ACTIVITY IDENTIFICATION AND LOCAL LINEAR CONVERGENCE OF FORWARD–BACKWARD-TYPE METHODS

Jingwei Liang, Jalal M. Fadili, Gabriel Peyré

To cite this version:

Jingwei Liang, Jalal M. Fadili, Gabriel Peyré. ACTIVITY IDENTIFICATION AND LOCAL LINEAR CONVERGENCE OF FORWARD–BACKWARD-TYPE METHODS. SIAM Journal on Optimization, Society for Industrial and Applied Mathematics, 2017. hal-01658850

HAL Id: hal-01658850
https://hal.archives-ouvertes.fr/hal-01658850
Submitted on 7 Dec 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract. In this paper, we consider a class of Forward–Backward (FB) splitting methods that includes several variants (e.g. inertial schemes, FISTA) for minimizing the sum of two proper convex and lower semi-continuous functions, one of which has a Lipschitz continuous gradient, and the other is partly smooth relative to a smooth active manifold \( \mathcal{M} \). We propose a unified framework, under which we show that, this class of FB-type algorithms (i) correctly identifies the active manifold in a finite number of iterations (finite activity identification), and (ii) then enters a local linear convergence regime, which we characterize precisely in terms of the structure of the underlying active manifold. We also establish and explain why FISTA (with convergent sequences) locally oscillates and can be locally slower than FB. These results may have numerous applications including in signal/image processing, sparse recovery and machine learning. Indeed, the obtained results explain the typical behaviour that has been observed numerically for many problems in these fields such as the Lasso, the group Lasso and the nuclear norm minimization to name only a few.

Key words. Forward–Backward, Inertial Methods, ISTA/FISTA, Partial Smoothness, Local Linear Convergence.

AMS subject classifications. 49J52, 65K05, 65K10, 90C25, 90C31.

1. Introduction.

1.1. Non-smooth optimization. In various fields of science and engineering, such as signal/image processing, inverse problems and machine learning, many problems can be cast as solving a structured composite non-smooth optimization problem of the sum of two functions, which usually reads

\[
\text{(P}_{\text{opt}}) \quad \min_{x \in \mathbb{R}^n} \Phi(x) \overset{\text{def}}{=} F(x) + R(x),
\]

where

(H.1) \( R \in \Gamma_0(\mathbb{R}^n) \), the set of proper convex and lower semi-continuous functions on \( \mathbb{R}^n \),

(H.2) \( F \in C^{1,1}(\mathbb{R}^n) \), and the gradient \( \nabla F \) is \((1/\beta)\)-Lipschitz continuous.

(H.3) \( \text{Argmin}(\Phi) \neq \emptyset \), i.e. the set of minimizers is non-empty.

From now on, we suppose that assumptions (H.1)-(H.3) hold. Problem (P_{opt}) is closely related to finding solutions of the monotone inclusion problem

\[
\text{(P}_{\text{inc}}) \quad \text{Find } x \in \mathbb{R}^n \text{ such that } 0 \in A(x) + B(x),
\]

where

(H.4) \( A : \mathbb{R}^n \rightrightarrows \mathbb{R}^n \) is a set-valued maximal monotone operator (see (A.1)).

(H.5) \( B : \mathbb{R}^n \rightarrow \mathbb{R}^n \) is maximal monotone and \( \beta \)-cocoercive (see (A.2)).

(H.6) \( \text{zer}(A + B) \neq \emptyset \), i.e. the set of zeros of \( A + B \) is non-empty.

For problem (P_{opt}), given a global minimizer \( x^* \in \text{Argmin}(\Phi) \), then the corresponding first-order optimality condition reads

\[ 0 \in \partial R(x^*) + \nabla F(x^*), \]

where \( \partial R \) denotes the sub-differential of \( R \) at \( x^* \). Clearly, if we let \( A = \partial R \) and \( B = \nabla F \), then (P_{opt}) is simply a special case of (P_{inc}).

This work has been partly supported by the European Research Council (ERC project SIGMA-Vision). JF was partly supported by Institut Universitaire de France.

†Normandie Univ, ENSICAEN, CNRS, GREYC (Jingwei.Liang, Jalal.Fadili@ensicaen.fr).

‡CNRS, DMA, Ecole Normale Supérieure (Gabriel.Peyre@ens.fr).
In this paper, our main focus is the non-smooth optimization problem \( (\mathcal{P}_{\text{opt}}) \). Though some of our results are also valid for the monotone inclusion problem \( (\mathcal{P}_{\text{inc}}) \), in particular Algorithm 1 and its global convergence analysis, see Section 2.

### 1.2. Forward–Backward-type splitting methods

The Forward–Backward (FB) splitting method [38] is a powerful tool for solving optimization problems \( (\mathcal{P}_{\text{opt}}) \) with the additively separable and “smooth + non-smooth” structure. The standard (non-relaxed) version of FB implements the iterative scheme

\[
(1.1) \quad x_{k+1} = \text{prox}_{\gamma_k R}(x_k - \gamma_k \nabla F(x_k)), \quad \gamma_k \in [\xi, 2\beta - \bar{\epsilon}],
\]

where \( \xi, \bar{\epsilon} > 0 \), and \( \text{prox}_{\gamma R} \) denotes the proximity operator of \( R \) which is defined as

\[
\text{prox}_{\gamma R}(\cdot) \text{ def } \min_{x \in \mathbb{R}^n} \frac{1}{2}\|x - \cdot\|^2 + \gamma R(x).
\]

Global convergence of the sequence \( (x_k)_{k \in \mathbb{N}} \) generated by the FB method is well-established in the literature, based on the property that the composed operator \( \text{prox}_{\gamma R}(\text{Id} - \gamma \nabla F) \) is so-called averaged non-expansive [12]. Moreover, sub-linear \( O(1/k) \) convergence rate of the sequence of objective values of FB is also well-known, e.g. [45, 16, 14].

*Inertial schemes and FISTA.* In the literature, different variants of FB method were studied, and a popular trend is the inertial schemes which aim at speeding up the convergence properties of FB. In [49], a two-step algorithm called the “heavy-ball with friction” method is studied for solving \( (\mathcal{P}_{\text{opt}}) \) with \( R = 0 \). It can be seen as an explicit discretization of a nonlinear second-order dynamical system (oscillator with viscous damping). This dynamical approach to iterative methods in optimization has motivated increasing attention in recent years. For instance, in real Hilbert spaces, it is used in [4] for solving \( (\mathcal{P}_{\text{opt}}) \) with \( F = 0 \) and [5] for solving \( (\mathcal{P}_{\text{inc}}) \) with \( B = 0 \) yielding an inertial PPA method. The authors in [42, 8, 39] propose different inertial versions of the FB method for solving \( (\mathcal{P}_{\text{opt}}) \) and/or \( (\mathcal{P}_{\text{inc}}) \).

On the other hand, in the context of convex optimization, the accelerated FISTA method was proposed in [14], based on the seminal work [43], which achieves \( O(1/k^2) \) convergence rate for the sequence of objective functions. However, while iterates generated by the FB are convergent, the convergence of FISTA iterates has been an open problem until recently. This question was first settled in [18], then followed by [9] in the continuous dynamical system case. More precisely, for \( \gamma_k \in [0, \beta] \) and a sequence of inertial parameter that converges at an appropriate rate (i.e. in Algorithm 1, set \( a_k = b_k = \frac{1}{k+q}, q > 2 \)), these authors established (weak in infinite-dimensional Hilbert spaces) convergence of the iterates sequence while maintaining the \( O(1/k^2) \) rate on the objective values. The rate is actually even \( o(1/k^2) \) as proved in [7].

---

**Algorithm 1 A General Inertial Forward–Backward splitting**

**Initial:** \( \bar{a} \leq a \leq b \leq 1, \xi, \bar{\epsilon} > 0 \) such that \( \xi \leq 2\beta - \bar{\epsilon} \). \( x_0 \in \mathbb{R}^n, x_{-1} = x_0 \).

Let \( a_k \in [0, \bar{a}], b_k \in [0, \bar{b}], \gamma_k \in [\xi, 2\beta - \bar{\epsilon}] \). Repeat

\[
(1.2) \quad y_{a,k} = x_k + a_k(x_k - x_{k-1}), \quad y_{b,k} = x_k + b_k(x_k - x_{k-1}),
\]

\[
(1.3) \quad x_{k+1} = \text{prox}_{\gamma_k R}(y_{a,k} - \gamma_k \nabla F(y_{b,k})).
\]

---

In this paper, we propose a general inertial Forward–Backward splitting method (iFB), see Algorithm 1. Based on the choice of the inertial parameters \( a_k \) and \( b_k \), the proposed method recovers the following special cases:

- \( a_k = 0, b_k = 0 \): this is the original FB method [38];
- \( a_k \in [0, \bar{a}], b_k = 0 \): this is the case studied in [42] for \( (\mathcal{P}_{\text{inc}}) \). In the context of optimization with \( R = 0 \), one recovers the heavy ball method with friction...
1.3. Contributions. The study of (local) linear convergence of FB-type methods in the absence of strong convexity has attracted increasing interest in recent years, see the related work below for details. In general, most of the existing work focuses on some special cases (e.g. $R = \| \cdot \|_1$ in $(P_{\text{opt}})$), and the proofs of the results heavily rely on the specific structure of the function $R$, which makes them rather difficult to extend to other cases. Therefore, it is important to present a unified analysis framework, and possibly with stronger claims. This is one of the main motivations of this work. To be more precise, this paper delivers the following contributions:

A general class of inertial algorithms. We present a unified iFB splitting class of algorithms for solving $(P_{\text{opt}})$. It can be viewed as a versatile explicit-implicit discretization of a nonlinear second-order dynamical system with viscous damping, and thus covers existing methods as special cases. We establish global convergence of the iterates, and also stability to errors.

Finite activity identification. Under the additional assumption that function $R$ is partly smooth at $x^* \in \text{Argmin}(\Phi)$ relative to a $C^2$-smooth manifold $\mathcal{M}_{x^*}$ (see Definition 5) and a non-degeneracy condition at $x^*$, we show that any FB-type method to solve $(P_{\text{opt}})$ has the finite time activity identification property. Meaning that, after a finite number of iterations, say $K$, the iterates $x_k \to x^*$ built by the FB-type method belong to $\mathcal{M}_{x^*}$ for all $k \geq K$.

Local linear convergence. Exploiting this identification property, we then show that the FB-type methods, locally along the manifold $\mathcal{M}_{x^*}$, exhibit a linear convergence regime. We characterize this regime and the corresponding rates precisely depending on the structure of the active manifold $\mathcal{M}_{x^*}$. For instance, we provide sharp estimates for the convergence rate. For the sequence convergent FISTA method, we draw two major conclusions:

- Locally, FISTA can be slower than the FB method (e.g. see Figure 1).
- We provide an explanation of the local oscillatory behaviour of FISTA and provide the exact oscillation period (e.g. see Figure 2).

This gives an enlightening explanation of the usefulness of the so-called restarting method to locally accelerate the convergence of FISTA used by many authors, for instance in sparse recovery [25, 46, 24]: the algorithm is restarted after a certain
number of iterations (set more or less empirically), where the inertial sequence $a_k = b_k$
is reset to 0.

We also discuss some practical acceleration procedures. Indeed, once finite iden-
tification happens, the globally non-smooth convex problem ($P_{opt}$) becomes (locally)
equivalent to a $C^2$-smooth one along the (possibly non-convex) active manifold $M_x$.
In turn, this opens the door to acceleration, especially using higher order methods
such as Newton or non-linear conjugate gradient, see Section 4.5 and Figure 2.

### 1.4. Related work.
Finite support identification and local linear convergence
of FB for solving a special instance of ($P_{opt}$) where $R$ is the $\ell_1$-norm is established
in [16, 26]. The same question has been recently addressed for FISTA under some
constraints on the inertial parameter in [54, 32]. [3] proved local linear convergence
of FB to solve ($P_{opt}$) for $R$ being a so-called convex decomposable regularizer. Local
linear convergence of FB is studied in [31] for $R$ the nuclear norm and $F$ locally
strongly convex. All these previous functions are subclass of partly smooth functions,
and their results are thus covered by ours under weaker assumptions. The proposed
work is also a deeper and sharper extension of our previous results on FB [37]. Finite
identification of active manifolds associated to partly smooth functions has been shown
in [28, 29, 27] for the (sub)gradient projection method, Newton-like methods, the
proximal point algorithm and the algorithm in [55]. Their work extends that of e.g.
[58] on identifiable surfaces (see references therein for related work of Dunn, and Burke
and Moré). However, in all these works, the local linear convergence behaviour was
not addressed.

### 1.5. Notations.
Throughout the paper, $\text{Id}$ denotes the identity operator on $\mathbb{R}^n$.
For a nonempty convex set $\Omega \subset \mathbb{R}^n$, $\text{ri}(\Omega)$ and $\text{rbd}(\Omega)$ denote its relative interior
and boundary respectively, $\text{aff}(\Omega)$ is its affine hull, and $\text{par}(\Omega) = \mathbb{R}(\Omega - \Omega)$ is the
subspace parallel to it. Denote $\iota_\Omega$ the indicator function of $\Omega$, $\sigma_\Omega$ its support function
and $P_\Omega$ the orthogonal projector onto $\Omega$. For a matrix $M$, $\text{ker}(M)$ is its null-space.
The subdifferential of a function $R \in \Gamma_0(\mathbb{R}^n)$ is the set-valued operator $\partial R : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, $x \mapsto \{ u \in \mathbb{R}^n | R(z) \geq R(x) + \langle u, z - x \rangle, \forall z \in \mathbb{R}^n \}$.

**Paper organization.** The rest of the paper is organized as follows. Global con-
vergence of the proposed iFB method is presented in Section 2. Then in Section 3,
we introduce the concept of partial smoothness, and prove the finite activity identi-
ﬁcation property of the FB-type methods. We then turn to local linear convergence
analysis in Section 4. Some numerical results are reported in Section 5.

### 2. Global convergence of the inertial Forward–Backward.
In this section,
we establish the global convergence of the iterates provided by the iFB method with
possible errors. We will state our results (Theorem 3 and 4) for the finite dimension-
al optimization problem ($P_{opt}$). In fact, our global convergence results can handle the
more general monotone inclusion problem ($P_{inc}$) in an inﬁnite dimensional real Hilbert
space, where weak convergence of the iterates sequence can be obtained. The proofs
given in Section A are written for this general setting.

We consider the case where $\partial R(x)$ and $\nabla F(x)$ are computed approximately. Tow-
ward this goal, we recall the notion of $\varepsilon$-enlargement.

**Definition 2 ($\varepsilon$-enlargement).** Let $A : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a set-valued maximal mono-
tone operator, $\varepsilon \geq 0$. Then the $\varepsilon$-enlargement of $A$ is defined as,

$$A^\varepsilon(x) \overset{\text{def}}{=} \{ v \in \mathbb{R}^n | \langle u - v, y - x \rangle \geq -\varepsilon, \forall y \in \mathbb{R}^n, u \in A(y) \}.$$ 

Denote $\partial^\varepsilon R$ the $\varepsilon$-enlargement of $\partial R$. We now consider an inexact form of the
ifB algorithm where step (1.3) is replaced by finding \( x_{k+1} \) such that

\[(2.1) \quad y_{a,k} - \gamma_k(\nabla F(y_{b,k}) + \xi_k) - x_{k+1} \in \gamma_k \partial^{\varepsilon} R(x_{k+1}), \]

where \( \xi_k \in \mathbb{R}^n \) is the error in the evaluation of the gradient operator \( \nabla F \). Observe that since the \( \varepsilon \)-approximate subdifferential of a proper closed convex function is contained in the \( \varepsilon \)-enlargement of its sub-differential [17], our setting also handles the case of approximate sub-differentials.

