Forward-Backward Envelope for the Sum of Two Nonconvex Functions: Further Properties and Nonmonotone Linesearch Algorithms

Themelis, Andreas
Department of Electrical Engineering (ESAT-STADIUS), KU Leuven

Stella, Lorenzo
Department of Electrical Engineering (ESAT-STADIUS), KU Leuven

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

出版情報：SIAM Journal on Optimization. 28 (3), pp.2274-2303, 2018-08-21. Society for Industrial and Applied Mathematics
バージョン：
権利関係：

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

ANDREAS THEMELIS AND LORENZO STELLA AND PANOS PATRINOS †

Abstract. We propose ZeroFPR, a nonmonotone linesearch algorithm for minimizing the sum of two nonconvex functions, one of which is smooth and the other possibly nonsmooth. ZeroFPR is the first algorithm that, despite being fit for fully nonconvex problems and requiring only the black-box oracle of forward-backward splitting (FBS) — namely evaluations of the gradient of the smooth term and of the proximity operator of the nonsmooth one — achieves superlinear convergence rates under mild assumptions at the limit point when the linesearch directions satisfy a Dennis-Moré condition, and we show that this is the case for Broyden's quasi-Newton directions. Our approach is based on the forward-backward envelope (FBE), an exact and strictly continuous penalty function for the original cost. Extending previous results we show that, despite being nonsmooth for fully nonconvex problems, the FBE still enjoys favorable first- and second-order properties which are key for the convergence results of ZeroFPR. Our theoretical results are backed up by promising numerical simulations. On large-scale problems, by computing linesearch directions using limited-memory quasi-Newton updates our algorithm greatly outperforms FBS and its accelerated variant (AFBS).

Key words. Nonsmooth optimization, nonconvex optimization, forward-backward splitting, linesearch methods, quasi-Newton methods, prox-regularity.

AMS subject classifications. 90C06, 90C25, 90C26, 90C53, 49J52, 49J53.

1. Introduction.

In this paper we deal with optimization problems of the form

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

under the following requirements, which will be assumed without further mention.

Assumption I (Basic assumption)**.** *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 (see Section 2.1);*

(iii) a solution exists, that is, $\mathbf{argmin}\,\varphi \neq \emptyset$.

Both f and g are allowed to be nonconvex, making (1.1) prototypic for a plethora of applications spanning signal and image processing, machine learning, statistics, control and system identification. A well known algorithm addressing (1.1) is *forward-backward splitting* (FBS), 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, also known as *fast* forward-backward splitting (FFBS), can be derived thanks to the work of Nesterov [9, 36], that only require minimal additional computations per iteration but achieve the provably optimal global convergence rate of order $o(1/k^2)$ [6].

The work in [41] pioneered an alternative acceleration technique. The method is based on an exact, real-valued penalty function for the original problem (1.1), namely

*This work was supported by: KU Leuven internal funding: StG/15/043 Fonds de la Recherche Scientifique – FNRS and the Fonds Wetenschappelijk Onderzoek – Vlaanderen under EOS Project no 30468160 (SeLMA) FWO projects: G086318N; G086518N

†Department of Electrical Engineering (ESAT-STADIUS) & Optimization in Engineering Center (OPTEC) – KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
andreas.themelis@esat.kuleuven.be, lorenzostella@gmail.com, panos.patrinos@esat.kuleuven.be

the *forward-backward envelope* (FBE), defined as follows

$$\varphi_\gamma(x) = \varphi_\gamma^{f,g}(x) \coloneqq \inf_{z\in\mathbb{R}^n} \left\{ f(x) + \langle \nabla f(x), z-x\rangle + \tfrac{1}{2\gamma}\|z-x\|^2 + g(z) \right\} \tag{1.2}$$

where $\gamma > 0$ is a given parameter. We will adopt the simpler notation φ_γ without superscript whenever f and g are clear from context.

The name *forward-backward envelope* comes from the fact that $\varphi_\gamma(x)$ is the value of the minimization problem that defines the *forward-backward* step and alludes to the kinship that it has with the Moreau *envelope*. These claims will be addressed more in detail in Section 4. When f is sufficiently smooth and both f and g are convex, the FBE was shown to be continuously differentiable and amenable to be minimized with generalized Newton methods. More recently, [50] proposed a linesearch algorithm based on (L-)BFGS quasi-Newton directions for minimizing the FBE. The curvature information exploited by Newton-like methods acts as an online preconditioner, enabling superlinear rates of convergence, under some assumptions. However, unlike plain (F)FBS schemes, such methods require accessing second-order information of the smooth term f (needed for the evaluation of $\nabla\varphi_\gamma$), and are well defined only as long as the nonsmooth term g is convex. On the contrary, FBS only requires first-order information on f and *prox-boundedness* of g, in which case all accumulation points are *stationary* for φ, *i.e.*, they satisfy the first order necessary conditions [5].

Contributions. In this paper we propose ZeroFPR, a nonmonotone linesearch algorithm that, to the best of our knowledge, is the first that (1) addresses the same range of problems as FBS, (2) requires the same black-box oracle as FBS (gradient of one function and *proximity operator* of the other), (3) yet achieves superlinear rates if some assumptions (only) at the limit point are met. Though related to minFBE algorithm [50], ZeroFPR is conceptually different, mainly because it is *gradient-free*, in the sense that it does not require the gradient of the FBE. Moreover,

- 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, such an analysis was previously made only for the proximal point algorithm [45], for a special case of the projected gradient method [7, 8] and for difference-of-convex minimization problems [40].
- The analysis of the FBE, previously studied only in the case of f being $C^2(\mathbb{R}^n)$ and g convex [50], is extended to f and g as in Assumption I. In particular, we discuss properties of f and g that ensure (1) continuous differentiabilty of the FBE around critical points, (2) (strict) twice differentiability at critical points, and (3) equivalence of strong local minimality for the original function and the FBE.
- Exploiting the investigated properties of the FBE and of critical points we prove that ZeroFPR with monotone linesearch converges (1) globally if φ_γ has the *Kurdyka-Łojasiewicz* property [33, 34, 27], and (2) superlinearly when quasi-Newton Broyden directions are employed, under additional requirements at the limit point.

Organization of the paper. In Section 2 we introduce some notation and list known facts about FBS. In Section 3 we define and explore notions of stationarity and criticality for the investigated problem and relate them with properties of the forward-backward operator. In Section 4 we extend the results of [50] about the fundamental properties of the FBE to the more general setting addressed in this paper; for the sake of readability, some of the proofs are deferred to Appendix A. Section 5 addresses the core contribution of the paper, ZeroFPR; although arbitrary directions can be chosen, we specialize the results on superlinear convergence to a quasi-Newton Broyden method so as to truely maintain the same black-box oracle as FBS. Some ancillary

results needed for the proofs are listed in Appendix B. Finally, Section 6 illustrates numerical results obtained with the proposed method.

2. Preliminaries.

2.1. Notation. The identity $n \times n$ matrix is denoted as I, and the extended real line as $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. The open and closed ball of radius $r \geq 0$ centered in $x \in \mathbb{R}^n$ is denoted as $\mathbf{B}(x; r)$ and $\overline{\mathbf{B}}(x; r)$, respectively. Given a set E and a sequence $(x^k)_{k\in\mathbb{N}}$, we write $(x^k)_{k\in\mathbb{N}} \subset E$ with the obvious meaning of $x^k \in E$ for all $k \in \mathbb{N}$. The (possibly empty) set of cluster points of $(x^k)_{k\in\mathbb{N}}$ is denoted as $\omega\big((x^k)_{k\in\mathbb{N}}\big)$, or simply as $\omega(x^k)$ whenever the indexing is clear from context. We say that $(x^k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ is *summable* if $\sum_{k\in\mathbb{N}} \|x^k\|$ is finite, and *square-summable* if $(\|x^k\|^2)_{k\in\mathbb{N}}$ is summable.

A function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *level-bounded* if for all $\alpha \in \mathbb{R}$ the *level-set* $\mathbf{lev}_{\leq\alpha} h \coloneqq \{x \in \mathbb{R}^n \mid h(x) \leq \alpha\}$ is bounded. Following the terminology of [49], we say that a function $f : \mathbb{R}^n \to \mathbb{R}$ is *strictly continuous at* $\bar{x}$ if $\mathbf{lim\,sup}_{\substack{y,z\to\bar{x}\\ y\neq z}} \frac{|f(y)-f(z)|}{\|y-z\|}$ is finite, and *strictly differentiable at* $\bar{x}$ if $\nabla f(\bar{x})$ exists and $\mathbf{lim}_{\substack{y,z\to\bar{x}\\ y\neq z}} \frac{f(y)-f(z)-\langle\nabla f(\bar{x}),y-z\rangle}{\|y-z\|} = 0$. The set of functions $\mathbb{R}^n \to \mathbb{R}$ with Lipschitz continuous gradient is denoted as $C^{1,1}(\mathbb{R}^n)$, and for $f \in C^{1,1}(\mathbb{R}^n)$ we write L_f to indicate the Lipschitz modulus of ∇f.

For a proper, closed function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$, a vector $v \in \partial g(x)$ is a *subgradient* of g at x, where the *subdifferential* $\partial g(x)$ is considered in the sense of [49, Def. 8.3]

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

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

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

We have $\partial\varphi(x) = \nabla f(x) + \partial g(x)$ and $\hat{\partial}\varphi(x) = \nabla f(x) + \hat{\partial} g(x)$ [49, Ex. 8.8(c)].

Given a parameter value $\gamma > 0$, the *Moreau envelope* function g^γ and the *proximal mapping* $\mathbf{prox}_{\gamma g}$ are defined by

$$g^\gamma(x) \coloneqq \inf_z \Big\{g(z) + \tfrac{1}{2\gamma}\|z - x\|^2\Big\}, \tag{2.1}$$

$$\mathbf{prox}_{\gamma g}(x) \coloneqq \operatorname*{argmin}_z \Big\{g(z) + \tfrac{1}{2\gamma}\|z - x\|^2\Big\}. \tag{2.2}$$

We now summarize some properties of g^γ and $\mathbf{prox}_{\gamma g}$; the interested reader is referred to [49] for a detailed discussion. A function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *prox-bounded* if there exists $\gamma > 0$ such that $g + \frac{1}{2\gamma}\|\cdot\|^2$ is bounded below on $\mathbb{R}^n$. The supremum of all such γ is the *threshold* γ_g *of prox-boundedness* for g. In particular, if g is convex or bounded below then $\gamma_g = \infty$. In general, for any $\gamma \in (0, \gamma_g)$ the proximal mapping $\mathbf{prox}_{\gamma g}$ is nonempty- and compact-valued, and the Moreau envelope g^γ finite [49, Thm. 1.25].

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 $\delta_S(x) = 0$ 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* $x \mapsto \mathbf{argmin}_{z\in S} \|z - x\|$. Proximal mappings can be seen as generalized projections, due to the relation $\mathbf{\Pi}_S = \mathbf{prox}_{\gamma\delta_S}$ for any $\gamma > 0$.

For a set-valued mapping $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ we let $\mathbf{gph}\, T = \{(x, y) \mid y \in T(x)\}$ denote its *graph*, $\mathbf{zer}\, T = \{x \in \mathbb{R}^n \mid 0 \in T(x)\}$ 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*.

2.2. Forward-backward iterations. Due to the quadratic upper bound

$$f(z) \le f(x) + \langle \nabla f(x), z - x\rangle + \tfrac{L_f}{2}\|z - x\|^2 \tag{2.3}$$

holding for all $x, z \in \mathbb{R}^n$ [11, Prop. A.24], for any $\gamma \in (0, 1/L_f)$ the function

$$\ell_\gamma^{f,g}(z; x) := f(x) + \langle \nabla f(x), z - x\rangle + \tfrac{1}{2\gamma}\|z - x\|^2 + g(z) \tag{2.4}$$

furnishes a majorization model for φ, in the sense that

- $\ell_\gamma^{f,g}(z; x) \ge \varphi(z)$ for all $x, z \in \mathbb{R}^n$, and
- $\ell_\gamma^{f,g}(x; x) = \varphi(x)$ for all $x \in \mathbb{R}^n$.

Given a point $x \in \mathbb{R}^n$, one iteration of *forward-backward splitting* (FBS) for problem (1.1) consists in the minimization of the majorizing function $\ell_\gamma^{f,g}$, namely, in selecting

$$x^+ \in T_\gamma^{f,g}(x) := \mathbf{argmin}_z \ell_\gamma^{f,g}(z; x), \tag{2.5}$$

where $\gamma \in \big(0, \mathbf{min}\{\gamma_g, 1/L_f\}\big)$ is the stepsize parameter. The (set-valued) *forward-backward operator* $T_\gamma^{f,g}$ can be equivalently expressed as

$$T_\gamma^{f,g}(x) = \mathbf{prox}_{\gamma g}(x - \gamma\nabla f(x)), \tag{2.6a}$$

which motivates the bound $\gamma < \gamma_g$ in (2.5) to ensure the existence of x^+ for any x. We also introduce the corresponding (set-valued) *forward-backward residual*, namely

$$R_\gamma^{f,g}(x) := \tfrac{1}{\gamma}\big(x - T_\gamma^{f,g}(x)\big). \tag{2.6b}$$

Whenever no ambiguity occurs, we will omit the superscript and write simply ℓ_γ, T_γ and R_γ in place of $\ell_\gamma^{f,g}$, $T_\gamma^{f,g}$ and $R_\gamma^{f,g}$, respectively.

The inclusion (2.5) emphasizes that FBS is a *majorization-minimization* algorithm (MM), a class of methods which has been thoroughly analyzed when the majorizing function is strongly convex in the first argument [14] (for ℓ_γ, this is the case when g is convex). MM algorithms are of interest whenever minimizing the surrogate function $\ell_\gamma(\,\cdot\,; x)$ is significantly easier than directly addressing the non structured minimization of φ. For FBS this translates into simplicity of $\mathbf{prox}_{\gamma g}$ and ∇f operations, cf. (2.6a). Under very mild assumptions FBS iterations (2.5) converge to a *critical point* (see §3) independently of the choice of x^+ in the set $T_\gamma(x)$ [5]. The key is the following *sufficient decrease* property, whose proof can be found in [15, Lem. 2].

Lemma 2.1 (Sufficient decrease)**.** *For any $\gamma \in (0, \gamma_g)$, $x \in \mathbb{R}^n$ and $\bar{x} \in T_\gamma(x)$ it holds that $\varphi(\bar{x}) \le \varphi(x) - \frac{1-\gamma L_f}{2\gamma}\|x - \bar{x}\|^2$.*

3. Stationary and critical points. Unless φ is convex, the *stationarity* condition $0 \in \hat{\partial}\varphi(x^\star)$ in problem (1.1) is only necessary for the optimality of $x^\star$ [49, Thm. 10.1]. In this section we define different concepts of (sub)optimality and show how they are related for generic functions $\varphi = f + g$ as in Assumption I.

