Forward-backward quasi-Newton methods for nonsmooth optimization problems

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

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

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

出版情報：Computational Optimization and Applications. 67 (3), pp.443-487, 2017-04-10. Springer
バージョン：
権利関係：

Noname manuscript No.
(will be inserted by the editor)

Forward-backward quasi-Newton methods for nonsmooth optimization problems

Lorenzo Stella · **Andreas Themelis** · **Panagiotis Patrinos**

Received: date / Accepted: date

Abstract The forward-backward splitting method (FBS) for minimizing a nonsmooth composite function can be interpreted as a (variable-metric) gradient method over a continuously differentiable function which we call forward-backward envelope (FBE). This allows to extend algorithms for smooth unconstrained optimization and apply them to nonsmooth (possibly constrained) problems. Since the FBE and its gradient can be computed by simply evaluating forward-backward steps, the resulting methods rely on the very same black-box oracle as FBS. We propose an algorithmic scheme that enjoys the same global convergence properties of FBS when the problem is convex, or when the objective function possesses the Kurdyka-Łojasiewicz property at its critical points. Moreover, when using quasi-Newton directions the proposed method achieves superlinear convergence provided that usual second-order sufficiency conditions on the FBE hold at the limit point of the generated sequence. Such conditions translate into milder requirements on the original function involving generalized second-order differentiability. We show that BFGS fits our framework and that the limited-memory variant L-BFGS is well suited for large-scale problems, greatly outperforming FBS or its accelerated version in practice, as well as ADMM and other problem-specific solvers. The analysis of superlinear convergence is based on an extension of the Dennis and Moré theorem for the proposed algorithmic scheme.

Keywords Nonsmooth optimization · Forward-backward splitting · Line-search methods · Quasi-Newton · Kurdyka-Łojasiewicz

This work was supported by the KU Leuven Research Council under BOF/STG-15-043.

Lorenzo Stella[1,2] ✉ lorenzo.stella@imtlucca.it
Andreas Themelis[1,2] ✉ andreas.themelis@imtlucca.it
Panagiotis Patrinos[1] ✉ panos.patrinos@esat.kuleuven.be

[1]KU Leuven, Department of Electrical Engineering (ESAT-STADIUS) & Optimization in Engineering Center (OPTEC), Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium.
[2]IMT School for Advanced Studies Lucca, Piazza San Francesco 19, 55100 Lucca, Italy.

1 Introduction

In this paper we focus on nonsmooth optimization problems over $\mathbb{R}^n$ of the form

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

where f is a smooth (possibly nonconvex) function, while g is a proper, closed, convex (possibly nonsmooth) function with cheaply computable proximal mapping [1]. Problems of this form appear in several application fields such as control, system identification, signal and image processing, machine learning and statistics.

Perhaps the most well known algorithm to solve problem (1.1) is the forward-backward splitting (FBS), also known as proximal gradient method [2, 3], which generalizes the classical gradient method to problems involving an additional nonsmooth term. Convergence of the iterates of FBS to a critical point of problem (1.1) has been shown, in the general nonconvex case, for functions φ having the Kurdyka-Łojasiewicz property [4–7]. This assumption was used to prove convergence of many other algorithms [7–11]. The global convergence rate of FBS is known to be sublinear of order $O(1/k)$ in the convex case, where k is the iteration count, and can be improved to $O(1/k^2)$ with techniques based on the work of Nesterov [12–15]. Therefore, FBS is usually efficient for computing solutions with small to medium precision only and, just like all first order methods, suffers from ill-conditioning of the problem at hand. A remedy to this is to add second-order information in the computation of the forward and backward steps, so to better scale the problem and achieve superlinear asymptotic convergence. As proposed by several authors [16–18], this can be done by computing the gradient steps and proximal steps according to the Q-norm rather than the Euclidean norm, where Q is the Hessian of f or some approximation to it. This approach has the severe limitation that, unless Q has a very particular structure, the backward step becomes now very hard and requires an inner iterative procedure to be computed.

In the present paper we follow a different approach. We define a function, which we call *forward-backward envelope* (FBE) that serves as a real-valued, continuously differentiable, exact penalty function for the original problem. Furthermore, forward-backward splitting is shown to be equivalent to a (variable-metric) gradient method applied to the problem of minimizing the FBE. The value and gradient of the FBE can be computed solely based on the evaluation of a forward-backward step at the point of interest. For these reasons, the FBE works as a surrogate of the Moreau envelope [1] for composite problems of the form (1.1). Most importantly, this opens up the possibility of using well known smooth unconstrained optimization algorithms, with faster asymptotic convergence properties than the gradient method, to minimize the FBE and thus solve (1.1), which is nonsmooth and possibly constrained. This approach was first explored in [19], where two Newton-type methods were proposed, and combines and extends ideas stemming from the literature on merit functions for *variational inequalities* (VIs) and *complementarity problems* (CPs), specifically the reformulation of a VI as a constrained continuously differentiable optimization problem via the regularized gap function [20] and as an unconstrained continuously differentiable optimization problem via the D-gap function [21] (see [22, §10] for a survey

and [23,24] for applications to constrained optimization and model predictive control of dynamical systems).

Then we propose an algorithmic scheme, based on line-search methods, to minimize the FBE. In particular, when descent steps are taken along quasi-Newton directions, superlinear convergence can be achieved when usual nonsingularity assumptions hold at the limit point of the sequence of iterates. The asymptotic analysis is based on an analogous of the Dennis and Moré theorem [25] for the proposed algorithmic scheme, and the BFGS quasi-Newton method is shown to fit this framework. Its limited memory variant L-BFGS, which is suited for large scale problems, is also analyzed. At the same time, we show that our algorithm enjoys the same global convergence properties of FBS under the same assumptions on the original function φ, despite our method operates on the FBE. Unlike the approaches of [16–18], our algorithm does not require the solution to any inner problem.

The contributions of this work can be summarized as follows:

- We give an interpretation of forward-backward splitting as a (variable-metric) gradient method over a C^1 function, the forward-backward envelope (FBE). We analyze the fundamental properties of the FBE, including second-order properties around the solutions to (1.1) under mild assumptions on g.
- We propose an algorithmic scheme for solving problem (1.1) based on line-search methods applied to the problem of minimizing the FBE, and prove that it converges globally to a critical point when φ is convex or has the Kurdyka-Łojasiewicz property. This is a crucial feature of our approach: in fact, the FBE is nonconvex in general, and there exist examples showing how classical line-search methods need not converge to critical points for nonconvex functions [26–29]. When φ is convex, in addition, global sublinear convergence of order $O(1/k)$ (in the objective value) is proved.
- We show that when the directions of choice satisfy the Dennis-Moré condition the method converges superlinearly, under appropriate assumptions, and illustrate when this is the case for BFGS. The resulting algorithm has the same global convergence properties as FBS but, despite relying on the same black-box oracle, converges much faster in practice.

The paper is organized as follows. Section 2 introduces the forward-backward envelope function and illustrates its properties. In Section 3 we propose our algorithmic scheme and prove its global convergence properties. Linear convergence is also discussed. Section 4 is devoted to the asymptotic convergence analysis in the particular case where quasi-Newton directions are used, specializing the results to the case of BFGS. Limited-memory directions are also discussed. Finally, Section 5 illustrates numerical results obtained with the proposed method. Some of the proofs are deferred to the Appendix for the sake of readability and, for the reader's convenience, Appendix A will list some definitions and known results on generalized differentiability which are needed in the analysis.

1.1 Notation and background

Throughout the paper, $\langle \cdot, \cdot \rangle$ is an inner product over $\mathbb{R}^n$ and $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$ is the induced norm. The set of continuously differentiable functions on $\mathbb{R}^n$ having L-Lipschitz continuous gradient (also refferred to as L-smooth) is denoted by $C_L^{1,1}(\mathbb{R}^n)$. We denote the extended real line as $\overline{\mathbb{R}} \equiv \mathbb{R} \cup \{+\infty\}$. The set of proper, closed, convex functions from $\mathbb{R}^n$ with values in $\overline{\mathbb{R}}$ is referred to as $\Gamma_0(\mathbb{R}^n)$. Given a function h on $\mathbb{R}^n$, the subdifferential $\partial h(x)$ of h at x is considered in the sense of [30, Def. 8.3], that is

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

where

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

This includes the ordinary gradient in the case of continuously differentiable functions, while for $g \in \Gamma_0(\mathbb{R}^n)$ it is equivalent to

$$\partial g(x) = \{ v \in \mathbb{R}^n \mid g(y) \geq g(x) + \langle v, y - x \rangle, \text{ for all } y \in \mathbb{R}^n \}.$$

We denote the set of *critical points* associated with problem (1.1) as

$$\operatorname{zer} \partial \varphi = \{ x \in \mathbb{R}^n \mid 0 \in \partial \varphi(x) \} = \{ x \in \mathbb{R}^n \mid -\nabla f(x) \in \partial g(x) \}.$$

The second equality is due to [30, Ex. 8.8]. A necessary condition for a point x to be a local minimizer for (1.1) is that $x \in \operatorname{zer} \partial \varphi$ [30, Thm. 10.1]. If φ is convex (for example when f is convex) then the condition is also sufficient, and x is a global minimizer.

Given $g \in \Gamma_0(\mathbb{R}^n)$, its *proximal mapping* is defined by

$$\operatorname{prox}_{\gamma g}(x) = \operatorname*{argmin}_{u \in \mathbb{R}^n} \left\{ g(u) + \tfrac{1}{2\gamma} \|u - x\|^2 \right\}, \tag{1.2}$$

cf. [1]. The proximal mapping is a generalized projection, in the sense that if $g = \delta_C$ is the *indicator function* of a nonempty closed convex set $C \subseteq \mathbb{R}^n$, *i.e.*, $g(x) = 0$ for $x \in C$ and $+\infty$ otherwise, then $\operatorname{prox}_{\gamma g} = \Pi_C$ is the projection on C for any $\gamma > 0$. The value function of the optimization problem (1.2) defining the proximal mapping is called the *Moreau envelope* and is denoted by g^γ, *i.e.*,

$$g^\gamma(x) = \min_{u \in \mathbb{R}^n} \left\{ g(u) + \tfrac{1}{2\gamma} \|u - x\|^2 \right\}. \tag{1.3}$$

Properties of the Moreau envelope and the proximal mapping are well documented in the literature [3, 30–32]. For example, the proximal mapping is single-valued, continuous and nonexpansive (Lipschitz continuous with Lipschitz constant 1) and the envelope function g^γ is convex, continuously differentiable, with gradient

$$\nabla g^\gamma(x) = \gamma^{-1}(x - \operatorname{prox}_{\gamma g}(x)), \tag{1.4}$$

which is γ^{-1}-Lipschitz continuous [31, Prop. 12.29].

We will consider cases where g is *twice epi-differentiable* [30, Def. 13.6], and indicate with $\mathrm{d}^2 g(x|v)$ the second-order epi-derivative of g at x for v.

For a mapping $F : \mathbb{R}^n \to \mathbb{R}^m$ we will indicate by $DF(x)$ and $JF(x)$, respectively, its semiderivative and Jacobian at x, when these exist. The directional derivative of F at x along a direction d will then be denoted as $DF(x)[d]$ if F is semidifferentiable at x, and as $JF(x)[d] = JF(x)d$ if F is differentiable at x. For the basic notions about semidifferentiability, and its link with ordinary differentiability, we refer the reader to Appendix A and the references therein.

We will talk about the linear and superlinear convergence of the proposed algorithm according to the following definition (see also [33, Def. 2.3.1] and discussion thereafter).

Definition 1.1. *We say that $(x^k)_{k\in\mathbb{N}}$ converges to $x_\star$*

(i) Q-linearly with factor $\omega \in [0,1)$ if $\|x^{k+1} - x_\star\| \leq \omega \|x^k - x_\star\|$ for all $k \geq 0$;

(ii) Q-superlinearly if $\|x^{k+1} - x_\star\| / \|x^k - x_\star\| \to 0$.

The convergence rate is R-linear (respectively, R-superlinear) if $\|x^k - x_\star\| \leq a_k$ for all $k \geq 0$ and a sequence $(a^k)_{k\in\mathbb{N}}$ such that $a_k \to 0$ with Q-linear (Q-superlinear) rate.

1.2 The forward-backward splitting

In the rest of the paper we will work under the following

Assumption 1. *$\varphi = f + g$ with $f \in C^{1,1}_{L_f}(\mathbb{R}^n)$ for some $L_f > 0$ and $g \in \Gamma_0(\mathbb{R}^n)$.*

If f satisfies Assumption 1 then [34, Prop. A.24]

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{L_f}{2}\|y - x\|^2. \tag{1.5}$$

Given an initial point x^0 and $\gamma > 0$, forward-backward splitting (also known as proximal gradient method) seeks solutions to the problem (1.1) by means of the following iterations:

$$x^{k+1} = \operatorname{prox}_{\gamma g}(x^k - \gamma \nabla f(x^k)). \tag{1.6}$$

Under Assumption 1 the generated sequence $(x^k)_{k\in\mathbb{N}}$ satisfies [15, eq. (2.13)]

$$\varphi(x^{k+1}) - \varphi(x^k) \leq -\tfrac{2-\gamma L_f}{2\gamma}\|x^{k+1} - x^k\|^2.$$

If $\gamma \in (0, 2/L_f)$ and φ is lower bounded, it can be easily inferred that any cluster point x is stationary for φ, in the sense that it satisfies the necessary condition for optimality $x \in \operatorname{zer} \partial\varphi$. The existence of cluster points is ensured if $(x^k)_{k\in\mathbb{N}}$ remains bounded; due to the monotonic behavior of $(\varphi(x^k))_{k\in\mathbb{N}}$ for γ in the given range, this condition in turn is guaranteed if φ and the initial point x^0 satisfy the following requirement, which is a standard assumption for nonconvex problems (see e.g. [15]).

Assumption 2. *The level set $\{x \in \mathbb{R}^n \mid \varphi(x) \leq \varphi(x^0)\}$, which for conciseness we shall denote $\{\varphi \leq \varphi(x^0)\}$, is bounded. In particular, there exists $R > 0$ such that $\|x - z\| \leq R$ for all $x \in \{\varphi \leq \varphi(x^0)\}$ and $z \in \operatorname{argmin} \varphi$.*

The existence of such a uniform radius R is due to boundedness of $\operatorname{argmin}\varphi$, which in turn follows from the assumed boundedness of $\{\varphi \le \varphi(x^0)\}$.

Example 1.2. To see that $\operatorname{argmin}\varphi \neq \emptyset$ is not enough for preventing the generation of unbounded sequences, consider $\varphi = f + g : \mathbb{R} \to \overline{\mathbb{R}}$ where

$$g = \delta_{(-\infty,2]} \quad \text{and} \quad f(x) = \begin{cases} \exp(x) - 1 & \text{if } x < 0, \\ x - x^2 & \text{if } x \ge 0. \end{cases}$$

Assumption 1 is satisfied with $L_f = 2$ and $\operatorname{argmin}\varphi = \{2\}$. However, for any $\gamma \in (0,1)$ the sequence $(x^k)_{k\in\mathbb{N}}$ generated by (1.6) with $x^0 < 1/2$ diverges to $-\infty$, and $\varphi(x^k) \to -1 > -2 = \min\varphi$. This however cannot happen in the convex case [31, Thm. 25.8].

We use shorthands to denote the forward-backward mapping and the associated *fixed-point residual* in order to simplify the notation:

$$T_\gamma(x) = \operatorname{prox}_{\gamma g}(x - \gamma\nabla f(x)), \tag{1.7}$$
$$R_\gamma(x) = \gamma^{-1}(x - T_\gamma(x)), \tag{1.8}$$

so that iteration (1.6) can be written as $x^{k+1} = T_\gamma(x^k) = x^k - \gamma R_\gamma(x^k)$. The set $\operatorname{zer}\partial\varphi$ is easily characterized in terms of the fixed-point set of T_γ as follows:

$$x = T_\gamma(x) \iff x \in \operatorname{zer}\partial\varphi. \tag{1.9}$$

Note that $T_\gamma(x)$ can alternatively be expressed as the solution to the following partially linearized subproblem (see also Figure 1):

$$T_\gamma(x) = \operatorname*{argmin}_{u\in\mathbb{R}^n} \left\{ \ell_\varphi(u,x) + \tfrac{1}{2\gamma}\|u-x\|^2 \right\}, \tag{1.10a}$$
$$\ell_\varphi(u,x) = f(x) + \langle \nabla f(x), u - x\rangle + g(u). \tag{1.10b}$$

2 Forward-backward envelope

We now proceed to the reformulation of (1.1) as the minimization of an unconstrained continuously differentiable function. To this end, we consider the value function of problem (1.10a) defining the forward-backward mapping T_γ and give the following definition.

Definition 2.1 (Forward-backward envelope)**.** *Let f, g and φ be as in Assumption 1, and let $\gamma > 0$. The forward-backward envelope (FBE) of φ with parameter γ is*

$$\varphi_\gamma(x) = \min_{u\in\mathbb{R}^n} \left\{ \ell_\varphi(u,x) + \tfrac{1}{2\gamma}\|u-x\|^2 \right\}. \tag{2.1}$$

Using (1.10a) and (1.10b) it is easy to verify that (2.1) can be equivalently expressed as

$$\varphi_\gamma(x) = f(x) + g(T_\gamma(x)) - \gamma\langle \nabla f(x), R_\gamma(x)\rangle + \tfrac{\gamma}{2}\|R_\gamma(x)\|^2 \tag{2.2}$$

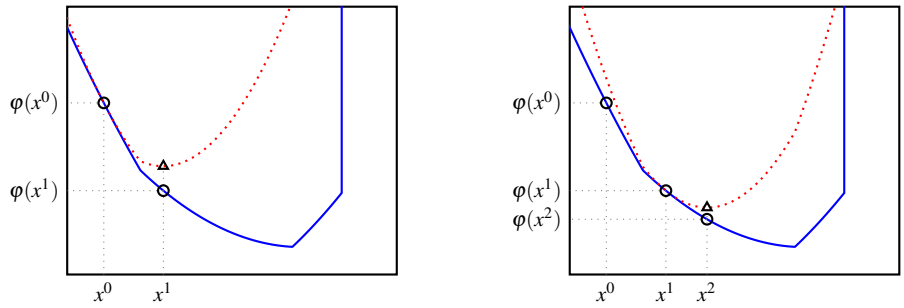

Fig. 1: When γ is small enough forward-backward splitting minimizes, at every step, a convex majorization (red, dotted lines) of the original cost φ (blue, solid line), cf. (1.10a).

