On the acceleration of forward-backward splitting via an inexact Newton method

Andreas Themelis, Masoud Ahookhosh and Panagiotis Patrinos

Abstract We propose a Forward-Backward Truncated-Newton method (FBTN) for minimizing the sum of two convex functions, one of which smooth. Unlike other proximal Newton methods, our approach does not involve the employment of variable metrics, but is rather based on a reformulation of the original problem as the unconstrained minimization of a continuously differentiable function, the forward-backward envelope (FBE). We introduce a generalized Hessian for the FBE that symmetrizes the generalized Jacobian of the nonlinear system of equations representing the optimality conditions for the problem. This enables the employment of conjugate gradient method (CG) for efficiently solving the resulting (regularized) linear systems, which can be done inexactly. The employment of CG prevents the computation of full (generalized) Jacobians, as it requires only (generalized) directional derivatives. The resulting algorithm is globally (subsequentially) convergent, $Q$-linearly under an error bound condition, and up to $Q$-superlinearly and $Q$-quadratically under regularity assumptions at the possibly non-isolated limit point.

Key words: Forward-backward splitting, linear Newton approximation, truncated-Newton method, backtracking linesearch, error bound, superlinear convergence

AMS 2010 Subject Classification: 49J52, 49M15, 90C06, 90C25, 90C30

1 Introduction

In this work we focus on convex composite optimization problems of the form
minimize $\varphi(x) \equiv f(x) + g(x)$, \hspace{1cm} (1.1)

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex, twice continuously differentiable and with $L_f$-Lipschitz-continuous gradient, and $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ has a cheaply computable proximal mapping [51]. To ease the notation, throughout the paper we indicate

\[ \varphi_* := \inf \varphi \quad \text{and} \quad X_* := \arg\min \varphi. \]

Problems of the form (1.1) are abundant in many scientific areas such as control, signal processing, system identification, machine learning and image analysis, to name a few. For example, when $g$ is the indicator of a convex set then (1.1) becomes a constrained optimization problem, while for $f(x) = \frac{1}{2}\|Ax - b\|^2$ and $g(x) = \lambda\|x\|_1$, it becomes the $\ell_1$-regularized least-squares problem (lasso) which is the main building block of compressed sensing. When $g$ is equal to the nuclear norm, then (1.1) models low-rank matrix recovery problems. Finally, conic optimization problems such as linear, second-order cone, and semidefinite programs can be brought into the form of (1.1), see [31].

Perhaps the most well-known algorithm for problems in the form (1.1) is the forward-backward splitting (FBS) or proximal gradient method [40, 16], that interleaves gradient descent steps on the smooth function and proximal steps on the non-smooth one, see §3.1. Accelerated versions of FBS, based on the work of Nesterov [54, 5, 77], have also gained popularity. Although these algorithms share favorable global convergence rate estimates of order $O(\varepsilon^{-1})$ or $O(\varepsilon^{-1/2})$ (where $\varepsilon$ is the solution accuracy), they are first-order methods and therefore usually effective at computing solutions of low or medium accuracy only. An evident remedy is to include second-order information by replacing the Euclidean norm in the proximal mapping with that induced by the Hessian of $f$ at $x$ or some approximation of it, mimicking Newton or quasi-Newton methods for unconstrained problems [6, 32, 42]. However, a severe limitation of the approach is that, unless $Q$ has a special structure, the computation of the proximal mapping becomes very hard. For example, if $\varphi$ models a lasso problem, the corresponding subproblem is as hard as the original problem.

In this work we follow a different approach by reformulating the nonsmooth constrained problem (1.1) into the smooth unconstrained minimization of the forward-backward envelope (FBE) [57], a real-valued, continuously differentiable, exact penalty function for $\varphi$. Although the FBE might fail to be twice continuously differentiable, by using tools from nonsmooth analysis we show that one can design Newton-like methods to address its minimization, that achieve $Q$-superlinear asymptotic rates of convergence under nondegeneracy and (generalized) smoothness conditions on the set of solutions. Furthermore, by suitably interleaving FBS and Newton-like iterations the proposed algorithm also enjoys good complexity guarantees provided by a global (non-asymptotic) convergence rate. Unlike the approaches of [6, 32], where the corresponding subproblems are expensive to solve, our algorithm only requires the inexact solution of a linear system to compute the Newton-type direction, which can be done efficiently with a memory-free CG method.
Our approach 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 [23] and as an unconstrained continuously differentiable optimization problem via the $D$-gap function [79] (see [19, §10] for a survey and [38], [58] for applications to constrained optimization and model predictive control of dynamical systems).

1.1 Contributions

We propose an algorithm that addresses problem (1.1) by means of a Newton-like method on the FBE. Differently from a direct application of the classical Newton method, our approach does not require twice differentiability of the FBE (which would impose additional properties on $f$ and $g$), but merely twice differentiability of $f$. This is possible thanks to the introduction of an approximate generalized Hessian which only requires access to $\nabla^2 f$ and to the generalized (Clarke) Jacobian of the proximal mapping of $g$, as opposed to third-order derivatives and classical Jacobian, respectively. Moreover, it allows for inexact solutions of linear systems to compute the update direction, which can be done efficiently with a truncated CG method; in particular, no computation of full (generalized) Hessian matrices is necessary, as only (generalized) directional derivatives are needed. The method is thus particularly appealing when the Clarke Jacobians are sparse and/or well structured, so that the implementation of CG becomes extremely efficient. Under an error bound condition and a (semi)smoothness assumption at the limit point, which is not required to be isolated, the algorithm exhibits asymptotic $Q$-superlinear rates. For the reader’s convenience we collect explicit formulas of the needed Jacobians of the proximal mapping for a wide range of frequently encountered functions, and discuss when they satisfy the needed semismoothness requirements that enable superlinear rates.

1.2 Related work

This work is a revised version of the unpublished manuscript [59] and extends ideas proposed in [57], where the FBE is first introduced. Other FBE-based algorithms are proposed in [69, 75, 71]; differently from the truncated-CG type of approximation proposed here, they all employ quasi-Newton directions to mimic second-order information. The underlying ideas can also be extended to enhance other popular proximal splitting algorithms: the Douglas Rachford splitting (DRS) and the alternating direction method of multipliers (ADMM) [74], and for strongly convex problems also the alternating minimization algorithm (AMA) [70].
The algorithm proposed in this paper adopts the recent techniques investigated in [75, 71] to enhance and greatly simplify the scheme in [59]. In particular, $Q$-linear and $Q$-superlinear rates of convergence are established under an error bound condition, as opposed to uniqueness of the solution. The proofs of superlinear convergence with an error bound pattern the arguments in [83, 82], although with less conservative requirements.

1.3 Organization

The work is structured as follows. In Section 2 we introduce the adopted notation and list some known facts on generalized differentiability needed in the sequel. Section 3 offers an overview on the connections between FBS and the proximal point algorithm, and serves as a prelude to Section 4 where the forward-backward envelope function is introduced and analyzed. Section 5 deals with the proposed truncated-Newton algorithm and its convergence analysis. In Section 6 we collect explicit formulas for the generalized Jacobian of the proximal mapping of a rich list of nonsmooth functions, needed for computing the update directions in the proposed algorithm. Finally, Section 7 draws some conclusions.

2 Preliminaries

2.1 Notation and known facts

Our notation is standard and follows that of convex analysis textbooks [2, 8, 28, 65]. For the sake of clarity we now properly specify the adopted conventions, and briefly recap known definitions and facts in convex analysis. The interested reader is referred to the above-mentioned textbooks for the details.

Matrices and vectors. The $n \times n$ identity matrix is denoted as $I_n$, and the $\mathbb{R}^n$ vector with all elements equal to 1 is as $\mathbf{1}_n$; whenever $n$ is clear from context we simply write $I$ or $\mathbf{1}$, respectively. We use the Kronecker symbol $\delta_{i,j}$ for the $(i,j)$-th entry of $I$. Given $v \in \mathbb{R}^n$, with $\text{diag} v$ we indicate the $n \times n$ diagonal matrix whose $i$-th diagonal entry is $v_i$. With $\mathcal{S}(\mathbb{R}^n)$, $\mathcal{S}_+(\mathbb{R}^n)$ and $\mathcal{S}_{++}(\mathbb{R}^n)$ we denote respectively the set of symmetric, symmetric positive semidefinite, and symmetric positive definite matrices in $\mathbb{R}^{n \times n}$.

The minimum and maximum eigenvalues of $H \in \mathcal{S}(\mathbb{R}^n)$ are denoted as $\lambda_{\min}(H)$ and $\lambda_{\max}(H)$, respectively. For $Q, R \in \mathcal{S}(\mathbb{R}^n)$ we write $Q \succeq R$ to indicate that $Q - R \in \mathcal{S}_+(\mathbb{R}^n)$, and similarly $Q > R$ indicates that $Q - R \in \mathcal{S}_{++}(\mathbb{R}^n)$. Any matrix $Q \in \mathcal{S}_+(\mathbb{R}^n)$ induces the semi-norm $\| \cdot \|_Q$ on $\mathbb{R}^n$, where $\|x\|^2_Q := \langle x, Qx \rangle$; in case $Q = I$, that is, for the Euclidean norm, we omit the subscript and simply write $\| \cdot \|$. 
No ambiguity occurs in adopting the same notation for the induced matrix norm, namely $||M|| := \max\{||Mx|| | x \in \mathbb{R}^n, ||x|| = 1\}$ for $M \in \mathbb{R}^{n \times n}$.

**Topology.** The convex hull of a set $E \subseteq \mathbb{R}^n$, denoted as $\text{conv} E$, is the smallest convex set that contains $E$ (the intersection of convex sets is still convex). The affine hull $\text{aff} E$ and the conic hull $\text{cone} E$ are defined accordingly. Specifically,

\[
\text{conv} E := \left\{ \sum_{i=1}^{k} \alpha_i x_i \mid k \in \mathbb{N}, x_i \in E, \alpha_i \geq 0, \sum_{i=1}^{k} \alpha_i = 1 \right\},
\]

\[
\text{cone} E := \left\{ \sum_{i=1}^{k} \alpha_i x_i \mid k \in \mathbb{N}, x_i \in E, \alpha_i \geq 0 \right\},
\]

\[
\text{aff} E := \left\{ \sum_{i=1}^{k} \alpha_i x_i \mid k \in \mathbb{N}, x_i \in E, \alpha_i \in \mathbb{R}, \sum_{i=1}^{k} \alpha_i = 1 \right\}.
\]

The closure and interior of $E$ are denoted as $\text{cl} E$ and $\text{int} E$, respectively, whereas its relative interior, namely the interior of $E$ as a subspace of $\text{aff} E$, is denoted as $\text{relint} E$. With $B(x; r)$ and $\overline{B}(x; r)$ we indicate, respectively, the open and closed balls centered at $x$ with radius $r$.

**Sequences.** The notation $(d^k)_{k \in K}$ represents a sequence indexed by elements of the set $K$, and given a set $E$ we write $(d^k)_{k \in K} \subseteq E$ to indicate that $d^k \in E$ for all indices $k \in K$. We say that $(d^k)_{k \in K} \subseteq \mathbb{R}^n$ is summable if $\sum_{k \in K} ||d^k||$ is finite, and square-summable if $(||d^k||^2)_{k \in K}$ is summable. We say that the sequence converges to a point $a \in \mathbb{R}^n$ superlinearly if either $d^k = a$ for some $k \in \mathbb{N}$, or $||d^{k+1} - a||/||d^k - a|| \rightarrow 0$; if $||d^{k+1} - a||/||d^k - a||^q$ is bounded for some $q > 1$, then we say that the sequence converges superlinearly with order $q$, and in case $q = 2$ we say that the convergence is quadratic.

**Extended-real valued functions.** The extended-real line is $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. Given a function $h : \mathbb{R}^n \rightarrow [-\infty, \infty]$, its epigraph is the set

\[
\text{epi} h := \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} \mid h(x) \leq \alpha\},
\]

while its domain is

\[
\text{dom} h := \{x \in \mathbb{R}^n \mid h(x) < \infty\},
\]

and for $\alpha \in \overline{\mathbb{R}}$ its $\alpha$-level set is

\[
\text{lev}_{\leq \alpha} h := \{x \in \mathbb{R}^n \mid h(x) \leq \alpha\}.
\]

Function $h$ is said to be lower semicontinuous (lsc) if $\text{epi} h$ is a closed set in $\mathbb{R}^{n+1}$ (equivalently, $h$ is said to be closed); in particular, all level sets of an lsc function are closed. We say that $h$ is proper if $h > -\infty$ and $\text{dom} h \neq \emptyset$, and that it is level bounded if for all $\alpha \in \overline{\mathbb{R}}$ the level set $\text{lev}_{\leq \alpha} h$ is a bounded subset of $\mathbb{R}^n$.

**Continuity and smoothness.** A function $G : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is $\theta$-Hölder continuous for some $\theta > 0$ if there exists $L \geq 0$ such that

\[
||G(x) - G(x')|| \leq L||x - x'||^\theta
\]

for all $x, x'$. In case $\theta = 1$ we say that $G$ is $(L)$-Lipschitz continuous. $G$ is strictly differentiable at $\bar{x} \in \mathbb{R}^n$ if the Jacobian matrix $J G(\bar{x}) := \left( \frac{\partial G_i}{\partial x_j} (\bar{x}) \right)_{i,j}$ exists and
The class of functions $h : \mathbb{R}^n \to \mathbb{R}$ that are $k$ times continuously differentiable is denoted as $C^k(\mathbb{R}^n)$. We write $h \in C^{1,1}(\mathbb{R}^n)$ to indicate that $h \in C^1(\mathbb{R}^n)$ and that $\nabla h$ is Lipschitz continuous with modulus $L_h$. To simplify the terminology, we will say that such an $h$ is $L_h$-smooth. If $h$ is $L_h$-smooth and convex, then for any $u, v \in \mathbb{R}^n$

$$0 \leq h(v) - [h(u) + \langle \nabla h(u), v - u \rangle] \leq \frac{L_h}{2} \|v - u\|^2.$$  \quad (2.1)

Moreover, having $h \in C^{1,1}(\mathbb{R}^n)$ and $\mu_h$-strongly convex is equivalent to having

$$\mu_h \|v - u\|^2 \leq \langle \nabla h(v) - \nabla h(u), v - u \rangle \leq L_h \|v - u\|^2$$  \quad (2.2)

for all $u, v \in \mathbb{R}^n$.

**Set-valued mappings.** We use the notation $H : \mathbb{R}^n \Rightarrow \mathbb{R}^m$ to indicate a point-to-set function $H : \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^m)$, where $\mathcal{P}(\mathbb{R}^m)$ is the set of all subsets of $\mathbb{R}^m$. The graph of $H$ is the set

$$\text{gph } H := \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^m \mid y \in H(x)\},$$

while its domain is

$$\text{dom } H := \{x \in \mathbb{R}^n \mid H(x) \neq \emptyset\}.$$  

We say that $H$ is outer semicontinuous (osc) at $\bar{x} \in \text{dom } H$ if for any $\varepsilon > 0$ there exists $\delta > 0$ such that $H(x) \subseteq H(\bar{x}) + B(0; \varepsilon)$ for all $x \in B(\bar{x}, \delta)$. In particular, this implies that whenever $(x^k)_{k \in \mathbb{N}} \subseteq \text{dom } H$ converges to $x$ and $(y^k)_{k \in \mathbb{N}}$ converges to $y$ with $y^k \in H(x^k)$ for all $k$, it holds that $y \in H(x)$. We say that $H$ is osc (without mention of a point) if $H$ is osc at every point of its domain or, equivalently, if gph $H$ is a closed subset of $\mathbb{R}^n \times \mathbb{R}^m$ (notice that this notion does not reduce to lower semicontinuity for a single-valued function $H$).

**Convex analysis.** The [indicator function](https://en.wikipedia.org/wiki/Indicator_function) of a set $S \subseteq \mathbb{R}^n$ is denoted as $\delta_S : \mathbb{R}^n \to \mathbb{R}$, namely

$$\delta_S(x) = \begin{cases} 0 & \text{if } x \in S, \\ \infty & \text{otherwise}. \end{cases}$$

If $S$ is nonempty closed and convex, then $\delta_S$ is proper convex and lsc, and both the projection $P_S : \mathbb{R}^n \to \mathbb{R}^n$ and the distance $\text{dist}(\cdot, S) : \mathbb{R}^n \to [0, \infty)$ are well-defined functions, given by $P_S(x) = \text{arg min}_{z \in S} \|z - x\|$ and $\text{dist}(x, S) = \text{min}_{z \in S} \|z - x\|$, respectively.

The subdifferential of $h$ is the set-valued mapping $\partial h : \mathbb{R}^n \Rightarrow \mathbb{R}^n$ defined as

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

A vector $v \in \partial h(x)$ is called a subgradient of $h$ at $x$. It holds that $\text{dom } \partial h \subseteq \text{dom } h$, and if $h$ is proper and convex, then $\text{dom } \partial h$ is a nonempty convex set containing $\text{relint } \text{dom } h$, and $\partial h(x)$ is convex and closed for all $x \in \mathbb{R}^n$. 

\[\lim_{x, x' \to \bar{x} \atop x \neq x'} \frac{\|G(x') - JG(\bar{x})(x' - \bar{x}) - G(x)\|}{\|x' - \bar{x}\|} = 0.\]
A function \( h \) is said to be \( \mu \)-strongly convex for some \( \mu \geq 0 \) if \( h - \frac{\mu}{2}\|x\|^2 \) is convex. Unless differently specified, we allow for \( \mu = 0 \) which corresponds to \( h \) being convex but not strongly so. If \( \mu > 0 \), then \( h \) has a unique (global) minimizer.

### 2.2 Generalized differentiability

Due to its inherent nonsmooth nature, classical notions of differentiability may not be directly applicable in problem (1.1). This subsection contains some definitions and known facts on generalized differentiability that will be needed later on in the paper. The interested reader is referred to the textbooks [15, 19, 64] for the details.

**Definition 2.1** (Bouligand and Clarke subdifferentials). Let \( G : \mathbb{R}^n \to \mathbb{R}^m \) be locally Lipschitz continuous, and let \( C_G \subseteq \mathbb{R}^n \) be the set of points at which \( G \) is differentiable (in particular \( \mathbb{R}^n \setminus C_G \) has measure zero). The B-subdifferential (also known as Bouligand or limiting Jacobian) of \( G \) at \( \bar{x} \) is the set-valued mapping \( \partial_B G : \mathbb{R}^n \to \mathbb{R}^m \times \mathbb{R}^n \) defined as

\[
\partial_B G(\bar{x}) = \left\{ H \in \mathbb{R}^{m \times n} \mid \exists (x^k)_{k \in \mathbb{N}} \subseteq C_G \text{ with } x^k \to \bar{x}, JG(x^k) \to H \right\},
\]

whereas the (Clarke) generalized Jacobian of \( G \) at \( \bar{x} \) is \( \partial_C G : \mathbb{R}^n \to \mathbb{R}^{m \times n} \) given by

\[
\partial_C G(\bar{x}) = \text{conv}(\partial_B G(\bar{x})).
\]

If \( G : \mathbb{R}^n \to \mathbb{R}^m \) is locally Lipschitz on \( \mathbb{R}^n \), then \( \partial_C G(\bar{x}) \) is a nonempty, convex and compact subset of \( \mathbb{R}^{m \times n} \) matrices, and as a set-valued mapping it is osc at every \( \bar{x} \in \mathbb{R}^n \). Semismooth functions [60] are precisely Lipschitz-continuous mappings for which the generalized Jacobian (and consequently the B-subdifferential) furnishes a first-order approximation.

