arXiv:1606.06256v1 [math.OC] 20 Jun 2016

FORWARD-BACKWARD ENVELOPE FOR THE SUM OF TWO NONCONVEX FUNCTIONS: FURTHER PROPERTIES AND NONMONOTONE LINE-SEARCH ALGORITHMS

ANDREAS THEMELIS,[1,2] LORENZO STELLA,[1,2] PANAGIOTIS PATRINOS[1]

Abstract. We consider the problem of minimizing the sum of two nonconvex functions, one of which is smooth and the other prox-bounded and possibly nonsmooth. Such weak requirements on the functions significantly widen the range of applications compared to traditional splitting methods which typically rely on convexity. Our approach is based on the forward-backward envelope (FBE), namely a strictly continuous function whose stationary points are a subset of those of the original cost function. We analyze first- and second-order properties of the FBE under assumptions of prox-regularity of the nonsmooth term in the cost. Although the FBE in the present setting is nonsmooth we propose a globally convergent derivative-free nonmonotone line-search algorithm which relies on exactly the same oracle as the forward-backward splitting method (FBS). Moreover, when the line-search directions satisfy a Dennis-Moré condition, the proposed method converges superlinearly under generalized second-order sufficiency conditions. Our theoretical results are backed up by promising numerical simulations. On large-scale problems, computing line-search directions using limited-memory quasi-Newton updates our algorithm greatly outperforms FBS and, in the convex case, its accelerated variant (FISTA).

1. Introduction

In this paper we deal with optimization problems of the form

$$\underset{x\in\mathbb{R}^n}{\textbf{minimize}}\ \ \varphi(x) \equiv f(x) + g(x), \tag{1.1}$$

where f is smooth (differentiable with Lipschitz continuous gradient) and g is a proper, closed function, possibly nonsmooth and extended-real-valued. Both f and g are allowed to be nonconvex, making (1.1) prototypic for a very wide range of applications. Problems of this form arise in signal and image processing, machine learning, statistics, control and system identification. Moreover, cone programs can be formulated as (1.1), see [1].

A well known algorithm addressing (1.1) is the *forward-backward splitting* (FBS) method, also known as *proximal gradient method*. FBS has been thoroughly analyzed under the assumption of g being convex. If moreover f is convex, then FBS is known to converge globally with rate $O(1/k)$ in terms of objective value, where k is the iteration count. In this case, accelerated variants of FBS can be derived thanks

[1]KU Leuven, Department of Electrical Engineering (ESAT-STADIUS) and Optimization in Engineering Center (OPTEC), Kasteelpark Arenberg 10, 3001 Leuven, Belgium, (panos.patrinos@esat.kuleuven.be).

[2]IMT School for Advanced Studies Lucca, Piazza S. Francesco 19, 55100 Lucca, Italy, ({andreas.themelis, lorenzo.stella}@imtlucca.it).

Date: June 21, 2016.

1991 *Mathematics Subject Classification.* 90C06, 90C25, 90C26, 90C53, 49J52, 49J53.

Key words and phrases. Nonsmooth optimization, nonconvex optimization, forward-backward splitting, line-search methods, quasi-Newton methods, prox-regularity.

to the work of Nesterov [2, 3, 4, 5], that only require minimal additional computations per iteration but achieve the provably optimal global convergence rate of order $O(1/k^2)$.

In the present paper, though, we focus in the most general case where both f and g are allowed to be nonconvex: under the assumption of *prox-boundedness* of g the iterations of FBS are always well defined, and properties of the algorithm this case were studied in [6]. The limit points of FBS are all stationary points of φ, *i.e.*, they satisfy the first order necessary conditions, if the stepsize γ is small enough.

We propose an algorithm for (1.1) which is based on an exact, real-valued penalty function for the problem, the *forward-backward envelope* function (FBE), indicated by φ_γ. The FBE was first introduced in [7] and analyzed under the assumption that both f and g are convex. In [8] the analysis was extended to the case where only g is required to be convex. In these cases the FBE is continuously differentiable provided that $f \in C^2$. This furnishes an equivalence between smooth unconstrained and nonsmooth optimization problems. As a result, well known algorithms from unconstrained optimization can be used to solve (1.1). In particular, [8] proposes a globally convergent line-search based algorithm: when the limit point of the iterates is a strong local minimum of φ then superlinear local convergence is achieved (under assumptions) if the search directions are computed according to quasi-Newton formulas. Very recently, further properties of the FBE were illustrated in [9] and the convergence of a line-search method is analyzed. When g is nonconvex, the FBE is not differentiable and the algorithms discussed in [7, 8, 9] cannot be applied: in this case new methods need to be devised that do not require the *gradient* of the FBE.

1.1. **Contributions.** In this paper:

- We provide the necessary theoretical background linking the concepts of stationarity of a point for problem (1.1), criticality and optimality. To the best of our knowledge, with the sole exception of a special case in [10] this is the first occurrence of the extension of such results, previously shown in the context of the proximal point algorithm [11], to the context of composite objectives and the forward-backward splitting method.
- The analysis of the FBE, previously studied only in the case of g being convex, is extended to the case of g being nonconvex. In particular, under a mild assumption of *prox-regularity* of g at a critical point $x^\star$, we show that T_γ is single-valued and the FBE is continuously differentiable around $x^\star$. Based on this, we provide assumptions on f and g that guarantee that $x^\star$ is a strong local minimum for φ and φ_γ. This is key for obtaining local linear and superlinear convergence of the proposed algorithm. Furthermore, (strict) differentiability of T_γ and its connection to (strict) twice differentiability of φ_γ is proved.
- We introduce Algorithm 1, based on line-search over the FBE, that globally converges provided that φ_γ has the *Kurdyka-Łojasiewicz* property [12, 13, 14] at the cluster points of the sequence of iterates. For related works exploiting this property in proving convergence of optimization algorithms such as FBS we refer the reader to [15, 16, 6, 17, 18]. The algorithm we present here substantially differs from the one proposed in [8], in that it converges under significantly less restrictive assumptions and therefore applies to a wider range of problems. Furthermore, Algorithm 1 is derivative-free, *i.e.*, it doesn't require the computation of $\nabla\varphi_\gamma$ or even its existence. Since computing values of the FBE essentially requires the evaluation of T_γ then our algorithm is based on exactly the same operations as FBS. Finally, the convergence properties of the method are proved under the

assumption that a *nonmonotone* line-search is performed at every iterations, which allows to save computations in many cases.

- We show that the search directions in Algorithm 1 can be selected so as to satisfy an analogous of the Dennis-Moré condition [19], which guarantees that the iterates converge to critical points with superlinear rate. Such condition holds, as we show, when the directions are taken according to the Broyden method for approximating the Jacobian of the fixed-point residual mapping $R_\gamma = \mathrm{Id} - T_\gamma$.

1.2. **Organization of the paper.** In Section 2 we introduce the notation used throughout the paper, and frame the problem rigorously under our assumptions. In Section 3 we explore notions on stationarity and criticality for the investigated problem and relate them with properties of the forward-backward operator. In Section 4 we analyze the fundamental properties of the FBE in the case where g is nonconvex, including first- and second-order differentiability. In Section 5 we describe Algorithm 1, a line-search based method for finding critical points of φ, discuss its global and local linear convergence. Section 6 is devoted to the selection of directions to be implemented in our algorithmic framework, with special attention to quasi-Newton methods for which superlinear convergence of the algorithm is proven. Finally, in Section 7 we illustrate numerical results obtained with the proposed method.

2. Preliminaries

2.1. **Notation.** We denote the extended real line with $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. For a proper, closed function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$, the *subdifferential* $\partial g(x)$ of g at x is considered in the sense of [20, Def. 8.3], that is

$$\partial g(x) = \Big\{ v \in \mathbb{R}^n \mid \exists (x^k)_{k\in\mathbb{N}} \to x, (v^k \in \widehat{\partial} g(x^k))_{k\in\mathbb{N}} \to v \text{ s.t. } g(x^k) \to g(x) \Big\},$$

where $\widehat{\partial} g(x)$ is the set of *regular subgradients* of g at x, namely

$$\widehat{\partial} g(x) = \big\{ v \in \mathbb{R}^n \mid g(z) \geq g(x) + \langle v,\, z - x\rangle + o(\|z - x\|),\ \forall z \in \mathbb{R}^n \big\}.$$

A vector $v \in \partial g(x)$ is a *subgradient* of g at x. Given a parameter value $\gamma > 0$, the *Moreau envelope* function g^γ and the *proximal mapping* $\mathbf{prox}_{\gamma g}$ are defined by

$$g^\gamma(x) := \inf_w \Big\{ g(w) + \tfrac{1}{2\gamma}\|w - x\|^2 \Big\}, \tag{2.1}$$

$$\mathbf{prox}_{\gamma g}(x) := \operatorname*{argmin}_w \Big\{ g(w) + \tfrac{1}{2\gamma}\|w - x\|^2 \Big\}. \tag{2.2}$$

We now summarize some properties of g^γ and $\mathbf{prox}_{\gamma g}$; the interested reader is referred to [20] for a detailed discussion. A function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *prox-bounded* if there exists $\gamma > 0$ such that $g^\gamma(x) > -\infty$ for some $x \in \mathbb{R}^n$. The supremum of the set of all such γ is the *threshold* γ_g *of prox-boundedness* for g. Equivalently, g is prox-bounded if for some $r \in \mathbb{R}$, $g + \frac{r}{2}\|\cdot\|^2$ is bounded below on $\mathbb{R}^n$. If r_g is the infimum of all such r, the proximal threshold of g is $\gamma_g = 1/\max\{0, r_g\}$, where we let $\frac{1}{0} = \infty$ by convention. In particular, if g is convex or bounded below, then it is prox-bounded with $\gamma_g = \infty$. In general, [20, Thm. 1.25]

$$\mathbf{prox}_{\gamma g}(x) \neq \emptyset \quad \text{and} \quad g^\gamma(x) > -\infty \qquad \forall \gamma \in (0, \gamma_g), x \in \mathbb{R}^n. \tag{2.3}$$

Given a nonempty closed set $S \subseteq \mathbb{R}^n$ we let $\delta_S : \mathbb{R}^n \to \overline{\mathbb{R}}$ denote its *indicator function*, namely the proper lower-semicontinuous function such that $\delta_S(x) = 0$ if $x \in S$ and $\delta_S(x) = \infty$ otherwise, and $\mathbf{\Pi}_S : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ the (set-valued) *projection* mapping $x \mapsto \mathbf{argmin}_{z\in S} \|z - x\|$. Proximal mappings can be seen as generalized projections, due to the relation $\mathbf{\Pi}_S = \mathbf{prox}_{\gamma\delta_S}$ which holds for any $\gamma > 0$.

We say that a function $f : \mathbb{R}^n \to \mathbb{R}$ is *strictly differentiable* at x if it is differentiable there and such that $f(y) = f(z) + \langle \nabla f(x),\, y - z\rangle + o(\|y - z\|)$. We denote with $C^{1,1}(\mathbb{R}^n)$ and $C^{1+}(\mathbb{R}^n)$ the set of functions $f \in C^1(\mathbb{R}^n)$ with globally and locally Lipschitz continuous gradient, respectively, and for $f \in C^{1,1}(\mathbb{R}^n)$ we indicate with L_f the Lipschitz modulus of ∇f.

For a set-valued mapping $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ we let $\mathbf{zer}\, T = \{x \in \mathbb{R}^n \mid 0 \in T(x)\}$ denote the set of its *zeros* and $\mathbf{fix}\, T = \{x \in \mathbb{R}^n \mid x \in T(x)\}$ the set of its *fixed-points*. Clearly, $\mathbf{fix}\, T = \mathbf{zer}(\mathrm{Id} - T)$ and $\mathbf{zer}\, T = \mathbf{fix}(\mathrm{Id} - T)$, where Id is the identity.

We indicate with $\mathbf{B}(x; r)$ the open ball of radius r centered in x.

Given a sequence $(x^j)_{j\in\mathbb{N}}$ we let $\omega(x^j) = \omega\big((x^j)_{j\in\mathbb{N}}\big)$ denote the (possibly empty) set of its cluster points.

2.2. **Problem setup and forward-backward iterations.** We proceed under the following assumption.

Assumption I. *In problem* (1.1)

(i) $f \in C^{1,1}(\mathbb{R}^n)$ *(differentiable with* L_f*-Lipschitz continuous gradient);*

(ii) $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ *is proper, closed and* γ_g*-prox-bounded.*

Because of Assumption I, the subdifferential of φ decomposes as $\partial\varphi(x) = \nabla f(x) + \partial g(x)$ and $\widehat{\partial}\varphi(x) = \nabla f(x) + \widehat{\partial} g(x)$ [20, Ex. 8.8(c)].

Given a point $x \in \mathbb{R}^n$, one iteration of the forward-backward splitting algorithm (FBS) for problem (1.1) consists of selecting

$$x^+ \in \mathbf{prox}_{\gamma g}\left(x - \gamma\nabla f(x)\right) = \operatorname*{\mathbf{argmin}}_{w\in\mathbb{R}^n} \left\{\ell_\varphi(w;\, x) + \tfrac{1}{2\gamma}\|w - x\|^2\right\}, \tag{2.4}$$

where

$$\ell_\varphi(w;\, x) \coloneqq f(x) + \langle \nabla f(x),\, w - x\rangle + g(w) \tag{2.5}$$

is the linear approximation of φ with respect to the smooth part f at x. Trivially,

$$\ell_\varphi(x;\, x) = \varphi(x). \tag{2.6}$$

As shorthands we introduce the following set-valued mappings

$$T_\gamma(x) = \mathbf{prox}_{\gamma g}\left(x - \gamma\nabla f(x)\right), \tag{2.7a}$$

$$R_\gamma(x) = x - T_\gamma(x), \tag{2.7b}$$

namely the *forward-backward operator* and its *fixed-point residual*.

Proposition 2.1. *Suppose that* f *and* g *satisfy Assumption I and let* $\gamma \in (0, \gamma_g)$, $x \in \mathbb{R}^n$ *and* $\bar{x} \in T_\gamma(x)$ *be fixed. Let* $L > 0$ *be such that*

$$f(\bar{x}) \le f(x) + \langle \nabla f(x),\, \bar{x} - x\rangle + \tfrac{L}{2}\|x - \bar{x}\|^2. \tag{2.8}$$

Then,

$$\varphi(\bar{x}) \le \varphi(x) - \tfrac{1-\gamma L}{2\gamma}\|x - \bar{x}\|^2 \tag{2.9}$$

and in particular

$$\varphi(\bar{x}) \le \varphi(x) - \tfrac{1-\gamma L_f}{2\gamma}\|x - \bar{x}\|^2 \qquad \forall x \in \mathbb{R}^n, \bar{x} \in T_\gamma(x). \tag{2.10}$$

Proof. Pretty much the same statement is proven in [6, §5]; for the sake of preciseness we here reformulate the proof using our notation.

Plugging the bound (2.8) into the definition (2.5) of ℓ_φ we get

$$\begin{aligned}
\varphi(\bar{x}) &\le \ell_\varphi(\bar{x};\, x) + \tfrac{L}{2}\|\bar{x} - x\|^2 = \ell_\varphi(\bar{x};\, x) + \tfrac{1}{2\gamma}\|\bar{x} - x\|^2 - \tfrac{1-\gamma L}{2\gamma}\|\bar{x} - x\|^2 \\
&\overset{(2.4)}{=} \inf_w \left\{\ell_\varphi(w;\, x) + \tfrac{1}{2\gamma}\|w - x\|^2\right\} - \tfrac{1-\gamma L}{2\gamma}\|\bar{x} - x\|^2 \\
&\le \ell_\varphi(x;\, x) + \tfrac{1}{2\gamma}\|x - x\|^2 - \tfrac{1-\gamma L}{2\gamma}\|\bar{x} - x\|^2 \overset{(2.6)}{=} \varphi(x) - \tfrac{1-\gamma L}{2\gamma}\|\bar{x} - x\|^2
\end{aligned}$$

proving (2.9). Since $L = L_f$ satisfies (2.8) globally, *i.e.*, for any $x, \bar{x} \in \mathbb{R}^n$ [21, Prop. A.24], then (2.10) follows. □

Inequality (2.10) shows that FBS is a descent algorithm if L_f is known and γ is small enough. Inequality (2.9) together with the fact that any $L \geq L_f$ satisfies (2.8) ensures that even when L_f is not known it can be retrieved adaptively, e.g., by performing local backtrackings.

3. Stationary and critical points

In this section the assumption of Lipschitz continuity of ∇f is often superfluous. In these circumstances, for the sake of generality we therefore keep the requirements on f and g separated.

Definition 3.1. *Suppose that $f \in C^1(\mathbb{R}^n)$ and that g satisfies Assumption I(ii). Relative to problem (1.1), we say that a point $x^\star \in \mathbf{dom}\,\varphi$ is*

(i) stationary *if $x^\star \in \mathbf{zer}\,\partial\varphi$;*

(ii) critical *if it is γ-critical for some $\gamma \in (0, \gamma_g)$, i.e., if $x^\star \in \mathbf{fix}\,T_\gamma = \mathbf{zer}\,R_\gamma$;*

(iii) optimal *if $x^\star \in \mathbf{argmin}\,\varphi$, i.e., if it solves (1.1).*

The notion of criticality was already discussed in [10] under the name of L-stationarity (L plays the role of $1/\gamma$) for the special case of $g = \delta_{B \cap C_s}$ where B is a convex set and C_s is the (nonconvex) set of vectors with at most s nonzero entries.

When g is convex it holds that $\gamma_g = \infty$ and we may talk of criticality without mention of γ; in fact, γ-criticality is equivalent to γ'-criticality for any $\gamma, \gamma' > 0$. When g only satisfies Assumption I*(ii)* it then makes sense to define a *criticality threshold* $\Gamma : \mathbb{R}^n \to [0, \gamma_g]$ as follows (the fact that $\Gamma \leq \gamma_g$ is due to empty-valuedness of $\mathbf{prox}_{\gamma g}$ for $\gamma > \gamma_g$).

Definition 3.2. *Relative to problem (1.1) with $f \in C^1(\mathbb{R}^n)$ and g satisfying Assumption I(ii), the* critical threshold *of a point $x \in \mathbb{R}^n$ is*

$$\Gamma(x) := \mathbf{sup}\big(\{\gamma > 0 \mid x \text{ is } \gamma\text{-critical}\} \cup \{0\}\big). \tag{3.1}$$

In particular, $\Gamma(x) > 0$ iff x is a critical point.

Example 3.3. Let us consider $\varphi = f + g$ for $f(x) = \frac{1}{2}x^2$ and $g(x) = \delta_{\{0,1\}}(x)$. Clearly, $\gamma_g = +\infty$ (as g is lower-bounded), $L_f = 1$ and $x^\star = 0$ is the unique optimal point. Since $\partial\varphi(0) = \partial\varphi(1) = \mathbb{R}$ and the subdifferential is clearly empty elsewhere, both $x^\star$ and $x = 1$ are stationary. $\mathbf{prox}_{\gamma g}$ is the (set-valued) projection on $\{0, 1\}$, therefore the forward-backward operator is $T_\gamma(x) = \mathbf{\Pi}_{\{0,1\}}((1-\gamma)x)$. According to Def. 3.1*(ii)*, a point x is γ-critical iff it satisfies $x \in T_\gamma(x)$. As a consequence, the only γ-critical points are the optimal solution $x^\star = 0$, for any $\gamma > 0$, and $x = 1$ for $\gamma \leq 1/2$. Indeed,

$$T_\gamma(0) = \{0\} \quad \text{for all } \gamma > 0, \quad \text{and} \quad T_\gamma(1) = \begin{cases} \{1\} & \text{if } \gamma < 1/2 \\ \{0, 1\} & \text{if } \gamma = 1/2 \\ \{0\} & \text{if } \gamma > 1/2. \end{cases}$$

It follows that $\Gamma(x^\star) = \infty$ and $\Gamma(1) = 1/2$. □

Next, we consider an example with finite prox-boundedness threshold γ_g.