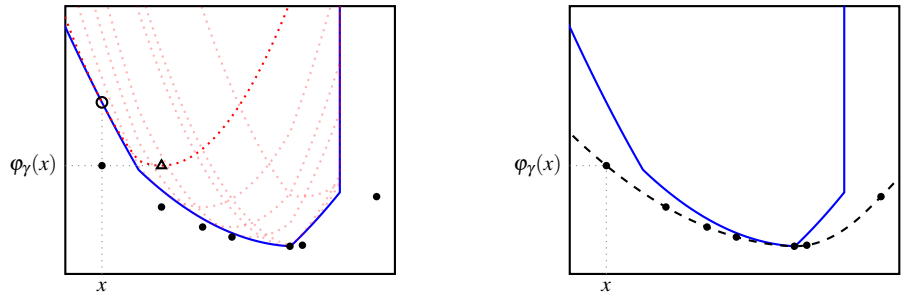

Fig. 2: The forward-backward envelope φ_γ (black, dashed line) is obtained by considering the optimal values of problems (1.10a) (dotted lines), and serves as a real-valued lower bound for the original objective φ (blue, solid line).

or, by the definition of Moreau envelope, as

$$\varphi_\gamma(x) = f(x) - \tfrac{\gamma}{2}\|\nabla f(x)\|^2 + g^\gamma(x - \gamma\nabla f(x)). \tag{2.3}$$

The geometrical construction of φ_γ is depicted in Figure 2. One distinctive feature of φ_γ is the fact that it is real-valued, despite the fact that φ can be extended-real-valued. Function φ_γ has other favorable properties which we now summarize.

2.1 Basic inequalities

The following result states the fundamental inequalities relating φ_γ to φ.

Proposition 2.2. *Suppose Assumption 1 is satisfied. Then, for all $x \in \mathbb{R}^n$*

(i) $\varphi_\gamma(x) \le \varphi(x) - \frac{\gamma}{2}\|R_\gamma(x)\|^2$ *for all* $\gamma > 0$;

(ii) $\varphi(T_\gamma(x)) \le \varphi_\gamma(x) - \frac{\gamma}{2}\left(1 - \gamma L_f\right)\|R_\gamma(x)\|^2$ *for all* $\gamma > 0$;

(iii) $\varphi(T_\gamma(x)) \le \varphi_\gamma(x)$ *for all* $\gamma \in (0, 1/L_f]$.

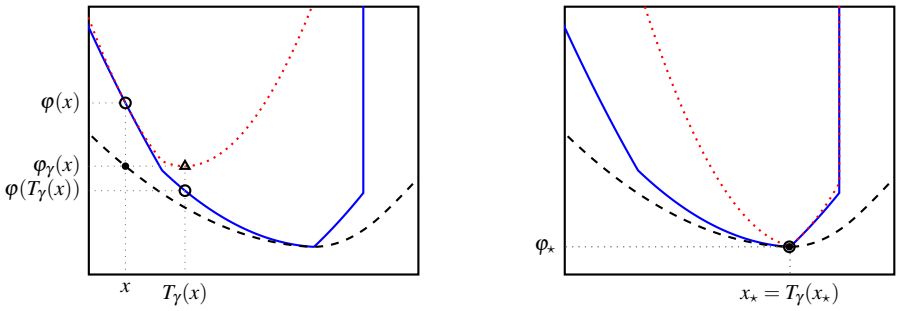

Fig. 3: on the left, by Proposition 2.2 $\varphi_\gamma(x)$ is upper bounded by $\varphi(x)$ and, when γ is small enough, lower bounded by $\varphi(T_\gamma(x))$. On the right, by Proposition 2.3(i) the two bounds coincide in correspondence of critical points.

Proof. Regarding 2.2(i), from the optimality condition for (1.10a) we have

$$R_\gamma(x) - \nabla f(x) \in \partial g(T_\gamma(x)),$$

i.e., $R_\gamma(x) - \nabla f(x)$ is a subgradient of g at $T_\gamma(x)$. From subgradient inequality

$$\begin{aligned} g(x) &\geq g(T_\gamma(x)) + \langle R_\gamma(x) - \nabla f(x), x - T_\gamma(x)\rangle \\ &= g(T_\gamma(x)) - \gamma\langle \nabla f(x), R_\gamma(x)\rangle + \gamma\|R_\gamma(x)\|^2. \end{aligned}$$

Adding $f(x)$ to both sides and considering (2.2) proves the claim. For 2.2(ii), we have

$$\begin{aligned} \varphi_\gamma(x) &= f(x) + \gamma\langle \nabla f(x), R_\gamma(x)\rangle + g(T_\gamma(x)) + \tfrac{\gamma}{2}\|R_\gamma(x)\|^2 \\ &\geq f(T_\gamma(x)) + g(T_\gamma(x)) - \tfrac{L_f}{2}\|T_\gamma(x) - x\|^2 + \tfrac{\gamma}{2}\|R_\gamma(x)\|^2. \end{aligned}$$

where the inequality follows by (1.5). 2.2(iii) then trivially follows. □

A consequence of Proposition 2.2 is that, whenever γ is small enough, the problems of minimizing φ and φ_γ are equivalent.

Proposition 2.3. *Suppose Assumption 1 is satisfied. Then,*

(i) $\varphi(z) = \varphi_\gamma(z)$ *for all* $\gamma > 0$ *and* $z \in \operatorname{zer}\partial\varphi$*;*

(ii) $\inf\varphi = \inf\varphi_\gamma$ *and* $\operatorname{argmin}\varphi \subseteq \operatorname{argmin}\varphi_\gamma$ *for* $\gamma \in (0, 1/L_f]$*;*

(iii) $\operatorname{argmin}\varphi = \operatorname{argmin}\varphi_\gamma$ *for all* $\gamma \in (0, 1/L_f)$.

Proof. 2.3(i) follows from (1.9), Propositions 2.2(i) and 2.2(ii).

Suppose now $\gamma \in (0, 1/L_f]$. In particular, 2.3(i) holds for any $x_\star \in \operatorname{argmin}\varphi$, so

$$\varphi_\gamma(x_\star) = \varphi(x_\star) \leq \varphi(T_\gamma(x)) \leq \varphi_\gamma(x) \qquad \text{for all } x \in \mathbb{R}^n$$

where the first inequality follows from optimality of $x_\star$ for φ, and the second from Proposition 2.2(iii). Therefore, every $x_\star \in \operatorname{argmin}\varphi$ is also a minimizer of φ_γ, and $\min\varphi = \min\varphi_\gamma$ provided that the former is attained. It remains to show the case $\operatorname{argmin}\varphi = \emptyset$. By Proposition 2.2(i) we have $\inf\varphi_\gamma \leq \inf\varphi$. If there exists $x \in \mathbb{R}^n$

such that $\varphi_\gamma(x) \leq \inf\varphi$, then Proposition 2.2(ii) implies that $\varphi(T_\gamma(x)) \leq \inf\varphi$, contradicting $\operatorname{argmin}\varphi = \emptyset$. Therefore $\inf\varphi_\gamma = \inf\varphi$, proving 2.3(ii).

Suppose now $\gamma \in (0, 1/L_f)$, and let $x_\star \in \operatorname{argmin}\varphi_\gamma$. From Propositions 2.2(i) and 2.2(ii) we get that

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

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

$$\varphi_\gamma(x_\star) = \varphi_\gamma(T_\gamma(x_\star)) \leq \varphi(x_\star) \leq \varphi_\gamma(x_\star).$$

Since $\varphi_\gamma \leq \varphi$ and $x_\star$ minimizes φ_γ, it follows that $x_\star \in \operatorname{argmin}\varphi$. Therefore, the sets of minimizers of φ and φ_γ coincide, proving 2.3(iii). □

Example 2.4. To see that the bounds on γ in Proposition 2.3 are tight, consider the convex problem

$$\operatorname*{minimize}_{x\in\mathbb{R}^n} \varphi(x) \equiv \overbrace{\tfrac{1}{2}\|x\|^2}^{f(x)} + \overbrace{\delta_{\mathbb{R}^n_+}(x)}^{g(x)}$$

where $\mathbb{R}^n_+ = \{x \in \mathbb{R}^n \mid x_i \geq 0, i = 1\ldots n\}$ is the nonnegative orthant. Assumption 1 is satisfied with $L_f = 1$, and the only stationary point for φ is the unique minimizer $x_\star = 0$. Using (2.3) we can explicitly compute the FBE: for any $\gamma > 0$ we have

$$\varphi_\gamma(x) = \tfrac{1-\gamma}{2}\|x\|^2 + \tfrac{1}{2\gamma}\|(1-\gamma)x - [(1-\gamma)x]_+\|^2,$$

where $[x]_+ = \Pi_{\mathbb{R}^n_+}(x) = \max\{x, 0\}$, the last expression being meant componentwise. For any $\gamma > 0$ we have that $\varphi_\gamma(x_\star) = \varphi(x_\star)$, as ensured by Proposition 2.3(i), and as long as $\gamma < 1 = 1/L_f$ all properties in Proposition 2.3 do hold. For $\gamma = 1$ we have that $\varphi_\gamma \equiv 0$, showing the inclusion in Proposition 2.3(ii) to be proper, yet satisfying $\min\varphi_\gamma = \min\varphi$.

However, for $\gamma > 1$ the FBE φ_γ is not even lower bounded, as it can be easily deduced by observing that, letting $x^k = (-k, 0\ldots 0)$ for $k \in \mathbb{N}$, $\varphi_\gamma(x^k) = \frac{1-\gamma}{2}k^2$ is arbitrarily negative. □

Proposition 2.3 implies, using Proposition 2.2(i), that an ε-optimal solution x of φ is automatically ε-optimal for φ_γ and, using Proposition 2.2(ii), from an ε-optimal for φ_γ we can directly obtain an ε-optimal solution for φ if $\gamma \in (0, 1/L_f]$:

$$\begin{aligned} \varphi(x) - \inf\varphi \leq \varepsilon &\implies \varphi_\gamma(x) - \inf\varphi \leq \varepsilon \\ \varphi_\gamma(x) - \inf\varphi_\gamma \leq \varepsilon &\implies \varphi(T_\gamma(x)) - \inf\varphi \leq \varepsilon \end{aligned}$$

Proposition 2.3 also highlights the first apparent similarity between the concepts of FBE and Moreau envelope (1.3): the latter is indeed itself a lower bound for the original function, sharing with it its minimizers and minimum value. In fact, the two are directly related as we now show. In particular, the following result implies that if φ is convex (e.g. if f is) and $\gamma \in (0, 1/L_f)$, then the possibly nonconvex φ_γ is upper and lower bounded by convex functions.

Proposition 2.5. *Suppose Assumption 1 is satisfied. Then,*

(i) $\varphi_\gamma \le \varphi^{\frac{\gamma}{1+\gamma L_f}}$ *for all* $\gamma > 0$*;*

(ii) $\varphi^{\frac{\gamma}{1-\gamma L_f}} \le \varphi_\gamma$ *for all* $\gamma \in (0, 1/L_f)$*;*

(iii) $\varphi_\gamma \le \varphi^\gamma$ *if* f *is convex.*

Proof. (1.5) implies the following bounds concerning the partial linearization:

$$-\tfrac{L_f}{2}\|u-x\|^2 \le \varphi(u) - \ell_\varphi(u,x) \le \tfrac{L_f}{2}\|u-x\|^2.$$

Combined with the definition of the FBE, cf. (2.1), this proves 2.5(i) and 2.5(ii).

If f is convex, the lower bound can be strengthened to $0 \le \varphi(u) - \ell_\varphi(u,x)$. Adding $\frac{1}{2\gamma}\|u-x\|^2$ to both sides and minimizing with respect to u yields 2.5(iii). □

2.2 Differentiability

We now turn our attention to differentiability of φ_γ, which is fundamental in devising and analyzing algorithms for solving (1.1). To ensure continuous differentiability of φ_γ we will need the following

Assumption 3. *Function* f *is twice-continuously differentiable over* $\mathbb{R}^n$.

Under Assumption 3, the function

$$Q_\gamma : \mathbb{R}^n \to \mathbb{R}^{n\times n} \qquad \text{given by} \qquad Q_\gamma(x) = I - \gamma\nabla^2 f(x) \tag{2.4}$$

is well defined, continuous, and symmetric-valued.

Theorem 2.6 (Differentiability of φ_γ)**.** *Suppose that Assumptions 1 and 3 are satisfied. Then,* φ_γ *is continuously differentiable with*

$$\nabla\varphi_\gamma(x) = Q_\gamma(x)R_\gamma(x). \tag{2.5}$$

If $\gamma \in (0, 1/L_f)$ *then the set of stationary points of* φ_γ *equals* $\operatorname{zer}\partial\varphi$.

Proof. Consider expression (2.3) for φ_γ. The gradient of g^γ is given by (1.4), and since $f \in C^2$ we have

$$\begin{aligned}\nabla\varphi_\gamma(x) &= \nabla f(x) - \gamma\nabla^2 f(x)\nabla f(x) + \gamma^{-1}\left(I - \gamma\nabla^2(x)\right)(x - \gamma\nabla f(x) - T_\gamma(x)) \\ &= \left(I - \gamma\nabla^2(x)\right)(\nabla f(x) - \nabla f(x) + \gamma^{-1}(x - T_\gamma(x))).\end{aligned}$$

This proves (2.5). If $\gamma \in (0, 1/L_f)$ then $Q_\gamma(x)$ is nonsingular for all x, and therefore $\nabla\varphi_\gamma(x) = 0$ if and only if $R_\gamma(x) = 0$, which means that x is a critical point of φ by (1.9). □

Together with Proposition 2.3, Theorem 2.6 shows that if $\gamma \in (0, 1/L_f)$ the nonsmooth problem (1.1) is completely equivalent to the unconstrained minimization of the continuously differentiable function φ_γ, in the sense that the sets of minimizers and optimal values are equal. In particular, as remarked in the next statement, if φ is convex then the set of stationary points of φ_γ turns out to be equal to the set of its minimizers, hence of solutions to the problem, even though φ_γ may not be convex.

Corollary 2.7. *Suppose that Assumptions 1 and 3 are satisfied. If* φ *is convex (e.g. if* f *is), then* $\operatorname{argmin}\varphi = \operatorname{zer}\nabla\varphi_\gamma$ *for all* $\gamma \in (0, 1/L_f)$.

2.3 Second-order properties

The FBE is not everywhere twice continuously differentiable in general. For example, if g is real valued then $g^\gamma \in C^2$ if and only if $g \in C^2$ [35]. However, second order properties will only be needed at critical points of φ in our framework, and for this purpose we can rely on generalized second-order differentiability notions described in [30, Chapter 13].

Assumption 4. *Function g is twice epi-differentiable at $x \in \operatorname{zer}\partial\varphi$ for $-\nabla f(x)$, with second order epi-derivative generalized quadratic. That is,*

$$\mathrm{d}^2 g(x|-\nabla f(x))[d] = \langle d, Md\rangle + \delta_S(d), \quad \forall d \in \mathbb{R}^n \tag{2.6}$$

where $S \subseteq \mathbb{R}^n$ is a linear subspace, and $M \in \mathbb{R}^{n\times n}$ is symmetric, positive semidefinite, and such that $\operatorname{Im}(M) \subseteq S$ and $\operatorname{Ker}(M) \supseteq S^\perp$.

In some results we will need to assume the following slightly stronger property.

Assumption 5. *Function g satisfies Assumption 4 at $x \in \operatorname{zer}\partial\varphi$ and is strictly twice epi-differentiable at x for $-\nabla f(x)$.*

The properties of M in Assumption 4 cause no loss of generality. Indeed, letting Π_S denote the orthogonal projection onto S (Π_S is symmetric, see [36]), if $M \succeq 0$ satisfies (2.6) so does $M' = \Pi_S[\frac{1}{2}(M+M^\top)]\Pi_S$, which has the wanted properties.

Twice epi-differentiability of g is a mild requirement, and cases where $\mathrm{d}^2 g$ is actually generalized quadratic are abundant [37–40]. For example, if g is piecewise linear and $x \in \operatorname{zer}\partial\varphi$, then from [37, Thm. 3.1] it follows that (2.6) holds if and only if the normal cone $N_{\partial g(x)}(-\nabla f(x))$ is a linear subspace, which is equivalent to

$$-\nabla f(x) \in \operatorname{relint}\partial g(x)$$

where $\operatorname{relint}\partial g(x)$ is the relative interior of the convex set $\partial g(x)$.

Example 2.8 (Lasso)**.** Let $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and $\lambda > 0$. Consider $f(x) = \frac{1}{2}\|Ax-b\|^2$ and $g(x) = \lambda\|x\|_1$. Minimizing $\varphi = f+g$ is a frequent problem known as lasso, and attempts to find a sparse least squares solution to the linear system $Ax = b$. One has

$$[\partial g(x)]_i = \begin{cases} \{\lambda\} & x_i > 0 \\ \{-\lambda\} & x_i < 0 \\ [-\lambda,\lambda] & x_i = 0. \end{cases}$$

In this case $\mathrm{d}^2 g(x|-\nabla f(x))$ is generalized quadratic at a solution x as long as whenever $x_i = 0$ it holds that $|(A^T(Ax-b))_i| \neq \lambda$.

We begin by investigating differentiability of the residual mapping R_γ.