**Theorem 3 (Conditional convergence).** Consider Algorithm 1 with the inexact iteration (2.1). Suppose that \( a < 1 \), \( \sum_{k \in \mathbb{N}} \varepsilon_k < +\infty \) and \( \sum_{k \in \mathbb{N}} \| \xi_k \| < +\infty \). Then the generated sequence \( (x_k)_{k \in \mathbb{N}} \) is bounded. If moreover \( (a_k)_{k \in \mathbb{N}} \) and \( (b_k)_{k \in \mathbb{N}} \) are such that

\[(2.2) \quad \sum_{k \in \mathbb{N}} \max\{a_k, b_k\} \|x_k - x_{k-1}\|^2 < +\infty, \]

then, there exists \( x^* \in \text{Argmin}(\Phi) \) such that the sequence \( (x_k)_{k \in \mathbb{N}} \) converges to \( x^* \).

The proof of Theorem 3 is given in Section A. This result generalizes that of [42] who considered the case \( b_k \equiv 0 \) and \( \xi_k \equiv 0 \). In [10] the inexact sequence convergent FISTA with the same errors as ours was studied, i.e. \( \gamma_k \in [0, \beta] \), \( a_k = b_k = \frac{k-1}{k\beta q} \), \( q > 2 \).

The terminology “conditional convergence” used in Theorem 3 refers to the fact that for the convergence to occur, the sequences \( (a_k)_{k \in \mathbb{N}} \) and \( (b_k)_{k \in \mathbb{N}} \) can be chosen depending (conditionally) on \( (x_k)_{k \in \mathbb{N}} \) in such a way that (2.2) holds. This can be enforced easily by a simple online updating rule such as, given \( a \in [0, 1] \), \( b \in [0, 1] \),

\[(2.3) \quad a_k = \min\{a, c_{a,k}\}, \quad b_k = \min\{b, c_{b,k}\}, \]

where \( c_{a,k}, c_{b,k} > 0 \), and \( \max\{c_{a,k}, c_{b,k}\} \|x_k - x_{k-1}\|^2 \) is summable. For instance, one can choose \( c_{a,k} = \frac{\epsilon_k}{k^{1+\epsilon} \|x_k - x_{k-1}\|^2}, \quad c_{b,k} \geq 0, \beta > 0 \) and similarly for \( c_{b,k} \).

One can also devise choices of \( (a_k)_{k \in \mathbb{N}} \) and \( (b_k)_{k \in \mathbb{N}} \) that are independent of \( (x_k)_{k \in \mathbb{N}} \), and still guarantee global convergence. We dub this unconditional convergence. The following result generalizes those in [5, 42, 39].

**Theorem 4 (Unconditional convergence).** Consider Algorithm 1 with the inexact iteration (2.1). Assume that there exists a constant \( \tau > 0 \) such that one of the following holds,

\[(2.4) \quad \begin{cases} 
(1 + a_k) - \frac{2\tau}{\sqrt{\gamma}} (1 + b_k)^2 > \gamma \tau : a_k < \frac{\gamma}{2\sqrt{\beta}} b_k, \\
(1 - 3a_k) - \frac{\gamma}{\sqrt{\beta}} (1 - b_k)^2 > \gamma \tau : b_k \leq a_k \text{ or } \frac{\gamma}{2\sqrt{\beta}} b_k \leq a_k < b_k,
\end{cases} \]

and, moreover \( \sum_{k \in \mathbb{N}} \varepsilon_k < +\infty \) and \( \sum_{k \in \mathbb{N}} \| \xi_k \| < +\infty \). Then \( \sum_{k \in \mathbb{N}} \|x_k - x_{k-1}\|^2 < +\infty \), and there exists \( x^* \in \text{Argmin}(\Phi) \) such that the sequence \( (x_k)_{k \in \mathbb{N}} \) converges to \( x^* \).

See Section A for the proof.

### 3. Partial smoothness and finite time activity identification.

#### 3.1. Partial smoothness.** From now on, besides assumption (H.1), we assume that \( R \) in (Popt) is moreover partly smooth relative to a smooth manifold. The notion of partial smoothness is first introduced in [35]. This concept, as well as that of identifiable surfaces [58], captures the essential features of the geometry of non-smoothness which are along the so-called active/identifiable manifold. For convex functions, a closely related idea is developed in [34]. Loosely speaking, a partly smooth function behaves smoothly as we move on the identifiable submanifold, and sharply if we move normal to the manifold. In fact, the behaviour of the function and of its minimizers depend essentially on its restriction to this manifold, hence offering a powerful framework for algorithmic and sensitivity analysis theory.
Let $\mathcal{M}_x$ be a $C^2$-smooth embedded submanifold of $\mathbb{R}^n$ around a point $x$. To lighten terminology, henceforth we shall state $C^2$-manifold instead of $C^2$-smooth embedded submanifold of $\mathbb{R}^n$. The natural embedding of a submanifold $\mathcal{M}_x$ into $\mathbb{R}^n$ permits to define a Riemannian structure on $\mathcal{M}_x$, and we simply say $\mathcal{M}_x$ is a Riemannian manifold. $\mathcal{T}_{\mathcal{M}_x}(x')$ denotes the tangent space to $\mathcal{M}_x$ at any point $x'$ near $x$ in $\mathcal{M}_x$. More materials on manifolds are given in Section B.1.

We are now ready to state formally the class of partly smooth functions through its regularity properties.

**Definition 5 (Partly smooth function).** Let $R \in \Gamma_0(\mathbb{R}^n)$, $R$ is said to be partly smooth at $x$ relative to a set $\mathcal{M}_x$ containing $x$ if $\partial R(x) \neq \emptyset$, and moreover

(i) **Smoothness:** $\mathcal{M}_x$ is a $C^2$-manifold around $x$, $R$ restricted to $\mathcal{M}_x$ is $C^2$ near $x$;

(ii) **Sharpness:** The tangent space $\mathcal{T}_{\mathcal{M}_x}(x)$ coincides with $T_x \equiv \text{par}(\partial R(x))^\perp$;

(iii) **Continuity:** The set-valued mapping $\partial R$ is continuous at $x$ relative to $\mathcal{M}_x$.

The class of partly smooth functions at $x$ relative to $\mathcal{M}_x$ is denoted as $\text{PSF}_x(\mathcal{M}_x)$.

One can easily show that a function in $\Gamma_0(\mathbb{R}^n)$ which is locally polyhedral around $x$ is partly smooth at $x$ relative to $x + T_x$. Polyhedrality also implies that the sub-differential is locally constant around $x$ along $x + T_x$. Capitalizing on the results of [35], it can be shown that under mild transversality conditions, the set of proper lsc convex and partly smooth functions is closed under addition and pre-composition by a linear operator. Moreover, absolutely permutation-invariant convex and partly smooth functions of the singular values of a real matrix, i.e. spectral functions, are convex and partly smooth spectral functions of the matrix [22]. Many examples of partly smooth functions that are popular in signal processing, machine learning and statistics can be found in [57], see also Section 5.

FACT 6 (Local normal sharpness). If $R \in \text{PSF}_x(\mathcal{M}_x)$, then all $x' \in \mathcal{M}_x$ near $x$ satisfy $\mathcal{T}_{\mathcal{M}_x}(x') = T_x$. In particular, when $\mathcal{M}_x$ is affine or linear, then $T_x = T_x$.

We now give expressions of the Riemannian gradient and Hessian (see Section B.1 for definitions) for the case of partly smooth functions relative to a set $\mathcal{M}_x$. Many examples of partly smooth functions that are popular in signal processing, machine learning and statistics can be found in [57], see also Section 5.

FACT 7. If $R \in \text{PSF}_x(\mathcal{M}_x)$, then for any $x' \in \mathcal{M}_x$ near $x$

$$
\nabla_{\mathcal{M}_x} R(x') = P_{T_x'}(\nabla R(x')),
$$

and this does not depend on the smooth representation of $R$ on $\mathcal{M}_x$. In turn, for all $h \in T_{x'}$

$$
\nabla^2_{\mathcal{M}_x} G(x') h = P_{T_{x'}} \nabla^2 \tilde{R}(x') h + \mathcal{W}_x(h, P_{T_{x'}} \nabla \tilde{R}(x')),
$$

where $\tilde{R}$ is a smooth extension (representative) of $R$ on $\mathcal{M}_x$, and $\mathcal{W}_x(\cdot, \cdot) : T_x \times T_x^\perp \rightarrow T_x$ is the Weingarten map of $\mathcal{M}_x$ at $x$ (see Section B.1 for definitions).

### 3.2. Finite time activity identification

In this section, we state our result establishing that FB-type methods have the finite activity identification property.

**Theorem 8 (Finite activity identification).** Suppose that an FB-type method is used to create a sequence $(x_k)_{k \in \mathbb{N}}$ that converges to $x^* \in \text{Argmin}(\Phi)$ such that $R \in \text{PSF}_{x^*}(\mathcal{M}_{x^*})$, and moreover the non-degeneracy condition

$$
(\text{ND}) - \nabla F(x^*) \in \text{ri}(\partial R(x^*))
$$

holds. Then, there exists a large enough $K > 0$ such that for all $k \geq K$, $x_k \in \mathcal{M}_{x^*}$.

If moreover,
Since \( \liminf_{k \to \infty} x_k = x^* \) we get that \( x_0, x_1, \ldots, x_k, \ldots \) converges to some \( x^* \in \text{Argmin}(\Phi) \), and the latter is non-empty by assumption (H.3).

Now (1.3) is equivalent to

\[ y_{a,k} - \gamma_k \nabla F(y_{b,k}) - x_{k+1} \in \gamma_k \partial R(x_{k+1}). \]

By (H.2), we get

\[
\text{dist}\left(-\nabla F(x^*), \partial R(x_{k+1})\right) \\
\leq \frac{1}{\gamma_k} (y_{a,k} - x_{k+1} - \nabla F(y_{b,k}) + \nabla F(x^*)) \\
\leq \frac{1}{\gamma_k} (a_k \|x_k - x_{k-1}\| + \|x_{k+1} - x_k\|) + \|\nabla F(y_{b,k}) - \nabla F(x^*)\| \\
\leq \left( \frac{1}{\gamma_k} + \frac{1}{\beta} \right) \|x_k - x_{k-1}\| + \frac{1}{\gamma_k} \|x_{k+1} - x_k\| + \frac{1}{\beta} \|x_k - x^*\|.
\]

Since \( \liminf_{k \to \infty} \gamma_k = \xi > 0 \) and \( x_k \) converges to \( x^* \), we obtain \( \text{dist}\left(-\nabla F(x^*), \partial R(x_k)\right) \to 0 \). Owing to assumption (H.1), \( R \) is subdifferentially continuous at every point in its domain, and in particular at \( x^* \) for \( -\nabla F(x^*) \), which in turn entails \( R(x_k) \to R(x^*) \). Altogether, this shows that the conditions of [28, Theorem 5.3] are fulfilled on \( \langle \nabla F(x^*), \cdot \rangle + R \), and the result follows.

(i) When the active manifold \( M_{x^*} \) is an affine subspace, then \( M_{x^*} = x^* + T_{x^*} \), owing to the normal sharpness property and the claim follows immediately;

(ii) When \( R \) is locally polyhedral around \( x^* \), then \( M_{x^*} \) is an affine subspace and the identification of \( y_{a,k}, y_{b,k} \) follows from (i). For the rest, it is sufficient to observe that by polyhedrality, for any \( x \in M_{x^*} \) near \( x^* \), \( \partial R(x) = \partial R(x^*) \). Therefore, combining Fact 6 and Fact 7, we get the second conclusion. \( \square \)

A bound on the identification iteration. In Theorem 8, we have not provided an estimate \( K \geq 0 \) beyond which finite identification occurs. There is of course a situation where the answer is trivial, \textit{i.e.} \( R \) is the indicator function of an affine subspace. However, knowing \( K \) has practical interest, for instance, if one wants to switch to higher order acceleration (see Section 4.5). It is then legitimate to wonder whether such an estimate of \( K \) can be given. In the following, we shall give a bound in some important cases. For the sake of simplicity, we state the result for the case of FB (\textit{i.e.} \( a_k = b_k \equiv 0 \) in Algorithm 1). A similar reasoning can be easily generalized to the case of any converging FB-type method.
PROPOSITION 10. Suppose that the assumptions of Theorem 8 hold. Then the following holds.

(i) If the iterates are such that $\partial R(x_k) \subset \text{rbd}(\partial R(x^*))$ whenever $x_k \notin M_{x^*}$, then $x_k \in M_{x^*}$ for all $k \geq \frac{\|x_0 - x^*\|^2}{\epsilon^2 \text{dist}(-\nabla F(x^*), \text{rbd}(\partial R(x^*)))}$;

(ii) If $R$ is separable, i.e. $R(x) = \sum_{i=1}^{m} \sigma_{C_i}(b_i)$, where $\forall 1 \leq i \leq m, b_i \subset \{1, \ldots, n\}$, $\bigcup_{i=1}^{m} b_i = \{1, \ldots, n\}$, and $b_i \cap b_j = \emptyset, \forall i \neq j$, and $\dim(C_i) = |b_i|$, then identification of $M_{x^*}$ occurs for some $k$ larger than \(\frac{\|x_0 - x^*\|^2}{\epsilon^2 \sum_{i \in I^*_x} \text{dist}(-\nabla F(x^*)_b, \text{rbd}(C_i))}^2\),

where $I_x \triangleq \{i : x_i \neq 0\}$.

Proof. (i) By firm non-expansiveness of $\text{prox}_{\gamma_k - 1} R$, and non-expansiveness of $\text{Id} - \gamma_k - 1 \nabla F$, we have

\[
\|x_k - x^*\|^2 \leq \|(\text{Id} - \gamma_k - 1 \nabla F)(x_k) - (\text{Id} - \gamma_k - 1 \nabla F)(x^*)\|^2 \\
- \|x_k - 1 - \gamma_k - 1 \nabla F(x_k) - x_k + \gamma_k - 1 \nabla F(x^*)\|^2 \\
\leq \|x_k - 1 - x^*\|^2 - \epsilon^2 \|x_k - \nabla F(x^*)\|^2,
\]

where we denoted $u_k \triangleq (x_k - x_k)/\gamma_k - 1 - \nabla F(x_k)$. By definition, we have $u_k \in \partial R(x_k)$. Suppose that identification has not occurred at $k$, i.e. that $x_k \notin M_{x^*}$, and hence $u_k \in \partial R(x_k) \subset \text{rbd}(\partial R(x^*))$. Therefore, continuing the above inequality, we get

\[
\|x_k - x^*\|^2 \leq \|x_k - 1 - x^*\|^2 - \epsilon^2 \text{dist}(-\nabla F(x^*), \partial R(x_k))^2 \\
\leq \|x_k - 1 - x^*\|^2 - \epsilon^2 \text{dist}(-\nabla F(x^*), \text{rbd}(\partial R(x^*)))^2 \\
\leq \|x_0 - x^*\|^2 - \epsilon^2 \text{dist}(-\nabla F(x^*), \text{rbd}(\partial R(x^*)))^2,
\]

and $\text{dist}(-\nabla F(x^*), \text{rbd}(\partial R(x^*))) > 0$ owing to (ND). Taking $k$ as the largest integer such that the right hand is positive, we deduce that the number of iterations where identification has not occurred, does not exceed the given bound, whence our conclusion follows.