Example 3.4. Let $\varphi = f + g$ for $f(x) = x^2$ and $g(x) = x - \frac{1}{2}x^2$. We have $\gamma_g = 1$, $L_f = 1$, and unique solution $x^\star = -1$. Moreover, $\partial\varphi(x) = \nabla\varphi(x) = x + 1$, which implies that $x^\star$ is the unique stationary point.

The forward-backward operator

$$T_\gamma(x) = \begin{cases} \frac{(1-2\gamma)x - \gamma}{1-\gamma} & \text{if } \gamma < 1 \\ \emptyset & \text{if } \gamma \geq 1 \end{cases}$$

is either single- or empty-valued, depending on whether γ is below or exceeds the threshold γ_g. Solving the equation $T_\gamma(x) = x$ we deduce that the only γ-critical point is $x^\star$ for all $\gamma < 1$. As a consequence, its criticality threshold is $\Gamma(x^\star) = \gamma_g = 1$. □

In both examples above it holds that a critical point x is γ-critical for every γ smaller than its criticality threshold $\Gamma(x)$, and that its forward-backward operator T_γ is single-valued as long as $\gamma < \Gamma(x)$. Moreover, it also happens that any optimal point is γ-critical for any *feasible* γ, *i.e.*, for every γ smaller than the prox-boundedness threshold γ_g, a property which is not ensured for other nonoptimal critical points (e.g. $x = 1$ in Example 3.3).

In the next results we show that these are not mere coincidences. On the one hand this allows to derive necessary conditions for optimality stronger than stationarity; on the other hand, single-valuedness of T_γ for γ small enough is key for first- and second-order properties of T_γ and g^γ, and later also of the FBE defined in Section 4.

Theorem 3.5 (Properties of critical points)**.** *Suppose $f \in C^1(\mathbb{R}^n)$ and that g satisfies Assumption I(ii). Then, relative to problem (1.1), the following properties hold:*

(i) for $\gamma \in (0, \gamma_g)$, a point $x^\star$ is γ-critical iff

$$g(x) \geq g(x^\star) + \langle -\nabla f(x^\star),\, x - x^\star\rangle - \tfrac{1}{2\gamma}\|x - x^\star\|^2 \qquad \forall x \in \mathbb{R}^n;$$

(ii) if $x^\star$ and $y^\star$ are γ-critical, then $\langle \nabla f(y^\star) - \nabla f(x^\star), y^\star - x^\star\rangle \leq \frac{1}{\gamma}\|y^\star - x^\star\|^2$; in particular, if f is μ-strongly convex, then for all $\gamma \geq 1/\mu$ there exists at most one γ-critical point;

(iii) if $x^\star$ is γ-critical, then it is also γ'-critical for any $\gamma' \in (0, \gamma]$;

(iv) if $x^\star$ is critical, then it is γ-critical for all $\gamma \in (0, \Gamma(x^\star))$; moreover, if $\Gamma(x^\star) < \gamma_g$ then $x^\star$ is also $\Gamma(x^\star)$-critical;

(v) $T_\gamma(x^\star) = \{x^\star\}$ and $R_\gamma(x^\star) = \{0\}$ for any critical point $x^\star$ and $\gamma \in (0, \Gamma(x^\star))$.

Proof. For $\gamma \in (0, \gamma_g)$ by (2.3) $\mathbf{prox}_{\gamma g}$ is nonempty-valued and, by definition, $x^\star$ is γ-critical iff $w = x^\star$ minimizes the function in (2.2) for $x = x^\star - \gamma\nabla f(x^\star)$, *i.e.*, iff

$$g(x) + \tfrac{1}{2\gamma}\|x - (x^\star - \gamma\nabla f(x^\star))\|^2 \geq g(x^\star) + \tfrac{\gamma}{2}\|\nabla f(x^\star)\|^2 \qquad \forall x \in \mathbb{R}^n.$$

Expanding the squared norm and rearranging, we obtain 3.5*(i)*.

Suppose now that $x^\star$ and $y^\star$ are γ-critical. Plugging $x = y^\star$ in 3.5*(i)*, then interchanging $x^\star$ and $y^\star$ and summing the obtained inequalities yields the one in 3.5*(ii)*. If, additionally, f is μ-strongly convex then

$$\mu\|y^\star - x^\star\|^2 \leq \langle \nabla f(y^\star) - \nabla f(x^\star),\, y^\star - x^\star\rangle \leq \tfrac{1}{\gamma}\|y^\star - x^\star\|^2$$

where the first inequality is due to [22, Thm. 2.1.9]. Therefore, if $\gamma \geq 1/\mu$ then necessarily $x^\star = y^\star$.

3.5*(iii)* trivially follows from the characterization 3.5*(i)*, together with the fact that γ-criticality entails nonempty-valuedness of $\mathbf{prox}_{\gamma g}$, hence $\gamma \leq \gamma_g$.

We now show 3.5*(iv)*. The first claim trivially follows from 3.5*(iii)* and the definition (3.1) of $\Gamma(x^\star)$. Suppose that $\Gamma(x^\star) < \gamma_g$. Then, from what just proven and from 3.5*(i)* we have

$$g(x) \geq g(x^\star) + \langle -\nabla f(x^\star),\, x - x^\star\rangle - \tfrac{1}{2\gamma}\|x - x^\star\|^2 \quad \forall x \in \mathbb{R}^n, \gamma \in (0, \Gamma(x^\star)).$$

Fixing an arbitrary x and taking the limit as $\gamma \nearrow \Gamma(x^\star)$ we obtain that the inequality holds for $\Gamma(x^\star)$ as well, proving the claim in light of the characterization 3.5*(i)* (which holds due to the fact that $\Gamma(x^\star) < \gamma_g$).

We now show 3.5*(v)*. Let $x^\star$ be a critical point, and contrary to the claim suppose that there exist $\gamma < \Gamma(x^\star)$ and $y \neq x^\star$ such that $y \in T_\gamma(x^\star)$. Let $\gamma' \in (\gamma, \Gamma(x^\star))$. From 3.5*(iv)* we know that $x^\star$ is γ'-critical and from 3.5*(i)* it then follows that

$$\begin{aligned} g(y) &\geq g(x^\star) + \langle -\nabla f(x^\star),\, y - x^\star\rangle - \tfrac{1}{2\gamma'}\|y - x^\star\|^2 \\ &> g(x^\star) + \langle -\nabla f(x^\star),\, y - x^\star\rangle - \tfrac{1}{2\gamma}\|y - x^\star\|^2 \end{aligned} \tag{3.2}$$

where the last inequality is due to the fact that $y \neq x^\star$ and $\gamma' > \gamma$. Since $y, x^\star \in T_\gamma(x^\star)$, the function values in (2.2) coincide, yielding

$$\begin{aligned} g(x^\star) + \tfrac{\gamma}{2}\|\nabla f(x^\star)\|^2 &= g(y) + \tfrac{1}{2\gamma}\|y - x^\star + \gamma\nabla f(x^\star)\|^2 \\ &\overset{(3.2)}{>} g(x^\star) + \langle -\nabla f(x^\star),\, y - x^\star\rangle - \tfrac{1}{2\gamma}\|y - x^\star\|^2 \\ &\quad + \tfrac{1}{2\gamma}\|y - x^\star + \gamma\nabla f(x^\star)\|^2. \end{aligned}$$

Expanding the last square norm we arrive to the contradiction $0 > 0$. □

The inequality in Theorem 3.5*(i)* can be rephrased as the fact that the vector $-\nabla f(x^\star)$ is a "global" *proximal subgradient* for g at $x^\star$ as in [20, Def. 8.45], where "global" refers to the fact that δ can be taken $+\infty$ in the cited definition.

If φ is convex, then stationarity, criticality and optimality are all equivalent properties, whereas if g is convex, then stationary and criticality are equivalent. The second main challenge when dealing with nonconvex problems is that these three properties are well distinguished, as we observed in Example 3.3. However, in the next result we show that criticality is a halfway property between stationarity and optimality, the latter being the strongest. In light of these relations we shall seek "suboptimal" solutions which we characterize as critical points.

Proposition 3.6 (Optimality, criticality, stationarity)**.** *Suppose that $f \in C^1(\mathbb{R}^n)$ and that g satisfies Assumption I(ii). Then,*

(i) (criticality ⇒ stationarity) $\mathbf{fix}\, T_\gamma \subseteq \mathbf{zer}\, \partial\varphi$ *for all* $\gamma \in (0, \gamma_g)$;

Moreover, if f satisfies Assumption I(i), then letting $\bar\gamma = \mathbf{min}\{\gamma_g, 1/L_f\}$ *the following also hold*

(ii) (optimality ⇒ criticality) $\Gamma(x^\star) \geq \bar\gamma$ *for all* $x^\star \in \mathbf{argmin}\,\varphi$*; in particular,* $\mathbf{argmin}\,\varphi \subseteq \mathbf{fix}\, T_\gamma$ *for all* $\gamma \in (0, \bar\gamma)$*, and also for* $\gamma = 1/L_f$ *if* $\gamma_g > 1/L_f$*;*

(iii) $T_\gamma(x^\star) = \{x^\star\}$ *and* $R_\gamma(x^\star) = \{0\}$ *for all* $x^\star \in \mathbf{argmin}\,\varphi$ *and* $\gamma \in (0, \bar\gamma)$.

Proof. Let $\gamma \in (0, \gamma_g)$ and $x \in \mathbf{fix}\, T_\gamma$, then

$$x \in T_\gamma(x) \subseteq \big[\mathrm{Id} + \gamma\partial g\big]^{-1}(x - \gamma\nabla f(x))$$

where the inclusion follows from [20, Ex. 10.2]. Therefore,

$$x - \gamma\nabla f(x) \in \big[\mathrm{Id} + \gamma\partial g\big](x)$$

hence $0 \in \nabla f(x) + \partial g(x) = \partial\varphi$ which proves 3.6*(i)*.

Suppose now that ∇f is L_f-Lipschitz continuous and let $x^\star \in \mathbf{argmin}\,\varphi$. Showing that $\Gamma(x^\star) \geq \bar\gamma$ amounts to proving that $x^\star$ is γ-critical for any $\gamma \in (0, \bar\gamma)$. Fix $\gamma \in (0, \bar\gamma)$ and let $y \in T_\gamma(x^\star)$ (which exists as ensured by (2.3)). To arrive at a contradiction, suppose that $y \neq x^\star$. Then, from (2.10) it follows that

$$\varphi(y) \leq \varphi(x^\star) - \tfrac{1-\gamma L_f}{2\gamma}\|x^\star - y\|^2 < \varphi(x^\star)$$

which contradicts minimality of $\varphi(x^\star)$. Therefore, necessarily $y = x^\star \in T_\gamma(x^\star)$, *i.e.*, $x^\star$ is γ-critical as claimed. 3.6*(ii)* and 3.6*(iii)* then follow from Thm.s 3.5*(iv)* and 3.5*(v)*. □

We conclude the section showing that the property of criticality yields a favorable regularity property on the Moreau envelope.

Proposition 3.7. *Suppose that $f \in C^1(\mathbb{R}^n)$ and that g satisfies Assumption I(ii). Let $x^\star$ be a critical point for problem (1.1). Then, for all $\gamma \in (0, \Gamma(x^\star))$ the Moreau envelope g^γ is strictly differentiable at $x^\star - \gamma\nabla f(x^\star)$ with gradient $\nabla g^\gamma(x^\star - \gamma\nabla f(x^\star)) = -\nabla f(x^\star)$.*

Proof. It follows from [20, Ex. 10.32] that g^γ is strictly continuous with

$$\partial g^\gamma\big(x^\star - \gamma\nabla f(x^\star)\big) \subseteq \tfrac{1}{\gamma}\big[x^\star - \gamma\nabla f(x^\star) - T_\gamma(x^\star)\big] = \tfrac{1}{\gamma}\big[R_\gamma(x^\star) - \gamma\nabla f(x^\star)\big].$$

From Thm. 3.5*(v)* we then have that $\partial g^\gamma\big(x^\star - \gamma\nabla f(x^\star)\big) \subseteq \{-\nabla f(x^\star)\}$ provided that $\gamma \in (0, \Gamma(x^\star))$, and the proof follows from [20, Thm. 9.18(b)]. □

4. Forward-backward envelope

The *forward-backward envelope* (FBE) for problem (1.1) is

$$\varphi_\gamma(x) = \inf_{w\in\mathbb{R}^n} \Big\{\ell_\varphi(w; x) + \tfrac{1}{2\gamma}\|w - x\|^2\Big\}. \tag{4.1}$$

Namely, $\varphi_\gamma(x)$ is the value of the minimization problem in the definition of the forward-backward step (2.4). An alternative expression for the FBE is

$$\varphi_\gamma(x) = f(x) - \tfrac{\gamma}{2}\|\nabla f(x)\|^2 + g^\gamma(x - \gamma\nabla f(x)). \tag{4.2}$$

The FBE was introduced in [7] and further analyzed in [8, 9] in the case when g is convex. Its definition as well as the alternative expression (4.2) remain unchanged; in this section we generalize the results in [8] to our setting where g is not required to be convex. In Section 4.1 we provide bounds between the FBE φ_γ and the original function φ; in Section 4.2 we study first-order properties of φ_γ and see how those are closely related to continuity properties of the forward-backward operator T_γ; finally, in Section 4.3 we study second-order properties of φ_γ at critical points and relate them to first-order properties of T_γ.

We start with some regularity properties of φ_γ. In [8] similar (stronger) properties where already studied in the case of g being convex. In this more general setting, continuous differentiability of the FBE is replaced by strict continuity, a property which entails almost everywhere differentiability. Let

$$Q_\gamma(x) := x - \gamma\nabla^2 f(x) \tag{4.3}$$

which is well-defined (single-valued) and symmetric as long as f is of class C^2 around x. If, additionally, $f \in C^{1,1}(\mathbb{R}^n)$ and $\gamma < 1/L_f$, then it also holds that $Q_\gamma(x) \succ 0$.

Proposition 4.1. *Suppose that g satisfies Assumption I(ii), and $f \in C^1(\mathbb{R}^n)$ (resp. $f \in C^{1+}(\mathbb{R}^n)$). Then,*

(i) $\varphi_\gamma \in C^0(\mathbb{R}^n)$ (resp. $\varphi_\gamma \in C^{0+}(\mathbb{R}^n)$) for any $\gamma \in (0, \gamma_g)$.

Moreover, if $f \in C^2$ in a neighborhood of a point x, then

(ii) $\partial\varphi_\gamma(x) \subseteq \frac{1}{\gamma}Q_\gamma(x)R_\gamma(x)$ for any $\gamma \in (0, \gamma_g)$;

(iii) if x is critical and $\gamma \in (0, \Gamma(x))$, then φ_γ is strictly differentiable at x with $\nabla\varphi_\gamma(x) = 0$.

Proof. Both 4.1*(i)* and 4.1*(ii)* follow from [20, Ex. 10.32] by considering expression (4.2) for φ_γ. Suppose that x is critical and let $\gamma < \Gamma(x)$. From Thm. 3.5*(v)* we know that $R_\gamma(x) = \{0\}$, and from 4.1*(ii)* it then follows that $\partial\varphi_\gamma(x) \subseteq \{0\}$. Invoking [20, Thm. 9.18(b)] we obtain 4.1*(iii)*. □

4.1. **Basic properties.** In this section we provide bounds relating φ_γ to the original function φ. The upper bound in Proposition 4.2*(iii)* is unchanged with respect to the convex case treated in [8]; on the contrary, without convexity of g we cannot provide a stronger lower bound involving the squared forward-backward residual.

Proposition 4.2. *Suppose that g satisfies Assumption I(ii) and $f \in C^1(\mathbb{R}^n)$. Then, for all $\gamma \in (0, \gamma_g)$*

(i) $\varphi_\gamma \leq \varphi$.

Moreover, let L be as in (2.8) with respect to $x \in \mathbb{R}^n$ and $\bar{x} \in T_\gamma(x)$. Then,

(ii) $\varphi_\gamma(\bar{x}) \leq \varphi(\bar{x}) \leq \varphi_\gamma(x) - \frac{1-\gamma L}{2\gamma}\|x - \bar{x}\|^2$.

In particular, if f satisfies Assumption I(i), then for all $x \in \mathbb{R}^n$ and $\bar{x} \in T_\gamma(x)$

(iii) $\varphi_\gamma(\bar{x}) \leq \varphi(\bar{x}) \leq \varphi_\gamma(x) - \frac{1-\gamma L_f}{2\gamma}\|x - \bar{x}\|^2$.

Proof. 4.2*(i)* is obvious from the definition of the FBE (consider $w = x$ in (4.1)).

In the hypothesis of 4.2*(ii)* we have

$$\varphi_\gamma(x) = \underbrace{f(x) + \langle \nabla f(x), \bar{x} - x\rangle}_{\geq f(\bar{x}) - L/2\|\bar{x}-x\|^2} + g(\bar{x}) + \tfrac{1}{2\gamma}\|x - \bar{x}\|^2 \geq \varphi(\bar{x}) + \tfrac{1-\gamma L}{2\gamma}\|x - \bar{x}\|^2.$$

Toghether with 4.2*(i)* this proves 4.2*(ii)*. Similarly, using (2.10) we obtain 4.2*(iii)*. □

An immediate consequence is that the value of φ and φ_γ at critical points is the same, and the sets of minimizers of the two functions coincide for γ small enough, as shown in the next result.

Theorem 4.3. *Suppose that f and g satisfy Assumption I. Then,*

(i) $\varphi(x) = \varphi_\gamma(x)$ *for all* $\gamma \in (0, \gamma_g)$ *and* $x \in \mathbf{fix}\, T_\gamma$;

(ii) $\mathbf{inf}\, \varphi = \mathbf{inf}\, \varphi_\gamma$ *and* $\mathbf{argmin}\, \varphi = \mathbf{argmin}\, \varphi_\gamma \quad \forall \gamma \in \big(0, \bar{\gamma} := \mathbf{min}\,\{1/L_f, \gamma_g\}\big)$.

Proof. 4.3*(i)* follows from Prop. 4.2*(iii)* with $\bar{x} = x$. In particular, because of Prop. 3.6*(ii)* this holds for all $x^\star \in \mathbf{argmin}\, \varphi$ and $\gamma \in (0, \bar{\gamma})$. Therefore,

$$\varphi_\gamma(x^\star) = \varphi(x^\star) \leq \varphi(\bar{x}) \leq \varphi_\gamma(x) \qquad \forall x \in \mathbb{R}^n, \bar{x} \in T_\gamma(x)$$

where the first inequality follows from optimality of $x^\star$, and the second from Prop. 4.2*(iii)*. Therefore, $\mathbf{argmin}\, \varphi \subseteq \mathbf{argmin}\, \varphi_\gamma$ and $\mathbf{min}\, \varphi = \mathbf{min}\, \varphi_\gamma$ if the former is attained. To show the converse inclusion, let $y^\star \in \mathbf{argmin}\, \varphi_\gamma$ and $\bar{y}^\star \in T_\gamma(y^\star)$. From Prop. 4.2*(iii)* we have that

$$\varphi_\gamma(\bar{y}^\star) \leq \varphi(\bar{y}^\star) \leq \varphi_\gamma(y^\star) - \tfrac{1-\gamma L_f}{2}\|y^\star - \bar{y}^\star\|^2,$$

which implies $y^\star = \bar{y}^\star$, since $y^\star$ minimizes φ_γ and $\frac{1-\gamma L_f}{2} > 0$. Therefore, the following chain of inequalities holds

$$\varphi_\gamma(y^\star) = \varphi_\gamma(\bar{y}^\star) \leq \varphi(\bar{y}^\star) \leq \varphi_\gamma(y^\star).$$

Since $\varphi_\gamma \leq \varphi$ and $y^\star$ minimizes φ_γ, it follows that $y^\star \in \mathbf{argmin}\, \varphi$. Therefore, the sets of minimizers of φ and φ_γ coincide, as well as the minimum values provided either one of the two functions attains it.