Lemma 2.9. *Suppose that Assumptions 1 and 3 are satisfied, and that g satisfies Assumption 4 (Assumption 5) at a point $x \in \operatorname{zer}\partial\varphi$. Then, $\operatorname{prox}_{\gamma g}$ is (strictly) differentiable at $x-\gamma\nabla f(x)$, and R_γ is (strictly) differentiable at x with Jacobian*

$$JR_\gamma(x) = \gamma^{-1}(I - P_\gamma(x)Q_\gamma(x)), \tag{2.7}$$

where Q_γ is as in (2.4), and

$$P_\gamma(x) = J\,\mathrm{prox}_{\gamma g}(x - \gamma\nabla f(x)) = \Pi_S[I + \gamma M]^{-1}\Pi_S. \tag{2.8}$$

Moreover, $Q_\gamma(x)$ and $P_\gamma(x)$ are symmetric, $P_\gamma(x) \succeq 0$, $\|P_\gamma(x)\| \leq 1$, and if $\gamma \in (0, 1/L_f)$ then $Q_\gamma(x) \succ 0$.

Proof. See Appendix B. □

Next, we see that differentiability of the residual R_γ is equivalent to that of $\nabla\varphi_\gamma$. Mild additional assumptions on f extend this kinship to strict differentiability. Moreover, all strong (local) minimizers of the original problem, *i.e.*, of φ, are also strong (local) minimizers of φ_γ (and vice versa, due to the lower-bound property of φ_γ).

Theorem 2.10. *Suppose that Assumptions 1 and 3 are satisfied, and that g satisfies Assumption 4 at a point $x \in \operatorname{zer}\partial\varphi$. Then, φ_γ is twice differentiable at x, with symmetric Hessian given by*

$$\nabla^2\varphi_\gamma(x) = \gamma^{-1}Q_\gamma(x)(I - P_\gamma(x)Q_\gamma(x)), \tag{2.9}$$

where Q_γ and P_γ are as in Lemma 2.9. If moreover $\nabla^2 f$ is Lipschitz around x and g satisfies Assumption 5 at x, then φ_γ is strictly twice differentiable at x.

Proof. Recall from (2.5) that $\nabla\varphi_\gamma(x) = Q_\gamma(x)R_\gamma(x)$. The result follows from Lemma 2.9 and Proposition A.2 in the Appendix with $Q = Q_\gamma$ and $R = R_\gamma$. □

Theorem 2.11. *Suppose that Assumptions 1 and 3 are satisfied, and that g satisfies Assumption 4 at a point $x \in \operatorname{zer}\partial\varphi$. Then, for all $\gamma \in (0, 1/L_f)$ the following are equivalent:*

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

Proof. See Appendix B. □

2.4 Interpretations

An interesting observation is that the FBE provides a link between gradient methods and FBS, just like the Moreau envelope (1.3) does for the proximal point algorithm [41]. To see this, consider the problem

$$\text{minimize } g(x) \tag{2.10}$$

where $g \in \Gamma_0(\mathbb{R}^n)$. The proximal point algorithm for solving (2.10) is

$$x^{k+1} = \mathrm{prox}_{\gamma g}(x^k). \tag{2.11}$$

It is well known that the proximal point algorithm can be interpreted as a gradient method for minimizing the Moreau envelope of g, cf. (1.3). Indeed, due to (1.4), iteration (2.11) can be expressed as

$$x^{k+1} = x^k - \gamma \nabla g^\gamma(x^k).$$

This simple idea provides a link between nonsmooth and smooth optimization and has led to the discovery of a variety of algorithms for problem (2.10), such as semismooth Newton methods [42], variable-metric [43] and quasi-Newton methods [44–46], and trust-region methods [47], to name a few.

However, when dealing with composite problems, even if $\operatorname{prox}_{\gamma f}$ and $\operatorname{prox}_{\gamma g}$ are cheaply computable, computing the proximal mapping of $\varphi = f + g$ is usually as hard as solving (1.1) itself. On the other hand, forward-backward splitting takes advantage of the structure of the problem by operating separately on the two summands, cf. (1.6). The question that naturally arises is the following:

> *Is there a continuously differentiable function that provides an interpretation of FBS as a gradient method, just like the Moreau envelope does for the proximal point algorithm?*

The forward-backward envelope provides an affirmative answer. Specifically, whenever f is C^2, FBS can be interpreted as the following (variable-metric) gradient method on the FBE:

$$x^{k+1} = x^k - \gamma(I - \gamma \nabla^2 f(x^k))^{-1} \nabla \varphi_\gamma(x^k), \tag{2.12}$$

cf. Theorem 2.6. Furthermore, the following properties hold for the Moreau envelope

$$g^\gamma \leq g, \quad \inf g^\gamma = \inf g, \quad \operatorname{argmin} g^\gamma = \operatorname{argmin} g,$$

which correspond to Propositions 2.2(i) and 2.3 for the FBE. The relationship between Moreau envelope and forward-backward envelope is then apparent. This opens the possibility of extending FBS and devising new algorithms for problem (1.1) by simply reconsidering and appropriately adjusting methods for unconstrained minimization of continuously differentiable functions, the most well studied problem in optimization.

3 Forward-backward line-search methods

We consider line-search methods applied to the problem of minimizing φ_γ, hence solving (1.1). Requirements of such methods are often restrictive, including convexity or even strong convexity of the objective function, properties that unfortunately the FBE does not satisfy in general. As opposed to this, FBS possesses strong convergence properties and complexity estimates. We now show that it is possible to exploit the composite structure of (1.1) and devise line-search methods with the same global convergence properties and oracle information as FBS.

Algorithm 1 MINFBE

Input: $x^0 \in \mathbb{R}^n$, $\gamma_0 > 0$, $\sigma \in (0,1)$, $\beta \in [0,1)$, $k \leftarrow 0$
1: **if** $R_{\gamma_k}(x^k) = 0$ **then** stop
2: select d^k such that $\langle d^k, \nabla\varphi_{\gamma_k}(x^k)\rangle \leq 0$
3: select $\tau_k \geq 0$ and set $w^k \leftarrow x^k + \tau_k d^k$ such that $\varphi_{\gamma_k}(w^k) \leq \varphi_{\gamma_k}(x^k)$
4: **if** $f(T_{\gamma_k}(w^k)) > f(x^k) - \gamma_k\langle \nabla f(x^k), R_{\gamma_k}(x^k)\rangle + \frac{(1-\beta)\gamma_k}{2}\|R_{\gamma_k}(x^k)\|^2$ **then** $\gamma_k \leftarrow \sigma\gamma_k$, go to step 1
5: $x^{k+1} \leftarrow T_{\gamma_k}(w^k)$
6: $\gamma_{k+1} \leftarrow \gamma_k$
7: $k \leftarrow k+1$, go to step 1

Algorithm 1, which we call MINFBE, interleaves descent steps over the FBE with forward-backward steps. In particular, steps 2 and 3 provide fast asymptotic convergence when directions d^k are appropriately selected, while step 5 ensures global convergence: this is of central importance, as such properties are not usually enjoyed by standard line-search methods employed to minimize general nonconvex functions [26–29]. Moreover, in the convex case we are able to show global convergence rate results which are not typical for line-search methods with e.g. quasi-Newton directions. We anticipate some of the favorable properties that MINFBE shares with FBS:

- square-summability of the residuals for lower bounded φ (Proposition 3.4);
- global sublinear rate of the objective for convex φ with bounded level sets (Theorem 3.6);
- global convergence when φ has bounded level sets and satisfies the Kurdyka-Łojasiewicz at its stationary points (Theorem 3.10);
- local linear rate when φ has the Łojasiewicz property at its critical points (Theorem 3.11).

Moreover, unlike ordinary line-search methods applied to φ_γ, we will see in Proposition 3.4 that MINFBE is a descent method both for φ_γ and φ. Note that, despite the fact that the algorithm operates on φ_γ, all the above properties require assumptions or provide results on φ, *i.e.*, on the original problem.

The parameter γ defining the FBE is adjusted in step 4 so as to comply with the inequality in Proposition 2.2(ii), starting from an initial value γ_0 and decreasing it when necessary. The next result shows that γ_0 is decremented only a finite number of times along the iterations, and therefore γ_k is positive and eventually constant. In the rest of the paper we will denote γ_∞ such asymptotic value of γ_k.

Lemma 3.1. *Let $(\gamma_k)_{k\in\mathbb{N}}$ the sequence of stepsize parameters computed by MINFBE, and let $\gamma_\infty = \min_{i\in\mathbb{N}} \gamma_i$. Then for all $k \in \mathbb{N}$,*

$$\gamma_k \geq \gamma_\infty \geq \min\left\{\gamma_0, \sigma(1-\beta)/L_f\right\} > 0.$$

Proof. See Appendix C. □

Remark 3.2. In MINFBE:

(i) Selecting $\beta = 0$ and $d^k \equiv 0$, $\tau_k \equiv 0$ for all k yields the classical forward-backward splitting with backtracking on γ [14, Sec. 3].

(ii) Substituting step 5 with $x^{k+1} \leftarrow w^k$ yields a classical line-search method for the problem of minimizing φ_γ, where a suitable γ is adaptively determined. However, extensive numerical experience has shown that even though this variant seems to always converge, our choice $x^{k+1} \leftarrow T_{\gamma_k}(w^k)$ usually performs better in practice, in terms of number of forward-backward steps, cf. Section 5.

(iii) Step 5 comes at no additional cost once τ_k has been determined by means of a line-search. In fact, in order to evaluate $\varphi_{\gamma_k}(w^k)$ and test the condition in step 3, the evaluation of $T_{\gamma_k}(w^k)$ is required.

(iv) When L_f is known and $\gamma_0 \in (0, (1-\beta)/L_f]$, the condition in step 4 never holds, see Proposition 2.2(ii). In this case MINFBE reduces to Algorithm 2: without loss of generality we will focus the analysis on Algorithm 1.

Algorithm 2 MINFBE with constant γ

Input: $x^0 \in \mathbb{R}^n$, $\beta \in [0,1)$, $\gamma \in (0, (1-\beta)/L_f]$, $k \leftarrow 0$

1: **if** $R_{\gamma_k}(x^k) = 0$ **then** stop No \gamma_k, just \gamma
2: select d^k such that $\langle d^k, \nabla\varphi_\gamma(x^k)\rangle \leq 0$
3: select $\tau_k \geq 0$ and set $w^k \leftarrow x^k + \tau_k d^k$ such that $\varphi_\gamma(w^k) \leq \varphi_\gamma(x^k)$
4: $x^{k+1} \leftarrow T_\gamma(w^k)$
5: $k \leftarrow k+1$, go to step 1

Remark 3.3. In order to compute descent directions in MINFBE, one usually needs to evaluate $\nabla\varphi_\gamma$ at a sequence of points. In practice, this only requires to perform matrix-vector products with $\nabla^2 f$, see (2.4)-(2.5), and *not* the computation of the full Hessian. For example, if $f(x) = \frac{1}{2}\|Ax-b\|^2$ then $\nabla\varphi_\gamma(x) = R_\gamma(x) - A^\top[AR_\gamma(x)]$. For general nonlinear f, the product $\nabla^2 f(x)v$ can be approximated numerically using finite-differences formulas which only require one additional evaluation of ∇f. If f is analytic, then one can use a complex step [48] to overcome numerical cancellation problems, and compute $\nabla^2 f(x)v$ to machine precision at the cost of one evaluation of ∇f. Finally, automatic differentiation techniques can be used to evaluate such Hessian-vector products, that only require a small multiple of $2n$ operations in addition to those required to evaluate f, see [49, Sec. 8.2].

We denote by $\omega(x^0)$ the set of cluster points of the sequence $(x^k)_{k\in\mathbb{N}}$ produced by MINFBE started from $x^0 \in \mathbb{R}^n$. The following result states that MINFBE is a descent method both for the FBE φ_γ and for the original function φ, and, as it holds for FBS, that the sequence of fixed-point residuals is square-summable if the function is lower bounded.

Proposition 3.4 (Subsequential convergence)**.** *Suppose that Assumption 1 is satisfied. Then, the following hold for the sequences generated by MINFBE:*

(i) $\varphi(x^{k+1}) \leq \varphi(x^k) - \frac{\beta\gamma_k}{2}\|R_{\gamma_k}(w^k)\|^2 - \frac{\gamma_k}{2}\|R_{\gamma_k}(x^k)\|^2$ *for all* $k \in \mathbb{N}$*;*

(ii) either $(\|R_{\gamma_k}(x^k)\|)_{k\in\mathbb{N}}$ *is square summable, or* $\varphi(x^k) \to \inf\varphi = -\infty$*, in which case* $\omega(x^0) = \emptyset$*;*

(iii) $\omega(x^0) \subseteq \operatorname{zer} \partial\varphi$, i.e., *every cluster point of* $(x^k)_{k\in\mathbb{N}}$ *is critical;*

(iv) *if* $\beta > 0$, *then either* $(\|R_{\gamma_k}(w^k)\|)_{k\in\mathbb{N}}$ *is square summable and every cluster point of* $(w^k)_{k\in\mathbb{N}}$ *is critical, or* $\varphi_{\gamma_k}(w^k) \to \inf\varphi = -\infty$ *in which case* $(w^k)_{k\in\mathbb{N}}$ *has no cluster points.*

Proof. See Appendix C. □

An immediate consequence is the following result concerning the convergence of the fixed-point residual.

Theorem 3.5. *Suppose that Assumption 1 is satisfied, and consider the sequences generated by MINFBE. Then,*

$$\min_{i=0\cdots k} \|R_{\gamma_i}(x^i)\|^2 \leq \frac{2}{(k+1)} \frac{\varphi(x^0) - \inf\varphi}{\min\left\{\gamma_0, \sigma(1-\beta)/L_f\right\}}.$$

If $\beta > 0$, *then for all* $k \in \mathbb{N}$ *we also have*

$$\min_{i=0\cdots k} \|R_{\gamma_i}(w^i)\|^2 \leq \frac{2}{(k+1)} \frac{\varphi(x^0) - \inf\varphi}{\beta\min\left\{\gamma_0, \sigma(1-\beta)/L_f\right\}}.$$

Proof. See Appendix C. □

We now analyze the convergence properties of MINFBE. We first consider the case where f is convex. Then we discuss the general case under the assumption that φ has the Kurdyka-Łojasiewicz property: in this case $(d^k)_{k\in\mathbb{N}}$ must be uniformly bounded with respect to $(R_{\gamma_k}(x^k))_{k\in\mathbb{N}}$ in order to ensure convergence, see Theorem 3.10, condition which is not required in the convex case. When the directions are selected, say, according to a quasi-Newton scheme $d^k = -B_k^{-1}\nabla\varphi_\gamma(x^k)$, boundedness of $(B_k^{-1})_{k\in\mathbb{N}}$ will be necessary for the sake of global convergence when the Kurdyka-Łojasiewicz property holds for φ. The latter is however a milder assumption with respect to usual nonconvex line-search methods where $(B_k^{-1})_{k\in\mathbb{N}}$ is required to have bounded condition number or $(d^k)_{k\in\mathbb{N}}$ to be *gradient-oriented* (see [50] and the references therein).

3.1 Convergence in the convex case

We now prove that when f is convex MINFBE converges to the optimal objective value with the same sublinear rate as FBS. Notice that we require convexity of f (and g), and *not* that of φ_γ which may fail to be convex even when φ is.

Theorem 3.6 (Global sublinear convergence)**.** *Suppose that Assumptions 1 and 2 are satisfied, and that* f *is convex. Then, for the sequences generated by MINFBE, either* $\varphi(x^0) - \inf\varphi \geq R^2/\gamma_0$ *and*

$$\varphi(x^1) - \inf\varphi \leq \frac{R^2}{2\gamma_0}, \tag{3.1}$$

or for any $k \in \mathbb{N}$ *it holds*

$$\varphi(x^k) - \inf\varphi \leq \frac{2R^2}{k\min\left\{\gamma_0, \sigma(1-\beta)/L_f\right\}}. \tag{3.2}$$

Proof. See Appendix C. □

In the following result we see that the convergence rate of $(x^k)_{k\in\mathbb{N}}$ is linear when close to a strong local minimum.

Theorem 3.7 (Local linear convergence)**.** *Suppose that Assumption 1 is satisfied. Suppose further that f is convex and that $x_\star$ is a strong (global) minimum of φ,* i.e.*, there exist a neighborhood N of $x_\star$ and $c > 0$ such that*

$$\varphi(x) - \varphi(x_\star) \geq \tfrac{c}{2}\|x - x_\star\|^2, \quad \forall x \in N. \tag{3.3}$$

Then there is $k_0 \geq 0$ such that the subsequences $(\varphi(x^k))_{k\geq k_0}$ and $(\varphi_{\gamma_k}(w^k))_{k\geq k_0}$ produced by MINFBE converge Q-linearly to $\varphi(x_\star)$ with factor ω, where

$$\omega \leq \max\left\{\tfrac{1}{2}, 1 - \tfrac{c}{4}\min\left\{\gamma_0, \sigma(1-\beta)/L_f\right\}\right\} \in [\tfrac{1}{2}, 1),$$

and $(x^k)_{k\geq k_0}$ converges R-linearly to $x_\star$. Moreover, if $x_\star$ is a strong (global) minimum for φ_{γ_∞}, with γ_∞ as in Lemma 3.1, then also $(\varphi(w^k))_{k\geq k_0}$ converges R-linearly to $x_\star$.

Proof. See Appendix C. □

The introduction of γ_∞ in the statement above is due to the fact that γ_k may vary over the iterations. However, under the assumptions of Theorem 2.11, if $\gamma_\infty < 1/L_f$ then the requirement of $x_\star$ to be a strong local minimizer for φ_{γ_∞} is superfluous, as it is already implied by strong local minimimality of $x_\star$ for φ.

Corollary 3.8 (Global linear convergence)**.** *Suppose that Assumption 1 is satisfied, that f is convex and that φ is strongly convex (e.g. if f is strongly convex). Then, the sequences $(\varphi(x^k))_{k\in\mathbb{N}}$ and $(\varphi_{\gamma_k}(w^k))_{k\in\mathbb{N}}$ generated by MINFBE converge Q-linearly to $\varphi_\star$, while $(x^k)_{k\in\mathbb{N}}$ converges R-linearly to $x_\star$.*

Proof. In this case Theorem 3.7 applies with $N = \mathbb{R}^n$, $c = \mu_\varphi$ (the convexity modulus of φ) and $k_0 = 0$. □

3.2 Convergence under KL assumption

We now analyze the convergence of the iterates of MINFBE to a critical point under the assumption that φ satisfies the Kurdyka-Łojasiewicz (KL) property [4–6]. For related works exploiting this property in proving convergence of optimization algorithms such as FBS we refer the reader to [7–11].

