixed point $x_\star$ of $\mathbf{T}^{f,g}_{h/\gamma}$;*
- A2. *$h \in \mathcal{C}^2(C)$ with $\nabla^2 h \succ 0$;*
- A3. *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.*

Then, with $Q^f_{h/\gamma}$ as in (4.9)*, $\mathbf{T}^{f,g}_{h/\gamma}$ is (strictly) differentiable at $x_\star$, and $\varphi^{f,g}_{h/\gamma}$ is (strictly) twice differentiable at $x_\star$ with symmetric Hessian*

$$\nabla^2 \varphi^{f,g}_{h/\gamma}(x_\star) = Q^f_{h/\gamma}(x_\star)\big[\mathrm{I} - J\,\mathbf{T}^{f,g}_{h/\gamma}(x_\star)\big].$$

Proof. It follows from [73, Ex. 13.18 and 13.25] that φ is prox-regular and (strictly) twice epi-differentiable at $x_\star$ for $-\nabla \tilde{f}(x_\star)$, with generalized quadratic second-order epi-derivative. The assumptions of Theorem 3.11 are thus satisfied for $\hat{h} = \frac{1}{\gamma}h - f = (L_f h - f) + (\frac{1}{\gamma} - L_f)h$, and the proof then follows from Theorem 3.11 in light of Theorem 4.1. ∎

5. The BELLA algorithm. Having completed an in-depth analysis of the Bregman forward backward mapping and its envelope, we are now ready to introduce a new algorithm based on these building blocks. The purpose of the algorithm is to globalize the convergence of a fast local method for solving problem (P), exclusively by means of calls to the Bregman forward-backward operator $\mathbf{T}^{f,g}_{h/\gamma}$. Some methods have been proposed that adopt Armijo-type linesearch strategies for nonsmooth problems [84, 25, 66], which, however, operate along "directions of descent" and thus require some directional differentiability properties on the cost function. In response to this limitation, the BELLA algorithm proposed here offers a viable alternative that is suited to problem (P) in its full generality and, in fact, can cope with arbitrary update directions.

Having assessed, under Assumption 1, the continuity property of the BFBE and the inequality $\varphi^{f,g}_{h/\gamma}(\bar{x}) \le \varphi^{f,g}_{h/\gamma}(x) - \frac{1-\gamma L_f}{\gamma}\,\mathbf{D}_h(\bar{x}, x)$ holding for any $\bar{x} \in \mathbf{T}^{f,g}_{h/\gamma}(x)$, the algorithmic rationale is quite self-explanatory. At each iteration, once a direction d has been selected according to a user-defined criterion (ideally a "fast" direction, but virtually any choice works regardless), the candidate update direction $x + d$ is "pushed" toward the forward-backward step $\bar{x}$ until a descent inequality is satisfied on the BFBE, which serves as a continuous and real-valued Lyapunov function for the algorithm. This well definedness aspect will be better detailed in the dedicated subsection 5.1, where a qualitative measure of stationarity of the output point $\hat{x}$ is also given.

The oracle complexity of one iteration of BELLA is dictated by three operations: a Bregman forward-backward call at step 1, the computation of a direction d^k at

Algorithm 5.1 BELLA.

REQUIRE with Assumption 1 holding, select stepsize $\gamma \in (0, 1/L_f)$, initial point $x^0 \in C$, $\sigma \in (0, \frac{1-\gamma L_f}{\gamma})$, tolerance $\varepsilon > 0$, max number of backtrackings $i_{\max} \in \mathbb{N} \cup \{\infty\}$

INITIALIZE $k = 0$

1: choose $\bar{x}^k \in \mathbf{T}^{f,g}_{h/\gamma}(x^k)$
2: **if** $\mathbf{D}_h(\bar{x}^k, x^k) \leq \varepsilon$ **then return** $\hat{x} \coloneqq \bar{x}^k$ **end if**
3: choose a direction $d^k \in \mathbb{R}^n$ and set $\tau_k = 1$ and $i_k = 0$
4: $x^{k+1} = (1-\tau_k)\bar{x}^k + \tau_k(x^k + d^k)$
5: **if** $\varphi^{f,g}_{h/\gamma}(x^{k+1}) \leq \varphi^{f,g}_{h/\gamma}(x^k) - \sigma\,\mathbf{D}_h(\bar{x}^k, x^k)$ **then** ▷ Linesearch passed
6: $k \leftarrow k+1$ and go to step 1
7: **else if** $i_k = i_{\max}$ **then** ▷ Max #backtrackings: do plain BFBS step
8: $x^{k+1} = \bar{x}^k$, $k \leftarrow k+1$ and go to step 1
9: **else** ▷ Linesearch failed: backtrack and retry
10: $\tau_k \leftarrow \tau_k/2$, $i_k \leftarrow i_k + 1$ and go to step 4

step 3, and the evaluation of the BFBE at step 5. As discussed in subsection 5.4, for instance, directions of quasi-Newton type involve only direct linear algebra operations on already available quantities. Moreover, the evaluation of $\varphi^{f,g}_{h/\gamma}(x^{k+1})$ requires only one call to $\mathbf{T}^{f,g}_{h/\gamma}(x^{k+1})$ (which can be stored and reused at step 1), and consequently, apart from the freedom in choosing suitably inexpensive directions, each iteration requires one call to the Bregman forward-backward operator per backtracking trial at step 5. A bound on the number of these calls can be imposed by selecting a finite threshold $i_{\max}$, which, however, makes no difference from a theoretical standpoint. We also remark that BELLA includes known methods as special cases; by setting $d^k = \bar{x}^k - x^k$ it reduces to the Bregman forward-backward algorithm given in [24] (the linesearch condition (5.1) is satisfied regardless of the stepsize τ_k owing to Proposition 4.3(i) and (ii)), while for $h = \frac{1}{2}\|\cdot\|^2$ one obtains the PANOC algorithm given in [75]. It is also worth remarking that by considering $f = 0$ and suitably choosing h, BELLA offers a linesearch extension to any algorithm that can be interpreted as a Bregman proximal point scheme.

5.1. Well definedness and finite termination. As discussed in the beginning of the section, the rationale of the linesearch involved in BELLA is a simple consequence of basic properties of the BFBE. The next result validates this claim.

LEMMA 5.1 (well definedness of BELLA). *Let Assumption 1 hold, and let $\gamma \in (0, 1/L_f)$ and $\sigma \in (0, \frac{1-\gamma L_f}{\gamma})$ be fixed. Then, for any $x \in C$, $\bar{x} \in \mathbf{T}^{f,g}_{h/\gamma}(x) \setminus \{x\}$, and $d \in \mathbb{R}^n$ there exists $\bar{\tau} \in (0, 1]$ such that for any $\tau \in [0, \bar{\tau}]$ the point $x^+_\tau \coloneqq (1-\tau)\bar{x} + \tau(x+d)$ satisfies*

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

In particular, since $x^+_\tau \in C$, the iterates of BELLA *are well defined with linesearch at* step 5 *terminating after a finite number of backtrackings regardless of the choice of $(d^k)_{k\in\mathbb{N}}$ and whether or not a finite maximum number of backtrackings $i_{\max}$ is imposed.*

Proof. It follows from Proposition 4.3(i) and (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 $x' = \bar{x} \in C$. Continuity of the envelope $\varphi^{f,g}_{h/\gamma}$ and openness of its domain as asserted in Proposition 4.2 then ensure 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 $d \in \mathbb{R}^n$, i.e., $x^+_\tau \in \mathcal{U}$ for small enough τ. Therefore, the claimed inequality will be satisfied after a finite number of backtrackings, which is our desired result. ∎

Next, we offer an explicit bound on the number of iterations needed to satisfy the termination criterion at step 2 and provide a qualitative measure of stationarity of the output vector $\hat{x}$. We recall that the number of calls to the Bregman forward-backward operator $\mathbf{T}^{f,g}_{h/\gamma}$ (corresponding to the number of backtrackings at step 9) can artificially be bounded by means of selecting a finite parameter $i_{\max}$ at initialization. By doing so, one obtains that BELLA terminates with at most $i_{\max}$ times the number of iterations many calls to $\mathbf{T}^{f,g}_{h/\gamma}$ (although this is likely a massively loose estimate when sensible directions are employed).

THEOREM 5.2 (iteration complexity of BELLA). *Suppose that Assumption* 1 *holds. Then,*

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

(ii) *if* $\mathbf{dom}\, h = \mathbb{R}^n$ *and* h *is* $\sigma_{h,\mathcal{U}}$*-strongly convex and* $L_{h,\mathcal{U}}$*-Lipschitz differentiable on an open convex set* $\mathcal{U}$ *that contains all the iterates* x^k *and* $\bar{x}^k$ *(this being true if* $h \in \mathcal{C}^2$ *with* $\nabla^2 h \succ 0$*), then the point* $\hat{x}$ *returned by* BELLA *satisfies* $\mathbf{dist}(0, \hat{\partial}\varphi(\hat{x})) \leq \frac{1+\gamma L_f}{\gamma}\sqrt{\frac{2L^2_{h,\mathcal{U}}}{\sigma_{h,\mathcal{U}}}\varepsilon}$.