Definition 3.1. *We say that a point $x^\star \in \mathbf{dom}\,\varphi$ is*

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

(ii) critical *if it is γ-critical for some $\gamma \in (0, \gamma_g)$, i.e., if $x^\star \in T_\gamma(x^\star)$;*

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

The notion of criticality was already discussed in [7, 8] 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.

In [40] it is defined as d-stationarity, although the analysis is limited to difference-of-convex minimization problems; more precisely, it addresses problem (1.1) for a concave piecewise smooth function f and a convex function g.

If g is convex, then $\gamma_g = \infty$ and we may talk of criticality without mention of γ: in this case, γ-criticality and stationarity are equivalent properties regardless of the value of γ. For more general functions g, instead, the value of γ plays a role in determining whether a point is γ-critical or not, which legitimizes the following definition.

Definition 3.2. *The* criticality threshold *is the function* $\Gamma^{f,g} : \mathbb{R}^n \to [0, \gamma_g]$

$$\Gamma^{f,g}(x) := \mathbf{sup}\big(\{\gamma > 0 \mid x \in T_\gamma^{f,g}(x)\} \cup \{0\}\big) \quad \text{for } x \in \mathbb{R}^n. \tag{3.1}$$

As usual, whenever f and g are clear from the context we simply write Γ in place of $\Gamma^{f,g}$. The bound $\Gamma \leq \gamma_g$ is due to the fact that $\mathbf{prox}_{\gamma g}$ (and consequently T_γ) is everywhere empty-valued for $\gamma > \gamma_g$. Considering also $\gamma = 0$ forces the set in the definition to be nonempty, and the lower-bound $\Gamma \geq 0$ in particular; more precisely, observe that, by definition, $\Gamma(x) > 0$ iff x is a critical point.

Example 3.3. Let us consider $\varphi = f + g$ for $f(x) = \frac{1}{2}x^2$ and $g = \delta_C$ where $C = \{\pm 1\}$. Clearly, $\gamma_g = +\infty$ (as g is lower-bounded), $L_f = 1$ and ± 1 are both (unique) optima. Since $\hat{\partial}\varphi(x) = \mathbb{R}$ for $x \in C$ and $\hat{\partial}\varphi$ is clearly empty elsewhere, all points in C are stationary. $\mathbf{prox}_{\gamma g}$ is the (set-valued) projection on C, therefore the forward-backward operator is $T_\gamma(x) = \mathbf{\Pi}_C((1-\gamma)x)$. We have

$$T_\gamma(-1) = \begin{cases} \{-1\} & \text{if } \gamma < 1 \\ \{\pm 1\} & \text{if } \gamma = 1 \\ \{1\} & \text{if } \gamma > 1 \end{cases} \quad \text{and} \quad T_\gamma(1) = \begin{cases} \{1\} & \text{if } \gamma < 1 \\ \{\pm 1\} & \text{if } \gamma = 1 \\ \{-1\} & \text{if } \gamma > 1. \end{cases}$$

In particular, $\Gamma(1) = \Gamma(-1) = 1$. □

We now list some properties of critical and optimal points which will be used to derive regularity properties of T_γ and g^γ.

Theorem 3.4 (Properties of critical points)**.** *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 $x^\star$ is critical, then it is γ-critical for all $\gamma \in (0, \Gamma(x^\star))$; moreover, $x^\star$ is also $\Gamma(x^\star)$-critical provided that $\Gamma(x^\star) < \gamma_g$;

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

Proof.

♠ 3.4*(i)*: by definition, $x^\star$ is γ-critical iff $\ell_\gamma(x^\star; x^\star) \leq \ell_\gamma(x; x^\star)$ for all x, *i.e.*, iff

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

By suitably rearranging, the claim readily follows.

♠ 3.4*(ii)*: since $x^\star$ is γ-critical, due to 3.4*(i)* apparently it is also γ'-critical for any $\gamma' \in (0, \gamma]$. From the definition (3.1) of the criticality threshold $\Gamma(x^\star)$, it then follows that $x^\star$ is γ-critical for any $\gamma \in (0, \Gamma(x^\star))$. Suppose now that $\Gamma(x^\star) < \gamma_g$. Then, due to 3.4*(i)* for all $\gamma \in (0, \Gamma(x^\star))$ we have

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

By 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 3.4*(i)*.

♠ 3.4*(iii)*: let $x^\star$ be a critical point, and let $x \in T_\gamma(x^\star)$ for some $\gamma < \Gamma(x^\star)$. Fix $\gamma' \in (\gamma, \Gamma(x^\star))$. From 3.4*(i)* and 3.4*(ii)* it then follows that

$$g(x) \geq g(x^\star) + \langle -\nabla f(x^\star), x - x^\star \rangle - \tfrac{1}{2\gamma'}\|x - x^\star\|^2. \tag{3.2}$$

Since $x, x^\star \in T_\gamma(x^\star)$, it holds that $\ell_\gamma(x^\star; x^\star) = \ell_\gamma(x; x^\star)$, *i.e.*,

$$g(x^\star) = \langle \nabla f(x^\star), x - x^\star \rangle + \tfrac{1}{2\gamma}\|x - x^\star\|^2 + g(x) \overset{(3.2)}{\geq} g(x^\star) + \big(\tfrac{1}{2\gamma} - \tfrac{1}{2\gamma'}\big)\|x - x^\star\|^2.$$

Since $\frac{1}{2\gamma} - \frac{1}{2\gamma'} > 0$, necessarily $x = x^\star$. □

The inequality in Theorem 3.4*(i)* can be rephrased as the fact that the vector $-\nabla f(\bar{x})$ is a "global" *proximal subgradient* for g at $\bar{x}$ as in [49, Def. 8.45], where "global" refers to the fact that δ can be taken $+\infty$ in the cited definition. An interesting consequence is that the definition of criticality depends solely on φ and *not* on the considered decomposition $f + g$; in fact, it is only the threshold Γ that depends on it. To see this, let $\tilde{f} = f - h$ and $\tilde{g} = g + h$ for some $h \in C^{1,1}(\mathbb{R}^n)$, and consider a point $x^\star$ which is γ-critical with respect to the decomposition $f + g$, *i.e.*, such that $x^\star \in T_\gamma^{f,g}(x^\star)$. Combining Theorem 3.4*(i)* with the quadratic bound (2.3) for h, we obtain

$$\tilde{g}(x) \geq \tilde{g}(x^\star) - \langle \nabla \tilde{f}(x^\star), x - x^\star \rangle - \tfrac{1}{2\frac{\gamma}{1+\gamma L_h}}\|x - x^\star\|^2 \quad \text{for all } x \in \mathbb{R}^n.$$

Again from the characterization of Theorem 3.4*(i)*, we deduce that $x^\star \in T_{\tilde{\gamma}}^{\tilde{f},\tilde{g}}(x^\star)$, where $\tilde{\gamma} = \frac{\gamma}{1+\gamma L_h}$. In particular, considering $h = -f$ we infer that a point $x^\star$ is critical iff $x^\star \in T_\gamma^{0,\varphi}(x^\star) = \mathbf{prox}_{\gamma\varphi}(x^\star)$ for some $\gamma > 0$, which legitimizes the notion of criticality without mentioning a specific decomposition.

In the next result we show that criticality is a halfway property between stationarity and optimality. In light of these relations we shall seek "suboptimal" solutions which we characterize as critical points.

Proposition 3.5 (Optimality, criticality, stationarity)**.** *Let* $\bar{\gamma} \coloneqq \mathbf{min}\{\gamma_g, 1/L_f\}$.

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

(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* $\gamma_g > 1/L_f$;

Proof.

♠ 3.5*(i)*: let $\gamma \in (0, \gamma_g)$ and $x \in \mathbf{fix}\, T_\gamma$. Since x minimizes $g + \frac{1}{2\gamma}\|\cdot - x + \gamma\nabla f(x)\|^2$, we have $0 \in \hat{\partial}\big[g + \frac{1}{2\gamma}\|\cdot - x + \gamma\nabla f(x)\|^2\big](x) = \hat{\partial} g(x) + \nabla f(x) = \hat{\partial}\varphi(x)$, where the first inclusion follows from [49, Thm. 10.1] and the equalities from [49, Thm. 8.8(c)]. This proves that x is stationary.

♠ 3.5*(ii)*: Fix $\gamma \in (0, \bar{\gamma})$, $x^\star \in \mathbf{argmin}\, \varphi$ and $y \in T_\gamma(x^\star)$. Necessarily $y = x^\star$, otherwise, due to Lem. 2.1, $\varphi(y)$ would contradict minimality of $\varphi(x^\star)$. Therefore, $x^\star$ is γ-critical and the claim follows from the arbitrarity of $\gamma \in (0, \bar{\gamma})$. □

As already seen in Example 3.3, the bound $\Gamma(x^\star) \geq \mathbf{min}\{\gamma_g, 1/L_f\}$ at optimal points in Proposition 3.5*(ii)* is tight, and clearly the implication *"optimality ⇒ criticality"* cannot be reversed (consider, *e.g.*, the point $x^\star = 0$ for $\varphi = \mathbf{cos}$). The next example shows that the other implication is also proper.

Example 3.6 (Stationarity ⇏ criticality)**.** Let $f(x) = \frac{1}{2}x^2$ and $g(x) = x^{5/3}$. We have $\gamma_g = +\infty$, $L_f = 1$, and for $x^\star = 0$ it holds that $\hat{\partial}\varphi(x^\star) = \{\nabla\varphi(x^\star)\} = \{0\}$. Thus, $x^\star$ is

stationary; however, $T_\gamma(x^\star) = \mathbf{prox}_{\gamma g}(0) = \{-(5\gamma/3)^3\}$, and in particular $x^\star \notin T_\gamma(x^\star)$ for any $\gamma > 0$, proving $x^\star$ to be non critical. ∎

4. Forward-backward envelope. The FBE (1.2) was introduced in [41] and further analyzed in [50, 32] in the case when g is convex. Under such assumption the FBE was shown to be continuously differentiable, which made it possible to derive minimization algorithms based on its gradient. In the general setting addressed in this paper the FBE might fail to be (continuously) differentiable, and as such we need to resort to methods that do not need first-order information of the FBE. This task will be addressed in Section 5 where Algorithm ZeroFPR will be proposed; other than being applicable to a wider range of problems, the proposed scheme is entirely based on the same oracle of forward-backward iterations, unlike the approaches in [41, 50, 32] which instead require the computation of $\nabla^2 f$. All this will be possible thanks to continuity properties of the FBE, and to the behavior of φ_γ at critical points. We now focus on its continuity, while the other property will be addressed shortly after in Theorem 4.4.

Remark 4.1 (Alternative expressions for φ_γ)**.** By expanding the square and rearranging the terms in the definition (1.2), φ_γ can equivalently be expressed as

$$\varphi_\gamma(x) = \inf_{z\in\mathbb{R}^n} \left\{ f(x) - \tfrac{\gamma}{2}\|\nabla f(x)\|^2 + g(z) + \tfrac{1}{2\gamma}\|z - x + \gamma\nabla f(x)\|^2 \right\}.$$

Comparing with (2.5), it is apparent that the set of minimizers z in the above expression coincides with $T_\gamma(x)$, the forward-backward operator at x. Moreover, taking out the constant term $f(x) - \frac{\gamma}{2}\|\nabla f(x)\|^2$ from the infimum we immediately obtain the following expression involving the Moreau envelope of g:

$$\varphi_\gamma(x) = f(x) - \tfrac{\gamma}{2}\|\nabla f(x)\|^2 + g^\gamma(x - \gamma\nabla f(x)). \tag{4.1}$$ ∎

Other than providing an explicit way of computing the FBE, (4.1) emphasizes how φ_γ inherits the regularity properties of the Moreau envelope of g. In particular, the next key property follows from the strict continuity of g^γ [49, Ex. 10.32].

Proposition 4.2 (Strict continuity of φ_γ)**.** *For any $\gamma \in (0, \gamma_g)$, the FBE φ_γ is a real-valued and strictly continuous function on $\mathbb{R}^n$.*

4.1. Connections with the Moreau envelope. For the special case $f = 0$, FBS iterations (2.5) reduce to the *proximal point algorithm* (PPA) $x^+ \in \mathbf{prox}_{\gamma\varphi}(x)$, first introduced in [35] for convex functions φ and later generalized for functions with convex majorizing surrogate $\ell_\gamma^{0,\varphi}(\,\cdot\,; x) = \varphi(\,\cdot\,) + \frac{1}{2\gamma}\|\,\cdot - x\|^2$, see *e.g.*, [26]. Similarly, the FBE reduces to the Moreau envelope $\varphi^\gamma = \varphi_\gamma^{0,\varphi}$. In fact, the FBE extends the connection between PPA and Moreau envelope

$$\varphi^\gamma(x) = \mathbf{min}_z \ell_\gamma^{0,\varphi}(z; x) \quad \leftrightarrow \quad \mathbf{prox}_{\gamma\varphi}(x) = \mathbf{argmin}_z \ell_\gamma^{0,\varphi}(z; x), \tag{4.2a}$$

holding for $f = 0$ in (2.4), to majorizing functions $\ell_\gamma^{f,g}$ with arbitrary $f \in C^{1,1}(\mathbb{R}^n)$

$$\varphi_\gamma(x) = \mathbf{min}_z \ell_\gamma^{f,g}(z; x) \quad \leftrightarrow \quad T_\gamma(x) = \mathbf{argmin}_z \ell_\gamma^{f,g}(z; x). \tag{4.2b}$$

In the next section we will see the fundamental qualitative similarities between the FBE and the Moreau envelope. Namely, for γ small enough both φ^γ and φ_γ are lower bounds for the original function φ with same minimizers and minimum; in particular the minimization of φ is equivalent to that of φ^γ or φ_γ. Similarly, the identity

$$\varphi(\bar{x}) = \varphi^\gamma(x) - \tfrac{1}{2\gamma}\|x - \bar{x}\|^2 \qquad \text{for } \bar{x} \in \mathbf{prox}_{\gamma\varphi}(x)$$

will be extended to the inequality

$$\varphi(\bar{x}) \le \varphi_\gamma(x) - \tfrac{1-\gamma L_f}{2\gamma}\|x-\bar{x}\|^2 \quad \text{for } \bar{x} \in T_\gamma(x).$$

4.2. Basic properties. We now provide bounds relating φ_γ to the original function φ that extend the well known inequalities involving the Moreau envelope.

Proposition 4.3. *Let $\gamma \in (0, \gamma_g)$ be fixed. Then*

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

(ii) $\varphi(\bar{x}) \le \varphi_\gamma(x) - \tfrac{1-\gamma L_f}{2\gamma}\|x-\bar{x}\|^2$ *for all* $x \in \mathbb{R}^n$ *and* $\bar{x} \in T_\gamma(x)$.