(ii) We have $\partial \sigma_{C_i}(x^*_k) = C_i, \forall i \in I^*_x$. In turn, by separability, $R$ is partly smooth at $x^*$ relative to $M_{x^*} = \times_{i=1}^{m} M_{x^*_i}$, where $M_{x^*_i} = \emptyset$ if $i \in I^*_x$, and $M_{x^*_i} \neq \emptyset$ otherwise. Suppose that at iteration $k$, $I^*_x \cap I_{x_k} \neq \emptyset$. Denote $h_{k-1} = x_{k-1} - \gamma_{k-1} \nabla F(x_{k-1})$, and $h^*_k = x^* - \gamma_{k-1} \nabla F(x^*)$. Thus for any $i \in I^*_x \cap I_{x_k}$, we have

\[
x_{k,b_i} - x^*_k = h_{k-1,b_i} - P_{\gamma_{k-1} C_i}(h_{k-1,b_i}) \\
= (h_{k-1,b_i} - h^*_b) - (P_{\gamma_{k-1} C_i}(h_{k-1,b_i}) - P_{\gamma_{k-1} C_i}(h^*_i))
\]

where we used Moreau identity in the first equality. Since $i \in I_{x_k} \cap I^*_x$, we have $h_{k-1,b_i} \notin \gamma_{k-1} C_i$ and $h^*_b \in \gamma_{k-1} C_i$, or equivalently, that $P_{\gamma_{k-1} C_i}(h_{k-1,b_i}) \in \gamma_{k-1} \text{rbd}(C_i) = \gamma_{k-1} \text{rbd}(\partial \sigma_{C_i}(x^*_k))$ and $P_{\gamma_{k-1} C_i}(h^*_i) = h^*_i$. Combining this with the fact that the orthogonal projector on $\gamma_{k-1} C_i$ is firmly non-expansive, we get

\[
x_{k,b_i} - x^*_k \leq \|h_{k-1,b_i} - h^*_b\|^2 - \|P_{\gamma_{k-1} C_i}(h_{k-1,b_i}) - h^*_i\|^2 \\
= \|h_{k-1,b_i} - h^*_b\|^2 - \|P_{\gamma_{k-1} C_i}(h_{k-1,b_i}) + \gamma_{k-1} \nabla F(x^*)_b\|^2 \\
\leq \|h_{k-1,b_i} - h^*_b\|^2 - 2 \gamma_{k-1} \text{dist}(-\nabla F(x^*)_b, \text{rbd}(C_i))^2 \\
\leq \|h_{k-1,b_i} - h^*_b\|^2 - \epsilon^2 \text{dist}(-\nabla F(x^*)_b, \text{rbd}(C_i))^2.
\]

This manuscript is for review purposes only.
This bound together with non-expansiveness of prox$_{\gamma_{k-1}C_i}$ and Id $-$ $\gamma_{k-1} \nabla F$ yield
\[
\|x_k - x^*\|^2 = \sum_{i \in I_{x_k}^*} \|x_{k,b_i} - x_b^*\|^2 + \sum_{j \in I_{x_k}} \|x_{k,b_j} - x_j^*\|^2
\]
\[
\leq \|h_{k-1} - h^*\|^2 - \xi \sum_{i \in I_{x_k}^*} \text{dist}(-\nabla F(x^*_b), \partial \text{bd}(C_i))^2
\]
\[
\leq \|x_{k-1} - x^*\|^2 - \xi \sum_{i \in I_{x_k}^*} \text{dist}(-\nabla F(x^*_b), \partial \text{bd}(C_i))^2
\]
\[
\leq \|x_0 - x^*\|^2 - h_k \xi \sum_{i \in I_{x}^*} \text{dist}(-\nabla F(x^*_b), \partial \text{bd}(C_i))^2,
\]
where the last term in the right hand side is strictly positive by (ND). Taking $k$ as the largest integer such that the right hand side is positive, we deduce that the number of iterations where $I_{x_k}^* \cap I_{x_k} \neq \emptyset$ does not exceed the given bound. We then conclude that beyond this bound, there is no $i$ such that $\mathcal{M}_{x_k,b_i} \neq 0$ while $\mathcal{M}_{x_k} = 0$. The proof is complete.}

Note that, as intuitively expected, this bound increases as the non-degeneracy condition (ND) becomes more stringent. However, as it depends on $x^*$, it is only of theoretical interest. In the separable case, observe that $\sum_{i \in I_{x_k}^*} \text{dist}(-\nabla F(x^*_b), \partial \text{bd}(C_i))^2 = \text{dist}(-\nabla F(x^*), \partial \text{bd}(C)^2)$ when $\partial C_i$ is differentiable at $x^*_b$ for all $i \in I_{x_k}^*$. The case of the $\ell_1$-norm considered in [26] is recovered in the second situation of Proposition 10 with $C_i \equiv [-\lambda, \lambda]$ for some $\lambda > 0$.

### 3.3. Stability to errors.
Consider the inexact version (2.1) with $\varepsilon_k \equiv 0$. Assume that $(x_k)_{k \in \mathbb{N}}$ is such that $(x_k)_{k \in \mathbb{N}}$ converges to some $x^* \in \text{Argmin}(\Phi)$ (see typically the summability conditions in Theorem 3(i)-(ii)). Then, since $\xi_k \to 0$, it can be easily seen from the proof of Theorem 8 that the activity identification property holds true for the above inexact iteration.

However, one cannot afford in general having non-zero errors $\varepsilon_k$ in the implicit step as in (2.1), even summable. The deep reason behind this is that in the exact case, under condition (ND), the proximal mappings of $R$ and $R + \iota_{\mathcal{M}}$, locally agree nearby $x^*$. This property is clearly violated if approximate proximal mappings are involved. Here is a simple example.

**Example 11.** Let $F : x \in \mathbb{R} \to \frac{1}{2}|x|^2$, with $\delta \in [-1,1]$, and $R : x \in \mathbb{R} \mapsto |x|$. $\Phi \in \Gamma_0(\mathbb{R})$ and has a unique minimizer $x^* = \text{prox}_{1/|\delta|}(\delta) = 0$. Moreover, $\Phi$ is partly smooth at $x^*$ relative to $\mathcal{M}_x = \{0\}$, and $\delta - x^* = \delta \in \text{ri}(\partial R(x^*)) = [-1,1]$. Consider the inexact version of the FB algorithm
\[
x_{k+1} \in (\text{Id} + \partial^F|\cdot|^{-1})(\delta),
\]
where we set $\gamma_k \equiv 1$, since $\nabla F$ is 1-Lipschitz. From [17, Example 5.2.5], we have
\[
\partial^F|\cdot|(x) = \begin{cases} 
[1 - \varepsilon/x, 1] & \text{if } x > \varepsilon/2 \\
[-1,1] & \text{if } |x| \leq \varepsilon/2 \\
[-1, -1 - \varepsilon/x] & \text{if } x < -\varepsilon/2,
\end{cases}
\]
whence the graph of $(\text{Id} + \partial^F|\cdot|^{-1})$, a set-valued operator, can be easily deduced. Thus, depending on $\varepsilon_k$ and the choice made in the inclusion (3.1), $x_k$ may never vanish, i.e. $x_k \notin \mathcal{M}_{x^*}$, for any finite $k$.

### 4. Local linear convergence of FB-type methods.

We are now in position to present the local linear convergence result for FB-type methods, and all the proofs in this section are collected in Section B. Throughout this section, $x^*$ is a global minimizer of problem (P$_{opt}$) to which the sequence $(x_k)_{k \in \mathbb{N}}$ provided by the FB-type method converges. $\mathcal{M}_x$ is the partial smoothness manifold of $R$ at $x^*$, and $T_{x^*}$ the corresponding tangent space.
Restricted injectivity. In addition to (H.2), in the rest of the paper, we also assume that \( F \) is locally \( C^2 \) around \( x^* \), and its Hessian fulfills the following restricted injectivity condition,

\[
\ker(\nabla^2 F(x^*)) \cap T_{x^*} = \{0\}.
\]

Local continuity of the Hessian of \( F \) then implies that there exist \( \alpha \geq 0 \) and \( \epsilon > 0 \), such that \( \forall h \in T_{x^*}, \)

\[
(4.1) \quad \langle h, \nabla^2 F(x)h \rangle > \alpha \|h\|^2, \forall x \in B_{\epsilon}(x^*) = \{x \in \mathbb{R}^n : \|x - x^*\| \leq \epsilon\}.
\]

It turns out that under conditions (ND) and (RI), one can show that problem \( (\mathcal{P}_{opt}) \) admits a unique minimizer, and local quadratic growth of \( \Phi \) if \( R \) is moreover partly smooth. Recall that a function \( \Phi \) grows quadratically locally around \( x^* \) if \( \exists c > 0 \) such that \( \Phi(x) \geq \Phi(x^*) + c\|x - x^*\|^2 \), \( \forall x \) near \( x^* \).

**Proposition 12** (Uniqueness of the minimizer). Under the assumptions (H.1)-(H.3), let \( x^* \in \text{Argmin}(\Phi) \) be a global minimizer of \( (\mathcal{P}_{opt}) \) such that \( F \) is locally \( C^2 \) around \( x^* \). If conditions (ND) and (RI) are also fulfilled, then

(i) \( x^* \) is the unique minimizer of \( (\mathcal{P}_{opt}) \).

(ii) If moreover \( R \in \text{PSF}_x(\mathcal{M}_{x^*}) \), then \( \Phi \) has at least a quadratic growth near \( x^* \).

**4.1. Locally linearized iteration.** Define the following matrices which are all symmetric,

\[
(4.2) \quad H \overset{\text{def}}{=} \gamma P_{T_{x^*}}, \nabla^2 F(x^*) P_{T_{x^*}}, \ G \overset{\text{def}}{=} \text{Id} - H, \ U \overset{\text{def}}{=} \gamma \nabla^2 \Phi(x^*) P_{T_{x^*}} - H,
\]

where \( \nabla^2 \Phi(x^*) \) is the Riemannian Hessian of \( \Phi \) on the manifold \( \mathcal{M}_{x^*} \) (see Fact 7).

**Lemma 13.** For problem \( (\mathcal{P}_{opt}) \), let (H.1)-(H.3) hold and \( x^* \in \text{Argmin}(\Phi) \) such that \( R \in \text{PSF}_x(\mathcal{M}_{x^*}) \) and \( F \) is locally \( C^2 \) around \( x^* \). Then \( U \) is symmetric positive semi-definite under either of the following circumstances:

(i) \( (\text{ND}) \) holds.

(ii) \( \mathcal{M}_{x^*} \) is an affine subspace.

In turn, \( \text{Id} + U \) is invertible, and \( W \overset{\text{def}}{=} (\text{Id} + U)^{-1} \) is symmetric positive definite with eigenvalues in \([0, 1]\).

The following simple lemma gathers important properties of the matrices in (4.2).

**Lemma 14.** For the matrices in (4.2) and \( W \),

(i) Under (H.2) and (RI),

(a) \( H \) is symmetric positive definite with eigenvalues in \([\gamma \alpha, \frac{\gamma}{\beta}]\).

(b) For \( \gamma \in [\xi, 2\beta - \tau], \xi \) and \( \tau > 0 \), \( G \) has eigenvalues in \([-1 + \frac{\tau}{\beta}, 1 - \alpha \xi] \subset \{-1, 1\} \).

(c) For \( \gamma \in [\xi, \beta], G \) is also symmetric positive semi-definite with eigenvalues in \([0, 1 - \alpha \xi] \subset [0, 1]\).

(ii) If both the assumptions of Lemma 13 and (i) hold, then \( W \) has real eigenvalues lying in \([-1, 1]\). If moreover \( \gamma \in [\xi, \beta], \) then \( W \) has eigenvalues lying in \([0, 1]\).

Let \( a \in [0, \bar{a}], b \in [0, \bar{b}], \gamma \in [\xi, 2\beta - \tau], \) define \( r_k \overset{\text{def}}{=} x_k - x^*, d_k \overset{\text{def}}{=} (r_k r_k^{-1}), \) and matrix

\[
(4.3) \quad M \overset{\text{def}}{=} \begin{bmatrix}
(a - b)W + (1 + b)WG & -(a - b)W - bWG \\
\text{Id} & 0
\end{bmatrix}.
\]

Our interest in the vector \( d_k \) is inspired by the convergence rate analysis of the heavy ball method [50, Section 3.2]. We now show that once the active manifold is identified, FB-type iteration locally linearizes.
Moreover, we have 
\begin{equation}
\eta = \text{eigenvector of } M \quad \text{and} \quad W \eta = \text{Id}
\end{equation} 

We pose that an FB-type method is used to create a sequence \( (x_k)_{k \in \mathbb{N}} \) that converges to \( x^* \in \text{Argmin}(\Phi) \) such that (ND) and (RI) hold. If moreover, 
\begin{equation}
\rho \left( \begin{array}{c}
a_k \\ b_k
\end{array} \right) \rightarrow a \in [0, 1], \quad b \rightarrow b \in [0, 1], \quad \gamma_k \rightarrow \gamma \in [\epsilon, 2\beta - \tau],
\end{equation}
then for \( k \) large enough, we have 
\begin{equation}
d_{k+1} = Md_k + o(||d_k||).
\end{equation}
The \( o(\cdot) \) term disappears when \( R \) is locally polyhedral around \( x^* \) and \((\gamma_k, a_k, b_k)\) are chosen constant.

**Remark 16.**

(i) Condition (4.4) asserts that both the inertial parameters \((a_k, b_k)\) and the step-size \(\gamma_k\) should converge to some limit points, and cannot be relaxed in general.

(ii) For the FB method (i.e. \( a_k = b_k \equiv 0 \)), (4.3) can be further simplified, and the corresponding linearized iteration can be given in terms of \( r_k \) directly,
\begin{equation}
r_{k+1} = WGr_k + o(||r_k||).
\end{equation}

(iii) Proposition 15 also covers the sequence convergent FISTA method \([18, 9]\), i.e. \( a_k = b_k = \frac{k-1}{k+q}, q > 2 \) and \( \gamma_k \in [0, \beta] \). In this case, we have indeed \( a_k \rightarrow a = b = 1 \).

4.2. **Spectral properties of** \( M \). Our aim now is to establish local linear convergence of FB-type schemes. For this, given the structure of the locally linearized iteration (4.5), it is sufficient to strictly upper-bound by 1 the spectral radius of \( M \), and conclude using standard arguments. This is what we are about to do.

The rationale is to start by relating explicitly the eigenvalues of \( M \) to those of \( G \) or \( W \), and then use Lemma 14 to upper-bound the spectral radius of \( M \). However, given the structure of \( M \), this is a challenging linear algebra problem, and can only be done for some cases: \( a \) and \( b \) possibly different but the the function \( R \) is locally polyhedral, or \( R \) is a general partly smooth function but \( a = b \). These situations are not restrictive at all and cover all interesting applications we have in mind.