Proof. We begin by observing that for every $k \in \mathbb{N}$ it holds that

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

The first inequality owes to Proposition 4.3(i) and (ii), whereas the second one is apparent when the condition at step 5 is satisfied and follows from Proposition 4.3(i) and (ii) together with the fact that $\sigma < \frac{1-\gamma L_f}{\gamma}$ otherwise (that is, when the maximum number of backtrackings is reached and the nominal step $x^{k+1} = \bar{x}^k$ is taken).

Theorem 5.2(i). By telescoping the inequality (5.1) over the first $K > 0$ iterations we have

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

where the last inequality follows from Proposition 4.3(i) and (iii). Since all the iterates up to the $(K-1)$th satisfy $\mathbf{D}_h(\bar{x}^k, x^k) > \varepsilon$, if $\varepsilon > 0$ necessarily $K \leq \frac{\varphi(x^0) - \mathbf{inf}\,\varphi}{\sigma\varepsilon}$ as claimed.

Theorem 5.2(ii). Let $\mathcal{M}(z, x) \coloneqq f(x) + \langle \nabla f(x), z - x\rangle + g(z) + \frac{1}{\gamma}\mathbf{D}_h(z, x)$ be the function in the minimization problem (4.1) defining the proximal-gradient operator $\mathbf{T}^{f,g}_{h/\gamma}$. For any $x \in \mathcal{U}$ it holds that the difference $\delta_x(w) \coloneqq \mathcal{M}(w, x) - \varphi(w)$ satisfies

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

By using convexity of $L_f h \pm f$ as in the proof of Proposition 2.5(ii), it is easy to verify that the gradient of the (convex) function $\frac{1}{\gamma}h - f$ is $\frac{1+\gamma L_f}{\gamma} L_{h,\mathcal{U}}$-Lipschitz continuous on $\mathcal{U}$, hence so is $\nabla\delta_x$ independently of x. The proof can now trace that of [80, Lem. 2.15]. Since $\nabla\delta_x(x) = 0$, for any $\bar{x} \in \mathcal{U}$ one has $\|\nabla\delta_x(\bar{x})\| \leq \frac{1+\gamma L_f}{\gamma} L_{h,\mathcal{U}} \|x - \bar{x}\|$. In particular, for $\bar{x} \in \mathbf{T}^{f,g}_{h/\gamma}(x) \cap \mathcal{U}$ one has

$$0 \in \hat{\partial}[\mathcal{M}(\,\cdot\,, x)](\bar{x}) = \hat{\partial}\varphi(\bar{x}) + \nabla\delta_x(\bar{x}),$$

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

$$\begin{aligned}\mathbf{dist}(0, \hat{\partial}\varphi(\bar{x}^k)) &\leq \|\nabla\delta_{x^k}(\bar{x}^k)\| \leq \tfrac{1+\gamma L_f}{\gamma} L_{h,\mathcal{U}} \|x^k - \bar{x}^k\| \\ &\leq \tfrac{1+\gamma L_f}{\gamma} \sqrt{\tfrac{2L^2_{h,\mathcal{U}}}{\sigma_{h,\mathcal{U}}} \mathbf{D}_h(\bar{x}^k, x^k)} \leq \tfrac{1+\gamma L_f}{\gamma} \sqrt{\tfrac{2L^2_{h,\mathcal{U}}}{\sigma_{h,\mathcal{U}}} \varepsilon}\end{aligned}$$

as claimed. ∎

We remark that although the proof of Theorem 5.2(ii) still works even when h does not have full domain, the result is not very informative in the constrained case $\overline{C} \neq \mathbb{R}^n$, owing to the fact that $0 \in \partial\varphi(x_\star)$ is not a necessary condition for optimality when $x_\star$ is on the boundary of $\overline{C}$. That the result holds regardless is not a contradiction, since the Lipschitz constant $L_{h,\mathcal{U}}$ involved in the proof grows bigger and bigger as the boundary is approached.

5.2. Subsequential convergence. The rest of the paper is devoted to asymptotic analyses of BELLA, corresponding to a null tolerance $\varepsilon = 0$. To rule out trivialities, we will implicitly assume $\bar{x}^k \neq x^k$ for every $k \in \mathbb{N}$, so that the algorithm runs infinitely many iterations.

THEOREM 5.3 (asymptotic analysis). *Suppose that Assumption* 1 *holds and consider the iterates generated by* BELLA *with tolerance* $\varepsilon = 0$. *Then,*

(i) $\sum_{k\in\mathbb{N}} \mathbf{D}_h(\bar{x}^k, x^k)$ *is finite;*
(ii) *the real-valued sequences* $(\varphi(\bar{x}^k))_{k\in\mathbb{N}}$ *and* $(\varphi^{f,g}_{h/\gamma}(x^k))_{k\in\mathbb{N}}$ *converge to a finite value* $\varphi_\star$, *the latter monotonically decreasing;*
(iii) *if* $\varphi + \delta_C$ *is level bounded, then* $(x^k)_{k\in\mathbb{N}}$ *and* $(\bar{x}^k)_{k\in\mathbb{N}}$ *are bounded;*
(iv) *if* $(x^k)_{k\in\mathbb{N}}$ *and* $(\bar{x}^k)_{k\in\mathbb{N}}$ *are bounded and either* h *is locally strongly convex or* $\mathbf{dom}\, h = \mathbb{R}^n$, *then* $(x^k)_{k\in\mathbb{N}}$ *and* $(\bar{x}^k)_{k\in\mathbb{N}}$ *have the same set of limit points* ω, *which is compact and such that* $\mathbf{dist}(x^k, \omega) \to 0$ *and* $\mathbf{dist}(\bar{x}^k, \omega) \to 0$ *as* $k \to \infty$.

Proof.

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

Theorem 5.3(ii). It follows from (5.1) that $(\varphi^{f,g}_{h/\gamma}(x^k))_{k\in\mathbb{N}}$ is decreasing, hence it admits a limit, be it $\varphi_\star$, which due to Proposition 4.3(iii) is lower bounded by $\mathbf{inf}_C\, \varphi$ and is thus finite. In turn, also $\varphi(\bar{x}^k) \to \varphi_\star$ (although not necessarily monotonically), as it follows from Proposition 4.3(ii) and the fact that $\mathbf{D}_h(\bar{x}^k, x^k) \to 0$.

Theorem 5.3(iii). Follows from Proposition 4.3(iv) together with the observation that both sequences are contained both in C and in the sublevel set $[\varphi^{f,g}_{h/\gamma} \leq \varphi^{f,g}_{h/\gamma}(x^0)]$, as is apparent from (5.1).

Theorem 5.3(iv). If $\mathbf{dom}\, h = \mathbb{R}^n$, continuity of $\mathbf{D}_h$ on $\mathbb{R}^n \times \mathbb{R}^n$ implies through assertion 5.3(i) that a subsequence $(x^k)_{k\in K}$ converges to a point $x_\star$ iff so does $(\bar{x}^k)_{k\in K}$.

The same holds if h is locally strongly convex, owing to the inequality $\frac{\sigma_{h,\mathcal{U}}}{2}\|x^k-\bar{x}^k\|^2 \leq \mathbf{D}_h(\bar{x}^k, x^k)$ where $\sigma_{h,\mathcal{U}}$ is a strong convexity modulus for h on a (convex) compact set $\mathcal{U} \subseteq \mathbf{dom}\, h$ that contains $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$. In particular, $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$ have same set of limit points, be it ω. The claimed properties on ω hold generally for any limit set of a bounded sequence. ∎

In the rest of the paper we will need to work under the assumption that h has full domain. To the best of our knowledge, no work in the literature dealing with a fully nonconvex setup as that of problem (P) can circumvent this requirement even for proving stationarity of the limit points. For the sake of maintaining the full generality of our problem setup we will not include a separate analysis for a (hypo)convex setting and will thus stick to this assumption, which we formulate next for future reference.

Assumption 2. The Legendre kernel h has full domain, i.e., $\mathbf{dom}\, h = \mathbb{R}^n$.

THEOREM 5.4 (subsequential convergence). *Suppose that Assumptions* 1 *and* 2 *hold and that φ is level bounded. Then, additionally to all the claims of* Theorem 5.3*, any point in the set ω satisfies the fixed-point inclusion $x_\star \in \mathbf{T}^{f,g}_{h/\gamma}(x_\star)$ and in particular is stationary for φ. Moreover, φ is constant on ω and equals $\varphi_\star$ as in* Theorem 5.3(ii).