Proof. 4.3*(i)* is obvious from the definition of the FBE (consider $z = x$ in (1.2)). As to 4.3*(ii)*, since the set of minimizers in (1.2) is $T_\gamma(x)$ (cf. (4.2b)), (2.3) yields

$$\begin{aligned}\varphi_\gamma(x) &= f(x) + \langle \nabla f(x), \bar{x} - x\rangle + g(\bar{x}) + \tfrac{1}{2\gamma}\|x-\bar{x}\|^2 \\ &\ge f(\bar{x}) - {}^{L_f}\!/_2\|\bar{x}-x\|^2 + g(\bar{x}) + \tfrac{1}{2\gamma}\|x-\bar{x}\|^2 = \varphi(\bar{x}) + \tfrac{1-\gamma L_f}{2\gamma}\|x-\bar{x}\|^2. \qquad \square\end{aligned}$$

With respect to the inequalities holding for convex g treated in [50], the lower bound in Proposition 4.3 is weaker, while the upper bound unchanged. Regardless, an immediate consequence of the result is that the value of φ and φ_γ at critical points is the same, and minimizers and infima of the two functions coincide for γ small enough.

Theorem 4.4. *The following hold*

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

(ii) $\inf \varphi = \inf \varphi_\gamma$ *and* $\operatorname{argmin} \varphi = \operatorname{argmin} \varphi_\gamma$ *for all* $\gamma \in \big(0, \min\{{}^1\!/_{L_f}, \gamma_g\}\big)$.

The bound $\gamma < {}^1\!/_{L_f}$ in Theorem 4.4*(ii)* is tight even when f and g are convex, as the counterexample with $f(x) = \frac{1}{2}x^2$ and $g = \delta_{\mathbb{R}_+}$ shows (see [50, Ex. 2.4] for details).

Although we will address problem (1.1) by simply exploiting the *continuity* of the FBE, nevertheless φ_γ enjoys favorable properties which are key for the efficacy of the method which will be discussed in Section 5. Firstly, observe that, due to *strict* continuity, φ_γ is almost everywhere differentiable, as it follows from Rademacher's theorem. The same applies to the mapping $x \mapsto x - \gamma\nabla f(x)$, its Jacobian being

$$Q_\gamma(x) \coloneqq \mathrm{I} - \gamma\nabla^2 f(x) \tag{4.3}$$

which is symmetric wherever it exists [49, Cor. 13.42 and Prop. 13.34]. However, in order to show that the proposed method achieves fast convergence we need additional regularity properties, namely (strict) twice differentiability at critical points and continuous differentiability around. The rest of the section is dedicated to this task.

4.3. Prox-regularity and first-order properties. In the favorable case in which g is convex and $f \in C^2(\mathbb{R}^n)$, the FBE enjoys global continuous differentiability [50]. In our setting, *prox-regularity* acts as a surrogate of convexity; the interested reader is referred to [49, §13.F] for a detailed discussion.

Definition 4.5 (Prox-regularity)**.** *Function g 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 $g(x') \ge g(x) + \langle v, x' - x\rangle - \frac{\rho}{2}\|x'-x\|^2$.

Prox-regularity is a mild requirement enjoyed globally and for any subgradient by all convex functions, with $\varepsilon = +\infty$ and $\rho = 0$. 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 $x_0 + \gamma v_0$ [45]. To our purposes, when needed, prox-regularity of g will be required only at critical points $x^\star$, and only for the subgradient $-\nabla f(x^\star)$. Therefore, with a slight abuse of terminology we define prox-regularity *of* critical points as follows.

Definition 4.6 (Prox-regularity *of* critical points)**.** *We say that a critical point $x^\star$ is* prox-regular *if g is prox-regular at $x^\star$ for $-\nabla f(x^\star)$.*

Examples where a critical point fails to be prox-regular are of challenging construction; before illustrating a cumbersome such instance in Example 4.9, we first prove an important result that connects prox-regularity with first-order properties of the FBE.

Theorem 4.7 (Continuous differentiability of φ_γ)**.** *Suppose that f is of class C^2 around a prox-regular critical point $x^\star$. 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 strictly continuous, and in particular single-valued;

(ii) $\varphi_\gamma \in C^1$ with $\nabla\varphi_\gamma = Q_\gamma R_\gamma$, where Q_γ is as in (4.3).

Proof. For $\gamma' \in (\gamma, \Gamma(x^\star))$, using Thm.s 3.4*(i)* and 3.4*(iii)* we obtain that

$$g(x) \geq g(x^\star) - \langle \nabla f(x^\star), x - x^\star \rangle - \tfrac{1}{2\gamma'}\|x - x^\star\|^2 \qquad \forall x \in \mathbb{R}^n. \tag{4.4}$$

Replacing γ' with γ in the above expression, the inequality is strict for all $x \neq x^\star$. From [45, 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 strictly continuous and g^γ is of class C^{1+} with $\nabla g^\gamma(x) = \gamma^{-1}\big(x - \mathbf{prox}_{\gamma g}(x)\big)$ for all $x \in V$. Since f is C^2 around $x^\star$ and ∇f is continuous, by possibly narrowing $U_{x^\star}$ we may assume that $f \in C^2(U_{x^\star})$ and $x - \gamma\nabla f(x) \in V$ for all $x \in U_{x^\star}$. Part 4.7*(ii)* then follows from (4.1) and the chain rule of differentiation, and 4.7*(i)* from the fact that strict continuity is preserved by composition. ∎

When $f = 0$, Theorem 4.7 restates the known fact that if g is prox-regular at $x^\star$ for $0 \in \partial g(x^\star)$, then g^γ is continuously differentiable around $x^\star$ with $\nabla g^\gamma(x) = \frac{1}{\gamma}(x - \mathbf{prox}_{\gamma g}(x))$. Notice that the bound $\gamma < \Gamma(x^\star)$ is tight: in general, for $\gamma = \Gamma(x^\star)$ no continuity of T_γ nor continuous differentiability of φ_γ around $x^\star$ can be guaranteed. In fact, even when $x^\star$ is $\Gamma(x^\star)$-critical, T_γ might even fail to be single-valued and φ_γ differentiable at $x^\star$, as the following counterexample shows.

Example 4.8 (Necessity of $\gamma \neq \Gamma(x^\star)$ in first-order properties)**.** Consider $f = \frac{1}{2}x^2$ and $g = \delta_S$ where $S = \{0, 1\}$. Then, $L_f = 1$, $\gamma_g = +\infty$, $T_\gamma(x) = \mathbf{\Pi}_S((1-\gamma)x)$ and the FBE is $\varphi_\gamma(x) = \frac{1-\gamma}{2}\|x\|^2 + \frac{1}{2\gamma}\,\mathbf{dist}((1-\gamma)x, S)^2$. At the critical point $x = 1$, which satisfies $\Gamma(1) = 1/2$, g is prox-regular for any subgradient. 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 $T_\gamma(1) = \mathbf{\Pi}_S(1/2) = \{0, 1\}$. ∎

Example 4.9 (Prox-nonregularity of critical points)**.** Consider $\varphi = f + g$ where $f(x) = \frac{1}{2}x^2$, $g(x) = \delta_S(x)$ and $S = \{1/n \mid n \in \mathbb{N}_{\geq 1}\} \cup \{0\}$. For $x_0 = 0$ we have $\Gamma(x_0) = +\infty$, however g fails to be prox-regular at x_0 for $v_0 = 0 = -\nabla f(x_0)$. 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 $\mathbf{\Pi}_{\mathbf{gph}\, g}(P_n) = \{1/n, 1/n+1\}$. By considering a large n, P_n can be

made arbitrarily close to $(0, -1/\rho)$ and at the same time its projection(s) arbitrarily close to $(0,0)$. It follows that g cannot be prox-regular at 0 for 0, for otherwise such projections would be single-valued close enough to $(0,0)$ [45, Cor. 3.4 and Thm. 3.5]. 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 spike. ∎

To underline how unfortunate the situation depicted in Example 4.9 is, notice that adding a linear term λx to f for any $\lambda \neq 0$, yet leaving g unchanged, restores the desired 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 $\lambda \in (0, -\infty)$, while for any $\lambda < 0$ the point 0 is not critical.

4.4. Second-order properties. In this section we discuss sufficient conditions for twice-differentiability of the FBE at critical points. Additionally to prox-regularity, which is needed for local continuous differentiability, we will also need generalized second-order properties of g. The interested reader is referred to [49, §13] for an extensive discussion on *epi-differentiability*.

Assumption II. *With respect to a given critical point $x^\star$*

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

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

$$\mathrm{d}^2 g(x^\star|-\nabla f(x^\star))[d] = \langle d, Md\rangle + \delta_S(d), \quad \forall d \in \mathbb{R}^n \tag{4.5}$$

where $S \subseteq \mathbb{R}^n$ is a linear subspace and $M \in \mathbb{R}^{n\times n}$. Without loss of generality we take M symmetric, and such that $\mathbf{Im}(M) \subseteq S$ and $\mathbf{ker}(M) \supseteq S^\perp$.[1]

We say that the assumptions are "strictly" satisfied if the stronger conditions in parenthesis hold.

Twice epi-differentiability of g is a mild requirement, and cases where $\mathrm{d}^2 g$ is generalized quadratic are abundant [47, 48, 43, 44]. Moreover, prox-regular and C^2-partly smooth functions g (see [29, 19]) comprise a wide class of functions that strictly satisfy Assumption II*(ii)* at a critical point $x^\star$ provided that *strict complementarity* holds, namely if $-\nabla f(x^\star) \in \mathbf{relint}\,\partial g(x^\star)$. In fact, it follows from [19, Thm. 28] applied to the *tilted* function $\tilde{g} = g + \langle \nabla f(x^\star), \cdot\,\rangle$ (which is still C^2-partly smooth and prox-regular at $x^\star$ [29, Cor. 4.6], [49, Ex. 13.35]) that $\mathbf{prox}_{\gamma\tilde{g}}$ is continuously differentiable around $x^\star$ for γ small enough (in fact, for $\gamma < \Gamma(x^\star)$). From [42, Thm 4.1(g)] we then obtain that $\tilde{g}$ is strictly twice epi-differentiable at $x^\star$ with generalized quadratic second-order epiderivative, and the claim follows by *tilting* back to g.

We now show that the quite common properties required in Assumption II are all that is needed for ensuring first-order properties of the proximal mapping and second-order properties of the FBE at critical points. The result generalizes the one in [50] by allowing nonconvex functions g. Although the proof is quite similar, we include it for the sake of self-inclusiveness.

Theorem 4.10 (Twice differentiability of φ_γ)**.** *Suppose that Assumption II is (strictly) 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 symmetric and positive semidefinite Jacobian*

$$P_\gamma(x^\star) := J\,\mathbf{prox}_{\gamma g}(x^\star - \gamma\nabla f(x^\star)); \tag{4.6}$$

[1] This can indeed be done without loss of generality: if M and S satisfy (4.5), then it suffices to replace M with $M' = \frac{1}{2}\,\mathbf{\Pi}_S(M + M^\top)\,\mathbf{\Pi}_S$ to ensure the desired properties.

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

$$JR_\gamma(x^\star) = \tfrac{1}{\gamma}[\mathrm{I} - P_\gamma(x^\star)Q_\gamma(x^\star)], \tag{4.7}$$

where Q_γ *is as in* (4.3) *and* P_γ *as in* (4.6)*;*

(iii) φ_γ *is (strictly) twice differentiable at* $x^\star$ *with symmetric Hessian*

$$\nabla^2\varphi_\gamma(x^\star) = Q_\gamma(x^\star)JR_\gamma(x^\star). \tag{4.8}$$

Proof. See Appendix A. ∎

Again, when $f \equiv 0$ Theorem 4.10 covers the differentiability properties of the proximal mapping (and consequently the second-order properties of the Moreau envelope, due to the identity $\nabla g^\gamma(x) = \frac{1}{\gamma}(x - \mathbf{prox}_{\gamma g}(x))$) as discussed in [42].

We now provide a key result that links nonsingularity of the Jacobian of the forward-backward residual R_γ to strong (local) minimality for the original cost φ and for the FBE φ_γ, under the generalized second-order properties of Assumption II.

Theorem 4.11 (Conditions for strong local minimality)**.** *Suppose that Assumption II is satisfied with respect to a critical point* $x^\star$*, and let* $\gamma \in (0, \mathbf{min}\{\Gamma(x^\star), 1/L_f\})$*. 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) the (symmetric) matrix $\nabla^2\varphi_\gamma(x^\star)$ *is positive definite;*

(d) $x^\star$ *is a strong local minimum for* φ_γ*;*

(e) $x^\star$ *is a local minimum for* φ_γ *and* $JR_\gamma(x^\star)$ *is nonsingular.*

Proof. See Appendix A. ∎

5. ZeroFPR algorithm. The first algorithmic framework exploiting the FBE for solving composite minimization problems was studied in [41], and other schemes have been recently investigated in [50, 32]. All such methods tackle the problem by looking for a (local) minimizer of the FBE, exploiting the equivalence of (local) minimality for the original function φ and for the FBE φ_γ, for γ small enough. To do so, they all employ the concept of directions of descent, thus requiring the gradient of the FBE to be well defined everywhere. In the more general framework addressed in this paper, such basic requirement is not met, which is why we approach the problem from a different perspective. This leads to ZeroFPR, the first algorithm, to the best of our knowledge, that despite requiring only the black-box oracle of FBS and being suited for fully nonconvex problems it achieves superlinear convergence rates.

5.1. Overview. Instead of directly addressing the minimization of φ or φ_γ, we seek solutions of the following nonlinear inclusion (*generalized equation*)

$$\text{find} \quad x^\star \in \mathbb{R}^n \quad \text{such that} \quad 0 \in R_\gamma(x^\star). \tag{5.2}$$

By doing so we address the problem from the same perspective of FBS, that is, finding fixed points of the forward-backward operator T_γ or, equivalently, zeros of its residual R_γ. Despite R_γ might be quite irregular when g is nonconvex, it enjoys favorable properties at the very solutions to (5.2) — *i.e.*, at γ-critical points — starting from single-valuedness, cf. Theorem 3.4*(iii)*. If some assumptions are met, R_γ turns out to be continuous around and even differentiable at critical points (cf. Theorems 4.7 and

Algorithm ZeroFPR generalized forward-backward with nonmonotone linesearch

REQUIRE $\gamma \in (0, \mathbf{min}\{1/L_f, \gamma_g\})$, $\beta, p_{\min} \in (0,1)$, $\sigma \in (0, \gamma\frac{1-\gamma L_f}{2})$, $x^0 \in \mathbb{R}^n$.
INITIALIZE $\bar\Phi_0 = \varphi_\gamma(x^0)$, $k = 0$.

1: Select $\bar x^k \in T_\gamma(x^k)$ and set $r^k = \frac{1}{\gamma}(x^k - \bar x^k)$
2: **if** $\|r^k\| = 0$, **then** stop; **end if**
3: Select a direction $d^k \in \mathbb{R}^n$
4: Let $\tau_k \in \{\beta^m \mid m \in \mathbb{N}\}$ be the largest such that $x^{k+1} = \bar x^k + \tau_k d^k$ satisfies

$$\varphi_\gamma(x^{k+1}) \le \bar\Phi_k - \sigma\|r^k\|^2 \tag{5.1}$$

5: $\bar\Phi_{k+1} = (1-p_k)\bar\Phi_k + p_k\varphi_\gamma(x^{k+1})$ for some $p_k \in [p_{\min}, 1]$
$k \leftarrow k+1$ and go to step 1.

