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

JINGWEI LIANG†, JALAL FADILI†, AND GABRIEL PEYRÉ‡

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, the fused 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

$$(\mathcal{P}_{\mathrm{opt}}) \qquad \min_{x\in\mathbb{R}^n} \ \Phi(x) \stackrel{\mathrm{def}}{=} F(x) + R(x),$$

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 ∇F is $(1/\beta)$-Lipschitz continuous.
- **(H.3)** $\mathrm{Argmin}(\Phi) \neq \emptyset$, *i.e.* the set of minimizers is non-empty.

From now on, we suppose that assumptions (**H.1**)-(**H.3**) hold. Problem ($\mathcal{P}_{\mathrm{opt}}$) is closely related to finding solutions of the *monotone inclusion problem*

$$(\mathcal{P}_{\mathrm{inc}}) \qquad \text{Find } x \in \mathbb{R}^n \text{ such that } 0 \in A(x) + B(x),$$

where

- **(H.4)** $A : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a set-valued maximal monotone operator (see (A.1)).
- **(H.5)** $B : \mathbb{R}^n \to \mathbb{R}^n$ is maximal monotone and β-cocoercive (see (A.2)).
- **(H.6)** $\mathrm{zer}(A + B) \neq \emptyset$, *i.e.* the set of zeros of $A + B$ is non-empty.

For problem ($\mathcal{P}_{\mathrm{opt}}$), given a global minimizer $x^\star \in \mathrm{Argmin}(\Phi)$, then the corresponding first-order optimality condition reads

$$0 \in \partial R(x^\star) + \nabla F(x^\star),$$

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

*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, École 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

$$x_{k+1} = \text{prox}_{\gamma_k R}\big(x_k - \gamma_k \nabla F(x_k)\big),\ \gamma_k \in [\underline{\epsilon}, 2\beta - \overline{\epsilon}], \tag{1.1}$$

where $\underline{\epsilon}, \overline{\epsilon} > 0$, and $\text{prox}_{\gamma R}$ denotes the *proximity operator* of R which is defined as

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

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 $a_k = b_k = \frac{k-1}{k+q}$, $q > 2$), 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: $\bar{a} \le 1$, $\bar{b} \le 1$, $\underline{\epsilon}, \overline{\epsilon} > 0$ such that $\underline{\epsilon} \le 2\beta - \overline{\epsilon}$. $x_0 \in \mathbb{R}^n$, $x_{-1} = x_0$.

Let $a_k \in [0, \bar{a}]$, $b_k \in [0, \bar{b}]$, $\gamma_k \in [\underline{\epsilon}, 2\beta - \overline{\epsilon}]$. Repeat

$$y_{a,k} = x_k + a_k(x_k - x_{k-1}),\ \ y_{b,k} = x_k + b_k(x_k - x_{k-1}), \tag{1.2}$$

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

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

[49];
- $a_k \in [0, \bar{a}]$, $b_k = a_k$: this corresponds to the work of [39] for solving ($\mathcal{P}_{\mathrm{inc}}$). If moreover restrict $\gamma_k \in]0, \beta]$ and let $a_k \to 1$, then Algorithm 1 specializes to FISTA-type methods [14, 18, 9, 7] developed for optimization.

When a_k, b_k satisfy $a_k \in [0, \bar{a}]$, $b_k \in]0, \bar{b}]$, $a_k \neq b_k$, Algorithm 1 is new in the literature to the best of our knowledge.

REMARK 1. *Though Algorithm 1 is stated for the optimization problem* ($\mathcal{P}_{\mathrm{opt}}$)*, it readily extends to the monotone inclusion problem* ($\mathcal{P}_{\mathrm{inc}}$)*, for which step* (1.3) *reads*

$$x_{k+1} = J_{\gamma_k A}\big(y_{a,k} - \gamma_k B(y_{b,k})\big), \tag{1.4}$$

where $J_{\gamma A} \stackrel{\mathrm{def}}{=} (\mathrm{Id} + \gamma A)^{-1}$ *denotes the resolvent of* γA.

For the rest of the paper, we use the terminology *FB-type methods* for any scheme in the form of Algorithm 1 such that the sequence $(x_k)_{k\in\mathbb{N}}$ converges. This will encompass the inertial schemes (denoted iFB) that we propose, and the sequence convergent FISTA method [18, 9] that corresponds to the specific choice of intertial sequences $a_k = b_k = \frac{k-1}{k+q}$, $q > 2$. It should be noted, however, that our global convergence analysis to be presented in Section 2 does not cover the case of FISTA, which requires a specific proof strategy as developed in [18, 9].

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

Local linear convergence. Exploiting this identification property, we then show that the FB-type methods, locally along the manifold $\mathcal{M}_{x^\star}$, exhibit a linear convergence regime. We characterize this regime and the corresponding rates precisely depending on the structure of the active manifold $\mathcal{M}_{x^\star}$. 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 identification happens, the globally non-smooth convex problem ($\mathcal{P}_{\mathrm{opt}}$) becomes (locally) equivalent to a C^2-smooth one along the (possibly non-convex) active manifold $\mathcal{M}_{x^\star}$. 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 ($\mathcal{P}_{\mathrm{opt}}$) where R is the ℓ_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 ($\mathcal{P}_{\mathrm{opt}}$) 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, Id denotes the identity operator on $\mathbb{R}^n$. For a nonempty convex set $\Omega \subset \mathbb{R}^n$, $\mathrm{ri}(\Omega)$ and $\mathrm{rbd}(\Omega)$ denote its relative interior and boundary respectively, $\mathrm{aff}(\Omega)$ is its affine hull, and $\mathrm{par}(\Omega) = \mathbb{R}(\Omega - \Omega)$ is the subspace parallel to it. Denote ι_Ω the indicator function of Ω, σ_Ω its support function and P_Ω the orthogonal projector onto Ω. For a matrix M, $\ker(M)$ 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 convergence 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 identification 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 dimensional optimization problem ($\mathcal{P}_{\mathrm{opt}}$). In fact, our global convergence results can handle the more general monotone inclusion problem ($\mathcal{P}_{\mathrm{inc}}$) 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. Toward this goal, we recall the notion of ε-enlargement.

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

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

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

$$y_{a,k} - \gamma_k(\nabla F(y_{b,k}) + \xi_k) - x_{k+1} \in \gamma_k \partial^{\varepsilon_k} R(x_{k+1}), \tag{2.1}$$

where $\xi_k \in \mathbb{R}^n$ is the error in the evaluation of the gradient operator ∇F. Observe that since the ε-approximate subdifferential of a proper closed convex function is contained in the ε-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 $\bar{a} < 1$, $\sum_{k\in\mathbb{N}} \varepsilon_k < +\infty$ 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*

$$\textstyle\sum_{k\in\mathbb{N}} \max\{a_k, b_k\} \|x_k - x_{k-1}\|^2 < +\infty, \tag{2.2}$$

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

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.* $\gamma_k \in]0, \beta]$, $a_k = b_k = \frac{k-1}{k+q}$, $q > 2$.

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 $a \in [0, 1], b \in [0, 1]$,

$$a_k = \min\{a, c_{a,k}\}, \quad b_k = \min\{b, c_{b,k}\}, \tag{2.3}$$

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 $c_{a,k} = \frac{c_a}{k^{1+\delta}\|x_k - x_{k-1}\|^2}$, $c_a > 0, \delta > 0$ 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,*

$$\begin{cases} (1 + a_k) - \frac{\gamma_k}{2\beta}(1 + b_k)^2 > \tau : a_k < \frac{\gamma_k}{2\beta} b_k, \\ (1 - 3a_k) - \frac{\gamma_k}{2\beta}(1 - b_k)^2 > \tau : b_k \le a_k \ \text{ or } \ \frac{\gamma_k}{2\beta} b_k \le a_k < b_k, \end{cases} \tag{2.4}$$

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 $x^\star \in \mathrm{Argmin}(\Phi)$ such that the sequence $(x_k)_{k\in\mathbb{N}}$ converges to $x^\star$.

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 ($\mathcal{P}_{\mathrm{opt}}$) 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 $T_x \stackrel{\mathrm{def}}{=} \mathrm{par}(\partial R(x))^\perp$;*

(iii) ***Continuity**: The set-valued mapping ∂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 $\mathrm{PSF}_x(\mathcal{M}_x)$.

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 subdifferential 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.

[35, Proposition 2.10] allows to prove the following fact.

FACT 6 (Local normal sharpness). *If $R \in \mathrm{PSF}_x(\mathcal{M}_x)$, 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 C^2 submanifold. This is summarized in the following fact which follows by combining (B.2), (B.3), Definition 5, Fact 6 and [23, Proposition 17] (or [40, Lemma 2.4]).