Proof. It follows from Theorem 5.3(iii) and (iv) that $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$ are bounded and with same set of limit points, be it ω. More precisely, as shown in the proof of Theorem 5.3(iv), for any infinite set $K \subset \mathbb{N}$ it holds that $(x^k)_{k\in K} \to x_\star$ iff $(\bar{x}^k)_{k\in K} \to x_\star$, which implies through Proposition 4.2 that any $x_\star \in \omega$ satisfies $x_\star \in \mathbf{T}^{f,g}_{h/\gamma}(x_\star)$. Let $x_\star \in \omega$ be fixed, and let $K \subseteq \mathbb{N}$ be such that $x^k, \bar{x}^k \to x_\star$ as $K \ni k \to \infty$. We have

$$\varphi(x_\star) \leq \lim_{K\ni k\to\infty} \varphi(\bar{x}^k) \overset{5.3(ii)}{=} \lim_{K\ni k\to\infty} \varphi^{f,g}_{h/\gamma}(x^k) \overset{4.2}{=} \varphi^{f,g}_{h/\gamma}(x_\star) \overset{4.3(i)}{\leq} \varphi(x_\star),$$

where the first inequality owes to the fact that φ is lsc. The arbitrarity of $x_\star \in \omega$ and the fact that $\lim_{k\to\infty} \varphi(\bar{x}^k) = \varphi_\star$ imply that $\varphi(x_\star) = \varphi_\star$ for all $x_\star \in \omega$. Let now $\hat{h} \coloneqq \frac{1}{\gamma}h - f$ so that $\bar{x}^k \in \mathbf{prox}^{\hat{h}}_{\varphi}$; cf. Theorem 4.1. The optimality conditions for $\bar{x}^k$ read

$$0 \in \partial\varphi(\bar{x}^k) + \hat{h}(\bar{x}^k) - \nabla\hat{h}(x^k), \tag{5.3}$$

owing to [73, Ex. 8.8(c)]. Continuity of $\nabla\hat{h}$ implies that $\hat{h}(\bar{x}^k) - \nabla\hat{h}(x^k) \to 0$ as $K \ni k \to \infty$, and since $\varphi(\bar{x}^k) \to \varphi(x_\star)$ from φ-attentive osc of $\partial\varphi$ we conclude that $0 \in \partial\varphi(x_\star)$. ∎

Remark 5.5 (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 by 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 will 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 5.4 remain valid. In order to replicate the proof of Theorem 5.4, 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 ensured only for $\gamma < 1/L_f$ (Proposition 4.3(iii)). 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 $\varphi_\star \ge \inf\varphi$.

5.3. Global and linear convergence. Similarly to the descent algorithms studied in [8], BELLA exhibits global (as opposed to subsequential) convergence when the cost function φ satisfies the so-called KL property, a mild requirement enjoyed by a large class of functions.

DEFINITION 5.6 (KL property). *A proper lsc function $F:\mathbb{R}^n\to\overline{\mathbb{R}}$ is said to have the KL property at $u_\star\in\operatorname{dom}F$ if there exist a continuous 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 $\mathcal{C}^1$ on $(0,\eta)$;*

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

$$\psi'(F(u)-F(u_\star))\,\mathbf{dist}(0,\partial F(u))\ge 1. \tag{5.4}$$

The first inequality of this type is given in the seminal work of Łojasiewicz [53, 54] for analytic functions, which we nowadays call Łojasiewicz's gradient inequality. Kurdyka [44] showed that this inequality is valid for $\mathcal{C}^1$ functions whose graph belongs to an *o-minimal structure* [87, 86], a result which was later extended in [18, 19, 20] for lsc *tame* functions, a wide category including semialgebraic functions as a special case. In the following result we adapt the convergence analysis of [7, 8, 65] to our framework, showing that BELLA converges to a (unique) limit point under a KL property assumption.

THEOREM 5.7 (global convergence). *Let Assumptions* 1 *and* 2 *hold and consider the iterates generated by* BELLA *with tolerance $\varepsilon=0$. Suppose further that*

A1. *φ is level bounded;*

A2. *$f,h\in\mathcal{C}^2$ with $\nabla^2 h\succ 0$;*

A3. *either $(\|d^k\|)_{k\in\mathbb{N}}$ has finite sum, or there exists $D\ge 0$ such that $\|d^k\|\le D\|x^k-\bar{x}^k\|$ for all k;*

A4. *f,g,h are tame functions* [87, 86] *(e.g., semialgebraic).*

Then, $(\|x^k-\bar{x}^k\|)_{k\in\mathbb{N}}$ has finite sum (in fact, regardless of whether or not requirement 5.7.A3 *holds), and $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$ converge to a stationary point $x_\star$.*

Proof. Let $\hat{h}:=\frac{1}{\gamma}h-f$ so that $\bar{x}^k=\mathbf{prox}^{\hat{h}}_{\varphi}(x^k)=\mathbf{arg\,min}_w\,\mathcal{M}(w,x)$, where $\mathcal{M}(w,x):=\varphi(w)+\mathbf{D}_{\hat{h}}(w,x)$ (cf. Theorem 4.1). It follows from Theorem 5.3(ii) that $\varphi^{f,g}_{h/\gamma}(x^k)=\mathcal{M}(\bar{x}^k,x^k)$ converges strictly decreasing to $\varphi_\star$. Since tame functions are closed under derivation and basic algebraic operations [87, sect. 2.1], $\mathcal{M}$ is a tame function and consequently satisfies the KL property [19, Thm. 14]. Moreover, $\mathcal{M}$ is constant (and equals $\varphi_\star$) on $\Omega:=\{(x_\star,x_\star)\mid x_\star\in\omega\}$. The properties of the set of accumulation points ω of the sequences $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$ asserted in Theorem 5.3(iv) ensure through [23, Lem. 6] the existence of a *uniformized KL function* on Ω, namely a function ψ satisfying properties 5.6.P1, 5.6.P2 and 5.6.P3 with $F=\mathcal{M}$ for all $u_\star\in\Omega$ and $u\in\mathbb{R}^n\times\mathbb{R}^n$ such that $\mathbf{dist}(u,\Omega)<\varepsilon$ and $\varphi_\star<\mathcal{M}(u)<\varphi_\star+\eta$. Up to possibly discarding the first iterates, we may assume that $u=(\bar{x}^k,x^k)$ satisfies these conditions for all $k\in\mathbb{N}$. Let $\Delta_k:=\psi(\mathcal{M}(\bar{x}^k,x^k)-\varphi_\star)=\psi(\varphi^{f,g}_{h/\gamma}(x^k)-\varphi_\star)$, and observe that

$$\partial\mathcal{M}(w,x)=\begin{pmatrix}\partial\varphi(w)+\nabla\hat{h}(w)-\nabla\hat{h}(x)\\ \nabla^2\hat{h}(x)(x-w)\end{pmatrix},\ \text{hence}\ \begin{pmatrix}0\\ \nabla^2\hat{h}(x^k)(x^k-\bar{x}^k)\end{pmatrix}\in\partial\mathcal{M}(\bar{x}^k,x^k) \tag{5.5}$$

as follows from (5.3). We have

$$
\begin{aligned}
1 &\le \psi'(\mathcal{M}(\bar{x}^k, x^k) - \varphi_\star)\, \mathbf{dist}(0, \partial \mathcal{M}(\bar{x}^k, x^k)) && \text{KL inequality (5.4)} \\
&\le \psi'(\mathcal{M}(\bar{x}^k, x^k) - \varphi_\star) \|\nabla^2 \hat{h}(x^k)\| \|x^k - \bar{x}^k\| && \text{inclusion (5.5)} \\
&\le \frac{\Delta_k - \Delta_{k+1}}{\mathcal{M}(\bar{x}^k, x^k) - \mathcal{M}(\bar{x}^{k+1}, x^{k+1})} \|\nabla^2 \hat{h}(x^k)\| \|x^k - \bar{x}^k\| && \text{concavity of } \psi \\
&\le \frac{\Delta_k - \Delta_{k+1}}{\sigma\, \mathbf{D}_h(\bar{x}^k, x^k)} \|\nabla^2 \hat{h}(x^k)\| \|x^k - \bar{x}^k\| && \mathcal{M}(\bar{x}^j, x^j) = \varphi^{f,g}_{h/\gamma}(x^j) \text{ and (5.1)} \\
&\le \frac{2L_{\hat{h},\mathcal{U}}}{\sigma \sigma_{\hat{h},\mathcal{U}}} \frac{\Delta_k - \Delta_{k+1}}{\|x^k - \bar{x}^k\|}, && (5.6)\text{–}(5.7)
\end{aligned}
$$

where $L_{\hat{h},\mathcal{U}}$ and $\sigma_{\hat{h},\mathcal{U}}$ are, respectively, smoothness and strong convexity moduli of $\hat{h}$ on a (convex) compact set $\mathcal{U}$ that contains all the iterates x^k and $\bar{x}^k$. Therefore,