To conclude, it remains to show the case $\mathbf{argmin}\, \varphi = \emptyset$ which, as just proven, corresponds to $\mathbf{argmin}\, \varphi_\gamma = \emptyset$. From Prop. 4.2*(i)* we have $\mathbf{inf}\, \varphi_\gamma \leq \mathbf{inf}\, \varphi$. Suppose that the inequality is strict, *i.e.*, that there exists $x \in \mathbb{R}^n$ such that $\varphi_\gamma(x) \leq \mathbf{inf}\, \varphi$. Then, Prop. 4.2*(iii)* would imply $\varphi(\bar{x}) \leq \mathbf{inf}\, \varphi$ for any $\bar{x} \in T_\gamma(x)$, which contradicts $\mathbf{argmin}\, \varphi = \emptyset$. This concludes the proof of 4.3*(ii)*. □

The bounds on γ in Theorem 4.3 are tight even in the convex case, as shown in [8, Ex. 2.4]. Combining Proposition 4.1*(ii)* and Theorem 3.5*(v)* with the fact that Q_γ as in (4.3) is positive definite if $\gamma < 1/L_f$, we immediately obtain the following result.

Proposition 4.4. *Suppose that f and g satisfy Assumption I and that $f \in C^2$ around a point x. Then, the following hold*

(i) if x is critical, then for all $\gamma < \Gamma(x)$ it is stationary for φ_γ;

(ii) if x is stationary for φ_γ, then it is γ-critical;

(iii) for $\gamma \in \big(0, \mathbf{min}\,\{1/L_f, \gamma_g\}\big)$, x is γ-critical iff it is stationary for φ_γ.

Proposition 4.2*(ii)* and Thm. 4.3*(ii)* imply that an ε-optimal solution x of φ is automatically ε-optimal for φ_γ, and that from an ε-optimal solution for φ_γ we can directly obtain an ε-optimal solution for φ, if $\gamma < \bar\gamma := \mathbf{min}\,\{1/L_f, \gamma_g\}$:

$$\begin{aligned} \varphi(x) - \mathbf{inf}\,\varphi \le \varepsilon \;&\Rightarrow\; \varphi_\gamma(x) - \mathbf{inf}\,\varphi \le \varepsilon && \forall \gamma \in (0, \gamma_g), \\ \varphi_\gamma(x) - \mathbf{inf}\,\varphi_\gamma \le \varepsilon \;&\Rightarrow\; \varphi(\bar x) - \mathbf{inf}\,\varphi \le \varepsilon && \forall \bar x \in T_\gamma(x), \gamma \in (0, \bar\gamma). \end{aligned}$$

Moreover, Theorem 4.3*(ii)* highlights the first apparent similarity between the concepts of FBE and Moreau envelope: the latter is indeed itself a lower bound for the original function, sharing with it its minimizers and minimum value.

4.2. **Prox-regularity and first-order properties.** In this section we provide sufficient conditions for the differentiability of φ_γ to hold not only *at* a critical point, but also around it and in a continuous fashion. In the favorable case in which g is convex, the FBE enjoys global continuous differentiability, (cf. Corollary 4.7). In our setting, *prox-regularity* acts as a surrogate of convexity, a mild requirement enjoyed globally by all convex functions with $\varepsilon = +\infty$ and $\rho = 0$ in the definition below. The interested reader is referred to [20, §13.F] for a detailed discussion.

Definition 4.5 (Prox-regularity)**.** *A function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ satisfying Assumption I(ii) is said to be* prox-regular *at x_0 for $v_0 \in \partial g(x_0)$ if there exist $\rho, \varepsilon > 0$ such that for all $x' \in \mathbf{B}(x_0; \varepsilon)$ and*

$$(x, v) \in \mathbf{gph}\,\partial g \quad s.t. \quad x \in \mathbf{B}(x_0; \varepsilon),\ v \in \mathbf{B}(v_0; \varepsilon),\ \text{and } g(x) \le g(x_0) + \varepsilon$$

it holds that

$$g(x') \ge g(x) + \langle v,\, x' - x\rangle - \tfrac{\rho}{2}\|x' - x\|^2. \tag{4.4}$$

To help better visualize this definition we consider the local geometrical property that it entails on the function's epigraph [11, Cor. 3.4, Thm. 3.5]. If g is prox-regular at x_0 for v_0 for some constants $\varepsilon, \rho > 0$ as in Definition 4.5, then there exists a neighborhood of $(x_0 + v_0/\rho, g(x_0) - 1/\rho)$ in which the projection on $\mathbf{epi}\, g \cap \big(\mathbf{B}(x_0; \varepsilon) \times \mathbf{B}(v_0; \varepsilon)\big)$ is single-valued.

When g is prox-regular at x_0 for v_0, then for sufficiently small $\gamma > 0$ the Moreau envelope g^γ is continuously differentiable in a neighborhood of $x_0 + \gamma v_0$. To our purposes, this property is needed at γ-critical points x_0 and only with respect to the vector $v_0 = -\nabla f(x_0) \in \partial g(x_0)$ (the inclusion is due to Proposition 3.6*(i)*). Prox-regularity for critical points as it will be assumed in Theorem 4.6 is really a mild requirement, also considering that the fact of being critical itself entails some regularity properties as shown in Proposition 3.7. Examples where this fails to hold are of challenging construction and of unlikely practical utility; before illustrating a cumbersome such instance in Example 4.10, we first prove an important result that connects prox-regularity with first-order properties of the FBE.

Theorem 4.6 (First-order properties under prox-regularity of g)**.** *Suppose that g satisfies Assumption I(ii) and is prox-regular at a critical point $x^\star$ around which*

f is of class C^2 (resp. of class C^{2+}). Then, for all $\gamma \in (0, \Gamma(x^\star))$ there exists a neighborhood $U_{x^\star}$ of $x^\star$ on which the following properties hold:

(i) T_γ and R_γ are Lipschitz continuous, and in particular single-valued;

(ii) $\varphi_\gamma \in C^1$ (resp. $\varphi_\gamma \in C^{1+}$) with $\nabla\varphi_\gamma = \frac{1}{\gamma} Q_\gamma R_\gamma$ where Q_γ is as in (4.3).

Moreover, if $\gamma \in (0, \mathbf{min}\{\Gamma(x^\star), 1/L_f\})$ then

(iii) $x^\star$ is a (strong) local minimum for φ iff it is a (strong) local minimum for φ_γ.

Proof. For $\gamma' \in (\gamma, \Gamma(x^\star))$, using Thm.s 3.5*(i)* and 3.5*(v)* we obtain that

$$g(x) \geq g(x^\star) + \langle -\nabla f(x^\star),\, x - x^\star\rangle - \tfrac{1}{2\gamma'}\|x - x^\star\|^2 \qquad \forall x \in \mathbb{R}^n. \tag{4.5}$$

Replacing γ' with γ in the above expression, the inequality is strict for all $x \neq x^\star$. From [11, Thm. 4.4] applied to the "tilted" function $x \mapsto g(x + x^\star) - g(x^\star) - \langle \nabla f(x^\star),\, x\rangle$ it follows that there is a neighborhood V of $x^\star - \gamma\nabla f(x^\star)$ in which $\mathbf{prox}_{\gamma g}$ is Lipschitz continuous and g^γ is of class C^{1+} with

$$\nabla g^\gamma(x) = \gamma^{-1}\big(x - \mathbf{prox}_{\gamma g}(x)\big) \qquad \forall x \in V.$$

It follows from the assumptions on f that by possibly narrowing $U_{x^\star}$ we may assume that $f \in C^2(U_{x^\star})$ (resp. $f \in C^{2+}(U_{x^\star})$) and $x - \gamma\nabla f(x) \in V$ for all $x \in U_{x^\star}$. 4.6*(ii)* then follows from (4.2) and the chain rule of differentiation, and 4.6*(i)* from the fact that Lipschitz continuity is preserved by composition (since $f \in C^2$ implies $f \in C^{1+}$).

We now show 4.6*(iii)*. From Prop. 4.2*(i)* and Thm. 4.3*(i)* it follows that

$$\varphi(x) - \varphi(x^\star) \geq \varphi_\gamma(x) - \varphi_\gamma(x^\star) \qquad \forall x \in \mathbb{R}^n,$$

and therefore any (strong) local minimum of φ_γ is also a (strong) local minimum of φ. Conversely, suppose that $\gamma < 1/L_f$ and that there exist $\varepsilon \geq 0$ and a neighborhood $U_{x^\star}$ of $x^\star$ such that

$$\varphi(x) - \varphi(x^\star) \geq \tfrac{\varepsilon}{2}\|x - x^\star\| \quad \text{for all } x \in U_{x^\star},$$

where we let $\varepsilon \geq 0$ so as to include the case of nonstrong minimality of $x^\star$. From (Lipschitz-) continuity of T_γ and since $T_\gamma(x^\star) = x^\star$, there exists a neighborhood $V_{x^\star}$ of $x^\star$ such that $T_\gamma(x) \in U_{x^\star}$ for all $x \in V_{x^\star}$. Then, using Thm. 4.3*(i)* and Prop. 4.2*(ii)*, for all $x \in V_{x^\star}$ we have

$$\begin{aligned}\varphi_\gamma(x) - \varphi_\gamma(x^\star) &\geq \varphi(T_\gamma(x)) - \varphi(x^\star) + \tfrac{1-\gamma L_f}{2\gamma}\|x - T_\gamma(x)\|^2 \\ &\geq \tfrac{\varepsilon}{2}\|T_\gamma(x) - x^\star\|^2 + \tfrac{1-\gamma L_f}{2\gamma}\|x - T_\gamma(x)\|^2;\end{aligned}$$

letting $\varepsilon' = \mathbf{min}\left\{\varepsilon, \frac{1-\gamma L_f}{\gamma}\right\}$, which is positive iff ε is,

$$\geq \tfrac{\varepsilon'}{4}\|x - x^\star\|^2$$

where in the last inequality we used the fact that $\|u\|^2 + \|v\|^2 \geq \frac{1}{2}\|u - v\|^2$ for any $u, v \in \mathbb{R}^n$. This concludes the proof of 4.6*(iii)*. □

Since prox-regularity is enjoyed globally by convex functions, the following special case is a straightforward consequence. We refer to [8, Thm. 3.6] for a detailed proof.

Corollary 4.7 (First-order properties for convex g)**.** *Suppose that f and g satisfy Assumption I, with g convex and $f \in C^2(\mathbb{R}^n)$ (resp. $f \in C^{2+}(\mathbb{R}^n)$). Then, for all $\gamma > 0$ all the properties in Theorem 4.6 hold globally (i.e., for all $x^\star \in \mathbb{R}^n$ with $U_{x^\star} = \mathbb{R}^n$).*

Proof. 4.6*(i)* follows from the fact that $\mathbf{prox}_{\gamma g}$ is globally 1-Lipschitz continuous when g is convex. The proof of point 4.6*(ii)* then follows from [8, Thm. 2.6]. Finally, for proving 4.6*(iii)* we may trace the proof of the nonconvex case keeping in mind that Lipschitz continuity of T_γ holds globally. □

To our purposes, when needed, prox-regularity of g will be required only at critical points $x^\star$ and only for the subgradient $-\nabla f(x^\star)$. To streamline the notation we therefore introduce the following slight abuse of terminology.

Definition 4.8 (Prox-regularity of critical points)**.** *Relative to problem* (1.1) *we say that a critical point $x^\star$ is* prox-regular *if g is prox-regular at $x^\star$ for $-\nabla f(x^\star)$.*

Clearly, if g is convex then any critical point is prox-regular, due to the fact that φ is everywhere globally prox-regular. The reason for requiring $\gamma \lneqq \Gamma(x^\star)$ in the first-order properties of Theorem 4.6 is motivated in the following example.

Example 4.9 (Why $\gamma \neq \Gamma(x^\star)$ in first-order properties)**.** Consider f and g as in Example 3.3. We know that $\Gamma(1) = 1/2$, and it is easy to verify that φ is prox-regular at $x = 1$ for any subgradient (cf. Proposition 4.11 with $g_1 = \delta_{\{0\}}$ and $g_2 = \delta_{\{1\}}$). Letting $S = \{0, 1\}$, the FBE is

$$\varphi_\gamma(x) = \tfrac{1-\gamma}{2}\|x\|^2 + \tfrac{1}{2\gamma}\,\mathbf{dist}((1-\gamma)x, S)^2.$$

For any $\gamma \in (0, 1/2)$ it is easy to see that φ_γ is differentiable in a neighborhood of $x = 1$. However, for $\gamma = 1/2$ the distance function has a first-order singularity in $x = 1$, due to the 2-valuedness of $\mathbf{\Pi}_S(1/2)$. □

Example 4.10 (Prox-nonregularity of critical points)**.** Consider $\varphi = f + g$ where

$$f(x) = \tfrac{1}{2}x^2 \quad \text{and} \quad g(x) = \delta_S(x) \quad \text{where} \quad S = \{1/n \mid n \in \mathbb{N}_{\geq 1}\} \cup \{0\}.$$

For $x_0 = 0$ we have $\Gamma(x_0) = +\infty$, however g fails to be prox-regular at x_0 for $v_0 = 0 = -\nabla f(x_0)$. This can be easily seen with the geometrical interpretation previously described, in that for any $\rho > 0$ and for any neighborhood V of $(0, 0)$ in $\mathbf{gph}\, g$ it is always possible to find a point arbitrarily close to $(0, -1/\rho)$ with multi-valued projection on V. Specifically, the midpoint $P_n = \big(\frac{1}{2}(\frac{1}{n} + \frac{1}{n+1}),\ -1/\rho\big)$ has 2-valued projection on $\mathbf{gph}\, g$ for any $n \in \mathbb{N}_{\geq 1}$, being it $\mathbf{\Pi}_{\mathbf{gph}\, g}(P_n) = \big\{\frac{1}{n}, \frac{1}{n+1}\big\}$. By considering large n, P_n can be made arbitrarily close to $(0, -1/\rho)$ and at the same time its projection(s) arbitrarily close to $(0, 0)$.

As a result, $g^\gamma(x) = \frac{1}{2\gamma}\,\mathbf{dist}(x, S)^2$ is not differentiable around $x = 0$, and indeed at each midpoint $\frac{1}{2}(\frac{1}{n} + \frac{1}{n+1})$ for $n \in \mathbb{N}_{\geq 1}$ it has a nonsmooth upward spike.

However, it can be easily seen that all the other critical points, *i.e.*, the ones in $S \setminus \{0\}$, are prox-regular, and with basic calculus it can be verified that g^γ and φ_γ are differentiable at 0 as ensured by Propositions 4.1*(iii)* and 3.7. □

To underline how unfortunate the situation depicted in Example 4.10 is, notice that adding a linear term λx to f for any $\lambda \neq 0$, yet leaving g unchanged, restores the sought prox-regularity of each critical point. Indeed, this is trivially true for any nonzero critical point; besides, g is prox-regular at 0 for any $\lambda \in (0, -\infty)$, and for $\lambda < 0$ we have that 0 is nomore critical for any $\gamma > 0$. The reason why prox-regularity fails to hold in the above example is due to the density of isolated points close to 0.

Proposition 4.11 (Prox-regularity of *convex-cluster-domained* functions)**.** *Let $g = \mathbf{min}_{i\in I}\, g_i$ for a family of pairwise disjoint-domained proper, closed, lower semicontinuous and convex functions $\{g_i \mid i \in I\}$. Suppose that*

$$\mathbf{mesh}(g) := \mathbf{inf}\,\{\mathbf{dist}(\mathbf{dom}\, g_i, \mathbf{dom}\, g_j) \mid i, j \in I,\ i \neq j\} > 0,$$

then g is everywhere prox-regular for all subgradients, with constants $r = 0$ and $\varepsilon = \frac{1}{2}\,\mathbf{mesh}(g)$ as in Definition 4.5.

Proof. Let $(x_0, v_0) \in \mathbf{gph}\, \partial g$, then there exists a unique $i \in I$ such that $(x_0, v_0) \in \mathbf{gph}\, \partial g_i$. By convexity we have

$$g_i(x') \geq g_i(x) + \langle v,\, x' - x\rangle \qquad \forall x, x' \in \mathbb{R},\ v \in \partial g_i(x).$$

The inequality above still holds if we replace g_i with g and constrain x and x in the ball $\mathbf{B}(x_0; \frac{1}{2}\,\mathbf{mesh}(g))$, which implies (4.4) with $\rho = 0$ and $\varepsilon = \frac{1}{2}\,\mathbf{mesh}(g)$. □

A practical employment of g as in Proposition 4.11 concerns problems where the optimization variable is constrained in the disjoint union of convex sets with nonvanishing mesh. Possibly the most appropriate example regards binary problems where $g = \delta_{\{0,1\}}$ or integer programming, in which case $g = \delta_{\mathbb{Z}}$, where no matter what $f \in C^{1,1}(\mathbb{R}^n)$ is, all critical points of $\varphi = f + g$ are prox-regular.

4.3. **Second-order properties.** Similarly to the role of prox-regularity as a surrogate of convexity to ensure first-order properties, when it comes to second-order theory we need generalized notions of second-order differentiability of g. We list now some properties that will be assumed for g and refer the interested reader to [20, §13] for an extensive discussion.

Assumption II. *With respect to a given critical point $x^\star$ for problem (1.1),*

(i) f is of class C^2 around $x^\star$;

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

$$\mathrm{d}^2 g(x^\star|-\nabla f(x^\star))[d] = \langle d,\, Md\rangle + \delta_S(d), \quad \forall d \in \mathbb{R}^n \tag{4.6}$$

where $S \subseteq \mathbb{R}^n$ is a linear subspace and $M \in \mathbb{R}^{n\times n}$ is symmetric and satisfies $\mathbf{Im}(M) \subseteq S$ and $\mathbf{ker}(M) \supseteq S^\perp$.

The properties of M in Assumption II*(ii)* cause no loss of generality. Indeed, letting Π_S denote the orthogonal projector onto the linear space S (which is symmetric [23]), if M satisfies (4.6), then so does $M' = \Pi_S[\frac{1}{2}(M + M^\intercal)]\Pi_S$, which has the wanted properties.

Twice epi-differentiability of g is a mild requirement, and cases where $\mathrm{d}^2 g$ is actually generalized quadratic are abundant [24, 25, 26, 27]. In some results the following additional properties on f and g are needed.

Assumption III. *With respect to a given critical point $x^\star$ for problem (1.1),*

(i) f is of class C^{2+} around $x^\star$;

(ii) g is strictly twice epi-differentiable at $x^\star$ for $-\nabla f(x^\star)$.

We begin by investigating differentiability of the residual mapping R_γ.