**Definition 2.2** (Semismooth mappings). Let \( G : \mathbb{R}^n \to \mathbb{R}^m \) be locally Lipschitz continuous at \( \bar{x} \). We say that \( G \) is semismooth at \( \bar{x} \) if

\[
\limsup_{H \in \partial_C G(\bar{x}), x \to \bar{x}} \frac{\|G(x) + H(\bar{x} - x) - G(\bar{x})\|}{\|x - \bar{x}\|} = 0. \quad (2.3a)
\]

We say that \( G \) is \( \vartheta \)-order semismooth for some \( \vartheta > 0 \) if the condition can be strengthened to

\[
\limsup_{H \in \partial_C G(\bar{x}), x \to \bar{x}} \frac{\|G(x) + H(\bar{x} - x) - G(\bar{x})\|}{\|x - \bar{x}\|^{1+\vartheta}} < \infty, \quad (2.3b)
\]

and in case \( \vartheta = 1 \) we say that \( G \) is strongly semismooth.

To simplify the notation, we adopt the small-\( o \) and big-\( O \) convention to write expressions as (2.3a) in the compact form \( G(x) + H(\bar{x} - x) - G(\bar{x}) = o(\|x - \bar{x}\|) \), and similarly (2.3b) as \( G(x) + H(\bar{x} - x) - G(\bar{x}) = O(\|x - \bar{x}\|^{1+\vartheta}) \). We remark that
the original definition of semismoothness given by [49] requires $G$ to be directionally differentiable at $x$. The definition given here is the one employed by [25]. It is also worth remarking that $\partial_c G(x)$ can be replaced with the smaller set $\partial_B G(x)$ in Definition 2.2. Fortunately, the class of semismooth mappings is rich enough to include many functions arising in interesting applications. For example piecewise smooth (PC$^1$) mappings are semismooth everywhere. Recall that a continuous mapping $G : \mathbb{R}^n \to \mathbb{R}^m$ is PC$^1$ if there exists a finite collection of smooth mappings $G_i : \mathbb{R}^n \to \mathbb{R}^m, i = 1, \ldots, N$, such that $G(x) \in [G_1(x), \ldots, G_N(x)] \quad \forall x \in \mathbb{R}^n$.

The definition of PC$^1$ mapping given here is less general than the one of, e.g., [66, §4] but it suffices for our purposes. For every $x \in \mathbb{R}^n$ we introduce the set of essentially active indices

$$I^c_G(x) := \{i \mid x \in \text{cl}(\text{int} \{w \mid G(w) = G_i(w)\})\}.$$ 

In other words, $I^c_G(x)$ contains only indices of the pieces $G_i$ for which there exists a full-dimensional set on which $G$ agrees with $G_i$. In accordance to Definition 2.1, the generalized Jacobian of $G$ at $x$ is the convex hull of the Jacobians of the essentially active pieces, i.e., [66, Prop. 4.3.1]

$$\partial_c G(x) = \text{conv} \{JG_i(x) \mid i \in I^c_G(x)\}.$$ 

(2.4)

The following definition is taken from [19, Def. 7.5.13].

**Definition 2.3** (Linear Newton approximation). Let $G : \mathbb{R}^n \to \mathbb{R}^m$ be continuous on $\mathbb{R}^n$. We say that $G$ admits a linear Newton approximation (LNA) at $\bar{x} \in \mathbb{R}^n$ if there exists a set-valued mapping $H : \mathbb{R}^n \rightrightarrows \mathbb{R}^m \times \mathbb{R}^n$ that has nonempty compact images, is outer semicontinuous at $\bar{x}$, and

$$\limsup_{x \to \bar{x}} \frac{\|G(x) + H(\bar{x} - x) - G(\bar{x})\|}{\|x - \bar{x}\|} = 0.$$ 

If for some $\vartheta > 0$ the condition can be strengthened to

$$\limsup_{x \to \bar{x}} \frac{\|G(x) + H(\bar{x} - x) - G(\bar{x})\|}{\|x - \bar{x}\|^{1+\vartheta}} < \infty,$$

then we say that $H$ is a $\vartheta$-order LNA, and if $\vartheta = 1$ we say that $H$ is a strong LNA.

Functions $G$ as in Definition 2.3 are also referred to as $H$-semismooth in the literature, see e.g., [78], however we prefer to stick to the terminology of [19] and rather say that $H$ is a LNA for $G$. Arguably the most notable example of a LNA for semismooth mappings is the generalized Jacobian, cf. Definition 2.1. However, semismooth mappings can admit LNAs different from the generalized Jacobian. More importantly, mappings that are not semismooth may also admit a LNA.
Lemma 2.4 ([19, Prop. 7.4.10]). Let $h \in C^1(\mathbb{R}^n)$ and suppose that $\mathcal{H} : \mathbb{R}^n \Rightarrow \mathbb{R}^{\partial h}$ is a LNA for $\nabla h$ at $\bar{x}$. Then,

$$\lim_{\bar{x} \in \mathcal{H}(x)} \frac{h(x) - h(\bar{x}) - (\nabla h(\bar{x}), x - \bar{x}) - \frac{1}{2} (H(x - \bar{x}), x - \bar{x})}{\|x - \bar{x}\|^2} = 0.$$ 

We remark that although [19, Prop. 7.4.10] assumes semismoothness of $\nabla h$ at $\bar{x}$ and uses $\partial_C(\nabla h)$ in place of $\mathcal{H}$; however, exactly the same arguments apply for any LNA of $\nabla h$ at $\bar{x}$ even without the semismoothness assumption.

3 Proximal algorithms

3.1 Proximal point and Moreau envelope

The proximal mapping of a proper closed and convex function $h : \mathbb{R}^n \rightarrow \mathbb{R}$ with parameter $\gamma > 0$ is $\text{prox}_{\gamma h} : \mathbb{R}^n \rightarrow \mathbb{R}^n$, given by

$$\text{prox}_{\gamma h}(x) := \arg \min_{w \in \mathbb{R}^n} \left\{ h(w) + \frac{1}{2\gamma} \|w - x\|^2 \right\}. \quad (3.1)$$

The majorization model $\mathcal{M}^\gamma(x; \cdot)$ is a proper and strongly convex function, and therefore has a unique minimizer. The value function, as opposed to the minimizer, defines the Moreau envelope $h^\gamma : \mathbb{R}^n \rightarrow \mathbb{R}$, namely

$$h^\gamma(x) := \min_{w \in \mathbb{R}^n} \left\{ h(w) + \frac{1}{2\gamma} \|w - x\|^2 \right\}. \quad (3.2)$$

which is real valued and Lipschitz differentiable, despite the fact that $h$ might be extended-real valued. Properties of the Moreau envelope and the proximal mapping are well documented in the literature, see e.g., [2, §24]. For example, $\text{prox}_{\gamma h}$ is nonexpansive (Lipschitz continuous with modulus 1) and is characterized by the implicit inclusion

$$\hat{x} = \text{prox}_{\gamma h}(x) \iff \frac{1}{\gamma} (x - \hat{x}) \in \partial h(\hat{x}). \quad (3.3)$$

For the sake of a brief recap, we now list some other important known properties. Theorem 3.1 provides some relations between $h$ and its Moreau envelope $h^\gamma$, which we informally refer to as sandwich property for apparent reasons, cf. Figure 1.1. Theorem 3.2 highlights that the minimization of a (proper, lsc and) convex function can be expressed as the convex smooth minimization of its Moreau envelope.

Theorem 3.1 (Moreau envelope: sandwich property [2, 12]). For all $\gamma > 0$ the following hold for the cost function $\varphi$:

(i) $\varphi^\gamma(x) \leq \varphi(x) - \frac{1}{2\gamma} \|x - \hat{x}\|^2$ for all $x \in \mathbb{R}^n$ where $\hat{x} := \text{prox}_{\gamma h}(x)$;

(ii) $\varphi(\hat{x}) = \varphi^\gamma(x) - \frac{1}{2\gamma} \|x - \hat{x}\|^2$ for all $x \in \mathbb{R}^n$ where $\hat{x} := \text{prox}_{\gamma h}(x)$;
(iii) \( \varphi^\gamma(x) = \varphi(x) \) iff \( x \in \arg\min \varphi \).

**Proof.**

\( \blacklozenge \) **3.1(i).** This fact is shown in [12, Lem. 3.2] for a more general notion of proximal point operator; namely, the square Euclidean norm appearing in (3.1) and (3.2) can be replaced by arbitrary Bregman divergences. In this simpler case, since \( \frac{1}{\gamma}(x - \hat{x}) \) is a subgradient of \( \varphi \) at \( \hat{x} \), cf. (3.3), we have

\[
\varphi(x) \geq \varphi(\hat{x}) + \frac{1}{\gamma} \langle x - \hat{x}, \nabla \varphi(x) \rangle = \varphi(\hat{x}) + \frac{1}{\gamma} \|x - \hat{x}\|^2.
\]

The claim now follows by subtracting \( \frac{1}{2} \|x - \hat{x}\|^2 \) from both sides.

\( \blacklozenge \) **3.1(ii).** Follows by definition, cf. (3.1) and (3.2).

\( \blacklozenge \) **3.1(iii).** See [2, Prop. 17.5]. \( \square \)

**Theorem 3.2** (Moreau envelope: convex smooth minimization equivalence [2]). For all \( \gamma > 0 \) the following hold for the cost function \( \varphi \):

(i) \( \varphi^\gamma \) is convex and smooth with \( L_{\varphi^\gamma} = \gamma^{-1} \) and \( \nabla \varphi^\gamma(x) = \gamma^{-1}(x - \text{prox} \varphi^\gamma(x)) \);

(ii) \( \inf \varphi = \inf \varphi^\gamma \);

(iii) \( x_\star \in \arg\min \varphi \) iff \( x_\star \in \arg\min \varphi^\gamma \) iff \( \nabla \varphi^\gamma(x_\star) = 0 \).

**Proof.**

\( \blacklozenge \) **3.2(i).** See [2, Prop. 12.15 and 12.30].

\( \blacklozenge \) **3.2(ii).** See [2, Prop. 12.9(iii)].

\( \blacklozenge \) **3.2(iii).** See [2, Prop. 17.5]. \( \square \)

**Fig. 1.1** Moreau envelope of the function \( \varphi(x) = \frac{1}{4} x^3 + x^2 - x + 1 + \delta_{[0,\infty)}(x) \) with parameter \( \gamma = 0.2 \).

At each point \( x \), the Moreau envelope \( \varphi^\gamma \) is the minimum of the quadratic majorization model \( M^\gamma_x = \varphi + \frac{1}{\gamma^2} (\cdot - x)^2 \), the unique minimizer being, by definition, the proximal point \( \hat{x} := \text{prox} \varphi^\gamma(x) \). It is a convex smooth lower bound to \( \varphi \), despite the fact that \( \varphi \) might be extended-real valued. Function \( \varphi \) and its Moreau envelope \( \varphi^\gamma \) have same \( \inf \) and \( \arg\min \); in fact, the two functions agree (only) on the set of minimizers. In general, \( \varphi^\gamma \) is sandwiched as \( \varphi \leq \text{prox} \varphi^\gamma \leq \varphi^\gamma \leq \varphi \).

As a consequence of Theorem 3.2, one can address the minimization of the convex but possibly nonsmooth and extended-real-valued function \( \varphi \) by means of gradient descent on the smooth envelope function \( \varphi^\gamma \) with stepsize \( 0 < \tau < \frac{2}{L_{\varphi^\gamma}} = 2\gamma \).
As first noticed by Rockafellar [63], this simply amounts to (relaxed) fixed-point iterations of the proximal point operator, namely

$$x^+ = (1 - \lambda)x + \lambda \text{prox}_{\gamma g}(x),$$  \hfill (3.4)

where \( \lambda = \frac{\gamma}{\tau} \in (0, 2) \) is a possible relaxation parameter. The scheme, known as proximal point algorithm (PPA) and first introduced by Martinet [45], is well covered by the broad theory of monotone operators, where convergence properties can be easily derived with simple tools of Fejérian monotonicity, see e.g., [2, Thm.s 23.41 and 27.1]. Nevertheless, not only does the interpretation as gradient method provide a beautiful theoretical link, but it also enables the employment of acceleration techniques exclusively stemming from smooth unconstrained optimization, such as Nesterov’s extrapolation [26] or quasi-Newton schemes [13], see also [7] for extensions to the dual formulation.

### 3.2 Forward-backward splitting

While it is true that every convex minimization problem can be smoothened by means of the Moreau envelope, unfortunately it is often the case that the computation of the proximal operator (which is needed to evaluate the envelope) is as hard as solving the original problem. For instance, evaluating the Moreau envelope of the cost of modeling a convex QP at one point amounts to solving another QP with same constraints and augmented cost. To overcome this limitation there comes the idea of splitting schemes, which decompose a complex problem in small components which are easier to operate onto. A popular such scheme is the forward-backward splitting (FBS), which addresses minimization problems of the form (1.1).

Given a point \( x \in \mathbb{R}^n \), one iteration of forward-backward splitting (FBS) for problem (1.1) with stepsize \( \gamma > 0 \) and relaxation \( \lambda > 0 \) consists in

$$x^+ = (1 - \lambda)x + \lambda T_\gamma(x),$$  \hfill (3.5)

where

$$T_\gamma(x) := \text{prox}_{\gamma g}(x - \gamma \nabla f(x))$$  \hfill (3.6)

is the forward-backward operator, characterized as

$$\bar{x} = T_\gamma(x) \iff \frac{1}{\gamma}(x - \bar{x}) - (\nabla f(x) - \nabla f(\bar{x})) \in \partial \phi(\bar{x}),$$  \hfill (3.7)

as it follows from (3.3). FBS interleaves a gradient descent step on \( f \) and a proximal point step on \( g \), and as such it is also known as proximal gradient method. If both \( f \) and \( g \) are (lsc, proper and) convex, then the solutions to (1.1) are exactly the fixed points of the forward-backward operator \( T_\gamma \). In other words,

$$x_* \in \text{arg min} \phi \iff R_\gamma(x_*) = 0,$$  \hfill (3.8)
where
\[ R_\gamma(x) := \frac{1}{\gamma}(x - \text{prox}_{\gamma g}(x - \gamma \nabla f(x))) \quad (3.9) \]
is the forward-backward residual.\(^1\) FBS iterations (3.5) are well known to converge to a solution to (1.1) provided that \( f \) is smooth and that the parameters are chosen as \( \gamma \in (0, 2/L_f) \) and \( \lambda \in (0, 2 - \gamma L_f / 2) \) \([2, Cor. 28.9]\) (\( \lambda = 1 \), which is always feasible, is the typical choice).

### 3.3 Error bounds and quadratic growth

We conclude the section with some inequalities that will be useful in the sequel.

**Lemma 3.3.** Suppose that \( X_\star \) is nonempty. Then,

\[ \varphi(x) - \varphi_\star \leq \text{dist}(0, \partial \varphi(x)) \text{dist}(x, X_\star) \quad \forall x \in \mathbb{R}^n. \]

**Proof.** From the subgradient inequality it follows that for all \( x_\star \in X_\star \) and \( v \in \partial \varphi(x) \) we have

\[ \varphi(x) - \varphi_\star = \varphi(x) - \varphi(x_\star) \leq \langle v, x - x_\star \rangle \leq ||v|| ||x - x_\star|| \]

and the claimed inequality follows from the arbitrariness of \( x_\star \) and \( v \). \( \square \)

**Lemma 3.4.** Suppose that \( X_\star \) is nonempty. For all \( x \in \mathbb{R}^n \) and \( \gamma > 0 \) the following holds

\[ ||R_\gamma(x)|| \geq \frac{1}{1 + \gamma L_f} \text{dist}(0, \partial \varphi(T_\gamma(x))) \]

**Proof.** Let \( \bar{x} := T_\gamma(x) \). The characterization (3.7) of \( T_\gamma \) implies that

\[ ||R_\gamma(x)|| \geq \text{dist}(0, \partial \varphi(\bar{x})) - ||\nabla f(x) - \nabla f(\bar{x})|| \geq \text{dist}(0, \partial \varphi(\bar{x})) - \gamma L_f ||R_\gamma(x)||. \]

After trivial rearrangements the sought inequality follows. \( \square \)

Further interesting inequalities can be derived if the cost function \( \varphi \) satisfies an error bound, which can be regarded as a generalization of strong convexity that does not require uniqueness of the minimizer. The interested reader is referred to \([43, 55, 3, 17]\) and references therein for extensive discussions.

**Definition 3.5** (Quadratic growth and error bound). Suppose that \( X_\star \neq \emptyset \). Given \( \mu, \nu > 0 \), we say that

(a) \( \varphi \) satisfies the quadratic growth with constants \( (\mu, \nu) \) if

\[ \varphi(x) - \varphi_\star \geq \frac{\mu}{2} \text{dist}(x, X_\star)^2 \quad \forall x \in \text{lev}_{\partial \varphi, \nu} \varphi; \quad (3.10) \]

\( \mu \) Due to apparent similarities with gradient descent iterations, having \( x^+ = x - \gamma R_\gamma(x) \) in FBS, \( R_\gamma \) is also referred to as (generalized) gradient mapping, see e.g., \([17]\). In particular, if \( g = 0 \) then \( R_\gamma = \nabla f \) whereas if \( f = 0 \) then \( R_\gamma = \nabla g^\prime \). The analogy will be supported by further evidence in the next section where we will see that, up to a change of metric, indeed \( R_\gamma \) is the gradient of the forward-backward envelope function.
(b) \( \varphi \) satisfies the error bound with constants \((\mu, \nu)\) if
\[
\text{dist}(0, \partial \varphi(x)) \geq \frac{\mu}{2} \text{dist}(x, \mathcal{X}_*) \quad \forall x \in \text{lev}_{\geq \nu} \varphi.
\]  
(3.11)

In case \( \nu = \infty \) we say that the properties are satisfied globally.

**Theorem 3.6** ([17, Thm. 3.3]). For a proper convex and lsc function, the quadratic growth with constants \((\mu, \nu)\) is equivalent to the error bound with same constants.