4.10), and as a consequence the *inclusion* problem (5.2) reduces to a well behaved system of *equations*, as opposed to *generalized equations*, when close to solutions.

This motivates addressing problem (5.2) with fast methods for nonlinear equations. Newton-like schemes are iterative methods that prescribe updates of the form

$$x^+ = x - HR_\gamma(x) \tag{5.3}$$

which essentially amount to selecting $H = H(x)$, a linear operator that ideally carries information of the geometry of R_γ around x, in the attempt to yield an optimal iterate x^+. For instance, when R_γ is sufficiently regular Newton method corresponds to selecting H as the inverse of an element of the generalized Jacobian of R_γ at x, enabling fast convergence when close to a solution under some assumptions. However, selecting H as in Newton method would require information additional to the forward-backward oracle T_γ, and as such it goes beyond the scope of the paper. For this reason we focus instead on *quasi-Newton* schemes, in which H are linear operators recursively defined with low-rank updates that satisfy the *(inverse) secant condition*

$$H^+ y = s, \quad \text{where} \quad s = x^+ - x \ \text{ and } \ y \in R_\gamma(x^+) - R_\gamma(x). \tag{5.4}$$

A famous result [21] states that, under some assumptions and starting sufficiently close to a solution $x^\star$, updates as in (5.3) are superlinearly convergent to $x^\star$ iff the *Dennis-Moré condition* holds, namely the limit $\frac{\|(H^{-1} - JR_\gamma(x^\star))s\|}{\|s\|} \to 0$, see also [22] for a thorough survey. More recently, in [23] the result was extended to generalized equations of the form $f(x) + G(x) \ni 0$, where f is smooth and G possibly set-valued. The study focuses on Josephy-Newton methods where the update x^+ is the solution of the inner problem $f(x) - Bx \in Bx^+ + G(x^+)$, where $B = H^{-1}$, which can be interpreted as a forward-backward step in the metric induced by B. In particular, differently from the proposed ZeroFPR, the method in [23] has the crucial limitation that, unless the operator B has a very particular structure, the *backward* step $(B + G)^{-1}$ may be prohibitely challenging. The same remark applies to proximal (quasi-) Newton-type methods, in which each iteration requires the computation of a scaled proximal gradient step, see [28] and the references therein.

5.1.1. Globalization strategy. Quasi-Newton schemes are extremely handy and widely used methods. However, it is well known that they are effective only when close enough to a solution and might even diverge otherwise. To cope with this crucial downside there comes the need of a globalization strategy; this is usually addressed

by means of a linesearch over a suitable *merit function* ψ, along directions of descent for ψ so as to ensure sufficient decrease for small enough stepsizes. Unfortunately, the potential choice $\psi(x) = \frac{1}{2}\|R_\gamma(x)\|^2$ is not regular enough for a 'direction of descent' to be everywhere defined. The proposed Algorithm ZeroFPR bypasses this limitation by exploiting the favorable properties of the FBE. In Theorem 5.10 we will see that ZeroFPR achieves superlinear convergence, provided that f and g enjoy some regularity requirements at the limit point and the directions satisfy a Dennis-Moré condition. However, regardless of whether or not any of such conditions is met, the algorithm has the same convergence guarantees of FBS (cf. Thm. 5.6).

ZeroFPR globalizes the convergence of any fast local method, and requires exactly the same oracle of FBS. Conceptually, the algorithm is really elementary; for simplicity, let us first consider the monotone case, *i.e.*, with $p_k \equiv 1$ so that $\bar\Phi_k = \varphi_\gamma(x^k)$ (cf. step 5). The following steps are executed for updating the iterate x^k:

1) first, at step 1 a nominal forward-backward call yields an element $\bar x^k \in T_\gamma(x^k)$ that decreases the value of φ_γ by at least $\gamma\frac{1-\gamma L_f}{2}\|r^k\|^2$ (Prop. 4.3*(i)*);
2) then, at step 3 an update direction d^k at $\bar x^k$ (*not* at x^k!) is selected;
3) because of the sufficient decrease $x^k \mapsto \bar x^k$ on φ_γ and the continuity of φ_γ, at step 4 a stepsize τ_k can be found with finite many backtrackings $\tau_k \leftarrow \beta\tau_k$ that ensures a decrease for φ_γ of at least $\sigma\|r^k\|^2$ in the update $x^k \mapsto \bar x^k + \tau_k d^k$, for any $\sigma < \gamma\frac{1-\gamma L_f}{2}$.

In order to reduce the number of backtrackings, $p_k < 1$ can be selected resulting in a *nonmonotone* linesearch. The sufficient decrease is enforced with respect to a parameter $\bar\Phi_k \geq \varphi_\gamma(x^k)$ (cf. Lem. 5.1), namely a convex combination of $\{\varphi_\gamma(x^i)\}_{i=0}^k$. For the sake of convergence, $(p_k)_{k\in\mathbb{N}}$ can be selected arbitrarily in $(0, 1]$ as long as it is bounded away from 0, hence the role of the user-set lower bound $p_{\min}$. Consequently, small values of σ and p_k concur in reducing conservatism in the linesearch by favoring larger stepsizes.

Lemma 5.1 (Nonmonotone linesearch globalization)**.** *For all $k \in \mathbb{N}$ the iterates generated by* ZeroFPR *satisfy*

$$\varphi_\gamma(\bar x^k) \leq \varphi(\bar x^k) \leq \varphi_\gamma(x^k) \leq \bar\Phi_k \tag{5.5}$$

and there exists $\bar\tau_k > 0$ such that

$$\varphi_\gamma(\bar x^k + \tau d^k) \leq \bar\Phi_k - \sigma\|r^k\|^2 \qquad \forall \tau \in [0, \bar\tau_k]. \tag{5.6}$$

In particular, the number of backtrackings at step 4 is finite.

Proof. The first two inequalities in (5.5) are due to Prop.s 4.3*(i)* and 4.3*(ii)*, respectively. Moreover,

$$\bar\Phi_{k+1} = (1-p_k)\bar\Phi_k + p_k\varphi_\gamma(x^{k+1}) \geq (1-p_k)\varphi_\gamma(x^{k+1}) + p_k\varphi_\gamma(x^{k+1}) = \varphi_\gamma(x^{k+1}),$$

where the inequality follows by the linesearch condition (5.1); this proves the last inequality in (5.5). As to (5.6), 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 $\varphi_\gamma(x_\varepsilon) > \varphi_\gamma(x^k) - \sigma\|r^k\|^2$. By taking the limit for $\varepsilon \to 0^+$, so that $x_\varepsilon \to \bar x^k$, we have

$$\varphi_\gamma(\bar x^k) = \lim_{\varepsilon\to 0^+} \varphi_\gamma(x_\varepsilon) \geq \varphi_\gamma(x^k) - \sigma\|r^k\|^2 \geq \varphi(\bar x^k) + \big(\gamma\tfrac{1-\gamma L_f}{2} - \sigma\big)\|r^k\|^2 > \varphi(\bar x^k)$$

which contradicts Prop. 4.3*(i)*. Here, the equality follows from the continuity of φ_γ (Prop. 4.2), the first inequality from the property of x_ε, the second one from Prop.

4.3*(ii)*, and the last one from the fact that $r^k \neq 0$ and $\gamma\frac{1-\gamma L_f}{2} > \sigma$. Therefore, there exists $\bar{\tau}_k > 0$ such that

$$\varphi_\gamma(\bar{x}^k + \tau d^k) \leq \varphi_\gamma(x^k) - \sigma\|r^k\|^2 \overset{(5.5)}{\leq} \bar{\Phi}_k - \sigma\|r^k\|^2 \qquad \text{for all } \tau \in [0, \bar{\tau}_k]. \qquad \square$$

The existence of $\bar{\tau}_k > 0$ as in (5.6) ensures that, for any direction d^k, a stepsize τ_k is found at step 3 with finitely many backtrackings. In Section 5.4 we will also see that ZeroFPR returns solutions of problem (5.2), and that convergence is superlinear when the directions are chosen according to a modified Broyden's quasi-Newton scheme, if some properties are satisfied at the limit point. In Section 6 we will then confirm the theoretical findings with numerical simulations, which however seem to agree that also BFGS is extremely well performing in practice, although not supported by the theory (of superlinear convergence). A tentative explanation of this fact will be hinted in the conclusive remarks. Before going into the theory, we briefly discuss how to compute such quasi-Newton directions.

5.1.2. Choice of the directions: quasi-Newton methods. As already emphasized, fast convergence of ZeroFPR will be obtained thanks to the employment of Newton-like directions d^k. Differently from the classical Newton-like step (5.3), when stepsize 1 is accepted, the update in ZeroFPR is of the form $x^+ = \bar{x} + d$ rather than $x^+ = x + d$, where $\bar{x}$ is an element of $T_\gamma(x)$. Consequently, d needs to be a Newton-like direction at $\bar{x}$, and *not* at x, namely

$$d^k = -H_k \bar{r}^k \quad \text{for some } \bar{r}^k \in R_\gamma(\bar{x}^k) \tag{5.7}$$

(as opposed to $\bar{r}^k \in R_\gamma(x^k)$).

Broyden's method. We consider a *modified* Broyden's scheme [46] that performs rank-one updates of the form

$$H_{k+1} = H_k + \frac{s_k - H_k y_k}{\langle s_k, (1/\vartheta_k - 1)s_k + H_k y_k\rangle} s_k^\top H_k \quad \text{with} \quad \begin{cases} s_k = x^{k+1} - \bar{x}^k \\ y_k = r^{k+1} - \bar{r}^k, \end{cases} \tag{5.8a}$$

for a sequence $(\vartheta_k)_{k\in\mathbb{N}} \subset (0, 2]$. The original Broyden formula [17] corresponds to selecting $\vartheta_k \equiv 1$, whereas for other values of ϑ_k the secant condition (5.4) is *drifted* to $H^+\tilde{y} = s$, where $\tilde{y} = (1-\vartheta)H^{-1}s + \vartheta y$. In particular, [46] suggests

$$\vartheta_k := \begin{cases} 1 & \text{if } |\gamma_k| \geq \bar{\vartheta} \\ \frac{1 - \mathbf{sgn}(\gamma_k)\bar{\vartheta}}{1-\gamma_k} & \text{if } |\gamma_k| < \bar{\vartheta} \end{cases} \qquad \text{where} \quad \gamma_k := \frac{\langle H_k y^k, s^k\rangle}{\|s^k\|^2} \tag{5.8b}$$

and $\bar{\vartheta} \in (0, 1)$ is a fixed parameter, with the convention that $\mathbf{sgn}\, 0 = 1$. Starting from an invertible matrix H_0, this specific selection ensures that all matrices H_k are invertible.

BFGS method. BFGS method consists of the following update rule for matrices H_k in (5.7): starting from a symmetric and positive definite H_0,

$$H_{k+1} = \big(I - \rho_k s_k y_k^\top\big) H_k \big(I - \rho_k y_k s_k^\top\big) + \rho_k s_k s_k^\top, \quad \rho_k = \begin{cases} \frac{1}{\langle s_k, y_k\rangle} & \text{if } \langle s_k, y_k\rangle > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{5.9}$$

with $s_k = x^{k+1} - \bar{x}^k$ and $y_k = r^{k+1} - \bar{r}^k$, see *e.g.*, [38, §6.1]. BFGS is the most popular quasi-Newton scheme; it is based on rank-two updates that enforce symmetry,

additionally to the secant condition. In fact, BFGS is guaranteed to satisfy the Dennis-Moré condition provided that the Jacobian of the nonlinear system at the limit point is symmetric [18]. Although this is not the case for $JR_\gamma(x^\star)$, we observed in practice that BFGS directions (5.9) perform extremely well.

Limited-memory variants. Ultimately, instead of storing and operating on dense $m \times m$ matrices, *limited-memory* variants of quasi-Newton schemes keep in memory only a few (usually 3 to 20) most recent pairs (s^k, y^k) implicitly representing the approximate inverse Jacobian. Their employment considerably reduces storage and computations over the full-memory counterparts, and as such they are the methods of choice for large-scale problems. The most popular limited-memory method is L-BFGS: based on BFGS, it efficiently computes matrix-vector products with the approximate inverse Jacobian using a *two-loop recursion* procedure [31, 37, 38].

5.2. Connections with other methods. The first algorithmic framework exploiting the FBE was studied in [41], where two semismooth Newton methods were analyzed for convex f and g with $f \in C^{2,1}(\mathbb{R}^n)$ (twice continuously differentiable with Lipschitz continuous gradient). A generalization of the scheme was then studied in [50] under less restrictive assumptions, with particular attention to quasi-Newton directions in place of semismooth Newton methods. The proposed algorithm interleaves descent steps over the FBE with forward-backward steps. The study [32] then analyzed global and linear convergence properties of a generic linesearch algorithmic framework for minimizing the FBE based on gradient-related directions, for analytic f and subanalytic, convex, and lower bounded g.

Though apparently closely related, the approach that we provide in this paper presents major conceptual differences from any of the ones above. Apart from the significantly less restrictive assumptions, the crucial distinction is that our method does not require the gradient of the FBE. 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. Moreover, not only does the method have the same worst-case convergence properties of FBS, but it also has a certificate of superlinear convergence if some mild requirements are met.

5.3. Main remarks. In this section we list a few observations that come in handy when implementing ZeroFPR.

Remark 5.2 (Adaptive variant when L_f is unknown)**.** In practice, no prior knowledge of the global Lipschitz constant L_f is required for ZeroFPR. In fact, replacing L_f with an initial estimate $L > 0$ and fixing a backtracking ratio $\alpha \in (0, 1)$, after step 2 the following instruction can be added:

2bis: **if** $f(\bar{x}^k) > f(x^k) - \langle \nabla f(x^k), x^k - \bar{x}^k \rangle + \frac{L}{2}\|x^k - \bar{x}^k\|^2$ **then**
$\gamma \leftarrow \alpha\gamma$, $L \leftarrow L/\alpha$, $\sigma \leftarrow \alpha\sigma$, $\bar{\Phi}_k \leftarrow \varphi_\gamma(x^k)$ and go to step 1.
end if

Whenever the quadratic bound (2.3) is violated with L in place of L_f, the estimated Lipschitz constant L is increased and γ decreased accordingly; as a consequence, the FBE φ_γ changes and the nonmonotone linesearch is restarted. Since replacing L_f with any $L \geq L_f$ still satisfies (2.3), it follows that L is incremented only a finite number of times. Therefore, there exists an iteration k_0 starting from which γ and σ are constant; in particular, all the results of the paper remain valid starting from iteration k_0, at latest. ∎

Remark 5.3 (Support for locally Lipschitz ∇f)**.** If $\mathbf{dom}\, g$ is bounded and, as it is

reasonable, the directions $(d^k)_{k\in\mathbb{N}}$ selected at step 3 do not diverge, then Assumption I*(i)* on f can be relaxed to ∇f being *locally* Lipschitz.