Lemma 4.12. *Suppose that f and g satisfy Assumption I, and that Assumption II (together with III) is satisfied with respect to a critical point $x^\star$. Then, for any $\gamma \in (0, \Gamma(x^\star))$*

(i) $\mathbf{prox}_{\gamma g}$ is (strictly) differentiable at $x^\star - \gamma\nabla f(x^\star)$ with Jacobian

$$P_\gamma(x^\star) \coloneqq J\,\mathbf{prox}_{\gamma g}(x^\star - \gamma\nabla f(x^\star)) = \Pi_S[I + \gamma M]^{-1}\Pi_S; \tag{4.7}$$

(ii) R_γ is (strictly) differentiable at $x^\star$ with Jacobian

$$JR_\gamma(x^\star) = \mathrm{Id} - P_\gamma(x^\star)Q_\gamma(x^\star) \tag{4.8}$$

where Q_γ is as in (4.3);

(iii) $P_\gamma(x^\star)$ is symmetric, and $P_\gamma(x^\star) \succeq 0$.

Proof. Once we show that $[I+\gamma M]^{-1}$ is well defined, the proof is exactly as in [8, Lem. 2.9], since the assumption of convexity of g can be replaced by prox-boundedness of g and prox-regularity of the critical point $x^\star$ (cf. the references in the cited proof).[1] For $\gamma' \in (\gamma, \Gamma(x^\star))$, from (4.5) we obtain that for all $w \in \mathbb{R}^n$

$$\mathrm{d}^2 g(x^\star|-\nabla f(x^\star))[w] = \liminf_{\substack{w'\to w \\ \tau\to 0^+}} \frac{g(x^\star+\tau w') - g(x^\star) + \tau\langle \nabla f(x^\star), w\rangle}{\tau^2/2} \geq -\tfrac{1}{\gamma'}\|w\|^2.$$

The expression (4.6) of the second-order epi-derivative then implies

$$\langle Mw,\, w\rangle \geq -1/\gamma' \|w\|^2 \qquad \forall w \in \mathbb{R}^n$$

since for $w \in S^\perp$ we have $Mv = 0$. Therefore (recall that M is symmetric), the minimum eigenvalue of M satisfies $\lambda_{\min}(M) \geq -1/\gamma' > -1/\gamma$ proving $I+\gamma M$ to be positive definite. As anticipated, we may now invoke the proof of [8, Lem. 2.9]. □

Next, we see that differentiability of the residual R_γ at critical points is equivalent to that of $\nabla\varphi_\gamma$ provided that $f \in C^2$ around the point. Mild additional assumptions on f extend this kinship to strict differentiability. Moreover, we extend the equivalence of strong (local) minimality already discussed in Theorem 4.6*(iii)* to first-order conditions of the residual and second-order conditions of the FBE.

Theorem 4.13. *Suppose that f and g satisfy Assumption I, and that Assumption II (together with III) is satisfied with respect to a critical point $x^\star$. Then, for any $\gamma \in (0, \Gamma(x^\star))$ φ_γ is (strictly) twice differentiable at $x^\star$ with symmetric Hessian*

$$\nabla^2\varphi_\gamma(x^\star) = \tfrac{1}{\gamma} Q_\gamma(x^\star)(\mathrm{Id} - P_\gamma(x^\star)Q_\gamma(x^\star)) = \tfrac{1}{\gamma} Q_\gamma(x^\star) J R_\gamma(x^\star) \tag{4.9}$$

where Q_γ and P_γ are as in Lemma 4.12.

Proof. Recall from Thm. 4.6*(ii)* that $\nabla\varphi_\gamma(x) = \frac{1}{\gamma}Q_\gamma(x)R_\gamma(x)$ is well defined around $x^\star$. The result follows from Lem. 4.12 together with the comments in its proof, and invoking [8, Thm. 2.10]. □

Theorem 4.14. *Suppose that f and g satisfy Assumption I, and that Assumption II is satisfied with respect to a critical point $x^\star$. Then, for all $\gamma \in (0, \mathbf{min}\,\{\Gamma(x^\star), 1/L_f\})$ the following are equivalent:*

(a) $x^\star$ is a strong local minimum for φ;
(b) $x^\star$ is a local minimum for φ and $JR_\gamma(x^\star)$ is nonsingular;
(c) $\langle d,\, (\nabla^2 f(x^\star) + M)d\rangle > 0$ for all $d \in S$;
(d) $JR_\gamma(x^\star)$ is similar to a symmetric and positive definite matrix;
(e) $\nabla^2\varphi_\gamma(x^\star) \succ 0$;
(f) $x^\star$ is a strong local minimum for φ_γ;
(g) $x^\star$ is a local minimum for φ_γ and $JR_\gamma(x^\star)$ is nonsingular.

Proof. The equivalence of 4.14*(a)*, 4.14*(c)*, 4.14*(d)*, 4.14*(e)*, and 4.14*(f)* follows from Lem. 4.12 and Thm. 4.13 applying the proof of [8, Thm. 2.11].

4.14*(e)*⇔4.14*(g)* follows from the expression (4.9) of $\nabla^2\varphi_\gamma(x^\star)$.

4.14*(g)*⇔4.14*(b)* trivially follows from Thm. 4.6*(iii)*.

4.14*(b)*⇒4.14*(e)* to arrive at a contradiction, suppose that $\nabla^2\varphi_\gamma(x^\star) \not\succ 0$. Then, since $JR_\gamma(x^\star)$ is nonsingular, from (4.9) it follows that $\nabla^2\varphi_\gamma(x^\star)$ is indefinite, which contradicts local minimality of $x^\star$ for φ_γ, which is due to Thm. 4.6*(iii)*. □

Algorithm 1 Generalized forward-backward with nonmonotone line-search

Require	$L_0 > 0$	$\gamma_0 \in (0, \mathbf{min}\{1/L_0, \gamma_g\})$	$\alpha, \beta, p_{\min} \in (0,1)$
	$x^0 \in \mathbb{R}^n$	$\sigma_0 \in (0, \frac{1-\gamma_0 L_0}{2\gamma_0})$	$(p_k)_{k\in\mathbb{N}} \subset [p_{\min}, 1]$
Initialize	$k = 0$	$\bar{\Phi}_0 = \varphi_{\gamma_0}(x^0)$	

1: Select $\bar{x}^k \in T_{\gamma_k}(x^k)$ and set $r^k = x^k - \bar{x}^k$
2: **if** $\|r^k\| = 0$, **then** stop; **end if**
3: **if** $f(\bar{x}^k) > f(x^k) - \langle \nabla f(x^k), r^k\rangle + \frac{L_k}{2}\|r^k\|^2$ **then**
$\gamma_k \leftarrow \alpha\gamma_k$, $L_k \leftarrow L_k/\alpha$, $\sigma_k \leftarrow \sigma_k/\alpha$, $\bar{\Phi}_k \leftarrow \varphi_{\gamma_k}(x^k)$, go to step 1.
end if
4: Select a direction $d^k \in \mathbb{R}^n$
5: Let $\tau_k = \beta^{m_k}$ with $m_k \in \mathbb{N}$ being the smallest such that
(5.1) $\varphi_{\gamma_k}(\bar{x}^k + \tau_k d^k) \le \bar{\Phi}_k - \sigma_k\|r^k\|^2$
6: $x^{k+1} = \bar{x}^k + \tau_k d^k$
7: $\bar{\Phi}_{k+1} = (1-p_k)\bar{\Phi}_k + p_k\varphi_{\gamma_k}(x^{k+1})$
$\gamma_{k+1} = \gamma_k$, $L_{k+1} = L_k$, $\sigma_{k+1} = \sigma_k$, $k \leftarrow k+1$ and go to step 1.

5. The algorithm

We now address the issue of devising an iterative method that converges to critical points. As shown in Proposition 3.6 and Example 3.3, though weaker than optimality, criticality is a stronger property than stationarity. The problem of finding γ-critical points boils down to

$$\text{find } x^\star \in \mathbf{zer}\, R_\gamma = \mathbf{fix}\, T_\gamma. \tag{5.2}$$

We propose Algorithm 1, a line-search method for solving (5.2) using φ_γ as merit function so as to ensure global convergence. This will be possible thanks to the real-valuedness of φ_γ on the whole space $\mathbb{R}^n$, to Theorem 4.3*(i)* which ensures that the original function φ agrees with φ_γ on the solutions to (5.2), and to Proposition 4.2*(ii)* which ensures that a *sufficient decrease* can always be enforced when far from the solutions to (5.2).

5.1. **Connections with other methods based on the FBE.** The first algorithmic framework exploiting the FBE was studied in [7], the same work in which the FBE was first introduced. There, for convex f and g with $f \in C^{2,1}(\mathbb{R}^n)$ (twice continuously differentiable with Lipschitz continuous gradient) two semismooth Newton methods were analyzed.

A similar approach was then studied in [8], where however major relaxations of assumptions were made. Convexity of the smooth function f is here not required, and the employment of quasi-Newton methods in place of semismooth Newton is considered. This results in major computational simplifications, as no linear system needs be solved nor (generalized) first-order information of $\mathbf{prox}_{\gamma g}$ are required. The proposed algorithm interleaves descent steps over the FBE with forward-backward steps.

Pretty much at the same time, in [9] classical line-search methods applied to the problem of minimizing φ_γ are studied. Global convergence is proved under the additional assumption of f being analytic and g being subanalytic and lower bounded.

Though apparently closely related and all comprising the classical forward-backward scheme as a special case, the approach that we provide in this paper presents

[1]Clearly, the extra property $\|P_\gamma(x^\star)\| \le 1$ shown in the cited proof does not apply, but with the same reasoning it can be inferred that $\|P_\gamma(x^\star)\| \le (1 - \gamma/\Gamma(x^\star))^{-1}$ in our case.

major conceptual differences from any of the ones above. Apart from the significantly less restrictive assumptions, the crucial distinction is that our method is *derivative-free*, *i.e.*, it does not require computing the gradient of the FBE nor even its existence. As a consequence, no computation nor the existence of $\nabla^2 f$ is required, resulting in a method that, differently from the others, truly relies on the very same oracle information of the forward-backward operator T_γ. Furthermore, in our setting the update directions d^k can be *arbitrary*, as no concept of "descent" is required. Directions apart, from the abstract algorithmic standpoint the difference from the schemes presented in [7, 8] lies in the order in which update direction and forward-backward steps are operated: roughly speaking, in the cited papers $x^+ = T_\gamma(x + \tau d)$ while in Algorithm 1 $x^+ \in T_\gamma(x) + \tau d$ (here, the inclusion sign is necessary due to possible nonconvexity of g).

5.2. **Well-definedness.** In this subsection we prove the well-definedness of Algorithm 1, namely that γ_0 is decreased a finite number of times and that the line-search (5.1) always ends in finitely many loops.

Lemma 5.1 (Backtracking on L_k)**.** *Suppose that f and g satisfy Assumption I, and consider the iterates generated by Algorithm 1. Then,*

$$\varphi_{\gamma_k}(\bar{x}^k) \le \varphi(\bar{x}^k) \le \varphi_{\gamma_k}(x^k) - \tfrac{1-\gamma_k L_k}{2\gamma_k}\|x^k - \bar{x}^k\|^2 \quad \forall k. \tag{5.3}$$

Moreover, there exist $L_\infty \le \max\{L_0, {}^{L_f}\!/_{\alpha}\}$, $\gamma_\infty \ge \frac{L_0}{L_\infty}\gamma_0$, $\sigma_\infty \le \frac{L_\infty}{L_0}\sigma_0$, and an iteration $k_0 \in \mathbb{N}$ such that

$$\gamma_k = \gamma_\infty,\ L_k = L_\infty,\ \sigma_k = \sigma_\infty \qquad \forall k \ge k_0.$$

In particular, the number of times i that condition at step 3 holds is finite and satisfies $i < 1 + \log_\alpha(\max\{1, {}^{L_0}\!/_{L_f}\})$. Finally, the ratio $\frac{\sigma_k}{(1-\gamma_k L_k)/2\gamma_k}$ is constant, and in particular $\sigma_k < \frac{1-\gamma_k L_k}{2\gamma_k}$ holds for all k.

Proof. (5.3) follows from Prop. 4.2*(ii)* since after step 3 is passed $L = L_k$ satisfies (2.8) with respect to $x = x^k$ and $\bar{x} = \bar{x}^k$.

From [21, Prop. A.24] we know that for any $L \ge L_f$ the condition at step 3 fails, hence $L_k = L_0$ if $L_0 \ge L_f$, or $L_k < {}^{L_f}\!/_{\alpha}$ otherwise. Passing i times the test at step 3 amounts to decrementing γ_0 and incrementing L_0 and σ_0 i times, and the provided upper bound for i follows solving the inequality ${}^{L_0}\!/_{\alpha^i} < {}^{L_f}\!/_{\alpha}$. Plugging this bound in $\alpha^i\gamma_0$ and ${}^{\sigma_0}\!/_{\alpha^i}$ proves the bounds on γ_k and σ_k as well.

Fix now an iteration k and suppose that the condition at step 3 has been passed j times. Letting $\rho = \frac{\sigma_0}{(1-\gamma_0 L_0)/2\gamma_0} \in (0,1)$ we have $\sigma_k = \frac{\sigma_0}{\alpha^j} = \rho\frac{1-\gamma_0 L_0}{2\gamma_0\alpha^j} = \rho\frac{1-\gamma_k L_k}{2\gamma_k}$ proving the last claim on σ_k. □

Remark 5.2 (Cost per iteration)**.** Evaluating φ_{γ_k} essentially amounts to one evaluation of T_{γ_k}, cf. (2.4) and (4.1). Therefore, computing $\varphi_{\gamma_k}(\bar{x}^k + \tau_k d^k)$ in step 5 of Algorithm 1 yields an element of T_{γ_k} required in step 1, since $x^{k+1} = \bar{x}^k + \tau_k d^k$ at every iteration. As a consequence, in the best case of $\tau_k = 1$ being accepted, the algorithm only requires *one* evaluation of T_{γ_k} per iteration. In general, one evaluation of T_{γ_k} per backtracking step is required.

Remark 5.3 (Non adaptive case: when L_f is known)**.** When the Lipschitz modulus L_f of ∇f is known, then by selecting $L_0 \ge L_f$ the condition at step 3 is never satisfied and L_k, γ_k and σ_k are constant. Algorithm 1 then reduces to Algorithm 2 below. However, though the scheme below is a special case, Lemma 5.1 ensures that for $\sigma = \sigma_\infty$, $\gamma = \gamma_\infty$ and by possibly enlarging L_f, eventually the former scheme does reduce to the latter. For the sake of generality, in the rest of the paper we analyze Algorithm 1.

Algorithm 2 Generalized forward-backward with nonmonotone line-search and nonadaptive γ

Require	$x^0 \in \mathbb{R}^n$	$\gamma \in (0, \mathbf{min}\{1/L_f, \gamma_g\})$	$\alpha, \beta, p_{\min} \in (0,1)$
		$\sigma \in (0, \frac{1-\gamma L_f}{2\gamma})$	$(p_k)_{k\in\mathbb{N}} \subset [p_{\min}, 1]$
Initialize	$k = 0$	$\bar{\Phi}_0 = \varphi_\gamma(x^0)$	

1: Select $\bar{x}^k \in T_\gamma(x^k)$ and set $r^k = x^k - \bar{x}^k$
2: **if** $\|r^k\| = 0$, **then** stop; **end if**
3: Select a direction $d^k \in \mathbb{R}^n$
4: Let $\tau_k = \beta^{m_k}$ with $m_k \in \mathbb{N}$ being the smallest such that
$$\varphi_\gamma(\bar{x}^k + \tau_k d^k) \le \bar{\Phi}_k - \sigma\|r^k\|^2$$
5: $x^{k+1} = \bar{x}^k + \tau_k d^k$
6: $\bar{\Phi}_{k+1} = (1-p_k)\bar{\Phi}_k + p_k\varphi_\gamma(x^{k+1})$
$k \leftarrow k+1$ and go to step 1.

So far we showed that the backtrackings on γ_k at step 3 are feasible and are performed only finitely many times. In the next result we show that a similar claim holds also for τ_k at (5.1).

Lemma 5.4 (Nonmonotone backtracking on τ_k)**.** *Suppose that f and g satisfy Assumption I, and consider the iterates generated by Algorithm 1. Then, for each iteration k it holds that*

$$\varphi_{\gamma_k}(\bar{x}^k) \le \varphi(\bar{x}^k) \le \varphi_{\gamma_k}(x^k) \le \bar{\Phi}_k \tag{5.4}$$

and there exists $\bar{\tau}_k > 0$ such that

$$\varphi_{\gamma_k}(\bar{x}^k + \tau d^k) \le \varphi_{\gamma_k}(x^k) - \sigma_k\|x^k - \bar{x}^k\|^2 \qquad \forall \tau \in [0, \bar{\tau}_k].$$

In particular, the number of backtrackings m_k at (5.1) is finite.

Proof. The first two inequalities in (5.4) are due to (5.3); we now proceed by induction to prove the third one. For $k = 0$ we have that $\bar{\Phi}_0 = \varphi_{\gamma_0}(x^0)$, since so is how $\bar{\Phi}_0$ is initialized and so $\bar{\Phi}_0$ is set in case γ_0 is decreased at step 3. Suppose now that (5.4) holds for $k \ge 0$ and consider iteration $k+1$. If the test at step 3 is passed, then $\bar{\Phi}_{k+1} = \varphi_{\gamma_{k+1}}(x^{k+1})$. Otherwise, $\gamma_{k+1} = \gamma_k$ and

$$\begin{aligned}
\bar{\Phi}_{k+1} &= (1-p_k)\bar{\Phi}_k + p_k\varphi_{\gamma_{k+1}}(x^{k+1}) \\
(\gamma_{k+1} = \gamma_k \text{ \& induction}) \quad &\ge (1-p_k)\varphi_{\gamma_k}(x^k) + p_k\varphi_{\gamma_k}(x^{k+1}) \\
&\overset{(5.1)}{\ge} (1-p_k)\varphi_{\gamma_k}(x^{k+1}) + p_k\varphi_{\gamma_k}(x^{k+1}) \\
(\gamma_{k+1} = \gamma_k) \quad &= \varphi_{\gamma_{k+1}}(x^{k+1})
\end{aligned}$$

which concludes the proof of (5.4).

As to the following inequality, let k be fixed and contrary to the claim suppose that for all $\varepsilon > 0$ there exists $\tau_\varepsilon \in [0, \varepsilon]$ such that the point $x_\varepsilon = \bar{x}^k + \tau_\varepsilon d^k$ satisfies

$$\varphi_{\gamma_k}(x_\varepsilon) > \varphi_{\gamma_k}(x^k) - \sigma_k\|x^k - \bar{x}^k\|^2.$$

Taking the limit for $\varepsilon \to 0^+$, continuity of φ_γ as ensured by Prop. 4.1*(i)* yields

$$\varphi_{\gamma_k}(\bar{x}^k) \ge \varphi_{\gamma_k}(x^k) - \sigma_k\|x^k - \bar{x}^k\|^2 > \varphi_{\gamma_k}(x^k) - \tfrac{1-\gamma_k L_k}{2\gamma_k}\|x^k - \bar{x}^k\|^2$$

where the last inequality is due to Lem. 5.1 and to the fact that $x^k \ne \bar{x}^k$. Since L_k satisfies (2.8) for x^k and $\bar{x}^k \in T_{\gamma_k}(x^k)$ (having failed the test at step 3), this contradicts Prop. 4.2*(ii)*, and the claim follows.