$$
\sum_{k\in\mathbb{N}} \|x^k - \bar{x}^k\| \le \frac{2L_{\hat{h},\mathcal{U}}}{\sigma \sigma_{\hat{h},\mathcal{U}}} \sum_{k\in\mathbb{N}} (\Delta_k - \Delta_{k+1}) \le \frac{2L_{\hat{h},\mathcal{U}}}{\sigma \sigma_{\hat{h},\mathcal{U}}} \Delta_0 < \infty, \tag{5.8}
$$

where the last inequality follows from the fact that $\psi \ge 0$. Therefore,

$$
\sum_{k\in\mathbb{N}} \|x^{k+1} - x^k\| = \sum_{k\in\mathbb{N}} \|(1-\tau_k)(\bar{x}^k - x^k) + \tau_k d^k\| \le \sum_{k\in\mathbb{N}} \|\bar{x}^k - x^k\| + \sum_{k\in\mathbb{N}} \|d^k\| \overset{(5.8)}{<} \infty
$$

as ensured by either conditions in requirement 5.7.A3. In particular, $(x^k)_{k\in\mathbb{N}}$ is a Cauchy sequence and thus has a limit $x_\star$, which is also the limit of $(\bar{x}^k)_{k\in\mathbb{N}}$ and is stationary for φ as it follows from Theorem 5.4. ∎

Requirement 5.7.A3 imposes a mild and reasonable consistency criterion on the directions d^k being used (which, however, is not needed for subsequential convergence; cf. Theorem 5.4). It simply reflects the idea that shorter steps should be taken when close to solutions and uses the fixed-point residual $\|x^k - \bar{x}^k\|$ to quantify the proximity. Though apparently quite abstract a requirement, in the next subsection a more practical understanding of this condition will be given when showing its connection to standard theory of quasi-Newton schemes.

When f, g, h are semialgebraic (and thus so is the model $\mathcal{M}$ in the proof of Theorem 5.7 as a byproduct), then the desingularizing function for $\mathcal{M}$ can be taken of the form $\psi(s) = \varrho s^\vartheta$ for some $\varrho > 0$ and $\vartheta \in (0, 1)$ [7], in which case we say that it satisfies the KL property *with exponent* $1 - \vartheta$. Such exponent has a cardinal role in determining asymptotic rates of convergence; in particular, when $\vartheta \ge 1/2$ linear convergence rates can be easily shown. We remark that our approach based on the model $\mathcal{M}$ (both in Theorem 5.7 and in Theorem 5.8 that follows) is largely inspired by the similar approach in [49] for the global and linear convergence of the Douglas–Rachford algorithm in the nonconvex setting.

THEOREM 5.8 (linear convergence). *Let Assumptions* 1 *and* 2 *hold and consider the iterates generated by* BELLA *with tolerance* $\varepsilon = 0$. *Suppose further that*

- A1. φ *is level bounded;*
- A2. $f, h \in \mathcal{C}^2$ *with* $\nabla^2 h \succ 0$*;*
- A3. *either* $(\|d^k\|)_{k\in\mathbb{N}}$ *converges R-linearly to* 0 *or there exists* $D \ge 0$ *such that* $\|d^k\| \le D\|x^k - \bar{x}^k\|$ *for all* k*;*

A4. *f, g, h are semialgebraic, and the KL exponent of $\mathcal{M}(w,x) \coloneqq \varphi(w)+\mathbf{D}_{\hat{h}}(w,x)$ with $\hat{h} \coloneqq \frac{1}{\gamma}h - f$ is $1-\vartheta \le 1/2$.*

Then, $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$ generated by BELLA *converge at R-linear rate to a stationary point.*

Proof. As shown in Theorem 5.7, the sequences converge to a stationary point $x_\star$. Since $x^{k+1}-x^k = (1-\tau_k)(\bar{x}^k - x^k) + \tau_k d^k$, if $\|d^k\| \le D\|x^k - \bar{x}^k\|$ for all k, then defining $B_k \coloneqq \sum_{i\ge k}\|\bar{x}^i - x^i\|$ one has $\|x^k - x_\star\| \le (1+D)B_k$, and similarly $\|\bar{x}^k - x_\star\| \le (3+D)B_k$ owing to the inequality

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

If, instead, there exist $c' > 0$ and $\rho \in (0,1)$ such that $\|d^k\| \le c'\rho^k$ for all $k \in \mathbb{N}$, then $\|x^k - x_\star\| \le B_k + c'\rho^k$ and $\|\bar{x}^k - x_\star\| \le 3B_k + c'\rho^k$. Either way, it suffices to show that the sequence $(B_k)_{k\in\mathbb{N}}$ converges with asymptotic linear rate. Let Δ_k, $L_{\hat{h},\mathcal{U}}$, and $\sigma_{\hat{h},\mathcal{U}}$ be as in the proof of Theorem 5.7. The KL inequality (5.4) with $\psi(s) = \varrho s^\vartheta$ reads

$$1 \le \varrho\vartheta L_{\hat{h},\mathcal{U}}(\mathcal{M}(\bar{x}^k, x^k) - \varphi_\star)^{\vartheta-1}\|x^k - \bar{x}^k\| = \varrho\vartheta L_{\hat{h},\mathcal{U}}(\varphi^{f,g}_{h/\gamma}(x^k) - \varphi_\star)^{\vartheta-1}\|x^k - \bar{x}^k\|.$$

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, $\Delta_k = \psi(\varphi^{f,g}_{h/\gamma}(x^k) - \varphi_\star)$ satisfies

(5.9)
$$\Delta_k = \varrho(\varphi^{f,g}_{h/\gamma}(x^k) - \varphi_\star)^\vartheta \le \varrho(\varrho\vartheta L_{\hat{h},\mathcal{U}})^{\frac{\vartheta}{1-\vartheta}}\|x^k - \bar{x}^k\|^{\frac{\vartheta}{1-\vartheta}} \le \varrho^{\frac{1}{1-\vartheta}}(\vartheta L_{\hat{h},\mathcal{U}})^{\frac{\vartheta}{1-\vartheta}}\|x^k - \bar{x}^k\|,$$

where the last inequality uses the fact that $\frac{\vartheta}{1-\vartheta} \ge 1$ and that $\|x^k - \bar{x}^k\| \le 1$. Hence,

$$B_k = \sum_{i\ge k}\|x^i - \bar{x}^i\| \overset{(5.7)}{\le} \frac{2L_{\hat{h},\mathcal{U}}}{\sigma\sigma_{\hat{h},\mathcal{U}}}\sum_{i\ge k}(\Delta_i - \Delta_{i+1}) \overset{\Delta_i\ge 0}{\le} \frac{2L_{\hat{h},\mathcal{U}}}{\sigma\sigma_{\hat{h},\mathcal{U}}}\Delta_k \overset{(5.9)}{\le} c\|x^k - \bar{x}^k\|$$

for some constant $c > 0$. Therefore, $B_k \le c\|x^k - \bar{x}^k\| = c(B_k - B_{k+1})$, leading to the sought asymptotic linear rate $B_{k+1} \le (1 - 1/c)B_k$. ∎

5.4. Superlinear convergence. Although BELLA is "robust" to any choice of directions, a suitable selection stemming, for instance, from Newton-type methods can cause a remarkable speed-up. As already discussed in Theorem 4.8, the Bregman forward-backward mapping $\mathbf{T}^{f,g}_{h/\gamma}$ behaves nicely around nondegenerate fixed points under some regularity assumptions. This motivates the quest to derive the direction d^k in BELLA by a Newton-type scheme for solving the inclusion $x \in \mathbf{T}^{f,g}_{h/\gamma}(x)$ or, equivalently, the nonlinear (generalized) equation $0 \in \mathbf{R}^{f,g}_{h/\gamma}(x)$ where $\mathbf{R}^{f,g}_{h/\gamma} \coloneqq \mathrm{id} - \mathbf{T}^{f,g}_{h/\gamma}$ is the fixed-point residual mapping. Newton-type methods for such a nonlinear equation prescribe updates of the form

$$x^+ = x - H(x)\,\mathbf{R}^{f,g}_{h/\gamma}(x), \tag{5.10}$$

where ideally $H(x)$ should well approximate the Jacobian of $\mathbf{R}^{f,g}_{h/\gamma}(x)$. 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,

$$H^+ y = s, \quad \text{where } s = x^+ - x,\ y \in \mathbf{R}^{f,g}_{h/\gamma}(x^+) - \mathbf{R}^{f,g}_{h/\gamma}(x). \tag{5.11}$$

In these scenarios, the condition $\|d^k\| \leq D\|x^k - \bar{x}^k\|$ appearing in requirement 5.7.A3 is verified when the operators H are bounded, a frequent assumption in the convergence analysis of quasi-Newton methods. A well-known result characterizing the superlinear convergence of this type of scheme is based on the Dennis–Moré condition [31, 32], 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 [33] for the extension to generalized equations. In Theorem 5.13 we will see that, in fact, directions satisfying this condition 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. We will make use of the notion of *nonisolated superlinear directions*, extending the similar definition of Facchinei and Pang [35, eq. (7.5.2)].