Definition 3.9 (KL property [10, Def. 3])**.** *A proper lower semi-continuous function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ has the Kurdyka-Łojasiewicz property (KL) at $x_\star \in \operatorname{dom}\partial\varphi$ if there exist $\eta \in (0, +\infty]$, a neighborhood U of $x_\star$, and a continuous concave function $\psi : [0, \eta] \to [0, +\infty)$ such that:*

(i) $\psi(0) = 0$,

(ii) ψ *is* C^1 *on* $(0, \eta)$,

(iii) $\psi'(s) > 0$ *for all* $s \in (0, \eta)$,

(iv) for every $x \in U \cap \{x \in \mathbb{R}^n \mid \varphi(x_\star) < \varphi(x) \leq \varphi(x_\star) + \eta\}$,

$$\psi'(\varphi(x) - \varphi(x_\star))\operatorname{dist}(0, \partial\varphi(x)) \geq 1.$$

We say that φ *has the KL property on* $S \subseteq \mathbb{R}^n$ *it has the KL property on every* $x \in S$.

Function ψ in the previous definition is usually called *desingularizing function*. All subanalytic functions which are continuous over their domain have the KL property [51]. Under the KL assumption we are able to prove the following convergence result. Once again, we remark that such property is required on the original function φ, rather than on the surrogate φ_γ.

Theorem 3.10. *Suppose that Assumptions 1 and 2 are satisfied, and that* φ *satisfies the KL property on* $\omega(x^0)$ *(e.g. if it has it on* $\operatorname{zer}\partial\varphi$*). Suppose further that in MINFBE* $\beta > 0$*, and that there exist* $\bar{\tau}, c > 0$ *such that* $\tau_k \leq \bar{\tau}$ *and* $\|d^k\| \leq c\|R_{\gamma_k}(x^k)\|$ *for all* $k \in \mathbb{N}$*. Then, the sequence of iterates* $(x^k)_{k\in\mathbb{N}}$ *is either finite and ends with* $R_{\gamma_k}(x^k) = 0$*, or converges to a critical point* $x_\star$ *of* φ*.*

Proof. See Appendix C. □

In case where φ is subanalytic, the desingularizing function can be taken of the form $\psi(s) = \sigma s^{1-\theta}$, for $\sigma > 0$ and $\theta \in [0, 1)$ [51]. In this case, the condition in Definition 3.9(iv) is referred to as Łojasiewicz inequality. Depending on the value of θ we can derive local convergence rates for MINFBE.

Theorem 3.11 (Local linear convergence)**.** *Suppose that Assumptions 1 and 2 are satisfied, and that* φ *satisfies the KL property on* $\omega(x^0)$ *(e.g. if it has it on* $\operatorname{zer}\partial\varphi$*) with*

$$\psi(s) = \sigma s^{1-\theta} \qquad \text{for some } \sigma > 0 \text{ and } \theta \in (0, \tfrac{1}{2}]. \tag{3.4}$$

Suppose further that in MINFBE $\beta > 0$*, and that there exist* $\bar{\tau}, c > 0$ *such that* $\tau_k \leq \bar{\tau}$ *and* $\|d^k\| \leq c\|R_{\gamma_k}(x^k)\|$ *for all* $k \in \mathbb{N}$*. Then, the sequence of iterates* $(x^k)_{k\in\mathbb{N}}$ *converges to a point* $x_\star \in \operatorname{zer}\partial\varphi$ *with R-linear rate.*

Proof. See Appendix C. □

4 Quasi-Newton methods

We now turn our attention to choices of the direction d^k in MINFBE. Applying classical quasi-Newton methods [52] to the problem of minimizing φ_γ yields, starting from a given x^0,

$$\begin{aligned} d^k &= -B_k^{-1}\nabla\varphi_\gamma(x^k), \\ x^{k+1} &= x^k + \tau_k d^k, \end{aligned}$$

where B_k is nonsingular and chosen so as to approximate (in some sense) the Hessian of φ_γ at x^k, and stepsize $\tau_k > 0$ is selected with a line-search procedure enforcing a sufficient decrease condition. However, the convergence properties of quasi-Newton

methods are quite restrictive. The BFGS algorithm is guaranteed to converge, in the sense that

$$\liminf_{k\to\infty} \|\nabla\varphi_\gamma(x^k)\| = 0,$$

when the objective is convex [53]. Its limited memory variant, L-BFGS, requires strong convexity to guarantee convergence, and in that case the cost is shown to converge *R*-linearly to the optimal value [54]. Moreover, there exist examples of nonconvex function for which quasi-Newton methods need not converge to critical points [26–29].

To overcome this, we consider quasi-Newton directions in the setting of MINFBE. The resulting methods enjoy the same global convergence properties illustrated in Section 3 and superlinear asymptotic convergence under standard assumptions: we will assume, as it is usual, (strict) differentiability of $\nabla\varphi_\gamma$ and nonsingularity of $\nabla^2\varphi_\gamma$ at a critical point. Properties of f and g that guarantee these requirements were discussed in Theorems 2.10 and 2.11: if $\gamma = \gamma_\infty$ is as in Lemma 3.1, then (strict) differentiability of $\nabla\varphi_\gamma$ at $x_\star \in \operatorname{zer}\partial\varphi$ and positive definiteness of $\nabla^2\varphi_\gamma(x_\star)$ are ensured if Assumption 4 (Assumption 5) holds, $x_\star$ is a strong local minimum for φ, and $\gamma < 1/L_f$.

The following result gives for the proposed algorithmic scheme the analogous of the Dennis-Moré condition, see [25, Thm. 2.2] and [55, Thm. 3.3]. Differently from the cited results, we fit the analysis to our algorithmic framework where an additional forward-backward step is operated. Furthermore, in Theorem 4.2 we will see how achieving superlinear convergence is possible without the need to ensure *sufficient* decrease in the objective, or even to consider direction of strict descent, but simply with the nonincrease conditions of steps 2 and 3. This contrasts with the usual requirements of classical line-search methods, where instead a sufficient decrease must be enforced in order for the sequence of iterates to converge. In MINFBE, in fact, such decrease is guaranteed by the final update in step 5.

Theorem 4.1. *Suppose that Assumption 1 is satisfied, and let $\gamma > 0$. Suppose that $\nabla\varphi_\gamma$ is strictly differentiable at $x_\star$, and that $\nabla^2\varphi_\gamma(x_\star)$ is nonsingular. Let $(B_k)_{k\in\mathbb{N}}$ be a sequence of nonsingular $\mathbb{R}^{n\times n}$-matrices and suppose that for some $x^0 \in \mathbb{R}^n$ the sequences $(x^k)_{k\in\mathbb{N}}$ and $(w^k)_{k\in\mathbb{N}}$ generated by*

$$w^k = x^k - B_k^{-1}\nabla\varphi_\gamma(x^k) \quad \text{and} \quad x^{k+1} = T_\gamma(w^k)$$

converge to $x_\star$. If $x^k, w^k \notin \operatorname{zer}\partial\varphi$ for all $k \geq 0$ and

$$\lim_{k\to\infty} \frac{\|(B_k - \nabla^2\varphi_\gamma(x_\star))(w^k - x^k)\|}{\|w^k - x^k\|} = 0, \tag{4.1}$$

then $(x^k)_{k\in\mathbb{N}}$ and $(w^k)_{k\in\mathbb{N}}$ converge Q-superlinearly to $x_\star$.

Proof. See Appendix D. □

To obtain superlinear convergence of MINFBE when quasi-Newton directions are used and condition (4.1) on the sequence $(B_k)_{k\in\mathbb{N}}$ holds, we must verify that eventually $\varphi_\gamma(x^k + d^k) \leq \varphi_\gamma(x^k)$, so that the stepsize $\tau_k = 1$ is accepted in step 3 and the iterations reduce to those described in Theorem 4.1.

Theorem 4.2. *Suppose that Assumption 1 is satisfied, and that in MINFBE direction d^k is set as*

$$d^k = -B_k^{-1}\nabla\varphi_{\gamma_k}(x^k)$$

for a sequence of nonsingular matrices $(B_k)_{k\in\mathbb{N}}$ satisfying (4.1), with $\tau_k = 1$ being tried first in step 3. Let $\gamma = \gamma_\infty$ as in Lemma 3.1, and suppose further that the sequences $(x^k)_{k\in\mathbb{N}}$ and $(w^k)_{k\in\mathbb{N}}$ converge to a critical point $x_\star$ at which $\nabla\varphi_\gamma$ is continuously semidifferentiable with $\nabla^2\varphi_\gamma(x_\star) \succ 0$. Then, $(x^k)_{k\in\mathbb{N}}$ and $(w^k)_{k\in\mathbb{N}}$ converge Q-superlinearly to $x_\star$.

Proof. See Appendix D. □

4.1 BFGS

The sequence $(B_k)_{k\in\mathbb{N}}$ can be computed using BFGS updates: starting from $B_0 \succ 0$, use vectors

$$s^k = w^k - x^k, \quad y^k = \nabla\varphi_\gamma(w^k) - \nabla\varphi_\gamma(x^k), \tag{4.2a}$$

to compute

$$B_{k+1} = \begin{cases} B_k + \frac{y^k(y^k)^\top}{\langle y^k, s^k\rangle} - \frac{B_k s^k (B_k s^k)^\top}{\langle s^k, B_k s^k\rangle} & \text{if } \langle s^k, y^k\rangle > 0, \\ B_k & \text{otherwise.} \end{cases} \tag{4.2b}$$

Note that in this way $B_k \succ 0$, for all $k \geq 0$, and $d^k = -B^{-1}\nabla\varphi_\gamma(x^k)$ is always a direction of descent for φ_γ. No matrix inversion is needed to compute d^k in practice, since it is possible to perform the inverse updates of (4.2b) directly producing the sequence $(B_k^{-1})_{k\in\mathbb{N}}$, see [49, 52].

In light of the convergence results for MINFBE given in Section 3 we now proceed under either of the following assumptions.

Assumption 6. *Function φ satisfies either of the following:*

(i) it is convex and such that $\varphi(x) - \varphi(x_\star) \geq \frac{c}{2}\|x - x_\star\|^2$, for some $c > 0$ and all x close enough to $x_\star$, the unique minimizer of φ;

(ii) it has the KL property on $\omega(x^0)$ with $\psi(s) = \sigma s^{1-\theta}$, where $\sigma > 0$ and $\theta \in (0, \frac{1}{2}]$.

Theorem 4.3. *Suppose that Assumption 1 is satisfied, and that in MINFBE directions d^k are set as*

$$d^k = -B_k^{-1}\nabla\varphi_{\gamma_k}(x^k) \qquad \text{with } B_k \text{ as in (4.2),}$$

and with $\tau_k = 1$ being tried first in step 3. Let $\gamma = \gamma_\infty$ as in Lemma 3.1, and suppose further that the sequences $(x^k)_{k\in\mathbb{N}}$ and $(w^k)_{k\in\mathbb{N}}$ converge to a critical point $x_\star$ at which $\nabla\varphi_\gamma$ is calmly semidifferentiable (see Proposition A.5 in the Appendix) with $\nabla^2\varphi_\gamma(x_\star) \succ 0$. Then, $(x^k)_{k\in\mathbb{N}}$ and $(w^k)_{k\in\mathbb{N}}$ converge Q-superlinearly to $x_\star$.

Proof. See Appendix D. □

4.2 L-BFGS

When dealing with a large number of variables, storing (and updating) approximations of the Hessian matrix (or its inverse) may be impractical. Limited-memory quasi-Newton methods remedy this by storing, instead of a dense $n \times n$ matrix, only a few most recent pairs (s^k, y^k) implicitly representing such approximation. The limited-memory BFGS method (L-BFGS) is probably the most widely used method of this class, and was first introduced in [54]. It is based on the BFGS update, but uses at iteration k only the most recent $\tilde{m} = \min\{m, k\}$ pairs (here m is a parameter, usually $m \in \{3, \dots, 20\}$) to compute a descent direction: d^k is obtained using a procedure known as *two-loop recursion* [56], so that no matrix storage is required, and in fact only $O(n)$ operations are needed. For this reason L-BFGS is better suited for large scale applications. Similarly to BFGS, a safeguard is used to make sure that $\langle s^k, y^k \rangle > 0$, so that d^k is always a descent direction for φ_{γ_k}.

Remark 4.4. In both BFGS and L-BFGS, the condition $\langle s^k, y^k \rangle > 0$ is sufficient to ensure the positive definiteness of the Hessian approximation, hence the fact that d^k is a descent direction. Therefore, in MINFBE one can simply check such condition and discard the update when it does not hold. Other methods were proposed in the literature to ensure convergence of quasi-Newton methods in the nonconvex case, by Powell (see [49, Sec. 18.3]) and Li, Fukushima [57]. In our experience, no significant advantage is gained when using these techniques in MINFBE. Moreover, no such care is required for MINFBE to converge to a critical point, and under the assumptions of Theorem 4.2 the condition $\langle s^k, y^k \rangle > 0$ will eventually always hold (see the proof of Theorem 4.2 for details).

5 Simulations

We now present numerical results obtained with the proposed method. In all the results, we indicate in parenthesis the choice of directions for MINFBE. We set $\beta = 0.05$ in MINFBE, therefore if L_f is known then we set a constant $\gamma = 0.95/L_f$. To determine the stepsize τ_k in MINFBE we use backtracking, starting with $\tau_k = 1$ and reducing it until $\varphi_{\gamma_k}(x^k + \tau_k d^k) \leq \varphi_{\gamma_k}(x^k)$ holds.

Among the other algorithms, for each choice descent directions we also compare MINFBE with the corresponding classical line-search method, see Remark 3.2(ii). In this case we use a line-search procedure, inspired by [58, Sec. II.3.3], enforcing the usual Wolfe conditions: although simpler, in our tests this strategy performed favorably with respect to other algorithms, see for example [34, Sec. 1.2], [49, Sec. 3], [59, Sec. 2.6].

We always set the memory parameter $m = 5$ when computing L-BFGS directions.

All experiments were performed in MATLAB, and the implementation of the methods used in the tests are available.[1]

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

5.1 Lasso

The problem is to find a sparse representation of a vector $b \in \mathbb{R}^m$ as combination of the columns of $A \in \mathbb{R}^{m\times n}$. This is done by minimizing $\varphi = f + g$ where

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

The proximal mapping of g is the soft-thresholding operation, while the computationally relevant operation here is the computation of f and ∇f, which involves matrix-vector products with A and $A^\top$. Parameter λ modulates between a small least squares residual and a sparse solution vector $x_\star$, *i.e.*, the larger the λ the more zero coefficients $x_\star$ has. In particular, $\lambda_{\max} = \|A^\top b\|_\infty$ is the minimum value such that for $\lambda \geq \lambda_{\max}$ the solution is $x_\star = 0$. We have $L_f = \|A^\top A\|$, which can be quickly approximated using power iteration, therefore we applied MINFBE with fixed stepsize $\gamma = 0.95/L_f$.

In Figure 4 the performance of MINFBE(BFGS) is shown in a small dimensional instance taken from the SPEAR datasets.[2] It is apparent that our method greatly improves over FBS, its accelerated version, and classical BFGS applied to the problem of minimizing φ_γ.

Then we considered larger instances from the same dataset. In this case we applied L-BFGS and the nonlinear conjugate gradient method by Dai and Yuan (CG-DY, see [60]), which always produces descent directions when a line-search satisfying the Wolfe conditions is employed. The same formulas were used in the context of MINFBE: in this case CG-DY does not necessarily produce descent directions, therefore we restart the memory of the method every time an ascent direction is encountered. We also compare against SpaRSA [61], a proximal gradient algorithm using the Barzilai-Borwein method to determine the stepsize and a nonmonotone line-search to guarantee convergence, and FPC_AS [62], which is an active-set type of algorithm. These are *ad-hoc* solvers for ℓ_1-regularization problems, in contrast to our approach which is for general problems of the form (1.1). Both SpaRSA and FPC_AS adopt a *continuation* strategy to warm-start the problem and accelerate convergence. For the sake of fairness we ran also the other methods (fast FBS, L-BFGS, CG-DY and MINFBE) in a similar continuation scheme: we solve a sequence of problems, with a large initial value of λ (close to $\lambda_{\max}$) which is successively reduced until the target value is reached, using the solution to each problem as initial iterate for the successive. As it is apparent from the results in Figure 5, MINFBE(L-BFGS) and MINFBE(CG-DY) are able to solve all the instances we considered and generally outperform the other methods, including the corresponding classical line-search methods. Therefore, the additional forward-backward step performed by MINFBE after the descent step indeed pays off.

5.2 Sparse logistic regression

The composite objective function consists of

$$f(x) = \sum_{i=1}^{m} \log(1 + e^{-b_i\langle a_i, x\rangle}), \qquad g(x) = \lambda\|x\|_1.$$

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

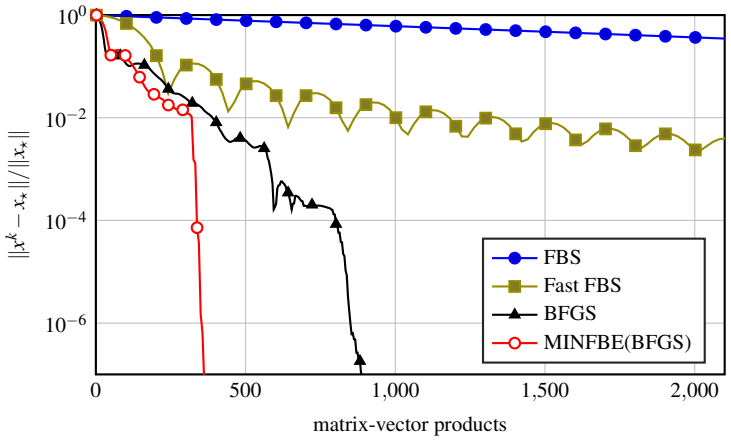

Fig. 4: Lasso: algorithms applied to the `spear_inst_1`, with $m = 512$ samples and $n = 1024$ variables, where $\lambda = 0.05\lambda_{\max}$ was used. MINFBE converges superlinearly to the global minimum when BFGS directions are used, and faster then the classical BFGS algorithm applied to the problem of minimizing φ_γ.