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

$$\nabla_{\mathcal{M}_x} R(x') = \mathrm{P}_{T_{x'}}(\partial R(x')),$$

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

$$\nabla^2_{\mathcal{M}_x} G(x')h = \mathrm{P}_{T_{x'}} \nabla^2 \widetilde{R}(x')h + \mathfrak{W}_{x'}\big(h, \mathrm{P}_{T_{x'}^\perp} \nabla \widetilde{R}(x')\big),$$

where $\widetilde{R}$ is a smooth extension (representative) of R on $\mathcal{M}_x$, and $\mathfrak{W}_x(\cdot,\cdot) : T_x \times T_x^\perp \to T_x$ 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 $x^\star \in \mathrm{Argmin}(\Phi)$ such that $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}_{x^\star})$, and moreover the non-degeneracy condition*

$$\text{(ND)} \qquad -\nabla F(x^\star) \in \mathrm{ri}\big(\partial R(x^\star)\big),$$

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

If moreover,

(i) *$\mathcal{M}_{x^\star}$ is an affine subspace, then $\mathcal{M}_{x^\star} = x^\star + T_{x^\star}$ and $y_{a,k}, y_{b,k} \in \mathcal{M}_{x^\star}, \forall k > K$;*
(ii) *R is locally polyhedral around $x^\star$, then $y_{a,k}, y_{b,k} \in \mathcal{M}_{x^\star} = x^\star + T_{x^\star}$ for all $k > K$, $\nabla_{\mathcal{M}_{x^\star}} R(x_k) = \nabla_{\mathcal{M}_{x^\star}} R(x^\star)$, and $\nabla^2_{\mathcal{M}_{x^\star}} R(x_k) = 0$, $\forall k \geq K$.*

REMARK 9.

(i) *If F is also locally C^2 around $x^\star$, the smooth perturbation rule of partly smooth functions [35, Corollary 4.7], ensures that $\Phi \in \mathrm{PSF}_{x^\star}(\mathcal{M}_{x^\star})$.*
(ii) *The iFB is convergent under the assumptions of Theorem 3 or Theorem 4. The FISTA method is sequence convergent for $a_k = b_k = \frac{k-1}{k+q}$, $q > 2$, and $\gamma_k \equiv \gamma \in]0, \beta]$ [18, 9]. Thus, Theorem 8 holds true for all these instances.*
(iii) *The non-degeneracy condition (ND) can be viewed as a geometric generalization of the* strict complementarity *of non-linear programming. Building on the arguments of [29], it is almost a necessary condition for the finite identification of $\mathcal{M}_{x^\star}$. Relaxing it in general is a challenging problem.*
(iv) *When R is locally polyhedral around $x^\star$, in addition with the finite identification of $\mathcal{M}_{x^\star} = x^\star + T_{x^\star}$, we also have $\nabla_{\mathcal{M}_{x^\star}} \Phi(x_k) = \nabla_{\mathcal{M}_{x^\star}} \Phi(x^\star)$, hence $\nabla^2_{\mathcal{M}_{x^\star}} \Phi(x_k) = 0$, for k large enough.*

Proof. By assumption, the sequence $(x_k)_{k\in\mathbb{N}}$ created by any FB-type method converges to some $x^\star \in \mathrm{Argmin}(\Phi)$, 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

$$\begin{aligned}
&\mathrm{dist}\big(-\nabla F(x^\star), \partial R(x_{k+1})\big) \\
&\leq \|\frac{1}{\gamma_k}\big(y_{a,k} - x_{k+1}\big) - \nabla F(y_{b,k}) + \nabla F(x^\star)\| \\
&\leq \frac{1}{\gamma_k}\big(a_k\|x_k - x_{k-1}\| + \|x_{k+1} - x_k\|\big) + \|\nabla F(y_{b,k}) - \nabla F(x^\star)\| \\
&\leq \big(\frac{1}{\gamma_k} + \frac{1}{\beta}\big)\|x_k - x_{k-1}\| + \frac{1}{\gamma_k}\|x_{k+1} - x_k\| + \frac{1}{\beta}\|x_k - x^\star\|.
\end{aligned}$$

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

(i) When the active manifold $\mathcal{M}_{x^\star}$ is an affine subspace, then $\mathcal{M}_{x^\star} = x^\star + T_{x^\star}$ owing to the normal sharpness property and the claim follows immediately;
(ii) When R is locally polyhedral around $x^\star$, then $\mathcal{M}_{x^\star}$ 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 $x \in \mathcal{M}_{x^\star}$ near $x^\star$, $\partial R(x) = \partial R(x^\star)$. Therefore, combining Fact 6 and Fact 7, we get the second conclusion. ∎

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, *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 (*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 $\partial R(x_k) \subset \mathrm{rbd}(\partial R(x^\star))$ *whenever* $x_k \notin \mathcal{M}_{x^\star}$*, then* $x_k \in \mathcal{M}_{x^\star}$ *for all* $k \geq \frac{\|x_0 - x^\star\|^2}{\underline{\epsilon}^2 \mathrm{dist}\big(-\nabla F(x^\star), \mathrm{rbd}(\partial R(x^\star))\big)^2}$;

(ii) If R *is separable, i.e.* $R(x) = \sum_{i=1}^m \sigma_{C_i}(x_{b_i})$*, 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* $\mathcal{M}_{x^\star}$ *occurs for some* k *larger than* $\frac{\|x_0 - x^\star\|^2}{\underline{\epsilon}^2 \sum_{i \in I^c_{x^\star}} \mathrm{dist}\big(-\nabla F(x^\star)_{b_i}, \mathrm{rbd}(C_i)\big)^2}$*, where* $I_x \stackrel{\mathrm{def}}{=} \big\{i : x_{b_i} \neq 0\big\}$.

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

$$\begin{aligned}\|x_k - x^\star\|^2 &\leq \|(\mathrm{Id} - \gamma_{k-1}\nabla F)(x_{k-1}) - (\mathrm{Id} - \gamma_{k-1}\nabla F)(x^\star)\|^2 \\ &\quad - \|x_{k-1} - \gamma_{k-1}\nabla F(x_{k-1}) - x_k + \gamma_{k-1}\nabla F(x^\star)\|^2 \\ &\leq \|x_{k-1} - x^\star\|^2 - \underline{\epsilon}^2\|u_k - \nabla F(x^\star)\|^2,\end{aligned}$$

where we denoted $u_k \stackrel{\mathrm{def}}{=} (x_{k-1} - x_k)/\gamma_{k-1} - \nabla F(x_{k-1})$. By definition, we have $u_k \in \partial R(x_k)$. Suppose that identification has not occurred at k, *i.e.* that $x_k \notin \mathcal{M}_{x^\star}$, and hence $u_k \in \partial R(x_k) \subset \mathrm{rbd}(\partial R(x^\star))$. Therefore, continuing the above inequality, we get

$$\begin{aligned}\|x_k - x^\star\|^2 &\leq \|x_{k-1} - x^\star\|^2 - \underline{\epsilon}^2 \mathrm{dist}\big(-\nabla F(x^\star), \partial R(x_k)\big)^2 \\ &\leq \|x_{k-1} - x^\star\|^2 - \underline{\epsilon}^2 \mathrm{dist}\big(-\nabla F(x^\star), \mathrm{rbd}(\partial R(x^\star))\big)^2 \\ &\leq \|x_0 - x^\star\|^2 - k\underline{\epsilon}^2 \mathrm{dist}\big(-\nabla F(x^\star), \mathrm{rbd}(\partial R(x^\star))\big)^2,\end{aligned}$$

and $\mathrm{dist}(-\nabla F(x^\star), \mathrm{rbd}(\partial R(x^\star))) > 0$ 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 $\partial \sigma_{C_i}(x^\star_{b_i}) = C_i, \forall i \in I^c_{x^\star}$. In turn, by separability, R is partly smooth at $x^\star$ relative to $\mathcal{M}_{x^\star} = \bigtimes_{i=1}^m \mathcal{M}_{x^\star_{b_i}}$, where $\mathcal{M}_{x^\star_{b_i}} = 0$ if $i \in I^c_{x^\star}$ and $\mathcal{M}_{x^\star_{b_i}} \neq 0$ otherwise. Suppose that at iteration k, $I^c_{x^\star} \cap I_{x_k} \neq \emptyset$. Denote $h_{k-1} = x_{k-1} - \gamma_{k-1}\nabla F(x_{k-1})$, and $h^\star = x^\star - \gamma_{k-1}\nabla F(x^\star)$. Thus for any $i \in I^c_{x^\star} \cap I_{x_k}$, we have

$$\begin{aligned}x_{k,b_i} - x^\star_{b_i} &= h_{k-1,b_i} - \mathrm{P}_{\gamma_{k-1}C_i}(h_{k-1,b_i}) \\ &= (h_{k-1,b_i} - h^\star_{b_i}) - (\mathrm{P}_{\gamma_{k-1}C_i}(h_{k-1,b_i}) - \mathrm{P}_{\gamma_{k-1}C_i}(h^\star_{b_i}))\end{aligned}$$