DEFINITION 5.9 (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*

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

with $\mathcal{X}_\star \coloneqq \{x \in \mathbb{R}^n \mid x \in \mathbf{T}^{f,g}_{h/\gamma}(x)\}$ *the set of fixed points of the forward-backward operator* $\mathbf{T}^{f,g}_{h/\gamma}$.

The set $\mathcal{X}_\star$ in the definition above corresponds to the possible limit points of the sequences $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$ generated by BELLA when h has full domain, as shown in Theorem 5.4. In fact, our notion of superlinear directions extends the one given in [35, eq. (7.5.2)] to cases in which $\mathcal{X}_\star$ is not a singleton. Despite the importance of nonisolated critical points in nonsmooth nonconvex optimization, there has been little attention to superlinear directions for such problems. In the convex setting, some studies have shown the potential of variants of regularized Newton [48, 81] and semismooth Newton methods [50, 82] under a local error bound. In the smooth nonconvex setting, there are many works relying on Levenberg–Marquardt [3, 4, 36, 89], cubic regularization [91], and regularized Newton [85] methods under variants of local error bounds and Hölder metric subregularity.

5.4.1. A nonlinear error bound. Our result hinges on three key ingredients: (i) the differentiability of the BFBE around a limit point of BELLA, (ii) the KL property (with exponent) on the BFBE, and (iii) a nonlinear error bound relating the KL function to the distance appearing in Definition 5.9. Sufficient conditions ensuring the first ingredient have already been discussed in Theorem 4.7. As to the second one, it has been shown in [90] that whenever the Legendre kernel h is twice continuously differentiable and (locally) strongly convex, if the function φ satisfies the KL property with exponent $\vartheta \geq 1/2$, then so does the Bregman envelope φ^h. 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 $\vartheta' \in [\vartheta, 1)$. Combined with the relation existing among the BFBE and the Bregman–Moreau envelope shown in Theorem 4.1, the following is obtained.

LEMMA 5.10 (equivalence of Łojasiewicz property [90, Thm. 5.2]). *Additionally to Assumption* 1*, suppose that* $f, h \in \mathcal{C}^2(\mathbf{int}\,\mathbf{dom}\,h)$ *with* $\nabla^2 h \succ 0$*. If* φ *has the KL property with exponent* $\vartheta \in (0, 1)$*, then so does* $\varphi^{f,g}_{h/\gamma}$ *with exponent* $\vartheta' = \mathbf{max}\{1/2, \vartheta\}$.

The last ingredient involves a desingularizing property stronger than the KL inequality which is investigated in [9], namely with $\mathbf{dist}(0, \partial F(u))$ being replaced by the *strong slope* $|\nabla F|(u) \coloneqq \limsup_{u \neq z \to u} \frac{F(u)-F(z)}{\|u-z\|}$ in (5.4). Of particular interest to our scope is [9, Thm. 4.1], which connects this property to a nonlinear growth condition of the form $\psi(F(u) - F(u_\star)) \geq \mathbf{dist}(u, [F \leq F(u_\star)])$. The key observation is that whenever F is continuously differentiable around $u_\star$, both the strong slope $|\nabla F|(u)$ and the minimum norm subgradient $\mathbf{dist}(0, \partial F(u))$ coincide and are equal to $\|\nabla F(u)\|$. The combination of [9, Thm. 4.1] with Theorem 4.7 thus leads to the following milestone of our analysis.

LEMMA 5.11 (nonlinear error bound [9, Thm. 4.1]). *Suppose that Assumption* 1 *holds and let $x_\star$ be a nondegenerate fixed point of $\mathbf{T}^{f,g}_{h/\gamma}$ at which g is prox-regular for $-\nabla f(x_\star)$. Suppose further that $f, h \in \mathcal{C}^2(\mathbf{int}\,\mathbf{dom}\, h)$ with $\nabla^2 h \succ 0$ and that $\varphi^{f,g}_{h/\gamma}$ has the KL property at $x_\star$ with desingularizing function ψ. Then, denoting $\varphi_\star \coloneqq \varphi^{f,g}_{h/\gamma}(x_\star)$ there exist $\varepsilon, \eta > 0$ such that*

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

5.4.2. Superlinear convergence to nonisolated local minima. We now have all the ingredients to address the superlinear convergence analysis of BELLA to nonisolated local minima. The next is a cardinal result of our methodology, as it shows that BELLA does not suffer from the *Maratos effect* [57], a well-known obstacle for fast local methods that inhibits the acceptance of the unit stepsize. On the contrary, we will show that under mild assumptions whenever the directions $(d^k)_{k\in\mathbb{N}}$ in BELLA are superlinear, then unit stepsize is eventually always accepted and the algorithm converges superlinearly even if the limit point belongs to a flat region of local minima. As detailed in the proof, local minimality (as opposed to the more general property of being a fixed point of the forward-backward mapping $\mathbf{T}^{f,g}_{h/\gamma}$) enables the identity $\mathbf{dist}\big(x, [\varphi^{f,g}_{h/\gamma} \leq \varphi_\star]\big) = \mathbf{dist}(x, \mathcal{X}_\star)$ in a neighborhood of the limit point, which in turns allows us to connect the notion of superlinear directions of Definition 5.9 to the nonlinear error bound of Lemma 5.11.

THEOREM 5.12 (acceptance of the unit stepsize and superlinear convergence). *Consider the iterates generated by* BELLA*, and additionally to Assumptions* 1 *and* 2 *suppose that the following requirements hold:*

- A1. *φ is level bounded;*
- A2. *$f, h \in \mathcal{C}^2(\mathbb{R}^n)$ with $\nabla^2 h \succ 0$;*
- A3. *either φ or $\varphi^{f,g}_{h/\gamma}$ has the KL property with exponent $1 - \vartheta \in (0, 1)$ (as is the case when f, g, h are semialgebraic);*
- A4. *d^k are superlinear directions with order $q \geq \mathbf{max}\{1, 1/2\vartheta\}$ (cf. Definition* 5.9*);*
- A5. *the sequence $(x^k)_{k\in\mathbb{N}}$ converges to a nondegenerate fixed point $x_\star$ of $\mathbf{T}^{f,g}_{h/\gamma}$, which is a (not necessarily isolated) local minimum for φ, and at which g is prox-regular for $-\nabla f(x_\star)$.*

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

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

In particular,

(i) *eventually stepsize $\tau = 1$ is always accepted at step* 5 *(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* $\mathcal{X}_\star$ *is as in Definition* 5.9.

Proof. First, in either cases of requirement 5.12.A3, Lemma 5.10 ensures that $\psi(s) := \varrho s^{\mathbf{min}\,\{\vartheta, 1/2\}}$ for some $\varrho > 0$ is a desingularizing function for $\varphi^{f,g}_{h/\gamma}$ at $x_\star$. Denoting $\varphi_\star := \varphi(x_\star) = \varphi^{f,g}_{h/\gamma}(x_\star)$, the equivalence of local minimality asserted in Theorem 4.4 ensures that for small enough $\varepsilon > 0$ it holds that

(5.12)
$$\varphi^{f,g}_{h/\gamma}(x) \geq \varphi_\star \;\text{ and }\; \mathbf{dist}\big(x, [\varphi^{f,g}_{h/\gamma} \leq \varphi_\star]\big) = \mathbf{dist}\big(x, [\varphi^{f,g}_{h/\gamma} = \varphi_\star]\big) \;\; \forall x \in \mathbf{B}(x_\star; \varepsilon).$$

Moreover, since $\varphi^{f,g}_{h/\gamma}(x^k)$ converges strictly decreasing to $\varphi_\star$ (cf. Theorem 5.3(ii)), up to possibly discarding the first iterates and restricting ε we may assume that $\varphi^{f,g}_{h/\gamma}(x^k) \leq \varphi_\star + \eta$, with $\eta, \varepsilon > 0$ as in Definition 5.6 of the KL function. Notice further that any point $x \in \mathbf{B}(x_\star; \varepsilon)$ and such that $\varphi^{f,g}_{h/\gamma}(x) = \varphi_\star$ necessarily satisfies $x \in \mathcal{X}_\star$ owing to Proposition 4.3(i) and local minimality of $x_\star$. Conversely, it follows from Theorem 4.7 that $\varphi^{f,g}_{h/\gamma}$ is $\mathcal{C}^1$ around $x_\star$ with $\nabla\varphi^{f,g}_{h/\gamma}(x) = Q^f_{h/\gamma}(x)(x - \mathbf{T}^{f,g}_{h/\gamma}(x))$, where $Q^f_{h/\gamma} \succ 0$ is as in (4.9), and in particular $\nabla\varphi^{f,g}_{h/\gamma}(x) = 0$ for any $x \in \mathcal{X}_\star$ close to $x_\star$. Combined with the KL inequality (5.4), we conclude that close to $x_\star$ a point x belongs to $\mathcal{X}_\star$ iff $\varphi(x) = \varphi^{f,g}_{h/\gamma}(x) = \varphi_\star$. Thus, up to possibly further restricting ε,

$$\mathbf{dist}\big(x, [\varphi^{f,g}_{h/\gamma} \leq \varphi_\star]\big) \overset{(5.12)}{=} \mathbf{dist}\big(x, [\varphi^{f,g}_{h/\gamma} = \varphi_\star]\big) = \mathbf{dist}\big(x, \mathcal{X}_\star\big) \quad \forall x \in \overline{\mathbf{B}}(x_\star; \varepsilon). \tag{5.13}$$