Let \( \eta \) and \( \sigma \) be an eigenvalue of \( W \) and \( M \) respectively. We denote \( \eta, \bar{\eta} \) the smallest and largest (signed) eigenvalues of \( W \), and \( \rho(M) \) the spectral radius of \( M \).

**Locally polyhedral case.** When \( R \) is locally polyhedral around \( x^* \), \( U \) vanishes and \( W = \text{Id} \), and \( M \) in (4.3) simplifies.

**Proposition 17.** Suppose that \( R \) is locally polyhedral around \( x^* \). If \( \left( \begin{array}{c}T_1 \\ T_2 \end{array} \right) \) is an eigenvector of \( M \) corresponding to an eigenvalue \( \sigma \), then it must satisfy \( r_1 = \sigma r_2 \). Moreover, we have

(i) \( r_2 \) is an eigenvector of \( G \) associated to an eigenvalue \( \eta \), where \( \eta \) and \( \sigma \) satisfy the relation
\begin{equation}
\sigma^2 - ((a - b) + (1 + b)\eta)\sigma + (a - b) + b\eta = 0.
\end{equation}

(ii) Given any \( (a, b) \in [0, 1]^2 \), then \( \rho(M) < 1 \) if, and only if, 
\begin{equation}
2(b - a) - 1) / (1 + 2b) < \eta.
\end{equation}

**Remark 18.** It can be shown that, given \( a \) and \( b \), \( \rho(M) \) is determined only by \( \eta \) and \( \bar{\eta} \). These extreme eigenvalues lie in \( ]-1, 1[ \) \((\gamma \in [0, 2\beta]) \) or even in \([0, 1[ \) \((\gamma \in [0, \beta]) \)
by Lemma 14(i)(b)(c).

**General partly smooth case.** When \( R \) is a general partly smooth function, then \( U \) is nontrivial, and the spectral analysis of (4.3) becomes a generalized eigenvalue problem which is much more complex. Therefore, we assume \( b = a \). We have the following corollary of Proposition 17.
Corollary 19. Let \( b = a \). If \( \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} \) be an eigenvector of \( M \) corresponding to an eigenvalue \( \sigma \), then it must satisfy \( r_1 = \sigma r_2 \). Moreover \( r_2 \) is an eigenvector of \( G \) related to eigenvalue \( \eta \), where \( \eta \) and \( \sigma \) satisfy the relation
\[
\sigma^2 - (1 + a)\eta \sigma + a\eta = 0,
\]
and \( \rho(M) < 1 \) if, and only if,
\[
1/(1 + 2a) < \eta.
\]

Remark 20. Condition (4.10) holds naturally for \( \gamma \in [0, \beta] \), since by Lemma 14(ii),
for such \( \gamma \), \( \eta \geq 0 \).

4.3. Local linear convergence of FB-type methods. We start with the case
where \( \mathcal{R} \) is locally polyhedral around \( x^* \).

Theorem 21. Suppose (H.1)-(H.3) hold, and an FB-type method generates a
sequence \( x_k \to x^* \in \text{Argmin}(\Phi) \) such that \( \mathcal{R} \) is locally polyhedral around \( x^* \), \( F \) is \( C^2 \)
near \( x^* \), and conditions (ND), (RI) are satisfied. If moreover (4.4) and (4.8) hold,
then \( (x_k)_{k \in \mathbb{N}} \) converges locally linearly to \( x^* \). More precisely, given any \( \rho \in [\rho(M), 1] \),
there exists \( K > 0 \) and a constant \( C > 0 \), such that for all \( k \geq K \), there holds
\[
\|x_k - x^*\| \leq C\rho^{k-K}\|x_K - x^*\|.
\]

Proof. Combining Proposition 15, Proposition 17 and [50, Section 2.1.2, Theorem 1],
leads to the claimed result.

Remark 22. \( \rho(M) \) is the optimal rate. Indeed, when \( a_k \equiv a, b_k \equiv b \) and \( \gamma_k \equiv \gamma \),
the \( o(\cdot) \) term vanishes in (4.5) and thus, \( \rho = \rho(M) \).

Let’s turn to the case \( \mathcal{R} \) is a general partly smooth function, but \( b = a \in [0, \bar{a}] \).

Theorem 23. Suppose assumptions (H.1)-(H.3) hold, and the FB-type methods
generate a sequence \( x_k \to x^* \in \text{Argmin}(\Phi) \) such that \( \mathcal{R} \in \text{PSF}_{x_0}(\mathcal{M}_{x^*}) \), \( F \) is \( C^2 \)
near \( x^* \), and conditions (ND), (RI) are satisfied. If moreover (4.4) holds with \( b = a \), and
(4.10) is satisfied, then \( (x_k)_{k \in \mathbb{N}} \) converges locally linearly to \( x^* \). More precisely, given
any \( \rho \in [\rho(M), 1] \), there exists \( K > 0 \) and a constant \( C > 0 \), such that for all \( k \geq K \),
there holds
\[
\|x_k - x^*\| \leq C\rho^{k-K}\|x_K - x^*\|.
\]

Proof. This follows by combining Proposition 15, Corollary 19 and [50, Section 2.1.2, Theorem 1].

Remark 24.
(i) The limit \( b = a \) in (4.4) does not mean that we should set \( b_k = a_k, \forall k \in \mathbb{N} \)
along the iterations.
(ii) In contrast to our previous work [37], which addresses the case of FB method,
the rate estimates that we provide here are much sharper in general, and
both estimates only coincide when \( \mathcal{R} \) is locally polyhedral (see the numerical
experiments for more details). The main reasons underlying this is that, here,
our rate estimates relies on the locally linearized iteration in Proposition 15 and
the spectral properties of \( M \), which takes into account the geometry of the
identified submanifold (its curvature for instance). This is not the case in our
former work.
(iii) The obtained results can be readily extended to the variable metric FB splitting
method [21], where a rate under an appropriate metric can be obtained.
However for the sake of brevity, we do not pursue this further.
(iv) In our proof of local linear convergence, convexity does play a crucial role. For instance, it was only needed to show that the matrix $U$ is positive semi-definite. This suggests that our local linear convergence claims can be extended to the non-convex case, provided that the Riemannian Hessian of $R$ is assumed positive semi-definite at $x^*$. In addition, to guarantee finite identification in the non-convex setting, we need global convergence of $iFB$ to a critical point, which can be ensured if for instance $\Phi$ satisfies the (non-smooth) Kurdyka-Lojasiewicz inequality [15]. This will be left to a forthcoming paper.

The restricted injectivity condition (RI) plays an important role in our local convergence rate analysis and in general cannot be relaxed. However, for some special cases, such as when $R$ is locally polyhedral, it can be removed, at the price of less sharp rate estimation. This is formalized in the following statement.

**Theorem 25.** Suppose that (H.1)-(H.3) hold, and an FB-type method creates a sequence $x_k \rightarrow x^* \in \text{Argmin}(\Phi)$ such that $R$ is locally polyhedral around $x^*$, $F$ is $C^2$ near $x^*$, and condition (ND) holds. If moreover there exists $\epsilon > 0$ and a subspace $V$ such that

$$\ker(P_{x^*} \nabla^2 F(x)P_{x^*}) = V, \ \forall x \in \mathbb{B}(x^*) \cap (x^* + T_{x^*}).$$

Then $(x_k)_{k \in \mathbb{N}}$ converges locally linearly to $x^*$.

The expression of the local rate can be found by inspecting the proof.

**4.4. Discussion.** We here summarize some main conclusions on the local linear convergence behaviour of FB-types methods. Recall that $\alpha$ from (4.1) and $1/\beta$ is the Lipschitz constant of $\nabla F$.

FB is locally faster than FISTA. For the sake of brevity (the same conclusions hold true in the general case), we consider $b_k = a_k = a \in [0, 1]$ and $\gamma_k = \gamma \in [0, \beta]$ is fixed, in which case $\gamma \geq 0$ (see Lemma 14(ii)), and thus condition (4.10) is in force. Moreover $\gamma$ is also the local convergence rate of the FB method, and $\rho(M)$ depends solely on $\gamma$ and the value of $a$. Recall that $\rho(M)$ is the best local linear convergence rate (see Theorem 23 and 21).

Figure 1 shows $\rho(M)$ as a function of $a$ for fixed $\gamma$. One can make the the following observations:

(i) When $a \in [0, \gamma]$, we have $\rho(M) \leq \gamma$. This entails that if iFB is used with such a choice of inertial parameter, it will converge locally lineally faster than FB. For $a \in [\gamma, 1]$, the situation reverses as $\rho(M) \geq \gamma$, and iFB becomes slower than FB.

(ii) In particular, as $a = 1$ for FISTA, we have $\rho(M) = \sqrt{\gamma} > \gamma$. In plain words, though FISTA is known to be globally faster (in terms of the objective) than FB, attaining the optimal $O(1/k^2)$ rate, locally, the situation radically changes as FISTA will always ends up being locally slower than FB. A similar observation is made in [54] for the special case of FISTA used to solve the LASSO problem. This explains in particular why many authors [25, 46] resort to restarting to accelerate local convergence of FISTA, which consists in resetting periodically the scheme to $a = 0$ which is more favorable to FISTA.

Our predictions in Figure 1 gives clues on when to restart (i.e. detect the point in red on the rate curve).

(iii) $\rho(M)$ attains its minimal value at $a = \frac{(1-\sqrt{1-\gamma})^2}{\gamma}$, and this is the best convergence rate that can be achieved locally for FB-type methods.

Oscillation of the FISTA method. A typical feature of the FISTA method is that it is not monotone and locally oscillates [13], which makes the local convergence even
Fig. 1: Let $b = a$, and assume $\eta, \overline{\eta}$ are known and also close enough such that the spectral radius $\rho(M)$ is only affected by $\eta$, then $\rho(M)$ is a function of $a$.

slower, see Figure 2 and [54] for a FISTA applied to the LASSO problem. In fact, the iFB scheme shares this property as well when the inertial parameters are large.

Such oscillatory behaviour is due to the fact that, for those inertial parameters, the eigenvalue $\sigma_{\text{max}}$ such that $\left|\sigma_{\text{max}}\right| = \rho(M)$ is complex. It can then be shown that the oscillation period of $\|x_k - x^*\|$ is exactly $\frac{\pi}{\theta}$, where $\theta$ is the argument of $\sigma_{\text{max}}$.

For the parameter settings used in Figure 1, i.e. $b = a$ and $\gamma [0, \beta]$, we have

$$
\begin{cases}
  a \in [0, ((1 - \sqrt{1 - \eta})^2)/\overline{\eta}] : \sigma_{\text{max}} \text{ is real}, \\
  a \in \left\{ ((1 - \sqrt{1 - \eta})^2)/\overline{\eta}, 1 \right\} : \sigma_{\text{max}} \text{ is complex},
\end{cases}
$$

then as long as $a > (1 - \sqrt{1 - \eta})^2/\overline{\eta}$, the iFB method locally oscillates.

4.5. Acceleration. The finite time activity identification property (Theorem 8) implies that, the globally convex but non-smooth problem eventually becomes locally $C^2$-smooth, but possibly non-convex, constrained on the activity manifold. This opens the door to acceleration, and even finite termination, exploiting the structure of the objective and that of the identified manifold. There are several ways to achieve this goal as we explain hereafter.

Optimal first-order method. In this case, the idea is to keep the scheme implemented in Algorithm 1, and to refine the parameters to minimize the local convergence rate established in Section 4. Indeed, as shown in Figure 1 and the discussion that follows, there is a proper choice of the inertial parameters $a$ and $b$ that minimizes $\rho(M)$. More precisely, choose $\gamma [0, \beta]$, then $\overline{\eta} = 1 - \alpha \gamma \geq \eta \geq 1 - \gamma / \beta \geq 0$, and $\rho(M)$ depends only on $\eta, a$ and $b$. Then with fixed $\gamma$ (hence $\overline{\eta}$), $\rho(M)$ attains its minimal value for $a$ and $b$ satisfying

$$
\begin{cases}
  b = a : a = ((1 - \sqrt{1 - \eta})^2)/\overline{\eta} = (1 - \sqrt{\alpha \gamma})/(1 + \sqrt{\alpha \gamma}), \\
  b \neq a : a = (1 - \sqrt{1 - \eta})^2 + b(1 - \eta) = (1 - \sqrt{\alpha \gamma})^2 + b a \gamma,
\end{cases}
$$

and the optimal value $\rho^*$ of $\rho(M)$ reads

$$
\rho^* = 1 - \sqrt{1 - \eta} = 1 - \sqrt{\gamma \alpha},
$$

where the second equality comes from (4.2) and Lemma 14. This is a decreasing function of $\gamma$, and $\rho^* = 1 - \sqrt{\alpha \beta}$ is then the minimal rate attained for $\gamma = \beta$. This rate is in agreement with that [44, Theorem 2.2.2]. If one can afford $\gamma \geq \beta$ as in our iFB schemes, owing to the result of [50, Section 3.2.1], the best local linear rate is actually

$$
\rho^* = \frac{1 - \sqrt{\alpha \beta}}{1 + \sqrt{\alpha \beta}} \quad \text{for} \quad \gamma = \frac{4 \beta}{(1 + \sqrt{\alpha \beta})^2}, \quad a = \left(\frac{1 - \sqrt{\alpha \beta}}{1 + \sqrt{\alpha \beta}}\right)^2 \quad \text{and} \quad b = 0.
$$

This manuscript is for review purposes only.
This is known to be the optimal rate that matches the lower complexity bounds for first-order methods to solve the class of problems \((P_{\text{opt}})\) if \(F\) were also \(\alpha\)-strongly convex [44, Theorem 2.1.13]. In comparison, for the FB method \((i.e. a = b = 0)\), the optimal rate is \(\rho^* = \eta^* = \frac{1-\alpha^2}{1+\alpha^2}\) attained for \(\gamma = \frac{2\beta}{1+\alpha\beta}\).

High-order acceleration: Newton method. Once the activity manifold has been identified, one can switch to Newton-type methods for locally minimizing \(\Phi\). This can be done either using local parameterizations obtained from \(U\)-Lagrangian theory or from Riemannian geometry [34, 40, 52]. One can also use the Riemannian version of the non-linear conjugate gradient method [52]. For these schemes, one can also show respectively quadratic and superlinear convergence since \(\nabla^2_{\mathcal{M}_x^*}\Phi(x^*)\) is positive definite by Proposition 12(ii).

5. Numerical experiments. In this section, we illustrate the obtained results by some popular examples originating from linear inverse problems in signal processing and machine learning. We consider the linear model \(y = Lx_{\text{ob}} + w\), where \(y \in \mathbb{R}^m\), \(L : \mathbb{R}^n \to \mathbb{R}^m\) is some linear operator, and \(w \in \mathbb{R}^m\) stands for noise. Solving such a linear inverse problem can be cast as the optimization problem

\[
(P_{\lambda}) \quad \min_{x \in \mathbb{R}^n} \frac{1}{2} \|y - Lx\|^2 + \lambda R(x),
\]

where \(\lambda > 0\) is the tradeoff parameter, \(R \in \Gamma_0(\mathbb{R}^n)\) promotes objects similar to \(x_{\text{ob}}\).