Combining this with (5.4) we obtain that $\varphi_{\gamma_k}(\bar{x}^k + \tau d^k) \le \bar{\Phi}_k - \sigma_k\|x^k - \bar{x}^k\|^2$ for all $\tau \in [0, \bar{\tau}_k]$ as well, which concludes the proof. □

Remark 5.5 (Extension of FBS)**.** Observe that selecting null directions $d^k \equiv 0$ in Algorithm 1, the method reduces to the classical FBS algorithm. To see this, for simplicity we shall omit the backtracking on L_f and analyze the nonadaptive variant of Algorithm 2. Proposition 4.2*(iii)* combined with the relation $\phi_\gamma(x^k) \leq \bar{\Phi}_k$ due to (5.4) shows that the condition at step 4 is always statisfied (with $\tau_k = 1$). Therefore, $x^{k+1} = \bar{x}^k + d^k = \bar{x}^k \in T_\gamma(x^k)$ for all k, which is FBS, cf. (2.4).

5.3. **Subsequential convergence.** In this section we analyze the properties of cluster points of the iterates generated by Algorithm 1. Specifically, we show that every cluster point of $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$ is critical. We stick to the notation of Algorithm 1 so that $r^k = x^k - \bar{x}^k$ is an element of $R_\gamma(x^k)$, and we assume that $r^k \neq 0$ for all k so to exclude the trivial case in which the optimality condition $r^k = 0$ is achieved in a finite number of iterations.

Theorem 5.6 (Criticality of cluster points)**.** *Suppose that f and g satisfy Assumption I. Let $\gamma = \gamma_\infty$, $\sigma = \sigma_\infty$ and k_0 be as in Lemma 5.1. The following hold for the iterates generated by Algorithm 1:*

(i) $\bar{\Phi}_{k+1} \leq \bar{\Phi}_k - \sigma p_{\min}\|r^k\|^2$ *for all* $k \geq k_0$*;*

(ii) either $(\|r^k\|)_{k\in\mathbb{N}}$ *is square-summable, or* $\varphi(\bar{x}^k) \to \inf\varphi = -\infty$ *in which case* $(x^k)_{k\in\mathbb{N}}$ *is unbounded and* $\omega(x^j) = \emptyset$*;*

(iii) $\omega(x^j) = \omega(\bar{x}^j) \subseteq \operatorname{fix} T_\gamma \subseteq \operatorname{zer}\partial\varphi$ *and* $\varphi \equiv \varphi_\gamma$ *on* $\omega(x^j)$*;*

(iv) the (real-valued) sequences $(\bar{\Phi}_k)_{k\geq k_0}$ *and* $(\varphi_\gamma(x^k))_{k\geq k_0}$ *converge to the same value* $\varphi_\star$*, the former strictly decreasing;*

(v) $\limsup_{k\to\infty} \varphi_\gamma(\bar{x}^k) \leq \limsup_{k\to\infty} \varphi(\bar{x}^k) \leq \varphi_\star$*, the limits being attained if either* $\varphi_\star = \inf\varphi = -\infty$ *or* $(\bar{x}^k)_{k\in\mathbb{N}}$ *is bounded;*

(vi) boundedness of $(x^k)_{k\in\mathbb{N}}$ *is equivalent to that of* $(\bar{x}^k)_{k\in\mathbb{N}}$*.*

Proof. The well-definedness of the algorithm has already been discussed in §5.2. Let $k \geq k_0$, then

$$\bar{\Phi}_{k+1} = (1-p_k)\bar{\Phi}_k + p_k\varphi_\gamma(x^{k+1}) \overset{(5.1)}{\leq} \bar{\Phi}_k - \sigma p_k\|r^k\|^2 \leq \bar{\Phi}_k - \sigma p_{\min}\|r^k\|^2$$

which is 5.6*(i)*. Telescoping the inequality we get

$$\bar{\Phi}_{k+1} - \bar{\Phi}_{k_0} = \sum_{i=k_0}^{k} \left[\bar{\Phi}_{i+1} - \bar{\Phi}_i\right] \leq -\sigma p_{\min} \sum_{i=k_0}^{k} \|r^i\|^2. \tag{5.5}$$

Since $(\bar{\Phi}_k)_{k\geq k_0}$ is monotonically decreasing, its limit, be it $\varphi_\star$, exists and from (5.5) it straightforwardly follows that either $(\|r^k\|^2)_{k\in\mathbb{N}}$ is summable or $\varphi_\star = -\infty$, the latter condition entailing unboundedness of $(x^k)_{k\in\mathbb{N}}$ due to continuity of φ_γ. Moreover, from (5.4) we have that the upper limits of the sequences $(\varphi(x^k))_{k\in\mathbb{N}}$, $(\varphi_\gamma(\bar{x}^k))_{k\in\mathbb{N}}$ and $(\varphi(\bar{x}^k))_{k\in\mathbb{N}}$ cannot be larger than $\varphi_\star$.

If $\varphi_\star = -\infty$, then all the sequences tend to $-\infty$ as well, and necessarily $\omega(x^j) = \omega(\bar{x}^j) = \emptyset$. Indeed, if $(x^{k'})_{k'\in K'} \to x'$, then by continuity we would get

$$\varphi_\gamma(x') = \lim_{K'\ni k'\to\infty} \varphi_\gamma(x^{k'}) = \lim_{k\to\infty} \varphi_\gamma(x^k) = -\infty$$

which contradicts the real-valuedness of φ_γ, and similarly if $(\bar{x}^{k'})_{k'\in K'} \to x'$, by using the fact that $\varphi_\gamma(\bar{x}^{k'}) \leq \varphi_\gamma(x^{k'})$. This concludes the proof of 5.6*(ii)*.

Otherwise, if $\varphi_\star \in \mathbb{R}$, since

$$0 \leftarrow \bar{\Phi}_k - \bar{\Phi}_{k+1} = p_k\big(\bar{\Phi}_k - \varphi_\gamma(x^{k+1})\big) \geq p_{\min}\big(\bar{\Phi}_{k+1} - \varphi_\gamma(x^{k+1})\big) \geq 0$$

by letting $k \to \infty$ we obtain again that $\varphi_\gamma(x^{k+1}) \to \varphi_\star$.

Suppose now that $(\bar{x}^k)_{k\in\mathbb{N}}$ is bounded, and let L be the Lipschitz modulus of φ_γ on a compact set that contains it (cf. Prop. 4.1*(i)*). Then,

$$\big\|\varphi_\gamma(\bar{x}^k) - \varphi_\gamma(x^k)\big\| \leq L\big\|r^k\big\| \to 0$$

due to 5.6*(ii)*. Combining with (5.4), this concludes the proof of 5.6*(iv)* and 5.6*(v)*.

Suppose now that $\omega(x^j) \neq \emptyset$; from 5.6*(ii)* it follows that $r^k \to 0$. Consider a subsequence $(x^{k'})_{k'\in K'}$ and suppose that $x^{k'} \to x'$ for some $x' \in \mathbb{R}^n$. Since $\varepsilon_{k'} := \|\bar{x}^{k'} - x^{k'}\| = \|r^{k'}\| \to 0$, if $(x^{k'})_{k'\in K'}$ converges to a point x', then so does $(\bar{x}^{k'})_{k'\in K'}$, and viceversa. Moreover,

$$x^{k'} \in \mathbf{B}(\bar{x}^{k'};\varepsilon_{k'}) \subseteq \mathbf{prox}_{\gamma g}\left(x^{k'} - \gamma\nabla f(x^{k'})\right) + \mathbf{B}(0;\varepsilon_{k'})$$

where the set sum is meant in the Minkowski sense. By continuity of ∇f it follows that the forward steps $x^{k'} - \gamma\nabla f(x^{k'})$ converge to $x' - \gamma\nabla f(x')$ and outer semicontinuity of $\mathbf{prox}_{\gamma g}$ [20, Ex. 5.23(b)] then implies that $x' \in \mathbf{prox}_{\gamma g}(x' - \gamma\nabla f(x'))$, proving x' to be γ-critical. The proof of 5.6*(iii)* follows invoking Prop. 3.6*(i)* and Thm. 4.3*(i)*.

It remains to show 5.6*(vi)*. If $(x^k)_{k\in\mathbb{N}}$ is bounded, then also $(\bar{x}^k)_{k\in\mathbb{N}}$ is, due to compact-valuedness of $\mathbf{prox}_{\gamma g}$ [20, Thm. 1.25]. If $(x^k)_{k\in\mathbb{N}}$ is unbounded, then from 5.6*(ii)* it follows that $\omega(x^j) = \emptyset$, hence $\omega(\bar{x}^j) = \emptyset$ due to 5.6*(iii)*, and therefore $(\bar{x}^k)_{k\in\mathbb{N}}$ cannot be bounded. The proof is complete. □

Remark 5.7 (Existence of cluster points)**.** Notice that boundedness of $(\bar{x}^k)_{k\in\mathbb{N}}$ (hence that of $(x^k)_{k\in\mathbb{N}}$ due to Thm. 5.6*(vi)*) is ensured if φ (and *not* necessarily the FBE), has bounded level sets. Indeed, for $\gamma = \gamma_\infty$ and $k_0 \in \mathbb{N}$ as in Lemma 5.1 it holds that

$$\varphi(\bar{x}^k) \leq \varphi_\gamma(x^k) \leq \varphi_\gamma(x^{k_0}) \quad \text{for all } k \geq k_0$$

and therefore $(\bar{x}^k)_{k\geq k_0} \subset \{\varphi \leq \varphi_\gamma(x^{k_0})\}$. Notice that unless γ_k is constant (e.g. if $L_0 \geq L_f$), the sole boundedness of the starting level set $\{\varphi \leq \varphi(\bar{x}^0)\}$ is not enough, as it does not take into account the fact that γ_k may change possibly resulting in the case that eventually $\varphi(\bar{x}^k) > \varphi(\bar{x}^0)$.

We conclude this section showing how in Theorem 5.6 the requirement of $f \in C^{1,1}$ can be relaxed to $f \in C^{1+}$ when g has bounded domain.

Proposition 5.8 (Subsequential convergence for $f \in C^{1+}(\mathbb{R}^n)$)**.** *Suppose that $f \in C^{1+}(\mathbb{R}^n)$ and that $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a bounded-domained function satisfying Assumption I. Consider the iterates generated by Algorithm 1, and suppose that the sequence of directions $(d^k)_{k\in\mathbb{N}}$ selected at step 4 is bounded. Then, all the claims of Theorem 5.6 hold and, additionally, $\omega(x^j) \neq \emptyset$.*

Proof. By definition of proximal mapping, necessarily $(\bar{x}^k)_{k\in\mathbb{N}} \subset \mathbf{dom}\, g$. Therefore, letting $D = \mathbf{sup}(\|d^k\|)_{k\in\mathbb{N}}$ and denoting with $\Omega = \mathbf{dom}\, g + \overline{\mathbf{B}}(0;D)$ the compact D-enlargement of $\mathbf{dom}\, g$, since $x^{k+1} = \bar{x}^k + \tau_k d^k$ and $\tau_k \in [0,1]$, it also holds that $(x^k)_{k\in\mathbb{N}} \subset \Omega$ (by possibly enlarging D, without loss of generality we may assume that $x^0 \in \Omega$ too). Observe that Lem.s 5.1 and 5.4 apply by replacing L_f with the Lipschitz constant of ∇f on the compact set Ω. We may now trace the proof of Thm. 5.6 and come to the same conclusions. Finally, that $\omega(x^j) \neq \emptyset$ follows by boundedness of Ω. □

5.4. **Global convergence.** In the previous subsection we proved the (sub-) optimality of the cluster points of the sequence generated by Algorithm 1. We now add mild conditions that ensure the convergence to a unique limit point. One of these is prox-regularity of the cluster points, which was discussed in Section 4.2, so to ensure first-order properties at the candidate limit point; this is automatic if g is convex. Then we need to impose some constraints on the norm of the arbitrarily selected

directions d^k, as in (5.6). The last ingredient is the so-called *Kurdyka-Łojasiewicz* property on φ_γ, a mild requirement that we will restate for the reader's convenience in Definition 5.11. The recent work [9, §3] discusses in detail some properties of the original functions f and g that ensure the FBE to satisfy such requirement.

We begin by showing some properties of the cluster points of the sequence generated by Algorithm 1; to do so, we need the following simple result.

Lemma 5.9. *Suppose that f and g satisfy Assumption I, and consider the iterates generated by Algorithm 1. Suppose further that*

$$\text{there exists } D \geq 0 \text{ such that } \|d^k\| \leq D\|r^k\| \quad \text{for all } k. \tag{5.6}$$

Then,

(i) $\|x^{k+1} - x^k\| \leq (1+D)\|r^k\|$

(ii) $\|\bar{x}^{k+1} - \bar{x}^k\| \leq \|r^{k+1}\| + (2+D)\|r^k\|$

In particular, if $\limsup_{k\to\infty} \varphi(\bar{x}^k) > -\infty$ *(e.g. if φ is lower bounded or the sequence $(x^k)_{k\in\mathbb{N}}$ is bounded), then*

(iii) $\|x^{k+1} - x^k\|$ *and* $\|\bar{x}^{k+1} - \bar{x}^k\|$ *converge to 0.*

Proof. For all k we have

$$\begin{aligned}\|x^{k+1} - x^k\| &= \|\bar{x}^k + \tau_k d^k - x^k\| = \|r^k + \tau_k d^k\| \\ &\leq \|r^k\| + \tau_k\|d^k\| \leq (1+D)\|r^k\|\end{aligned}$$

where in the last inequality we used the fact that $\tau_k \in (0,1]$. This proves 5.9*(i)*, and 5.9*(ii)* trivially follows by the triangular inequality

$$\|\bar{x}^{k+1} - \bar{x}^k\| \leq \|x^{k+1} - x^k\| + \|r^{k+1}\| + \|r^k\|.$$

Using this, 5.9*(iii)* follows from Thm. 5.6*(ii)*. □

Lemma 5.10. *Suppose that f and g satisfy Assumption I and consider the iterates generated by Algorithm 1. Suppose further that (5.6) is satisfied and that the sequence $(\bar{x}^k)_{k\in\mathbb{N}}$ is bounded. Then,*

(i) $\omega(x^j) = \omega(\bar{x}^j)$ *are nonempty, compact, and connected;*

(ii) $\varphi \equiv \varphi_\gamma$ *are constant on* $\omega(x^j)$*, where* $\gamma = \gamma_\infty$ *is as in Lemma 5.1;*

(iii) $\lim_{k\to\infty} \operatorname{dist}(x^k, \omega(x^j)) = \lim_{k\to\infty} \operatorname{dist}(\bar{x}^k, \omega(x^j)) = 0$.

Proof. That $\omega(x^j) = \omega(\bar{x}^j)$ and $\varphi \equiv \varphi_\gamma$ on $\omega(x^j)$ was proven in Thm. 5.6*(iii)*. Moreover, we know from Thm. 5.6*(iv)* that $\varphi_\gamma(x^k) \to \varphi_\star$ for some finite value $\varphi_\star$. Suppose that $(x^{k'})_{k'\in K'} \to x' \in \omega(x^j)$, then by continuity of φ_γ we have $\varphi(x^{k'}) \to \varphi(x') = \varphi_\star$, and 5.10*(ii)* follows from the arbitrarity of $x' \in \omega(x^j)$.

Finally, under the boundedness assumption we have that $\omega(x^j) \neq \emptyset$ and Thm. 5.6*(ii)* ensures $r^k \to 0$. The rest of the proof then follows from Lem. 5.9*(iii)* and [17, Lem. 5(ii)-(iii) and Rem. 5]. □

We now analyze the convergence of the iterates of Algorithm 1 to a critical point under the assumption that φ satisfies the Kurdyka-Łojasiewicz property [12, 13, 14]. For related works exploiting this property in proving convergence of optimization algorithms such as FBS we refer the reader to [15, 16, 6, 17, 18].

Definition 5.11 (KL property)**.** *A proper lower semicontinuous function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ has the Kurdyka-Łojasiewicz property (KL-property) at $x^\star \in \operatorname{dom}\partial\varphi$ if there exist $\eta \in (0, +\infty]$, a neighborhood $U_{x^\star}$ of $x^\star$, and a continuous concave function $\psi : [0,\eta] \to [0,+\infty)$ such that*

(i) $\psi(0) = 0$*;*

(ii) ψ *is* C^1 *with* $\psi' > 0$ *on* $(0,\eta)$*;*

(iii) for all $x \in U_{x^\star}$ *s.t.* $\varphi(x^\star) < \varphi(x) < \varphi(x^\star) + \eta$ *it holds that*

$$\psi'\big(\varphi(x) - \varphi(x^\star)\big)\,\mathbf{dist}\big(0, \partial\varphi(x)\big) \geq 1. \tag{5.7}$$

Function ψ in the previous definition is usually called *desingularizing function*. All subanalytic functions which are continuous on their domain have the KL property [28].

Lemma 5.12 (Uniformized KL property)**.** *Let* $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ *be a proper lower semicontinuous function and let* $K \subseteq \mathbf{dom}\,\varphi$ *be a nonempty compact set. Suppose that the function* φ *is constant on* K *and has the KL-property at every* $x^\star \in K$*. Then, there exist* $\eta, \varepsilon \in (0, +\infty]$ *and a continuous concave function* $\psi : [0, \eta] \to [0, +\infty)$ *satisfying Definition 5.11(i), 5.11(ii), and*

(iii') for every $x^\star \in K$ *and* x *s.t.* $\mathbf{dist}(x, K) < \varepsilon$ *and* $\varphi(x^\star) < \varphi(x) < \varphi(x^\star) + \eta$ *it holds that*

$$\psi'\big(\varphi(x) - \varphi(x^\star)\big)\,\mathbf{dist}\big(0, \partial\varphi(x)\big) \geq 1.$$

Proof. See [17, Lem. 6]. □

Theorem 5.13 (Global convergence)**.** *Suppose that* f *and* g *satisfy Assumption I, that* $f \in C^2(\mathbb{R}^n)$*, and that the directions* $(d^k)_{k\in\mathbb{N}}$ *are selected so as to satisfy* (5.6)*. Consider the iterates generated by Algorithm 1 with* p_k *eventually always selected equal to* 1*, and let* $\gamma = \gamma_\infty$*,* $\sigma = \sigma_\infty$ *and* $k_0 \in \mathbb{N}$ *be as in Lemma 5.1. Suppose that* $(x^k)_{k\in\mathbb{N}}$ *remains bounded, that* φ_γ *has the KL property on* $\omega(x^j)$*, and that every cluster point is prox-regular. Then,* $(x^k)_{k\in\mathbb{N}}$ *and* $(\bar{x}^k)_{k\in\mathbb{N}}$ *are convergent to the same point* $x^\star$*, and the sequence of residuals* $(r^k)_{k\in\mathbb{N}}$ *is summable.*

Proof. Let η, ε and ψ be as in Lem. 5.12 with respect to the KL-property of φ_γ on the set $\omega(x^j)$, which is compact and over which φ_γ is constant as shown in Lem.s 5.10*(i)* and 5.10*(ii)*. Let $\varphi_\star := \lim_{k\to\infty} \varphi_\gamma(x^k)$, which exists and is finite as ensured by Thm.s 5.6*(ii)*, 5.6*(iv)* and 5.6*(v)*, and let $k_1 \geq k_0$ be such that $p_k = 1$ for all $k \geq k_1$. Then, for all such k it holds that (cf. step 7 and (5.1))

$$\bar{\Phi}_k = \varphi_k(x^k) \quad \text{and} \quad \varphi_\gamma(x^k) > \varphi_\gamma(x^{k+1}) > \varphi_\star \qquad \forall k \geq k_1. \tag{5.8}$$

By possibly restricting ε, from Thm. 4.6*(ii)* and since $\omega(x^j)$ is compact it follows that φ_γ is differentiable in an ε-enlargement of $\omega(x^j)$. By Lem.s 5.10*(iii)* and 5.1 there exists $k_2 \geq k_1$ such that for all $k \geq k_2$ we have $\varphi_\star < \varphi_\gamma(x^k) < \varphi_\star + \eta$ and $\mathbf{dist}(x^k, \omega(x^j)) < \varepsilon$. For all such k, by Thm. 4.6*(ii)* we have

$$\partial\varphi_\gamma(x^k) = \nabla\varphi_\gamma(x^k) = \tfrac{1}{\gamma}[\mathrm{Id} - \gamma\nabla^2 f(x^k)]r^k$$

and the uniformized KL property yields

$$\psi'\big(\varphi_\gamma(x^k) - \varphi_\star\big) \geq \frac{1}{\|\nabla\varphi_\gamma(x^k)\|} \geq \frac{\gamma}{(1+\gamma L_f)\|r^k\|}. \tag{5.9}$$

Letting $\Delta_k := \psi\big(\varphi_\gamma(x^k) - \varphi_\star\big) > 0$, by concavity of ψ and (5.8) it follows that

$$\begin{aligned}
\Delta_k - \Delta_{k+1} &\geq \psi'\big(\varphi_\gamma(x^k) - \varphi_\star\big)\big(\varphi_\gamma(x^k) - \varphi_\gamma(x^{k+1})\big) \\
&\geq \gamma\frac{\varphi_\gamma(x^k) - \varphi_\gamma(x^{k+1})}{(1+\gamma L_f)\|r^k\|} = \gamma\frac{\bar{\Phi}_k - \bar{\Phi}_{k+1}}{(1+\gamma L_f)\|r^k\|} \\
&\geq \frac{\gamma\sigma p_{\min}}{1+\gamma L_f}\|r^k\| \qquad (5.10)
\end{aligned}$$

where the last inequality follows from Thm. 5.6*(i)*. Telescoping the inequality it follows that $(\|r^k\|)_{k\in\mathbb{N}}$ is summable, hence, due to Lem. 5.9*(i)*, also $(\|x^{k+1} - x^k\|)_{k\in\mathbb{N}}$ is. Therefore, $(x^k)_{k\in\mathbb{N}}$ is a Cauchy sequence and admits a limit, this being also the limit of $(\bar{x}^k)_{k\in\mathbb{N}}$ in light of Thm. 5.6*(iii)* and boundedness of $(\bar{x}^k)_{k\in\mathbb{N}}$ due to Thm. 5.6*(vi)*. □

In case where φ is subanalytic, the desingularizing function can be taken of the form $\psi(s) = \sigma s^{1-\theta}$, for $\sigma > 0$ and $\theta \in [0, 1)$ [28]. In this case, the condition in Definition 5.11*(iii)* is referred to as Łojasiewicz inequality. Depending on the value of θ we can derive local convergence rates for Algorithm 1.