Combined with the error bound in Lemma 5.11 with $\psi(s) = \varrho s^{\mathbf{min}\,\{\vartheta, 1/2\}}$, we obtain

$$\varphi^{f,g}_{h/\gamma}(x^k) - \varphi_\star \geq \big(\varrho^{-1}\,\mathbf{dist}(x^k, \mathcal{X}_\star)\big)^{\mathbf{max}\,\{2, 1/\vartheta\}} \quad \forall k \in \mathbb{N}. \tag{5.14}$$

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 $x^k_\star \in \mathbf{Proj}_{\mathcal{X}_\star}(x^k + d^k)$, hence such that $\varphi(x^k_\star) = \varphi_\star$ as motivated before. In particular,

$$\varphi^{f,g}_{h/\gamma}(x^k + d^k) \leq \varphi(x^\star_k) + \mathbf{D}_{\hat{h}}(x^\star_k, x^k + d^k) \leq \varphi_\star + \tfrac{L_{\hat{h},\mathcal{U}}}{2}\,\mathbf{dist}(x^k + d^k, \mathcal{X}_\star)^2,$$

where $\hat{h} := \frac{1}{\gamma}h - f \in \mathcal{C}^2(\mathbb{R}^n)$, $L_{\hat{h},\mathcal{U}}$ is a Lipschitz modulus of $\nabla\hat{h}$ on $\mathcal{U} = \overline{\mathbf{B}}(x_\star; \varepsilon)$, and the first inequality follows from Theorem 4.1. Together with (5.14), this implies

$$\varepsilon_k := \frac{\varphi^{f,g}_{h/\gamma}(x^k + d^k) - \varphi_\star}{\varphi^{f,g}_{h/\gamma}(x^k) - \varphi_\star} \leq \frac{L_{\hat{h},\mathcal{U}}}{2}\left(\frac{\mathbf{dist}(x^k + d^k, \mathcal{X}_\star)}{\big(\varrho^{-1/2}\,\mathbf{dist}(x^k, \mathcal{X}_\star)\big)^{\mathbf{max}\,\{1, 1/2\vartheta\}}}\right)^2 \to 0 \;\text{ as } k \to \infty. \tag{5.15}$$

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

$$\begin{aligned}
\varphi^{f,g}_{h/\gamma}(x^k + d^k) - \varphi^{f,g}_{h/\gamma}(x^k) &= \big(\varphi^{f,g}_{h/\gamma}(x^k + d^k) - \varphi_\star\big) - \big(\varphi^{f,g}_{h/\gamma}(x^k) - \varphi_\star\big) \\
&= (\varepsilon_k - 1)\big(\varphi^{f,g}_{h/\gamma}(x^k) - \varphi_\star\big) \\
\text{(since } \varepsilon_k \leq 1,\;\; \varphi(\bar{x}^k) \geq \varphi_\star) \;\; &\leq (\varepsilon_k - 1)\big(\varphi^{f,g}_{h/\gamma}(x^k) - \varphi(\bar{x}^k)\big) \\
\text{(use Proposition 4.3)} \;\; &\leq (\varepsilon_k - 1)\tfrac{1 - \gamma L_f}{\gamma}\,\mathbf{D}_h(\bar{x}^k, x^k) \leq -\,\sigma\,\mathbf{D}_h(\bar{x}^k, x^k)
\end{aligned}$$

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

5.4.3. Superlinear convergence to strong minima under the Dennis–Moré condition. Quasi-Newton methods constitute an important class of directions widely used in optimization. Superlinear convergence of this type of direction 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 5.13 (superlinear convergence under Dennis–Moré condition). *Consider the iterates generated by* BELLA. *Additionally to Assumptions* 1 *and* 2, *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. $f, h \in \mathcal{C}^2(\mathbb{R}^n)$ *with* $\nabla^2 h \succ 0$;

A3. $\mathbf{R}^{f,g}_{h/\gamma}(x) \coloneqq x - \mathbf{T}^{f,g}_{h/\gamma}(x)$ *is strictly differentiable at* $x_\star$ *(see Theorem* 4.8 *for sufficient conditions) with* $J\,\mathbf{R}^{f,g}_{h/\gamma}(x_\star)$ *nonsingular;*

A4. $(d^k)_{k\in\mathbb{N}}$ *satisfy the Dennis–Moré condition*

$$\lim_{k\to\infty} \frac{\mathbf{R}^{f,g}_{h/\gamma}(x^k) + J\,\mathbf{R}^{f,g}_{h/\gamma}(x_\star)d^k}{\|d^k\|} = 0. \tag{5.16}$$

Then, $(d^k)_{k\in\mathbb{N}}$ *are superlinear directions with order* $q = 1$, *and all the claims of Theorem* 5.12 *hold.*

Proof. Since $\mathbf{R}^{f,g}_{h/\gamma}(x_\star) = 0$ as it follows from Theorem 5.4, the Dennis–Moré condition (5.16) and strict differentiability at $x_\star$ imply that

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

Moreover, nonsingularity of $J\,\mathbf{R}^{f,g}_{h/\gamma}(x_\star)$ entails the existence of $\alpha > 0$ 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 $x_\star$. 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. To conclude, observe that $\varphi^{f,g}_{h/\gamma}$ is twice differentiable at $x_\star$ (with Hessian $Q^f_{h/\gamma}(x_\star)J\,\mathbf{R}^{f,g}_{h/\gamma}(x_\star)$ owing to Theorem 4.7 and [74, Prop. 6.2]), hence with $\nabla^2\varphi^{f,g}_{h/\gamma}(x_\star)$ positive definite at its strong local minimum $x_\star$ (Theorem 4.4). Any such function satisfies the KL property at $x_\star$ with exponent $1 - \vartheta = 1/2$ (cf. proof of [83, Lem. B.3]), hence all the assertions of Theorem 5.12 hold. ∎

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 line-search 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 belongs to a flat region of local minima.

In future work we plan to address the following issues: (i) reducing the working assumptions by also accounting for boundary points, (ii) extending existing superlinear direction schemes such as those proposed in [48, 85, 3, 81] for either convex or smooth problems to the more general setting of this paper, (iii) 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 (iv) guaranteeing saddle point avoidance, in the spirit of [67, 46, 52].

Acknowledgments. The authors are grateful to the associate editor and to the three anonymous referees for their helpful comments and suggestions that substantially improved the quality of the paper.

REFERENCES

[1] M. Ahookhosh, *Optimal subgradient methods: Computational properties for large-scale linear inverse problems*, Optim. Eng., 19 (2018), pp. 815–844, https://doi.org/10.1007/s11081-018-9378-5.

[2] M. Ahookhosh, *Accelerated first-order methods for large-scale convex optimization: Nearly optimal complexity under strong convexity*, Math. Methods Oper. Res., 89 (2019), pp. 319–353, https://doi.org/10.1007/s00186-019-00674-w.

[3] M. Ahookhosh, F. J. A. Artacho, R. M. Fleming, and P. T. Vuong, *Local convergence of the Levenberg–Marquardt method under Hölder metric subregularity*, Adv. Comput. Math., 45 (2019), pp. 2771–2806, https://doi.org/10.1007/s10444-019-09708-7.

[4] M. Ahookhosh, R. M. Fleming, and P. T. Vuong, *Finding zeros of Hölder metrically subregular mappings via globally convergent Levenberg–Marquardt methods*, Optim. Methods Softw. (2020), https://doi.org/10.1080/10556788.2020.1712602.

[5] M. Ahookhosh, L. T. K. Hien, N. Gillis, and P. Patrinos, *Multi-block Bregman Proximal Alternating Linearized Minimization and Its Application to Orthogonal Nonnegative Matrix Factorization*, arXiv:1908.01402, 2019.

[6] M. Ahookhosh, L. T. K. Hien, N. Gillis, and P. Patrinos, *A Block Inertial Bregman Proximal Algorithm for Nonsmooth Nonconvex Problems with Application to Nonnegative Matrix Tri-factorization*, arXiv:2003.03963, 2020.