We use three functions \(R\): the \(\ell_1\)-norm \((R(x) = \|x\|_1 \text{ def } = \sum_{i=1}^n |x_i|)\), the \(\ell_{1,2}\)-norm \((R(x) = \|x\|_{1,2} \text{ def } = \sum_{b \in B} \|x_b\|\) for a uniform disjoint partition of \(\{1, \ldots, n\}\) in blocks \(B\), and the nuclear norm \((R(x) = \|x\|_* \text{ def } = \|\sigma(x)\|_1\) where \(\sigma(x) \in (\mathbb{R}_+ \setminus \{0\})^r\) is the vector of singular values of the rank-\(r\) matrix \(x \in \mathbb{R}^{n_1 \times n_2}\). Both the \(\ell_1\) and \(\ell_{1,2}\)-norms are partly smooth relative to subspaces [57] \((\ell_1\) is polyhedral), and the nuclear norm is partly smooth relative to the constant rank-\(r\) manifold [22].

In all tests, the entries of \(L\) are independent copies of a mean-zero and standard Gaussian random variable. We consider the following settings of \(x_{\text{ob}}\):

- \(\ell_1\)-norm: \((m, n) = (48, 128)\), \(\|x_{\text{ob}}\|_0 = 8\);
- \(\ell_{1,2}\)-norm: \((m, n) = (60, 128)\), \(x_{\text{ob}}\) has 3 non-zero blocks of size 4;
- Nuclear norm: \((m, n) = (1425, 2500)\), \(x_{\text{ob}} \in \mathbb{R}^{50 \times 50}\) and rank \((x_{\text{ob}}) = 5\).

One can show that with the number of measurements \(m\) in the above cases, if \(\lambda\) and \(\|w\|\) are set properly, then with high probability on \(L\), \((P_{\lambda})\) admits a unique solution \(x^*\) with \(\mathcal{M}_{x^*} = \mathcal{M}_{x_{\text{ob}}}\), and \(x^*\) satisfies both \((\text{ND})\) and \((\text{RI})\).

Parameter settings. We choose \(\gamma_k \equiv \beta\) for FISTA. For FB/ifB methods, two choices of \(\gamma_k\) are considered: \(\gamma_k \equiv \beta\) and \(\gamma_k \equiv 1.5\beta\). The inertial parameter of iFB and FISTA are:

- FISTA: \(a_k = b_k = (k-1)/(k+q)\), with \(q = 2\) and \(q = 50\);
- iFB: \(\gamma_k \equiv \beta: a_k = b_k \equiv \sqrt{5} - 2 - 10^{-3}\) such that Theorem 4 applies;
- iFB: \(\gamma_k \equiv 1.5\beta: a_k, b_k\) are chosen according to (2.3) such that Theorem 3 applies.

The convergence profiles of \(\|x_k - x^*\|\) are shown in Figure 2. As demonstrated by all the plots, identification and local linear convergence occurs after finite time. The solid lines (denoted as “P”) represent the observed profiles, while dashed ones (denoted as “T”) stand for the theoretically predicted ones. The positions of the green points (or the starting points of the dashed lines) stand for the iteration at which \(\mathcal{M}_{x^*}\) has been identified.

Tightness of predicted rates. For the \(\ell_1\)-norm, our predicted rates coincide exactly with the observed ones (same slopes for the dashed and solid lines). This is due to the fact that they are all polyhedral and \(F\) is quadratic. Note that for FISTA, which is non-monotone, the prediction coincides with the envelope of the oscillations. For
the $\ell_{1,2}$-norm, though it is not polyhedral, our predicted rates still are very tight, due to the fact that the Riemannian Hessian is taken into account. For the nuclear norm, whose active manifold is not anymore a subspace, our estimation becomes slightly less sharp compared to the other examples, though barely visible on the plots. Our predicted rates for FB are much sharper than in our previous work [37].

Comparison of the methods. From the numerical results, we can infer the following observations.

(i) Comparison of FB/iFB and FISTA under $\gamma_k \equiv \beta$:
   - Globally, FISTA $q = 50$ is the fastest while $q = 2$ is the slowest. FB and iFB are in between them with iFB being faster.
   - For the finite identification, however, FISTA $q = 2$ in general shows the fastest identification, and FB is the slowest.
   - Locally, similar to the global convergence, FISTA $q = 50$ has the fastest rate and $q = 2$ is the slowest. Again, FB and iFB are between them with iFB being faster than FB.

(ii) $\gamma_k \equiv \beta$ vs $\gamma_k \equiv 1.5\beta$:
   - For FB, larger $\gamma_k$ leads to faster global convergence and activity identification. However, this does not mean that the bigger the better locally. As we discussed in Section 4.5, the best choice to get the optimal local linear rate is $2\beta/(1 + \alpha\beta)$.
   - iFB is faster than FB under the same choice of $\gamma_k$. FISTA $q = 50$ is no longer the fastest one, while it is outperformed by iFB $\gamma_k \equiv 1.5\beta$ for the first 2 examples.

It can be concluded from the above remarks that, in practice, FISTA with $q = 2$ is not a wise choice if high accuracy solutions are needed. Indeed, under this choice, $a_k$ converges to 1 too fast, and this hampers its local behaviour as the discussions we anticipated in Section 4.4 (see Figure 1). In fact, such behaviour of $a_k$ can be avoided by choosing relatively bigger $q$, and this is exactly what the difference between $q = 2$ and $q = 50$ implies. In our tests, $q \in [50, 100]$ seems to a good trade-off, even bigger $q$ is not recommended since it may lead to a much slower activity identification.

However, it should be pointed out that the local rate of FISTA $q = 50$ being faster than FB does not contradict with our claim in Section 4.4 that FB is faster than FISTA locally. The reason is that we are limited by machine accuracy, and bigger value of $q$ delays the speed at which $a_k$ approaches to 1 which actually makes FISTA behaviour similar to the iFB method.

Acceleration. For the $\ell_1$-norm which is polyhedral, we applied the first-order acceleration described in (4.11) for $\gamma_k \equiv \beta$ and $\gamma_k \equiv 1.5\beta$ respectively (Figure 2(a)). In fact, acceleration is not even needed in this case and one can access a closed-form
solution of $x^*$ once identification occurs. This can be easily achieved by projection the
first-order minimality condition on $\mathcal{M}_{x^*} = x^* + T_{x^*}$, which boils down to solving an
overdetermined linear system which has a unique solution under the restricted inject-
tivity condition (RI). For the $\ell_{1,2}$-norm, we applied the Riemannian Newton method
which converges quadratically, leading to a dramatic acceleration as can be seen in
Figure 2(b). For the nuclear norm, a non-linear conjugate gradient method is ap-
plicated, leading again to a much faster (super-linear) local convergence. To summarize,
in practice, the inertial+higher-order method hybrid strategy is an ideal choice for
solving ($\mathcal{P}_{\text{opt}}$).

Acknowledgements. This work has been partly supported by the European
Research Council (ERC project SIGMA-Vision). JF is partly supported by Institut
Universitaire de France.

Appendix A. Proofs of Section 2.

Throughout this section, $\mathcal{H}$ denotes a real Hilbert space. Let $A : \mathcal{H} \rightrightarrows \mathcal{H}$ be a
set-valued operator. The graph of $A$ is the set $\text{gph} A = \{(x, y) \in \mathcal{H} \times \mathcal{H} | y \in A(x)\}$,
and its zero set is $\text{zer} A = \{x \in \mathcal{H} | 0 \in A(x)\}$. Recall that a set-valued operator
$A : \mathcal{H} \rightrightarrows \mathcal{H}$ is monotone if
\begin{equation}
(\forall (x, v) \in \text{gph} A), (\forall (y, u) \in \text{gph} A), \langle x - y, v - u \rangle \geq 0.
\end{equation}

It is moreover maximal monotone if $\text{gph} A$ cannot be contained in the graph of any
other monotone operator. Let $\beta \in [0, +\infty[; B : \mathcal{H} \to \mathcal{H}$, then $B$ is $\beta$-cocoercive if
\begin{equation}
(\forall x, y \in \mathcal{H}), \beta \|Bx - By\|^{2} \leq \langle Bx - By, x - y \rangle.
\end{equation}

\textbf{Proof (Theorem 3).} Define the following quantities
\begin{equation}
\varphi_{k} = \frac{1}{2}\|x_{k} - x^{*}\|^{2}, \Delta_{k} = \frac{1}{2}\|x_{k} - x_{k-1}\|^{2}, E_{b, k} = \frac{1}{2}\|y_{b, k} - x_{k+1}\|^{2}.
\end{equation}

Let $x^{*} \in \text{zer}(A + B)$, i.e. a solution ($\mathcal{P}_{\text{inc}}$), which exists thanks to (H.6). Recall from
(1.4) and (2.1) that
\begin{equation}
-B(x^{*}) \in A(x^{*}) \quad \text{and} \quad y_{a, k} - \gamma_{k} B(y_{b, k}) - \gamma_{k} \xi_{k} - x_{k+1} \in \gamma_{k} A^{\ast}(x_{k+1}).
\end{equation}

Thus, we get
\begin{equation}
\langle y_{a, k} - x_{k+1} - \gamma_{k} B(y_{b, k}) - \gamma_{k} \xi_{k} - x_{k+1}, x_{k+1} - x^{*} \rangle \geq -\gamma_{k} \varepsilon_{k}.
\end{equation}

Combining this with the definition of $y_{a, k}$, we obtain
\begin{equation}
\varphi_{k} - \varphi_{k+1} = \frac{1}{2}(x_{k} - x^{*} + x_{k+1} - x^{*}, x_{k} - x_{k+1})
= \Delta_{k+1} + \langle y_{a, k} - x_{k+1}, x_{k+1} - x^{*} \rangle - a_{k}(x_{k} - x_{k-1}, x_{k+1} - x^{*})
\geq \Delta_{k+1} + \gamma_{k} \langle B(y_{b, k}) - B(x^{*}) + \xi_{k}, x_{k+1} - x^{*} \rangle
- a_{k}(x_{k} - x_{k-1}, x_{k+1} - x^{*}) - \gamma_{k} \varepsilon_{k}.
\end{equation}

For $\langle x_{k} - x_{k-1}, x_{k+1} - x^{*} \rangle$, we have
\begin{equation}
\langle x_{k} - x_{k-1}, x_{k+1} - x^{*} \rangle = \langle x_{k} - x_{k-1}, x_{k+1} - x_{k} \rangle + (\Delta_{k} + \varphi_{k} - \varphi_{k-1}),
\end{equation}
where we applied the usual Pythagoras relation to $\langle x_{k} - x_{k-1}, x_{k} - x^{*} \rangle$. Putting (A.5)
back into (A.4) yields
\begin{equation}
\varphi_{k+1} - \varphi_{k} - a_{k}(\varphi_{k} - \varphi_{k-1}) \leq -\Delta_{k+1} + \gamma_{k} \langle B(y_{b, k}) - B(x^{*}) + \xi_{k}, x_{k+1} - x^{*} \rangle
+ a_{k}(x_{k} - x_{k-1}, x_{k+1} - x_{k}) + a_{k} \Delta_{k} + \gamma_{k} \varepsilon_{k}.
\end{equation}

Since $B$ is $\beta$-cocoercive, Young’s inequality yields
\begin{equation}
\langle B(y_{b, k}) - B(x^{*}), x_{k+1} - x^{*} \rangle \geq \beta \|B(y_{b, k}) - B(x^{*})\|^{2} + \langle B(y_{b, k}) - B(x^{*}), x_{k+1} - y_{b, k} \rangle = -\frac{1}{2 \beta} E_{b, k}.
\end{equation}
Denote \( \mu_k = 1 - \frac{\gamma_k}{2\beta} \in [\frac{\gamma_k}{2\beta}, 1 - \frac{\gamma_k}{2\beta}] \), \( \nu_k = a_k - \frac{\gamma_k b_k}{2\beta} \) and \( v_k = x_{k+1} - x_k - \frac{\mu_k}{2\beta} (x_k - x_{k-1}) \).

Substituting (A.7) back into (A.6), and since \( \Delta_k = \Delta_{k+1} + b_k^2 \Delta_k + b_k (x_k - x_{k+1}, x_k - x_{k-1}) \), we get (A.8)

\[
\varphi_{k+1} - \varphi_k - a_k (\varphi_k - \varphi_{k-1}) \\
\leq -\Delta_{k+1} + \frac{\gamma_k}{2\beta} E_{b,k} + a_k (x_k - x_{k-1}, x_{k+1} - x_k) + a_k \Delta_k + \gamma_k \varepsilon_k - \gamma_k (\xi_k, x_{k+1} - x^*) \\
\leq -\frac{\mu_k}{2} ||v_k||^2 + (a_k + \frac{\nu_k^2}{\mu_k} + \frac{\gamma_k b_k^2}{2\beta}) \Delta_k + \gamma_k (\xi_k + \sqrt{2||\varphi_k||}) \\
\leq -\frac{\mu_k}{2} ||v_k||^2 + \left( \frac{2a_k}{\mu_k} + \frac{\gamma_k b_k^2}{2\beta} \right) \Delta_k + \gamma_k (\xi_k + \sqrt{2||\varphi_k||}) \\
\leq -\frac{\mu_k}{2} ||v_k||^2 + \left( \frac{4\beta}{\beta} a_k + (1 - \frac{\gamma_k}{2\beta}) b_k \right) \Delta_k + \gamma_k (\xi_k + \sqrt{2||\varphi_k||}).
\]

where \( \gamma = (2\beta - \gamma) \). Denote \( \theta_k = \varphi_k - \varphi_{k-1} \) and \( \delta_k = \left( \frac{4\beta}{\beta} a_k + (1 - \frac{\gamma_k}{2\beta}) b_k \right) \Delta_k \). We then arrive at the following key estimate

\[
\theta_{k+1} \leq -\frac{\mu_k}{2} ||v_k||^2 + a_k \theta_k + \delta_k + \gamma \varepsilon_k + \sqrt{2\gamma} ||\xi_k|| \sqrt{\varphi_{k+1}}
\]

(A.9)

\[
\leq \prod_{j=1}^{k} a_j \theta_j + \sum_{j=1}^{k} \left( \sum_{l=1}^{j} a_{j-l} \right) \left( \delta_j + \gamma \varepsilon_j + \sqrt{2\gamma} ||\xi_j|| \sqrt{\varphi_{j+1}} \right)
\]

\[
\leq \bar{a}^k \varphi_1 + \sum_{j=1}^{k} \bar{a}^{k-j} \left( \delta_j + \gamma \varepsilon_j + \sqrt{2\gamma} ||\xi_j|| \sqrt{\varphi_{j+1}} \right).
\]