**Lemma 3.7** (Globality of quadratic growth). Suppose that \( \varphi \) satisfies the quadratic growth with constants \((\mu, \nu)\). Then, for every \( \nu' > \nu \) it satisfies the quadratic growth with constants \((\mu', \nu')\), where
\[
\mu' := \frac{\mu}{2} \min\{1, \frac{\nu}{\nu'-\nu}\}.
\]

**Proof.** Let \( \nu' > \nu \) be fixed, and let \( x \in \text{lev}_{\geq \nu} \varphi \) be arbitrary. Since \( \mu' \leq \mu \), the claim is trivial if \( \varphi(x) \leq \varphi_* + \nu \); we may thus suppose that \( \varphi(x) > \varphi_* + \nu \). Let \( z \) be the projection of \( x \) onto the (nonempty closed and convex) level set \( \text{lev}_{\geq \nu} \varphi \), and observe that \( \varphi(z) = \varphi_* + \nu \). With Lem. 3.3 and Thm. 3.6 we can upper bound \( \nu \) as
\[
\nu = \varphi(z) - \varphi_* \leq \text{dist}(0, \partial \varphi(z)) \text{dist}(z, \mathcal{X}_*) \leq \frac{2}{\mu} \text{dist}(0, \partial \varphi(z))^2.
\]  
(3.12)

Moreover, it follows from [28, Thm. 1.3.5] that there exists a subgradient \( v \in \partial \varphi(z) \) such that \((\nu, x - z) = \|v\|\|x - z\|\). Then,
\[
\varphi(x) \geq \varphi(z) + (\nu, x - z) = \varphi(z) + \|v\|\|x - z\| \geq \varphi(z) + \text{dist}(0, \partial \varphi(z))\|x - z\|
\]  
(3.12)
\[
\geq \varphi(z) + \sqrt{\frac{\mu}{2}}\|x - z\|.
\]  
(3.13)

By substracting \( \varphi(z) \) from the first and last terms we obtain
\[
\|x - z\| \leq \sqrt{\frac{2}{\mu}}(\varphi(x) - \varphi(z)) \leq \sqrt{\frac{2}{\mu}}(\nu' - \nu),
\]
which implies
\[
\|x - z\| \geq \sqrt{\frac{\mu}{2}} \frac{1}{\nu'-\nu} \|x - z\|^2.
\]  
(3.14)

Thus,
\[
\varphi(x) - \varphi_* \geq \varphi(z) - \varphi_* + \sqrt{\frac{\mu}{2}}\|x - z\|
\]
using the quadratic growth at \( z \) and the inequality (3.14)
\[
\geq \frac{\mu}{2} \text{dist}(z, \mathcal{X}_*)^2 + \frac{\mu}{2(\nu'-\nu)}\|x - z\|^2
\geq \frac{\mu}{2} \min\{1, \frac{\nu}{\nu'-\nu}\} \left[\text{dist}(z, \mathcal{X}_*)^2 + \|x - z\|^2\right].
\]

By using the fact that \( a^2 + b^2 \geq \frac{1}{2}(a + b)^2 \) for any \( a, b \in \mathbb{R} \) together with the triangular inequality \( \text{dist}(x, \mathcal{X}_*) \leq \|x - z\| + \text{dist}(z, \mathcal{X}_*) \), we conclude that \( \varphi(x) - \varphi_* \geq \frac{\mu}{2} \min\{1, \frac{\nu}{\nu'-\nu}\} \left[\text{dist}(z, \mathcal{X}_*)^2 + \|x - z\|^2\right]. \)
\[ \frac{\mu'}{2} \text{dist}(x, X_\star)^2, \] with \( \mu' \) as in the statement. Since \( \mu' \) depends only on \( \mu, \nu, \) and \( \nu' \), from the arbitrariness of \( x \in \text{lev}_{\leq \varphi} \) the claim follows. \( \square \)

**Theorem 3.8** ([17, Cor. 3.6]). Suppose that \( \varphi \) satisfies the quadratic growth with constants \((\mu, \nu)\). Then, for all \( \gamma \in (0, \frac{1}{L_f}) \) and \( x \in \text{lev}_{\leq \varphi} \) \( \varphi \) we have

\[
\text{dist}(x, X_\star) \leq (\gamma + \frac{2}{\mu})(1 + \gamma L_f)\|R_\gamma(x)\|.
\]

### 4 Forward-backward envelope

There are clearly infinite ways of representing the (proper, lsc and) convex function \( \varphi \) in (1.1) as the sum of two convex functions \( f \) and \( g \) with \( f \) smooth, and each of these choices leads to a different FBS operator \( T_\gamma \). If \( f = 0 \), for instance, then \( T_\gamma \) reduces to \( \text{prox}_{\gamma \varphi} \), and consequently FBS (3.5) to the PPA (3.4). A natural question then arises, whether a function exists that serves as “envelope” for FBS in the same way that \( \varphi_\gamma \) does for \( \text{prox}_{\gamma \varphi} \). We will now provide a positive answer to this question by reformulating the nonsmooth problem (1.1) as the minimization of a differentiable function. To this end, the following requirements on \( f \) and \( g \) will be assumed throughout the paper without further mention.

**Assumption I** (Basic requirements). In problem (1.1),

(i) \( f : \mathbb{R}^n \to \mathbb{R} \) is convex, twice continuously differentiable and \( L_f \)-smooth;

(ii) \( g : \mathbb{R}^n \to \overline{\mathbb{R}} \) is lsc, proper and convex.

Compared to the classical FBS assumptions, the only additional requirement is twice differentiability of \( f \). This ensures that the forward operator \( x \mapsto x - \gamma \nabla f(x) \) is differentiable; we denote its Jacobian as \( Q_\gamma : \mathbb{R}^n \to \mathbb{R}^{n \times n} \), namely

\[
Q_\gamma(x) := I - \gamma \nabla^2 f(x). \tag{4.1}
\]

Notice that, due to the bound \( \nabla^2 f(x) \preceq L_f I \) (which follows from \( L_f \)-smoothness of \( f \), see [53, Lem. 1.2.2]) \( Q_\gamma(x) \) is invertible (in fact, positive definite) whenever \( \gamma < \frac{1}{L_f} \). Moreover, due to the chain rule and Theorem 3.2(ii) we have that

\[
\nabla[(\gamma^{-1}Q_\gamma(x)[x - \gamma \nabla f(x) - \text{prox}_{\gamma g}(x - \gamma \nabla f(x))])
= Q_\gamma(x)[R_\gamma(x) - \nabla f(x)].
\]

Rearranging,
\[ Q_\gamma(x)R_\gamma(x) = \nabla f(x) - \gamma \nabla^2 f(x) \nabla f(x) + \nabla [g^\gamma \circ (\text{id} - \gamma \nabla f)](x) \]
\[ = \nabla f(x) - \nabla \left[ \frac{\gamma}{2} \|\nabla f\|^2 \right](x) + \nabla [g^\gamma \circ (\text{id} - \gamma \nabla f)](x) \]
\[ = \nabla \left[ f - \frac{\gamma}{2} \|\nabla f\|^2 + g^\gamma \circ (\text{id} - \gamma \nabla f) \right](x) \]

we obtain the gradient of a real-valued function, which we define as follows.

**Definition 4.1 (Forward-backward envelope).** The forward-backward envelope (FBE) for the composite minimization problem (1.1) is the function \( \varphi_\gamma : \mathbb{R}^n \to \mathbb{R} \) defined as
\[
\varphi_\gamma(x) := f(x) - \frac{\gamma}{2} \|\nabla f(x)\|^2 + g^\gamma(x - \gamma \nabla f(x)).
\] (4.2)

In the next section we discuss some of the favorable properties enjoyed by the FBE.

### 4.1 Basic properties

We already verified that the FBE is differentiable with gradient
\[
\nabla \varphi_\gamma(x) = Q_\gamma(x)R_\gamma(x). \tag{4.3}
\]

In particular, for \( \gamma < 1/L_f \) one obtains that a FBS step is a (scaled) gradient descent step on the FBE, similarly as the relation between Moreau envelope and PPA; namely,
\[
T_\gamma(x) = x - \gamma Q_\gamma(x)^{-1} \nabla \varphi_\gamma(x). \tag{4.4}
\]

To take the analysis of the FBE one step further, let us consider the equivalent expression of the operator \( T_\gamma \) as
\[
T_\gamma(w) = \arg \min_{w \in \mathbb{R}^p} \left( f(x) + \langle \nabla f(x), w - x \rangle + \frac{1}{2\gamma} \|w - x\|^2 + g(w) \right). \tag{4.5}
\]

Differently from the quadratic model \( M_\gamma^f \) in (3.1), \( M_\gamma^{fg} \) replaces the differentiable component \( f \) with a linear approximation. Building upon the idea of the Moreau envelope, instead of the minimizer \( T_\gamma(x) \) we consider the value attained in the subproblem (4.5), and with simple algebra one can easily verify that this gives rise once again to the FBE:
\[
\varphi_\gamma(x) = \min_{w \in \mathbb{R}^p} \left( f(x) + \langle \nabla f(x), w - x \rangle + \frac{1}{2\gamma} \|w - x\|^2 + g(w) \right). \tag{4.6}
\]

Starting from this expression we can easily mirror the properties of the Moreau envelope stated in Theorems 3.1 and 3.2. These results appeared in the independent works [54] and [57], although the former makes no mention of an “envelope” function and simply analyzes the majorization-minimization model \( M_\gamma^{fg} \).

**Theorem 4.2 (FBE: sandwich property).** Let \( \gamma > 0 \) and \( x \in \mathbb{R}^n \) be fixed, and denote \( \tilde{x} = T_\gamma(x) \). The following hold:
\[(i) \quad \varphi(x) \leq \varphi(x) - \frac{1}{2\gamma} \|x - \bar{x}\|^2; \]

\[(ii) \quad \varphi(x) - \frac{1}{2\gamma} \|x - \bar{x}\|^2 \leq \varphi(\bar{x}) \leq \varphi(x) - \frac{1-\gammaL}{2\gamma} \|x - \bar{x}\|^2. \]

In particular,

\[(iii) \quad \varphi(x) = \varphi(x) \quad \text{iff} \quad x \in \text{arg min} \varphi. \]

In fact, the assumption of twice continuous differentiability of \(f\) can be dropped.

Proof.

\[\blacktriangleleft 4.2(i) \quad \text{Since the minimum in (4.6) is attained at} \ w = \bar{x}, \text{cf. (4.5), we have} \]

\[\varphi(x) = f(x) + \langle \nabla f(x), \bar{x} - x \rangle + \frac{1}{2\gamma} \|\bar{x} - x\|^2 + g(\bar{x}) \]

\[\leq f(x) + \langle \nabla f(x), \bar{x} - x \rangle + \frac{1}{2\gamma} \|\bar{x} - x\|^2 + g(x) + \langle \frac{1}{\gamma}(x - \bar{x}) - \nabla f(x), \bar{x} - x \rangle \]

\[= f(x) + g(x) - \frac{1}{2\gamma} \|x - \bar{x}\|^2 \]

where in the inequality we used the fact that \(\frac{1}{\gamma}(x - \bar{x}) - \nabla f(x) \in \partial g(\bar{x}), \text{cf. (3.7)}. \]

\[\blacktriangleleft 4.2(ii) \quad \text{Follows by using (2.1) (with} \ h = f, \text{and} \ u = x \text{and} \ v = \bar{x} \text{in (4.7)}. \]

\[\blacktriangleleft 4.2(iii) \quad \text{Follows by 4.2(i) and the optimality condition (3.8).} \]

Notice that by combining Theorems 4.2(i) and 4.2(ii) we recover the “sufficient decrease” condition of (convex) FBS [54, Thm. 1], that is

\[\varphi(\bar{x}) \leq \varphi(x) - \frac{2-\gammaL}{2\gamma} \|x - \bar{x}\|^2 \quad (4.8)\]

holding for all \(x \in \mathbb{R}^n\) with \(\bar{x} = T_\gamma(x)\).

**Theorem 4.3 (FBE: smooth minimization equivalence).** For all \(\gamma > 0\)

\[(i) \quad \varphi \in C^1(\mathbb{R}^n) \text{ with} \ \nabla \varphi = Q\gamma R\gamma. \]

Moreover, if \(\gamma \in (0, 1/L)\) then the following also hold:

\[(ii) \quad \inf \varphi = \inf \varphi; \]

\[(iii) \quad x_* \in \text{arg min} \varphi \quad \text{iff} \quad x_* \in \text{arg min} \varphi \quad \text{iff} \quad \nabla \varphi(x_*) = 0. \]

Proof.

\[\blacktriangleleft 4.3(i) \quad \text{Since} \ f \in C^2(\mathbb{R}^n) \text{ and} \ g^\gamma \in C^1(\mathbb{R}^n) \text{ (cf. Thm. 3.2(i)), from the definition (4.2) it is apparent that} \ \varphi \text{ is continuously differentiable for all} \ \gamma > 0. \text{ The formula for the gradient was already shown in (4.3).} \]

Suppose now that \(\gamma < 1/L\).

\[\blacktriangleleft 4.3(ii) \quad \inf \varphi \leq \inf_{x \in \mathbb{R}^n} \varphi(T_\gamma(x)) \leq \inf_{x \in \mathbb{R}^n} \varphi(x) = \inf \varphi \leq \inf \varphi. \]
Suppose now that \( x_\star \in \mathbf{arg \, min}_\gamma \varphi \). Since \( \varphi_\gamma \in C^1(\mathbb{R}^n) \), the first-order necessary condition reads \( \nabla \varphi_\gamma = 0 \), and from the equivalence proven above we conclude that \( \mathbf{arg \, min}_\gamma \varphi \subseteq \mathbf{arg \, min} \varphi \). Conversely, if \( x_\star \in \mathbf{arg \, min} \varphi \) then

\[
\varphi_\gamma(x_\star) = \varphi(x_\star) = \inf \varphi = \inf \varphi_\gamma,
\]

proving \( x_\star \in \mathbf{arg \, min}_\gamma \varphi_\gamma \), hence the inclusion \( \mathbf{arg \, min}_\gamma \varphi_\gamma \supseteq \mathbf{arg \, min} \varphi \). \( \square \)

**Fig. 1.2** FBE of the function \( \varphi \) as in Fig. 1.1 with same parameter \( \gamma = 0.2 \), relative to the decomposition as the sum of \( f(x) = x^2 + x - 1 \) and \( g(x) = \frac{7}{2} x^2 + \delta_{[0, \infty)}(x) \). For \( \gamma < \frac{1}{L_f} \), \( L_f = 2 \) in this example) at each point \( x \) the FBE \( \varphi_\gamma \) is the minimum of the quadratic majorization model \( M^{\gamma}_{f,g}(\cdot, x) \) for \( \varphi \), the unique minimizer being the proximal gradient point \( \bar{x} = T_{\varphi}(x) \). The FBE is a differentiable lower bound to \( \varphi \) and since \( f \) is quadratic in this example, it is also smooth and convex (cf. Thm. 4.6). In any case, its stationary points and minimizers coincide, and are equivalent to the minimizers of \( \varphi \).

**Proposition 4.4** (FBE and Moreau envelope [54, Thm. 2]). For any \( \gamma \in (0, \frac{1}{L_f}) \), it holds that

\[
\varphi = f + g
\]

\[
\varphi_\gamma \leq \varphi \gamma \leq \varphi.
\]

**Proof.** We have

\[
\varphi_\gamma(x) = \min_{w \in \mathbb{R}^n} \left[ f(x) + \langle \nabla f(x), w - x \rangle + \frac{1}{2\gamma} \|w - x\|^2 + g(w) \right]
\]

\[
\leq \min_{w \in \mathbb{R}^n} \left[ f(w) - \frac{\gamma L_f}{2} \|w - x\|^2 + \frac{1}{2\gamma} \|w - x\|^2 + g(w) \right]
\]

\[
= \min_{w \in \mathbb{R}^n} \left[ f(w) + g(w) + \frac{1 - \gamma L_f}{2\gamma} \|w - x\|^2 \right] = \varphi_\gamma(x).
\]

Using the upper bound in (2.1) instead, yields the other inequality. \( \square \)

Since \( \varphi_\gamma \) is upper bounded by the \( \gamma^{-1} \)-smooth function \( \varphi_\gamma \) with which it shares the set of minimizers \( X_\star \), from (2.1) we easily infer the following quadratic upper bound.
Corollary 4.5 (Global quadratic upper bound). If $X_\ast \neq \emptyset$, then

$$\varphi_\gamma(x) - \varphi_\ast \leq \frac{1}{\gamma^2} \text{dist}(x, X_\ast)^2 \quad \forall x \in \mathbb{R}^n.$$ 

Although the FBE may fail to be convex, for $\gamma < 1/L_f$ its stationary points and minimizers coincide and are the same as those of the original function $\varphi$. That is, the minimization of $\varphi$ is equivalent to the minimization of the differentiable function $\varphi_\gamma$.

This is a clear analogy with the Moreau envelope, which in fact is the special case of the FBE corresponding to $f \equiv 0$ in the decomposition of $\varphi$. In the next result we tighten the claims of Theorem 4.3(i) when $f$ is a convex quadratic function, showing that in this case the FBE is convex and smooth and thus recover all the properties of the Moreau envelope.