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

$$\begin{aligned}\|x_{k,b_i} - x^\star_{b_i}\|^2 &\leq \|h_{k-1,b_i} - h^\star_{b_i}\|^2 - \|\mathrm{P}_{\gamma_{k-1}C_i}(h_{k-1,b_i}) - h^\star_{b_i}\|^2 \\ &= \|h_{k-1,b_i} - h^\star_{b_i}\|^2 - \|\mathrm{P}_{\gamma_{k-1}C_i}(h_{k-1,b_i}) + \gamma_{k-1}\nabla F(x^\star)_{b_i}\|^2 \\ &\leq \|h_{k-1,b_i} - h^\star_{b_i}\|^2 - \gamma^2_{k-1}\mathrm{dist}\big(-\nabla F(x^\star)_{b_i}, \mathrm{rbd}(C_i)\big)^2 \\ &\leq \|h_{k-1,b_i} - h^\star_{b_i}\|^2 - \underline{\epsilon}^2\mathrm{dist}\big(-\nabla F(x^\star)_{b_i}, \mathrm{rbd}(C_i)\big)^2.\end{aligned}$$

This bound together with non-expansiveness of $\mathrm{prox}_{\gamma_{k-1}C_i}$ and $\mathrm{Id}-\gamma_{k-1}\nabla F$ yield

$$
\begin{aligned}
\|x_k - x^\star\|^2 &= \textstyle\sum_{i\in I^c_{x^\star}} \|x_{k,b_i} - x^\star_{b_i}\|^2 + \sum_{j\in I_{x^\star}} \|x_{k,b_j} - x^\star_{b_j}\|^2 \\
&\le \|h_{k-1} - h^\star\|^2 - \underline{\epsilon}^2 \textstyle\sum_{i\in I^c_{x^\star}} \mathrm{dist}\big(-\nabla F(x^\star)_{b_i}, \mathrm{rbd}(C_i)\big)^2 \\
&\le \|x_{k-1} - x^\star\|^2 - \underline{\epsilon}^2 \textstyle\sum_{i\in I^c_{x^\star}} \mathrm{dist}\big(-\nabla F(x^\star)_{b_i}, \mathrm{rbd}(C_i)\big)^2 \\
&\le \|x_0 - x^\star\|^2 - k\underline{\epsilon}^2 \textstyle\sum_{i\in I^c_{x^\star}} \mathrm{dist}\big(-\nabla F(x^\star)_{b_i}, \mathrm{rbd}(C_i)\big)^2,
\end{aligned}
$$

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 $I^c_{x^\star} \cap I_{x_k} \neq \emptyset$ 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 $\mathcal{M}_{x^\star_{b_i}} = 0$. 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 $x^\star$, it is only of theoretical interest. In the separable case, observe that $\sum_{i\in I^c_{x^\star}} \mathrm{dist}(-\nabla F(x^\star)_{b_i}, \mathrm{rbd}(C_i))^2 = \mathrm{dist}(-\nabla F(x^\star), \partial R(x^\star))^2$ when σ_{C_i} is differentiable at $x^\star_{b_i}$ for all $i \in I_{x^\star}$. The case of the ℓ_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 $(\xi_k)_{k\in\mathbb{N}}$ is such that $(x_k)_{k\in\mathbb{N}}$ converges to some $x^\star \in \mathrm{Argmin}(\Phi)$ (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 ε_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 $R + \iota_{\mathcal{M}_{x^\star}}$ locally agree nearby $x^\star$. This property is clearly violated if approximate proximal mappings are involved. Here is a simple example.

EXAMPLE 11. *Let $F : x \in \mathbb{R} \mapsto \frac{1}{2}|\delta - x|^2$, with $\delta \in]-1, 1[$, and $R : x \in \mathbb{R} \mapsto |x|$. $\Phi \in \Gamma_0(\mathbb{R})$ and has a unique minimizer $x^\star = \mathrm{prox}_{|\cdot|}(\delta) = 0$. Moreover, Φ is partly smooth at $x^\star$ relative to $\mathcal{M}_{x^\star} = \{0\}$, and $\delta - x^\star = \delta \in \mathrm{ri}\big(\partial R(x^\star)\big) =]-1, 1[$. Consider the inexact version of the FB algorithm*

$$x_{k+1} \in (\mathrm{Id} + \partial^{\varepsilon_k}|\cdot|)^{-1}(\delta), \tag{3.1}$$

where we set $\gamma_k \equiv 1$, since ∇F is 1-Lipschitz. From [17, Example 5.2.5], we have

$$
\partial^\varepsilon|\cdot|(x) = \begin{cases} [1 - \varepsilon/x, 1] & \text{if } x > \varepsilon/2 \\ [-1, 1] & \text{if } |x| \le \varepsilon/2 \\ [-1, -1 - \varepsilon/x] & \text{if } x < -\varepsilon/2, \end{cases}
$$

whence the graph of $(\mathrm{Id} + \partial^\varepsilon|\cdot|)^{-1}$, a set-valued operator, can be easily deduced. Thus, depending on ε_k and the choice made in the inclusion (3.1), x_k may never vanish, i.e. $x_k \notin \mathcal{M}_{x^\star}$, 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, $x^\star$ is a global minimizer of problem ($\mathcal{P}_{\mathrm{opt}}$) to which the sequence $(x_k)_{k\in\mathbb{N}}$ provided by the FB-type method converges. $\mathcal{M}_{x^\star}$ is the partial smoothness manifold of R at $x^\star$, and $T_{x^\star}$ 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 $x^\star$, and its Hessian fulfills the following *restricted injectivity* condition,

$$\text{(RI)} \qquad \ker\big(\nabla^2 F(x^\star)\big) \cap T_{x^\star} = \{0\}.$$

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

$$\langle h,\, \nabla^2 F(x)h\rangle > \alpha\|h\|^2, \forall x \in \mathbb{B}_\epsilon(x^\star) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : \|x - x^\star\| \leq \epsilon\}. \tag{4.1}$$

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

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

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

(ii) *If moreover $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}_{x^\star})$, then Φ has at least a quadratic growth near $x^\star$.*

4.1. Locally linearized iteration.

Define the following matrices which are all *symmetric*,

$$H \stackrel{\text{def}}{=} \gamma \mathrm{P}_{T_{x^\star}} \nabla^2 F(x^\star) \mathrm{P}_{T_{x^\star}},\;\; G \stackrel{\text{def}}{=} \mathrm{Id} - H,\;\; U \stackrel{\text{def}}{=} \gamma \nabla^2_{\mathcal{M}_{x^\star}} \Phi(x^\star) \mathrm{P}_{T_{x^\star}} - H, \tag{4.2}$$

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

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

(i) (ND) *holds.*

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

In turn, $\mathrm{Id} + U$ is invertible, and $W \stackrel{\text{def}}{=} (\mathrm{Id} + U)^{-1}$ 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 $\gamma \in [\underline{\epsilon}, 2\beta - \overline{\epsilon}]$, $\underline{\epsilon}$ and $\overline{\epsilon} > 0$, G has eigenvalues in $[-1 + \frac{\overline{\epsilon}}{\beta}, 1 - \alpha\underline{\epsilon}[\subset]-1, 1[$.*

  (c) *For $\gamma \in [\underline{\epsilon}, \beta]$, G is also symmetric positive semi-definite with eigenvalues in $[0, 1 - \alpha\underline{\epsilon}[\subset [0, 1[$.*

(ii) *If both the assumptions of Lemma 13 and (i) hold, then WG has real eigenvalues lying in $]-1, 1[$. If moreover $\gamma \in [\underline{\epsilon}, \beta]$, then WG has eigenvalues lying in $[0, 1[$.*

Let $a \in [0, \bar{a}], b \in [0, \bar{b}], \gamma \in [\underline{\epsilon}, 2\beta - \overline{\epsilon}]$, define $r_k \stackrel{\text{def}}{=} x_k - x^\star, d_k \stackrel{\text{def}}{=} \binom{r_k}{r_{k-1}}$, and matrix

$$M \stackrel{\text{def}}{=} \begin{bmatrix} (a-b)W + (1+b)WG & -(a-b)W - bWG \\ \mathrm{Id} & 0 \end{bmatrix}. \tag{4.3}$$

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.