In fact, it follows from the definition of proximal mapping that $(\bar{x}^k)_{k\in\mathbb{N}} \subseteq \mathbf{dom}\, g$, and if the directions are bounded then there exists a compact domain $\Omega \supseteq \mathbf{dom}\, g$ such that $(x^k)_{k\in\mathbb{N}} \subseteq \Omega$. Then, all results of the paper apply by replacing L_f with $\mathbf{lip}_\Omega \nabla f$, the (finite) Lipschitz constant of ∇f on Ω. ▯

Remark 5.4 (Cost per iteration)**.** Evaluating φ_γ essentially amounts to one evaluation of T_γ; this is evident from the expression (4.1), together with the observation that $g^\gamma(x-\gamma\nabla f(x)) = g(\bar{x}) + \frac{1}{2\gamma}\|x-\gamma\nabla f(x)-\bar{x}\|^2$ for any $\bar{x} \in T_\gamma(x)$. Therefore, computing $\varphi_\gamma(\bar{x}^k + \tau_k d^k)$ at step 4 yields an element $\bar{x}^{k+1} \in T_\gamma(x^{k+1})$ required in step 1, since $x^{k+1} = \bar{x}^k + \tau_k d^k$ at every iteration. In general, one evaluation of T_γ per backtracking step is required. If the directions d^k are computed with Broyden or BFGS methods (5.8) and (5.9), then one additional evaluation of T_γ is required for retrieving d^k; in the best case of $\tau_k = 1$ being accepted, which asymptotically happens if some assumptions are met (cf. Thm. 5.11), the algorithm then requires exactly *two* evaluations of T_γ per iteration. ▯

Remark 5.5 (Extension of FBS)**.** Observe that by selecting $d^k \equiv 0$ the condition at step 4 is always statisfied with $\tau_k = 1$ (in fact, for any τ_k), since for any $\sigma < \frac{1-\gamma L_f}{2\gamma}$ it holds that $\varphi_\gamma(\bar{x}^k) \le \varphi_\gamma(x^k) - \frac{1-\gamma L_f}{2\gamma}\|x^k-\bar{x}^k\|^2 \le \bar{\Phi}_k - \sigma\|x^k-\bar{x}^k\|^2$, where the inequalities follow from Prop. 4.3*(ii)* and (5.5), respectively. ZeroFPR then reduces to the classical FBS algorithm (cf. (2.5)), as $x^{k+1} = \bar{x}^k + d^k = \bar{x}^k \in T_\gamma(x^k)$ for any k. ▯

5.4. Convergence results. In this section we analyze the properties of cluster points of the iterates generated by ZeroFPR. Specifically,

- every cluster point of $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$ solves problem (5.2) (Theorem 5.6);
- if the linesearch is (eventually) monotone and some assumptions are met, then global and linear convergence are achieved (Theorems 5.8 and 5.9);
- directions satisfying the Dennis-Moré condition, such as Broyden's, enable superlinear rates under mild assumptions (Theorems 5.10 and 5.11).

In what follows, in order to exclude the trivial case in which the optimality condition $r^k = 0$ is achieved in a finite number of iterations we assume $r^k \neq 0$ for all k's.

Theorem 5.6 (Criticality of cluster points)**.** *The following hold for the iterates generated by* ZeroFPR*:*

(i) $r^k \to 0$ *square-summably, and all cluster points of* $(x^k)_{k\in\mathbb{N}}$ *and* $(\bar{x}^k)_{k\in\mathbb{N}}$ *are critical; more precisely,* $\omega(x^k) = \omega(\bar{x}^k) \subseteq \mathbf{fix}\, T_\gamma$*;*

(ii) $(\varphi_\gamma(x^k))_{k\in\mathbb{N}}$ *converges to a (finite) value* $\varphi_\star$*, and so does* $(\varphi(\bar{x}^k))_{k\in\mathbb{N}}$ *if* $(x^k)_{k\in\mathbb{N}}$ *is bounded.*

Proof.

♠ 5.6*(i)*: For all iterates k we have

$$\bar{\Phi}_{k+1} = (1-p_k)\bar{\Phi}_k + p_k\varphi_\gamma(x^{k+1}) \overset{(5.1)}{\le} \bar{\Phi}_k - \sigma p_k\|r^k\|^2 \le \bar{\Phi}_k - \sigma p_{\min}\|r^k\|^2. \tag{5.10}$$

By telescoping the above inequality and using (5.5), we obtain

$$\bar{\Phi}_k - \mathbf{inf}\,\varphi \ge \bar{\Phi}_0 - \bar{\Phi}_{k+1} = \textstyle\sum_{i=0}^k \left[\bar{\Phi}_i - \bar{\Phi}_{i+1}\right] \ge \sigma p_{\min} \sum_{i=0}^k \|r^i\|^2, \tag{5.11}$$

proving $r^k \to 0$ square-summably. Now, let $K \subseteq \mathbb{N}$ be such that $(x^k)_{k\in K} \to x'$ for some $x' \in \mathbb{R}^n$. Then, since $\|\bar{x}^k - x^k\| = \gamma\|r^k\| \to 0$, in particular $(\bar{x}^k)_{k\in K} \to x'$ as well. Due to the arbitrarity of the cluster point x' it follows that $\omega(x^k) \subseteq \omega(\bar{x}^k)$, and

a similar reasoning proves the converse inclusion, hence $\omega(x^k) = \omega(\bar{x}^k)$. Moreover, we have $x^k \in \overline{\mathbf{B}}(\bar{x}^k; \gamma\|r^k\|) \subseteq \mathbf{prox}_{\gamma g}\big(x^k - \gamma\nabla f(x^k)\big) + \overline{\mathbf{B}}(0; \gamma\|r^k\|)$ and since $(x^k - \gamma\nabla f(x^k))_{k\in K} \to x' - \gamma\nabla f(x')$, from the outer semicontinuity of $\mathbf{prox}_{\gamma g}$ [49, Ex. 5.23(b)] it follows that $x' \in \mathbf{prox}_{\gamma g}(x' - \gamma\nabla f(x'))$, *i.e.*, $x' \in \mathbf{fix}\, T_\gamma$.

♠ 5.6*(ii)*: from (5.10) it follows that $(\bar{\Phi}_k)_{k\in\mathbb{N}}$ is decreasing, and in particular its limit exists, be it $\varphi_\star$. Notice that $\varphi_\star = \inf \bar{\Phi}_k \geq \inf \varphi > -\infty$, where the first inequality is due to (5.5). Therefore

$$0 \leftarrow \bar{\Phi}_k - \bar{\Phi}_{k+1} = p_k\big(\bar{\Phi}_k - \varphi_\gamma(x^{k+1})\big) \overset{(5.1)}{\geq} p_{\min}\sigma\|r^k\|^2 \geq 0,$$

and since $\bar{\Phi}_k$ converges to $\varphi_\star$, then so does $\varphi_\gamma(x^{k+1})$. If $(x^k)_{k\in\mathbb{N}}$ is bounded, then so is $(\bar{x}^k)_{k\in\mathbb{N}}$ due to compact-valuedness of $\mathbf{prox}_{\gamma g}$ [49, Thm. 1.25]. Due to local Lipschitz continuity of the FBE (Prop. 4.2) and boundedness of $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$, φ_γ is Lipschitz continuous on a compact set that contains the sequences with modulus, say, $L > 0$. Then,

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

where the first two inequalities follow from Prop. 4.3. This shows that $\varphi(\bar{x}^k) \to \varphi_\star$. □

5.4.1. Global and linear convergence. Due to (5.5) and the fact that the sequence $(\bar{\Phi}_k)_{k\in\mathbb{N}}$ is decreasing (cf. (5.10)), the iterates of ZeroFPR satisfy $\varphi(\bar{x}^k) \leq \bar{\Phi}_0 = \varphi(\bar{x}^0)$. As a consequence, a sufficient condition for ensuring that the sequence $(\bar{x}^k)_{k\in\mathbb{N}}$ does not diverge, and consequently nor does $(x^k)_{k\in\mathbb{N}}$ (provided that the sequence of directions $(d^k)_{k\in\mathbb{N}}$ is bounded), is that the level set $\mathbf{lev}_{\leq\varphi(\bar{x}^0)}\varphi$ is compact. In the adaptive variant discussed in Remark 5.2, this translates to boundedness of the level set $\mathbf{lev}_{\leq\varphi(\bar{x}^{k_0})}\varphi$, where k_0 denotes the iteration starting from which γ is constant. Since the iterate k_0 and the point $\bar{x}^{k_0}$ are unknown a priori, the sufficient condition needs be strengthened to φ having bounded level sets.

We now show that if φ_γ is well-behaved at cluster points, then the whole sequence generated by ZeroFPR is convergent. Good behavior involves the existence of a *desingularizing function*, that is, φ_γ needs to possess the *Kurdyka-Łojasiewicz* property, a standard requirement that we restate here for the reader's convenience.

Definition 5.7 (KL property)**.** *A proper and lower semicontinuous function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ has the Kurdyka-Łojasiewicz property (KL property) at $x^\star \in \mathbf{dom}\,\partial h$ if there exist a concave* desingularizing function *(or* KL function*) $\psi : [0, \eta] \to [0, +\infty)$ for some $\eta > 0$ and a neighborhood $U_{x^\star}$ of $x^\star$, 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. $h(x^\star) < h(x) < h(x^\star) + \eta$ it holds that

$$\psi'\big(h(x) - h(x^\star)\big)\,\mathbf{dist}\big(0, \partial h(x)\big) \geq 1. \tag{5.12}$$

The KL property is a mild requirement enjoyed by semi-algebraic functions and by subanalytic functions which are continuous on their domain [13, 12] see also [33, 34, 27]. Moreover, since semi-algebraic functions are closed under parametric minimization, from the expression (1.2) it is apparent that φ_γ is semi-algebraic provided that f and g are. More precisely, in all such cases the desingularizing function can be taken of the form $\psi(s) = \rho s^\theta$ for some $\rho > 0$ and $\theta \in (0, 1]$, in which case it is usually referred to as a Łojasiewicz function. This property has been extensively

exploited to provide convergence rates of optimization algorithms such as FBS, see [3, 4, 5, 15, 24, 39]. Further properties of f and g that ensure φ_γ to satisfy such requirement are discussed in [32].

We first show how the KL property on φ_γ ensures global convergence of the iterates of ZeroFPR if the linesearch is eventually monotone, *i.e.*, if $p_k = 1$ for k sufficiently large, and then show that linear convergence is attained when the KL function is actually a Łojasiewicz function with large enough exponent.

Theorem 5.8 (Global convergence (monotone LS))**.** *Consider the iterates generated by* ZeroFPR *with $p_k = 1$ for k's large enough, and with directions satisfying*

$$\|d^k\| \le D\|r^k\| \quad \text{for all } k \tag{5.13}$$

for some $D \ge 0$. Suppose that $(x^k)_{k\in\mathbb{N}}$ remains bounded, that φ_γ has the KL property on $\omega(x^k)$, and that every cluster point is prox-regular. If f is of class C^2 in a neighborhood of $\omega(x^k)$, then $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$ are convergent to (the same γ-critical point) $x^\star$, and the sequence of residuals $(r^k)_{k\in\mathbb{N}}$ is summable.

Proof. From Lem. B.2 we know that φ_γ is constant on the (nonempty) compact set $\omega(x^k)$. It then follows from [15, Lem. 6] that there exist $\eta, \varepsilon > 0$ and a *uniformized KL function*, namely a function ψ satisfying Def.s 5.7*(i)*, 5.7*(ii)* and 5.7*(iii)* for all $x^\star \in \omega(x^k)$ and x such that $\mathbf{dist}(x, \omega(x^k)) < \varepsilon$ and $\varphi(x^\star) < \varphi(x) < \varphi(x^\star) + \eta$. Let $\varphi_\star := \mathbf{lim}_{k\to\infty} \varphi_\gamma(x^k)$, which exists and is finite (cf. Thm. 5.6), and let $k_1 \in \mathbb{N}$ be such that $p_k = 1$ for all $k \ge k_1$. Then we have (cf. step 5 and (5.1))

$$\bar{\Phi}_k = \varphi_\gamma(x^k) \quad \text{and} \quad \varphi_\gamma(x^k) > \varphi_\gamma(x^{k+1}) > \varphi_\star \qquad \forall k \ge k_1. \tag{5.14}$$

By possibly restricting ε, from Thm. 4.7*(ii)* and since $\omega(x^k)$ is compact, it follows that φ_γ is differentiable in an ε-enlargement of $\omega(x^k)$. Since $(\varphi_\gamma(x^k))_{k\ge k_1}$ converges to $\varphi_\star$ strictly decreasing (cf. (5.14)), and since $\mathbf{dist}(x^k, \omega(x^k)) \to 0$ as shown in Lem. B.2, there exists $k_2 \ge k_1$ such that for all $k \ge k_2$ we have $\varphi_\star < \varphi_\gamma(x^k) < \varphi_\star + \eta$ and $\mathbf{dist}(x^k, \omega(x^k)) < \varepsilon$. For all such k, by Thm. 4.7*(ii)* we have $\nabla\varphi_\gamma(x^k) = Q_\gamma(x^k)R_\gamma(x^k) = \big[\mathrm{I} - \gamma\nabla^2 f(x^k)\big]r^k$ and the uniformized KL property yields

$$\psi'\big(\varphi_\gamma(x^k) - \varphi_\star\big) \ge \tfrac{1}{\|\nabla\varphi_\gamma(x^k)\|} \ge \tfrac{1}{(1+\gamma L_f)\|r^k\|}. \tag{5.15}$$

Let $\Delta_k := \psi\big(\varphi_\gamma(x^k) - \varphi_\star\big) > 0$. Then,

$$\begin{aligned}\Delta_k - \Delta_{k+1} &\ge \psi'\big(\varphi_\gamma(x^k) - \varphi_\star\big)\big(\varphi_\gamma(x^k) - \varphi_\gamma(x^{k+1})\big) \\ &\overset{(5.15)}{\ge} \frac{\varphi_\gamma(x^k) - \varphi_\gamma(x^{k+1})}{(1+\gamma L_f)\|r^k\|} \overset{(5.14)}{=} \frac{\bar{\Phi}_k - \bar{\Phi}_{k+1}}{(1+\gamma L_f)\|r^k\|} \overset{(5.10)}{\ge} \frac{\sigma p_{\min}}{1+\gamma L_f}\|r^k\|\end{aligned} \tag{5.16}$$

where the first inequality follows from the concavity of ψ, and the second uses the fact that $\psi' \ge 0$. Since $\varphi_\gamma(x^k) \to \varphi_\star$ and ψ is continuous, it follows that $\Delta_k \to \psi(0) = 0$. Hence, by telescoping the inequality it follows that $(\|r^k\|)_{k\in\mathbb{N}}$ is summable. In turn, due to Lem. B.1*(i)*, $(\|x^{k+1} - x^k\|)_{k\in\mathbb{N}}$ is also summable. We conclude that $(x^k)_{k\in\mathbb{N}}$ is a Cauchy sequence and as such it admits a limit. It follows from Thm. 5.6*(i)* that the limit point is also the limit of $(\bar{x}^k)_{k\in\mathbb{N}}$ and that it is γ-critical. ∎