Theorem 5.14 (Local linear rate under Łojasiewicz property)**.** *Suppose that the assumptions of Theorem 5.13 are satisfied, and that the KL function can be taken as*

$$\psi(s) = \rho s^{1-\theta} \qquad \exists \rho > 0, \theta \in (0, 1/2]. \tag{5.11}$$

Then, $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$ converge with asymptotic R-linear rate.

Proof. From Thm. 5.13 we know that $(x^k)_{k\in\mathbb{N}}$ converges to a point, be it $x^\star$, and let $\varphi_\gamma(x^k) \to \varphi_\star$. From Lem.s 5.9*(i)* and 5.9*(ii)* we have

$$\|x^k - x^\star\| \leq \sum_{i\geq k} \|x^{i+1} - x^i\| \leq (1+D) \sum_{i\geq k} \|r^i\| \quad \text{and} \quad \|\bar{x}^k - x^\star\| \leq (3+D) \sum_{i\geq k} \|r^i\|.$$

Therefore, the whole proof follows once we show that $B_k := \sum_{i\geq k} \|r^i\|$ converges with asymptotic Q-linear rate. With ψ as in (5.11), inequality (5.9) reads

$$\varphi_\gamma(x^k) - \varphi_\star \leq \left[\tfrac{1+\gamma L_f}{\gamma} \rho (1-\theta) \|r^k\|\right]^{1/\theta}$$

and therefore,

$$\Delta_k = \psi\big(\varphi_\gamma(x^k) - \varphi_\star\big) = \rho\big[\varphi_\gamma(x^k) - \varphi_\star\big]^{1-\theta} \leq \rho\left[\tfrac{1+\gamma L_f}{\gamma} \rho (1-\theta) \|r^k\|\right]^{\frac{1-\theta}{\theta}}.$$

Since $r^k \to 0$, for large k the quantity in square brackets is smaller than 1 and when $\theta \in (0, 1/2]$ the whole rightmost-hand side is maximezed for $\theta = 1/2$, which yields

$$\Delta_k \leq \tfrac{1+\gamma L_f}{\gamma} \rho^2 (1-\theta) \|r^k\|. \tag{5.12}$$

On the other hand, from (5.10) we get

$$B_k = \sum_{i\geq k} \|r^i\| \leq \frac{1+\gamma L_f}{\gamma \sigma p_{\min}} \sum_{i\geq k} (\Delta_i - \Delta_{i+1}) \leq \frac{1+\gamma L_f}{\gamma \sigma p_{\min}} \Delta_k.$$

Combining with (5.12) we obtain that there exists a constant $C > 0$ such that $B_k \leq C\|r^k\| = C(B_k - B_{k+1})$ for k large enough, *i.e.*, $B_{k+1} \leq (1 - 1/C) B_k$, proving asymptotic Q-linear convergence of B_k, hence the claim on $(x^k)_{k\in\mathbb{N}}$. □

Finally, when g has generalized second-order properties at a strong local minimum for φ, then φ_γ has the Łojasiewicz property with $\theta = 1/2$.

Proposition 5.15 (Łojasiewicz property of φ_γ for strong local minima)**.** *Suppose that f and g satisfy Assumption I, and that Assumption II is satisfied with respect to a strong local minimum $x^\star$ of φ. Then, for all $\gamma < \mathbf{min}\,\{\Gamma(x^\star), 1/L_f\}$ φ_γ has the Łojasiewicz property (5.11) with $\theta = 1/2$.*

Proof. From Thm. 4.14*(e)* it follows that $x^\star$ is a strong local minimum for φ_γ at which φ_γ is twice differentiable with positive definite Hessian. In fact, we prove that the claim holds for all functions satisfying these properties. Letting

$$\lambda := \lambda_{\min}\big(\nabla^2 \varphi_\gamma(x^\star)\big) > 0 \quad \text{and} \quad \Lambda := \lambda_{\max}\big(\nabla^2 \varphi_\gamma(x^\star)\big) > 0,$$

since $\nabla \varphi_\gamma(x^\star) = 0$, from a second-order expansion of φ_γ and a first-order expansion of $\nabla \varphi_\gamma$ we obtain that there exists a neighborhood $U_{x^\star}$ of $x^\star$ such that

$$\varphi_\gamma(x) - \varphi_\gamma(x^\star) \leq \tfrac{\Lambda}{4} \|x - x^\star\|^2 \quad \text{and} \quad \|\nabla \varphi_\gamma(x)\| \geq \tfrac{\lambda}{2} \|x - x^\star\| \qquad \forall x \in U_{x^\star}.$$

Let now $\psi(s) = \sigma\sqrt{s}$ for some $\sigma > 0$ to be determined. For all $x \in U_{x^\star}$ we have

$$\psi'(\varphi_\gamma(x) - \varphi_\gamma(x^\star))\|\nabla\varphi_\gamma(x)\| = \frac{\sigma}{2\sqrt{\varphi_\gamma(x) - \varphi_\gamma(x^\star)}}\|\nabla\varphi_\gamma(x)\| \geq \frac{\sigma\lambda}{2\sqrt{\Lambda}}.$$

Letting $\sigma = \frac{2\sqrt{\Lambda}}{\sigma\lambda}$ we obtain that ψ is a KL-function for φ_γ at $x^\star$. □

Corollary 5.16. *Suppose that f and g are convex and satisfy Assumption I, and that Assumption II is satisfied with respect to a strong local minimum $x^\star$ of φ. Consider the sequences generated by Algorithm 1. Then the sequences $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$ converge to $x^\star$. If, moreover, $p_k = 1$ is eventually always selected and $\gamma < 1/L_f$ where $\gamma = \gamma_\infty$ as in Lemma 5.1, then $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$ converge to $x^\star$ with asymptotic R-linear rate.*

Proof. Due to convexity and strong minimality, φ has bounded level sets, $x^\star$ is the unique critical point and $\Gamma(x^\star) = +\infty$. Rem. 5.7 and Thm.s 5.6*(ii)* and 5.6*(iii)* then ensure that $(x^k)_{k\in\mathbb{N}}$ is bounded with $\omega(x^j) = \{x^\star\}$, which proves the first statement.

The proof of the second statement follows from Prop. 5.15. □

6. Choice of the directions

We now focus on the selection of the directions $(d^k)_{k\in\mathbb{N}}$ in Algorithm 1. Ideally, we look for fast directions for solving the system of nonlinear equations $R_\gamma(x) = 0$, as in (5.2). Under the assumptions of Theorem 4.6*(i)*, the residual R_γ is strictly continuous around critical points; if R_γ is also semismooth, then generalized Newton methods can be implemented that achieve fast local convergence [29, §7.5]. This is a consequence of the results that will be proven in Section 6.1. However, this specific approach is beyond the scope of the paper, which instead aims at an algorithmic scheme based on the exact same oracle as the forward-backward operator T_γ. For this purpose we analyze in detail the employment of *quasi-Newton* directions, and then discuss *limited-memory* variants suited for large-scale problems.

6.1. Conditions for superlinear convergence.

We remind that $(x^k)_{k\in\mathbb{N}}$ with $x^k \neq x^\star$ for all k is said to be *superlinearly convergent* to $x^\star$ if

$$\lim_{k\to\infty} \frac{\|x^{k+1} - x^\star\|}{\|x^k - x^\star\|} = 0.$$

The following definition is taken from [29, §7.5].

Definition 6.1 (Superlinearly convergent directions)**.** *Let $(x^k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ be a sequence converging to a point $x^\star$. We say that a sequence $(d^k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ is* superlinearly convergent with respect to $(x^k)_{k\in\mathbb{N}}$ *if*

$$\lim_{k\to\infty} \frac{\|x^k + d^k - x^\star\|}{\|x^k - x^\star\|} = 0.$$

For Algorithm 1, the next result gives a Dennis-Moré condition ensuring that the iterates converge superlinearly.

Theorem 6.2. *Suppose that f and g satisfy Assumption I, and consider the iterates generated by Algorithm 1. Let $\gamma = \gamma_\infty$ be as in Lemma 5.1 and suppose that the sequence $(x^k)_{k\in\mathbb{N}}$ converges to a point $x^\star$ at which R_γ is strictly differentiable with $JR_\gamma(x^\star)$ nonsingular and around which it is Lipschitz continuous. If the directions $(d^k)_{k\in\mathbb{N}}$ satisfy the Dennis-Moré condition*

$$\lim_{k\to\infty} \frac{\|\bar{r}^k + JR_\gamma(x^\star)d^k\|}{\|d^k\|} = 0 \tag{6.1}$$

for some $\bar{r}^k \in \bar{x}^k - T_\gamma(\bar{x}^k)$, then

(i) $(d^k)_{k\in\mathbb{N}}$ *are superlinearly convergent with respect to* $(\bar{x}^k)_{k\in\mathbb{N}}$*;*

(ii) if $\tau_k = 1$ *is eventually always accepted, then* $x^k \to x^\star, \bar{x}^k \to x^\star, r^k \to 0$ *and* $\bar{r}^k \to 0$ *with superlinear rate.*

Proof. Let $k_0 \in \mathbb{N}$ be as in Lem. 5.1, and since $\bar{x}^k \to x^\star$ due to Thm. 5.6*(iii)*, by possibly enlarging k_0 we may assume that x^k and $\bar{x}^k$ belong to a convex neighborhood of $x^\star$ in which R_γ is Lipschitz continuous with modulus, say, L_R. By single-valuedness of R_γ, for all $k \geq k_0$ we may write $R_\gamma(x^k)$ and $R_\gamma(\bar{x}^k)$ in place of r^k and $\bar{r}^k$, respectively. From Thm.s 5.6*(ii)* and 5.6*(iii)* we know that $R_\gamma(\bar{x}^k) \to 0$, therefore from (6.1) necessarily $d^k \to 0$. Let $x_0^{k+1} := \bar{x}^k + d^k$; then, from (6.1) we have

$$0 \leftarrow \frac{R_\gamma(\bar{x}^k) + JR_\gamma(x^\star)d^k}{\|d^k\|} = \frac{R_\gamma(\bar{x}^k) + JR_\gamma(x^\star)(x_0^{k+1} - \bar{x}^k) - R_\gamma(x_0^{k+1})}{\|x_0^{k+1} - \bar{x}^k\|} + \frac{R_\gamma(x_0^{k+1})}{\|x_0^{k+1} - \bar{x}^k\|}$$

and since $x_0^{k+1} - \bar{x}^k = d^k \to 0$, from strict differentiability of R_γ at $x^\star$ applied on the first term on the right-hand side it follows that

$$\lim_{k\to\infty} \frac{\|R_\gamma(x_0^{k+1})\|}{\|x_0^{k+1} - \bar{x}^k\|} = 0. \tag{6.2}$$

Since $JR_\gamma(x^\star)$ is nonsingular, for some neighborhood $U_{x^\star}$ of $x^\star$ and $\alpha > 0$ it holds that

$$\|R_\gamma(x)\| \geq \alpha\|x - x^\star\| \qquad \text{for all } x \in U_{x^\star}.$$

Since $x^k \to x^\star$, eventually $x_0^{k+1} \in U_{x^\star}$. From (6.2) we obtain

$$0 \leftarrow \frac{\|R_\gamma(x_0^{k+1})\|}{\|x_0^{k+1} - \bar{x}^k\|} \geq \alpha\frac{\|x_0^{k+1} - x^\star\|}{\|x_0^{k+1} - \bar{x}^k\|} \geq \alpha\frac{\|x_0^{k+1} - x^\star\|}{\|x_0^{k+1} - x^\star\| + \|\bar{x}^k - x^\star\|} = \alpha\frac{\|x_0^{k+1} - x^\star\|/\|\bar{x}^k - x^\star\|}{1 + \|x_0^{k+1} - x^\star\|/\|\bar{x}^k - x^\star\|}$$

and therefore

$$\lim_{k\to\infty} \frac{\|x_0^{k+1} - x^\star\|}{\|\bar{x}^k - x^\star\|} = 0. \tag{6.3}$$

Since $x_0^{k+1} = \bar{x}^k + d^k$, 6.2*(i)* follows.

Suppose now that there exists $k_1 \geq k_0$ such that $\tau_k = 1$ for all $k \geq k_1$. Equation (6.3) then reads $\lim_{k\to\infty} \|x^{k+1} - x^\star\|/\|\bar{x}^k - x^\star\| = 0$, and from the inequality

$$\begin{aligned}\|\bar{x}^k - x^\star\| &= \|\bar{x}^k - x^k + x^k - x^\star\| \leq \|R_\gamma(x^k)\| + \|x^k - x^\star\| \\ &= \|R_\gamma(x^k) - R_\gamma(x^\star)\| + \|x^k - x^\star\| \leq (L_R + 1)\|x^k - x^\star\|\end{aligned}$$

superlinear convergence of $(x^k)_{k\in\mathbb{N}}$ follows. Since $\|r^k\| \leq L_R\|x^k - x^\star\|$, then $r^k \to 0$ superlinearly; in turn, this concludes the proof of 6.2*(ii)* due to the inequalities $\|\bar{x}^k - x^\star\| \leq \|r^k\| + \|x^k - x^\star\|$ and $\|\bar{r}^k\| \leq L_R\|\bar{x}^k - x^\star\|$. □

Next, we show that under Assumptions II and III the step-size $\tau_k = 1$ in Algorithm 1 is eventually always accepted if the directions $(d^k)_{k\in\mathbb{N}}$ satisfy the Dennis-Moré condition (6.1) and the sequence converges to a strong local minimum. This fact is quite surprising considering that no "descent" property is required on the directions, a property upon which classical line-search methods strongly rely.

Theorem 6.3. *Suppose that f and g satisfy Assumption I, and that Assumptions II and III are satisfied with respect to a strong local minimum $x^\star$. Consider the iterates generated by Algorithm 1 and let $\gamma = \gamma_\infty$ be as in Lemma 5.1. Suppose that by selecting directions $(d^k)_{k\in\mathbb{N}}$ satisfying the Dennis-Moré condition (6.1) the sequence $(x^k)_{k\in\mathbb{N}}$ converges to $x^\star$. If $\gamma < \mathbf{min}\,\{\Gamma(x^\star), 1/L_f\}$, then eventually step-size $\tau_k = 1$ is always accepted and $x^k \to x^\star, \bar{x}^k \to x^\star, r^k \to 0$ and $\bar{r}^k \to 0$ with superlinear rate.*

Proof. From Thm. 4.6*(i)* and Lem. 4.12*(ii)* we know that R_γ is strictly differentiable at $x^\star$ and Lipschitz continuous around. Therefore, in light of Thm. 6.2*(ii)* it suffices to show that eventually step-size $\tau_k = 1$ is always accepted.

From Thm. 4.13 we know that $\nabla\varphi_\gamma$ is strictly differentiable at $x^\star$, and that

$$G_\star := \nabla^2\varphi_\gamma(x^\star) = \tfrac{1}{\gamma}Q_\gamma(x^\star)JR_\gamma(x^\star)$$

is symmetric and positive definite due to Thm. 4.14*(e)*, where $Q_\gamma \succ 0$ is as in (4.3).

Let $k_0 \in \mathbb{N}$ be as in Lem. 5.1. Due to Thm.s 4.6*(i)* and 4.6*(ii)* and Ass. III*(i)*, there exists $k_1 \geq k_0$ such that for all $k \geq k_1$ x^k and $\bar{x}^k$ belong to a convex neighborhood in which φ_γ is differentiable and Q_γ and R_γ are Lipschitz continuous with, say, L_Q and L_R as respective moduli.