Theorem 4.6 (FBE: convexity & smoothness for quadratic $f$ [24, Prop. 4.4]). Suppose that $f$ is convex quadratic, namely $f(x) = \frac{1}{2}(x, Hx) + (h, x)$ for some $H \in S_n(\mathbb{R}^n)$ and $h \in \mathbb{R}^n$. Then, for all $\gamma \in (0, 1/L_f]$ the FBE $\varphi_\gamma$ is convex and smooth, with

$$L_{\varphi_\gamma} = \frac{1 - \gamma \mu_f}{\gamma} \quad \text{and} \quad \mu_{\varphi_\gamma} = \min \left\{ \mu_f(1 - \gamma \mu_f), L_f(1 - \gamma L_f) \right\},$$

where $L_f = \lambda_{\max}(H)$ and $\mu_f = \lambda_{\min}(H)$. In particular, when $f$ is $L_f$-strongly convex the strong convexity of $\varphi_\gamma$ is maximized for $\gamma = \frac{1}{\mu_f + L_f}$, in which case

$$L_{\varphi_\gamma} = L_f \quad \text{and} \quad \mu_{\varphi_\gamma} = \frac{L_f \mu_f}{\mu_f + L_f}.$$

Proof. Letting $Q := I - \gamma H$, we have that $Q_\gamma \equiv Q$ and $x - \gamma \nabla f(x) = Qx - \gamma h$. Therefore,

$$\gamma(\nabla \varphi_\gamma(x) - \nabla \varphi_\gamma(y), x - y) \stackrel{(4.3)}{=} \langle Q(R_\gamma(x) - R_\gamma(y)), x - y \rangle \quad \stackrel{(4.3)}{=} \langle Q(x - y), x - y \rangle - \langle Q(T_\gamma(x) - T_\gamma(y)), x - y \rangle \quad \stackrel{(4.3)}{=} ||x - y||_Q^2 - \langle \text{prox}_{\gamma g}(Qx - \gamma h) - \text{prox}_{\gamma g}(Qy - \gamma h), Q(x - y) \rangle.$$

From the firm nonexpansiveness of $\text{prox}_{\gamma g}$ (see [2, Props 4.35(iii) and 12.28]) it follows that

$$0 \leq \langle \text{prox}_{\gamma g}(Qx - \gamma h) - \text{prox}_{\gamma g}(Qy - \gamma h), Q(x - y) \rangle \leq ||Q(x - y)||_Q^2.$$

By combining with the previous inequality, we obtain

$$\frac{1}{\gamma} ||x - y||_Q^2 \leq \langle \nabla \varphi_\gamma(x) - \nabla \varphi_\gamma(y), x - y \rangle \leq \frac{1}{\gamma} ||x - y||_Q^2.$$

Since $\lambda_{\min}(Q) = 1 - \gamma L_f$ and $\lambda_{\max}(Q) = 1 - \gamma \mu_f$, from Lem. A.2 we conclude that

$$\mu_{\varphi_\gamma} ||x - y||_Q^2 \leq \langle \nabla \varphi_\gamma(x) - \nabla \varphi_\gamma(y), x - y \rangle \leq L_{\varphi_\gamma} ||x - y||_Q^2$$

with $\mu_{\varphi_\gamma}$ and $L_{\varphi_\gamma}$ as in the statement, hence the claim, cf. (2.2).
Proposition 4.9. Suppose that $\varphi$ has the quadratic growth with constants ($\mu, \nu$), and let $\varphi_* := \min \varphi$. Then, for all $\gamma \in (0, 1/L_{\gamma})$ and $x \in \text{lev}_{\varphi_* + \gamma \varphi}$ it holds that

$$\varphi_{\gamma}(x) - \varphi_* \leq \gamma \left[ \frac{1}{2} + \frac{\gamma}{2 \mu} \right] \left( 1 + \frac{\gamma L_{\gamma}}{2} \right)^2 \|R_{\gamma}(x)\|^2.$$  

Proof. Fix $x \in \text{lev}_{\varphi_* + \gamma \varphi}$ and let $\bar{x} := T_{\gamma}(x)$. We have

$$\varphi_{\gamma}(x) - \varphi_* \leq \frac{\gamma}{2} \|R_{\gamma}(x)\|^2 + \varphi(\bar{x}) - \varphi_*$$

$$\leq \frac{\gamma}{2} \|R_{\gamma}(x)\|^2 + \text{dist}(\bar{x}, X_*) \text{dist}(0, \partial \varphi(\bar{x}))$$

$$\leq \left[ \frac{\gamma}{2} \|R_{\gamma}(x)\| + (1 + \gamma L_{\gamma}) \text{dist}(\bar{x}, X_*) \right] \|R_{\gamma}(x)\|$$

and since $\bar{x} \in \text{lev}_{\varphi_* + \gamma \varphi}$ (cf. Thm. 4.2(ii)), from Thm. 3.8 we can bound the quantity $\text{dist}(\bar{x}, X_*)$ in terms of the residual as

$$\leq \left[ \frac{\gamma}{2} \|R_{\gamma}(x)\| + (1 + \gamma L_{\gamma}) \|R_{\gamma}(\bar{x})\| \right] \|R_{\gamma}(x)\|.$$

The proof now follows from the inequality $\|R_{\gamma}(\bar{x})\| \leq \|R_{\gamma}(x)\|$, see [4, Thm. 10.12], after easy algebraic manipulations. $\square$

4.2 Further equivalence properties

Proposition 4.8 (Equivalence of level boundedness). For any $\gamma \in (0, 1/L_{\gamma})$, $\varphi$ has bounded level sets iff $\varphi_{\gamma}$ does.

Proof. Thm. 4.2 implies that $\text{lev}_{\varphi} \subseteq \text{lev}_{\varphi_{\gamma}}$ for all $\alpha \in \mathbb{R}$, therefore level boundedness of $\varphi_{\gamma}$ implies that of $\varphi$. Conversely, suppose that $\varphi_{\gamma}$ is not level bounded, and consider $(x_k)_{k \in \mathbb{N}} \subseteq \text{lev}_{\varphi_{\gamma}}$ with $\|x_k\| \to \infty$. Then from Thm. 4.2 it follows that $\varphi(x_k) \leq \varphi(x_k) - \frac{\gamma}{2} \|x_k - \bar{x}_k\|^2 \leq \alpha - \frac{\gamma}{2} \|x_k - \bar{x}_k\|^2$, where $\bar{x}_k = T_{\gamma}(x_k)$. In particular, $(\bar{x}_k)_{k \in \mathbb{N}} \subseteq \text{lev}_{\varphi_{\gamma}}$. If $(\bar{x}_k)_{k \in \mathbb{N}}$ is bounded, then $\inf \varphi = -\infty$; otherwise, $\text{lev}_{\varphi_{\gamma}}$ contains the unbounded sequence $(\bar{x}_k)_{k \in \mathbb{N}}$. Either way, $\varphi$ cannot be level bounded. $\square$

Proposition 4.9 (Equivalence of quadratic growth). Let $\gamma \in (0, 1/L_{\gamma})$ be fixed. Then,

(i) if $\varphi$ satisfies the quadratic growth condition with constants ($\mu, \nu$), then so does $\varphi_{\gamma}$ with constants ($\mu', \nu$), where $\mu' := \frac{1 - \gamma L_{\gamma}}{1 + \gamma L_{\gamma}} \frac{\mu}{(1 + \gamma L_{\gamma})^2} \mu$;

(ii) conversely, if $\varphi_{\gamma}$ satisfies the quadratic growth condition, then so does $\varphi$ with same constants.

Proof. Since $\varphi$ and $\varphi_{\gamma}$ have same infimum and minimizers (cf. Thm. 4.3), 4.9(ii) is a straightforward consequence of the fact that $\varphi_{\gamma} \leq \varphi$ (cf. Thm. 4.2(ii)).

Conversely, suppose that $\varphi$ satisfies the quadratic growth with constants ($\mu, \nu$). Then, for all $x \in \text{lev}_{\varphi_* + \gamma \varphi}$ we have that $\bar{x} := T_{\gamma}(x) \in \text{lev}_{\varphi_* + \gamma \varphi}$, therefore

$$\varphi_{\gamma}(x) - \varphi_* \geq \varphi(\bar{x}) - \varphi_* + \gamma \frac{1 - \gamma L_{\gamma}}{2} \|R_{\gamma}(x)\|^2 \geq \frac{\mu'}{2} \text{dist}(x, X_*),$$

where $\mu' := \frac{1 - \gamma L_{\gamma}}{1 + \gamma L_{\gamma}} \frac{\mu}{(1 + \gamma L_{\gamma})^2} \mu$. $\square$
where in the last inequality we discarded the term \( \varphi(\bar{x}) - \varphi_* \geq 0 \) and used Thm. 3.8 to lower bound \( \| R_\varphi(x) \| \).

\[ \square \]

**Corollary 4.10** (Equivalence of strong minimality). For all \( \gamma \in (0, 1/L) \), a point \( x_* \) is a (locally) strong minimizer for \( \varphi \) iff it is a (locally) strong minimizer for \( \varphi_\gamma \).

Lastly, having showed that for convex functions the quadratic growth can be extended to arbitrary level sets (cf. Lem. 3.7), an interesting consequence of Proposition 4.9 is that, although \( \varphi_\gamma \) may fail to be convex, it enjoys the same property.

**Corollary 4.11** (FBE: globality of quadratic growth). Let \( \gamma \in (0, 1/L) \) and suppose that \( \varphi_\gamma \) satisfies the quadratic growth with constants \( (\mu, \nu) \). Then, for every \( v' > v \) there exists \( \mu' > 0 \) such that \( \varphi_\gamma \) satisfies the quadratic growth with constants \( (\mu', \nu') \).

### 4.3 Second-order properties

Although \( \varphi_\gamma \) is continuously differentiable over \( \mathbb{R}^n \), it fails to be \( C^2 \) in most cases; since \( g \) is nonsmooth, its Moreau envelope \( g^\gamma \) is hardly ever \( C^2 \). For example, if \( g \) is real valued then \( g^\gamma \) is \( C^2 \) (and \( \text{prox}_{g^\gamma} \) is \( C^1 \)) if and only if \( g \) is \( C^2 \) \([33]\). Therefore, we hardly ever have the luxury of assuming continuous differentiability of \( \nabla \varphi_\gamma \) and we must resort to generalized notions of differentiability stemming from nonsmooth analysis. Specifically, our analysis is largely based on generalized differentiability properties of \( \text{prox}_{g^\gamma} \) which we study next.

**Theorem 4.12.** For all \( x \in \mathbb{R}^n \), \( \partial_C(\text{prox}_{g^\gamma})(x) \neq \emptyset \) and any \( P \in \partial_C(\text{prox}_{g^\gamma})(x) \) is a symmetric positive semidefinite matrix that satisfies \( ||P|| \leq 1 \).

**Proof.** Nonempty-valuedness of \( \partial_C(\text{prox}_{g^\gamma}) \) is due to Lipschitz continuity of \( \text{prox}_{g^\gamma} \). Moreover, since \( g \) is convex, its Moreau envelope is a convex function as well, therefore every element of \( \partial_C(\nabla g^\gamma)(x) \) is a symmetric positive semidefinite matrix (see \( e.g., [19, \S 8.3.3] \)). Due to Thm. 3.2(i), we have that \( \text{prox}_{g^\gamma}(x) = x - \gamma \nabla g^\gamma(x) \), therefore

\[ \partial_C(\text{prox}_{g^\gamma})(x) = I - \gamma \partial_C(\nabla g^\gamma)(x). \quad (4.9) \]

The last relation holds with equality (as opposed to inclusion in the general case) due to the fact that one of the summands is continuously differentiable. Now, from (4.9) we easily infer that every element of \( \partial_C(\text{prox}_{g^\gamma})(x) \) is a symmetric matrix. Since \( \nabla g^\gamma(x) \) is Lipschitz continuous with Lipschitz constant \( \gamma^{-1} \), using \([15, \text{Prop. 2.6.2(d)}]\), we infer that every \( H \in \partial_C(\nabla g^\gamma)(x) \) satisfies \( ||H|| \leq \gamma^{-1} \). Now, according to (4.9) it holds that

\[ P \in \partial_C(\text{prox}_{g^\gamma})(x) \iff P = I - \gamma H, \quad H \in \partial_C(\nabla g^\gamma)(x). \]

Therefore, for every \( d \in \mathbb{R}^n \) and \( P \in \partial_C(\text{prox}_{g^\gamma})(x) \),

\[ \langle d, Pd \rangle = ||d||^2 - \gamma \langle d, Hd \rangle \geq ||d||^2 - \gamma \gamma^{-1} ||d||^2 = 0. \]
On the other hand, since \( \operatorname{prox}_{\gamma g} \) is Lipschitz continuous with Lipschitz constant 1, using \([15, \text{Prop. 2.6.2(d)}]\) we obtain that \( \|P\| \leq 1 \) for all \( P \in \partial_C(\operatorname{prox}_{\gamma g})(x) \).

We are now in a position to construct a generalized Hessian for \( \varphi \), that will allow the development of Newton-like methods with fast asymptotic convergence rates. An obvious route to follow would be to assume that \( \nabla \varphi \) is semismooth and employ \( \partial_C(\nabla \varphi) \) as a generalized Hessian for \( \varphi \). However, this approach would require extra assumptions on \( f \) and involve complicated operations to evaluate elements of \( \partial_C(\nabla \varphi) \). On the other hand, what is really needed to devise Newton-like algorithms with fast local convergence rates is a linear Newton approximation (LNA), cf. \textbf{Definition 2.3}, at some stationary point of \( \varphi \), which by \textbf{Theorem 4.3(iii)} is also a minimizer of \( \varphi \), provided that \( \gamma \in (0,1/\ell_x) \).

The approach we follow is largely based on \([72], [19, \text{Prop. 10.4.4}]\). Without any additional assumptions we can define a set-valued mapping \( \hat{\partial}^2 \varphi : \mathbb{R}^n \to \mathbb{R}^{n \times n} \) with full domain and whose elements have a simpler form than those of \( \partial C(\nabla \varphi) \), which serves as a LNA for \( \nabla \varphi \), at any stationary point \( x^* \) provided \( \operatorname{prox}_{\gamma g} \) is semismooth at \( x^* - \gamma \nabla f(x^*) \). We call it approximate generalized Hessian of \( \varphi \) and it is given by

\[
\hat{\partial}^2 \varphi(x) := \{ \gamma^{-1}Q_y(x)(I - PQ_y(x)) \mid P \in \partial_C(\operatorname{prox}_{\gamma g})(x - \gamma \nabla f(x)) \}. \tag{4.10}
\]

Notice that if \( f \) is quadratic, then \( \hat{\partial}^2 \varphi \equiv \partial C \nabla \varphi \); more generally, the key idea in the definition of \( \hat{\partial}^2 \varphi \), reminiscent of the Gauss-Newton method for nonlinear least-squares problems, is to omit terms vanishing at \( x^* \) that contain third-order derivatives of \( f \).

\textbf{Proposition 4.13.} Let \( \bar{x} \in \mathbb{R}^n \) and \( \gamma > 0 \) be fixed. If \( \operatorname{prox}_{\gamma g} \) is \((\theta\text{-order})\) semismooth at \( \bar{x} - \gamma \nabla f(\bar{x}) \) and \( \nabla^2 f \) is \( \theta \)-Hölder continuous around \( \bar{x} \), then

\[
\mathcal{R}_\gamma(x) := \{ \gamma^{-1}(I - PQ_y(x)) \mid P \in \partial_C \operatorname{prox}_{\gamma g}(x - \gamma \nabla f(x)) \} \tag{4.11}
\]

is a \((\theta\text{-order})\) LNA for \( \mathcal{R}_\gamma \) at \( \bar{x} \).

\textbf{Proof.} We shall prove only the \( \theta \)-order semismooth case, as the other one is shown by simply replacing all occurrences of \( O(\| \cdot \|^{1+\theta}) \) with \( o(|| \cdot ||) \) in the proof. Let \( q_y = \text{id} - \gamma \nabla f \) be the forward operator, so that the forward-backward operator \( T_\gamma \) can be expressed as \( T_\gamma = \operatorname{prox}_{\gamma g} \circ q_y \). With a straightforward adaptation of the proof of \([19, \text{Prop. 7.2.9}]\) to include the \( \theta \)-Hölderian case, it can be shown that

\[
q_y(x) - q_y(\bar{x}) - Q_y(x)(\bar{x} - x) = O(||x - \bar{x}||^{1+\theta}). \tag{4.12}
\]

Moreover, since \( \nabla f \) is Lipschitz continuous and thus so is \( q_y \), we also have

\[
q_y(x) - q_y(\bar{x}) = O(||x - \bar{x}||). \tag{4.13}
\]

Let \( U_x \in \mathcal{R}_\gamma(x) \) be arbitrary; then, there exists \( P_x \in \partial_C \operatorname{prox}_{\gamma g}(x - \gamma \nabla f(x)) \) such that \( U_x = \gamma^{-1}(I - P_x Q_y(x))(\bar{x} - x) \). We have
\[ R_{\gamma}(x) + U_{\gamma}(\bar{x} - x) - R_{\gamma}(\bar{x}) = R_{\gamma}(x) + \gamma^{-1}(1 - P_{\gamma}Q_{\gamma}(x))(\bar{x} - x) - R_{\gamma}(\bar{x}) = \gamma^{-1}[(\text{prox}_{\gamma q_{\gamma}} - \text{prox}_{\gamma q_{\gamma}} - P_{\gamma}Q_{\gamma}(x)(\bar{x} - x)] \]
due to $\theta$-order semismoothness of $\text{prox}_{\gamma q_{\gamma}}$ at $q_{\gamma}(\bar{x})$,
\[ = \gamma^{-1}P_{\gamma}[q_{\gamma}(\bar{x}) - q_{\gamma}(x) + O(\|q_{\gamma}(\bar{x}) - q_{\gamma}(x)\|^{1+\theta}) - Q_{\gamma}(x)(\bar{x} - x)] \]
\[ = \gamma^{-1}P_{\gamma}[q_{\gamma}(\bar{x}) - q_{\gamma}(x) - Q_{\gamma}(\bar{x})(\bar{x} - x) + O(\|ar{x} - x\|^{1+\theta})] \]
\[ = \gamma^{-1}P_{\gamma}O(\|ar{x} - x\|^{1+\theta}) = O(\|ar{x} - x\|^{1+\theta}), \]
where in the last equality we used the fact that $\|P_{\gamma}\| \leq 1$, cf. Thm. 4.12. \qed

**Corollary 4.14.** Let $\gamma \in (0, 1/L_f)$ and $x_{\star} \in X_{\ast}$. If $\text{prox}_{\gamma q_{\gamma}}$ is $(\theta$-order) semismooth at $x_{\star} - \gamma \nabla f(x_{\star})$ (and $\nabla^2 f$ is locally $\theta$-Hölder continuous around $x_{\star}$), then $\hat{\hat{\partial}}^2 \varphi_{\gamma}$ is a $(\theta$-order) LNA for $\nabla \varphi_{\gamma}$ at $x_{\star}$.

**Proof.** Let $H_{\epsilon} \in \hat{\hat{\partial}}^2 \varphi_{\gamma}(x) = \{Q_{\gamma}(x)U \mid U \in \mathcal{R}_{\gamma}(x)\}$, so that $H_{\epsilon} = Q_{\gamma}(x)U_{\epsilon}$ for some $U_{\epsilon} \in \mathcal{R}_{\gamma}(x)$. Then,
\[ \|\nabla \varphi_{\gamma}(x) + H_{\epsilon}(x_{\star} - x) - \nabla \varphi_{\gamma}(x_{\star})\| = \|Q_{\gamma}(x)R_{\gamma}(x) + Q_{\gamma}(x)U_{\epsilon}(x_{\star} - x_{\star})\| \]
\[ = \|Q_{\gamma}(x)[R_{\gamma}(x) + U_{\epsilon}(x_{\star} - x_{\star}) - R_{\gamma}(x_{\star})]\| \]
\[ \leq \|R_{\gamma}(x) + U_{\epsilon}(x_{\star} - x_{\star}) - R_{\gamma}(x_{\star})\|, \]
where in the equalities we used the fact that $\nabla \varphi_{\gamma}(x_{\star}) = R_{\gamma}(x_{\star}) = 0$, and in the inequality the fact that $\|Q_{\gamma}\| \leq 1$. Since $\mathcal{R}_{\gamma}$ is a $(\theta$-order) LNA of $R_{\gamma}$ at $x_{\star}$, the last term is $o(\|x - x_{\star}\|)$ (resp. $O(\|x - x_{\star}\|^{1+\theta})$).

As shown in the next result, although the FBE is in general not convex, for $\gamma$ small enough every element of $\hat{\hat{\partial}}^2 \varphi_{\gamma}(x)$ is a (symmetric and) positive semidefinite matrix. Moreover, the eigenvalues are lower and upper bounded uniformly over all $x \in \mathbb{R}^n$.

**Proposition 4.15.** Let $\gamma \leq 1/L_f$ and $H \in \hat{\hat{\partial}}^2 \varphi_{\gamma}(x)$ be fixed. Then, $H \in \mathcal{S}_{\mu_f}(\mathbb{R}^n)$ with
\[ \lambda_{\min}(H) = \min\left\{ (1 - \gamma \mu_f)\mu_f, (1 - \gamma L_f)L_f \right\} \text{ and } \lambda_{\max}(H) = \gamma^{-1}(1 - \gamma \mu_f), \]
where $\mu_f \geq 0$ is the modulus of strong convexity of $f$.