PROPOSITION 15 (Locally linearized iteration). *Let* (**H.1**)-(**H.3**) *hold, and suppose that an FB-type method is used to create a sequence* $(x_k)_{k\in\mathbb{N}}$ *that converges to* $x^\star \in \operatorname{Argmin}(\Phi)$ *such that* (ND) *and* (RI) *hold. If moreover,*

$$a_k \to a \in [0,1],\ b_k \to b \in [0,1],\ \gamma_k \to \gamma \in [\underline{\epsilon}, 2\beta - \overline{\epsilon}], \tag{4.4}$$

then for k large enough, we have

$$d_{k+1} = M d_k + o(\|d_k\|). \tag{4.5}$$

The $o(\cdot)$ *term disappears when* R *is locally polyhedral around* $x^\star$ *and* (γ_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* γ_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,*

$$r_{k+1} = WGr_k + o(\|r_k\|). \tag{4.6}$$

(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* $a_k \to a = b = 1$*.*

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 WG, 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 η and σ be an eigenvalue of WG and M respectively. We denote $\underline{\eta}, \overline{\eta}$ the smallest and largest (signed) eigenvalues of WG, and $\rho(M)$ the spectral radius of M.

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

PROPOSITION 17. *Suppose that* R *is locally polyhedral around* $x^\star$*. If* $\binom{r_1}{r_2}$ *is an eigenvector of* M *corresponding to an eigenvalue* σ*, then it must satisfy* $r_1 = \sigma r_2$*. Moreover, we have*

(i) r_2 *is an eigenvector of* G *associated to an eigenvalue* η*, where* η *and* σ *satisfy the relation*

$$\sigma^2 - \big((a-b) + (1+b)\eta\big)\sigma + (a-b) + b\eta = 0. \tag{4.7}$$

(ii) *Given any* $(a,b) \in [0,1[^2$*, then* $\rho(M) < 1$ *if, and only if,*

$$\big(2(b-a) - 1\big)/(1+2b) < \underline{\eta}. \tag{4.8}$$

REMARK 18. *It can be shown that, given* a *and* b*,* $\rho(M)$ *is determined only by* $\underline{\eta}$ *and* $\overline{\eta}$*. 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 $\binom{r_1}{r_2}$ be an eigenvector of M corresponding to an eigenvalue σ, then it must satisfy $r_1 = \sigma r_2$. Moreover r_2 is an eigenvector of G related to eigenvalue η, where η and σ satisfy the relation*

$$\sigma^2 - (1+a)\eta\sigma + a\eta = 0, \tag{4.9}$$

and $\rho(M) < 1$ if, and only if,

$$-1/(1+2a) < \underline{\eta}. \tag{4.10}$$

REMARK 20. *Condition (4.10) holds naturally for $\gamma \in]0, \beta]$, since by Lemma 14(ii), for such γ, $\underline{\eta} \geq 0$.*

4.3. Local linear convergence of FB-type methods. We start with the case where R is locally polyhedral around $x^\star$.

THEOREM 21. *Suppose* (**H.1**)-(**H.3**) *hold, and an FB-type method generates a sequence $x_k \to x^\star \in \mathrm{Argmin}(\Phi)$ such that R is* locally polyhedral *around $x^\star$, F is C^2 near $x^\star$, 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 $x^\star$. 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*

$$\|x_k - x^\star\| \leq C\rho^{k-K}\|x_K - x^\star\|.$$

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 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 $x_k \to x^\star \in \mathrm{Argmin}(\Phi)$ such that $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}_{x^\star})$, F is C^2 near $x^\star$, 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 $x^\star$. 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*

$$\|x_k - x^\star\| \leq C\rho^{k-K}\|x_K - x^\star\|.$$

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 R is locally polyhedral (see the numerical experiments for more details). The main reasons underlying this is that, here, our rate estimate relies on the locally linearized iteration in Proposition 15 and the spectral properties of M, which takes intro 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 $x^\star$. 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 Φ satisfies the (non-smooth) Kurdyka-Łojasiewicz 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 $x_k \to x^\star \in \mathrm{Argmin}(\Phi)$ such that R is locally polyhedral around $x^\star$, F is C^2 near $x^\star$, and condition* (ND) *holds. If moreover there exists $\epsilon > 0$ and a subspace V such that*

$$\ker(\mathrm{P}_{T_x}\nabla^2 F(x)\mathrm{P}_{T_x}) = V, \ \ \forall x \in \mathbb{B}_\epsilon(x^\star) \cap (x^\star + T_{x^\star}).$$

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

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 α from (4.1) and $1/\beta$ is the Lipschitz constant of ∇F.

FB is locally faster than FISTA. For the sake of brevity (the same conclusions hold true in the general case), we consider $b_k = a_k \equiv a \in [0,1]$ and $\gamma_k \equiv \gamma \in]0,\beta]$ is fixed, in which case $\overline{\eta} \geq \underline{\eta} \geq 0$ (see Lemma 14(ii)), and thus condition (4.10) is in force. Moreover $\overline{\eta}$ is also the local convergence rate of the FB method, and $\rho(M)$ depends solely on $\overline{\eta}$ 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 $\overline{\eta}$. One can make the the following observations:

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

(ii) In particular, as $a = 1$ for FISTA, we have $\rho(M) = \sqrt{\overline{\eta}} > \overline{\eta}$. 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 $a = \frac{(1-\sqrt{1-\overline{\eta}})^2}{\overline{\eta}}$, 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

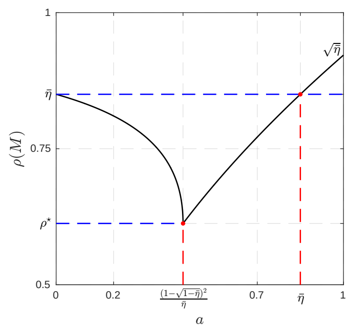


Fig. 1: Let $b = a$, and assume $\underline{\eta}, \overline{\eta}$ are known and also close enough such that the spectral radius $\rho(M)$ is only affected by $\overline{\eta}$, then $\rho(M)$ is a function of a.

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 $\sigma_{\max}$ such that $|\sigma_{\max}| = \rho(M)$ is complex. It can then be shown that the oscillation period of $\|x_k - x^\star\|$ is exactly $\frac{\pi}{\theta}$, where θ is the argument of $\sigma_{\max}$.

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

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

then as long as $a > (1-\sqrt{1-\overline{\eta}})^2/\overline{\eta}$, 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 $\gamma \in]0, \beta]$, then $\overline{\eta} = 1 - \alpha\gamma \geq \underline{\eta} \geq 1 - \gamma/\beta \geq 0$, and $\rho(M)$ depends only on $\overline{\eta}$, a and b. Then with fixed γ (hence $\overline{\eta}$), $\rho(M)$ attains its minimal value for a and b satisfying

$$\begin{cases} b = a : a = \big((1-\sqrt{1-\overline{\eta}})^2\big)/\overline{\eta} = (1-\sqrt{\alpha\gamma})/(1+\sqrt{\alpha\gamma}), \\ b \neq a : a = (1-\sqrt{1-\overline{\eta}})^2 + b(1-\overline{\eta}) = (1-\sqrt{\alpha\gamma})^2 + b\alpha\gamma, \end{cases} \tag{4.11}$$

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

$$\rho^\star = 1 - \sqrt{1-\overline{\eta}} = 1 - \sqrt{\gamma\alpha}, \tag{4.12}$$

where the second equality comes from (4.2) and Lemma 14. This is a decreasing function of γ, and $\rho^\star = 1 - \sqrt{\alpha\beta}$ 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

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

This is known to be the optimal rate that matches the lower complexity bounds for first-order methods to solve the class of problems ($\mathcal{P}_{\rm opt}$) if F were also α-strongly convex [44, Theorem 2.1.13]. In comparison, for the FB method (*i.e.* $a = b = 0$), the optimal rate is $\rho^\star = \overline{\eta}^\star = \frac{1-\alpha\beta}{1+\alpha\beta}$ 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 Φ. This can be done either using local parameterizations obtained from $\mathcal{U}$-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 $\nabla^2_{\mathcal{M}_{x^\star}} \Phi(x^\star)$ 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 $y = Lx_{\rm ob} + w$, 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

$$(\mathcal{P}_\lambda) \qquad \min_{x\in\mathbb{R}^n} \frac{1}{2}\|y - Lx\|^2 + \lambda R(x),$$

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

We use three functions R: the ℓ_1-norm ($R(x) = \|x\|_1 \stackrel{\rm def}{=} \sum_{i=1}^n |x_i|$), the $\ell_{1,2}$-norm ($R(x) = \|x\|_{1,2} \stackrel{\rm def}{=} \sum_{b\in\mathcal{B}} \|x_b\|$, for a uniform disjoint partition of $\{1, \ldots, n\}$ in blocks $\mathcal{B}$), and the nuclear norm ($R(x) = \|x\|_* \stackrel{\rm def}{=} \|\sigma(x)\|_1$, 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 ℓ_1 and $\ell_{1,2}$-norms are partly smooth relative to subspaces [57] (ℓ_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 $x_{\rm ob}$:

ℓ_1-norm: $(m, n) = (48, 128)$, $\|x_{\rm ob}\|_0 = 8$;
$\ell_{1,2}$-norm: $(m, n) = (60, 128)$, $x_{\rm ob}$ has 3 non-zero blocks of size 4;
Nuclear norm: $(m, n) = (1425, 2500)$, $x_{\rm ob} \in \mathbb{R}^{50\times 50}$ and $\mathrm{rank}(x_{\rm ob}) = 5$.