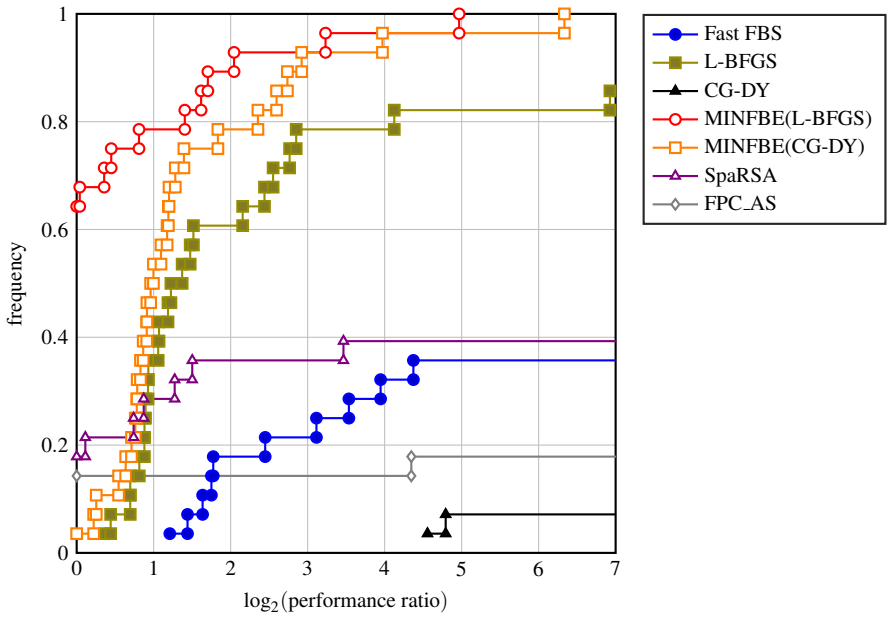

Fig. 5: Lasso: performance profile of the CPU time, for the problems in the SPEAR dataset ranging from `spear_inst_173` to `spear_inst_200`, and $\lambda = 10^{-3}\lambda_{\max}$. All algorithms use a continuation technique to warm-start the problem solution. Each method was stopped as soon as $\varphi(x^k) - \varphi_\star \leq 10^{-6}(1 + |\varphi_\star|)$. Methods not meeting this condition in 10^4 iterations were assigned a performance ratio of $+\infty$.

			Fast FBS				L-BFGS				MINFBE(L-BFGS)			
ID	$\lambda/\lambda_{\max}$	$\mathrm{nnz}(x_\star)$	it.	f	A	time (s)	it.	f	A	time (s)	it.	f	A	time (s)
rcv1	$2\cdot10^{-1}$	25	134	269	403	1.57	58	144	386	1.37	29	87	198	0.94
$m=20242$	$1\cdot10^{-1}$	70	261	523	784	2.91	132	305	843	3.68	51	168	367	1.46
$n=44504$	$5\cdot10^{-2}$	141	406	813	1219	4.49	170	386	1075	4.65	46	152	332	1.30
$\mathrm{nnz}(A)=910K$	$2\cdot10^{-2}$	287	885	1771	2656	9.75	230	530	1459	6.32	76	239	539	2.13
	$1\cdot10^{-2}$	470	1189	2379	3568	14.62	356	787	2220	8.48	105	304	720	2.93
real-sim	$2\cdot10^{-1}$	19	123	247	370	4.62	43	115	296	2.86	15	35	91	1.38
$m=72201$	$1\cdot10^{-1}$	52	200	401	601	7.09	72	176	472	4.75	20	56	132	1.54
$n=20958$	$5\cdot10^{-2}$	111	325	651	976	14.18	93	215	595	5.79	29	83	195	2.42
$\mathrm{nnz}(A)=1.5M$	$2\cdot10^{-2}$	251	577	1155	1732	21.05	154	352	976	9.46	48	139	327	3.42
	$1\cdot10^{-2}$	448	824	1649	2473	33.46	220	499	1388	13.68	72	227	511	7.09
news20	$2\cdot10^{-1}$	47	793	1590	2383	84.03	179	427	1162	50.65	79	264	573	32.76
$m=19954$	$1\cdot10^{-1}$	98	1131	2265	3396	125.86	341	789	2172	96.70	127	401	902	51.74
$n=1355191$	$5\cdot10^{-2}$	208	3106	6216	9322	320.49	409	944	2599	125.49	193	646	1411	82.85
$\mathrm{nnz}(A)=3.7M$	$2\cdot10^{-2}$	422	6647	13298	19945	673.90	1082	2481	6829	352.46	440	1499	3252	204.97

Table 1: Sparse logistic regression: performance of the algorithms on three datasets, for different values of λ. We used $\varphi(x^k)-\varphi_\star \leq 10^{-8}|\varphi_\star|$ as termination criteria.

Here vector $a_i \in \mathbb{R}^n$ contains the features of the i-th instance, and $b_i \in \{-1,1\}$ indicates the correspondent class. The ℓ_1-regularization enforces sparsity in the solution. Indicating by A the matrix having a_i as i-th row, we have $\lambda_{\max} = \frac{1}{2}\|A^\top b\|_\infty$, so that for $\lambda \geq \lambda_{\max}$ the optimal solution is $x_\star = 0$.

We ran the algorithms one three datasets,[3] and recorded the number of iterations, calls to f and ∇f, matrix-vector products with A and $A^\top$, and the running time needed to reach $\varphi(x^k)-\varphi_\star \leq 10^{-8}(1+|\varphi_\star|)$. Unlike the previous example, here a tight Lipschitz constant for ∇f is not readily available: in this case we applied MINFBE (as well as fast FBS) with backtracking on parameter γ. The results are in Table 1: MINFBE significantly reduces the number of operations needed to solve the problems. Since directions are computed according to L-BFGS, which is able to scale to large dimensional problems, CPU time is reduced analogously.

5.3 Group lasso

Let vector x be partitioned as $x = (x_1,\ldots,x_N)$, where each $x_i \in \mathbb{R}^{n_i}$, and $\sum_i n_i = n$. We consider the ℓ_2-regularized least squares problem having

$$f(x) = \frac{1}{2}\|Ax-b\|_2^2, \quad g(x) = \lambda\sum_{i=1}^N \|x_i\|_2,$$

where $x = (x_1,\ldots,x_N)$ and $x_i \in \mathbb{R}^{n_i}$ for $i = 1,\ldots,N$. The ℓ_2 terms enforce sparsity at the block level, so that for sufficiently large λ we expect many of the x_i's to be zero. Partitioning the A by columns as $A = (A_1,\ldots,A_N)$, with the same block structure at x, then for $\lambda \geq \lambda_{\max} = \max\left\{\|A_1^\top b\|_2,\ldots,\|A_N^\top b\|_2\right\}$ the optimal solution is $x_\star = 0$.

To test the methods we generated a random instance as follows: we set $m = 200$, $N = 2000$ and $n_1 = \ldots = n_N = 100$, and generated A as a sparse matrix with normally distributed entries, density 10^{-2} and condition number 10^2 using MATLAB's `sprandn` command. Then we chose x_{true} with 10 nonzero blocks, and computed

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

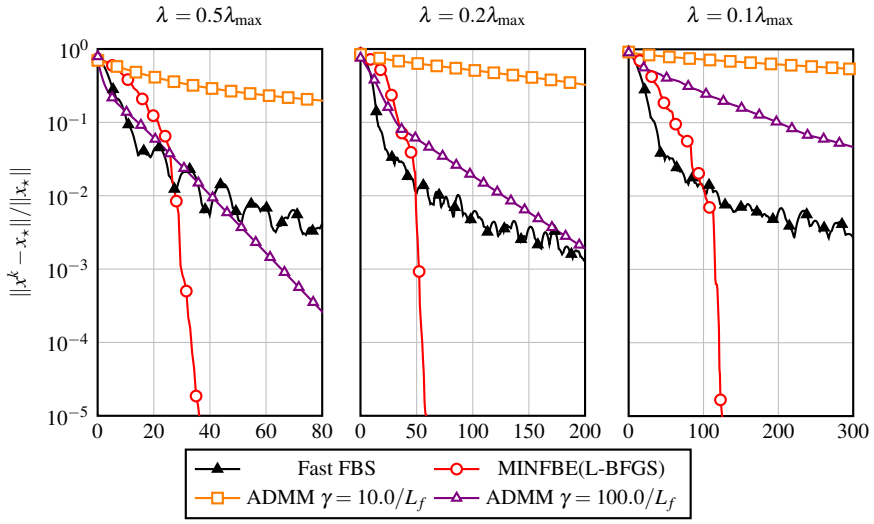

Fig. 6: Group lasso: performance of the proposed method on a random sparse problem with $m = 200$ data points and $n = 2 \cdot 10^5$ variables. Horizontal axis is time in seconds.

$b = Ax_{\text{true}} + v$, where v is a Gaussian noise vector with standard deviation 0.1. Just like in the case of lasso, the Lipschitz constant L_f can be easily estimated using power iterations. We compared fast FBS, MINFBE(L-BFGS) and ADMM (with two different stepsize parameters γ), on such an instance. As it is shown in Figure 6, MINFBE exhibits fast asymptotic convergence, and approaches the solution much faster then fast FBS and ADMM. Unlike ADMM, no tuning of γ is needed in MINFBE to obtain fast convergence.

5.4 Matrix completion

We consider the problem of recovering the entries of an m-by-n matrix, which is known to have small rank, from a sample of them. One may refer to [63] for a detailed theoretical analysis of the problem. The decision variable is now a matrix $x = (x_{ij}) \in \mathbb{R}^{m\times n}$, and the problem has the form

$$f(x) = \tfrac{1}{2}\|\mathscr{A}(x) - b\|^2, \qquad g(x) = \lambda\|x\|_*,$$

where $\mathscr{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^k$ is a linear mapping selecting k entries from x, vector $b \in \mathbb{R}^k$ contains the known entries, and $\|x\|_*$ indicates the nuclear norm of x, which is the sum of its singular values. In this case $L_f = 1$, therefore we applied MINFBE with constant $\gamma = 0.95$.

The most computationally expensive operation here is the proximal step, requiring a singular value decomposition (SVD). Computing the full SVD becomes infeasible as m and n grow, therefore we use the following partial decomposition strategy

in evaluating $\mathrm{prox}_{\gamma g}$: start with $\nu_0 = 10$, and the i-th time $\mathrm{prox}_{\gamma g}$ is evaluated compute only the largest ν_i singular values $\sigma_1 \geq \ldots \geq \sigma_{\nu_i}$, and

$$\mathrm{prox}_{\gamma g}(x) \approx U\tilde{\Sigma}_+ V^T, \quad \tilde{\Sigma}_+ = \mathrm{diag}(\max\{0, \sigma_i - \gamma\lambda\}, i = 1, \ldots, \nu_i),$$

Then set ν_{i+1} according to the following rule

$$\nu_{i+1} = \begin{cases} \min\{j \mid \sigma_j \leq \gamma\} & \text{if } \sigma_{\nu_i} \leq \gamma\lambda \\ \nu_i + 5 & \text{otherwise.} \end{cases}$$

The same technique for approximately evaluating the singular value thresholing is used in other algorithms for nuclear norm regularization problems [64]. The partial singular value decompositions were performed using PROPACK software package.[4]

We compared fast FBS, L-BFGS, MINFBE(L-BFGS) and ADMM on the MovieLens100k dataset.[5] This consists of 10^5 ratings of 1682 movies from 943 users, so that the problem has ≈ 1.6 millions variables. The results of the simulations, for decreasing values of λ, are in Figure 7. Unlike the previous example, in this case MINFBE performs very similarly to standard L-BFGS: they both converge considerably faster than the accelerated FBS, and generally faster than ADMM, especially for smaller values of the regularization parameter. Note also that, just like in the previous example, the performance of ADMM is very sensitive to the value of parameter γ. In our experiment we identified $\gamma = 10$ as a good value by hand-tuning. Such tuning is not required in MINFBE, where the selection of a suitable γ is automatic.

5.5 Image restoration

As a nonconvex example we consider the restoration of a noisy blurred $M \times N$ image. The formulation we use is similar to that in [65], although here we consider the ℓ_1 norm in place of the ℓ_0 norm as regularization term. Specifically, we set

$$f(x) = \sum_{i=1}^{MN} \psi((Ax-b)_i), \qquad g(x) = \lambda\|Wx\|_1.$$

Here, b denotes the noisy blurred image, A is a Gaussian blur operator and W is a discrete Haar wavelet transform with four levels, while $\psi(t) = \log(1+t^2)$, therefore here ∇f has Lipschitz constant 2. Since $W^\top W = WW^\top = I$, the proximal mapping of g can be computed as

$$\mathrm{prox}_{\gamma g}(x) = W^\top \mathrm{prox}_{\gamma\|\cdot\|_1}(Wx). \tag{5.1}$$

We applied MINFBE to a 256×256-pixel black-and-white image. We distorted the original image with a Gaussian blur operator 9×9 with standard deviation 4, and with Gaussian noise with standard deviation 10^{-3}. The regularization parameter in (5.1) was set as $\lambda = 10^{-4}$. Results of the simulations are shown in Figures 8 and 9.

[4] http://sun.stanford.edu/~rmunk/PROPACK/

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

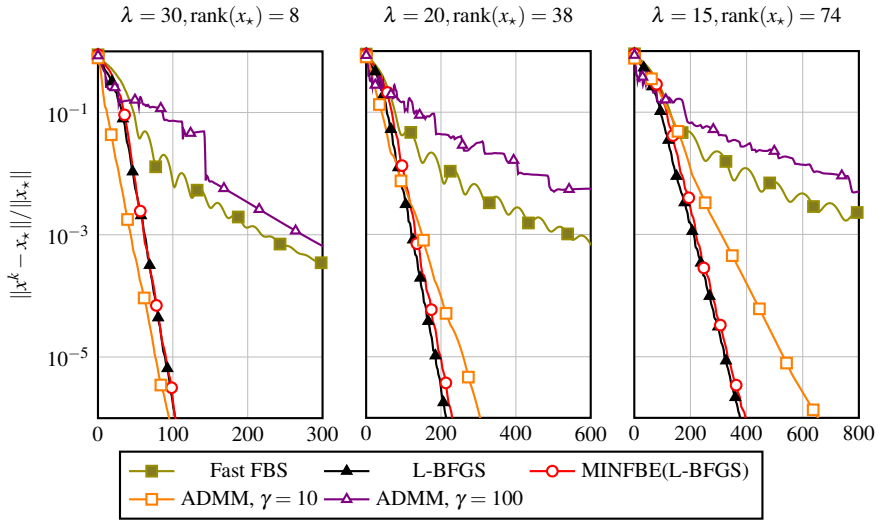

Fig. 7: Matrix completion: performance of MINFBE on the MovieLens100k dataset, for different values of λ. Horizontal axis is time in seconds.

6 Conclusions

The forward-backward splitting (FBS) algorithm for minimizing $\varphi = f + g$, where f is smooth and g is convex, is equivalent to a variable-metric gradient method applied to a continuously differentiable objective, which we called forward-backward envelope (FBE), when $f \in C^2$. Therefore, we can adopt advanced smooth unconstrained minimization algorithms, such as quasi-Newton and limited-memory methods, to the problem of minimizing the FBE and thus solving the original, nonsmooth problem. We propose to implement them in an algorithmic scheme, which we call MINFBE, which is appealing in that (i) it relies on the very same black-box oracle as FBS (evaluations of f, its gradient, g and its proximal mapping) and is therefore suited for large scale applications, (ii) it does not require the knowledge of global information such as Lipschitz constant L_f, but can adaptively estimate it. The proposed method exploits the composite structure of φ, and alternates line-search steps over descent directions and forward-backward steps. For this reason, MINFBE possesses the same global convergence properties of FBS, under the assumptions that φ has the Kurdyka-Łojasiewicz properties at its critical points, and a global convergence rate $O(1/k)$ in case φ is convex. This is a peculiar feature of our approach, since line-search methods do not converge to stationary points, in general, when applied to nonconvex functions. Moreover, we proved that when quasi-Newton directions are used in MINFBE, and the FBE is twice differentiable with nonsingular Hessian at the limit point of the sequence of iterates, superlinear asymptotic convergence is achieved. Our theoretical results are supported by numerical experiments. These show that MINFBE with (limited-memory) quasi-Newton directions improves the asymptotic convergence of

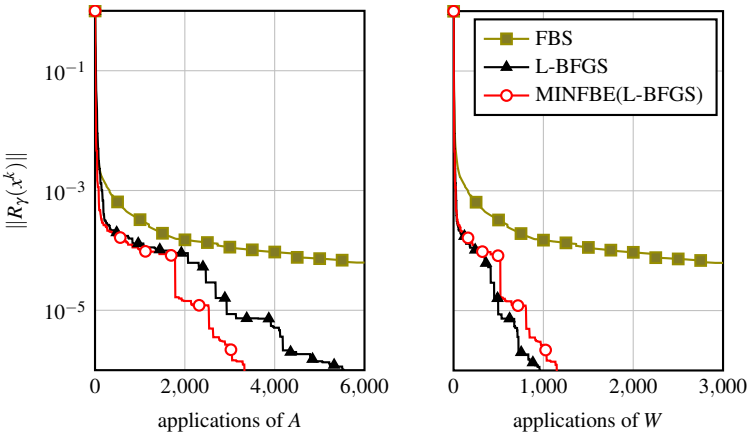

Fig. 8: (Nonconvex) image restoration: performance of MINFBE compared with FBS. On the horizontal axis, number of calls to the blur operator (left plot) and Haar operator (right plot); on the vertical axis the fixed-point residual R_γ. Original, noisy/blurred, and recovered images are shown in Figure 9.

(a) Original image (b) Noisy blurred image

(c) FBS (d) L-BFGS (e) Alg. 2 (L-BFGS)

Fig. 9: (Nonconvex) image restoration: recovered images obtained with the three considered algorithms.

FBS (and its accelerated variant when φ is convex), and usually converges faster than the corresponding classical line-search method applied to the problem of minimizing the FBE.