We now adapt part of the proof of [29, Prop. 8.3.18] to our case, where semismoothness is replaced with strict twice differentiability at the solution. Notice first that from Thm. 6.2*(i)* and [29, Prop. 7.5.7] we have

$$\lim_{k\to\infty} \frac{\|\bar{x}^k - x^\star\|}{\|d^k\|} = 1 \quad \text{and therefore} \quad \lim_{k\to\infty} \frac{\|\bar{x}^k + d^k - x^\star\|}{\|d^k\|} = 0. \tag{6.4}$$

For $k \geq k_1$ we have

$$\begin{aligned}
\frac{\|\nabla\varphi_\gamma(\bar{x}^k) + G_\star d^k\|}{\|d^k\|} &= \frac{\|Q_\gamma(\bar{x}^k)R_\gamma(\bar{x}^k) + Q_\gamma(x^\star)JR_\gamma(x^\star)d^k\|}{\gamma\|d^k\|} \\
&\leq \frac{\|Q_\gamma(\bar{x}^k) - Q_\gamma(x^\star)\|\|R_\gamma(\bar{x}^k)\|}{\gamma\|d^k\|} \\
&\quad + \frac{\|Q_\gamma(x^\star)\|\|R_\gamma(\bar{x}^k) + JR_\gamma(x^\star)d^k\|}{\gamma\|d^k\|} \\
&\leq \frac{L_Q L_R\|\bar{x}^k - x^\star\|^2}{\gamma\|d^k\|} + \frac{\|Q_\gamma(x^\star)\|\|R_\gamma(\bar{x}^k) + JR_\gamma(x^\star)d^k\|}{\gamma\|d^k\|}.
\end{aligned}$$

In the last-hand side, as $k \to \infty$ the first term converges to 0 due to (6.4) and the fact that $\bar{x}^k \to x^\star$, and the second vanishes due to (6.1). In particular, the first-hand side converges to 0. Using Cauchy-Schwartz inequality we then obtain

$$\begin{aligned}
0 \leftarrow \frac{\left\|\nabla\varphi_\gamma(\bar{x}^k) + G_\star d^k\right\|}{\|d^k\|} &\geq \frac{\left|\left\langle \nabla\varphi_\gamma(\bar{x}^k),\, d^k\right\rangle + \left\langle G_\star d^k,\, d^k\right\rangle\right|}{\|d^k\|^2} \\
&\geq \frac{\left\langle \nabla\varphi_\gamma(\bar{x}^k),\, d^k\right\rangle}{\|d^k\|^2} + \lambda_{\min}(G_\star).
\end{aligned}$$

Therefore, letting $\alpha := \frac{1}{2}\lambda_{\min}(G_\star) > 0$, there exists $k_2 \geq k_1$ such that

$$\left\langle \nabla\varphi_\gamma(\bar{x}^k),\, d^k\right\rangle \leq -\alpha\|d^k\|^2 \quad \text{for all } k \geq k_2.$$

Using (5.3) and letting $\sigma = \sigma_\infty$ as in Lem. 5.1, for $k \geq k_2$ we have

$$\begin{aligned}
\varphi_\gamma(\bar{x}^k + d^k) - \varphi_\gamma(x^k) &= \varphi_\gamma(\bar{x}^k + d^k) - \varphi_\gamma(\bar{x}^k) + \varphi_\gamma(\bar{x}^k) - \varphi_\gamma(x^k) \\
&\leq \varphi_\gamma(\bar{x}^k + d^k) - \varphi_\gamma(\bar{x}^k) - \sigma\|r^k\|^2.
\end{aligned} \tag{6.5}$$

From strict differentiability of $\nabla\varphi_\gamma$ at $x^\star$, and since $\nabla\varphi_\gamma(x^\star) = 0$ we have the quadratic expansions

$$\begin{aligned}\varphi_\gamma(\bar{x}^k + d^k) &= \varphi_\gamma(x^\star) + \tfrac{1}{2}\langle G_\star(\bar{x}^k + d^k - x^\star), \bar{x}^k + d^k - x^\star\rangle + o(\|\bar{x}^k + d^k - x^\star\|^2)\\ \varphi_\gamma(\bar{x}^k) &= \varphi_\gamma(x^\star) + \tfrac{1}{2}\langle G_\star(\bar{x}^k - x^\star), \bar{x}^k - x^\star\rangle + o(\|\bar{x}^k - x^\star\|^2).\end{aligned}$$

Substracting the second from the first and using (6.4) we obtain

$$\begin{aligned}\varphi_\gamma(\bar{x}^k + d^k) - \varphi_\gamma(\bar{x}^k) &= -\tfrac{1}{2}\langle G_\star(x^k - x^\star),\, x^k - x^\star\rangle + o(\|x^k - x^\star\|^2)\\ &\le -\alpha\|x - x^\star\|^2 + o(\|x^k - x^\star\|^2).\end{aligned}$$

Therefore, there exists $k_3 \ge k_2$ such that $\varphi_\gamma(\bar{x}^k + d^k) \le \varphi_\gamma(\bar{x}^k)$ for all $k \ge k_3$. Combined with (6.5), it follows that for all $k \ge k_3$ we have

$$\varphi_\gamma(\bar{x}^k + d^k) \le \varphi_\gamma(x^k) - \sigma\|r^k\|^2 \overset{(5.4)}{\le} \bar{\Phi}_k - \sigma\|r^k\|^2.$$

Therefore, for all $k \ge k_3$ the line-search condition (5.1) holds with $\tau_k = 1$, proving unitary step-size to be eventually always accepted. □

6.2. **Quasi-Newton directions.** We now specialize the results of Section 6.1 to the case in which directions d^k in Algorithm 1 are selected as

$$d^k = -B_k^{-1}\bar{r}^k \qquad \text{for some} \quad \bar{r}^k \in \bar{x}^k - T_\gamma(\bar{x}^k) \tag{6.6a}$$

with respect to nonsingular matrices B_k satisfing the *secant condition*

$$B_{k+1}s^k = y^k \quad \text{with} \quad \begin{cases} s^k = x^{k+1} - \bar{x}^k \\ y^k = r^{k+1} - \bar{r}^k. \end{cases} \tag{6.6b}$$

For this specific selection, the Dennis-Moré condition becomes as in (6.7), which yields the following straightforward special instance of Theorem 6.2.

Theorem 6.4. *Suppose that f and g satisfy Assumption I, and that Assumptions II and III are satisfied with respect to a strong local minimum $x^\star$. Consider the iterates generated by Algorithm 1 and let $\gamma = \gamma_\infty$ be as in Lemma 5.1. Suppose that by selecting directions $(d^k)_{k\in\mathbb{N}}$ as in (6.6) the sequence $(x^k)_{k\in\mathbb{N}}$ converges to $x^\star$. If $\gamma < \mathbf{min}\{\Gamma(x^\star), 1/L_f\}$ and $(d^k)_{k\in\mathbb{N}}$ satisfies the Dennis-Moré condition*

$$\lim_{k\to\infty} \frac{\|(B_k - JR_\gamma(x^\star))s^k\|}{\|s^k\|} = 0, \tag{6.7}$$

then eventually step-size $\tau_k = 1$ is always accepted and $x^k \to x^\star, \bar{x}^k \to x^\star, r^k \to 0$ and $\bar{r}^k \to 0$ with superlinear rate.

There are many quasi-Newton updates available, the BFGS formula being the most popular, but many of them are provably superlinearly convergent only under the assumption of symmetricity (and nonsingularity) of the Jacobian at the solution. Unfortunately, the Jacobian of the residual R_γ in meaningful applications of problem (1.1) that we intend to solve does not enjoy this property. Broyden's method offers a universal alternative that does not necessitate symmetricity, but whose well performance is heavily dependant on some scalar products involved in its updates to be non-null. To bypass this issue, we propose the implementation of a modified Broyden scheme first introduced in [30] and later further developed in [31].

6.2.1. *Broyden.* The modified Broyden update as in [30] prescribes

$$B_{k+1} = B_k + \theta_k \frac{(y^k - B_k s^k)}{\|s^k\|^2}(s^k)^\intercal \tag{6.8a}$$

for a sequence $(\theta_k)_{k\in\mathbb{N}} \subset [0,2]$. Equivalently, letting $\widetilde{y}^k = (1-\theta_k)B_k s^k + \theta_k y^k$,

$$B_{k+1} = B_k + \frac{(\widetilde{y}^k - B_k s^k)}{\|s^k\|^2}(s^k)^\intercal. \tag{6.8b}$$

The original Broyden formula [32] corresponds to selecting $\theta_k \equiv 1$. Alternatively, [30] suggests to fix a parameter $\bar{\theta} \in (0,1)$ and select

$$\gamma_k := \frac{\langle B_k^{-1} y^k,\, s^k\rangle}{\|s^k\|^2} \quad \text{and} \quad \theta_k := \begin{cases} 1 & \text{if } |\gamma_k| \geq \bar{\theta} \\ \frac{1-\mathbf{sgn}(\gamma_k)\bar{\theta}}{1-\gamma_k} & \text{if } |\gamma_k| < \bar{\theta} \end{cases} \tag{6.9}$$

with the convention $\mathbf{sgn}\, 0 = 1$. Starting from an invertible matrix B_0 this specific selection ensures all the matrices B_k to be nonsingular.

From [19, Thm. 2.2] and [31, Thm. 2.4] it follows that the modified Broyden scheme satisfies the Dennis-Moré condition (6.7) provided that the mapping whose zeros are sought, R_γ in our case, is continuously differentiabile around the limiting point, be it $x^\star$. In [33] the requirements for the Dennis-Moré condition to be satisfied for quasi-Newton schemes are relaxed to semidifferentiability around $x^\star$ and *calmness* of the semiderivative at $x^\star$, namely

$$\|DR_\gamma(x)[d] - DR_\gamma(x^\star)[d]\| \leq L\|x - x^\star\| \qquad \text{for all } d \in \mathbb{R}^n \text{ with } \|d\| = 1.$$

Combining these results, superlinear (and global) convergence of Broyden method (6.8) in our framework can be stated as follows.

Theorem 6.5. *Suppose that f and g satisfy Assumption I, and that Assumptions II and III are satisfied with respect to a strong local minimum $x^\star$ at which the semiderivative DR_γ is calm. Consider the iterates generated by Algorithm 1 and let $\gamma = \gamma_\infty$ be as in Lemma 5.1. Suppose that by selecting directions $(d^k)_{k\in\mathbb{N}}$ as in (6.8) the sequence $(x^k)_{k\in\mathbb{N}}$ converges to $x^\star$. If $\gamma < \mathbf{min}\,\{\Gamma(x^\star), 1/L_f\}$, then*

(i) the Dennis-Moré condition (6.7) holds;

(ii) step-size $\tau_k = 1$ is eventually always accepted;

(iii) $x^k \to x^\star$, $\bar{x}^k \to x^\star$, $r^k \to 0$, and $\bar{r}^k \to 0$ at superlinear rate.

6.3. **Limited-memory quasi-Newton directions.** When the size of the problem is large, computing d^k using quasi-Newton methods as in (6.6) becomes more and more computationally infeasible as n grows. For this reason limited-memory quasi-Newton methods are usually employed in optimization.

The limited-memory BFGS method (L-BFGS) is probably the most widely used method of this class, and was first introduced in [34]. It is based on the BFGS update (see for example [35, §6.1]), but only the most recent $\widetilde{m} = \mathbf{min}\,\{m,k\}$ pairs (s^i, y^i) as in (6.6) are used at iteration k (here m is a parameter, usually $m \in \{3,\dots,20\}$) to update an initial approximation H_0^k of the inverse Hessian, and compute a descent direction. In practice, d^k is computed using a *two-loop recursion* [35, §7.2], so that no matrix storage is required, and in fact only $4mn$ multiplications are needed.

Applying L-BFGS with memory $m = 0$ and H_0^k as proposed in [34, Eq. (3.2)] yields the method of Barzilai and Borwein [36]. This consists of a line-search-free step-size selection for the gradient method. Its global convergence has been demonstrated in the strictly convex quadratic case [37] and a globally convergent variant using a nonmonotone line-search for smooth functions was proposed in [38]. Specifically, in the context of Algorithm 1 $d^k = -\alpha_k \bar{r}^k$ where $\alpha_k = \langle s^{k-1},\, y^{k-1}\rangle / \langle y^{k-1},\, y^{k-1}\rangle$. Only $2n$ multiplications are required in this case.

The convergence results for Algorithm 1, that were presented in Section 5.3 and Section 5.4, continue to hold even when limited-memory quasi-Newton directions are used. In particular, convergence of the iterates is guaranteed under the assumptions of Theorem 5.13, and the rate is locally linear under the assumptions of Theorem 5.14. Condition (5.6) can be enforced by appropriately restarting the method if necessary.

7. Simulations

In this section we present numerical simulations with the proposed algorithm. We compare Algorithm 1 with forward-backward splitting (FBS) and its accelerated variant (fast FBS) in the case of convex problems. In Algorithm 1 we set $\sigma_0 = (1-\gamma_0 L_0)/(4\gamma_0)$, $\alpha = \beta = 1/2$, $\gamma_0 = 0.95/L_0$, and used a numerical approximation L_0 to L_f when this is not available otherwise. For the nonmonotone line-search, we used the sequence $p_k = (\eta Q_k + 1)^{-1}$ where $Q_0 = 1$, $Q_{k+1} = \eta Q_k + 1$, $\eta = 0.85$: in this way $(p_k)_{k\in\mathbb{N}}$ is computed similarly to what is done in [39].

We performed experiments on small dimensional problems using the modified Broyden method (see Section 6.2.1) and BFGS to compute d^k in Algorithm 1. Although the JR_γ is nonsymmetric at the critical points in general, we observed that the symmetric updates of BFGS perform very well in practice. For large-scale problems, when storing the approximate inverse Hessians is infeasible, we used L-BFGS to compute search directions.

All experiments were performed in MATLAB. The implementation of the methods used in the tests are available online.[2]

7.1. **Lasso.** The problem is to find a sparse least square solution to a system which is usually underdetermined. This is done by minimizing $\varphi = f + g$ where

$$f(x) = \tfrac{1}{2}\|Ax - b\|_2^2, \quad g(x) = \lambda\|x\|_1,$$

and $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. In this case g is convex, and its (single-valued) proximal mapping is the soft-thresholding operation. The computationally relevant operation here is the computation of f and ∇f, which involves matrix-vector products with A and $A^\intercal$. Parameter λ modulates between a small residual $Ax^\star - b$ and a sparse solution vector $x^\star$, *i.e.*, the larger λ the more zero coefficients $x^\star$ has. In particular, $\lambda_{\mathbf{max}} = \|A^\intercal b\|_\infty$ is the minimum value such that for $\lambda \geq \lambda_{\mathbf{max}}$ the solution is $x^\star = 0$. We have $L_f = \|A^\intercal A\|$, which can be quickly approximated using power iterations.

We ran the algorithms on the SPEAR datasets.[3] First we considered Broyden and BFGS directions in Algorithm 1 on small dimensional instances, and the results are illustrated in Figure 1. We computed $x^\star$ (approximately) by running one of the methods with very restrictive termination criteria. Then we considered the largest instances implementing L-BFGS directions in Algorithm 1. Performance of the algorithms is detailed in Table 1. It is apparent that Algorithm 1 requires much less iterations and operations than fast FBS.

7.2. **Sparse logistic regression.** This is a convex problem, and the composite objective function consists of $f(x) = \sum_{i=1}^m \mathbf{log}(1 + e^{-b_i\langle a_i,\, x\rangle})$ and $g(x) = \lambda\|x\|_1$. Here vector $a_i \in \mathbb{R}^n$ contains the features of the i-th instance, and $b_i \in \{-1, 1\}$ indicates the correspondent class. The ℓ_1-regularization enforces sparsity in the solution. Indicating with $A \in \mathbb{R}^{m\times n}$ the matrix having rows $a_1, a_2, \ldots, a_m$, the optimal solution is $x^\star = 0$ if $\lambda \geq \lambda_{\mathbf{max}} := \frac{1}{2}\|A^\intercal b\|_\infty$. We ran the algorithms on

[2] http://github.com/kul-forbes/ForBES

[3] http://wwwopt.mathematik.tu-darmstadt.de/spear/

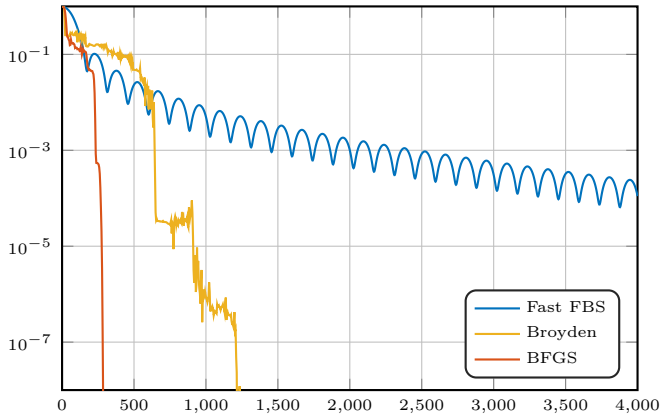


FIGURE 1. Lasso: performance on the `spear_inst_6` dataset, with $n = 1024$ and $m = 512$, for $\lambda = 0.1\lambda_{\max}$. On the x-axis is the number of matrix-vector products computed, on the y-axis is the relative error $\|x^k - x^\star\|_\infty / \|x^\star\|_\infty$.

ID	$\lambda/\lambda_{\max}$	$\mathrm{nnz}(x^\star)$	Fast FBS		L-BFGS	
			iterations	matvecs	iterations	matvecs
191	$1 \cdot 10^{-1}$	3	2260	4520	125	498
$m = 8192$	$5 \cdot 10^{-2}$	4	7529	15058	276	1102
$n = 32768$	$1 \cdot 10^{-2}$	7	11222	22444	592	2366
192	$9 \cdot 10^{-1}$	2	234	468	3	10
$m = 8192$	$7 \cdot 10^{-1}$	9	649	1298	28	110
$n = 32768$	$5 \cdot 10^{-1}$	21	2051	4102	120	478
193	$5 \cdot 10^{-1}$	3	2142	4284	182	726
$m = 8192$	$1 \cdot 10^{-1}$	4	2413	4826	479	1914
$n = 32768$	$5 \cdot 10^{-2}$	5	4761	9522	642	2566
194	$9 \cdot 10^{-1}$	6	338	676	16	62
$m = 8192$	$7 \cdot 10^{-1}$	29	1532	3064	223	890
$n = 32768$	$5 \cdot 10^{-1}$	51	2482	4964	385	1538
195	$5 \cdot 10^{-2}$	2	3215	6430	164	654
$m = 8192$	$1 \cdot 10^{-2}$	3	3833	7666	458	1830
$n = 32768$	$5 \cdot 10^{-3}$	4	10525	21050	688	2750
196	$7 \cdot 10^{-1}$	2	651	1302	12	46
$m = 8192$	$5 \cdot 10^{-1}$	4	1656	3312	28	110
$n = 32768$	$1 \cdot 10^{-1}$	11	3075	6150	395	1578
197	$1 \cdot 10^{-1}$	3	3863	7726	349	1394
$m = 8192$	$5 \cdot 10^{-2}$	4	4343	8686	483	1930
$n = 32768$	$1 \cdot 10^{-2}$	6	10894	21788	1397	5586
198	$9 \cdot 10^{-1}$	4	628	1256	49	194
$m = 8192$	$7 \cdot 10^{-1}$	17	2351	4702	219	874
$n = 32768$	$5 \cdot 10^{-1}$	25	3236	6472	252	1006
199	$1 \cdot 10^{-1}$	2	7163	14326	338	1350
$m = 8192$	$5 \cdot 10^{-2}$	3	2537	5074	554	2214
$n = 49152$	$1 \cdot 10^{-2}$	5	10460	20920	1271	5082
200	$9 \cdot 10^{-1}$	6	622	1244	71	282
$m = 8192$	$7 \cdot 10^{-1}$	10	2144	4288	166	662
$n = 49152$	$5 \cdot 10^{-1}$	19	2313	4626	298	1190

TABLE 1. Lasso: performance of the algorithms on the largest of the considered datasets. Column ID indicates the identification number of the dataset among the SPEAR instances. Results refer to iterations and matrix vector products needed to obtain $(\varphi(x^k) - \varphi_\star)/|\varphi_\star| \leq 10^{-6}$.

a series of datasets,[4] using full and limited-memory quasi-Newton directions. The results are displayed in Figure 2 and Figure 3.