(i) \( a_k \in [0, a] \): summing up the last inequality, we get

\[
\sum_{m=1}^{k} \theta_{m+1} = \varphi_{k+1} - \varphi_1 \\
\leq -\frac{1}{a} \varphi_1 + \sum_{m=1}^{k} \sum_{j=1}^{m} \bar{a}^{k-j} \left( \delta_j + \gamma \varepsilon_j + \sqrt{2\gamma} ||\xi_j|| \sqrt{\varphi_{j+1}} \right)
\]

\[
\leq -\frac{1}{a} \varphi_1 + \sum_{m=1}^{k} \left( \sum_{j=1}^{m} \bar{a}^{k-j} \right) \left( \delta_m + \gamma \varepsilon_m + \sqrt{2\gamma} ||\xi_m|| \sqrt{\varphi_{m+1}} \right)
\]

\[
\leq -\frac{1}{a} \left( \varphi_1 + \sum_{m=1}^{k} \left( \delta_m + \gamma \varepsilon_m + \sqrt{2\gamma} ||\xi_m|| \sqrt{\varphi_{m+1}} \right) \right),
\]

which entails (A.10)

\[
\varphi_{k+1} \leq c + \sqrt{2\gamma} \sum_{m=1}^{k} ||\xi_m|| \sqrt{\varphi_{m+1}} \leq c + \sqrt{2\gamma} \sum_{m=1}^{k+1} ||\xi_{m-1}|| \sqrt{\varphi_{m+1}},
\]

where \( c = \varphi_1 + \frac{1}{a} (\varphi_1 + \sum_{m=1}^{k} \delta_m + \gamma \varepsilon_m) \geq 0. \) By assumption on the sequences \( (\varepsilon_m)_{m=1}^{\infty} \) and \( (\delta_m)_{m=1}^{\infty} \), \( c \) is bounded. Using the fact that \( (||\xi_m||)_{m=1}^{\infty} \) is summable, it can be easily shown, e.g. [6, Lemma A.9], that since \( (\varphi_k)_{k=1}^{\infty} \) satisfies (A.10), it also obeys \( \varphi_k \leq \sqrt{c} + \sum_{j=1}^{\infty} ||\xi_j|| < +\infty. \)

Denote \( t = \sqrt{c} + \sum_{j=1}^{\infty} ||\xi_j|| \). Then, (A.9) becomes

\[
\theta_{k+1} \leq -\frac{\mu_k}{2} ||v_k||^2 + a_k \theta_k + \delta_k + \gamma \varepsilon_k + \sqrt{2\gamma} ||\xi_k||
\]

(A.11)

\[
\leq -\frac{\mu_k}{2} ||v_k||^2 + a_k [\theta_1]_+ + \delta_k + \gamma \varepsilon_k + \sqrt{2\gamma} ||\xi_k||
\]

where \( [\theta_1]_+ = \max \{ \theta_1, 0 \} \). As a result, we have

\[
[\theta_{k+1}]_+ \leq a[\theta_1]_+ + e_k,
\]

where \( e_k = \delta_k + \gamma \varepsilon_k + \sqrt{2\gamma} ||\xi_k|| \) is a summable sequence by assumption. Therefore, using that \( a < 1 \) and applying [20, Lemma 3.1(iv)], it follows that \( [\theta_k]_+ \) is summable. In turn,

\[
\varphi_{k+1} - \sum_{j=1}^{k} [\theta_j]_+ \leq \varphi_{k+1} - \theta_{k+1} - \sum_{j=1}^{k} [\theta_j]_+ = \varphi_k - \sum_{j=1}^{k} [\theta_j]_+.
\]

It then follows that the sequence \( (\varphi_k - \sum_{j=1}^{k} [\theta_j]_+)_{k=1}^{\infty} \) is decreasing and bounded from below, hence convergent, whence we deduce that \( \varphi_k \) is also convergent.
(ii) \( a_k \equiv 0 \): in this case, (A.9) reduces to
\[
\varphi_{k+1} \leq \varphi_k + \delta_k + \gamma \varepsilon_k + \sqrt{2\gamma} \|\xi_k\| \sqrt{\varphi_k}
\]
\[
\leq \varphi_1 + \sum_{j \in \mathbb{N}} \delta_j + \gamma \sum_{j \in \mathbb{N}} \varepsilon_j + \sqrt{2\gamma} \sum_{j=1}^k \|\xi_j\| \sqrt{\varphi_{j+1}}.
\]
Again, by virtue of [6, Lemma A.9] and the summability of the sequences 
\((\delta_j)_{k \in \mathbb{N}}, (\varepsilon_j)_{k \in \mathbb{N}}\) and \((\|\xi_j\|)_{k \in \mathbb{N}}\), we have \( \varphi_k \leq t = \sqrt{\varphi_1 + \sum_{j \in \mathbb{N}} (\delta_j + \gamma \varepsilon_j + \|\xi_j\|)} < +\infty \).
Consequently, we have
\[
\varphi_{k+1} \leq \varphi_k + \delta_k + \gamma \varepsilon_k + \sqrt{2\gamma} \|\xi_k\|.
\]

We then conclude that the sequence \((x_k)_{k \in \mathbb{N}}\) is quasi-Fejér monotone (of type III) relative to \(\text{zer}(A + B)\) [20, Definition 1.1(3)], and thus \(\varphi_k\) is convergent [20, Proposition 3.6].

In summary, for \( a_k \in [0, a] \), \( \lim_{k \to +\infty} \|x_k - x^*\| \) exists for any \( x^* \in \text{zer}(A + B) \), and \((x_k)_{k \in \mathbb{N}}\) is bounded.

By assumption (2.2), \( a_k(x_k - x_{k-1}) \to 0 \) and \( b_k(x_k - x_{k-1}) \to 0 \), and thus
\[
\frac{\varphi_k}{\mu_k} (x_k - x_{k-1}) \to 0,
\]
since \( \mu_k \geq \frac{c_0}{2\varepsilon_0} > 0 \). Moreover, from (A.11), we obtain
\[
\sum_{k \in \mathbb{N}} \|v_k\|^2 \leq \frac{4\beta}{\tau} \left( a \varphi_0 + \sum_{k \in \mathbb{N}} (\bar{a} \theta_k + \varepsilon_k) \right) < +\infty.
\]

Consequently, \( v_k \to 0 \). Combining this with (A.12), we get that \( x_{k+1} - x_k \to 0 \). In

turn, \( y_{a,k} - x_{k+1} \to 0 \) and \( y_{b,k} - x_{k+1} \to 0 \). Let \( \bar{x} \) be a weak cluster point of \((x_k)_{k \in \mathbb{N}}\), and

let us fix a subsequence, say \( x_{k_j} \to \bar{x} \). Denote \( u_{k_j} \equiv \frac{y_{a,k_j} - x_{k_j+1} - B(y_{b,k_j}) - \xi_{k_j}}{2} \).
Since \( B \) is cocoercive and \( y_{b,k_j} \to \bar{x} \), we have \( B(y_{b,k_j}) \to B(\bar{x}) \). In turn, \( u_{k_j} \to -B(\bar{x}) \) since \( \gamma_k \geq \xi > 0 \) and \( \xi_k \to 0 \). Since \((x_{k_j+1}, u_{k_j}) \in \text{gph} A^{k_j}\), and the graph of the enlargement of \( A \) is weakly-strongly sequentially closed in \( \mathbb{R}_+ \times \mathcal{H} \times \mathcal{H} \) [53, Proposition 3.4(b)], we get that \( -B(\bar{x}) \in A(\bar{x}) \), i.e. \( \bar{x} \) is a solution of \((P_{\text{inc}})\). Opial’s theorem [47] concludes the proof.

**Proof (Theorem 4).** In view of the imposed assumptions, we deduce from Theorem 3 that \((x_k)_{k \in \mathbb{N}}\) is bounded, and thus \( c = \sup_{k \in \mathbb{N}} \|x_k - x^*\| < +\infty \). From (A.8), we apply Young’s inequality to get
\[
\varphi_{k+1} - \varphi_k - a_k (\varphi_k - \varphi_{k-1})
\leq \left( \frac{\gamma_k}{2\gamma} - 1 \right) \Delta_{k+1} + |a_k - \frac{2\gamma b_k}{2\gamma}| (\Delta_{k+1} + \Delta_k) + \left( \frac{4\gamma b_k^2}{2\gamma} + a_k \right) \Delta_k + (\gamma_k \varepsilon_k + c \|\xi_k\|)
\]
\[
= s_k \Delta_{k+1} + t_k \Delta_k + \gamma (\varepsilon_k + c \|\xi_k\|),
\]
where \( s_k = \frac{\gamma_k}{2\gamma} - 1 + |a_k - \frac{2\gamma b_k}{2\gamma}|, t_k = \frac{4\gamma b_k^2}{2\gamma} + a_k + |a_k - \frac{2\gamma b_k}{2\gamma}| \). Suppose that \( a_k, b_k \) and \( \gamma_k \) are non-decreasing so that \( s_k, t_k \) are also non-decreasing. Denote \( \phi_k = \varphi_k - a_k \varphi_{k-1} + t_k \Delta_k \) and \( \delta_k = \gamma (\varepsilon_k + c \|\xi_k\|) \).
\[
\phi_{k+1} - \phi_k \leq (\varphi_{k+1} - \varphi_k) - a_k (\varphi_k - \varphi_{k-1}) + t_{k+1} \Delta_{k+1} - t_k \Delta_k
\leq s_k \Delta_{k+1} + t_k \Delta_k + t_{k+1} \Delta_{k+1} - t_k \Delta_k + \delta_k
\leq (s_k + t_{k+1}) \Delta_{k+1} + \delta_k.
\]

(i) \( a_k \in [0, a], b_k \in [0, b], b_k \leq a_k \). We have \( \frac{2\gamma b_k}{2\gamma} < a_k \), then from (A.13), and under the second condition in (2.4),
\[
\phi_{k+1} - \phi_k \leq (s_{k+1} + t_{k+1}) \Delta_{k+1}
\leq (3a_{k+1} - 1) + \frac{2\gamma b_{k+1}}{2\gamma} (1 - b_{k+1})^2 \Delta_{k+1} + \delta_k \leq -\tau \Delta_{k+1} + \delta_k.
\]
(ii) \( a_k \in [0, \bar{a}], b_k \in [0, \bar{b}], a_k < b_k \). Since \( s_k, t_k \) are non-decreasing, then from (A.13) we have,
\[
\phi_{k+1} - \phi_k \leq (s_{k+1} + t_{k+1}) \Delta_{k+1} + \delta_k \\
\leq \left( \frac{2k+1}{2\beta} - 1 + 2|a_{k+1} - \frac{2k+1}{2\beta} b_{k+1}| + \frac{2k+1}{2\beta} b_{k+1}^2 + a_{k+1} \right) \Delta_{k+1} + \delta_k.
\]

Next we discuss the relationship between \( a_{k+1} \) and \( \frac{2k+1}{2\beta} b_{k+1} \), which splits into two subcases.

(a) If \( \frac{2k+1}{2\beta} b_{k+1} \leq a_{k+1}, k \in \mathbb{N} \), then from the second condition in (2.4),
\[
\phi_{k+1} - \phi_k \leq \left( (3a_{k+1} - 1) + \frac{2k+1}{2\beta} (1 - b_{k+1})^2 \right) \Delta_{k+1} + \delta_k \leq -\tau \Delta_{k+1} + \delta_k.
\]

(b) If \( a_{k+1} < \frac{2k+1}{2\beta} b_{k+1}, k \in \mathbb{N} \), then from the first condition of (2.4),
\[
\phi_{k+1} - \phi_k \leq (1 + a_{k+1}) + \frac{2k+1}{2\beta} (1 + b_{k+1})^2 \Delta_{k+1} + \delta_k \leq -\tau \Delta_{k+1} + \delta_k.
\]

Under the assumptions of (i), we have from (A.14) (resp. (A.15) or (A.16)) that
\[
\sum_{j=1}^k \Delta_{j+1} \leq \frac{1}{\tau} (\phi_1 - \phi_{k+1}) + \sum_{j=1}^k \delta_j \leq \frac{1}{\tau} (\phi_1 + \bar{a} \varphi_k) + \sum_{j=1}^k \delta_j < +\infty.
\]

If the errors vanish, (A.14) (resp. (A.15) or (A.16)) indicate that \( \phi_k \) is non-increasing. Thus
\[
\sum_{j=1}^k \Delta_{j+1} \leq \frac{1}{\tau} (\phi_1 - \phi_{k+1}) \leq \frac{1}{\tau} (\phi_1 + \bar{a} \varphi_k) \leq \frac{1}{\tau} \left( \bar{a}_k \varphi_1 + \frac{\phi_1}{1 - \bar{a}} \right) < +\infty.
\]

In summary, the summability condition in (2.2) is satisfied. The claim follows from Theorem 3. \( \square \)

Appendix B. Proofs of Section 4.

B.1. Riemannian Geometry. Let \( \mathcal{M} \) be a \( C^2 \)-smooth embedded submanifold of \( \mathbb{R}^n \) around a point \( x \). With some abuse of terminology, we shall state \( C^2 \)-manifold instead of \( C^2 \)-smooth embedded submanifold of \( \mathbb{R}^n \). The natural embedding of a submanifold \( \mathcal{M} \) into \( \mathbb{R}^n \) permits to define a Riemannian structure and to introduce geodesics on \( \mathcal{M} \), and we simply say \( \mathcal{M} \) is a Riemannian manifold. Denote respectively \( T_\mathcal{M}(x) \) and \( N_\mathcal{M}(x) \) the tangent and normal space of \( \mathcal{M} \) at point near \( x \) in \( \mathcal{M} \).

Exponential map. Geodesics generalize the concept of straight lines in \( \mathbb{R}^n \), preserving the zero acceleration characteristic, to manifolds. Roughly speaking, a geodesic is locally the shortest path between two points on \( \mathcal{M} \). We denote by \( \exp(t; x, h) \) the value at \( t \in \mathbb{R} \) of the geodesic starting at \( \exp(0; x, h) = x \in \mathcal{M} \) with velocity \( \dot{\exp}(t; x, h) = \frac{d}{dt}(t; x, h) = h \in \mathcal{T}_\mathcal{M}(x) \) (which is uniquely defined). For every \( h \in \mathcal{T}_\mathcal{M}(x) \), there exists an interval \( I \) around 0 and a unique geodesic \( \exp(t; x, h) : I \to \mathcal{M} \) such that \( \exp(0; x, h) = x \) and \( \dot{\exp}(0; x, h) = h \). The mapping
\[
\exp_x : T_\mathcal{M}(x) \to \mathcal{M}, \ h \mapsto \exp_x(h) = \exp(1; x, h),
\]
is called Exponential map. Given \( x, z \in \mathcal{M} \), the direction \( h \in T_\mathcal{M}(x) \) we are interested in is such that
\[
\exp_x(h) = z = \exp(1; x, h).
\]

Parallel translation. Given two points \( x, z \in \mathcal{M} \), let \( T_\mathcal{M}(x), T_\mathcal{M}(z) \) be their corresponding tangent spaces. Define
\[
\tau : T_\mathcal{M}(x) \to T_\mathcal{M}(z),
\]
the parallel translation along the unique geodesic joining \( x \) to \( z \), which is isomorphism and isometry w.r.t. the Riemannian metric.
Riemannian gradient and Hessian. For a vector $v \in T_M(x)$, the Weingarten map of $M$ at $x$ is the operator $\mathfrak{W}_x(\cdot, v) : T_M(x) \to T_M(x)$ defined by

$$\mathfrak{W}_x(\cdot, v) = -P_{T_M(z)}dV[h],$$

where $V$ is any local extension of $v$ to a normal vector field on $M$. The definition is independent of the choice of the extension $V$, and $\mathfrak{W}_x(\cdot, v)$ is a symmetric linear operator which is closely tied to the second fundamental form of $M$, see [19, Proposition II.2.1].