References

1. J.-J. Moreau, "Proximité et dualité dans un espace Hilbertien," *Bull. Soc. Math. France*, vol. 93, pp. 273–299, 1965.
2. P.-L. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM Journal on Numerical Analysis*, vol. 16, no. 6, pp. 964–979, 1979.
3. P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp. 185–212, 2011.
4. S. Łojasiewicz, "Une propriété topologique des sous-ensembles analytiques réels," *Les équations aux dérivées partielles*, pp. 87–89, 1963.
5. ——, "Sur la géométrie semi-et sous-analytique," in *Annales de l'institut Fourier*, vol. 43, no. 5, 1993, pp. 1575–1595.
6. K. Kurdyka, "On gradients of functions definable in o-minimal structures," in *Annales de l'institut Fourier*, vol. 48, no. 3, 1998, pp. 769–783.
7. H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods," *Mathematical Programming*, vol. 137, no. 1-2, pp. 91–129, 2013.
8. H. Attouch and J. Bolte, "On the convergence of the proximal algorithm for nonsmooth functions involving analytic features," *Mathematical Programming*, vol. 116, no. 1-2, pp. 5–16, 2009.
9. H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality," *Mathematics of Operations Research*, vol. 35, no. 2, pp. 438–457, 2010.
10. J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, no. 1-2, pp. 459–494, 2014.
11. P. Ochs, Y. Chen, T. Brox, and T. Pock, "iPiano: Inertial proximal algorithm for nonconvex optimization," *SIAM Journal on Imaging Sciences*, vol. 7, no. 2, pp. 1388–1419, 2014.
12. Y. Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$," *Soviet Mathematics Doklady*, 1983.
13. P. Tseng, "On accelerated proximal gradient methods for convex-concave optimization," Department of Mathematics, University of Washington, Tech. Rep., 2008.
14. A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.
15. Y. Nesterov, "Gradient methods for minimizing composite functions," *Mathematical Programming*, vol. 140, no. 1, pp. 125–161, 2013.
16. S. Becker and M. J. Fadili, "A quasi-Newton proximal splitting method," in *Advances in Neural Information Processing Systems 25*, P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., 2012, vol. 1, pp. 2618–2626.
17. J. Lee, Y. Sun, and M. Saunders, "Proximal Newton-type methods for convex optimization," in *Advances in Neural Information Processing Systems*, 2012, pp. 836–844.
18. K. Scheinberg and X. Tang, "Practical inexact proximal quasi-Newton method with global complexity analysis," *arXiv preprint arXiv:1311.6547*, 2013.
19. P. Patrinos and A. Bemporad, "Proximal Newton methods for convex composite optimization," in *IEEE Conference on Decision and Control*, 2013, pp. 2358–2363.
20. M. Fukushima, "Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems," *Mathematical programming*, vol. 53, no. 1, pp. 99–110, 1992.
21. N. Yamashita, K. Taji, and M. Fukushima, "Unconstrained optimization reformulations of variational inequality problems," *Journal of Optimization Theory and Applications*, vol. 92, no. 3, pp. 439–456, 1997.
22. F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003, vol. II.
23. W. Li and J. Peng, "Exact penalty functions for constrained minimization problems via regularized gap function for variational inequalities," *Journal of Global Optimization*, vol. 37, pp. 85–94, 2007.

24. P. Patrinos, P. Sopasakis, and H. Sarimveis, "A global piecewise smooth Newton method for fast large-scale model predictive control," *Automatica*, vol. 47, pp. 2016–2022, 2011.
25. J. E. Dennis and J. J. Moré, "A characterization of superlinear convergence and its application to quasi-Newton methods," *Mathematics of computation*, vol. 28, no. 126, pp. 549–560, 1974.
26. Y.-H. Dai, "Convergence Properties of the BFGS Algoritm," *SIAM Journal on Optimization*, vol. 13, no. 3, pp. 693–701, 2002.
27. W. F. Mascarenhas, "The BFGS method with exact line searches fails for non-convex objective functions," *Mathematical Programming*, vol. 99, no. 1, pp. 49–61, 2004.
28. ——, "On the divergence of line search methods," *Computational & Applied Mathematics*, vol. 26, pp. 129 – 169, 2007.
29. Y. H. Dai, "A perfect example for the BFGS method," *Mathematical Programming*, vol. 138, pp. 501–530, 2013.
30. R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Springer, 2011, vol. 317.
31. H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.
32. P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling & Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005.
33. J. E. Dennis Jr and R. B. Schnabel, *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM, 1996, vol. 16.
34. D. Bertsekas, *Nonlinear programming*. Athena Scientific, 1999.
35. C. Lemaréchal and C. Sagastizábal, "Practical aspects of the Moreau–Yosida regularization: Theoretical preliminaries," *SIAM Journal on Optimization*, vol. 7, no. 2, pp. 367–385, 1997.
36. D. S. Bernstein, *Matrix mathematics: theory, facts, and formulas with application to linear systems theory*. Woodstock: Princeton University Press, 2009.
37. R. T. Rockafellar, "First- and second-order epi-differentiability in nonlinear programming," *Transactions of the American Mathematical Society*, no. 307, pp. 75–108, 1988.
38. R. Rockafellar, "Second-order optimality conditions in nonlinear programming obtained by way of epi-derivatives," *Mathematics of Operations Research*, vol. 14, no. 3, pp. 462–484, 1989.
39. R. A. Poliquin and R. T. Rockafellar, "Amenable functions in optimization," *Nonsmooth optimization: methods and applications (Erice, 1991)*, pp. 338–353, 1992.
40. ——, "Second-order nonsmooth analysis in nonlinear programming," *Recent advances in nonsmooth optimization*, pp. 322–349, 1995.
41. R. T. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM Journal on Control and Optimization*, vol. 14, no. 5, pp. 877–898, 1976.
42. M. Fukushima and L. Qi, "A globally and superlinearly convergent algorithm for nonsmooth convex minimization," *SIAM Journal on Optimization*, vol. 6, no. 4, pp. 1106–1120, 1996.
43. J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal, "A family of variable metric proximal methods," *Mathematical Programming*, vol. 68, no. 1, pp. 15–47, 1995.
44. R. Mifflin, D. Sun, and L. Qi, "Quasi-Newton bundle-type methods for nondifferentiable convex optimization," *SIAM Journal on Optimization*, vol. 8, no. 2, pp. 583–603, 1998.
45. X. Chen and M. Fukushima, "Proximal quasi-Newton methods for nondifferentiable convex optimization," *Mathematical Programming*, vol. 85, no. 2, pp. 313–334, 1999.
46. J. V. Burke and M. Qian, "On the superlinear convergence of the variable metric proximal point algorithm using Broyden and BFGS matrix secant updating," *Mathematical Programming*, vol. 88, no. 1, pp. 157–181, 2000.
47. N. Sagara and M. Fukushima, "A trust region method for nonsmooth convex optimization," *Management*, vol. 1, no. 2, pp. 171–180, 2005.
48. W. Squire and G. Trapp, "Using complex variables to estimate derivatives of real functions," *Siam Review*, vol. 40, no. 1, pp. 110–112, 1998.
49. J. Nocedal and S. Wright, *Numerical optimization*. Springer Science & Business Media, 2006.
50. D. Noll and A. Rondepierre, *Computational and Analytical Mathematics: In Honor of Jonathan Borwein's 60th Birthday*. New York, NY: Springer New York, 2013, ch. Convergence of Linesearch and Trust-Region Methods Using the Kurdyka–Łojasiewicz Inequality, pp. 593–611.
51. J. Bolte, A. Daniilidis, and A. Lewis, "The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems," *SIAM Journal on Optimization*, vol. 17, no. 4, pp. 1205–1223, 2007.
52. J. E. Dennis and J. J. Moré, "Quasi-Newton methods, motivation and theory," *SIAM review*, vol. 19, no. 1, pp. 46–89, 1977.

53. M. Powell, "Some global convergence properties of a variable metric algorithm for minimization without exact line searches," *Nonlinear programming*, vol. 9, pp. 53–72, 1976.
54. D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, vol. 45, no. 1-3, pp. 503–528, 1989.
55. C.-M. Ip and J. Kyparisis, "Local convergence of quasi-Newton methods for B-differentiable equations," *Mathematical Programming*, vol. 56, no. 1-3, pp. 71–89, 1992.
56. J. Nocedal, "Updating quasi-Newton matrices with limited storage," *Mathematics of computation*, vol. 35, no. 151, pp. 773–782, 1980.
57. D.-H. Li and M. Fukushima, "On the global convergence of the bfgs method for nonconvex unconstrained optimization problems," *SIAM Journal on Optimization*, vol. 11, no. 4, pp. 1054–1064, 2001.
58. J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms I: Part 1: Fundamentals*. Springer Science & Business Media, 1996, vol. 305.
59. R. Fletcher, *Practical methods of optimization*. John Wiley & Sons, 2013.
60. Y.-H. Dai and Y. Yuan, "A nonlinear conjugate gradient method with a strong global convergence property," *SIAM Journal on Optimization*, vol. 10, no. 1, pp. 177–182, 1999.
61. S. J. Wright, R. D. Nowak, and M. A. Figueiredo, "Sparse reconstruction by separable approximation," *Signal Processing, IEEE Transactions on*, vol. 57, no. 7, pp. 2479–2493, 2009.
62. Z. Wen, W. Yin, D. Goldfarb, and Y. Zhang, "A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation," *SIAM Journal on Scientific Computing*, vol. 32, no. 4, pp. 1832–1857, 2010.
63. E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.
64. K.-C. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific Journal of Optimization*, vol. 6, no. 3, pp. 615–640, 2010.
65. 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*, pp. 1–23, 2014.
66. J.-S. Pang, "Newton's method for B-differentiable equations," *Mathematics of Operations Research*, vol. 15, no. 2, pp. 311–341, 1990.
67. R. Poliquin and R. Rockafellar, "Generalized Hessian properties of regularized nonsmooth functions," *SIAM Journal on Optimization*, vol. 6, no. 4, pp. 1121–1137, 1996.
68. R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.
69. R. H. Byrd and J. Nocedal, "A tool for the analysis of quasi-Newton methods with application to unconstrained minimization," *SIAM Journal on Numerical Analysis*, vol. 26, no. 3, pp. 727–739, 1989.

Appendix A Definitions and known results

Given a differentiable mapping $G:\mathbb{R}^n\to\mathbb{R}^m$ we let $JG:\mathbb{R}^n\to\mathbb{R}^{m\times n}$ denote the *Jacobian* of G. When $m=1$ we indicate with $\nabla G = JG^\top$ the *gradient* of G and with $\nabla^2 G = J\nabla G^\top$ its *Hessian*, whenever it makes sense. We say that G is *strictly differentiable* at $\bar{x}$ if it satisfies the stronger limit

$$\lim_{\substack{(x,y)\to(\bar{x},\bar{x})\\ x\neq y}} \frac{\|G(y)-G(x)-JG(\bar{x})[y-x]\|}{\|y-x\|} = 0$$

The next result states that strict differentiability is preserved by composition; its proof is a trivial computation and is therefore omitted.

Proposition A.1. *Let $F:\mathbb{R}^n\to\mathbb{R}^m$, $P:\mathbb{R}^m\to\mathbb{R}^k$. Suppose that F and P are (strictly) differentiable at $\bar{x}$ and $F(\bar{x})$, respectively. Then the composition $T=P\circ F$ is (strictly) differentiable at $\bar{x}$.*

Similarly, the product of (strictly) differentiable functions is still (strictly) differentiable. However, if one of the two functions vanishes at one point, then we may relax some assumptions, as it is proved in the next result.

Proposition A.2. *Let $Q:\mathbb{R}^n\to\mathbb{R}^{m\times k}$ and $R:\mathbb{R}^n\to\mathbb{R}^k$, and suppose that $R(\bar{x})=0$. If Q is continuous at $\bar{x}$ (resp. Lipschitz-continuous around $\bar{x}$) and R is differentiable (resp. strictly differentiable) at $\bar{x}$, then their product $G:\mathbb{R}^n\to\mathbb{R}^m$ defined as $G(x)=Q(x)R(x)$ is differentiable (resp. strictly differentiable) at $\bar{x}$ with $JG(\bar{x})=Q(\bar{x})JR(\bar{x})$.*

Proof. Suppose first that Q is continuous at $\bar{x}$ and R is differentiable at $\bar{x}$. Then, expanding $R(x)$ at $\bar{x}$ and since $G(\bar{x}) = 0$, we obtain

$$\begin{aligned}\frac{G(x) - G(\bar{x}) - Q(\bar{x})JR(\bar{x})[x-\bar{x}]}{\|x-\bar{x}\|} &= \frac{Q(x)R(x) - Q(\bar{x})JR(\bar{x})[x-\bar{x}]}{\|x-\bar{x}\|} \\ &= \frac{(Q(x)-Q(\bar{x}))\,JR(\bar{x})[x-\bar{x}] + o(\|x-\bar{x}\|)}{\|x-\bar{x}\|}\end{aligned}$$

The quantity $JR(\bar{x})[\frac{x-\bar{x}}{\|x-\bar{x}\|}]$ is bounded, and continuity of Q at $\bar{x}$ implies that taking the limit for $\bar{x} \neq x \to \bar{x}$ yields 0. This proves that G is differentiable at $\bar{x}$.

Suppose now that Q is Lipschitz-continuous around $\bar{x}$, and that R is strictly differentiable at $\bar{x}$. Then, expanding $R(y)$ at x we obtain

$$\begin{aligned}\frac{G(y) - G(x) - Q(\bar{x})JR(\bar{x})[y-x]}{\|y-x\|} = &\quad \frac{(Q(y)-Q(\bar{x}))JR(\bar{x})[y-x]}{\|y-x\|} \\ &+ \frac{(Q(y)-Q(x))R(x) + Q(y)o(\|x-y\|)}{\|y-x\|}\end{aligned}$$

The quantity $JR(\bar{x})[\frac{y-x}{\|y-x\|}]$ is bounded, and by Lipschitz-continuity of Q at $\bar{x}$ so is $\frac{Q(x)-Q(y)}{\|x-y\|}$ for x, y sufficiently close to $\bar{x}$. Taking the limit for $(\bar{x},\bar{x}) \neq (x,y) \to (\bar{x},\bar{x})$ with $x \neq y$ in the above expression then yields 0, proving strict differentiability. Uniqueness of the Jacobian proves also the claimed form of $JG(\bar{x})$. □

Definition A.3. *A mapping $G : \mathbb{R}^n \to \mathbb{R}^m$ is said to be* semidifferentiable *(or B-*differentiable *[55, 66]) at a point $\bar{x} \in \mathbb{R}^n$ if there exists a positively homogeneous mapping $DG(\bar{x})[\,\cdot\,] : \mathbb{R}^n \to \mathbb{R}^m$ such that*

$$\lim_{x \to \bar{x}} \frac{\|G(x) - G(\bar{x}) - DG(\bar{x})[x-\bar{x}]\|}{\|x-\bar{x}\|} = 0.$$

It is strictly semidifferentiable *at $\bar{x}$ if the stronger limit holds*

$$\lim_{\substack{(x,y)\to(\bar{x},\bar{x}) \\ x \neq y}} \frac{\|G(y) - G(x) - DG(\bar{x})[y-x]\|}{\|y-x\|} = 0.$$

$DG(\bar{x})$ is called semiderivative *of G at $\bar{x}$. If G is (strictly) semidifferentiable at every point of a set S, then it is said to be (strictly) semidifferentiable in S.*

Proposition A.4 ([66, Thm. 2])**.** *Suppose that $G : \mathbb{R}^n \to \mathbb{R}^m$ is semidifferentiable in a neighborhood of $\bar{x} \in \mathbb{R}^n$. Then, the following are equivalent:*

(a) $DG(\,\cdot\,)[d]$ is continuous in its first argument at $\bar{x}$ for all $d \in \mathbb{R}^n$;

(b) G is strictly semidifferentiable at $\bar{x}$;

(c) G is strictly (Fréchet) differentiable at $\bar{x}$.

Proposition A.5. *Suppose that $G : \mathbb{R}^n \to \mathbb{R}^m$ is semidifferentiable in a neighborhood N of $\bar{x}$ and that DG is* calm *at $\bar{x}$,* i.e.*, there exists $L > 0$ such that, for all $x \in N$ and $d \in \mathbb{R}^n$ with $\|d\| = 1$,*

$$\|DG(x)[d] - DG(\bar{x})[d]\| \leq L\|x-\bar{x}\|.$$

Then,

$$\|G(x) - G(y) - DG(\bar{x})[x-y]\| \leq L \max\{\|x-\bar{x}\|, \|y-\bar{x}\|\}\,\|x-y\|$$

Proof. Follows from [55, Lem. 2.2] by observing that the assumption of Lipschitz-continuity may be relaxed to calmness. □

Appendix B Proofs of Section 2

Proof of Lemma 2.9.

We know from [67, Thms. 3.8, 4.1] that $\operatorname{prox}_{\gamma g}$ is (strictly) differentiable at $x-\gamma\nabla f(x)$ if and only if g satisfies Assumption 4 (Assumption 5) at x for $-\nabla f(x)$. Since $f\in C^2$ by assumption, then in particular ∇f is strictly differentiable. The formula (2.7) follows from Proposition A.1 with $P=\operatorname{prox}_{\gamma g}$ and $F(x)=x-\gamma\nabla f(x)$.

Matrix $Q_\gamma(x)$ is symmetric since $f\in C^2$ and positive definite if $\gamma<1/L_f$. To obtain an expression for $P_\gamma(x)=J\operatorname{prox}_{\gamma g}(x-\gamma\nabla f(x))$ we can apply [30, Ex. 13.45] to the *tilted* function $g+\langle\nabla f(x),\cdot\rangle$ so that, letting $\mathrm{d}^2 g=\mathrm{d}^2 g(x|-\nabla f(x))[\cdot]$ and Π_S the idempotent and symmetric projection matrix on S,

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

where † indicates the pseudo-inverse, and last equality is due to [36, Facts 6.4.12(i)-(ii) and 6.1.6(xxxii)] and the properties of M as stated in Assumption 4. Apparently $P_\gamma(x)\succeq 0$ is symmetric, with $\|P_\gamma(x)\|\leq 1$. □

Proof of Theorem 2.11.