**Proof.** Fix $x \in \mathbb{R}^n$ and let $Q := Q_{\gamma}(x)$. Any $H \in \hat{\hat{\partial}}^2 \varphi_{\gamma}(x)$ can be expressed as $H = \gamma^{-1}Q(1 - PQ)$ for some $P \in \partial_C(\text{prox}_{\gamma q_{\gamma}})(x - \gamma \nabla f(x))$. Since both $Q$ and $P$ are symmetric (cf. Thm. 4.12), it follows that so is $H$. Moreover, for all $d \in \mathbb{R}^n$
In any such case, there exist $\delta, \kappa > 0$ such that

\[ \langle H d, d \rangle = \gamma^{-1}(Q d, d) - \gamma^{-1}(P Q d, Q d) \]

\[ \geq \gamma^{-1}(Q d, d) - \gamma^{-1}\|Q d\|^2 \]

\[ = \langle (I - \gamma \nabla^2 f(x)) \nabla^2 f(x) d, d \rangle \]

\[ \geq \min \left\{ (1 - \gamma \mu_f) \mu_f, (1 - \gamma L_f) L_f \right\} \|d\|^2. \]

On the other hand, since $P \geq 0$ (cf. Thm. 4.12) and thus $QPQ \geq 0$, we can upper bound (4.14) as

\[ \langle H d, d \rangle \leq \gamma^{-1}(Q d, d) \leq \|Q\|\|d\|^2 \leq \gamma^{-1}(1 - \gamma \mu_f)\|d\|^2. \]

\[ \square \]

The next lemma links the behavior of the FBE close to a solution of (1.1) and a nonsingularity assumption on the elements of $\tilde{\partial}^2 \varphi_\gamma(x_\star)$. Part of the statement is similar to [19, Lem. 7.2.10]; however, here $\nabla \varphi_\gamma$ is not required to be locally Lipschitz around $x_\star$.

**Lemma 4.16.** Let $x_\star \in \text{arg min} \varphi$ and $\gamma \in (0, 1/L_f)$. If $\text{prox}_{g_d}$ is semismooth at $x_\star - \gamma \nabla f(x_\star)$, then the following conditions are equivalent:

(a) $x_\star$ is a locally strong minimum for $\varphi$ (or, equivalently, for $\varphi_\gamma$);

(b) every element of $\tilde{\partial}^2 \varphi_\gamma(x_\star)$ is nonsingular.

In any such case, there exist $\delta, \kappa > 0$ such that

\[ \|x - x_\star\| \leq \kappa \|R_\gamma(x)\| \text{ and } \max \left\{ \|H\|, \|H^{-1}\| \right\} \leq \kappa, \]

for any $x \in B(x_\star; \delta)$ and $H \in \tilde{\partial}^2 \varphi_\gamma(x)$.

**Proof.** Observe first that Cor. 4.14 ensures that $\tilde{\partial}^2 \varphi_\gamma$ is a LNA of $\nabla \varphi_\gamma$ at $x_\star$, thus semicontinuous and compact valued (by definition). In particular, the last claim follows from [19, Lem. 7.5.2].

\[ \blacklozenge \ 4.16(a) \Rightarrow 4.16(b) \] It follows from Cor. 4.10 that there exists $\mu, \delta > 0$ such that

\[ \varphi_\gamma(x) - \varphi_\gamma \geq \frac{\delta}{2} \|x - x_\star\|^2 \text{ for all } x \in B(x_\star; \delta). \]

In particular, for all $H \in \tilde{\partial}^2 \varphi_\gamma(x_\star)$ and $x \in B(x_\star; \delta)$ we have

\[ \frac{\delta}{2} \|x - x_\star\|^2 \leq \varphi_\gamma(x) - \varphi_\gamma = \frac{1}{2} \langle H(x - x_\star), x - x_\star \rangle + o(\|x - x_\star\|^2). \]

Let $v_{\min}$ be a unitary eigenvector of $H$ corresponding to the minimum eigenvalue $\lambda_{\min}(H)$. Then, for all $\varepsilon \in (-\delta, \delta)$ the point $x_\varepsilon = x_\star + \varepsilon v_{\min}$ is $\delta$-close to $x_\star$ and thus

\[ \frac{1}{2} \lambda_{\min}(H) \varepsilon^2 \geq \frac{\delta}{2} \varepsilon^2 + o(\varepsilon^2) \geq \frac{\delta}{2} \varepsilon^2, \]

where the last inequality holds up to possibly restricting $\delta$ (and thus $\varepsilon$). The claim now follows from the arbitrariness of $H \in \tilde{\partial}^2 \varphi_\gamma(x_\star)$.

\[ \blacklozenge \ 4.16(a) \Leftrightarrow 4.16(b) \] Easily follows by reversing the arguments of the other implication. \[ \square \]
5 Forward-backward truncated-Newton algorithm (FBTN)

Having established the equivalence between minimizing \( \varphi \) and \( \varphi_\gamma \), we may recast problem (1.1) into the smooth unconstrained minimization of the FBE. Under some assumptions the elements of \( \partial^2 \varphi_\gamma \) mimic second-order derivatives of \( \varphi_\gamma \), suggesting the employment of Newton-like update directions \( d = -(H+\delta I)^{-1} \nabla \varphi_\gamma (x) \) with \( H \in \partial^2 \varphi_\gamma (x) \) and \( \delta > 0 \) (the regularization term \( \delta I \) ensures the well definedness of \( d \), as \( H \) is positive semidefinite, see Prop. 4.15). If \( \delta \) and \( \epsilon \) are suitably selected, under some nondegeneracy assumptions updates \( x^\tau = x + d \) are locally superlinearly convergent. Since such \( d \)'s are directions of descent for \( \varphi_\gamma \), a possible globalization strategy is an Armijo-type linesearch. Here, however, we follow the simpler approach proposed in [71, 75] that exploits the basic properties of the FBE investigated in Section 4.1. As we will discuss shortly after, this is also advantageous from a computational point of view, as it allows an arbitrary warm starting for solving the underlying linear system.

Let us elaborate on the linesearch. To this end, let \( x \) be the current iterate; then, Thm. 4.2 ensures that \( \varphi_\gamma (T_\gamma (x)) \leq \varphi_\gamma (x) - \gamma \frac{1-\gamma L_\gamma}{2} \| R_\gamma (x) \|^2 \). Therefore, unless \( R_\gamma (x) = 0 \), in which case \( x \) would be a solution, for any \( \sigma \in (0, \gamma \frac{1-\gamma L_\gamma}{2}) \) the strict inequality \( \varphi_\gamma (T_\gamma (x)) < \varphi_\gamma (x) - \sigma \| R_\gamma (x) \|^2 \) is satisfied. Due to the continuity of \( \varphi_\gamma \), all points sufficiently close to \( T_\gamma (x) \) will also satisfy the inequality, thus so will the point \( x^\tau = (1-\tau)T_\gamma (x) + \tau (x + d) \) for small enough stepsizes \( \tau \). This fact can be used to enforce the iterates to sufficiently decrease the value of the FBE, cf. (5.1), which straightforwardly implies optimality of all accumulation points of the generated sequence. We defer the details to the proof of Theorem 5.2. In Theorems 5.6 and 5.7 we will provide conditions ensuring acceptance of unit stepsizes so that the scheme reduces to a regularized version of the (undamped) linear Newton method [19, Alg. 7.5.14] for solving \( \nabla \varphi_\gamma (x) = 0 \), which, under due assumptions, converges superlinearly.

In order to ease the computation of \( d^k \), we allow for inexact solutions of the linear system by introducing a tolerance \( \epsilon_k > 0 \) and requiring \( \|(H_k + \delta_k I) d^k + \nabla \varphi_\gamma (x^k)\| \leq \epsilon_k \). Since \( H_k + \delta_k I \) is positive definite, inexact solutions of the linear system can be efficiently retrieved by means of CG (Alg. 2), which only requires matrix-vector products and thus only (generalized) directional derivatives, namely, (generalized) derivatives (denoted as \( \overline{\partial} \)) of the single-variable functions \( t \mapsto \text{prox}_{\varphi_\gamma} (x + t \lambda) \) and \( t \mapsto \nabla f (x + t \lambda) \), as opposed to computing the full (generalized) Hessian matrix. To further enhance computational efficiency, we may warm start the CG method with the previously computed direction, as eventually subsequent update directions are expected to have a small difference. Notice that this warm starting does not ensure that the provided (inexact) solution \( d^k \) is a direction of descent for \( \varphi_\gamma \); either way, this property is not required by the adopted linesearch, showing a considerable advantage over classical Armijo-type rules. Putting all these facts together we obtain the proposed FBE-based truncated-Newton algorithm FBTN (Alg. 1) for convex composite minimization.
Algorithm 1 (FBTN) Forward-Backward Truncated-Newton method

| Require  | γ ∈ (0, 1/ξ); σ ∈ (0, \(\frac{1+\lambda \epsilon}{2}\)); \(\tilde{\eta}, \tilde{\zeta} \in (0, 1)\); \(\rho, \nu \in (0, 1)\) |
|----------|----------------------------------------------------------------------------------------------------------------------------------|
|          | initial point \(x_0 \in \mathbb{R}^n\); accuracy \(\epsilon > 0\)                                                                 |
| Provide  | \(\epsilon\)-suboptimal solution \(x^*\) (i.e., such that \(\|R_\gamma(x^*)\| \leq \epsilon\))                                                                 |
| Initialize | \(k \leftarrow 0\)                                                                                                             |
| 1.1: while | \(\|R_\gamma(x^*)\| > \epsilon\) do                                                                                           |
| 1.2:     | \(\delta_k \leftarrow \tilde{\zeta} \|\nabla_\gamma \phi_\gamma(x^*)\|^2\); \(\eta_k \leftarrow \min\{\tilde{\eta}, \|\nabla_\gamma \phi_\gamma(x^*)\|^2\}\); \(s_k \leftarrow \eta_k \|\nabla_\gamma \phi_\gamma(x^*)\|\) |
| 1.3:     | Apply CG (Alg. 2) to find an \(s_k\)-approximate solution \(d^k\) to \([H_k + \delta_k I]d^k = -\nabla_\gamma \phi_\gamma(x^*)\) for some \(H_k \in \partial^2 \phi_\gamma(x^*)\) |
| 1.4:     | Let \(\tau_k\) be the maximum in \(2^{-i} \mid i \in \mathbb{N}\) such that \(\phi_\gamma(x^{k+1}) \leq \phi_\gamma(x^k) - \sigma \|R_\gamma(x^k)\|^2\) |
|          | where \(x^{k+1} \leftarrow (1 - \tau_k) T_\gamma(x^k) + \tau_k [x^k + d^k]\)                                                                 |
| 1.5:     | \(k \leftarrow k + 1\) and go to step 1.2                                                                                     |
| 1.6: end while |

Remark 5.1 (Adaptive variant when \(L_f\) is unknown). In practice, no prior knowledge of the global Lipschitz constant \(L_f\) is required for FBTN. In fact, replacing \(L_f\) with an initial estimate \(L > 0\), the following instruction can be added at the beginning of each iteration, before step 1.2:

\(1.0: \quad \delta \leftarrow T_\gamma(x^k)\)

while \(f(\delta) > f(x^k) + \langle \nabla f(x^k), \delta - x^k \rangle + \frac{L}{2} \|\delta - x^k\|^2\) do

\(\gamma \leftarrow \gamma_2; \, L \leftarrow 2L; \, \delta \leftarrow T_\gamma(x^k)\)

Algorithm 2 (CG) Conjugate Gradient for computing the update direction

| Require  | \(\nabla_\gamma \phi_\gamma(x^k)\); \(\delta_k\); \(s_k\); \(d^{k-1}\) (set to 0 if \(k = 0\)) |
|----------|----------------------------------------------------------------------------------------------------------------------------------|
|          | (generalized) directional derivatives \(\lambda \mapsto \frac{\partial \phi_\gamma}{\partial \lambda}(x^k - \gamma \nabla f(x^k))\) and \(\lambda \mapsto \frac{\partial^2 \phi_\gamma}{\partial \lambda^2}(x^k)\) |
| Provide  | update direction \(d^k\)                                                                                                     |
| Initialize | \(e, p \leftarrow -\nabla_\gamma \phi_\gamma(x^k)\); warm start \(d^k \leftarrow d^{k-1}\)                                                                                     |
| 2.1: while | \(\|e\| > \epsilon_k\) do                                                                                                  |
| 2.2:     | \(u \leftarrow \frac{\partial \phi_\gamma}{\partial \lambda}(x^k)\)                                                             |
| 2.3:     | \(v \leftarrow p - \gamma u\)                                                                                               |
| 2.4:     | \(w \leftarrow p - \frac{\partial \phi_\gamma}{\partial \lambda^2}(x^k - \gamma \nabla f(x^k))\)                             |
| 2.5:     | \(z \leftarrow \delta_k p + w - \gamma \frac{\partial \phi_\gamma}{\partial \lambda^2}(x^k)\)                             |
| 2.6:     | \(\alpha \leftarrow \|e\|^2/(p, z)\)                                                                                         |
| 2.7:     | \(d^k \leftarrow d^k + \alpha p; \, e^* \leftarrow e - \alpha z\)                                                            |
| 2.8:     | \(p \leftarrow e^* + \left(\|e^*\|/\alpha\right)^2 p\)                                                                     |
| 2.9:     | \(e \leftarrow e^*\)                                                                                                       |
| 2.10: end while |
Moreover, since positive definiteness of $H_k + \delta_k I$ is ensured only for $\gamma \leq 1/L_f$ where $L_f$ is the true Lipschitz constant of $\nabla \varphi_\gamma$ (cf. Prop. 4.15), special care should be taken when applying CG in order to find the update direction $d_k$. Specifically, CG should be stopped prematurely whenever $\langle p, z \rangle \leq 0$ in step 2.8, in which case $\gamma \leftarrow \gamma/2$, $L \leftarrow 2L$ and the iteration should start again from step 1.2.

Whenever the quadratic bound (2.1) is violated with $L$ in place of $L_f$ in (2.1), the estimated Lipschitz constant $L$ is increased, $\gamma$ is decreased accordingly, and the proximal gradient point $\bar{x}_k$ with the new stepsize $\gamma$ is evaluated. Since replacing $L_f$ with any $L \geq L_f$ still satisfies (2.1), it follows that $L$ is incremented only a finite number of times. Therefore, there exists an iteration $k_0$ starting from which $\gamma$ and $L$ are constant; in particular, all the convergence results here presented remain valid starting from iteration $k_0$, at least. Moreover, notice that this step does not increase the complexity of the algorithm, since both $\bar{x}_k$ and $\nabla f(x_k)$ are needed for the evaluation of $\varphi_\gamma(x_k)$.

$\square$

### 5.1 Subsequential and linear convergence

Before going through the convergence proofs let us spend a few lines to emphasize that FBTN is a well-defined scheme. First, that a matrix $H_k$ as in step 1.2 exists is due to the nonemptyness of $\partial^2 \varphi_\gamma(x_k)$ (cf. §4.3). Second, since $\delta_k > 0$ and $H_k \succeq 0$ (cf. Prop. 4.15) it follows that $H_k + \delta_k I$ is (symmetric and) positive definite, and thus CG is indeed applicable at step 1.2.

Having clarified this, the proof of the next result falls as a simplified version of [75, Lem. 5.1 and Thm. 5.6]; we elaborate on the details for the sake of self-inclusiveness. To rule out trivialities, in the rest of the paper we consider the limiting case of infinite accuracy, that is $\varepsilon = 0$, and assume that the termination criterion $\|R_\gamma(x_k)\| = 0$ is never met. We shall also work under the assumption that a solution to the investigated problem (1.1) exists, thus in particular that the cost function $\varphi$ is lower bounded.

**Theorem 5.2** (Subsequential convergence). Every accumulation point of the sequence $(x_k)_{k \in \mathbb{N}}$ generated by FBTN (Alg. 1) is optimal.

**Proof.** Observe that

$$\varphi_\gamma(x_k - \gamma R_\gamma(x_k)) \leq \varphi_\gamma(x_k) - \gamma \frac{1 - L_f}{2} \|R_\gamma(x_k)\|^2 < \varphi_\gamma(x_k) - \sigma \|R_\gamma(x_k)\|^2$$

and that $x_k \rightarrow T_\gamma(x_k)$ as $\tau_k \rightarrow 0$. Continuity of $\varphi_\gamma$ ensures that for small enough $\tau_k$ the linesearch condition (5.1) is satisfied, in fact, regardless of what $d_k$ is. Therefore, for each $k$ the stepsize $\tau_k$ is decreased only a finite number of times. By telescoping the linesearch inequality (5.1) we obtain

$$\sigma \sum_{k \in \mathbb{N}} \|R_\gamma(x_k)\|^2 \leq \sum_{k \in \mathbb{N}} [\varphi_\gamma(x_k) - \varphi_\gamma(x^{k+1})] \leq \varphi_\gamma(x^0) - \varphi_* < \infty$$
and in particular $R_{\gamma}(x^k) \to 0$. Since $R_{\gamma}$ is continuous we infer that every accumulation point $x_* \in (x^k)_{k \in \mathbb{N}}$ satisfies $R_{\gamma}(x_*) = 0$, hence $x_* \in \arg \min \varphi$, cf. (3.8). □

**Remark 5.3.** Since FBTN is a descent method on $\varphi_{\gamma}$, as ensured by the linesearch condition (5.1), from Proposition 4.8 it follows that a sufficient condition for the existence of cluster points is having $\varphi$ with bounded level sets or, equivalently, having $\arg \min \varphi$ bounded (cf. Lem. A.1). □

As a straightforward consequence of Lemma 4.7, from the linesearch condition (5.1) we infer $Q$-linear decrease of the FBE along the iterates of FBTN provided that the original function $\varphi$ has the quadratic growth property. In particular, although the quadratic growth is a local property, $Q$-linear convergence holds globally, as described in the following result.

**Theorem 5.4** ($Q$-linear convergence of FBTN under quadratic growth). Suppose that $\varphi$ satisfies the quadratic growth with constants $(\mu, \nu)$. Then, the iterates of FBTN (Alg. 1) decrease $Q$-linearly the value of $\varphi_{\gamma}$ as

$$\varphi_{\gamma}(x^{k+1}) - \varphi_* \leq \left(1 - \frac{2\nu'}{\mu(1+\gamma L_f)}\right)(\varphi_{\gamma}(x^k) - \varphi_*) \quad \forall k \in \mathbb{N},$$

where

$$\mu' := \begin{cases} 
\mu & \text{if } \varphi_{\gamma}(x_0) \leq \varphi_* + \nu, \\
\frac{\nu}{\mu} & \min\left(1, \frac{\nu}{\varphi_{\gamma}(x_0)-\varphi_*-\nu}\right) & \text{otherwise}. 
\end{cases}$$

**Proof.** Since FBTN is a descent method on $\varphi_{\gamma}$, it holds that $(x^k)_{k \in \mathbb{N}} \subseteq \text{lev}_{\varphi_{\gamma}} \varphi_{\gamma}$ with $\alpha = \varphi_{\gamma}(x^0)$. It follows from Lemma 3.7 that $\varphi$ satisfies the quadratic growth condition with constants $(\mu', \varphi_{\gamma}(x^0))$, with $\mu'$ is as in the statement. The claim now follows from the inequality ensured by linesearch condition (5.1) combined with Lemma 4.7. □

### 5.2 Superlinear convergence

In this section we provide sufficient conditions that enable superlinear convergence of FBTN. In the sequel, we will make use of the notion of *superlinear directions* that we define next.