[7] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, *Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality*, Math. Oper. Res., 35 (2010), pp. 438–457, https://doi.org/10.1287/moor.1100.0449.

[8] H. Attouch, J. Bolte, and B. F. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods*, Math. Program., 137 (2013), pp. 91–129, https://doi.org/10.1007/s10107-011-0484-9.

[9] D. Azé and J.-N. Corvellec, *Nonlinear error bounds via a change of function*, J. Optim. Theory Appl., 172 (2017), pp. 9–32, https://doi.org/10.1007/s10957-016-1001-3.

[10] H. H. Bauschke, J. Bolte, J. Chen, M. Teboulle, and X. Wang, *On linear convergence of non-Euclidean gradient methods without strong convexity and Lipschitz gradient continuity*, J. Optim. Theory Appl., 182 (2019), pp. 1068–1087, https://doi.org/10.1007/s10957-019-01516-9.

[11] H. H. Bauschke, J. Bolte, and M. Teboulle, *A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications*, Math. Oper. Res., 42 (2016), pp. 330–348, https://doi.org/10.1287/moor.2016.0817.

[12] H. H. Bauschke, J. M. Borwein, and P. L. Combettes, *Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces*, Commun. Contemp. Math., 03 (2001), pp. 615–647, https://doi.org/10.1142/S0219199701000524.

[13] H. H. Bauschke, P. Combettes, and D. Noll, *Joint minimization with alternating Bregman proximity operators*, Pacific J. Optim., 2 (2005).

[14] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books in Math., Springer, New York, 2017, https://doi.org/10.1007/978-3-319-48311-5.

[15] H. H. Bauschke, M. N. Dao, and S. B. Lindstrom, *Regularizing with Bregman–Moreau envelopes*, SIAM J. Optim., 28 (2018), pp. 3208–3228, https://doi.org/10.1137/17M1130745.

[16] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202, https://doi.org/10.1137/080716542.

[17] S. Bhojanapalli, B. Neyshabur, and N. Srebro, *Global optimality of local search for low rank matrix recovery*, in Advances in Neural Information Processing Systems, 2016, pp. 3873–3881.

[18] J. Bolte, A. Daniilidis, and A. Lewis, *The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems*, SIAM J. Optim., 17 (2007), pp. 1205–1223, https://doi.org/10.1137/050644641.

[19] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota, *Clarke subgradients of stratifiable functions*, SIAM J. Optim., 18 (2007), pp. 556–572, https://doi.org/10.1137/060670080.

[20] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet, *Characterizations of Łojasiewicz inequalities: Subgradient flows, talweg, convexity*, Trans. Amer. Math. Soc., 362 (2010), pp. 3319–3363.

[21] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter, *From error bounds to the complexity of first-order descent methods for convex functions*, Math. Program., 165 (2017), pp. 471–507, https://doi.org/10.1007/s10107-016-1091-6.

[22] J. Bolte and E. Pauwels, *Majorization-minimization procedures and convergence of SQP methods for semi-algebraic and tame programs*, Math. Oper. Res., 41 (2016), pp. 442–465, https://doi.org/10.1287/moor.2015.0735.

[23] J. Bolte, S. Sabach, and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Math. Program., 146 (2014), pp. 459–494, https://doi.org/10.1007/s10107-013-0701-9.

[24] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd, *First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems*, SIAM J. Optim., 28 (2018), pp. 2131–2151, https://doi.org/doi.org/10.1137/17M1138558.

[25] S. Bonettini, I. Loris, F. Porta, and M. Prato, *Variable metric inexact line-search-based methods for nonsmooth optimization*, SIAM J. Optim., 26 (2016), pp. 891–921, https://doi.org/10.1137/15M1019325.

[26] R. I. Boţ and E. R. Csetnek, *An inertial Tseng's type proximal algorithm for nonsmooth and nonconvex optimization problems*, J. Optim. Theory Appl., 171 (2016), pp. 600–616, https://doi.org/10.1007/s10957-015-0730-z.

[27] R. I. Boţ, E. R. Csetnek, and S. C. László, *An inertial forward-backward algorithm for the minimization of the sum of two nonconvex functions*, EURO J. Comput. Optim., 4 (2016), pp. 3–25, https://doi.org/10.1007/s13675-015-0045-8.

[28] R. I. Boţ, E. R. Csetnek, and D.-K. Nguyen, *A proximal minimization algorithm for structured nonconvex and nonsmooth problems*, SIAM J. Optim., 29 (2019), pp. 1300–1328, https://doi.org/10.1137/18M1190689.

[29] L. M. Bregman, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR Comput. Math. Math. Phys., 7 (1967), pp. 200–217, https://doi.org/10.1016/0041-5553(67)90040-7.

[30] D. Davis, D. Drusvyatskiy, and K. J. MacPhee, *Stochastic Model-Based Minimization Under High-Order Growth*, arXiv:1807.00255, 2018.

[31] J. E. J. Dennis and J. J. Moré, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comp., 28 (1974), pp. 549–560, https://doi.org/10.2307/2005926.

[32] J. E. J. Dennis and J. J. Moré, *Quasi-Newton methods, motivation and theory*, SIAM Rev., 19 (1977), pp. 46–89, https://doi.org/10.1137/1019005.

[33] A. Dontchev, *Generalizations of the Dennis-Moré theorem*, SIAM J. Optim., 22 (2012), pp. 821–830, https://doi.org/10.1137/110833567.

[34] R.-A. Dragomir, A. d'Aspremont, and J. Bolte, *Quartic First-Order Methods for Low Rank Minimization*, arXiv:1901.10791, 2019.

[35] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Vol. II, Springer, New York, 2003, https://doi.org/10.1007/b97543.

[36] A. Fischer, *Local behavior of an iterative framework for generalized equations with non-isolated solutions*, Math. Program., 94 (2002), pp. 91–124, https://doi.org/10.1007/s10107-002-0364-4.

[37] M. Fukushima and H. Mine, *A generalized proximal point algorithm for certain non-convex minimization problems*, Internat. J. Systems Sci., 12 (1981), pp. 989–1000, https://doi.org/10.1080/00207728108963798.

[38] M. Fukushima and L. Qi, *A globally and superlinearly convergent algorithm for nonsmooth convex minimization*, SIAM J. Optim., 6 (1996), pp. 1106–1120, https://doi.org/10.1137/S1052623494278839.

[39] F. Hanzely and P. Richtárik, *Fastest Rates for Stochastic Mirror Descent Methods*, arXiv:1803.07374, 2018.

[40] F. Hanzely, P. Richtarik, and L. Xiao, *Accelerated Bregman Proximal Gradient Methods for Relatively Smooth Convex Optimization*, arXiv:1808.03045, 2018.

[41] A. F. Izmailov and M. V. Solodov, *Newton-Type Methods for Optimization and Variational Problems*, Springer, New York, 2014, https://doi.org/10.1007/978-3-319-04247-3.

[42] C. Kan and W. Song, *The Moreau envelope function and proximal mapping in the sense of the Bregman distance*, Nonlinear Anal., 75 (2012), pp. 1385–1399, https://doi.org/10.1016/j.na.2011.07.031.

[43] K. Kawaguchi, *Deep learning without poor local minima*, in Advances in Neural Information Processing Systems, 2016, pp. 586–594.

[44] K. Kurdyka, *On gradients of functions definable in o-minimal structures*, Ann. Inst. Fourier, 48 (1998), pp. 769–783, https://doi.org/10.5802/aif.1638.

[45] E. Laude, P. Ochs, and D. Cremers, *Bregman proximal mappings and Bregman–Moreau envelopes under relative prox-regularity*, J. Optim. Theory Appl., 184 (2020), pp. 724–761, https://doi.org/10.1007/s10957-019-01628-2.

[46] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht, *First-order methods almost always avoid strict saddle points*, Math. Program., 176 (2019), pp. 311–337, https://doi.org/10.1007/s10107-019-01374-3.

[47] C. Lemaréchal and C. Sagastizábal, *Practical aspects of the Moreau–Yosida regularization: Theoretical preliminaries*, SIAM J. Optim., 7 (1997), pp. 367–385, https://doi.org/10.1137/S1052623494267127.

[48] D.-H. Li, M. Fukushima, L. Qi, and N. Yamashita, *Regularized Newton methods for convex minimization problems with singular solutions*, Comput. Optim. Appl., 28 (2004), pp. 131–147, https://doi.org/10.1023/B:COAP.0000026881.96694.32.

[49] G. Li and T. K. Pong, *Douglas-Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems*, Math. Program., 159 (2016), pp. 371–401, https://doi.org/10.1007/s10107-015-0963-5.

[50] X. Li, D. Sun, and K.-C. Toh, *A highly efficient semismooth Newton augmented Lagrangian method for solving lasso problems*, SIAM J. Optim., 28 (2018), pp. 433–458, https://doi.org/10.1137/16M1097572.