If follows from Theorem 2.10 that the Hessian $\nabla^2\varphi_\gamma(x)$ exists and is symmetric. Moreover, from [30, Ex. 13.18] we know that for all $d\in\mathbb{R}^n$

$$\begin{aligned}
\mathrm{d}^2\varphi(x|0)[d] &= \langle d,\nabla^2 f(x)d\rangle+\mathrm{d}^2 g(x|-\nabla f(x))[d]\\
&= \langle d,\nabla^2 f(x)d\rangle+\langle d,Md\rangle+\delta_S(d).
\end{aligned}\tag{B.1}$$

2.11(a) $\Leftrightarrow$ 2.11(b): Follows directly from (B.1), using [30, Thm. 13.24(c)].

2.11(c) $\Leftrightarrow$ 2.11(d): Letting $Q=Q_\gamma(x)$, we see from (2.7) and (2.9) that $JR_\gamma(x)$ is similar to the symmetric matrix $Q^{-1/2}\nabla^2\varphi_\gamma(x)Q^{-1/2}$, which is positive definite if and only if $\nabla^2\varphi_\gamma(x)$ is.

2.11(b) $\Leftrightarrow$ 2.11(c): From the point above we know that $JR_\gamma(x)$ has all real eigenvalues, and it can be easily seen to be similar to $\gamma^{-1}(I-QP)$, where $P=P_\gamma(x)$. From [68, Theorem 7.7.3] it follows that $\lambda_{\min}(I-QP)>0$ if and only if $Q^{-1}\succ P$. For all $d\in S$, using (2.8) we have

$$\begin{aligned}
\langle d,(Q^{-1}-P)d\rangle &= \langle d,Q^{-1}d\rangle-\langle d,\Pi_S[I+\gamma M]^{-1}\Pi_S d\rangle\\
&= \langle d,Q^{-1}d\rangle-\langle \Pi_S d,[I+\gamma M]^{-1}\Pi_S d\rangle\\
&= \langle d,Q^{-1}d\rangle-\langle d,[I+\gamma M]^{-1}d\rangle
\end{aligned}$$

and last quantity is positive if and only if $I+\gamma M\succ Q$ on S. By definition of Q, we then have that this holds if and only if $\nabla^2 f(x)+M\succ 0$ on S, which is 2.11(b).

2.11(d) $\Leftrightarrow$ 2.11(e): Trivial since $\nabla^2\varphi_\gamma(x)$ exists. □

Appendix C Proofs of Section 3

The following results are instrumental in proving convergence of the iterates of MINFBE.

Lemma C.1. *Under Assumption 1, consider the sequences $(x^k)_{k\in\mathbb{N}}$ and $(w^k)_{k\in\mathbb{N}}$ generated by MINFBE. If there exist $\bar\tau, c > 0$ such that $\tau_k \le \bar\tau$ and $\|d^k\| \le c\|R_{\gamma_k}(x^k)\|$, then*

$$\|x^{k+1} - x^k\| \le \gamma_k \|R_{\gamma_k}(w^k)\| + \bar\tau c \|R_{\gamma_k}(x^k)\| \qquad \forall k \in \mathbb{N} \tag{C.1}$$

and, for k large enough,

$$\|x^{k+1} - x^k\| \le \gamma_k \|R_{\gamma_k}(w^k)\| + \bar\tau c(1+\gamma_k L_f)\|R_{\gamma_{k-1}}(w^{k-1})\| \tag{C.2}$$

Proof. Equation (C.1) follows simply by

$$\|x^{k+1} - x^k\| = \|x^{k+1} - w^k + \tau_k d^k\| \le \gamma_k \|R_{\gamma_k}(w^k)\| + \bar\tau c \|R_{\gamma_k}(x^k)\|.$$

Now, for k sufficiently large $\gamma_k = \gamma_{k-1} = \gamma_\infty > 0$, see Lemma 3.1, and

$$\begin{aligned}
\|R_{\gamma_k}(x^k)\| &= \gamma_k^{-1}\|x^k - T_{\gamma_k}(x^k)\| \\
&= \gamma_k^{-1}\|T_{\gamma_k}(w^{k-1}) - T_{\gamma_k}(x^k)\| \\
&\le \gamma_k^{-1}\|w^{k-1} - \gamma_k \nabla f(w^{k-1}) - x^k + \gamma_k \nabla f(x^k)\| \\
&\le \gamma_k^{-1}\|w^{k-1} - x^k\| + \|\nabla f(w^{k-1}) - \nabla f(x^k)\| \\
&\le (1+\gamma_k L_f)\|R_{\gamma_{k-1}}(w^{k-1})\|,
\end{aligned}$$

where the first inequality follows from nonexpansiveness of $\operatorname{prox}_{\gamma g}$, and the last one from Lipschitz continuity of ∇f. Putting this together with (C.1) gives (C.2). □

Lemma C.2. *Let $(\beta_k)_{k\in\mathbb{N}}$ and $(\delta_k)_{k\in\mathbb{N}}$ be real sequences satisfying $\beta_k \ge 0$, $\delta_k \ge 0$, $\delta_{k+1} \le \delta_k$ and $\beta_{k+1}^2 \le (\delta_k - \delta_{k+1})\beta_k$ for all $k \in \mathbb{N}$. Then $\sum_{k=0}^\infty \beta_k < \infty$.*

Proof. Taking the square root of both sides in $\beta_{i+1}^2 \le (\delta_i - \delta_{i+1})\beta_i$ and using

$$\sqrt{\zeta\eta} \le (\zeta + \eta)/2,$$

for any nonnegative numbers ζ, η, we arrive at $2\beta_{i+1} \le (\delta_i - \delta_{i+1}) + \beta_i$. Summing up the latter for $i = 0, \dots, k$, for any $k \in \mathbb{N}$,

$$\begin{aligned}
2\textstyle\sum_{i=0}^k \beta_{i+1} &\le \sum_{i=0}^k (\delta_i - \delta_{i+1}) + \sum_{i=0}^k \beta_i \\
&= \delta_0 - \delta_{k+1} + \beta_0 - \beta_{k+1} + \textstyle\sum_{i=0}^k \beta_{i+1} \\
&\le \delta_0 + \beta_0 + \textstyle\sum_{i=0}^k \beta_{i+1}.
\end{aligned}$$

Hence

$$\sum_{i=0}^\infty \beta_{i+1} \le \delta_0 + \beta_0 < \infty, \tag{C.3}$$

which concludes the proof. □

Proposition C.3. *Suppose Assumption 1 is satisfied and that φ is lower bounded, and consider the sequences generated by MINFBE. If $\beta \in (0,1)$ and there exist $\bar\tau, c > 0$ such that $\tau_k \le \bar\tau$ and $\|d^k\| \le c\|R_{\gamma_k}(x^k)\|$ then*

$$\sum_{k=0}^\infty \|x^{k+1} - x^k\|^2 < \infty. \tag{C.4}$$

If moreover $(x^k)_{k\in\mathbb{N}}$ is bounded, then

$$\lim_{k\to\infty} \operatorname{dist}_{\omega(x^0)}(x^k) = 0 \tag{C.5}$$

and $\omega(x^0)$ is a nonempty, compact and connected subset of $\operatorname{zer}\partial\varphi$ over which φ is constant.

Proof. (C.4) follows from (C.1), Propositions 3.4(ii) and 3.4(iv), and the fact that the sum of square-summable sequences is square summable.

If $(x^k)_{k\in\mathbb{N}}$ is bounded, that $\omega(x^0)$ is nonempty, compact and connected and $\lim_{k\to\infty}\operatorname{dist}_{\omega(x^0)}(x^k)=0$ follow by [10, Lem. 5(ii),(iii), Remark 5]. That φ is constant on $\omega(x^0)$ follows by a similar argument as in [10, Lem. 5(iv)]. □

The following is [10, Lem. 6], therefore we state it with no proof.

Lemma C.4 (Uniformized KL property)**.** *Let $K\subset\mathbb{R}^n$ be a compact set and suppose that the proper lower semi-continuous function $\varphi:\mathbb{R}^n\to\overline{\mathbb{R}}$ is constant on K and satisfies the KL property at every $x^\star\in K$. Then there exist $\varepsilon>0$, $\eta>0$, and a continuous concave function $\psi:[0,\eta]\to[0,+\infty)$ such that properties 3.9(i), 3.9(ii) and 3.9(iii) hold, and*
(iv') for all $x_\star\in K$ and x such that $\operatorname{dist}_K(x)<\varepsilon$ and $\varphi(x_\star)<\varphi(x)<\varphi(x_\star)+\eta$,

$$\psi'(\varphi(x)-\varphi(x_\star))\operatorname{dist}(0,\partial\varphi(x))\geq 1. \tag{C.6}$$

Proof of Lemma 3.1.

Let $(\gamma_k)_{k\in\mathbb{N}}$ be the sequence of stepsize parameters computed by MINFBE. To arrive to a contradiction, suppose that k_0 is the smallest element of $\mathbb{N}$ such that

$$\gamma_{k_0}<\min\left\{\gamma_0,\sigma(1-\beta)/L_f\right\}.$$

Clearly, $k_0\geq 1$. Moreover $\sigma^{-1}\gamma_{k_0}$ must satisfy the condition in step 4: for some $w\in\mathbb{R}^n$ (corresponding to $w^k=x^k+\tau_k d^k$ selected before going back to step 1 after the condition in step 4 is passed, which might differ from the final value of w^k after step 4 is passed)

$$\varphi(T_{\sigma^{-1}\gamma_{k_0}}(w))>\varphi_{\sigma^{-1}\gamma_{k_0}}(w)-\frac{\beta\sigma^{-1}\gamma_{k_0}}{2}\|R_{\sigma^{-1}\gamma_{k_0}}(w)\|^2.$$

But from Proposition 2.2(ii) we also have

$$\begin{aligned}\varphi(T_{\sigma^{-1}\gamma_{k_0}}(w))&\leq\varphi_{\sigma^{-1}\gamma_{k_0}}(w)-\frac{\sigma^{-1}\gamma_{k_0}}{2}(1-\sigma^{-1}\gamma_{k_0}L_f)\|R_{\sigma^{-1}\gamma_{k_0}}(w)\|^2\\&\leq\varphi_{\sigma^{-1}\gamma_{k_0}}(w)-\frac{\beta\sigma^{-1}\gamma_{k_0}}{2}\|R_{\sigma^{-1}\gamma_{k_0}}(w)\|^2,\end{aligned}$$

where last inequality follows from $\sigma^{-1}\gamma_{k_0}<(1-\beta)/L_f$. This leads to a contradiction, therefore $\gamma_k\geq\min\left\{\gamma_0,\sigma(1-\beta)/L_f\right\}$ as claimed. That γ_k is asymptotically constant follows since the sequence $(\gamma_k)_{k\in\mathbb{N}}$ is nonincreasing. □

Proof of Proposition 3.4.

We have

$$\begin{aligned}\varphi(x^{k+1})&\leq\varphi_{\gamma_k}(w^k)-\tfrac{\beta\gamma_k}{2}\|R_{\gamma_k}(w^k)\|^2\\&\leq\varphi_{\gamma_k}(x^k)-\tfrac{\beta\gamma_k}{2}\|R_{\gamma_k}(w^k)\|^2\\&\leq\varphi(x^k)-\tfrac{\beta\gamma_k}{2}\|R_{\gamma_k}(w^k)\|^2-\tfrac{\gamma_k}{2}\|R_{\gamma_k}(x^k)\|^2,\end{aligned}\tag{C.7}$$

where the first inequality comes from step 4, the second from step 3 and the third from Proposition 2.2(i). This shows 3.4(i). Let $\varphi_\star=\lim_{k\to\infty}\varphi(x^k)$, which exists since $(\varphi(x^k))_{k\in\mathbb{N}}$ is monotone. If $\varphi_\star=-\infty$, clearly $\inf\varphi=-\infty$ and $\omega(x^0)=\emptyset$ due to properness and lower semicontinuity of φ and to the monotonic behavior of $(\varphi(x^k))_{k\in\mathbb{N}}$. Otherwise, telescoping the inequality we get

$$\frac{1}{2}\sum_{i=0}^{k}\gamma_i\left(\beta\|R_{\gamma_i}(w^i)\|^2+\|R_{\gamma_i}(x^i)\|^2\right)\leq\varphi(x^0)-\varphi(x^{k+1})\leq\varphi(x^0)-\varphi_\star\tag{C.8}$$

and since γ_k is uniformly lower bounded by a positive number (see Lemma 3.1) 3.4(ii) follows, hence 3.4(iii). If $\beta>0$, observing that for k large enough such that $\gamma_k\equiv\gamma_\infty$ we have

$$\varphi_{\gamma_k}(w^{k+1})\overset{\text{step 3}}{\leq}\varphi_{\gamma_k}(x^{k+1})\overset{\text{step 5}}{=}\varphi_{\gamma_k}(T_k(w^k))\leq\varphi_{\gamma_k}(w^k),$$

similar argumentations as those for proving 3.4(ii) show 3.4(iv). □

Proof of Theorem 3.5.
If $\inf\varphi = -\infty$ there is nothing to prove. Otherwise, since the sequence $(\gamma_k)_{k\in\mathbb{N}}$ is nonincreasing, from (C.8) we get

$$\frac{(k+1)\gamma_k}{2}\left(\min_{i=0\dots k}\|R_{\gamma_i}(x^i)\|^2+\beta\min_{i=0\dots k}\|R_{\gamma_i}(w^i)\|^2\right) \le \varphi(x^0)-\inf\varphi.$$

Rearranging the terms and invoking Lemma 3.1 gives the result. □

Proof of Theorem 3.6.
The proof is similar to that of [15, Thm. 4]. By Proposition 2.5(iii) we know that $\varphi_\gamma \le \varphi^\gamma$ for any $\gamma > 0$. Combining this with (C.7) we get

$$\varphi(x^{k+1}) \le \min_{x\in\mathbb{R}^n}\left\{\varphi(x)+\tfrac{1}{2\gamma_k}\|x-x^k\|^2\right\}, \tag{C.9}$$

and in particular, for $x_\star \in \operatorname{argmin}\varphi$,

$$\begin{aligned}\varphi(x^{k+1}) &\le \min_{\alpha\in[0,1]}\left\{\varphi(\alpha x_\star+(1-\alpha)x^k)+\tfrac{\alpha^2}{2\gamma_k}\|x^k-x_\star\|^2\right\}\\ &\le \min_{\alpha\in[0,1]}\left\{\varphi(x^k)-\alpha(\varphi(x^k)-\inf\varphi)+\tfrac{R^2}{2\gamma_k}\alpha^2\right\},\end{aligned}$$

where the last inequality follows by convexity of φ. If $\varphi(x^0)-\inf\varphi \ge R^2/\gamma_0$, then the optimal solution of the latter problem for $k=0$ is $\alpha=1$ and we obtain (3.1). Otherwise, the optimal solution is

$$\alpha = \frac{\gamma_k(\varphi(x^k)-\inf\varphi)}{R^2} \le \frac{\gamma_k(\varphi(x^0)-\inf\varphi)}{R^2} \le 1,$$

and we obtain

$$\varphi(x^{k+1}) \le \varphi(x^k) - \frac{\gamma_k(\varphi(x^k)-\inf\varphi)^2}{2R^2}.$$

Letting $\lambda_k = \frac{1}{\varphi(x^k)-\inf\varphi}$ the latter inequality is expressed as

$$\frac{1}{\lambda_{k+1}} \le \frac{1}{\lambda_k} - \frac{\gamma_k}{2R^2\lambda_{k+1}^2}.$$

Multiplying both sides by $\lambda_k\lambda_{k+1}$ and rearranging

$$\lambda_{k+1} \ge \lambda_k + \frac{\gamma_k}{2R^2}\frac{\lambda_{k+1}}{\lambda_k} \ge \lambda_k + \frac{\gamma_k}{2R^2},$$

where the latter inequality follows from the fact that $(\varphi(x^k))_{k\in\mathbb{N}}$ is nonincreasing, cf. Proposition 3.4(i). Telescoping the inequality and using Lemma 3.1, we obtain

$$\lambda_k \ge \lambda_0 + \frac{k\min\{\gamma_0,\sigma(1-\beta)/L_f\}}{2R^2} \ge \frac{k\min\{\gamma_0,\sigma(1-\beta)/L_f\}}{2R^2}.$$

Rearranging, we arrive at (3.2). □

Proof of Theorem 3.7.
If (3.3) holds, then φ has bounded level sets and $\operatorname{zer}\partial\varphi = \{x_\star\}$. In particular, $\omega(x^0)\neq\emptyset$ and Proposition 3.4(iii) then ensures $x^k \to x_\star$. Therefore, there is $k_0\in\mathbb{N}$ such that $x^k \in N$ for all $k \ge k_0$. Inequality (C.9) holds, and in particular for $k \ge k_0$

$$\begin{aligned}\varphi(x^{k+1}) &\le \min_{\alpha\in[0,1]}\left\{\varphi(\alpha x_\star+(1-\alpha)x^k)+\tfrac{\alpha^2}{2\gamma_k}\|x_\star-x^k\|^2\right\}\\ &\le \min_{\alpha\in[0,1]}\left\{\varphi(x^k)+\alpha\left(\tfrac{\alpha}{c\gamma_k}-1\right)(\varphi(x^k)-\inf\varphi)\right\},\end{aligned}$$

where the second inequality follows by convexity of φ and (3.3). The minimum of last expression is achieved for $\alpha = \min\{1, \frac{c}{2}\gamma_k\}$. When $\gamma_k < 2c^{-1}$ we have the bound

$$\varphi(x^{k+1}) - \inf\varphi \leq (1 - \tfrac{c}{4}\gamma_k)(\varphi(x^k) - \inf\varphi).$$

When instead $\gamma_k \geq 2c^{-1}$ we have the bound

$$\varphi(x^{k+1}) - \inf\varphi \leq (c\gamma_k)^{-1}(\varphi(x^k) - \inf\varphi) \leq \tfrac{1}{2}(\varphi(x^k) - \inf\varphi).$$

Therefore $\varphi(x^{k+1}) - \inf\varphi \leq \omega(\varphi(x^k) - \inf\varphi)$, where

$$\begin{aligned}\omega &\leq \sup_k \max\{\tfrac{1}{2}, 1 - \tfrac{c}{4}\gamma_k\} \\ &\leq \max\{\tfrac{1}{2}, 1 - \tfrac{c}{4}\min\{\gamma_0, \sigma(1-\beta)/L_f\}\} \in [\tfrac{1}{2}, 1),\end{aligned}$$

last inequality following from Lemma 3.1. This proves the claim on the sequence $(\varphi(x^k))_{k\geq k_0}$ and using inequality (C.7) the same holds for $(\varphi_{\gamma_k}(w^k))_{k\geq k_0}$. From the error bound (3.3) we obtain that $x^k \to x_\star$ R-linearly. If the same error bound holds for φ_{γ_∞}, then also $w^k \to x_\star$ R-linearly. □

Proof of Theorem 3.10.