Theorem 5.9 (Linear convergence (monotone LS))**.** *Consider the iterates generated by* ZeroFPR*. Suppose that the assumptions of Theorem 5.8 are satisfied, and that the KL function can be taken of the form $\psi(s) = \rho s^\theta$ for some $\theta \in [1/2, 1]$. Then, $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$ are R-linearly convergent.*

Proof. As shown in Thm. 5.8, both $(x^k)_{k\in\mathbb{N}}$ and $(\bar{x}^k)_{k\in\mathbb{N}}$ converge to the same (γ-critical) point, be it $x^\star$. Defining $B_k := \sum_{i\geq k}\|r^i\|$, from Lem.s B.1*(i)* and B.1*(ii)* we have

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

The proof now reduces to showing that $(B_k)_{k\in\mathbb{N}}$ converges with asymptotic Q-linear rate. Inequality (5.15) reads $\varphi_\gamma(x^k)-\varphi_\star \leq \big[(1+\gamma L_f)\rho\theta\|r^k\|\big]^{\frac{1}{1-\theta}}$, and since $r^k\to 0$ for large enough k, we have

$$\Delta_k := \psi\big(\varphi_\gamma(x^k)-\varphi_\star\big) = \rho[\varphi_\gamma(x^k)-\varphi_\star]^\theta \leq \rho\big[\overbrace{(1+\gamma L_f)\rho\theta\|r^k\|}^{<1 \text{ for large } k}\big]^{\frac{\theta}{1-\theta}} \leq \rho^2(1+\gamma L_f)\|r^k\|.$$

Therefore, eventually $\Delta_k<1$, and from (5.16) we get

$$B_k = \textstyle\sum_{i\geq k}\|r^i\| \overset{(5.16)}{\leq} \frac{1+\gamma L_f}{\sigma p_{\min}}\sum_{i\geq k}(\Delta_i-\Delta_{i+1}) \leq \frac{1+\gamma L_f}{\sigma p_{\min}}\Delta_k \leq \frac{\rho^2(1+\gamma L_f)^2}{\sigma p_{\min}}\|r^k\| = C\|r^k\|$$

for some $C>0$. Thus, $B_k \leq C\|r^k\| = C(B_k - B_{k+1})$ for any k large enough *i.e.*, $B_{k+1}\leq(1-1/C)B_k$, proving the sought asymptotic Q-linear convergence of B_k. ∎

5.4.2. Superlinear convergence. In the next result we provide sufficient conditions ensuring that, if the directions satisfy a Dennis-Moré condition, ZeroFPR achieves superlinear convergence rates. Then, we show that the Broyden scheme (5.8) produces directions that satisfy such condition, and that due to the acceptance of unit stepsize $\tau_k=1$, eventually each iteration of ZeroFPR will require only two evaluations of T_γ (cf. Rem. 5.4). We remind that a sequence $(x^k)_{k\in\mathbb{N}}$ such that $x^k\neq x^\star$ for all k is said to be *superlinearly convergent* to $x^\star$ if $\|x^{k+1}-x^\star\|/\|x^k-x^\star\|\to 0$ as $k\to\infty$.

Theorem 5.10 (Superlinear convergence under Dennis-Moré condition)**.** *Suppose that Assumption II is strictly satisfied at a strong local minimum $x^\star$ of φ, and consider the iterates generated by ZeroFPR. Suppose that $(x^k)_{k\in\mathbb{N}}$ converges to $x^\star$ and that the directions $(d^k)_{k\in\mathbb{N}}$ satisfy the Dennis-Moré condition*

$$\lim_{k\to\infty}\frac{\|\bar{r}^k + JR_\gamma(x^\star)d^k\|}{\|d^k\|} = 0 \qquad \text{where } \bar{r}^k\in R_\gamma(\bar{x}^k). \tag{5.17}$$

Then, eventually the stepsize $\tau_k=1$ is always accepted and the sequences $(x^k)_{k\in\mathbb{N}}$, $(\bar{x}^k)_{k\in\mathbb{N}}$, and $(r^k)_{k\in\mathbb{N}}$, converge with superlinear rate.

Proof. From Thm.s 4.10*(ii)*, 4.10*(iii)*, 4.7 and 4.11 we know that $\nabla\varphi_\gamma$ and R_γ are strictly differentiable at $x^\star$, with $G_\star := \nabla^2\varphi_\gamma(x^\star) = Q_\gamma(x^\star)JR_\gamma(x^\star)\succ 0$, and that there exists a neighborhood $U_{x^\star}$ of $x^\star$ in which φ_γ is differentiable and R_γ Lipschitz continuous with modulus, say, L_R. Since $\bar{x}^k = x^k-\gamma r^k\to x^\star$ due to Thm. 5.6*(i)*, it holds that $x^k,\bar{x}^k\in U_{x^\star}$ for any k large enough. By single-valuedness of R_γ, for all such k we may write $R_\gamma(x^k)$ and $R_\gamma(\bar{x}^k)$ in place of r^k and $\bar{r}^k$, respectively. In particular, since $x^\star\in\mathbf{fix}\,T_\gamma$ (cf. Thm. 5.6*(i)*), necessarily $R_\gamma(\bar{x}^k)\to 0$. In turn, due to (5.17), it also holds that $d^k\to 0$. Let $x_0^{k+1} := \bar{x}^k + d^k$; then, from (5.17) we have

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

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

$$\lim_{k\to\infty}\|R_\gamma(x_0^{k+1})\|/\|x_0^{k+1}-\bar{x}^k\| = 0. \tag{5.18}$$

By possibly restricting $U_{x^\star}$, nonsingularity of $JR_\gamma(x^\star)$ ensures the existence of a constant $\alpha > 0$ such that $\|R_\gamma(x)\| \geq \alpha\|x - x^\star\|$ for all $x \in U_{x^\star}$. Since $\bar{x}^k + d^k \to x^\star$, eventually $x_0^{k+1} \in U_{x^\star}$. From (5.18) we obtain

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

which implies

$$\lim_{k\to\infty} \frac{\|x_0^{k+1} - x^\star\|}{\|\bar{x}^k - x^\star\|} = \lim_{k\to\infty} \frac{\|\bar{x}^k + d^k - x^\star\|}{\|\bar{x}^k - x^\star\|} = 0. \tag{5.19}$$

A second-order expansion of φ_γ at $x^\star$ yields

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

and

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

where the last equality uses the inclusion $\|\bar{x}^k + d^k - x^\star\| \in o(\|\bar{x}^k - x^\star\|)$ (which follows from (5.19)). Substracting,

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

where $\beta = \frac{1}{2}\lambda_{\min}(G_\star) > 0$. Hence there exists $k_0 \in \mathbb{N}$ such that $\varphi_\gamma(\bar{x}^k + d^k) \leq \varphi_\gamma(\bar{x}^k)$ for all $k \geq k_0$; in particular, for all $k \geq k_0$

$$\varphi_\gamma(\bar{x}^k + d^k) \leq \varphi_\gamma(\bar{x}^k) \leq \varphi_\gamma(x^k) - \gamma\tfrac{1-\gamma L_f}{2}\|r^k\|^2 \leq \bar{\Phi}_k - \sigma\|r^k\|^2,$$

where the second inequality follows from Prop. 4.3*(ii)*, and the last one from (5.5) and the fact that $\sigma < \gamma\frac{1-\gamma L_f}{2}$. Therefore, for $k \geq k_0$ the linesearch condition (5.1) holds with $\tau_k = 1$, and unitary stepsize is always accepted. In particular, the limit (5.19) reads $\mathbf{lim}_{k\to\infty} \|x^{k+1} - x^\star\|/\|\bar{x}^k - x^\star\| = 0$, and from the inequality

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

superlinear convergence of $(x^k)_{k\in\mathbb{N}}$ follows. Since $\|r^k\| = \|R_\gamma(x^k) - R_\gamma(x^\star)\| \leq L_R\|x^k - x^\star\|$, also the sequence $(r^k)_{k\in\mathbb{N}}$ converges superlinearly; in turn, since $\|\bar{x}^k - x^\star\| \leq \gamma\|r^k\| + \|x^k - x^\star\|$, so does the sequence $(\bar{x}^k)_{k\in\mathbb{N}}$. □

We conclude the section showing that employing Broyden directions (5.8) in ZeroFPR enables superlinear convergence rates, provided that R_γ is Lipschitz continuously *semidifferentiable* at the limit point (see [25]).

Theorem 5.11 (Superlinear convergence with Broyden directions)**.** *Suppose that Assumption II is strictly satisfied at a strong local minimum $x^\star$ of φ at which R_γ is Lipschitz-continuously semidifferentiable. Consider the iterates generated by ZeroFPR with directions d^k selected with Broyden method (5.8), and suppose that $x^k \to x^\star$.*

Then, the Dennis-Moré condition (5.17) is satisfied, and in particular all the claims of Theorem 5.10 hold.

Proof. It follows from the assumptions and Thm. 4.10 that R_γ is strictly differentiable at $x^\star$, and Lipschitz-continuously semidifferentiable there. Denoting $G_\star = JR_\gamma(x_\star)$,

$$\frac{\|y^k - G_\star s^k\|}{\|s^k\|} = \frac{\|R_\gamma(x^{k+1}) - R_\gamma(\bar{x}^k) - G_\star(x^{k+1} - \bar{x}^k)\|}{\|x^{k+1} - \bar{x}^k\|},$$

and since $x^k, \bar{x}^k \to x^\star$, due to [25, Lem. 2.2] there exists $L > 0$ such that $\frac{\|y^k - G_\star s^k\|}{\|s^k\|} \leq L \operatorname{\mathbf{max}}\{\|x^{k+1} - x^\star\|, \|\bar{x}^k - x^\star\|\}$ for k large enough. Consequently, due to Thm. 5.9 and Lem. B.3, $\frac{\|y_k - G_\star s_k\|}{\|s_k\|}$ is summable. Let $E_k = B_k - G_\star$ and let $\|\cdot\|_F$ denote the Frobenius norm. With a simple modification of the proofs of [25, Thm. 4.1] and [2, Lem. 4.4] that takes into account the scalar $\vartheta_k \in [\bar{\vartheta}, 2 - \bar{\vartheta}]$ we obtain

$$\|E_{k+1}\|_F \leq \left\|E_k\left(\mathrm{I} - \vartheta_k \tfrac{s_k (s_k)^\top}{\|s_k\|^2}\right)\right\|_F + \vartheta_k \frac{\|y_k - G_\star s_k\|}{\|s_k\|} \leq \|E_k\|_F - \frac{\bar{\vartheta}(2-\bar{\vartheta})}{2\|E_k\|_F} \frac{\|E_k s_k\|^2}{\|s_k\|^2}.$$

Consequently, $(\|E_k\|_F)_{k\in\mathbb{N}}$ is decreasing, and in particular $\bar{E} \coloneqq \operatorname{\mathbf{sup}}(\|E_k\|_F)_{k\in\mathbb{N}}$ is finite. By rearranging the inequality above we obtain

$$\frac{\bar{\vartheta}(2-\bar{\vartheta})}{2\bar{E}} \sum_{k\in\mathbb{N}} \frac{\|E_k s_k\|^2}{\|s_k\|^2} \leq \sum_{k\in\mathbb{N}} \frac{\bar{\vartheta}(2-\bar{\vartheta})}{2\|E_k\|_F} \frac{\|E_k s_k\|^2}{\|s_k\|^2} \leq \sum_{k\in\mathbb{N}} (\|E_k\|_F - \|E_{k+1}\|_F) \leq \|E_0\|_F.$$

Therefore, $(\frac{\|E_k s_k\|}{\|s_k\|})_{k\in\mathbb{N}} = (\frac{\|(B_k - G_\star)s_k\|}{\|s_k\|})_{k\in\mathbb{N}}$ is square summable, proving in particular the claimed Dennis-Moré condition (5.17). ∎

6. Simulations. We now present numerical results with the proposed method. In ZeroFPR we set $\beta = 1/2$, and for the nonmonotone linesearch 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 as in [52, 30].

We performed experiments with different choices of d^k in step 3. In particular,

- ZeroFPR(Broyden): $d^k = -H_k \bar{r}^k$, and H_k obtained by the Broyden method (5.8) with $\bar{\vartheta} = 10^{-4}$;
- ZeroFPR(BFGS): $d^k = -H_k \bar{r}^k$, where H_k is computed using BFGS updates (5.9);
- ZeroFPR(L-BFGS): d^k is computed using L-BFGS [38, Alg. 7.4] with memory 10.

We only show the results with full quasi-Newton updates (Broyden, BFGS) for one of the examples: for the other experiments we focus on L-BFGS, which is better suited for large-scale problems. Although JR_γ is nonsymmetric at the critical points in general, we observed that the symmetric updates of BFGS and L-BFGS perform very well in practice and outperform the Broyden method.

We compared ZeroFPR with the forward-backward splitting algorithm (denoted FBS), that is (2.5), the inertial FBS (denoted IFBS) proposed in [16, Eq. (7)] (with parameter $\beta = 0.2$), and the nonmonotone accelerated FBS (denoted AFBS) proposed in [30, Alg. 2] for fully nonconvex problems. All experiments were performed in MATLAB. The implementation of the methods used in the tests is available online.[2]

6.1. Nonconvex sparse approximation. Here we consider the problem of finding a sparse solution $x \in \mathbb{R}^n$ to a least-squares problem $Ax = b$, where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Sparsity can be induced by constraining or penalizing the ℓ_0 quasi-norm of x, namely the number of nonzero elements of x, but due to the challenges of

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

n	λ	FBS avg/max (s)	IFBS avg/max (s)	AFBS avg/max (s)	ZeroFPR(L-BFGS) avg/max (s)
500	0.10	0.141/0.405	0.159/0.449	0.135/0.221	0.037/0.088
	0.03	0.498/2.548	0.688/3.962	0.274/0.430	0.084/0.126
	0.01	1.305/5.445	1.721/4.942	0.570/1.157	0.152/0.560
1000	0.10	0.176/0.287	0.231/0.659	0.228/0.483	0.021/0.077
	0.03	0.576/2.756	0.645/4.165	0.382/0.841	0.091/0.275
	0.01	1.864/9.740	2.391/8.311	0.795/1.446	0.222/0.438
2000	0.10	0.291/0.599	0.392/0.719	0.393/0.640	0.025/0.055
	0.03	0.553/1.841	0.602/3.270	0.464/0.702	0.088/0.198
	0.01	2.108/10.934	2.439/8.010	0.979/1.411	0.271/0.464

TABLE 1

Nonconvex sparse approximation. Performance of FBS, IFBS, AFBS and ZeroFPR on problems with different values of n and λ. The table shows average and maximum CPU time required to reach $\|R_\gamma(x^k)\| \le 10^{-6}$ in 100 random experiments. Each algorithm was run on the same set of randomly generated problems, with $x^0 = 0$.

nonconvexity it is often the case that the ℓ_1 norm is used instead. As well explained and documented in [51], the use of the (square root of the) $\ell_{1/2}$ quasi-norm, namely $\|x\|_{1/2}^{1/2} = \sum_{i=1}^n |x_i|^{1/2}$, is in some sense optimal in trading-off representativeness of the solution and numerical simplicity of the ℓ_0 and ℓ_1 approaches, respectively. The problem then becomes

$$\underset{x\in\mathbb{R}^n}{\mathbf{minimize}} \ \tfrac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_{1/2}^{1/2}, \tag{6.1}$$

where $\lambda > 0$ is a regularization parameter. Function $\|x\|_{1/2}^{1/2}$ is separable, and its proximal mapping can be computed in closed form as follows, see [51, Thm. 1]:

$$\Big[\mathbf{prox}_{\gamma\|\cdot\|_{1/2}^{1/2}}(x)\Big]_i = \tfrac{2}{3}\Big(1+\cos\tfrac{2}{3}\Big(\pi - \arccos\tfrac{\gamma}{8}(|x_i|/3)^{-3/2}\Big)\Big)x_i, \quad i=1,\dots,n.$$

We ran numerical experiments consistently with the setting of [20, Sec. 8.2]. We considered different scenarios obtained by changing the regularization term λ and the size of A, keeping a constant column-to-row ratio of $n/m = 5$ for matrix A. Matrix A was generated with random Gaussian entries, with zero mean and variance $1/m$, while vector b was generated as $b = Ax_{\text{orig}} + v$ where $x_{\text{orig}} \in \mathbb{R}^n$ was randomly generated with $k = 5$ nonzero normally distributed entries, and $v \in \mathbb{R}^n$ is a noise vector with zero mean and variance $1/m$.