Let $G$ be a real-valued function which is $C^2$ along the $M$ around $x$. The covariant gradient of $G$ at $z \in M$ is the vector $\nabla_M G(z) \in T_M(z)$ defined by

$$\langle \nabla_M G(z), h \rangle = \frac{d}{dt} G(P_M(z + th)) \bigg|_{t=0}, \quad \forall h \in T_M(z),$$

where $P_M$ is the projection operator onto $M$. The covariant Hessian of $G$ at $z$ is the symmetric linear mapping $\nabla^2_M G(z)$ from $T_M(z)$ to itself which is defined as

$$\nabla^2_M G(z)h, h \rangle = \frac{d^2}{dt_1 dt_2} G(P_M(z + th_1 + t_2h_2)) \bigg|_{t_1 = t_2 = 0}, \quad \forall h \in T_M(z).$$

This definition agrees with the usual definition using geodesics or connections [40].

Now assume that $M$ is a Riemannian embedded submanifold of $\mathbb{R}^n$, and that a function $G$ has a $C^2$-smooth restriction on $M$. This can be characterized by the existence of a $C^2$-smooth extension (representative) of $G$, i.e. a $C^2$-smooth function $\tilde{G}$ on $\mathbb{R}^n$ such that $\tilde{G}$ agrees with $G$ on $M$. Thus, the Riemannian gradient $\nabla_M G(z)$ is also given by

$$\nabla_M G(z) = P_{T_M(z)} \nabla \tilde{G}(z),$$

and $\forall h \in T_M(z)$, the Riemannian Hessian reads

$$\nabla^2_M G(z)h = P_{T_M(z)}d(\nabla_M G(z))[h] = P_{T_M(z)}d(z \mapsto P_{T_M(z)}\nabla_M \tilde{G})[h]$$

$$= P_{T_M(z)}\nabla^2 \tilde{G}(z)h + \mathfrak{W}_x(h, P_{N_M(z)}\nabla \tilde{G}(z)),$$

where the last equality comes from [2, Theorem 1]. When $M$ is an affine or linear subspace of $\mathbb{R}^n$, then obviously $M = x + T_M(x)$, and $\mathfrak{W}_x(h, P_{N_M(z)}\nabla \tilde{G}(z)) = 0$, hence (B.3) reduces to

$$\nabla^2_M G(z) = P_{T_M(z)}\nabla^2 \tilde{G}(z)P_{T_M(z)}.$$

See [33, 19] for more materials on differential and Riemannian manifolds.

The following lemmas summarize two key properties that we will need throughout.

**Lemma 26.** Let $x \in M$, and $x_k$ a sequence converging to $x$ in $M$. Denote $\tau_k : T_M(x) \to T_M(x_k)$ be the parallel translation along the unique geodesic joining $x$ to $x_k$. Then, for any bounded vector $u \in \mathbb{R}^n$, we have

$$(\tau_k^{-1}P_{T_M(x_k)} - P_{T_M(x)})u = o(\|u\|).$$

**Proof.** From [1, Chapter 5], we deduce that for $k$ sufficiently large,

$$\tau_k^{-1} = P_{T_M(x)} + o(\|x_k - x\|).$$

In addition, locally near $x$ along $M$, the operator $x \mapsto P_{T_M(x)}$ is $C^1$, hence,

$$\lim_{k \to \infty} \frac{\| (\tau_k^{-1}P_{T_M(x_k)} - P_{T_M(x)})u \|}{\|u\|} \leq \lim_{k \to \infty} \frac{\| P_{T_M(x)} (P_{T_M(x_k)} - P_{T_M(x)}) \|}{\|u\|} + o(\|x_k - x\|)$$

$$\leq \lim_{k \to \infty} \| P_{T_M(x_k)} - P_{T_M(x)} \| + o(\|x_k - x\|) = 0.$$

**Lemma 27.** Let $x, z$ be two close points in $M$, denote $\tau : T_M(x) \to T_M(z)$ the parallel translation along the unique geodesic joining $x$ to $z$. The Riemannian Taylor expansion of $\Phi \in C^2(M)$ around $x$ reads,

$$\tau^{-1}\nabla_M \Phi(z) = \nabla_M \Phi(x) + \nabla^2_M \Phi(x)P_{T_M(x)}(z - x) + o(\|z - x\|).$$

This manuscript is for review purposes only.
Proof. Since $x, z \in \mathcal{M}$ are close, we have $z = \text{Exp}_x(h)$ for some $h \in T_{\mathcal{M}}(x)$ small enough, and thus, the Taylor expansion [52, Remark 4.2] of $\nabla_{\mathcal{M}} \Phi$ around $x$ reads

\begin{equation}
\tag{B.4}
\tau^{-1} \nabla_{\mathcal{M}} \Phi(z) = \nabla_{\mathcal{M}} \Phi(x) + \nabla_{\mathcal{M}}^2 \Phi(x) h + o(\|h\|).
\end{equation}

Moreover, from the proof of [40, Theorem 4.9], one can show that

\[ P_{T_{\mathcal{M}}(x)}(z) = P_{T_{\mathcal{M}}(x)}(\exp(h)) = P_{T_{\mathcal{M}}(x)}(x) + h + o(\|h\|^2). \]

Substituting back into (B.4) we get the claimed result. \qed

B.2. Proofs.

Proof (Proposition 12).

(i) Since $F$ is locally $C^2$ around $x^*$, there exists $\epsilon > 0$ sufficiently small such that for any $\delta \in \mathbb{B}_r(0)$, we have for some $t \in [0, 1[$,

\[ \Phi(x^* + \delta) - \Phi(x^*) = \frac{1}{2} \langle \delta, \nabla^2 F(x^* + t\delta) \rangle + R(x^* + \delta) - R(x^*) + \langle \nabla F(x^*), \delta \rangle. \]

Let $x_t = x^* + t \delta \in \mathbb{B}_r(x^*)$. We then distinguish two cases.

(a) $\delta \notin \text{ker}(\nabla^2 F(x_t))$. Since $F$ and $R$ are convex with $-\nabla F(x^*) \in \partial R(x^*)$, $\Phi(x^* + \delta) - \Phi(x^*) \geq \frac{1}{2} \langle \delta, \nabla^2 F(x_t) \rangle > 0$.

(b) $\delta \in \ker(\nabla^2 F(x_t)) \setminus \{0\}$. As $R \in \Gamma_0(\mathbb{R}^n)$, it is sub-differentially regular at $x^*$. Moreover $\partial R(x^*) \neq \emptyset$ ($-\nabla F(x^*)$ is in it), and thus the directional derivative $R'(x^*, \cdot)$ is proper and closed, and it is the support of $\partial R(x^*)$ [51, Theorem 8.30]. It then follows from the separation theorem [30, Theorem V.2.2.3] that

\[ -\nabla F(x^*) \in \text{ri}(\partial R(x^*)) \]

\[ \Leftrightarrow R'(x^*, \delta) > (\nabla F(x^*), \delta), \forall \delta \text{ s.t. } R'(x^*; \delta) + R'(x^* - \delta) > 0. \]

Since (RI) holds and $\nabla^2 F(x)$ depends continuously on $x \in \mathbb{B}_r(x^*), (4.1)$ holds for any such $x$, and in particular at $x_t$. Combining with the fact that $\ker(\text{ker}(R'(x^*; \cdot)) = T_{x^*}$, [56, Proposition 3(iii) and Lemma 10], we get

\[ -\nabla F(x^*) \in \text{ri}(\partial R(x^*)) \Leftrightarrow R'(x^*, \delta) > (\nabla F(x^*), \delta), \forall \delta \notin T_{x^*}, \]

\[ R'(x^*, \delta) > (\nabla F(x^*), \delta), \forall \delta \in \ker(\nabla^2 F(x_t)) \setminus \{0\}. \]

Thus, classical properties of the directional derivative of a convex function yield

\[ \Phi(x^* + \delta) - \Phi(x^*) = R(x^* + \delta) - R(x^*) + (\nabla F(x^*), \delta) \geq R'(x^*; \delta) + (\nabla F(x^*), \delta) > 0. \]

(ii) Let $\Psi$ as defined in the proof of Lemma 13. If $R \in \text{PSF}_{x^*}(\mathcal{M}_{x^*})$, the Riemannian Hessian of $\Phi$ reads

\[ \nabla_{\mathcal{M}_{x^*}} \Phi(x^*) = R_{T_{x^*}} \nabla F(x^*) P_{T_{x^*}} + \nabla_{\mathcal{M}_{x^*}}^2 \Psi(x^*). \]

In view of Lemma 13(i), $\nabla_{\mathcal{M}_{x^*}}^2 \Psi(x^*)$ is positive semi-definite on $T_{x^*}$. On the other hand, hypothesis (RI) entails positive definiteness of $P_{T_{x^*}} \nabla F(x^*) P_{T_{x^*}}$. Altogether, this shows that $\nabla_{\mathcal{M}_{x^*}}^2 \Phi(x^*)$ is positive definite on $T_{x^*} \setminus \{0\}$. Local quadratic growth of $\Phi$ near $x^*$ then follows by combining [35, Definition 5.4], [40, Theorem 3.4] and [28, Theorem 6.2]. \qed

Proof (Lemma 13). By definition of $U$, $Uh = 0$ for any $h \in T_{x^*}$. Thus, in the following we only examine the case $h \in T_{x^*}$. (i) Let $\Psi(x) = R(x) + (x, \nabla F(x^*)).$ From the smooth perturbation rule of partial smoothness [35, Corollary 4.7], $\Psi \in \text{PSF}_{x^*}(\mathcal{M}_{x^*})$. Moreover, from Fact 7 and normal sharpness, the Riemannian Hessian of $\Psi$ at $x^*$ is such that, $\forall h \in T_{x^*},$

\[ \gamma \nabla_{\mathcal{M}_{x^*}}^2 \Psi(x^*) h = \gamma R_{T_{x^*}} \nabla^2 \tilde{R}(x^*) h + \gamma 2 \mathbb{M}_{x^*} (h, P_{T_{x^*}^\perp} \nabla \Phi(x^*)) \]

\[ = \gamma \nabla_{\mathcal{M}_{x^*}}^2 \Phi(x^*) P_{T_{x^*}} h - Hh = Uh, \]
Since $-\nabla F(x^*) \in \text{ri}(\partial R(x^*))$, we have from [36, Corollary 5.4] that

$$\partial^2 R(x^*) - \nabla F(x^*) h = \begin{cases} \nabla^2_{M_{x^*}} \Psi(x^*) h + T^*_h, & h \in T_{x^*}, \\ \emptyset, & h \notin T_{x^*}, \end{cases}$$

where $\partial^2 R(x^*) - \nabla F(x^*)$ denotes the Mordukhovich generalized Hessian mapping of function $R$ at $(x^*, -\nabla F(x^*)) \in \text{gph}(\partial R)$ [41]. As $R \in \Gamma_0(\mathbb{R}^n)$, $\partial R$ is a maximal monotone operator, and in view of [48, Theorem 2.1] we have that the mapping $\partial^2 R(x^*) - \nabla F(x^*)$ is positive semi-definite, whence we conclude that $\forall h \in T_{x^*},$

$$0 \leq \gamma(\partial^2 R(x^*) - \nabla F(x^*) h, h) = \gamma(\nabla^2_{M_{x^*}} \Psi(x^*) h, h) = \langle U h, h \rangle.$$

(ii) From Lemma 13, we have

$$\gamma(\partial^2 R(x^*) - \nabla F(x^*) h, h) = \gamma(\nabla^2_{M_{x^*}} \Psi(x^*) h, h) = \langle U h, h \rangle.$$

Thus, $\forall u \in \partial R(x^*)$ and $\forall v \in \partial R(x_t)$

$$0 \leq t^{-2} \langle u, x_t - x^* \rangle = t^{-1} \langle P_{T_{x^*}} v - P_{T_{x^*}} u, h \rangle$$

(by Fact 7) \begin{align*}
&= (t^{-1}(\nabla M_{x^*} R(x_t) - \nabla M_{x^*} R(x^*)) h) \\
&= (\nabla R(x^* + tP_{T_{x^*}} h) - \nabla R(x^*)) h).
\end{align*}

Since $\widetilde{R}$ is $C^2$, passing to the limit as $t \to 0$ leads to the desired result. \hfill \Box

Proof (Lemma 14).

(i) (a) is proved using the assumptions and Rademacher theorem. (b) and (c) follow from simple linear algebra arguments.

(ii) From Lemma 13, we have $WG = W^{1/2} W^{1/2} G W^{1/2} W^{-1/2}$, meaning that $WG$ is similar to $W^{1/2} G W^{1/2}$. The latter is symmetric and obeys

$$\|W^{1/2} G W^{1/2}\| \leq \|W^{1/2}\| \|G\| W^{1/2} \| < 1,$$

where we used (i)-(b) to get the last inequality. Thus $W^{1/2} G W^{1/2}$ has real eigenvalues in $]-1, 1[$, and so does $WG$ by similarity. The last statement follows using (i)-(c). \hfill \Box

We define the iteration-dependent versions of the matrices in (4.2), i.e.

$$(B.5) \quad H_k = \gamma_k P_{T_{x^*}} \nabla^2 F(x^*) P_{T_{x^*}}, \quad G_k = \text{Id} - H_k, \quad U_k = \gamma_k \nabla^2_{M_{x^*}} \Phi(x^*) P_{T_{x^*}} - H_k,$$

$M_{k,1} = [(1 + b)W(G_k - G) - bW(G_k - G)],$ $M_{k,2} = [(a_k - b_k - (a - b))W + (b_k - b)WG_k - ((a_k - b_k) - (a - b))W - (b_k - b)WG_k].$

After identification, we have $x_k \in M_{x^*}$ for $x_k$ close enough to $x^*$. Let $T_{x_k}$ be their corresponding tangent spaces, and define $r_k : T_{x_k} \to T_{x_k}$ the parallel translation along the unique geodesic joining from $x_k$ to $x^*$.

Before proving Proposition 15, we first establish the following useful estimates.