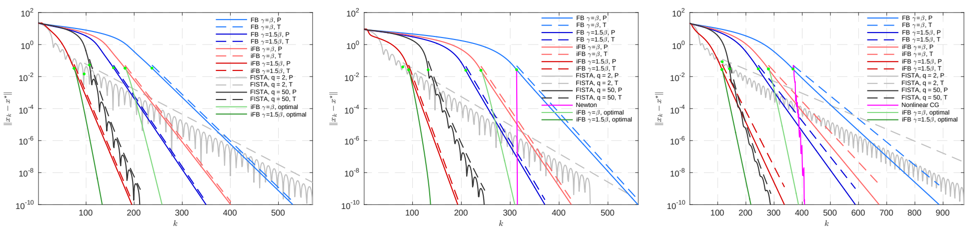
One can show that with the number of measurements m in the above cases, if λ and $\|w\|$ are set properly, then with high probability on L, ($\mathcal{P}_\lambda$) admits a unique solution $x^\star$ with $\mathcal{M}_{x^\star} = \mathcal{M}_{x_{\rm ob}}$, and $x^\star$ satisfies both (ND) and (RI).

Parameter settings. We choose $\gamma_k \equiv \beta$ for FISTA. For FB/iFB methods, two choices of γ_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 $\|x_k - x^\star\|$ 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 $\mathcal{M}_{x^\star}$ has been identified.

Tightness of predicted rates. For the ℓ_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



(a) ℓ_1-norm (b) $\ell_{1,2}$-norm (c) Nuclear norm

Fig. 2: Local linear convergence and comparison of the FB-type methods (FB, iFB and FISTA) in terms of $\|x_k - x^\star\|$. See text for description.

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 γ_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 γ_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 ℓ_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 $x^\star$ once identification occurs. This can be easily achieved by projection the first-order minimality condition on $\mathcal{M}_{x^\star} = x^\star + T_{x^\star}$, which boils down to solving an overdetermined linear system which has a unique solution under the restricted injectivity 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 applied, 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 $\operatorname{gph} A = \{(x, y) \in \mathcal{H} \times \mathcal{H} | y \in A(x)\}$, and its zeros set is $\operatorname{zer} A = \{x \in \mathcal{H} | 0 \in A(x)\}$. Recall that a set-valued operator $A : \mathcal{H} \rightrightarrows \mathcal{H}$ is monotone if

$$(\forall (x, v) \in \operatorname{gph} A), (\forall (y, u) \in \operatorname{gph} A),\ \langle x - y,\, v - u\rangle \geq 0. \tag{A.1}$$

It is moreover maximal monotone if $\operatorname{gph} A$ can not be contained in the graph of any other monotone operator. Let $\beta \in]0, +\infty[$, $B : \mathcal{H} \to \mathcal{H}$, then B is β-cocoercive if

$$(\forall x, y \in \mathcal{H}),\ \beta\|Bx - By\|^2 \leq \langle Bx - By,\, x - y\rangle. \tag{A.2}$$

Proof (Theorem 3). Define the following quantities

$$\varphi_k = \tfrac{1}{2}\|x_k - x^\star\|^2, \Delta_k = \tfrac{1}{2}\|x_k - x_{k-1}\|^2, E_{b,k} = \tfrac{1}{2}\|y_{b,k} - x_{k+1}\|^2. \tag{A.3}$$

Let $x^\star \in \operatorname{zer}(A + B)$, *i.e.* a solution $(\mathcal{P}_{\text{inc}})$, which exists thanks to (**H.6**). Recall from (1.4) and (2.1) that

$$-B(x^\star) \in A(x^\star) \text{ and } y_{a,k} - \gamma_k B(y_{b,k}) - \gamma_k \xi_k - x_{k+1} \in \gamma_k A^{\varepsilon_k}(x_{k+1}).$$

Thus, we get

$$\langle y_{a,k} - x_{k+1} - \gamma_k(B(y_{b,k}) - B(x^\star)) - \gamma_k \xi_k,\, x_{k+1} - x^\star\rangle \geq -\gamma_k \varepsilon_k.$$

Combining this with the definition of $y_{a,k}$, we obtain

$$\begin{aligned}\varphi_k - \varphi_{k+1} &= \tfrac{1}{2}\langle x_k - x^\star + x_{k+1} - x^\star,\, x_k - x_{k+1}\rangle \\ &= \Delta_{k+1} + \langle y_{a,k} - x_{k+1},\, x_{k+1} - x^\star\rangle - a_k\langle x_k - x_{k-1},\, x_{k+1} - x^\star\rangle \\ &\geq \Delta_{k+1} + \gamma_k\langle B(y_{b,k}) - B(x^\star) + \xi_k,\, x_{k+1} - x^\star\rangle \\ &\quad - a_k\langle x_k - x_{k-1},\, x_{k+1} - x^\star\rangle - \gamma_k\varepsilon_k.\end{aligned} \tag{A.4}$$

For $\langle x_k - x_{k-1},\, x_{k+1} - x^\star\rangle$, we have

$$\langle x_k - x_{k-1},\, x_{k+1} - x^\star\rangle = \langle x_k - x_{k-1},\, x_{k+1} - x_k\rangle + (\Delta_k + \varphi_k - \varphi_{k-1}), \tag{A.5}$$

where we applied the usual Pythagoras relation to $\langle x_k - x_{k-1},\, x_k - x^\star\rangle$. Putting (A.5) back into (A.4) yields

$$\begin{aligned}\varphi_{k+1} - \varphi_k - a_k(\varphi_k - \varphi_{k-1}) \leq &- \Delta_{k+1} - \gamma_k\langle B(y_{b,k}) - B(x^\star) + \xi_k,\, x_{k+1} - x^\star\rangle \\ &+ a_k\langle x_k - x_{k-1},\, x_{k+1} - x_k\rangle + a_k\Delta_k + \gamma_k\varepsilon_k.\end{aligned} \tag{A.6}$$

Since B is β-cocoercive, Young's inequality yields

$$\begin{aligned}&\langle B(y_{b,k}) - B(x^\star),\, x_{k+1} - x^\star\rangle \\ &\geq \beta\|B(y_{b,k}) - B(x^\star)\|^2 + \langle B(y_{b,k}) - B(x^\star),\, x_{k+1} - y_{b,k}\rangle = -\tfrac{1}{2\beta}E_{b,k}.\end{aligned} \tag{A.7}$$

Denote $\mu_k = 1-\frac{\gamma_k}{2\beta} \in [\frac{\bar\epsilon}{2\beta}, 1-\frac{\epsilon}{2\beta}]$, $\nu_k = a_k - \frac{\gamma_k b_k}{2\beta}$ and $v_k = x_{k+1}-x_k-\frac{\nu_k}{\mu_k}(x_k-x_{k-1})$. Substituting (A.7) back into (A.6), and since $E_{b,k} = \Delta_{k+1}+b_k^2\Delta_k+b_k\langle x_k-x_{k+1},\, x_k-x_{k-1}\rangle$, we get