For each scenario, we solved 100 randomly generated problems and compared the performance of all algorithms in terms of CPU time to reach an accuracy of $\|r^k\| \le 10^{-6}$. For all algorithms and problems, we used $x^0 = 0$ as the starting iterate. Average and worst-case performance of the algorithms in each of the nine scenarios are illustrated in Table 1; apparently, ZeroFPR is significantly faster than FBS, IFBS and AFBS, even in a worst-case-to-average comparison.

Figure 1 shows the convergence rates of the algorithms in one of the generated problems. Since ZeroFPR employs a linesearch, and therefore the complexity of each iteration is unknown *a priori*, we recorded the number of matrix-vector products by A and $A^\top$ performed during the iterations, and displayed it on the horizontal axis. Apparently, ZeroFPR with Broyden's directions achieves superlinear convergence, beating the linear of FBS, IFBS and AFBS. This comparison also confirms what previously announced, namely the great performance of (L-)BFGS directions.

6.2. Dictionary learning. Expressing large data by means of only few elements from a collection of vectors is an important problem in machine learning and signal

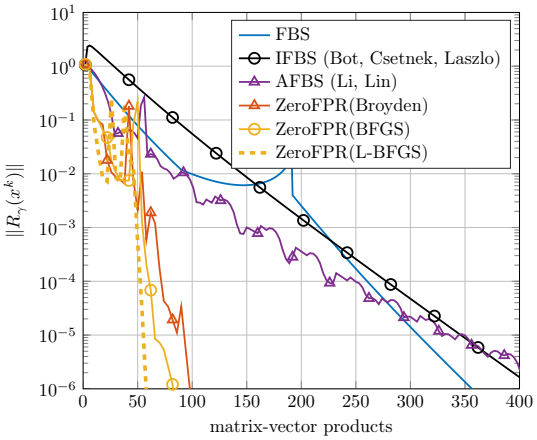

FIGURE 1. *Nonconvex sparse approximation. Convergence of fixed-point residual and cost in FBS, IFBS, AFBS and ZeroFPR, for different choices of the search directions and for* $n = 1500$, $\lambda = 0.03$.

processing. The challenge is finding such a collection of vectors, known as *dictionary*, that can accurately represent data signals in the sparsest way. In mathematical terms, given m signals $y_1, \dots, y_m \in \mathbb{R}^n$ we wish to find k *dictionary atoms* $d_1, \dots, d_k \in \mathbb{R}^n$ in such a way that each y_j can be represented, or accurately approximated, as a sparse linear combination of them. If we stack the data in a matrix $Y \in \mathbb{R}^{n\times m}$, and the dictionary atoms in a matrix $D \in \mathbb{R}^{n\times k}$ (to be found), the problem can be expressed as follows [1]

$$\underset{D,C}{\textbf{minimize}} \quad \tfrac{1}{2}\|Y - DC\|_F^2 \quad \textbf{subject to} \quad \begin{array}{ll} \|d_i\|_2 = 1 & i = 1, \dots, k, \\ \|c_j\|_0 \le N & j = 1, \dots, m, \\ \|c_j\|_\infty \le T & j = 1, \dots, m, \end{array} \tag{6.2}$$

where $C = [c_1, \dots, c_m] \in \mathbb{R}^{k\times m}$ is a matrix containing the sought coefficients, and $N \in \mathbb{N}$ and $T > 0$ are parameters. Differently from [1], we bound the set of feasible points by means of the ℓ_∞-norm constraint; this artificial constraint ensures that ∇f is globally Lipschitz continuous over the feasible domain (cf. Rem. 5.3). Moreover, we explicitly constrain the norm of the dictionary atoms: this causes no loss of generality, as the objective value of (6.2) is unchanged if the j-th atom d_j and the j-th row of C are scaled by reciprocal factors.

The problem can be expressed in the canonical form (1.1) by letting $f(D, C) = \frac{1}{2}\|Y - DC\|_F^2$ and $g(D, C) = \delta_S(D, C)$, where

$$S = \left\{ D \in \mathbb{R}^{n\times k} \mid \|d_j\|_2 = 1,\ j = 1 \dots k \right\} \times \left\{ C \in \mathbb{R}^{k\times m} \mid \begin{smallmatrix} \|c_j\|_0 \le N \\ \|c_j\|_\infty \le T, \end{smallmatrix} j = 1 \dots m \right\}$$

is the product of Euclidean spheres and box-constrained ℓ_0 balls. Both f and g are nonconvex in this case. The projection of (D, C) onto S is simple and column-wise separable: the columns d_j of D are scaled by their ℓ_2 norm, while the N largest coefficients (in absolute value) of the columns c_j of C are projected onto the box $[-T, T]$ and the other ones are set to zero, see *e.g.*, [7, Alg. 3 and Ex. 4.6].

We tested our algorithm on 50 problems with $N = 3$, $n = 20$, $m = 500$ and $k = 50$, for a total of 26000 variables each. We chose $T = 10^6$ as a large bound for ℓ_∞ norm of the columns of C. Problems were generated according to [1, §V.A]: first, a dictionary $D_{\text{gen}} \in \mathbb{R}^{20\times 50}$ was randomly generated with normal entries, and each column was normalized to one. Then, a matrix $C_{\text{gen}} \in \mathbb{R}^{50\times 500}$ was constructed with 3 normally distributed nonzero coefficients per column. Then we set $Y = C_{\text{gen}} D_{\text{gen}} + V$,

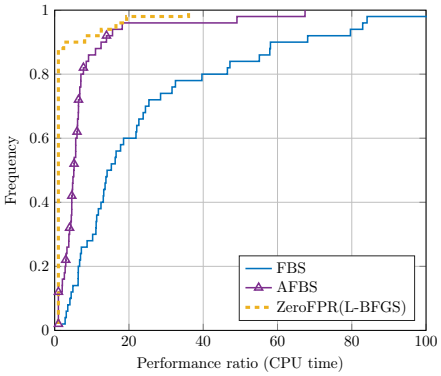

FIGURE 2. *Dictionary learning. Performance profiles of FBS, AFBS and* ZeroFPR(L-BFGS) *when applied to* 50 *randomly generated problems with* $n = 20$, $m = 500$, $k = 50$, $T = 10^6$ *and* $N = 3$. *The algorithms are executed until tolerance* $\|R_\gamma(x^k)\| \leq 10^{-4}$ *is reached. In the great majority of cases,* ZeroFPR(L-BFGS) *reaches a critical point significantly faster than FBS.*

where $V \in \mathbb{R}^{20\times 500}$ is a matrix with normally distributed entries with variance 10^{-2}.

We compared FBS, AFBS and ZeroFPR(L-BFGS), using the backtracking procedure discussed in Remark 5.2 to adaptively adjust the stepsize γ. IFBS could not be applied due to the lack of an adaptive stepsize-selection rule for the algorithm [16]. Moreover, we did not test ZeroFPR with Broyden and (full) BFGS directions because of the prohibitive overhead of storing and operating with 26000×26000 matrices.

Figure 2 shows the performance profile of the algorithms by comparing the time needed to reach an accuracy of $\|r^k\| \leq 10^{-4}$ starting from $(D^0, C^0) = (0, 0)$. In most of the cases, ZeroFPR(L-BFGS) exhibited a speedup of a factor 5-to-100 with respect to FBS, and 3-to-60 with respect to AFBS, at reaching a critical point.

7. 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 linesearch methods solving the system of nonlinear equations defining the stationary points of φ.

ZeroFPR 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 linesearch 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.

Numerical simulations with the proposed method on convex and nonconvex problems confirm our theoretical results. Using Broyden method, BFGS (in the case of small-scale problems) and L-BFGS (for large-scale problems) to compute directions in ZeroFPR greatly outperform FBS and 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.

Acknowledgements. We would like to express our sincere gratitude to the anonymous reviewers for their meticolous analysis and for the extremely constructive and insightful comments that significantly contributed to improving the paper.

REFERENCES

[1] M. AHARON, M. ELAD, AND A. BRUCKSTEIN, *K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation*, IEEE Transactions on signal processing, 54 (2006), pp. 4311–4322.

[2] F. J. Aragón Artacho, A. Belyakov, A. L. Dontchev, and M. López, *Local convergence of quasi-Newton methods under metric regularity*, Computational Optimization and Applications, 58 (2014), pp. 225–247, https://doi.org/10.1007/s10589-013-9615-y.
[3] H. Attouch and J. Bolte, *On the convergence of the proximal algorithm for nonsmooth functions involving analytic features*, Mathematical Programming, 116 (2009), pp. 5–16, https://doi.org/10.1007/s10107-007-0133-5.
[4] 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, 35 (2010), pp. 438–457.
[5] 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, 137 (2013), pp. 91–129, https://doi.org/10.1007/s10107-011-0484-9.
[6] H. Attouch and J. Peypouquet, *The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than* $1/k^2$, SIAM Journal on Optimization, 26 (2016), pp. 1824–1834, https://doi.org/10.1137/15M1046095.
[7] A. Beck and N. Hallak, *On the minimization over sparse symmetric sets: Projections, optimality conditions, and algorithms*, Math. Oper. Res., 41 (2016), pp. 196–223, https://doi.org/10.1287/moor.2015.0722.
[8] A. Beck and N. Hallak, *Proximal mapping for symmetric penalty and sparsity*, SIAM Journal on Optimization, 28 (2018), pp. 496–527, https://doi.org/10.1137/17M1116544.
[9] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 183–202.
[10] D. S. Bernstein, *Matrix mathematics: theory, facts, and formulas with application to linear systems theory*, Princeton University Press, Woodstock, 2009.
[11] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, 1995.
[12] J. Bochnak, M. Coste, and M.-F. Roy, *Real Algebraic Geometry*, Springer, 1998.
[13] J. Bolte, A. Daniilidis, and A. Lewis, *The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems*, SIAM Journal on Optimization, 17 (2007), pp. 1205–1223, https://doi.org/10.1137/050644641.
[14] J. Bolte and E. Pauwels, *Majorization-minimization procedures and convergence of SQP methods for semi-algebraic and tame programs*, Mathematics of Operations Research, 41 (2016), pp. 442–465, https://doi.org/10.1287/moor.2015.0735.
[15] J. Bolte, S. Sabach, and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Mathematical Programming, 146 (2014), pp. 459–494, https://doi.org/10.1007/s10107-013-0701-9.
[16] R. I. Boţ, E. R. Csetnek, and S. C. László, *An inertial forward-backward algorithm for the minimization of the sum of two nonconvex functions*, EURO Journal on Computational Optimization, 4 (2016), pp. 3–25.
[17] C. G. Broyden, *A class of methods for solving nonlinear simultaneous equations*, Mathematics of Computation, 19 (1965), pp. 577–593.
[18] R. H. Byrd and J. Nocedal, *A tool for the analysis of quasi-Newton methods with application to unconstrained minimization*, SIAM J. Numer. Anal., 26 (1989), pp. 727–739, https://doi.org/10.1137/0726042.
[19] A. Daniilidis, W. Hare, and J. Malick, *Geometrical interpretation of the predictor-corrector type algorithms in structured optimization problems*, Optimization, 55 (2006), pp. 481–503, https://doi.org/10.1080/02331930600815884.
[20] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, *Iteratively reweighted least squares minimization for sparse recovery*, Communications on Pure and Applied Mathematics, 63 (2010), pp. 1–38.
[21] J. E. Dennis and J. J. Moré, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Mathematics of computation, 28 (1974), pp. 549–560.
[22] J. E. J. Dennis and J. J. Moré, *Quasi-Newton methods, motivation and theory*, SIAM Review, 19 (1977), pp. 46–89, https://doi.org/10.1137/1019005.
[23] A. Dontchev, *Generalizations of the Dennis-Moré theorem*, SIAM Journal on Optimization, 22 (2012), pp. 821–830, https://doi.org/10.1137/110833567.
[24] P. Frankel, G. Garrigos, and J. Peypouquet, *Splitting methods with variable metric for Kurdyka-Łojasiewicz functions and general convergence rates*, Journal of Optimization Theory and Applications, 165 (2015), pp. 874–900, https://doi.org/10.1007/s10957-014-0642-3.
[25] C.-M. Ip and J. Kyparisis, *Local convergence of quasi-Newton methods for B-differentiable equations*, Mathematical Programming, 56 (1992), pp. 71–89.