The case where the sequence is finite does not deserve any further investigation, therefore we assume that $(x^k)_{k\in\mathbb{N}}$ is infinite. We then assume that $R_{\gamma_k}(x^k) \neq 0$ which implies through Proposition 3.4 that $\varphi(x^{k+1}) < \varphi(x^k)$. Due to (C.5), the KL property for φ, and Lemma C.4, there exist $\varepsilon, \eta > 0$ and a continuous concave function $\psi : [0, \eta] \to [0, +\infty)$ such that for all x with $\operatorname{dist}_{\omega(x^0)}(x) < \varepsilon$ and $\varphi(x^\star) < \varphi(x) < \varphi(x_\star) + \eta$ one has

$$\psi'(\varphi(x) - \varphi(x_\star))\operatorname{dist}(0, \partial\varphi(x)) \geq 1.$$

According to Proposition C.3 there exists a $k_1 \in \mathbb{N}$ such that $\operatorname{dist}_{\omega(x^0)}(x^k) < \varepsilon$ for all $k \geq k_1$. Furthermore, since $\varphi(x^k)$ converges to $\varphi(x_\star)$ there exists a k_2 such that $\varphi(x^k) < \varphi(x_\star) + \eta$ for all $k \geq k_2$. Take $\bar{k} = \max\{k_1, k_2\}$. Then for every $k \geq \bar{k}$ we have

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

From Proposition 3.4(i)

$$\varphi(x^{k+1}) \leq \varphi(x^k) - \tfrac{\beta\gamma_k}{2}\|R_{\gamma_k}(w^k)\|^2.$$

For every $k > 0$ let $\tilde{\nabla}\varphi(x^k) = \nabla f(x^k) - \nabla f(w^{k-1}) + R_{\gamma_{k-1}}(w^{k-1})$. Since $R_{\gamma_{k-1}}(w^{k-1}) \in \nabla f(w^{k-1}) + \partial g(x^k)$, then $\tilde{\nabla}\varphi(x^k) \in \partial\varphi(x^k)$ and

$$\begin{aligned}\|\tilde{\nabla}\varphi(x^k)\| &\leq \|\nabla f(x^k) - \nabla f(w^{k-1})\| + \|R_{\gamma_{k-1}}(w^{k-1})\| \\ &= (1 + \gamma_{k-1}L_f)\|R_{\gamma_{k-1}}(w^{k-1})\|.\end{aligned}$$

From (C.6)

$$\psi'(\varphi(x^k) - \varphi(x_\star)) \geq \frac{1}{\|\tilde{\nabla}\varphi(x^k)\|} \geq \frac{1}{(1 + \gamma_{k-1}L_f)\|R_{\gamma_{k-1}}(w^{k-1})\|}.$$

Let $\Delta_k = \psi(\varphi(x^k) - \varphi(x_\star))$. By concavity of ψ and Proposition 3.4(i)

$$\begin{aligned}\Delta_k - \Delta_{k+1} &\geq \psi'(\varphi(x^k) - \varphi(x_\star))(\varphi(x^k) - \varphi(x^{k+1})) \\ &\geq \frac{\beta\gamma_k}{2(1 + \gamma_{k-1}L_f)} \frac{\|R_{\gamma_k}(w^k)\|^2}{\|R_{\gamma_{k-1}}(w^{k-1})\|} \\ &\geq \frac{\beta\gamma_{\min}}{2(1 + \gamma_0 L_f)} \frac{\|R_{\gamma_k}(w^k)\|^2}{\|R_{\gamma_{k-1}}(w^{k-1})\|}\end{aligned}$$

where $\gamma_{\min} = \min\{\gamma_0, \sigma(1-\beta)/L_f\}$, see Lemma 3.1, or

$$\|R_{\gamma_k}(w^k)\|^2 \leq \alpha(\Delta_k - \Delta_{k+1})\|R_{\gamma_{k-1}}(w^{k-1})\| \tag{C.10}$$

where $\alpha = 2(1+\gamma_0 L_f)/(\beta\gamma_{\min})$. Applying Lemma C.2 with

$$\delta_k = \alpha\Delta_k, \quad \beta_k = \|R_{\gamma_{k-1}}(w^{k-1})\|,$$

we conclude that $\sum_{k=0}^{\infty}\|R_{\gamma_k}(w^k)\| < \infty$. From (C.2), using the fact that $\gamma_k \le \gamma_0$ for all k, then it follows that

$$\sum_{k=0}^{\infty}\|x^{k+1}-x^k\| < \infty.$$

Then $(x^k)_{k\in\mathbb{N}}$ is a Cauchy sequence, hence it converges to a point that, by Proposition 3.4, is a critical point $x_\star$ of φ. □

Proof of Theorem 3.11.

Theorem 3.10 ensures that $(x^k)_{k\in\mathbb{N}}$ converges to a critical point, be it $x_\star$. We know from Lemma 3.1 that eventually $\gamma_k = \gamma_\infty > 0$, therefore we assume k is large enough for this purpose and indicate γ in place of γ_k for simplicity. Denoting $A_k = \sum_{i=k}^{\infty}\|x^{i+1}-x^i\|$ clearly $A_k \ge \|x^k - x_\star\|$, so we will prove that A_k converges linearly to zero to obtain the result. Note that by (C.2) we know that

$$\|x^{i+1}-x^i\| \le \gamma\|R_\gamma(w^i)\| + \bar\tau c(1+\gamma L_f)\|R_\gamma(w^{i-1})\|.$$

Therefore we can upper bound A_k as follows

$$\begin{aligned} A_k &\le \bar\tau c(1+\gamma L_f)\|R_\gamma(w^{k-1})\| + \big(\gamma + \bar\tau c(1+\gamma L_f)\big)\textstyle\sum_{i=k}^{\infty}\|R_\gamma(w^i)\| \\ &\le \big(\gamma + \bar\tau c(1+\gamma L_f)\big)\textstyle\sum_{i=k-1}^{\infty}\|R_\gamma(w^i)\|, \end{aligned} \tag{C.11}$$

and reduce the problem to proving linear convergence of $B_k = \sum_{i=k}^{\infty}\|R_\gamma(w^i)\|$. When ψ is as in (3.4), for sufficiently large k the KL inequality reads

$$\varphi(x^k) - \varphi(x_\star) \le [\sigma(1-\theta)\|v^k\|]^{\frac{1}{\theta}}, \quad \forall v^k \in \partial\varphi(x^k).$$

Taking $v^k = \nabla f(x^k) - \nabla f(w^{k-1}) + R_\gamma(w^{k-1}) \in \partial\varphi(x^k)$, this in turn yields

$$\varphi(x^k) - \varphi(x_\star) \le \Big[\sigma(1-\theta)(1+\gamma L_f)\|R_\gamma(w^{k-1})\|\Big]^{\frac{1}{\theta}}, \tag{C.12}$$

(see the proof of Theorem 3.10). Inequality (C.10) holds, for sufficiently large k, with $\Delta_k = \sigma(\varphi(x^k) - \varphi(x_\star))^{1-\theta}$ in this case. Applying Lemma C.2 with

$$\delta_k = \alpha\Delta_k, \quad \beta_k = \|R_\gamma(w^{k-1})\| = B_{k-1} - B_k,$$

we obtain

$$\begin{aligned} B_k &\le (B_{k-1}-B_k) + \sigma(\varphi(x^k) - \varphi(x_\star))^{1-\theta} \\ &\le (B_{k-1}-B_k) + \sigma\left[\sigma(1-\theta)(1+\gamma L_f)(B_{k-1}-B_k)\right]^{\frac{1-\theta}{\theta}}, \end{aligned}$$

where the second inequality is due to (C.12). Since $B_{k-1} - B_k \to 0$, then for k large enough it holds that $\sigma(1+\gamma L_f)(B_{k-1}-B_k) \le 1$, and the last term in the previous chain of inequalities is increasing in θ when $\theta \in (0, \frac{1}{2}]$. Therefore B_k eventually satisfies

$$B_k \le C(B_{k-1} - B_k),$$

where $C > 0$, and so $B_k \le [C/(1+C)]B_{k-1}$, *i.e.*, B_k converges to zero Q-linearly. This in turn implies that $\|x^k - x_\star\|$ converges to zero with R-linear rate. Furthermore,

$$\begin{aligned} \|w^k - x_\star\| &= \|x^k - x_\star + \tau_k d^k\| \\ &\le \|x^k - x_\star\| + \bar\tau c\|R_{\gamma_k}(x^k)\| \\ &= \|x^k - x_\star\| + \bar\tau c\gamma_k^{-1}\|T_{\gamma_k}(x^k) - x^k\| \\ &\le (1+\bar\tau c\gamma_k^{-1})\|x^k - x_\star\| + \bar\tau c\gamma_k^{-1}\|T_{\gamma_k}(x^k) - T_{\gamma_k}(x_\star)\| \\ &\le (1+\bar\tau c\gamma_k^{-1})\|x^k - x_\star\| + \bar\tau c\gamma_k^{-1}\|x^k - \gamma_k\nabla f(x^k) - x_\star + \gamma_k\nabla f(x_\star)\| \\ &\le (1+\bar\tau c(2\gamma_k^{-1} + L_f))\|x^k - x_\star\|, \end{aligned}$$

where the last two inequalities follow by nonexpansiveness of $\mathrm{prox}_{\gamma g}$ and Lipschitz continuity of ∇f. Since γ_k is lower bounded by a positive quantity, then we deduce that also $\|w^k - x_\star\|$ converges R-linearly to zero. □

Appendix D Proofs of Section 4

Proof of Theorem 4.1.

Since $w^k = x^k - B_k^{-1}\nabla\varphi_\gamma(x^k)$, letting $k \to \infty$ and using (4.1) we have that

$$\begin{aligned} 0 \leftarrow \frac{(B_k - \nabla^2\varphi_\gamma(x_\star))(w^k - x^k)}{\|w^k - x^k\|} &= -\frac{\nabla\varphi_\gamma(x^k) + \nabla^2\varphi_\gamma(x_\star)(w^k - x^k)}{\|w^k - x^k\|} \\ &= -\frac{\nabla\varphi_\gamma(x^k) - \nabla\varphi_\gamma(w^k) + \nabla^2\varphi_\gamma(x_\star)(w^k - x^k)}{\|w^k - x^k\|} \\ &\quad - \frac{\nabla\varphi_\gamma(w^k)}{\|w^k - x^k\|}. \end{aligned}$$

By strict differentiability of $\nabla\varphi_\gamma$ at $x_\star$ we obtain

$$\lim_{k\to\infty} \frac{\|\nabla\varphi_\gamma(w^k)\|}{\|w^k - x^k\|} = 0 \tag{D.1}$$

By nonsingularity of $\nabla^2\varphi_\gamma(x_\star)$ and since $w^k \to x^\star$, there exist $\alpha > 0$ such that $\|\nabla\varphi_\gamma(x^k)\| \geq \alpha\|x^k - x_\star\|$ for k large enough. Therefore, for k sufficiently large,

$$\frac{\|\nabla\varphi_\gamma(w^k)\|}{\|w^k - x^k\|} \geq \frac{\alpha\|w^k - x_\star\|}{\|w^k - x^k\|} \geq \frac{\alpha\|w^k - x_\star\|}{\|w^k - x_\star\| + \|x^k - x_\star\|}.$$

Using (D.1) we get

$$\lim_{k\to\infty} \frac{\|w^k - x_\star\|}{\|w^k - x_\star\| + \|x^k - x_\star\|} = \lim_{k\to\infty} \frac{\|w^k - x_\star\|/\|x^k - x_\star\|}{\|w^k - x_\star\|/\|x^k - x_\star\| + 1} = 0,$$

from which we obtain

$$\lim_{k\to\infty} \frac{\|w^k - x_\star\|}{\|x^k - x_\star\|} = 0. \tag{D.2}$$

Finally,

$$\begin{aligned} \|x^{k+1} - x_\star\| &= \|T_\gamma(w^k) - T_\gamma(x_\star)\| \\ &= \left\|\mathrm{prox}_{\gamma g}(w^k - \gamma\nabla f(w^k)) - \mathrm{prox}_{\gamma g}(x_\star - \gamma\nabla f(x_\star))\right\| \\ &\leq \left\|w^k - \gamma\nabla f(w^k) - x_\star + \gamma\nabla f(x_\star)\right\| \\ &\leq (1 + \gamma L_f)\|w^k - x_\star\|, \end{aligned} \tag{D.3}$$

where the first inequality follows from nonexpansiveness of $\mathrm{prox}_{\gamma g}$ and the second from Lipschitz continuity of ∇f. Using (D.3) in (D.2) we obtain that $(x^k)_{k\in\mathbb{N}}$ and $(w^k)_{k\in\mathbb{N}}$ converge Q-superlinearly to $x_\star$. □

Proof of Theorem 4.2.

From Proposition A.4(a) it follows that $\nabla\varphi_\gamma$ is strictly differentiable and continuosly semidifferentiable at $x_\star$. Moreover, we know from Lemma 3.1 that eventually $\gamma_k = \gamma_\infty > 0$. Therefore we assume that k is large enough for this purpose and indicate γ in place of γ_k for simplicity. We denote for short $g^k = \nabla\varphi_\gamma(x^k)$. In MINFBE

$$w^k - x^k = \tau_k d^k = -\tau_k B_k^{-1} g^k,$$

and by (4.1) and Cauchy-Schwarz inequality

$$\frac{\|(B_k - \nabla^2\varphi_\gamma(x_\star))(w^k - x^k)\|}{\|w^k - x^k\|} = \frac{\|g^k + \nabla^2\varphi_\gamma(x_\star)d^k\|}{\|d^k\|} \geq \left|\frac{\langle d^k, g^k + \nabla^2\varphi_\gamma(x_\star)d^k\rangle}{\|d^k\|^2}\right| \to 0.$$

Therefore

$$-\langle g^k, d^k\rangle = \langle d^k, \nabla^2\varphi_\gamma(x_\star)d^k\rangle + o(\|d^k\|^2). \tag{D.4}$$

Since $\nabla^2\varphi_\gamma(x_\star)$ is positive definite, then there is $\eta > 0$ such that for sufficiently large k

$$-\langle g^k, d^k\rangle \geq \eta\|d^k\|^2. \tag{D.5}$$

Since $D\nabla\varphi_\gamma$ is continuous at $x_\star$ and $x^k \to x_\star$, we have

$$\|D\nabla\varphi_\gamma(x^k)[d^k] - \nabla^2\varphi_\gamma(x_\star)d^k\| = o(\|d^k\|). \tag{D.6}$$

Next, since $x^k \to x_\star$, for k large enough $\nabla\varphi_\gamma$ is semidifferentiable at x^k and we can expand φ_γ around x^k using [30, Ex. 13.7(c)] to obtain

$$\begin{aligned}\varphi_\gamma(x^k + d^k) - \varphi_\gamma(x^k) &= \langle g^k, d^k\rangle + \tfrac{1}{2}\langle d^k, D\nabla\varphi_\gamma(x^k)[d^k]\rangle + o(\|d^k\|^2)\\ &= \langle g^k, d^k\rangle + \tfrac{1}{2}\langle d^k, \nabla^2\varphi_\gamma(x_\star)d^k\rangle + o(\|d^k\|^2)\\ &= \tfrac{1}{2}\langle g^k, d^k\rangle + o(\|d^k\|^2),\end{aligned}$$

where the second equality is due to (D.6), and the last equality is due to (D.4). Therefore, using (D.5), for sufficiently large k

$$\varphi_\gamma(x^k + d^k) - \varphi_\gamma(x^k) \leq -\tfrac{\eta}{2}\|d^k\|^2 < 0.$$

i.e., $\tau_k = 1$ satisfies the non-increase condition. As a consequence, MINFBE eventually reduces to the iterations of Theorem 4.1 and the proof follows. □

Proof of Theorem 4.3.

Suppose that Assumption 6(i) holds. Since $x_\star \in \operatorname{zer}\partial\varphi$ and $\nabla^2\varphi_\gamma(x_\star) \succ 0$, it follows that $x_\star$ is a strong local minimizer of φ_γ, hence of φ in light of Propositions 2.2(i) and 2.3(i). Theorem 3.7 then ensures that $(x^k)_{k\in\mathbb{N}}$ and $(w^k)_{k\in\mathbb{N}}$ converge linearly to $x_\star$. If instead $(\|B_k^{-1}\|)_{k\in\mathbb{N}}$ is bounded and Assumption 6(ii) holds, then Theorem 3.11 applies and again $(x^k)_{k\in\mathbb{N}}$ and $(w^k)_{k\in\mathbb{N}}$ converge linearly to a critical point, be it $x_\star$. In both cases we can apply Proposition A.5 and for k sufficiently large

$$\frac{\|y^k - \nabla^2\varphi_\gamma(x_\star)s^k\|}{\|s^k\|} \leq L\max\left\{\|w^k - x_\star\|, \|x^k - x_\star\|\right\}. \tag{D.7}$$

Since the convergence is linear, then the right-hand side of (D.7) is summable. With similar arguments to those of [25, Lem. 3.2] we can see that eventually $\langle s^k, y^k\rangle > 0$. Therefore we can apply [69, Thm. 3.2], which ensures that condition (4.1) holds. The result follows then from Theorem 4.2. □