**Definition 5.5** (Superlinear directions). Suppose that $X_* \neq \emptyset$ and consider the iterates generated by FBTN (Alg. 1). We say that $(d^k)_{k \in \mathbb{N}} \subseteq \mathbb{R}^n$ are superlinearly convergent directions if

$$\lim_{k \to \infty} \frac{\text{dist}(x^k + d^k, X_*)}{\text{dist}(x^k, X_*)} = 0.$$

If for some $q > 1$ the condition can be strengthened to
then we say that \((d^k)_{k \in \mathbb{N}}\) are superlinearly convergent directions with order \(q\).

We remark that our definition of superlinear directions extends the one given in [19, §7.5] to cases in which \(X_*\) is not a singleton. The next result constitutes a key component of the proposed methodology, as it shows that the proposed algorithm does not suffer from the Maratos’ effect [44], a well-known obstacle for fast local methods that inhibits the acceptance of the unit stepsize. On the contrary, we will show that whenever the directions \((d^k)_{k \in \mathbb{N}}\) computed in FBTN are superlinear, then indeed the unit stepsize is eventually always accepted, and the algorithm reduces to a regularized version of the (undamped) linear Newton method [19, Alg. 7.5.14] for solving \(\nabla \varphi(x) = 0\) or, equivalently, \(R_\gamma(x) = 0\), and \(\text{dist}(x^k, X_*)\) converges superlinearly.

**Theorem 5.6** (Acceptance of the unit stepsize and superlinear convergence). Consider the iterates generated by FBTN (Alg. 1). Suppose that \(\varphi\) satisfies the quadratic growth (locally) and that \((d^k)_{k \in \mathbb{N}}\) are superlinearly convergent directions (with order \(q\)). Then, there exists \(\bar{k} \in \mathbb{N}\) such that

\[
\varphi_\gamma(x^\bar{k} + d^\bar{k}) \leq \varphi_\gamma(x^\bar{k}) - \sigma \|R_\gamma(x^\bar{k})\|^2 \quad \forall \bar{k} \geq \bar{k}.
\]

In particular, eventually the iterates reduce to \(x^{\bar{k} + 1} = x^\bar{k} + d^\bar{k}\), and \(\text{dist}(x^\bar{k}, X_*\) converges superlinearly (with order \(q\)).

**Proof.** Without loss of generality we may assume that \((x^k)_{k \in \mathbb{N}}\) and \((x^k + d^k)_{k \in \mathbb{N}}\) belong to a region in which quadratic growth holds. Denoting \(\varphi_* := \min \varphi\), since \(\varphi_\gamma\) also satisfies the quadratic growth (cf. Prop. 4.9(ii)) if follows that

\[
\mu' \frac{\text{dist}(x^k, X_*)^2}{\|R_\gamma(x^k)\|^2} \to 0 \quad \text{as} \quad k \to \infty.
\]

Moreover, we know from Lem. 4.7 that

\[
\varphi_\gamma(x^k + d^k) - \varphi_* \leq c \|R_\gamma(x^k + d^k)\|^2 \leq c' \text{dist}(x^k + d^k, X_*)^2
\]

for some constants \(c, c' > 0\), where in the second inequality we used Lipschitz continuity of \(R_\gamma\) (Lem. A.3) together with the fact that \(R_\gamma(x_*) = 0\) for all points \(x_* \in X_*\). By combining the last two inequalities, we obtain

\[
t_k := \frac{\varphi_\gamma(x^k + d^k) - \varphi_*}{\varphi_\gamma(x^k) - \varphi_*} \leq 2c' \frac{\text{dist}(x^k + d^k, X_*)^2}{\mu' \text{dist}(x^k, X_*)^2} \to 0 \quad \text{as} \quad k \to \infty. \tag{5.2}
\]

Moreover,

\[
\varphi_\gamma(x^k) - \varphi_* \geq \varphi_\gamma(x^k) - \varphi(T_\gamma(x^k)) \geq \frac{\gamma - \gamma L_*}{2} \|R_\gamma(x^k)\|^2. \tag{5.3}
\]

Thus,
\[ \varphi(x^k + d^k) - \varphi(x^k) = [\varphi(x^k + d^k) - \varphi_*] - [\varphi(x^k) - \varphi_*] = (t_k - 1)[\varphi(x^k) - \varphi_*] \]

and since \( t_k \to 0 \), eventually it holds that \( t_k \leq 1 - \frac{2\tau}{\|1 - \varphi\|} \in (0, 1) \), resulting in
\[ \leq - \sigma |R_k(x^k)|^2. \]

\[ \square \]

**Theorem 5.7.** Consider the iterates generated by FBTN (Alg. 1). Suppose that \( \varphi \) satisfies the quadratic growth (locally), and let \( x_* \) be the limit point of \((x^k)_{k \in \mathbb{N}}\).\(^2\) Then, \((d^k)_{k \in \mathbb{N}}\) are superlinearly convergent directions provided that

(i) either \( R_\gamma \) is strictly differentiable at \( x_* \)\(^3\) and there exists \( D > 0 \) such that
\[ ||d^k|| \leq D||\nabla \varphi_\gamma(x^k)|| \text{ for all } k \text{'s}, \]

(ii) or \( X_* = \{x_*\} \) and prox\( \varphi \) is semismooth at \( x_* - \gamma \nabla f(x_*). \) In this case, if prox\( \varphi \) is \( \theta \)-order semismooth at \( x_* - \gamma \nabla f(x_*) \) and \( \nabla^2 f \) is \( \theta \)-Hölder continuous close to \( x_* \), then the order of superlinear convergence is at least
\[ 1 + \min \{\rho, \vartheta, \nu\}. \]

**Proof.** Due to Prop. 4.9 and Thm. 5.4, if \( X_* = \{x_*\} \) then the sequence \((x^k)_{k \in \mathbb{N}}\) converges to \( x_* \). Otherwise, the hypothesis ensure that
\[ ||x^{k+1} - x^k|| = \tau_k ||d^k|| \leq D||\nabla \varphi_\gamma(x^k)|| \leq D||R_k(x^k)||, \]

from which we infer that \((||x^{k+1} - x^k||)_{k \in \mathbb{N}}\) is \( R \)-linearly convergent, hence that \((x^k)_{k \in \mathbb{N}}\) is a Cauchy sequence, and again we conclude that the limit point \( x_* \) indeed exists. Moreover, in light of Prop. 4.9 we have that \((x^k)_{k \in \mathbb{N}}\) is contained in a level set of \( \varphi \), where \( \varphi \) has quadratic growth. To establish a notation, let \( e^k := [H_k + \delta_k I]d^k + \nabla \varphi_\gamma(x^k) \) be the error in solving the linear system at step 1.2, so that
\[ ||e^k|| \leq \epsilon_k \leq ||\nabla \varphi_\gamma(x^k)||^{1+\rho}, \quad (5.4) \]

(cf. step 1.2), and let \( H_k = Q_\gamma(x^k)U_k \) for some \( U_k \in R_\gamma(x^k) \), see (4.11). Let us now analyze the two cases separately.

\[ \star 5.7(i) \] Let \( x_* \rightleftharpoons \mathcal{P}_{X_*} x^k \), so that dist\((x^k, X_*\) = \( ||x^k - x_*||\). Recall that \( \nabla \varphi_\gamma = Q_\gamma R_\gamma \) and that \( (1 - \gamma L_\gamma)I \leq Q_\gamma \leq I \). Since \( R_\gamma(x_*^k) = 0 \), from Lem. A.3 and Thm. 3.8 we infer that there exist \( r_1, r_2 > 0 \) such that
\[ ||R_\gamma(x^k)|| \geq r_1 \text{ dist}\((x^k, X_*\) \quad \text{and} \quad ||\nabla \varphi_\gamma(x^k)|| \leq r_2 \text{ dist}\((x^k, X_*\) \). \quad (5.5) \]

In particular, the assumption on \( d^k \) ensures that \( ||d^k|| = O(\text{dist}\((x^k, X_*\) \)). \]

---

\(^2\) As detailed in the proof, under the assumptions the limit point indeed exists.

\(^3\) From the chain rule of differentiation it follows that \( R_\gamma \) is strictly differentiable at \( x_* \) if \( \text{prox}_{\gamma \varphi} \) is strictly differentiable at \( x_* - \gamma \nabla f(x_*) \) (strict differentiability is closed under composition).
Moreover, due to Lem. 4.16 there exists semismoothness ensures through Prop. 4.13 that holds. Combined with the fact that δo
Since [83, 82] which implies Uk \rightarrow JR(x_\star). In order to bound (b), recall that δk = ζ∥∇ϕk(x_\star)∥" (cf. step 1.2). Then,
\[ \begin{align*}
\text{(a)} & \leq \|R_k(x_\star) - R_k(x_\star) - JR_k(x_\star)\|d_k\| + \|U_k - JR(x_\star)\|d_k\| \\
& = o(\text{dist}(x_\star, X_\star)),
\end{align*} \]
where we used strict differentiability and the fact that ∂\text{e}R_k(x_\star) = \{JR_k(x_\star)\} [15, Prop. 2.2.4] which implies Uk \rightarrow JR(x_\star). Moreover, due to Lem. 4.16 there exists c > 0 such that \|H_k + δ_k\| \leq c for all k’s. We have
\[ \begin{align*}
\|x_\star + d_k - x_\star\| & \leq \|x_\star + [H_k + δ_k\|^{-1}\|H_k + δ_k\|^{-1}(x_\star - x_\star) + e_k - \nablaϕ(x_\star)\| \\
& \leq \|H_k + δ_k\|^{-1}\|H_k + δ_k\|^{-1}\|H_k + δ_k\|^{-1}\|x_\star - x_\star\| + cδ_k\|x_\star - x_\star\| + c\|e_k\| \\
& = c\|Q_k(x_\star)(U_k(x_\star - x_\star) - R_k(x_\star))\| + cδ_k\|x_\star - x_\star\| + c\|e_k\|. 
\end{align*} \]
Since R_k is a LNA at x_\star, it follows that the quantity emphasized in the bracket is a o(∥x_\star - x_\star∥), whereas in case of a (θ-order) LNA the tighter estimate O(∥x_\star - x_\star∥) holds. Combined with the fact that δk = O(∥x_\star - x_\star∥) and ∥e_k∥ = O(∥x_\star - x_\star∥), we conclude that (d_k)_{k∈N} are superlinearly convergent directions, and with order at least 1 + \min{\{ρ, θ, ν\}} in case of θ-order semismoothness.

Problems where the residual is (θ-order) semismooth are quite common. For instance, piecewise affine functions are everywhere strongly semismooth, as it is the case for the residual in lasso problems [67]. On the contrary, when the solution is not unique the condition ∥d∥ ≤ D∥\nablaϕ(x_\star)∥ (or, equivalently, ∥d∥ ≤ D∥R_k(x_\star)∥) is trickier. As detailed in [83, 82], this bound on the directions is ensured if ρ = 1 and for all iterates x_k and points x close enough to the limit point the following smoothness condition holds:
\[ \|R_k(x_\star) + U_k(x - x_\star)\| \leq c\|x - x_\star\|^2 \] (5.6)
Therefore, is piecewise $\vartheta$

In particular, by exploiting this last condition it can be easily verified that if $R_\varphi$ requirement (5.6) in [83, 82] does not.

which is indeed $\parallel$ constant (equal to 0). To see this, notice that (5.7) implies that $\parallel x - x^k \parallel$ for some $c > 0$. In particular, mimicking the arguments in the cited references, since $H_k \geq 0$ and $\parallel Q_\gamma \parallel \leq 1$, observe that

\[
\parallel [H_k + \delta_k I]^{-1} Q_\gamma(x^k) \parallel \leq \parallel [H_k + \delta_k I]^{-1} \parallel \leq \delta_k^{-1}\tag{5.8a}
\]

and

\[
\parallel [H_k + \delta_k I]^{-1} H_k \parallel = \parallel I - \delta_k [H_k + \delta_k I]^{-1} \parallel \leq 2. \tag{5.8b}
\]

Therefore,

\[
\parallel \delta k \parallel = \parallel [H_k + \delta_k I]^{-1} (e^k - \nabla \varphi_\gamma(x^k)) \parallel \\
\leq \parallel [H_k + \delta_k I]^{-1} \parallel \parallel e^k \parallel + \parallel [H_k + \delta_k I]^{-1} Q_\gamma(x^k) (R_\varphi(x^k) + U_k(x^k - x^k)) \parallel \\
+ \parallel [H_k + \delta_k I]^{-1} H_k (x^k - x^k) \parallel \\
\leq \delta_k^{-1} \parallel \nabla \varphi_\gamma(x^k) \parallel^{1+\rho} + c \delta_k^{-1} \parallel x^k - x^k \parallel^{1+\vartheta} + 2 \parallel x^k - x^k \parallel \\
= \epsilon^{-1} \parallel \nabla \varphi_\gamma(x^k) \parallel^{1+\rho-\nu} + c \epsilon^{-1} \parallel x^k - x^k \parallel^{1+\vartheta-\nu} + 2 \parallel x^k - x^k \parallel \\
= O(\text{dist}(x^k, X_\star)^{1+\min(0,\rho-\nu,\vartheta-\nu)}),
\]

which is indeed $O(\parallel \nabla \varphi_\gamma(x^k) \parallel)$ whenever $\nu \leq \min \{ \vartheta, \rho \}$. Some comments are in order on condition (5.7).

1) If $X_\star = \{ x^k \}$ is a singleton, then $x^k$ is fixed to $x^k_\star$ and the requirement reduces to $\vartheta$-order semismoothness at $x^k_\star$.

2) This notion of uniformity is a local property: for any $\epsilon > 0$ the set $X_\star$ can be replaced by $X_\star \cap B(x^k_\star, \epsilon)$.

3) The condition $U \in R_\varphi(x)$ in the limit can be replaced by $U \in \hat{R}_\varphi(x) := \{ \gamma^{-1}(1 - P \gamma(x)) | P \in \partial_\varphi \text{prox}_{\gamma \varphi}(x - \gamma \nabla \varphi(x)) \}$, since $R_\varphi(x) = \text{conv}(\hat{R}_\varphi(x))$.

In particular, by exploiting this last condition it can be easily verified that if $R_\varphi$ is piecewise $\vartheta$-Hölder differentiable around $x^k_\star$, then (5.7) holds, yet the stronger requirement (5.6) in [83, 82] does not.
6 Generalized Jacobians of proximal mappings

In many interesting cases \( \text{prox}_{\gamma g} \) is \( PC^1 \) and thus semismooth. Piecewise quadratic (PWQ) functions comprise a special but important class of convex functions whose proximal mapping is \( PC^1 \). A convex function \( g \) is called PWQ if \( \text{dom} g \) can be represented as the union of finitely many polyhedral sets, relative to each of which \( g(x) \) is given by an expression of the form \( \frac{1}{2} \langle x, Hx \rangle + \langle q, x \rangle + c \) (\( H \in \mathbb{R}^{n \times n} \) must necessarily be symmetric positive semidefinite) [64, Def. 10.20]. The class of PWQ functions is quite general since it includes e.g., polyhedral norms, indicators and support functions of polyhedral sets, and it is closed under addition, composition with affine mappings, conjugation, inf-convolution and inf-projection [64, Props. 10.22 and 11.32]. It turns out that the proximal mapping of a PWQ function is piecewise affine (PWA) [64, 12.30] (\( \mathbb{R}^n \) is partitioned in polyhedral sets relative to each of which \( \text{prox}_{\gamma g} \) is an affine mapping), hence strongly semismooth [19, Prop. 7.4.7]. Another example of a proximal mapping that is strongly semismooth is the projection operator over symmetric cones [73].

A big class with semismooth proximal mapping is formed by the semi-algebraic functions. We remind that a set \( A \subseteq \mathbb{R}^n \) is semi-algebraic if it can be expressed as

\[
A = \bigcup_{i=1}^{p} \bigcap_{j=1}^{q} \{ x \in \mathbb{R}^m | P_{ij}(x) = 0, \ Q_{ij}(x) < 0 \}
\]

for some polynomial functions \( P_{ij}, Q_{ij} : \mathbb{R}^n \to \mathbb{R} \), and that a function \( h : \mathbb{R}^n \to \overline{\mathbb{R}^m} \) is semi-algebraic if \( \text{gph} \ h \) is a semi-algebraic subset of \( \mathbb{R}^{n \times m} \).

**Proposition 6.1.** If \( g : \mathbb{R}^n \to \overline{\mathbb{R}} \) is semi-algebraic, then so are \( g^\gamma \) and \( \text{prox}_{\gamma g} \). In particular, \( g^\gamma \) and \( \text{prox}_{\gamma g} \) are semismooth.

**Proof.** Since \( g^\gamma \) and \( \text{prox}_{\gamma g} \) are both Lipschitz continuous, semismoothness will follow once we show that they are semi-algebraic [11, Rem. 4]. Every polynomial is clearly semi-algebraic, and since the property is preserved under addition [10, Prop. 2.2.6(ii)], the function \( (x,w) \mapsto g(w) + \frac{1}{2\gamma} \| w - x \|^2 \) is semi-algebraic. Moreover, since parametric minimization of a semi-algebraic function is still semi-algebraic (see e.g., [1, §2]), it follows that the Moreau envelope \( g^\gamma \) is semi-algebraic and therefore so is \( h(x,w) := g(w) + \frac{1}{2\gamma} \| w - x \|^2 - g^\gamma(x) \). Notice that \( \text{prox}_{\gamma g}(x) = \{ w \in \mathbb{R}^n | h(x,w) \leq 0 \} \), therefore

\[
\text{gph} \ \text{prox}_{\gamma g} = \{(x,\bar{x}) \in \mathbb{R}^n \times \mathbb{R}^n | \ \text{prox}_{\gamma g}(x) = \bar{x} \}
\]

\[
= \{(x,\bar{x}) \in \mathbb{R}^n \times \mathbb{R}^n | h(x,\bar{x}) \leq 0 \}
\]

\[
= h^{-1}((-\infty,0])
\]

is a semi-algebraic set, since the interval \((-\infty,0]\) is clearly semi-algebraic and thus so is \( h^{-1}((-\infty,0]) \) [10, Prop. 2.2.7]. \( \square \)
In fact, with the same arguments it can be shown that the result still holds if ‘semi-algebraic’ is replaced with the broader notion of ‘tame’, see [11]. Other conditions that guarantee semismoothness of the proximal mapping can be found in [47, 48, 46, 50]. The rest of the section is devoted to collecting explicit formulas of $\partial C \text{prox}_g$ for many known useful instances of convex functions $g$.