[26] A. Kaplan and R. Tichatschke, *Proximal point methods and nonconvex optimization*, Journal of Global Optimization, 13 (1998), pp. 389–406, https://doi.org/10.1023/A:1008321423879.
[27] K. Kurdyka, *On gradients of functions definable in o-minimal structures*, Annales de l'institut Fourier, 48 (1998), pp. 769–783.
[28] J. D. Lee, Y. Sun, and M. A. Saunders, *Proximal Newton-type methods for minimizing composite functions*, SIAM Journal on Optimization, 24 (2014), pp. 1420–1443, https://doi.org/10.1137/130921428.
[29] A. S. Lewis, *Active sets, nonsmoothness, and sensitivity*, SIAM Journal on Optimization, 13 (2002), pp. 702–725, https://doi.org/10.1137/S1052623401387623.
[30] H. Li and Z. Lin, *Accelerated proximal gradient methods for nonconvex programming*, in Advances in neural information processing systems, 2015, pp. 379–387.
[31] D. C. Liu and J. Nocedal, *On the limited memory BFGS method for large scale optimization*, Mathematical Programming, 45 (1989), pp. 503–528.
[32] T. Liu and T. K. Pong, *Further properties of the forward-backward envelope with applications to difference-of-convex programming*, Computational Optimization and Applications, (2017), pp. 1–32, https://doi.org/10.1007/s10589-017-9900-2.
[33] S. Łojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels*, Les équations aux dérivées partielles, (1963), pp. 87–89.
[34] S. Łojasiewicz, *Sur la géométrie semi- et sous- analytique*, Annales de l'institut Fourier, 43 (1993), pp. 1575–1595.
[35] B. Martinet, *Brève communication. Régularisation d'inéquations variationnelles par approximations successives*, ESAIM: Modélisation Mathématique et Analyse Numérique, 4 (1970), pp. 154–158.
[36] Y. Nesterov, *Gradient methods for minimizing composite functions*, Mathematical Programming, 140 (2013), pp. 125–161.
[37] J. Nocedal, *Updating quasi-Newton matrices with limited storage*, Mathematics of computation, 35 (1980), pp. 773–782.
[38] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, New York, 2nd edition ed., Aug. 2006.
[39] P. Ochs, Y. Chen, T. Brox, and T. Pock, *iPiano: Inertial proximal algorithm for nonconvex optimization*, SIAM Journal on Imaging Sciences, 7 (2014), pp. 1388–1419, https://doi.org/10.1137/130942954.
[40] J.-S. Pang, M. Razaviyayn, and A. Alvarado, *Computing B-stationary points of nonsmooth DC programs*, Mathematics of Operations Research, 42 (2017), pp. 95–118, https://doi.org/10.1287/moor.2016.0795.
[41] P. Patrinos and A. Bemporad, *Proximal Newton methods for convex composite optimization*, in IEEE Conference on Decision and Control, 2013, pp. 2358–2363.
[42] R. Poliquin and R. Rockafellar, *Generalized Hessian properties of regularized nonsmooth functions*, SIAM Journal on Optimization, 6 (1996), pp. 1121–1137.
[43] R. A. Poliquin and R. T. Rockafellar, *Amenable functions in optimization*, Nonsmooth optimization: methods and applications, (1992), pp. 338–353.
[44] R. A. Poliquin and R. T. Rockafellar, *Second-order nonsmooth analysis in nonlinear programming*, Recent advances in nonsmooth optimization, (1995), pp. 322–349.
[45] R. A. Poliquin and R. T. Rockafellar, *Prox-regular functions in variational analysis*, Transactions of the American Mathematical Society, 348 (1996), pp. 1805–1838.
[46] M. Powell, *A hybrid method for nonlinear equations*, Numerical Methods for Nonlinear Algebraic Equations, (1970), pp. 87–144.
[47] R. T. Rockafellar, *First- and second-order epi-differentiability in nonlinear programming*, Transactions of the American Mathematical Society, 307 (1988), pp. 75–108.
[48] R. T. Rockafellar, *Second-order optimality conditions in nonlinear programming obtained by way of epi-derivatives*, Mathematics of Operations Research, 14 (1989), pp. 462–484.
[49] R. T. Rockafellar and R. J. Wets, *Variational analysis*, vol. 317, Springer, 2011.
[50] L. Stella, A. Themelis, and P. Patrinos, *Forward-backward quasi-Newton methods for nonsmooth optimization problems*, Computational Optimization and Applications, (2017), pp. 1–45, https://doi.org/10.1007/s10589-017-9912-y.
[51] Z. Xu, X. Chang, F. Xu, and H. Zhang, $l_{1/2}$ *regularization: a thresholding representation theory and a fast solver*, IEEE Transactions on neural networks and learning systems, 23 (2012), pp. 1013–1027.
[52] H. Zhang and W. W. Hager, *A nonmonotone line search technique and its application to unconstrained optimization*, SIAM Journal on Optimization, 14 (2004), pp. 1043–1056.

Appendix A. Proofs of Section 4.

Proof of Theorem 4.10. (Twice differentiability of φ_γ)
♠ 4.10*(i)*: It follows from [42, Thm.s 3.8 and 4.1] that $\mathbf{prox}_{\gamma g}$ is (strictly) differentiable at $x^\star - \gamma\nabla f(x^\star)$ iff g (strictly) satisfies Assumption II*(ii)*. Consequently, if f is of class C^2 around $x^\star$ (and in particular strictly differentiable at $x^\star$ [49, Cor. 9.19]), $R_\gamma(x) = x - \mathbf{prox}_{\gamma g}(x - \gamma\nabla f(x))$ is (strictly) differentiable at $x^\star$ with Jacobian as in (4.7) due to the chain rule of differentiation (and the fact that strict differentiability is preserved by composition). For $\gamma' \in (\gamma, \Gamma(x^\star))$ and $w \in \mathbb{R}^n$ we have

$$\mathrm{d}^2 g(x^\star|-\nabla f(x^\star))[w] = \liminf_{\substack{w'\to w\\ \tau\to 0^+}} \frac{g(x^\star+\tau w') - g(x^\star) + \tau\langle \nabla f(x^\star), w\rangle}{\tau^2/2} \overset{(4.4)}{\geq} -\tfrac{1}{\gamma'}\|w\|^2.$$

The expression (4.5) of the second-order epi-derivative then implies $\langle Mw, w\rangle \geq -\frac{1}{\gamma'}\|w\|^2$ for all $w \in \mathbb{R}^n$ (since $Mw = 0$ for $w \in S^\perp$). Therefore, $\lambda_{\min}(M) \geq -1/\gamma' > -1/\gamma$, proving $I + \gamma M$ to be positive definite, and in particular invertible.
The proof now is similar to that of [50, Lem. 2.9]. To obtain an expression for $P_\gamma(x^\star) = J\,\mathbf{prox}_{\gamma g}(x^\star - \gamma\nabla f(x^\star))$ we can apply [49, Ex. 13.45] to the function $g + \langle \nabla f(x^\star), \cdot\,\rangle$ so that, letting $\mathrm{d}^2 g = \mathrm{d}^2 g(x^\star|-\nabla f(x^\star))[\,\cdot\,]$ and $\mathbf{\Pi}_S$ the idempotent and symmetric projection matrix on S,

$$\begin{aligned} P_\gamma(x^\star)d &= \mathbf{prox}_{(\gamma/2)\mathrm{d}^2 g}(d) = \operatorname*{\mathbf{argmin}}_{d'\in S}\left\{\tfrac{1}{2}\langle d', Md'\rangle + \tfrac{1}{2\gamma}\|d'-d\|^2\right\} \\ &= \mathbf{\Pi}_S \operatorname*{\mathbf{argmin}}_{d'\in\mathbb{R}^n}\left\{\tfrac{1}{2}\langle \mathbf{\Pi}_S\, d', M\,\mathbf{\Pi}_S\, d'\rangle + \tfrac{1}{2\gamma}\|\mathbf{\Pi}_S\, d' - d\|^2\right\} \\ &= \mathbf{\Pi}_S\left(\mathbf{\Pi}_S[I+\gamma M]\,\mathbf{\Pi}_S\right)^\dagger \mathbf{\Pi}_S\, d \\ &= \mathbf{\Pi}_S[I+\gamma M]^{-1}\,\mathbf{\Pi}_S \qquad \text{(A.1)} \end{aligned}$$

where † indicates the pseudo-inverse, and last equality is due to [10, Facts 6.4.12(i)-(ii) and 6.1.6(xxxii)]. Apparently, $JP_\gamma(x^\star)$ is symmetric and positive semidefinite.
♠ 4.10*(ii)*: Since Q_γ is (strictly) continuous at $x^\star$ and R_γ is (strictly) differentiable at $x^\star$, from [50, Prop. 6.2] we have that $\nabla\varphi_\gamma = Q_\gamma R_\gamma$ is (strictly) differentiable at $x^\star$, and (4.7) follows by the chain rule.
♠ 4.10*(iii)*: A simple application of the chain rule proves (4.8); moreover, combined with (4.7) we obtain $\nabla^2\varphi_\gamma(x^\star) = \frac{1}{\gamma}[Q_\gamma(x^\star) - Q_\gamma(x^\star)P_\gamma(x^\star)Q_\gamma(x^\star)]$, and since both $Q_\gamma(x^\star)$ and $P_\gamma(x^\star)$ are symmetric, so is $\nabla^2\varphi(x^\star)$. □

Proof of Theorem 4.11. (Conditions for strong local minimality) We show that all conditions are equivalent to either one of the following

(f) $\langle d, (\nabla^2 f(x^\star) + M)d\rangle > 0\ \forall d \in S$, *where* M *and* S *are as in Assumption II;*
(g) $JR_\gamma(x^\star)$ *is similar to a symmetric and positive definite matrix.*

♠ 4.11*(c)* ⇔ 4.11*(d)*: trivial, since $\nabla^2\varphi_\gamma(x^\star)$ exists as shown in Thm. 4.10*(iii)*.
♠ 4.11*(a)* ⇔ 4.11*(f)*: follows from [49, Thm. 13.24(c)], since

$$\mathrm{d}^2\varphi(x^\star|0)[d] = \langle d, \nabla^2 f(x^\star)d\rangle + \mathrm{d}^2 g(x^\star|-\nabla f(x^\star))[d] = \langle d, (\nabla^2 f(x^\star) + M)d\rangle + \delta_S(d).$$

♠ 4.11*(c)* ⇔ 4.11*(e)*: if $\nabla^2\varphi_\gamma(x^\star) \succ 0$, then $x^\star$ is a (strong) local minimum for φ_γ and, due to (4.8), necessarily $JR_\gamma(x^\star)$ is invertible. Conversely, if $x^\star$ is a local minimum

for φ_γ, then $\nabla^2\varphi_\gamma(x^\star) \succeq 0$. If, additionally, $JR_\gamma(x^\star)$ is invertible, then due to (4.8) $\nabla^2\varphi_\gamma(x^\star)$ is also invertible, thus positive definite.

♠ 4.11*(c)* $\Leftrightarrow$ 4.11*(g)*: by comparing (4.7) and (4.8) we observe that $JR_\gamma(x^\star)$ is similar to $Q_\gamma(x^\star)^{-1/2}\nabla^2\varphi_\gamma(x^\star)Q_\gamma(x^\star)^{-1/2}$, which is positive definite iff so is $\nabla^2\varphi_\gamma(x^\star)$.

♠ 4.11*(f)* $\Leftrightarrow$ 4.11*(g)*: the proof is the same as that of [50, Thm. 2.11*(b)*$\Leftrightarrow$*(c)*].

♠ 4.11*(b)* $\Rightarrow$ 4.11*(g)*: with similar reasonings as in the proof of the implications "4.11*(a)* $\Leftrightarrow$ 4.11*(f)* $\Leftrightarrow$ 4.11*(g)*", we conclude that local minimality of $x^\star$ for φ entails $JR_\gamma(x^\star)$ being similar to a symmetric and positive *semi*definite matrix. Therefore, if $JR_\gamma(x^\star)$ is nonsingular, then it is similar to a symmetric and positive definite matrix.

♠ 4.11*(e)* $\Rightarrow$ 4.11*(b)*: trivial, since $\varphi_\gamma \le \varphi$ and $\varphi_\gamma(x^\star) = \varphi(x^\star)$ (cf. Prop. 4.3*(i)* and Thm. 4.4*(i)*). ∎

Appendix B. Additional results for Section 5.

Lemma B.1. *Consider the iterates generated by* ZeroFPR *and suppose that the directions $(d^k)_{k\in\mathbb{N}}$ are selected so as to satisfy (5.13). Then,*

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

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

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

Proof. For all k's we have

$$\|x^{k+1} - x^k\| = \|\bar{x}^k + \tau_k d^k - x^k\| = \|\tau_k d^k - \gamma r^k\| \le \gamma\|r^k\| + \tau_k\|d^k\| \le (\gamma + D)\|r^k\|$$

where in the last inequality we used the fact that $\tau_k \in (0, 1]$. This proves B.1*(i)*, and B.1*(ii)* trivially follows by the triangular inequality $\|\bar{x}^{k+1} - \bar{x}^k\| \le \|x^{k+1} - x^k\| + \gamma\|r^{k+1}\| + \gamma\|r^k\|$. Using this, B.1*(iii)* follows from Thm. 5.6*(i)*. ∎

Lemma B.2. *Consider the iterates generated by* ZeroFPR*. Suppose that (5.13) is satisfied and that the sequence $(x^k)_{k\in\mathbb{N}}$ is bounded. Then, $\omega(x^k) = \omega(\bar{x}^k)$ are nonempty compact and connected sets over which φ and φ_γ are constant and coincide. Moreover,*

$$\lim_{k\to\infty} \mathbf{dist}(x^k, \omega(x^k)) = \lim_{k\to\infty} \mathbf{dist}(\bar{x}^k, \omega(x^k)) = 0. \tag{B.1}$$

Proof. The sets of cluster points are nonempty because of boundedness of the sequences; in turn, connectedness and compactness as well as (B.1) are shown in [15, Rem. 5], which applies since $\|x^{k+1} - x^k\|$ and $\|\bar{x}^{k+1} - \bar{x}^k\|$ converge to 0 (cf. Lem. B.1*(iii)*). Moreover, since $(\varphi_\gamma(x^k))_{k\in\mathbb{N}}$ converges to some value $\varphi_\star \in \mathbb{R}$ and $\omega(x^k) = \omega(\bar{x}^k) \subseteq \mathbf{fix}\, T_\gamma$ as shown in Thm. 5.6, it follows Theorem 4.4*(i)* that φ and φ_γ coincide on $\omega(x^k)$ (and equal $\varphi_\star$). ∎

Lemma B.3. *Suppose that Assumption II is satisfied at a strong local minimum $x^\star$ of φ. Then, for any $\gamma \in (0, 1/L_f)$ the FBE φ_γ possesses the KL property at $x^\star$, and the desingularizing function ψ can be taken of the form $\psi(s) = \rho s^{1/2}$ for some $\rho > 0$.*

Proof. From Thm. 4.11*(c)* it follows that $x^\star$ is a strong local minimum for φ_γ at which φ_γ is twice differentiable with $H_\star := \nabla^2\varphi_\gamma(x^\star) \succ 0$. Let $\lambda := \lambda_{\min}(H_\star)$ and $\Lambda := \lambda_{\max}(H_\star)$. 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, for all $x \in U_{x^\star}$, $\varphi_\gamma(x) - \varphi_\gamma(x^\star) \le \frac{\Lambda}{4}\|x - x^\star\|^2$ and $\|\nabla\varphi_\gamma(x)\| \ge \frac{\lambda}{2}\|x - x^\star\|$, and in particular $\psi'(\varphi_\gamma(x) - \varphi_\gamma(x^\star))\|\nabla\varphi_\gamma(x)\| = \frac{\rho}{2\sqrt{\varphi_\gamma(x) - \varphi_\gamma(x^\star)}}\|\nabla\varphi_\gamma(x)\| \ge \frac{\rho\lambda}{2\sqrt{\Lambda}}$. Letting $\rho = \frac{2\sqrt{\Lambda}}{\lambda}$ we obtain that ψ is a KL function for φ_γ at $x^\star$. ∎