(A.8)

$$\begin{aligned}
&\varphi_{k+1}-\varphi_k-a_k(\varphi_k-\varphi_{k-1})\\
&\leq -\Delta_{k+1}+\tfrac{\gamma_k}{2\beta}E_{b,k}+a_k\langle x_k-x_{k-1},\, x_{k+1}-x_k\rangle + a_k\Delta_k+\gamma_k\varepsilon_k-\gamma_k\langle \xi_k,\, x_{k+1}-x^\star\rangle\\
&= -\tfrac{\mu_k}{2}\|v_k\|^2+\big(a_k+\frac{\nu_k^2}{\mu_k}+\frac{\gamma_k b_k^2}{2\beta}\big)\Delta_k+\gamma_k\big(\varepsilon_k+\sqrt{2}\|\xi_k\|\sqrt{\varphi_{k+1}}\big)\\
&\leq -\tfrac{\mu_k}{2}\|v_k\|^2+\big(\frac{2a_k}{\mu_k}+\frac{\gamma_k b_k}{2\beta}\big)\Delta_k+\gamma_k\big(\varepsilon_k+\sqrt{2}\|\xi_k\|\sqrt{\varphi_{k+1}}\big)\\
&\leq -\tfrac{\mu_k}{2}\|v_k\|^2+\big(\tfrac{4\beta}{\bar\epsilon}a_k+(1-\frac{\bar\epsilon}{2\beta})b_k\big)\Delta_k+\overline{\gamma}\big(\varepsilon_k+\sqrt{2}\|\xi_k\|\sqrt{\varphi_{k+1}}\big).
\end{aligned}$$

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

$$\begin{aligned}
\theta_{k+1} &\leq -\tfrac{\mu_k}{2}\|v_k\|^2+a_k\theta_k+\delta_k+\overline{\gamma}\varepsilon_k+\sqrt{2}\overline{\gamma}\|\xi_k\|\sqrt{\varphi_{k+1}}\\
&\leq \prod_{j=1}^k a_j\theta_1+\textstyle\sum_{j=1}^k\big(\prod_{l=j}^k a_{l-j}\big)\big(\delta_j+\overline{\gamma}\varepsilon_j+\sqrt{2}\overline{\gamma}\|\xi_j\|\sqrt{\varphi_{j+1}}\big) \qquad \text{(A.9)}\\
&\leq \bar a^k\varphi_1+\textstyle\sum_{j=1}^k\bar a^{k-j}\big(\delta_j+\overline{\gamma}\varepsilon_j+\sqrt{2}\overline{\gamma}\|\xi_j\|\sqrt{\varphi_{j+1}}\big).
\end{aligned}$$

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

$$\begin{aligned}
\textstyle\sum_{m=1}^k\theta_{m+1} &= \varphi_{k+1}-\varphi_1\\
&\leq \frac{1}{1-\bar a}\varphi_1+\textstyle\sum_{m=1}^k\sum_{j=1}^m\bar a^{k-j}\big(\delta_j+\overline{\gamma}\varepsilon_j+\sqrt{2}\overline{\gamma}\|\xi_j\|\sqrt{\varphi_{j+1}}\big)\\
&\leq \frac{1}{1-\bar a}\varphi_1+\textstyle\sum_{m=1}^k\big(\sum_{j=1}^{k-m}\bar a^j\big)\big(\delta_m+\overline{\gamma}\varepsilon_m+\sqrt{2}\overline{\gamma}\|\xi_m\|\sqrt{\varphi_{m+1}}\big)\\
&\leq \frac{1}{1-\bar a}\big(\varphi_1+\textstyle\sum_{m=1}^k\big(\delta_m+\overline{\gamma}\varepsilon_m+\sqrt{2}\overline{\gamma}\|\xi_m\|\sqrt{\varphi_{m+1}}\big)\big),
\end{aligned}$$

which entails

(A.10)

$$\varphi_{k+1}\leq c+\sqrt{2}\overline{\gamma}\textstyle\sum_{m=1}^k\|\xi_m\|\sqrt{\varphi_{m+1}}\leq c+\sqrt{2}\overline{\gamma}\sum_{m=1}^{k+1}\|\xi_{m-1}\|\sqrt{\varphi_m},$$

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

$$\begin{aligned}
\theta_{k+1} &\leq -\tfrac{\mu_k}{2}\|v_k\|^2+\bar a\theta_k+\delta_k+\overline{\gamma}\varepsilon_k+\sqrt{2t}\overline{\gamma}\|\xi_k\|\\
&\leq -\tfrac{\mu_k}{2}\|v_k\|^2+a_k[\theta_k]_+ +\delta_k+\overline{\gamma}\varepsilon_k+\sqrt{2t}\overline{\gamma}\|\xi_k\| \qquad \text{(A.11)}
\end{aligned}$$

where $[\theta]_+ = \max\{\theta,0\}$. As a result, we have

$$[\theta_{k+1}]_+\leq\bar a[\theta_k]_+ + e_k,$$

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

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

It then follows that the sequence $(\varphi_k-\sum_{j=1}^k[\theta_j]_+)_{k\in\mathbb{N}}$ is decreasing and bounded from below, hence convergent, whence we deduce that φ_k is also convergent.

(ii) $a_k \equiv 0$: in this case, (A.9) reduces to

$$\begin{aligned}\varphi_{k+1} &\le \varphi_k + \delta_k + \overline{\gamma}\varepsilon_k + \sqrt{2}\overline{\gamma}\|\xi_k\|\sqrt{\varphi_k} \\ &\le \varphi_1 + \textstyle\sum_{j\in\mathbb{N}}\delta_j + \overline{\gamma}\sum_{j\in\mathbb{N}}\varepsilon_j + \sqrt{2}\overline{\gamma}\sum_{j=1}^{k}\|\xi_j\|\sqrt{\varphi_{j+1}}.\end{aligned}$$

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 $\varphi_k \le t = \sqrt{\varphi_1 + \sum_{j\in\mathbb{N}}(\delta_j + \overline{\gamma}\varepsilon_j + \|\xi_j\|)} < +\infty$. Consequently, we have

$$\varphi_{k+1} \le \varphi_k + \delta_k + \overline{\gamma}\varepsilon_k + \sqrt{2t}\overline{\gamma}\|\xi_k\|.$$

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

In summary, for $a_k \in [0, \bar{a}]$, $\lim_{k\to+\infty}\|x_k - x^\star\|$ exists for any $x^\star \in \mathrm{zer}(A+B)$, 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

$$\tfrac{\nu_k}{\mu_k}(x_k - x_{k-1}) \to 0, \tag{A.12}$$

since $\mu_k \ge \frac{\overline{\epsilon}}{2\beta} > 0$. Moreover, from (A.11), we obtain

$$\textstyle\sum_{k\in\mathbb{N}}\|v_k\|^2 \le \frac{4\beta}{\overline{\epsilon}}\big(\bar{a}\varphi_0 + \sum_{k\in\mathbb{N}}(\bar{a}[\theta_k]_+ + e_k)\big) < +\infty.$$

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 $x_{k_j} \rightharpoonup \bar{x}$. Denote $u_{k_j} \stackrel{\text{def}}{=} \frac{y_{a,k_j} - x_{k_j+1}}{\gamma_{k_j}} - B(y_{b,k_j}) - \xi_{k_j}$. Since B is cocoercive and $y_{b,k_j} \rightharpoonup \bar{x}$, we have $B(y_{b,k_j}) \to B(\bar{x})$. In turn, $u_{k_j} \to -B(\bar{x})$ since $\gamma_k \ge \underline{\epsilon} > 0$ and $\xi_k \to 0$. Since $(x_{k_j+1}, u_{k_j}) \in \mathrm{gph}\, A^{\varepsilon_{k_j}}$, 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 ($\mathcal{P}_{\mathrm{inc}}$). 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 $c = \sup_{k\in\mathbb{N}}\|x_k - x^\star\| < +\infty$. From (A.8), we apply Young's inequality to get

$$\begin{aligned}&\varphi_{k+1} - \varphi_k - a_k(\varphi_k - \varphi_{k-1}) \\ &\le \big(\tfrac{\gamma_k}{2\beta} - 1\big)\Delta_{k+1} + |a_k - \tfrac{\gamma_k b_k}{2\beta}|(\Delta_{k+1} + \Delta_k) + \big(\tfrac{\gamma_k}{2\beta}b_k^2 + a_k\big)\Delta_k + \gamma_k\big(\varepsilon_k + c\|\xi_k\|\big) \\ &= s_k\Delta_{k+1} + t_k\Delta_k + \overline{\gamma}\big(\varepsilon_k + c\|\xi_k\|\big),\end{aligned}$$

where $s_k = \frac{\gamma_k}{2\beta} - 1 + |a_k - \frac{\gamma_k b_k}{2\beta}|, t_k = \frac{\gamma_k}{2\beta}b_k^2 + a_k + |a_k - \frac{\gamma_k b_k}{2\beta}|$. Suppose that a_k, b_k and γ_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 $\delta_k = \overline{\gamma}\big(\varepsilon_k + c\|\xi_k\|\big)$,

$$\begin{aligned}\phi_{k+1} - \phi_k &\le (\varphi_{k+1} - \varphi_k) - a_k(\varphi_k - \varphi_{k-1}) + t_{k+1}\Delta_{k+1} - t_k\Delta_k \\ &\le s_k\Delta_{k+1} + t_k\Delta_k + t_{k+1}\Delta_{k+1} - t_k\Delta_k + \delta_k \\ &= (s_k + t_{k+1})\Delta_{k+1} + \delta_k.\end{aligned} \tag{A.13}$$