[51] T. Liu and T. K. Pong, *Further properties of the forward-backward envelope with applications to difference-of-convex programming*, Comput. Optim. Appl., 67 (2017), pp. 489–520, https://doi.org/10.1007/s10589-017-9900-2.

[52] Y. Liu and W. Yin, *An envelope for Davis–Yin splitting and strict saddle-point avoidance*, J. Optim. Theory Appl., 181 (2019), pp. 567–587, https://doi.org/10.1007/s10957-019-01477-z.

[53] S. Łojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels*, in Les équations aux dérivées partielles, CNRS, Paris, 1963, pp. 87–89.

[54] S. Łojasiewicz, *Sur la géométrie semi- et sous- analytique*, Ann. Inst. Fourier, 43 (1993), pp. 1575–1595, https://doi.org/10.5802/aif.1384.

[55] H. Lu, R. M. Freund, and Y. Nesterov, *Relatively smooth convex optimization by first-order methods, and applications*, SIAM J. Optim., 28 (2018), pp. 333–354, https://doi.org/10.1137/16M1099546.

[56] J. Mairal, *Incremental majorization-minimization optimization with application to large-scale machine learning*, SIAM J. Optim., 25 (2015), pp. 829–855, https://doi.org/10.1137/140957639.

[57] N. Maratos, *Exact Penalty Function Algorithms for Finite Dimensional and Control Optimization Problems*, Ph.D. thesis, Imperial College London, 1978.

[58] R. Mifflin, L. Qi, and D. Sun, *Properties of the Moreau-Yosida regularization of a piecewise C^2 convex function*, Math. Program., 84 (1999), pp. 269–281, https://doi.org/10.1007/s10107980029a.

[59] J.-J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299, https://doi.org/10.24033/bsmf.1625.

[60] M. C. Mukkamala and P. Ochs, *Beyond alternating updates for matrix factorization with inertial Bregman proximal gradient algorithms*, in Advances in Neural Information Processing Systems, 2019, pp. 4268–4278.

[61] M. C. Mukkamala, P. Ochs, T. Pock, and S. Sabach, *Convex-concave backtracking for inertial Bregman proximal gradient algorithms in nonconvex optimization*, SIAM J. Math. Data Sci., 2 (2020), pp. 658–682, https://doi.org/10.1137/19M1298007.

[62] Y. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program., 140 (2013), pp. 125–161, https://doi.org/10.1007/s10107-012-0629-5.

[63] Y. Nesterov, *Implementable tensor methods in unconstrained convex optimization*, Math. Program. (2019), https://doi.org/10.1007/s10107-019-01449-1.

[64] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, New York, 2006, https://doi.org/10.1007/978-0-387-40065-5.

[65] D. Noll, *Convergence of non-smooth descent methods using the Kurdyka–Łojasiewicz inequality*, J. Optim. Theory Appl., 160 (2014), pp. 553–572, https://doi.org/10.1007/s10957-013-0391-8.

[66] P. Ochs, J. Fadili, and T. Brox, *Non-smooth non-convex Bregman minimization: Unification and new algorithms*, J. Optim. Theory Appl., 181 (2019), pp. 244–278, https://doi.org/10.1007/s10957-018-01452-0.

[67] M. O'Neill and S. J. Wright, *Behavior of accelerated gradient methods near critical points of nonconvex functions*, Math. Program., 176 (2019), pp. 403–427, https://doi.org/10.1007/s10107-018-1340-y.

[68] P. Patrinos and A. Bemporad, *Proximal Newton methods for convex composite optimization*, in Proceedings of the 52nd IEEE Conference on Decision and Control, 2013, pp. 2358–2363, https://doi.org/10.1109/CDC.2013.6760233.

[69] R. A. Poliquin and R. T. Rockafellar, *Second-order nonsmooth analysis in nonlinear programming*, in Recent Advances in Nonsmooth Optimization, World Scientific, River Edge, NJ, 1995, pp. 322–350, https://doi.org/10.1142/9789812812827_0018.

[70] R. A. Poliquin and R. T. Rockafellar, *Generalized Hessian properties of regularized nonsmooth functions*, SIAM J. Optim., 6 (1996), pp. 1121–1137, https://doi.org/10.1137/S1052623494279316.

[71] R. A. Poliquin and R. T. Rockafellar, *Prox-regular functions in variational analysis*, Trans. Amer. Math. Soc., 348 (1996), pp. 1805–1838, https://doi.org/10.1090/s0002-9947-96-01544-9.

[72] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970, https://doi.org/10.2307/j.ctt14bs1ff.

[73] R. T. Rockafellar and R. J. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer, New York, 1998, https://doi.org/10.1007/978-3-642-02431-3.

[74] L. Stella, A. Themelis, and P. Patrinos, *Forward-backward quasi-Newton methods for nonsmooth optimization problems*, Comput. Optim. Appl., 67 (2017), pp. 443–487, https://doi.org/10.1007/s10589-017-9912-y.

[75] L. Stella, A. Themelis, P. Sopasakis, and P. Patrinos, *A simple and efficient algorithm for nonlinear model predictive control*, in Proceedings of the 56th Annual Conference on Decision and Control, IEEE, 2017, pp. 1939–1944, https://doi.org/10.1109/CDC.2017.8263933.

[76] J. Sun, Q. Qu, and J. Wright, *A geometric analysis of phase retrieval*, Found. Comput. Math., 18 (2018), pp. 1131–1198, https://doi.org/10.1007/s10208-017-9365-9.

[77] J. Tanner and K. Wei, *Low rank matrix completion by alternating steepest descent methods*, Appl. Comput. Harmon. Anal., 40 (2016), pp. 417–429, https://doi.org/10.1016/j.acha.2015.08.003.

[78] M. Teboulle, *Entropic proximal mappings with applications to nonlinear programming*, Math. Oper. Res., 17 (1992), pp. 670–690, https://doi.org/10.1287/moor.17.3.670.

[79] M. Teboulle, *A simplified view of first order methods for optimization*, Math. Program., 170 (2018), pp. 67–96, https://doi.org/10.1007/s10107-018-1284-2.

[80] A. Themelis, *Proximal Algorithms for Structured Nonconvex Optimization*, Ph.D. thesis, KU Leuven/IMT Lucca, 2018, https://lirias2repo.kuleuven.be/bitstream/id/524341/.

[81] A. Themelis, M. Ahookhosh, and P. Patrinos, *On the acceleration of forward-backward splitting via an inexact Newton method*, in Splitting Algorithms, Modern Operator Theory, and Applications, R. Luke, H. Bauschke, and R. Burachik, eds., Springer, New York, 2019, pp. 363–412, https://doi.org/10.1007/978-3-030-25939-6_15.

[82] A. Themelis and P. Patrinos, *SuperMann: A superlinearly convergent algorithm for finding fixed points of nonexpansive operators*, IEEE Trans. Automat. Control, 64 (2019), https://doi.org/10.1109/TAC.2019.2906393.

[83] A. Themelis, L. Stella, and P. Patrinos, *Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms*, SIAM J. Optim., 28 (2018), pp. 2274–2303, https://doi.org/10.1137/16M1080240.

[84] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Math. Program., 117 (2009), pp. 387–423, https://doi.org/10.1007/s10107-007-0170-0.

[85] K. Ueda and N. Yamashita, *A regularized Newton method without line search for unconstrained optimization*, Comput. Optim. Appl., 59 (2014), pp. 321–351, https://doi.org/10.1007/s10589-014-9656-x.

[86] L. Van den Dries, *Tame Topology and O-minimal Structures*, London Math. Soc. Lect. Note Ser. 248, Cambridge University Press, Cambridge, UK, 1998, https://doi.org/10.1017/CBO9780511525919.

[87] L. Van den Dries and C. Miller, *Geometric categories and o-minimal structures*, Duke Math. J., 84 (1996), pp. 497–540, https://doi.org/10.1215/S0012-7094-96-08416-1.

[88] Q. Van Nguyen, *Forward-backward splitting with Bregman distances*, Vietnam J. Math., 45 (2017), pp. 519–539, https://doi.org/10.1007/s10013-016-0238-3.

[89] N. Yamashita and M. Fukushima, *On the rate of convergence of the Levenberg-Marquardt method*, in Topics in Numerical Analysis, Springer, New York, 2001, pp. 239–249, https://doi.org/10.1007/978-3-7091-6217-0_18.

[90] P. Yu, G. Li, and T. K. Pong, *Deducing Kurdyka-Łojasiewicz Exponent via Inf-Projection*, arXiv:1902.03635, 2019.

[91] M.-C. Yue, Z. Zhou, and A. Man-Cho So, *On the quadratic convergence of the cubic regularization method under a local error bound condition*, SIAM J. Optim., 29 (2019), pp. 904–932, https://doi.org/10.1137/18M1167498.