6.1 Properties

a. Separable functions.
Whenever $g$ is (block) separable, i.e., $g(x) = \sum_{i=1}^{N} g_i(x_i)$, $x_i \in \mathbb{R}^{n_i}$, $\sum_{i=1}^{N} n_i = n$, then every $P \in \partial C (\text{prox}_g)(x)$ is a (block) diagonal matrix. This has favorable computational implications especially for large-scale problems. For example, if $g$ is the $\ell_1$ norm or the indicator function of a box, then the elements of $\partial C \text{prox}_g(x)$ (or $\partial B \text{prox}_g(x)$) are diagonal matrices with diagonal elements in $[0, 1]$ (or in $\{0, 1\}$).

b. Convex conjugate.
With a simple application of the Moreau’s decomposition [2, Thm. 14.3(ii)], all elements of $\partial C \text{prox}_g^*$ are readily available as long as one can compute $\partial C \text{prox}_g/\gamma$. Specifically,

$$\partial C (\text{prox}_{g/\gamma})(x) = I - \partial C (\text{prox}_{\gamma})(\gamma x).$$

Section 6.2 offers a rich list of sets $D$ for which a close form expression exists.

c. Support function.
The support function of a nonempty closed and convex set $D$ is the proper convex and lsc function $\sigma_D(x) := \sup_{y \in D} \langle x, y \rangle$. Alternatively, $\sigma_D$ can be expressed as the convex conjugate of the indicator function $\delta_D$, and one can use the results of §6.1b to find that

$$\partial C (\text{prox}_{g/\gamma})(x) = I - \partial C (\mathcal{P}_D)(\gamma x).$$

d. Spectral functions.
The eigenvalue function $\lambda : S(\mathbb{R}^{n \times n}) \to \mathbb{R}^n$ returns the vector of eigenvalues of a symmetric matrix in nonincreasing order. Spectral functions are of the form

$$G := h \circ \lambda : S(\mathbb{R}^{n \times n}) \to \overline{\mathbb{R}},$$

where $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper, lsc, convex and symmetric, i.e., invariant under coordinate permutations [35]. Such $G$ inherits most of the properties of $h$ [36, 37]; in particular, its proximal mapping is [56, §6.7]

$$\text{prox}_{gG}(X) = Q \text{diag}(\text{prox}_{hG}(\lambda(X))) Q^T,$$
where $X = Q \text{diag}(\lambda(X))Q^T$ is the spectral decomposition of $X$ ($Q$ is an orthogonal matrix). If, additionally,

$$h(x) = g(x_1) + \cdots + g(x_N)$$

(6.2)

for some $g : \mathbb{R} \rightarrow \mathbb{R}$, then

$$\text{prox}_{gh}(x) = (\text{prox}_{g_1}(x_1), \ldots, \text{prox}_{g_N}(x_N)),$$

and therefore the proximal mapping of $G$ can be expressed as

$$\text{prox}_{y}G(x) = Q \text{diag}(\text{prox}_{g_1}(\lambda_1(X)), \ldots, \text{prox}_{g_N}(\lambda_N(X)))Q^T,$$

(6.3)

[9, Chap. V], [29, Sec. 6.2]. Now we can use the theory of nonsmooth symmetric matrix-valued functions developed in [14] to analyze differentiability properties of $\text{prox}_{y}G$. In particular, $\text{prox}_{y}G$ is (strongly) semismooth at $X$ iff $\text{prox}_{y}G$ is (strongly) semismooth at the eigenvalues of $X$ [14, Prop. 4.10]. Moreover, for any $X \in S(\mathbb{R}^{m,n})$ and $P \in \partial_B(\text{prox}_{y}G)(X)$ we have [14, Lem. 4.7]

$$P(X) = Q(Q^T \ominus (Q^T X Q))Q^T,$$

(6.4)

where $\ominus$ denotes the Hadamard product and for vectors $u, v \in \mathbb{R}^n$ we defined $Q_{u,v}^T$ as the $n \times n$ matrix

$$(Q_{u,v}^T)_{ij} := \begin{cases} \frac{\partial_B \text{prox}_{g_i}(u)}{\|u-v\|} & \text{if } i = j, \\ \frac{\text{prox}_{g_i}(u) - \text{prox}_{g_j}(v)}{u-v} & \text{otherwise}. \end{cases}$$

(6.5)

e. Orthogonally invariant functions. A function $G : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ is called orthogonally invariant if $G(UXV^T) = G(X)$ for all $X \in \mathbb{R}^{m \times n}$ and orthogonal matrices $U \in \mathbb{R}^{m \times m}, V \in \mathbb{R}^{n \times n}.$

A function $h : \mathbb{R}^q \rightarrow \mathbb{R}$ is absolutely symmetric if $h(Qx) = h(x)$ for all $x \in \mathbb{R}^q$ and any generalized permutation matrix $Q$, i.e., a matrix $Q \in \mathbb{R}^{q \times q}$ that has exactly one nonzero entry in each row and each column, that entry being $\pm 1$ [34]. There is a one-to-one correspondence between orthogonally invariant functions on $\mathbb{R}^{m \times n}$ and absolutely symmetric functions on $\mathbb{R}^q$. Specifically, if $G$ is orthogonally invariant then

$$G(X) = h(\sigma(X))$$

for the absolutely symmetric function $h(x) = G(\text{diag}(x))$. Here, for $X \in \mathbb{R}^{m \times n}$ and $q := \min\{m,n\}$ the spectral function $\sigma : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^q$ returns the vector of its singular values in nonincreasing order. Conversely, if $h$ is absolutely symmetric then $G(X) = h(\sigma(X))$ is orthogonally invariant. Therefore, convex analytic and generalized differentiability properties of orthogonally invariant functions

---

*In case of complex-valued matrices, functions of this form are known as unitarily invariant [34].*
can be easily derived from those of the corresponding absolutely symmetric functions [34]. For example, assuming for simplicity that \( m \leq n \), the proximal mapping of \( G \) is given by [56, Sec. 6.7]

\[
\text{prox}_{\gamma G}(X) = U \text{diag}(\text{prox}_{\gamma h}(\sigma(X)))V_1^T,
\]

where \( X = U \begin{bmatrix} \text{diag}(\sigma(X)) & 0 \end{bmatrix} V_1^T \) is the singular value decomposition of \( X \). If we further assume that \( h \) has a separable form as in (6.2), then

\[
\text{prox}_{\gamma G}(X) = U \Sigma_g(X) V_1^T, \tag{6.6}
\]

where

\[
\Sigma_g(X) = \text{diag}(\text{prox}_{\gamma_1}(\sigma_1(X)), \ldots, \text{prox}_{\gamma_n}(\sigma_n(X))).
\]

Functions of this form are called nonsymmetric matrix-valued functions. We also assume that \( g \) is a non-negative function such that \( g(0) = 0 \). This implies that \( \text{prox}_{\gamma g}(0) = 0 \) and guarantees that the nonsymmetric matrix-valued function (6.6) is well defined [80, Prop. 2.1.1]. Now we can use the results of [80, §2] to draw conclusions about generalized differentiability properties of \( \text{prox}_{\gamma G} \). For example, through [80, Thm. 2.27] we have that \( \text{prox}_{\gamma G} \) is continuously differentiable at \( X \) if and only if \( \text{prox}_{\gamma g} \) is continuously differentiable at the singular values of \( X \). Furthermore, \( \text{prox}_{\gamma G} \) is (strongly) semismooth at \( X \) if \( \text{prox}_{\gamma g} \) is (strongly) semismooth at the singular values of \( X \) [80, Thm. 2.3.11].

For any \( X \in \mathbb{R}^{m \times n} \) the generalized Jacobian \( \partial_B(\text{prox}_{\gamma G})(X) \) is well defined and nonempty, and any \( P \in \partial_B(\text{prox}_{\gamma G})(X) \) acts on \( H \in \mathbb{R}^{m \times n} \) as [80, Prop. 2.3.7]

\[
P(H) = U \left[ \Omega_{\sigma,\sigma}^g \circ \left( \frac{H_1 + H}{2} \right) + \Omega_{\sigma,\sigma}^g \circ \left( \frac{H_1 - H}{2} \right) \right] V,
\]

where \( V = [V_1 V_2] \), \( H_1 = U^T H V_1 \in \mathbb{R}^{m \times n} \), \( H_2 = U^T H V_2 \in \mathbb{R}^{m \times (n-m)} \) and matrices \( \Omega \) are as in (6.5).

### 6.2 Indicator functions

Smooth constrained convex problems

\[
\text{minimize } f(x) \quad \text{subject to } x \in D \tag{6.8}
\]

can be cast in the composite form (1.1) by encoding the feasible set \( D \) with the indicator function \( g = \delta_D \). Whenever \( \mathcal{P}_D \) is efficiently computable, then algorithms like the forward-backward splitting (3.5) can be conveniently considered. In the following we analyze the generalized Jacobian of some of such projections.

**a. Affine sets.** \( D = \{ x \in \mathbb{R}^n \mid Ax = b \} \) for some \( A \in \mathbb{R}^{m \times n} \) and \( b \in \mathbb{R}^m \).

In this case, \( \mathcal{P}_D(x) = x - A^\dagger (Ax - b) \) where \( A^\dagger \) is the Moore-Penrose pseudoinverse of \( A \). For example, if \( A \) is surjective (i.e., it has full row rank and thus
b. Polyhedral sets. \( D = \{ x \in \mathbb{R}^n \mid Ax = b, \ Cx \leq d \} \), for some \( A \in \mathbb{R}^{p \times n}, \ b \in \mathbb{R}^p, \ C \in \mathbb{R}^{q \times n} \) and \( d \in \mathbb{R}^q \).

It is well known that \( \mathcal{P}_D \) is piecewise affine. In particular, let
\[ I_D := \{ I \subseteq \{ 1 \ldots q \} \mid \exists x \in \mathbb{R}^n : Ax = b, \ C_i x = d_i, i \in I, \ C_j x < d_j \ j \notin I \}. \]

Then, the faces of \( D \) can be indexed with the elements of \( I \) [66, Prop. 2.1.3]: for each \( I \in I_D \) let
\[ F_I := \{ x \in D \mid C_i x = d_i, i \in I \} \]
be the \( I \)-th face of \( D \),
\[ S_I := \text{aff} \ F_I = \{ x \in \mathbb{R}^n \mid Ax = b, \ C_i x = d_i, i \in I \} \]
be the hyperplane containing the \( I \)-th face of \( D \),
\[ N_I := \text{range} A^T + \text{cone} \{ C_i \} \]
be the normal cone to any point in the relative interior of \( F_I \) [66, Eq. (2.44)],\(^5\) and
\[ R_I := F_I + N_I. \]

We then have \( \mathcal{P}_D(x) \in \{ \mathcal{P}_{S_I}(x) \mid I \in I_D \} \), i.e., \( \mathcal{P}_D \) is a piecewise affine function. The affine pieces of \( \mathcal{P}_D \) are the projections on the corresponding affine subspaces \( S_I \) (cf. §6.2a). In fact, for each \( x \in R_I \) we have \( \mathcal{P}_D(x) = \mathcal{P}_{S_I}(x) \), each \( R_I \) is full dimensional and \( \mathbb{R}^n = \bigcup_{I \in I_D} R_I \) [66, Prop.s 2.4.4 and 2.4.5]. For each \( I \in I_D \) let
\[ P_I := \nabla \mathcal{P}_{S_I} = I - \begin{pmatrix} A \end{pmatrix}^\top \begin{pmatrix} A & C_i \end{pmatrix}, \]
and for each \( x \in \mathbb{R}^n \)
\[ I_D(x) := \{ I \in I_D \mid x \in R_I \}. \]

Then,
\[ \partial_C(\mathcal{P}_D)(x) = \text{conv} \partial_B(\mathcal{P}_D)(x) = \text{conv} \{ P_I \mid I \in I_D(x) \}. \]

Therefore, an element of \( \partial_B \mathcal{P}_D(x) \) is \( P_I \) as in (6.9) where \( I = \{ i \mid C_i \bar{x} = d_i \} \) is the set of active constraints of \( \bar{x} = \mathcal{P}_D(x) \). For a more general analysis we refer the reader to [27, 39].

\(^5\) Consistently with the definition in [66], the polyhedron \( P \) can equivalently be expressed by means of only inequalities as \( P = \{ x \in \mathbb{R}^n \mid Ax \leq b, \ -Ax \leq -b, \ Cx \leq b \} \), resulting indeed in \( \text{cone}[A^\top, -A^\top, C^\top] = \text{range} A^\top + \text{cone} C^\top \).
c. **Halfspaces.** $H = \{ x \in \mathbb{R}^n \mid \langle a, x \rangle \leq b \}$ for some $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$.

Then, denoting the positive part of $r \in \mathbb{R}$ as $[r]_+ := \max \{0, r\}$,

$$P_H(x) = x - \frac{\langle a, x \rangle - b}{\|a\|^2}a$$

and

$$\partial_C(P_H)(x) = \begin{cases} 
\{I - \|a\|^2aa^T\} & \text{if } x \notin H, \\
\{I\} & \text{if } \langle a, x \rangle < b, \\
\text{conv} \{I, I - \|a\|^2aa^T\} & \text{if } \langle a, x \rangle = b.
\end{cases}$$

d. **Boxes.** $D = \{ x \in \mathbb{R}^n \mid \ell \leq x \leq u \}$ for some $\ell, u \in [-\infty, \infty]^n$.

We have

$$P_D(x) = \min \{\max \{x, \ell\}, u\},$$

and since the corresponding indicator function $\delta_D$ is separable, every element of $\partial_C(P_D)(x)$ is diagonal (cf. §6.1a)

$$\partial_C(P_D)(x)_i = \begin{cases} 
\{0\} & \text{if } x \notin D, \\
\{1\} & \text{if } \ell_i < x_i < u_i, \\
\{0, 1\} & \text{if } x_i \in \{\ell_i, u_i\}.
\end{cases}$$

e. **Unit simplex.** $D = \{ x \in \mathbb{R}^n \mid x \geq 0, \sum_{i=1}^n x_i = 1 \}$.

By writing down the optimality conditions for the corresponding projection problem, one can easily see that

$$P_D(x) = [x - \lambda 1]_+,$$

where $\lambda$ solves $\langle 1, [x - \lambda 1]_+ \rangle = 1$. Since the unit simplex is a polyhedral set, we are dealing with a special case of §6.2b, where $A = 1^T$, $b = 1$, $C = -1$ and $d = 0$. Therefore, in order to calculate an element of the generalized Jacobian of the projection, we first compute $P_D(x)$ and then determine the set of active indices $J := \{ i \mid P_D(x)_i = 0 \}$. An element $P \in \partial_B(P_D)(x)$ is given by

$$P_{ij} = \begin{cases} 
\delta_{i,j} - \frac{1}{n-1} & \text{if } i, j \notin J, \\
0 & \text{otherwise},
\end{cases}$$

where $|J|$ denotes the cardinality of the set $J$. Notice that $P$ is block diagonal after a permutation of rows and columns.

f. **Euclidean unit ball.** $B = \overline{B}(0; 1)$.

We have

$$P_B(x) = \begin{cases} 
x & \text{if } x \in B, \\
\frac{x}{\|x\|} & \text{otherwise},
\end{cases}$$

and
\( \partial_C(\mathbb{P}_B)(x) = \begin{cases} 
\{1\} & \text{if } ||x|| < 1, \\
\text{conv} \begin{bmatrix} ||x||^{-1}(I - ww^T) \end{bmatrix} & \text{if } ||x|| = 1, \\
\{||x||^{-1}(I - ww^T)\} & \text{if } x \notin B, 
\end{cases} \)

where \( w := \frac{x}{||x||} \).

**g. Second-order cone.** \( \mathcal{K} = \{(x_0, \bar{x}) \in \mathbb{R} \times \mathbb{R}^{n-1} \mid x_0 \geq ||\bar{x}||\} \).

Let \( x := (x_0, \bar{x}) \), and for \( w \in \mathbb{R}^n \) and \( \alpha \in \mathbb{R} \) define

\[
M_{w, \alpha} := \frac{1}{2} \begin{bmatrix} w \top (I - \alpha) I_{n-1} + \alpha w w^\top \end{bmatrix}.
\]

Then, \( \partial_{\mathcal{K}}(\mathbb{P}_B)(x) = \text{conv}(\partial_{\mathcal{K}}(\mathbb{P}_B)(x)) \) where, for \( \bar{w} := \frac{x}{||x||} \) and \( \bar{\alpha} := -\frac{x}{||x||} \), we have [30, Lem. 2.6]

\[
\partial_{\mathcal{K}}(\mathbb{P}_B)(x) = \begin{cases} 
\{0\} & \text{if } x_0 < -||\bar{x}||, \\
\{I_n\} & \text{if } x_0 > ||\bar{x}||, \\
\{M_{\bar{w}, \bar{\alpha}}\} & \text{if } -||\bar{x}|| < x_0 < ||\bar{x}||, \\
\{I_n, M_{\bar{w}, \bar{\alpha}}\} & \text{if } x_0 = ||\bar{x}|| 
eq 0, \\
\{0, M_{\bar{w}, \bar{\alpha}}\} & \text{if } x_0 = -||\bar{x}|| 
eq 0, \\
\{0, I_n\} \cup \{M_{\bar{w}, \bar{\alpha}} \mid ||\bar{x}|| \leq 1, ||w|| \leq 1\} & \text{if } x_0 = \bar{x} = 0.
\end{cases}
\]

**h. Positive semidefinite cone.** \( \mathcal{S}_+ = \mathcal{S}_+(\mathbb{R}^{n \times n}) \).

For any symmetric matrix \( M \) it holds that

\[
\mathbb{P}_{\mathcal{S}_+}(M) = Q[\text{diag}(\lambda)], Q^\top,
\]

where \( M = Q \text{diag}(\lambda) Q^\top \) is any spectral decomposition of \( M \). This coincides with (6.3), as \( \delta_{\mathcal{S}_+} \) can be expressed as in (6.4), where \( h \) has the separable form (6.2) with \( g = \delta_{\mathbb{R}_+} \), so that for \( r \in \mathbb{R} \) we have

\[
\text{prox}_{g}(r) = [r]_+ \quad \text{and} \quad \partial_B(\text{prox}_g(r)) = \begin{cases} 
\{0\} & \text{if } r < 0, \\
\{0, 1\} & \text{if } r = 0, \\
\{1\} & \text{if } r > 0.
\end{cases}
\]

An element of \( \partial_B \mathbb{P}_{\mathcal{S}_+}(\mathbb{R}^{n \times n}) \) is thus given by (6.4).