(i) $a_k \in [0, \bar{a}]$, $b_k \in [0, \bar{b}]$, $b_k \le a_k$. We have $\frac{\gamma_k}{2\beta}b_k < a_k$, then from (A.13), and under the second condition in (2.4),

$$\begin{aligned}\phi_{k+1} - \phi_k &\le (s_{k+1} + t_{k+1})\Delta_{k+1} \\ &= \big((3a_{k+1} - 1) + \tfrac{\gamma_{k+1}}{2\beta}(1 - b_{k+1})^2\big)\Delta_{k+1} + \delta_k \le -\tau\Delta_{k+1} + \delta_k.\end{aligned} \tag{A.14}$$

(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,

$$\begin{aligned}\phi_{k+1} - \phi_k &\leq (s_{k+1} + t_{k+1})\Delta_{k+1} + \delta_k \\ &\leq \big(\tfrac{\gamma_{k+1}}{2\beta} - 1 + 2|a_{k+1} - \tfrac{\gamma_{k+1}}{2\beta} b_{k+1}| + \tfrac{\gamma_{k+1}}{2\beta} b_{k+1}^2 + a_{k+1}\big)\Delta_{k+1} + \delta_k.\end{aligned}$$

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

(a) If $\frac{\gamma_{k+1}}{2\beta} b_{k+1} \leq a_{k+1}$, $k \in \mathbb{N}$, then from the second condition in (2.4),

$$\phi_{k+1} - \phi_k \leq \big((3a_{k+1} - 1) + \tfrac{\gamma_{k+1}}{2\beta}(1 - b_{k+1})^2\big)\Delta_{k+1} + \delta_k \leq -\tau\Delta_{k+1} + \delta_k. \tag{A.15}$$

(b) If $a_{k+1} < \frac{\gamma_{k+1}}{2\beta} b_{k+1}$, $k \in \mathbb{N}$, then from the first condition of (2.4),

$$\phi_{k+1} - \phi_k \leq \big(-(1 + a_{k+1}) + \tfrac{\gamma_{k+1}}{2\beta}(1 + b_{k+1})^2\big)\Delta_{k+1} + \delta_k \leq -\tau\Delta_{k+1} + \delta_k. \tag{A.16}$$

Under the assumptions of (i), we have from (A.14) (resp. (A.15) or (A.16)) that

$$\textstyle\sum_{j=1}^{k} \Delta_{j+1} \leq \frac{1}{\tau}(\phi_1 - \phi_{k+1}) + \sum_{j=1}^{k} \delta_j \leq \frac{1}{\tau}(\phi_1 + \bar{a}\varphi_k) + \sum_{j=1}^{k} \delta_j < +\infty.$$

If the errors vanish, (A.14) (resp. (A.15) or (A.16)) indicate that ϕ_k is non-increasing. Thus

$$\textstyle\sum_{j=1}^{k} \Delta_{j+1} \leq \frac{1}{\tau}(\phi_1 - \phi_{k+1}) \leq \frac{1}{\tau}(\phi_1 + \bar{a}\varphi_k) \leq \frac{1}{\tau}\big(\bar{a}^k \varphi_1 + \frac{\phi_1}{1 - \bar{a}}\big) < +\infty.$$

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

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 $\mathcal{T}_{\mathcal{M}}(x)$ and $\mathcal{N}_{\mathcal{M}}(x)$ 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 $\mathfrak{g}(t; x, h)$ the value at $t \in \mathbb{R}$ of the geodesic starting at $\mathfrak{g}(0; x, h) = x \in \mathcal{M}$ with velocity $\dot{\mathfrak{g}}(t; x, h) = \frac{d\mathfrak{g}}{dt}(t; x, h) = h \in \mathcal{T}_{\mathcal{M}}(x)$ (which is uniquely defined). For every $h \in \mathcal{T}_{\mathcal{M}}(x)$, there exists an interval I around 0 and a unique geodesic $\mathfrak{g}(t; x, h) : I \to \mathcal{M}$ such that $\mathfrak{g}(0; x, h) = x$ and $\dot{\mathfrak{g}}(0; x, h) = h$. The mapping

$$\mathrm{Exp}_x : \mathcal{T}_{\mathcal{M}}(x) \to \mathcal{M}, \ \ h \mapsto \mathrm{Exp}_x(h) = \mathfrak{g}(1; x, h),$$

is called *Exponential map*. Given $x, z \in \mathcal{M}$, the direction $h \in \mathcal{T}_{\mathcal{M}}(x)$ we are interested in is such that

$$\mathrm{Exp}_x(h) = z = \mathfrak{g}(1; x, h).$$

Parallel translation. Given two points $x, z \in \mathcal{M}$, let $\mathcal{T}_{\mathcal{M}}(x), \mathcal{T}_{\mathcal{M}}(z)$ be their corresponding tangent spaces. Define

$$\tau : \mathcal{T}_{\mathcal{M}}(x) \to \mathcal{T}_{\mathcal{M}}(z),$$

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 $v \in \mathcal{N}_{\mathcal{M}}(x)$, the Weingarten map of $\mathcal{M}$ at x is the operator $\mathfrak{W}_x(\cdot, v) : \mathcal{T}_{\mathcal{M}}(x) \to \mathcal{T}_{\mathcal{M}}(x)$ defined by

$$\mathfrak{W}_x(\cdot, v) = -\mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x)} \mathrm{d}V[h],$$

where V is any local extension of v to a normal vector field on $\mathcal{M}$. 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 $\mathcal{M}$, see [19, Proposition II.2.1].

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

$$\langle \nabla_{\mathcal{M}} G(z),\, h\rangle = \tfrac{d}{dt} G\big(\mathrm{P}_{\mathcal{M}}(z + th)\big)\big|_{t=0},\ \forall h \in \mathcal{T}_{\mathcal{M}}(z),$$

where $\mathrm{P}_{\mathcal{M}}$ is the projection operator onto $\mathcal{M}$. The covariant Hessian of G at z is the symmetric linear mapping $\nabla^2_{\mathcal{M}} G(z)$ from $\mathcal{T}_{\mathcal{M}}(z)$ to itself which is defined as

$$\langle \nabla^2_{\mathcal{M}} G(z)h,\, h\rangle = \tfrac{d^2}{dt^2} G\big(\mathrm{P}_{\mathcal{M}}(z + th)\big)\big|_{t=0},\ \forall h \in \mathcal{T}_{\mathcal{M}}(z). \tag{B.1}$$

This definition agrees with the usual definition using geodesics or connections [40]. Now assume that $\mathcal{M}$ is a Riemannian embedded submanifold of $\mathbb{R}^n$, and that a function G has a C^2-smooth restriction on $\mathcal{M}$. This can be characterized by the existence of a C^2-smooth extension (representative) of G, *i.e.* a C^2-smooth function $\widetilde{G}$ on $\mathbb{R}^n$ such that $\widetilde{G}$ agrees with G on $\mathcal{M}$. Thus, the Riemannian gradient $\nabla_{\mathcal{M}} G(z)$ is also given by

$$\nabla_{\mathcal{M}} G(z) = \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(z)} \nabla \widetilde{G}(z), \tag{B.2}$$

and $\forall h \in \mathcal{T}_{\mathcal{M}}(z)$, the Riemannian Hessian reads

$$\begin{aligned} \nabla^2_{\mathcal{M}} G(z)h &= \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(z)} \mathrm{d}(\nabla_{\mathcal{M}} G)(z)[h] = \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(z)} \mathrm{d}\big(z \mapsto \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(z)} \nabla_{\mathcal{M}} \widetilde{G}\big)[h] \\ &= \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(z)} \nabla^2 \widetilde{G}(z)h + \mathfrak{W}_z\big(h, \mathrm{P}_{\mathcal{N}_{\mathcal{M}}(z)} \nabla \widetilde{G}(z)\big), \end{aligned} \tag{B.3}$$

where the last equality comes from [2, Theorem 1]. When $\mathcal{M}$ is an affine or linear subspace of $\mathbb{R}^n$, then obviously $\mathcal{M} = x + \mathcal{T}_{\mathcal{M}}(x)$, and $\mathfrak{W}_z(h, \mathrm{P}_{\mathcal{N}_{\mathcal{M}}(z)} \nabla \widetilde{G}(z)) = 0$, hence (B.3) reduces to

$$\nabla^2_{\mathcal{M}} G(z) = \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(z)} \nabla^2 \widetilde{G}(z) \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(z)}.$$