**Proposition 28. Under the assumptions of Proposition 15, we have**

$$\|y_{a_k} - x^*\| = O(\|d_k\|), \quad \|y_{b_k} - x^*\| = O(\|d_k\|), \quad \|r_{k+1}\| = O(\|d_k\|),$$

$$(B.6) \quad (\gamma_{k+1}^{-1} P_{T_{x^*} - T_{x^*}})(\nabla F(y_{b_k}) - \nabla F(x_{k+1})) = o(\|d_k\|).$$

$$(B.7) \quad \|W(U_k - U) r_{k+1}\| = o(\|d_k\|), \quad \|M_{k,1} d_k\| = o(\|d_k\|) \text{ and } \|M_{k,2} d_k\| = o(\|d_k\|).$$

**Proof.** We have

$$\|y_{a_k} - x^*\| = \|(1 + a_k) r_k - a_k r_{k-1}\| \leq (1 + a_k) \|r_k\| + a_k \|r_{k-1}\|$$

$$\leq (1 + a_k)(\|r_k\| + \|r_{k-1}\|) \leq \sqrt{2}(1 + a_k) \|d_k\|,$$
whence we get the first and second estimates. In turn, we obtain
\[
\|r_{k+1}\| = \|\text{prox}_{\gamma_k R}(y_{a,k} - \gamma_k \nabla F(y_{b,k})) - \text{prox}_{\gamma_{k+1}}(x^* - \gamma_k \nabla F(x^*))\|
\leq \|(y_{a,k} - x^*) - \gamma_k (\nabla F(y_{b,k}) - \nabla F(x^*))\|
\leq (1 + a_k)\|r_k\| + a_k\|r_{k-1}\| + (1 + b_k)\frac{2\gamma_k}{\beta}\|r_k\| + \frac{b_k\gamma_k}{\beta}\|r_{k-1}\|
\leq (1 + a_k) + (1 + b_k)\frac{2\gamma_k}{\beta}\sqrt{2}\|d_k\|,
\]
where we used non-expansiveness of the proximity operator and assumption (H.2).
This yields the third estimate. Combining Lemma 20, assumption (H.2), (B.8) and (B.9), we get
\[
(\tau_{k+1}^{-1}P_{T_{k+1}} - P_{T_x})(\nabla F(y_{b,k}) - \nabla F(x_{k+1})) = o(\|\nabla F(y_{b,k}) - \nabla F(x_{k+1})\|)
= o(\|y_{b,k} - x^*\|) + o(\|r_{k+1}\|) = o(\|d_k\|).
\]
For (B.7), recall the function \(\Psi\) in the proof of Lemma 13(i). First, we have
\[
\lim_{k \to \infty} \|W(U_k - U)r_{k+1}\|/\|r_{k+1}\| = \lim_{k \to \infty} \|W(\gamma_k - \gamma)\nabla M_x, \Psi(x^*)P_{T_x}, r_{k+1}/\|r_{k+1}\|
\leq \lim_{k \to \infty} |\gamma_k - \gamma|\|\nabla M_x, \Psi(x^*)P_{T_x}\| = 0,
\]
which entails \(\|W(U_k - U)r_{k+1}\| = o(\|r_{k+1}\|) = o(\|d_k\|)\). Again, since \(\gamma_k \to \gamma\),
\[
\lim_{k \to \infty} \|M_{k,2}d_k\|/\|d_k\| \leq \lim_{k \to \infty} (1 + b)\|W\|\|G_k - G\|\|r_k\| + \|r_{k-1}\|)/\|d_k\|
\leq \lim_{k \to \infty} (1 + b)\|\gamma_k - \gamma\|\|P_{T_x}, \nabla^2 F(x^*)P_{T_x}\|\|d_k\|/\|d_k\| = 0.
\]
Similarly, for \(M_{k,2}\), since \(a_k \to a, b_k \to b\),
\[
\lim_{k \to \infty} \|M_{k,2}d_k\|/\|d_k\| \leq \lim_{k \to \infty} (|a_k - a| + |b_k - b|)\|W\|\|G_k\|\sqrt{2}\|d_k\|/\|d_k\| = 0,
\]
where \(W, G\) are bounded.

\textit{Proof (Proposition 15). (1.3) and the first-order optimality condition for problem (Popt) are respectively equivalent to
\[y_{a,k} - x_{k+1} - \gamma_k (\nabla F(y_{b,k}) - \nabla F(x_{k+1})) \in \gamma_k \partial \Phi(x_{k+1}) \text{ and } 0 \in \gamma_k \partial \Phi(x^*).
\]
Projecting into \(T_{x_{k+1}}\) and \(T_x\), respectively, and using Fact 7, leads to
\[
\gamma_k \nabla M_x, \Phi(x_{k+1}) = \tau_{k+1}^{-1}P_{T_{x_{k+1}}} \gamma_k \nabla F(y_{b,k}) - \nabla\Phi(x_{k+1})
\gamma_k \nabla M_x, \Phi(x^*) = 0.
\]
Adding both identities, and subtracting \(\tau_{k+1}^{-1}P_{T_{x_{k+1}}} x^*\) on both sides, we arrive at
\[
\tau_{k+1}^{-1}P_{T_{x_{k+1}}}, T_{x_{k+1}}(y_{a,k} - x_{k+1} - \gamma_k (\nabla F(y_{b,k}) - \nabla F(x_{k+1}))) = \tau_{k+1}^{-1}P_{T_{x_{k+1}}} \gamma_k (\nabla F(y_{b,k}) - \nabla F(x_{k+1})).
\]
In view of Lemma 26, we get
\[
\tau_{k+1}^{-1}P_{T_{x_{k+1}}}, T_{x_{k+1}}(y_{a,k} - x_{k+1} - \gamma_k (\nabla F(y_{b,k}) - \nabla F(x_{k+1}))) = \tau_{k+1}^{-1}P_{T_{x_{k+1}}} (y_{a,k} - x_{k+1} + o(\|r_{k+1}\|)).
\]
Using [37, Lemma 5.1], we have
\[
r_{k+1} = P_{T_x}, r_{k+1} + o(\|r_{k+1}\|)
\Rightarrow \tau_{k+1}^{-1}P_{T_{x_{k+1}}}, T_{x_{k+1}}(y_{a,k} - x_{k+1} + o(\|r_{k+1}\|)) = r_{k+1} + o(\|d_k\|),
\]
where we also used (B.6). Similarly
\[
\tau_{k+1}^{-1}P_{T_{x_{k+1}}} (y_{a,k} - x^*) = P_{T_x}, (y_{a,k} - x^*) + (\tau_{k+1}^{-1}P_{T_{x_{k+1}}} - P_{T_x})(y_{a,k} - x^*)
= P_{T_x}, (y_{a,k} - x^*) + o(\|y_{a,k} - x^*\|)
= P_{T_x}, (y_{a,k} - x^*) + o(\|d_k\|)
= (1 + a_k)P_{T_x}, r_{k+1} - a_k P_{T_x}, r_{k-1} + o(\|d_k\|)
= (1 + a_k) r_{k+1} - a_k r_{k-1} + o(\|r_k\|) + o(\|r_{k-1}\|) + o(\|d_k\|)
= (y_{a,k} - x^*) + o(\|d_k\|).
\]
Moreover owing to Lemma 27 and (B.6),
\[
\tau^{-1} \nabla_{\mathcal{M}_*} \Phi(x_{k+1}) - \nabla_{\mathcal{M}_*} \Phi(x^*) = \nabla_{\mathcal{M}_*}^2 \Phi(x^*) P_{r_{k+1}} + o(||r_{k+1}||)
\]
Therefore, inserting (B.11), (B.12) and (B.13) into (B.10), we obtain
\[
(Id + \gamma_r \nabla_{\mathcal{M}_*}^2 \Phi(x^*) P_{r_{k+1}}) r_{k+1} = (y_{a,k} - x^*) - \gamma_r \tau^{-1}_{k+1} P_{r_{k+1}} \left( \nabla F(y_{b,k}) - \nabla F(x_{k+1}) \right) + o(||d_k||).
\]
Owing to (B.6) and local $C^2$-smoothness of $F$, we have
\[
\tau_{k+1}^{-1} P_{r_{k+1}} \left( \nabla F(y_{b,k}) - \nabla F(x_{k+1}) \right)
\]
\[
(Id + \gamma_r \nabla_{\mathcal{M}_*}^2 \Phi(x^*) P_{r_{k+1}} - \gamma_r P_{r_{k+1}} \nabla^2 F(x^* P_{r_{k+1}}) r_{k+1}
\]
which can be further written as,
\[
(Id + U_k) r_{k+1} = (y_{a,k} - x^*) - H_k(y_{b,k} - x^*) + o(||d_k||)
\]
Inverting $Id + U$ (which is possible thanks to Lemma 13), we obtain
\[
r_{k+1} = W(U_k - U) r_{k+1} = [(a_k - b_k)W + (1 + b_k)WG_k - (a_k - b_k)W - b_kWG_k] d_k + o(||d_k||).
\]
Using the estimates (B.7), we get
\[
d_{k+1} = \left[ \begin{array}{c} (a_k - b_k)W + (1 + b_k)WG_k - (a_k - b_k)W - b_kWG_k \end{array} \right] d_k + o(||d_k||)
\]
\[
= \left( M + \begin{bmatrix} M_{k,1} \\ 0 \end{bmatrix} \right) d_k + o(||d_k||) = Md_k + o(||d_k||)
\]
Proof (Proposition 17).
(i) We have
\[
M \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} = \begin{bmatrix} (a-b)r_1 + (1+b)Gr_1 - (a-b)r_2 - bGr_2 \\ r_2 \end{bmatrix} = \sigma \begin{bmatrix} r_1 \\ r_2 \end{bmatrix},
\]
and thus $r_1 = \sigma r_2$. Inserting this in the first identity, we obtain
\[
\sigma^2 r_2 = (a-b)\sigma r_2 + (1+b)\sigma Gr_2 - (a-b)r_2 - bGr_2
\]
\[
\Rightarrow Gr_2 = \frac{((a-b)(1-\sigma) + \sigma^2)/((1+b)\sigma - b)}{r_2} = \eta r_2
\]
\[
\Rightarrow 0 = \sigma^2 - ((a-b) + (1+b)\eta)\sigma + (a-b) + b\eta.
\]
(ii) For this quadratic equation of $\sigma$, the two roots are
\[
\sigma_1 = ((a-b) + (1+b)\eta + \sqrt{\Delta_\sigma})/2, \quad \sigma_2 = ((a-b) + (1+b)\eta - \sqrt{\Delta_\sigma})/2.
\]
where $\Delta_\sigma = ((a - b) + (1 + b)\eta)^2 - 4((a - b) + b\eta)$ is the discriminant, which is a quadratic polynomial of three variables. Consider the following three linear functions of $a$

(B.18) \[ a_1 = (1 - \eta)b - \eta, \quad a_3 = (1 - \eta)b - (1 + \eta)/2 \]

(B.19) \[ a_2 = (1 - \eta)b + (1 - \sqrt{1 - \eta})^2 \Delta_\sigma \leq 0: a \in [a_2, (1 - \eta)b + (1 + \sqrt{1 - \eta})^2], \]

\[ \Delta_\sigma \geq 0: a \leq a_2. \]

Recall from Lemma 14(i) that $\eta \in [-1, 1]$. Thus, $a_1 \geq a_2$ when $\eta \in [-1, 0]$, $a_1 \leq a_2$ for $\eta \in [0, 1]$, and $a_3$ is smaller than both $a_1, a_2$ independently of $\eta$.

**Case $\eta \in [-1, 0]$:** We have $a_1 \geq a_2$.

**Subcase $a \in [a_2, 1]$: $\sigma_1, \sigma_2$ are complex, hence**

(B.19) \[ |\sigma|^2 = (((a - b) + (1 + b)\eta)^2 - \Delta_\sigma)/4 = a - b + b\eta. \]

As $a_2 \leq 1 \implies b \leq \frac{1-(1-\sqrt{1-\eta})^2}{1-\eta}$, then $(1-\sqrt{1-\eta})^2 \leq |\sigma|^2 \leq 1 + (\eta - 1)b < 1.$

**Subcase $a \in [0, a_2]$; $\Delta_\sigma \geq 0$ and $\sigma_2$ has the bigger absolute value, then**

(B.20) \[ |\sigma_2| < 1 \iff -(a - b) + (1 + b)\eta + \sqrt{\Delta_\sigma} < 2 \iff 2(b - a) - 1 < \eta, \]

which means $|\sigma_2| < 1$ for $a \in [a_3, a_2]$, and $|\sigma_2| > 1$ for $a \in [0, a_3]$. Moreover, $a_3 < 0$ for $b \in [0, \frac{1+\eta}{2(1-\eta)}]$, meaning that if $\eta \geq \frac{1}{2}$, $|\sigma_2| \leq 1$ for $a \in [0, a_2]$.

**Case $\eta \in [0, 1]$;** First we have $a_2 \geq a_1$, and moreover

\[ a_1 = 0 \iff b = \frac{\eta}{1-\eta} \leq 1: \eta \in [0, 0.5], \]

\[ \geq 1: \eta \in [0.5, 1]. \]

Obviously, we have $|\sigma| \leq 1$ holds for any $a \in [0, a_2]$ as long as $\eta \in [0.5, 1]$, though this situation is useless as $b \in [0, 1]$. In the subcases hereafter, we only consider $\eta \in [0, 0.5]$.

**Subcase $a \in [a_2, 1]$; same result as (B.19).**

**Subcase $a \in [a_1, a_2]$; $\sigma_1 \geq |\sigma_2|$, hence**

(B.21) \[ a_1 < 1 \iff (a - b) + (1 + b)\eta + \sqrt{\Delta_\sigma} < 2 \iff 0 < 4(1 - \eta). \]

**Subcase $a \in [0, a_1]$;** we have $|\sigma_2| \geq |\sigma_1|$, hence (B.20) applies and the result follows.

Summarizing this discussion yields the claimed result.

**Proof (Theorem 25).** Since $R$ is locally polyhedral, we have $\nabla_{M_*} \Phi(x_k)$ is locally constant along $M_* = x^* + T_{x^*}$, around $x^*$ (see Remark 9(iii)). Thus, embarking from (B.16) in the proof of Proposition 15, for $k$ large enough, we get

(B.22) \[ x_{k+1} - x^* = (y_{a,k} - x^*) - E_k(y_{b,k} - x^*), \]

where we used the mean-value theorem with $E_k = \gamma_k \int_0^1 \nabla^2 F(x^* + t(y_{b,k} - x^*))dt \geq 0$.

Using that $E_k$ is symmetric and $\text{Im}(E_k)^{-1} = V$, we have

\[ P_V(x_{k+1} - x^*) = P_V(y_{a,k} - x^*) = (1 + a_k)P_V(x_k - x^*) - a_k(x_{k-1} - x^*). \]

If $a_k = 0$, then $P_V(x_{k+1} - x^*) = P_V(x_k - x^*)$. Thus, in the rest, without loss of generality, we assume that $a_k > 0$ for $k$ large enough. The above iteration leads to

\[ \left( P_V(x_{k+1} - x^*) \right) = \left[ (1 + a_k)I_d \quad -a_kI_d \right] \left( P_V(x_k - x^*) \right) \]

It is straightforward to check that $N_k \overset{\text{def}}{=} \left[ (1 + a_k)I_d \quad -a_kI_d \right]$ is invertible and admits two eigenvalues $a_k > 0$ and 1 respectively. Iterating the above argument, and owing
to the fact that \( x_k, y_{a,k}, y_{b,k} \to x^* \), we get
\[
\binom{0}{0} = \left( \prod_{j=k}^{\infty} N_j \right) \left( \begin{array}{c} P_V(x_k - x^*) \\ P_V(x_{k-1} - x^*) \end{array} \right),
\]
and \( \prod_{j=k}^{\infty} N_j \) is invertible. Therefore, we obtain that \( x_k - x^* \in V^\perp \), and in turn, \( y_{a,k} - x^* \in V^\perp \) and \( y_{b,k} - x^* \in V^\perp \), for all large enough \( k \). Observe that \( V^\perp \subset T_{x^*} \), it then follows that
\[
x_{k+1} = y_{a,k} - x^* - P_{V^\perp} E_k P_{V^\perp} (y_{b,k} - x^*).
\]
By definition, \( P_{V^\perp} E_k P_{V^\perp} \) is symmetric positive definite. Thus, substituting this matrix for \( H_k \), and \( G \) and \( M \) accordingly in Lemma 14 and Corollary 19, and applying Theorem 21, leads to the result. \( \square \)

REFERENCES


This manuscript is for review purposes only.