### 6.3 Norms

**a. \( \ell_1 \text{ norm.} \)** \( g(x) = ||x||_1 \).

The proximal mapping is the well known soft-thresholding operator

\[
(\text{prox}_{g_y}(x))_i = \text{sgn}(x_i)(|x_i| - y)_+, \quad i = 1, \ldots, n.
\]

Function \( g \) is separable, and thus every element of \( \partial_B(\text{prox}_{g_y}) \) is a diagonal matrix, cf. §6.1a. Specifically, the nonzero elements are
1 \text{ On the acceleration of forward-backward splitting via an inexact Newton method} 39

\[ \partial C(\text{prox}_\gamma g)(x)_i = \begin{cases} \{1\} & \text{if } |x_i| > \gamma, \\ \{0, 1\} & \text{if } |x_i| = \gamma, \\ \{0\} & \text{if } |x_i| < \gamma. \end{cases} \]

We could also arrive to the same conclusion by applying the Moreau decomposition of §6.1b to the function of §6.2d with \( u = -\ell = 1_n \), since the \( \ell_1 \) norm is the conjugate of the indicator of the \( \ell_\infty \)-norm ball.

b. \( \ell_\infty \) norm. \( g(x) = \|x\|_\infty \).

Function \( g \) is the convex conjugate of the indicator of the unit simplex \( D \) analyzed in §6.2e. From the Moreau decomposition, see §6.1b, we obtain

\[ \partial C(\text{prox}_\gamma g)(x) = I - \partial C(\mathcal{P}_D)(\gamma). \]

Then, \( \mathcal{P}_D(\gamma) = [\gamma - \lambda 1]\_+ \) where \( \lambda \in \mathbb{R} \) solves \( \langle 1, [\gamma - \lambda 1]\_+ \rangle = 1 \). Let \( J = \{i \mid \mathcal{P}_D(\gamma)_i = 0\} \), then an element of \( \partial g(\text{prox}_\gamma g)(x) \) is given by

\[ P_{ij} = \begin{cases} \frac{1}{\gamma - |x_j|} & \text{if } i, j \not\in J, \\ \delta_{i,j} & \text{otherwise.} \end{cases} \]

c. Euclidean norm. \( g(x) = \|x\| \).

The proximal mapping is given by

\[ \text{prox}_\gamma g(x) = \begin{cases} (1 - \gamma \|x\|^{-1})x & \text{if } \|x\| \geq \gamma, \\ 0 & \text{otherwise.} \end{cases} \quad (6.10) \]

Since \( \text{prox}_\gamma g \) is a \( PC^1 \) mapping, its B-subdifferential can be computed by simply computing the Jacobians of its smooth pieces. Specifically, denoting \( w = \gamma \|x\|^{-1} \) we have

\[ \partial C(\text{prox}_\gamma g)(x) = \begin{cases} \{1 - \gamma \|x\|^{-1}(1 - \ell w)\} & \text{if } \|x\| > \gamma, \\ \{0\} & \text{if } \|x\| < \gamma, \\ \text{conv} \{1 - \gamma \|x\|^{-1}(1 - w^\top, 0)\} & \text{otherwise.} \end{cases} \]

d. Sum of Euclidean norms. \( g(x) = \sum_{s \in S} \|x_s\| \), where \( S \) is a partition of \( \{1, \ldots, n\} \).

Differently from the \( \ell_1 \)-norm which induces sparsity on the whole vector, this function serves as regularizer to induce group sparsity [81]. For \( s \in S \), the components of the proximal mapping indexed by \( s \) are

\[ (\text{prox}_\gamma g(x))_s = (1 - \gamma \|x_s\|^{-1})_+ x_s. \]

Any \( P \in \partial_B(\text{prox}_\gamma g)(x) \) is block diagonal with the \( s \)-block equal to

\[ P_s = \begin{cases} \{1 - \gamma \|x_s\|^{-1}(1 - \|x_s\|^{-2} x_s x_s^\top)\} & \text{if } \|x_s\| > \gamma, \\ 1 & \text{if } \|x_s\| < \gamma, \\ \text{any of these two matrices} & \text{if } \|x_s\| = \gamma. \end{cases} \]
e. **Matrix nuclear norm.** $G(X) = \|X\|_\star$ for $X \in \mathbb{R}^{m \times n}$.

The nuclear norm returns the sum of the singular values of a matrix $X \in \mathbb{R}^{m \times n}$, i.e., $G(X) = \sum_{i=1}^{m} \sigma_i(X)$ (for simplicity we are assuming that $m \leq n$). It serves as a convex surrogate for the rank, and has found many applications in systems and control theory, including system identification and model reduction [21, 20, 22, 41, 61]. Other fields of application include matrix completion problems arising in machine learning [68, 62] and computer vision [76, 52], and nonnegative matrix factorization problems arising in data mining [18].

The nuclear norm can be expressed as $G(X) = h(\sigma(X))$, where $h(x) = \|x\|_1$ is absolutely symmetric and separable. Specifically, it takes the form (6.2) with $g = |\cdot|$, for which $g(0) = 0$ and $0 \in \partial g(0)$, and whose proximal mapping is the soft-thresholding operator. In fact, since the case of interest here is $x \geq 0$ (because $\sigma_i(X) \geq 0$), we have $\text{prox}_{\gamma g}(x) = [x - \gamma]_+$, cf. (6.10). Consequently, the proximal mapping of $\|X\|_\star$ is given by (6.6) with

$$
\Sigma_g(X) = \text{diag}([\sigma_1(X) - \gamma]_+, \ldots, [\sigma_m(X) - \gamma]_+).
$$

For $x \in \mathbb{R}_+$ we have that

$$
\partial_C(\text{prox}_{\gamma g})(x) = \begin{cases} 
0 & \text{if } 0 \leq x < \gamma, \\
[0, 1] & \text{if } x = \gamma, \\
1 & \text{if } x > \gamma,
\end{cases}
$$

(6.11)

then $\partial_B(\text{prox}_{\gamma g})(X)$ takes the form as in (6.7).

7 Conclusions

A forward-backward truncated-Newton method (FBTN) is proposed, that minimizes the sum of two convex functions one of which Lipschitz continuous and twice continuously differentiable. Our approach is based on the forward-backward envelope (FBE), a continuously differentiable tight lower bound to the original (nonsmooth and extended-real valued) cost function sharing minima and minimizers. The method requires forward-backward steps, Hessian evaluations of the smooth function and Clarke Jacobians of the proximal map of the nonsmooth term. Explicit formulas of Clarke Jacobians of a wide variety of useful nonsmooth functions are collected from the literature for the reader’s convenience. The higher-order operations are needed for the computation of symmetric and positive semidefinite matrices that serve as surrogate for the Hessian of the FBE, allowing for a generalized (regularized, truncated-) Newton method for its minimization. The algorithm exhibits global $Q$-linear convergence under an error bound condition, and $Q$-superlinear or even $Q$-quadratic if an additional semismoothness assumption at the limit point is satisfied.
Appendix A. Auxiliary results

Lemma A.1. Any proper lsc convex function with nonempty and bounded set of minimizers is level bounded.

Proof. Let \( h \) be such function; to avoid trivialities we assume that \( \text{dom } h \) is unbounded. Fix \( x_* \in \arg \min h \) and let \( R > 0 \) be such that \( \arg \min h \subseteq B := B(x_*; R) \). Since \( \text{dom } h \) is closed, convex and unbounded, it holds that \( h \) attains a minimum on the compact set \( \text{bdry } B \), be it \( m \), which is strictly larger than \( h(x_*) \) (since \( \text{dist}(\arg \min h, \text{bdry } B) > 0 \) due to compactness of \( \arg \min h \) and openness of \( B \)). For \( x \notin B \), let \( s_x = x_* + R \frac{x_* - x}{\|x_* - x\|} \) denote its projection onto \( \text{bdry } B \), and let \( t_x := \frac{\|x_* - x\|}{R} \geq 1 \). Then,

\[
h(x) = h(x_*) + t_x(h(s_x) - h(x_*) + h(s_x) - h(x)) = h(x_*) + t_x(m - h(x_*))
\]

where in the first inequality we used the fact that \( t_x \geq 1 \). Since \( m - h(x_*) > 0 \) and \( t_x \to \infty \) as \( \|x\| \to \infty \), we conclude that \( h \) is coercive, and thus level bounded. \( \square \)

Lemma A.2. Let \( H \in \mathcal{S}_+(\mathbb{R}^n) \) with \( \lambda_{\max}(H) \leq 1 \). Then \( H - H^2 \in \mathcal{S}_+(\mathbb{R}^n) \) with

\[
\lambda_{\min}(H - H^2) = \min \{ \lambda_{\min}(H)(1 - \lambda_{\min}(H)), \lambda_{\max}(H)(1 - \lambda_{\max}(H)) \}.
\]

Proof. Consider the spectral decomposition \( H = S^T DS \) for some orthogonal matrix \( S \) and diagonal \( D \). Then, \( H - H^2 = S^T DS - D^2 \). Apparently, \( D \) is diagonal, hence the eigenvalues of \( H - H^2 \) are exactly \( \{ \lambda - \lambda^2 \mid \lambda \in \text{eigs}(H) \} \). The function \( \lambda \mapsto \lambda - \lambda^2 \) is concave, hence the minimum in \( \text{eigs}(H) \) is attained at one extremum, that is, either at \( \lambda = \lambda_{\min}(H) \) or \( \lambda = \lambda_{\max}(H) \), which proves the claim. \( \square \)

Lemma A.3. For any \( \gamma \in (0, 2/L_i) \) the forward-backward operator \( T_\gamma \) (3.6) is nonexpansive (in fact, \( 2/4\gamma L_i \)-averaged), and the residual \( R_\gamma \) is Lipschitz continuous with modulus \( 2/(4\gamma L_i) \).

Proof. By combining \([2, \text{Prop. 4.39 and Cor. 18.17}]\) it follows that the gradient descent operator \( x \mapsto x - \gamma \nabla f(x) \) is \( \gamma L_i/2 \)-averaged. Moreover, since the proximal mapping is \( 1/2 \)-averaged [2, Prop. 12.28] we conclude from [2, Prop. 4.44] that the forward-backward operator \( T_\gamma \) is \( \alpha \)-averaged with \( \alpha = \frac{2}{4\gamma L_i} \), thus nonexpansive [2, Rem. 4.34(i)]. Therefore, by definition of \( \alpha \)-averagedness there exists a \( 1 \)-Lipschitz continuous operator \( S \) such that \( T_\gamma = (1 - \alpha) \text{id} + \alpha S \) and consequently the residual \( R_\gamma = \frac{1}{\gamma}(\text{id} - T_\gamma) = \frac{\alpha}{\gamma}(\text{id} - S) \) is \( (2\gamma L_i) \)-Lipschitz continuous. \( \square \)
References

1. Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Mathematical Programming*, 137(1):91–129, Feb 2013.
2. Heinz H. Bauschke and Patrick L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics. Springer, 2017.
3. Heinz H. Bauschke, Dominikus Noll, and Hung M. Phan. Linear and strong convergence of algorithms involving averaged nonexpansive operators. *Journal of Mathematical Analysis and Applications*, 421(1):1–20, 2015.
4. Amir Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.
5. Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.
6. Stephen Becker and Jalal Fadili. A quasi-Newton proximal splitting method. In *Advances in Neural Information Processing Systems*, pages 2618–2626, 2012.
7. Dimitri P. Bertsekas. Constrained optimization and lagrange multiplier methods. *Computer Science and Applied Mathematics, Boston: Academic Press*, 1982, 1982.
8. Dimitri P. Bertsekas. *Convex Optimization Algorithms*. Athena Scientific, 2015.
9. Rajendra Bhatia. *Matrix Analysis*. Graduate Texts in Mathematics. Springer New York, 1997.
10. Jacek Bochnak, Michel Coste, and Marie-Francoise Roy. *Real Algebraic Geometry*. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics. Springer Berlin Heidelberg, 2013.
11. Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. Tame functions are semismooth. *Mathematical Programming*, 117(1):5–19, Mar 2009.
12. Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.
13. Xiaojun Chen and Masao Fukushima. Proximal quasi-Newton methods for nondifferentiable convex optimization. *Mathematical Programming*, 85(2):313–334, Jun 1999.
14. Xin Chen, Houduo Qi, and Paul Tseng. Analysis of nonsmooth symmetric-matrix-valued functions with applications to semidefinite complementarity problems. *SIAM Journal on Optimization*, 13(4):960–985, 2003.
15. Frank H. Clarke. *Optimization and Nonsmooth Analysis*. Society for Industrial and Applied Mathematics, 1990.
16. Patrick L. Combettes and Jean-Christophe Pesquet. *Proximal Splitting Methods in Signal Processing*, pages 185–212. Springer New York, New York, NY, 2011.
17. Dmitriy Drusvyatskiy and Adrian S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 2018.
18. Lars Eldén. *Matrix Methods in Data Mining and Pattern Recognition*. Society for Industrial and Applied Mathematics, 2007.
19. Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems, volume II*. Springer, 2003.
20. Maryam Fazel. *Matrix rank minimization with applications*. PhD thesis, Stanford University, 2002.
21. Maryam Fazel, Haitham Hindi, and Stephen P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference*, volume 6, pages 4734–4739, 2001.
22. Maryam Fazel, Haitham Hindi, and Stephen P. Boyd. Rank minimization and applications in system theory. In *Proceedings of the 2004 American Control Conference*, volume 4, pages 3273–3278 vol.4, June 2004.
23. Masao Fukushima. Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems. *Mathematical Programming*, 53(1):99–110, Jan 1992.
On the acceleration of forward-backward splitting via an inexact Newton method

24. Pontus Giselsson and Mattias Fält. Envelope functions: Unifications and further properties. *Journal of Optimization Theory and Applications*, Jun 2018.

25. M. Seetharama Gowda. Inverse and implicit function theorems for H-differentiable and semismooth functions. *Optimization Methods and Software*, 19(5):443–461, 2004.

26. Osman Güler. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992.

27. Jiye Han and Defeng Sun. Newton and quasi-Newton methods for normal maps with polyhedral sets. *Journal of Optimization Theory and Applications*, 94(3):659–676, Sep 1997.

28. Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Grundlehren Text Editions. Springer Berlin Heidelberg, 2004.

29. Adrian S. Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1):173–183, 1995.

30. Adrian S. Lewis and Hristo S. Sendov. Twice differentiable spectral functions. *SIAM Journal on Matrix Analysis and Applications*, 23(2):368–386, 2001.

31. Guanghui Lan, Zhaosong Lu, and Renato D. C. Monteiro. Primal-dual first-order methods with $O(1/\varepsilon)$ iteration-complexity for cone programming. *Mathematical Programming*, 126(1):1–29, Jan 2011.

32. Jason D. Lee, Yuekai Sun, and Michael Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.

33. Roger A. Horn, Roger A. Horn, and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1994.

34. Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, Mar 1993.
48. Fanwen Meng, Gongyun Zhao, Mark Goh, and Robert De Souza. Lagrangian-dual functions and Moreau-Yosida regularization. *SIAM Journal on Optimization*, 19(1):39–61, 2008.

49. Robert Mifflin. Semismooth and semiconvex functions in constrained optimization. *SIAM Journal on Control and Optimization*, 15(6):959–972, 1977.

50. Robert Mifflin, Liqun Qi, and Defeng Sun. Properties of the Moreau-Yosida regularization of a piecewise C² convex function. *Mathematical Programming*, 84(2):269–281, Feb 1999.

51. Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

52. Toshihiko Morita and Takeo Kanade. A sequential factorization method for recovering shape and motion from image streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(8):858–867, Aug 1997.

53. Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2003.

54. Yurii Nesterov. *Gradient methods for minimizing composite functions*. *Mathematical Programming*, 140(1):125–161, Aug 2013.

55. Jong-Shi Pang. Error bounds in mathematical programming. *Mathematical Programming*, 79(1):299–332, Oct 1997.

56. Neal Parikh and Stephen Boyd. Proximal algorithms. *Found. Trends Optim.*, 1(3):127–239, January 2014.

57. Panagiotis Patrinos and Alberto Bemporad. Proximal Newton methods for convex composite optimization. In *IEEE Conference on Decision and Control*, pages 2358–2363, 2013.

58. Panagiotis Patrinos, Pantelis Sopasakis, and Haralambos Sarimveis. A global piecewise smooth Newton method for fast large-scale model predictive control. *Automatica*, 47(9):2016–2022, 2011.

59. Panagiotis Patrinos, Lorenzo Stella, and Alberto Bemporad. Forward-backward truncated Newton methods for convex composite optimization. *ArXiv e-prints*, feb 2014.

60. Liqun Qi and Jie Sun. A nonsmooth version of Newton’s method. *Mathematical Programming*, 58(1):353–367, Jan 1993.

61. Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

62. Jasson D. M. Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML ’05, pages 713–719, New York, NY, USA, 2005. ACM.

63. R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

64. R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2011.

65. RT Rockafellar. Convex analysis. 1970.

66. Stefan Scholtes. *Piecewise Differentiable Functions*, pages 91–111. Springer New York, New York, NY, 2012.

67. Pantelis Sopasakis, Nikolaos Freris, and Panagiotis Patrinos. Accelerated reconstruction of a compressively sampled data stream. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 1078–1082, Aug 2016.

68. Nathan Srebro. *Learning with Matrix Factorizations*. PhD thesis, Cambridge, MA, USA, 2004.

69. Lorenzo Stella, Andreas Themelis, and Panagiotis Patrinos. Forward-backward quasi-Newton methods for nonsmooth optimization problems. *Computational Optimization and Applications*, 67(3):443–487, Jul 2017.

70. Lorenzo Stella, Andreas Themelis, and Panagiotis Patrinos. Newton-type alternating minimization algorithm for convex optimization. *IEEE Transactions on Automatic Control*, 2018.

71. Lorenzo Stella, Andreas Themelis, Pantelis Sopasakis, and Panagiotis Patrinos. A simple and efficient algorithm for nonlinear model predictive control. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 1939–1944, Dec 2017.
72. Defeng Sun, Masao Fukushima, and Liqun Qi. A computable generalized Hessian of the D-gap function and Newton-type methods for variational inequality problems. *Complementarity and Variational Problems: State of the Art, MC Ferris and JS Pang (eds.), SIAM, Philadelphia, PA*, pages 452–472, 1997.

73. Defeng Sun and Jie Sun. Semismooth matrix-valued functions. *Mathematics of Operations Research, 27*(1):150–169, 2002.

74. Andreas Themelis and Panagiotis Patrinos. Douglas-Rachford splitting and ADMM for non-convex optimization: tight convergence results. *ArXiv e-prints*, Sep 2017.

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

76. Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision, 9*(2):137–154, Nov 1992.

77. Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. Technical report, 2008.

78. Michael Ulbrich. *Optimization Methods in Banach Spaces*, pages 97–156. Springer Netherlands, Dordrecht, 2009.

79. Nobuo Yamashita, Kouichi Taji, and Masao Fukushima. Unconstrained optimization reformulations of variational inequality problems. *Journal of Optimization Theory and Applications, 92*(3):439–456, Mar 1997.

80. Zhe Yang. A study on nonsymmetric matrix-valued functions. Master’s thesis, National University of Singapore, 2009.

81. Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 68*(1):49–67, 2006.

82. Guanglu Zhou and Liqun Qi. On the convergence of an inexact Newton-type method. *Oper. Res. Lett., 34*(6):647–652, November 2006.

83. Guanglu Zhou and Kim-Chuan Toh. Superlinear convergence of a Newton-type algorithm for monotone equations. *Journal of Optimization Theory and Applications, 125*(1):205–221, 2005.