[4] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

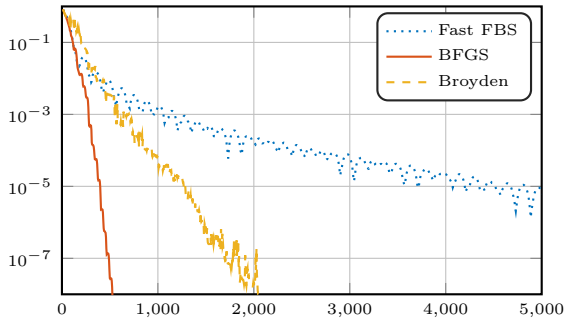
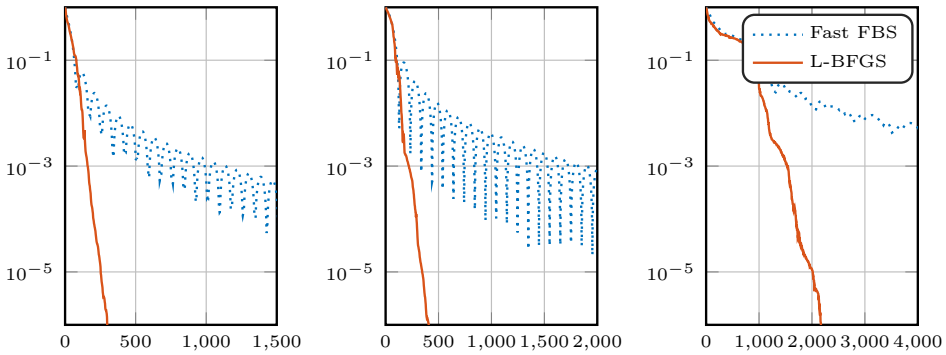


FIGURE 2. Sparse logistic regression: performance of the `w4a` dataset, with $n = 299$, $m = 6760$, using $\lambda = 0.05\lambda_{\max}$. On the x-axis is the number of matrix-vector products computed, on the y-axis is the relative error $\|x^k - x^\star\|_\infty / \|x^\star\|_\infty$.



(A) `real-sim` (B) `rcv1` (C) `news20`

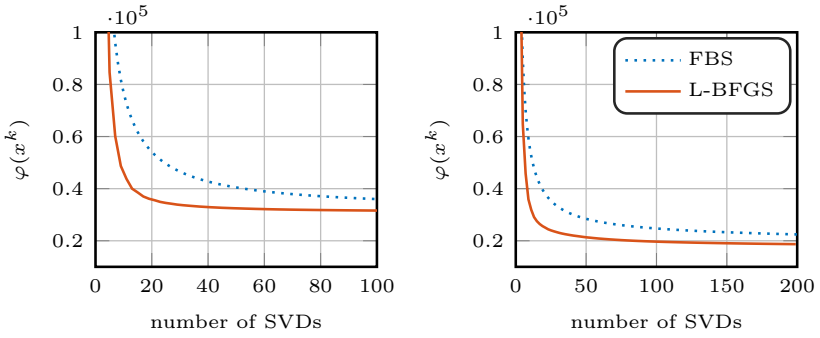
FIGURE 3. Sparse logistic regression: performance on large scale datasets, using $\lambda = 0.01\lambda_{\max}$. On the x-axis is the number of matrix-vector products computed, on the y-axis is the relative error $\|x^k - x^\star\|_\infty / \|x^\star\|_\infty$.

7.3. **Matrix completion.** The problem has the form

$$\underset{X \in \mathbb{R}^{m \times n}}{\textbf{minimize}} \ \tfrac{1}{2}\|\mathcal{A}(X) - b\|^2 \text{ subject to } \textbf{rank}(X) \leq r$$

where $\mathcal{A} : \mathbb{R}^{m \times n} \to \mathbb{R}^k$ is a linear mapping. In this case, $\mathcal{A}$ selects k entries from its matrix arguments and the goal is to find a low-rank matrix given a subset of its coefficients. This problem has application, for example, in recommender systems. The set of matrices of rank at most r is nonconvex and the projection onto it is given by $\mathbf{\Pi}_{\mathbf{rank} \leq r}(X) = U_r \, \mathbf{diag}(\sigma_1, \ldots, \sigma_r) V_r^T$. Here, $\sigma_1 \ldots \sigma_r$ are r largest singular values of X, and U_r, V_r are the matrices of left and right singular vectors, respectively: each computation of $\Pi_{\mathbf{rank} \leq r}$ entails the computation of a partial SVD, which is the most computational expensive operation in evaluating T_γ here. We ran the methods on the MovieLens100k dataset.[5] This consists of 10^5 ratings of 1682 movies from 943 users, so the problem has ≈ 1.6 millions variables. The results are in Figure 4.

[5] http://grouplens.org/datasets/movielens/



(A) $r = 5$ (B) $r = 15$

FIGURE 4. Matrix completion: convergence of the cost in FBS and Algorithm 1 with L-BFGS directions, for $r = 5, 15$.

		FBS		L-BFGS	
Graph	tol	∇f	avg. cut	∇f	avg. cut
G72	10^{-1}	22	$5.7436{\cdot}10^3$	17	$5.9479{\cdot}10^3$
$n = 10000$	10^{-2}	208	$6.1914{\cdot}10^3$	101	$6.2381{\cdot}10^3$
	10^{-3}	1674	$6.2514{\cdot}10^3$	620	$6.2625{\cdot}10^3$
	10^{-4}	15650	$6.2578{\cdot}10^3$	2702	$6.2616{\cdot}10^3$
G77	10^{-1}	22	$8.1506{\cdot}10^3$	17	$8.4333{\cdot}10^3$
$n = 14000$	10^{-2}	212	$8.7927{\cdot}10^3$	125	$8.8632{\cdot}10^3$
	10^{-3}	1784	$8.8700{\cdot}10^3$	846	$8.8909{\cdot}10^3$
	10^{-4}	17142	$8.8857{\cdot}10^3$	2868	$8.8861{\cdot}10^3$
G81	10^{-1}	22	$1.1511{\cdot}10^4$	17	$1.1918{\cdot}10^4$
$n = 20000$	10^{-2}	186	$1.2410{\cdot}10^4$	113	$1.2522{\cdot}10^4$
	10^{-3}	2284	$1.2548{\cdot}10^4$	978	$1.2562{\cdot}10^4$
	10^{-4}	17768	$1.2562{\cdot}10^4$	3036	$1.2570{\cdot}10^4$

TABLE 2. Max-cut: comparison between FBS and Algorithm 1 with L-BFGS directions. The "avg. cut" column shows the average value of 500 random cuts obtained from the computed solution. We used $r = 20$, and stopped the algorithms when $\|R_\gamma(x^k)\| \le \text{tol}$.

7.4. **Max-cut.** Given an undirected weighted graph with n nodes and weights collected in a symmetric matrix $W = (w_{ij}) \in \mathbb{R}^{n\times n}$, the goal is to partition (*cut*) the nodes into two subsets so that the sum of the weights connecting nodes in distinct subsets is maximized. This is a hard combinatorial problem for which the SDP relaxation in [40] provides a tight approximation to the global optimum. Instead [41] proposes to solve the following nonconvex but considerably smaller problem:

$$\underset{x_1 \dots x_n \in \mathcal{S}^r}{\textbf{minimize}} \sum_{i<j} w_{ij} \langle x_i,\, x_j \rangle, \tag{7.1}$$

where $\mathcal{S}^r$ denotes the unit sphere in $\mathbb{R}^r$. Provided that r is sufficiently large, (7.1) is equivalent to the SDP relaxation. From a solution to (7.1), a random cut can be computed by selecting a vector v uniformly at random on the unit sphere, and assigning each node i to either of the two subsets according to the $\mathbf{sgn}\,\langle x_i,\, v\rangle$. We applied Algorithm 1 with L-BFGS to the graphs of the *Gset* dataset[6] and $r = 20$. We considered more and more restrictive termination criteria based on $\|R_\gamma(x^k)\|$, and for each solution obtained we computed the average of 500 random cuts. The results are shown in Table 2.

7.5. **Matrix decomposition.** We consider the following problem, of approximating a given matrix $A \in \mathbb{R}^{m\times n}$ as the sum of a low-rank and a sparse component.

[6] http://web.stanford.edu/~yyye/yyye/Gset/

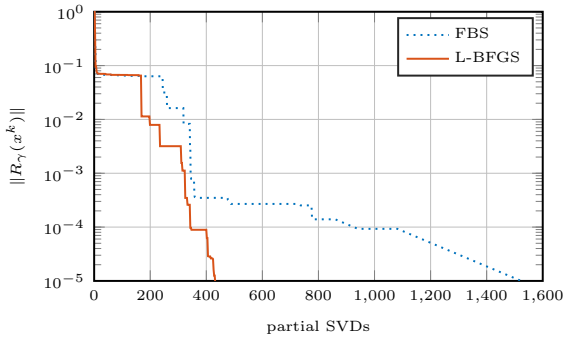


FIGURE 5. Matrix decomposition: convergence of the fixed-point residual of Algorithm 1 using L-BFGS directions compared to FBS on the *ShoppingMall* dataset.

Generally speaking, the problem can be formulated as follows

$$\underset{X_L, X_S \in \mathbb{R}^{m\times n}}{\textbf{minimize}} \ \tfrac{1}{2}\|A - X_L - X_S\|_F^2 + \Psi_L(X_L) + \Psi_S(X_S), \tag{7.2}$$

Here Ψ_L is a low-rank inducing regularization term, while Ψ_S is a sparsity inducing term. This problem has application, for example, is in the analysis of video imagery, specifically the separation of the background (fixed over time) scenery from the foreground (moving) objects in a series of video frames. In this case then, matrix A contains n video frames (columns), each consisting of m pixels, and X_L, X_S will contain the background scenery and foreground objects identified in each frame. For this task, we consider the nonconvex regularizers $\Psi_L = \delta_{\mathbf{rank}\leq r}$ and $\Psi_S = \lambda\|\cdot\|_0$. Therefore here $f(X_L, X_S) = \frac{1}{2}\|A - X_L - X_S\|_F^2$ and $L_f = 2$, while $g(X_L, X_S) = \delta_{\mathbf{rank}\leq r}(X_L) + \lambda\|X_S\|_0$. The proximal mapping of g is given by

$$\mathbf{prox}_{\gamma g}(X_L, X_S) = (\mathbf{\Pi}_{\mathbf{rank}\leq r}(X_L), \mathbf{prox}_{\gamma\lambda\|\cdot\|_0}(X_S)),$$

where $\mathbf{\Pi}_{\mathbf{rank}\leq r}$ is as in Section 7.3 and $\mathbf{prox}_{\gamma\lambda\|\cdot\|_0}$ is the hard-thresholding operation, defined componentwise for $i, j = 1 \dots m$ as

$$\left(\mathbf{prox}_{\gamma\lambda\|\cdot\|_0}(X_S)\right)_{ij} = \begin{cases} 0 & \text{if } \left(X_S\right)_{ij} \leq \sqrt{2\gamma\lambda}, \\ \left(X_S\right)_{ij} & \text{otherwise.} \end{cases}$$

We applied this technique to a sequence of 100 frames coming from the *ShoppingMall* dataset,[7] comparing the performance of FBS with Algorithm 1 using L-BFGS directions with memory 5. The results are shown in Figure 5 and Figure 6. For this experiment we considered $r = 1$ and $\lambda = 10^{-3}$.

8. CONCLUSIONS

The forward-backward envelope is a valuable tool for deriving efficient algorithms tackling nonsmooth and nonconvex problems of the form $\varphi = f + g$, as it can be used as a merit function to devise globally convergent line-search methods solving the system of nonlinear equations defining the stationary points of φ.

Algorithm 1 implements this idea, and we proved that it globally converges to a stationary point under the assumption that φ_γ has the Kurdyka-Łojasiewicz property. Furthermore, if the line-search directions satisfy the Dennis-Moré condition (for example, if they are determined according to the Broyden method), the convergence rate at strong local minima is superlinear.

[7] http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html

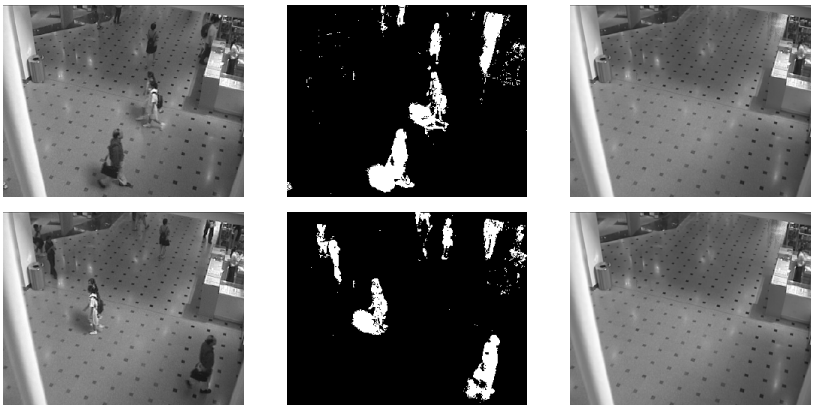


FIGURE 6. Matrix decomposition: separation of the background scenery from the foreground moving objects in the *ShoppingMall* dataset.

Numerical simulations with the proposed method on convex and nonconvex problems confirm our theoretical results. Using Broyden's method, BFGS (in the case of small-scale problems) and L-BFGS (for large-scale problems) to compute directions in Algorithm 1 greatly outperform FBS and (in the convex case) its accelerated variant. It is our belief that the surprising efficacy of (L-)BFGS is due to the fact that, under the appropriate assumptions, the Jacobian of R_γ at strong local minima is similar to a symmetric and positive definite matrix. Future investigation may better explain the effectiveness of symmetric update formulas in this framework.

REFERENCES

[1] G. Lan, Z. Lu, and R. D. Monteiro, "Primal-dual first-order methods with $O(1/\epsilon)$ iteration-complexity for cone programming," *Mathematical Programming*, vol. 126, no. 1, pp. 1–29, 2011.

[2] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $o(1/k^2)$," *Soviet Mathematics Doklady*, 1983.

[3] P. Tseng, "On accelerated proximal gradient methods for convex-concave optimization," Department of Mathematics, University of Washington, Tech. Rep., 2008.

[4] A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[5] Y. Nesterov, "Gradient Methods for Minimizing Composite Functions," *Mathematical Programming*, vol. 140, no. 1, pp. 125–161, 2013.

[6] 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," *Mathematical Programming*, vol. 137, no. 1, pp. 91–129, 2013.

[7] P. Patrinos and A. Bemporad, "Proximal Newton methods for convex composite optimization," in *IEEE Conference on Decision and Control*, 2013, pp. 2358–2363.

[8] L. Stella, A. Themelis, and P. Patrinos, "Forward-backward quasi-Newton methods for nonsmooth optimization problems," *ArXiv e-prints*, Apr. 2016.

[9] T. Liu and T. Kei Pong, "Further properties of the forward-backward envelope with applications to difference-of-convex programming," *ArXiv e-prints*, Apr. 2016.

[10] A. Beck and N. Hallak, "On the minimization over sparse symmetric sets: Projections, optimality conditions, and algorithms," *Math. Oper. Res.*, vol. 41, no. 1, pp. 196–223, 2016.

[11] R. A. Poliquin and R. T. Rockafellar, "Prox-regular functions in variational analysis," *Transactions of the American Mathematical Society*, vol. 348, no. 5, pp. 1805–1838, 1996.

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

[13] ——, "Sur la géométrie semi- et sous- analytique," *Annales de l'institut Fourier*, vol. 43, no. 5, pp. 1575–1595, 1993.

[14] K. Kurdyka, "On gradients of functions definable in o-minimal structures," *Annales de l'institut Fourier*, vol. 48, no. 3, pp. 769–783, 1998.

[15] H. Attouch and J. Bolte, "On the convergence of the proximal algorithm for nonsmooth functions involving analytic features," *Mathematical Programming*, vol. 116, no. 1-2, pp. 5–16, 2009.
[16] 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," *Mathematics of Operations Research*, vol. 35, no. 2, pp. 438–457, 2010.
[17] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, no. 1-2, pp. 459–494, 2014.
[18] P. Ochs, Y. Chen, T. Brox, and T. Pock, "iPiano: Inertial proximal algorithm for nonconvex optimization," *SIAM Journal on Imaging Sciences*, vol. 7, no. 2, pp. 1388–1419, 2014.
[19] J. E. Dennis and J. J. Moré, "A characterization of superlinear convergence and its application to quasi-Newton methods," *Mathematics of computation*, vol. 28, no. 126, pp. 549–560, 1974.
[20] R. T. Rockafellar and R. J. Wets, *Variational analysis*. Springer, 2011, vol. 317.
[21] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1995.
[22] Y. Nesterov, *Introductory lectures on convex optimization: a basic course*, ser. Applied optimization. Springer US, 2004.
[23] D. S. Bernstein, *Matrix mathematics: theory, facts, and formulas with application to linear systems theory*. Woodstock: Princeton University Press, 2009.
[24] R. T. Rockafellar, "First- and second-order epi-differentiability in nonlinear programming," *Transactions of the American Mathematical Society*, vol. 307, no. 1, pp. 75–108, 1988.
[25] ——, "Second-order optimality conditions in nonlinear programming obtained by way of epi-derivatives," *Mathematics of Operations Research*, vol. 14, no. 3, pp. 462–484, 1989.
[26] R. A. Poliquin and R. T. Rockafellar, "Amenable functions in optimization," *Nonsmooth optimization: methods and applications*, pp. 338–353, 1992.
[27] ——, "Second-order nonsmooth analysis in nonlinear programming," *Recent advances in nonsmooth optimization*, pp. 322–349, 1995.
[28] J. Bolte, A. Daniilidis, and A. Lewis, "The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems," *SIAM Journal on Optimization*, vol. 17, no. 4, pp. 1205–1223, 2007.
[29] F. Facchinei and J.-S. Pang, *Finite-imensional Variational Inequalities and Complementarity Problems*, ser. Springer Series in Operations Research and Financial Engineering. Springer New York, 2007.
[30] M. Powell, "A hybrid method for nonlinear equations," *Numerical Methods for Nonlinear Algebraic Equations*, pp. 87–144, 1970.
[31] J. J. Moré and J. A. Trangenstein, "On the global convergence of Broyden's method," *Mathematics of Computation*, vol. 30, no. 135, pp. 523–540, 1976.
[32] C. G. Broyden, "A class of methods for solving nonlinear simultaneous equations," *Mathematics of Computation*, vol. 19, no. 92, pp. 577–593, 1965.
[33] C.-M. Ip and J. Kyparisis, "Local convergence of quasi-Newton methods for B-differentiable equations," *Mathematical Programming*, vol. 56, no. 1-3, pp. 71–89, 1992.
[34] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, vol. 45, no. 1-3, pp. 503–528, 1989.
[35] J. Nocedal and S. J. Wright, *Numerical optimization*. Springer New York, 1999, vol. 2.
[36] J. Barzilai and J. M. Borwein, "Two-point step size gradient methods," *IMA Journal of Numerical Analysis*, vol. 8, Jan. 1988.
[37] M. Raydan, "On the Barzilai and Borwein choice of steplength for the gradient method," *IMA Journal of Numerical Analysis*, vol. 13, no. 3, pp. 321–326, 1993.
[38] ——, "The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem," *SIAM Journal on Optimization*, vol. 7, no. 1, pp. 26–33, 1997.
[39] H. Zhang and W. W. Hager, "A nonmonotone line search technique and its application to unconstrained optimization," *SIAM Journal on Optimization*, vol. 14, no. 4, pp. 1043–1056, 2004.
[40] M. X. Goemans and D. P. Williamson, "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming," *Journal Of The ACM*, vol. 42, no. 6, pp. 1115–1145, 1995.
[41] S. Burer and R. D. C. Monteiro, "A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization," *Mathematical Programming, Series B*, vol. 95, no. 2, pp. 329–357, 2003.