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 $x \in \mathcal{M}$, and x_k a sequence converging to x in $\mathcal{M}$. Denote $\tau_k : \mathcal{T}_{\mathcal{M}}(x) \to \mathcal{T}_{\mathcal{M}}(x_k)$ 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*

$$(\tau_k^{-1} \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x_k)} - \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x)})u = o(\|u\|).$$

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

$$\tau_k^{-1} = \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x)} + o(\|x_k - x\|).$$

In addition, locally near x along $\mathcal{M}$, the operator $x \mapsto \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x)}$ is C^1, hence,

$$\begin{aligned} \lim_{k\to\infty} \frac{\|(\tau_k^{-1} \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x_k)} - \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x)})u\|}{\|u\|} &\le \lim_{k\to\infty} \frac{\|\mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x)}(\mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x_k)} - \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x)})\|\|u\|}{\|u\|} + o(\|x_k - x\|) \\ &\le \lim_{k\to\infty} \|\mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x_k)} - \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x)}\| + o(\|x_k - x\|) = 0. \end{aligned}$$

∎

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

$$\tau^{-1} \nabla_{\mathcal{M}} \Phi(z) = \nabla_{\mathcal{M}} \Phi(x) + \nabla^2_{\mathcal{M}} \Phi(x) \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x)}(z - x) + o(\|z - x\|).$$

Proof. Since $x, z \in \mathcal{M}$ are close, we have $z = \mathrm{Exp}_x(h)$ for some $h \in \mathcal{T}_{\mathcal{M}}(x)$ small enough, and thus, the Taylor expansion [52, Remark 4.2] of $\nabla_{\mathcal{M}}\Phi$ around x reads

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

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

$$\mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x)}(z) = \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x)}(\mathrm{Exp}_x(h)) = \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x)}(x) + h + o(\|h\|^2).$$

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

B.2. Proofs.

Proof (Proposition 12).

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

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

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

(a) $\delta \notin \ker(\nabla^2 F(x_t))$. Since F and R are convex with $-\nabla F(x^\star) \in \partial R(x^\star)$,

$$\Phi(x^\star + \delta) - \Phi(x^\star) \geq \frac{1}{2}\langle \delta,\, \nabla^2 F(x_t)\delta\rangle > 0.$$

(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 $x^\star$. Moreover $\partial R(x^\star) \neq \emptyset$ ($-\nabla F(x^\star)$ is in it), and thus the directional derivative $R'(x^\star, \cdot)$ is proper and closed, and it is the support of $\partial R(x^\star)$ [51, Theorem 8.30]. It then follows from the separation theorem [30, Theorem V.2.2.3] that

$$\begin{aligned} &-\nabla F(x^\star) \in \mathrm{ri}(\partial R(x^\star)) \\ &\Leftrightarrow R'(x^\star, \delta) > -\langle \nabla F(x^\star),\, \delta\rangle,\ \forall \delta \text{ s.t. } R'(x^\star; \delta) + R'(x^\star; -\delta) > 0. \end{aligned}$$

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

$$\begin{aligned} -\nabla F(x^\star) \in \mathrm{ri}(\partial R(x^\star)) &\Leftrightarrow R'(x^\star; \delta) > -\langle \nabla F(x^\star),\, \delta\rangle, \forall \delta \notin T_{x^\star} \\ &\Rightarrow R'(x^\star; \delta) > -\langle \nabla F(x^\star),\, \delta\rangle, \forall \delta \in \ker(\nabla^2 F(x_t)) \setminus \{0\}. \end{aligned}$$

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

$$\begin{aligned} &\Phi(x^\star + \delta) - \Phi(x^\star) \\ &= R(x^\star + \delta) - R(x^\star) + \langle \nabla F(x^\star),\, \delta\rangle \geq R'(x^\star; \delta) + \langle \nabla F(x^\star),\, \delta\rangle > 0. \end{aligned}$$

(ii) Let Ψ as defined in the proof of Lemma 13. If $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}_{x^\star})$, the Riemannian Hessian of Φ reads

$$\nabla^2_{\mathcal{M}_{x^\star}}\Phi(x^\star) = \mathrm{P}_{T_{x^\star}}\nabla F(x^\star)\mathrm{P}_{T_{x^\star}} + \nabla^2_{\mathcal{M}_{x^\star}}\Psi(x^\star).$$

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

Proof (Lemma 13). By definition of U, $Uh = 0$ for any $h \in T_{x^\star}^\perp$. Thus, in the following we only examine the case $h \in T_{x^\star}$.

(i) Let $\Psi(x) \stackrel{\text{def}}{=} R(x) + \langle x,\, \nabla F(x^\star)\rangle$. From the smooth perturbation rule of partial smoothness [35, Corollary 4.7], $\Psi \in \mathrm{PSF}_{x^\star}(\mathcal{M}_{x^\star})$. Moreover, from Fact 7 and normal sharpness, the Riemannian Hessian of Ψ at $x^\star$ is such that, $\forall h \in T_{x^\star}$,

$$\begin{aligned} \gamma\nabla^2_{\mathcal{M}_{x^\star}}\Psi(x^\star)h &= \gamma\mathrm{P}_{T_{x^\star}}\nabla^2\widetilde{R}(x^\star)h + \gamma\mathfrak{W}_{x^\star}\big(h, \mathrm{P}_{T_{x^\star}^\perp}\nabla\widetilde{\Phi}(x^\star)\big) \\ &= \gamma\nabla^2_{\mathcal{M}_{x^\star}}\Phi(x^\star)\mathrm{P}_{T_{x^\star}}h - Hh = Uh, \end{aligned}$$

Since $-\nabla F(x^\star) \in \mathrm{ri}\big(\partial R(x^\star)\big)$, we have from [36, Corollary 5.4] that

$$\partial^2 R\big(x^\star | -\nabla F(x^\star)\big)h = \begin{cases} \nabla^2_{\mathcal{M}_{x^\star}} \Psi(x^\star)h + T_{x^\star}^\perp, & h \in T_{x^\star}, \\ \emptyset, & h \notin T_{x^\star}, \end{cases}$$

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

$$0 \le \gamma\langle \partial^2 R\big(x^\star | -\nabla F(x^\star)\big)h,\, h\rangle = \gamma\langle \nabla^2_{\mathcal{M}_{x^\star}} \Psi(x^\star)h,\, h\rangle = \langle Uh,\, h\rangle.$$

(ii) In this case, $U = \gamma \mathrm{P}_{T_{x^\star}} \nabla^2 \widetilde{R}(x^\star) \mathrm{P}_{T_{x^\star}}$. Let $x_t = x^\star + th$, $t > 0$, for any scalar t and $h \in T_{x^\star}$. Obviously, $x_t \in x^\star + T_{x^\star} = \mathcal{M}_{x^\star}$, and for t sufficiently small, by Fact 6, $T_{x_t} = T_{x^\star}$. Thus, $\forall u \in \partial R(x^\star)$ and $\forall v \in \partial R(x_t)$

$$\begin{aligned} 0 \le t^{-2}\langle v - u,\, x_t - x^\star\rangle &= t^{-1}\langle \mathrm{P}_{T_{x_t}} v - \mathrm{P}_{T_{x^\star}} u,\, h\rangle \\ \text{(by Fact 7)} \;&= \langle t^{-1}(\nabla_{\mathcal{M}_{x^\star}} R(x_t) - \nabla_{\mathcal{M}_{x^\star}} R(x^\star)),\, h\rangle \\ \text{(by (B.2))} \;&= \langle t^{-1}\mathrm{P}_{T_{x^\star}}(\nabla \widetilde{R}(x^\star + t\mathrm{P}_{T_{x^\star}} h) - \nabla \widetilde{R}(x^\star)),\, h\rangle. \end{aligned}$$

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

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 $WG = W^{1/2}W^{1/2}GW^{1/2}W^{-1/2}$, meaning that WG is similar to $W^{1/2}GW^{1/2}$. The latter is symmetric and obeys

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

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

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

(B.5)
$$\begin{aligned} &H_k = \gamma_k \mathrm{P}_{T_{x^\star}} \nabla^2 F(x^\star) \mathrm{P}_{T_{x^\star}},\; G_k = \mathrm{Id} - H_k,\; U_k = \gamma_k \nabla^2_{\mathcal{M}_{x^\star}} \Phi(x^\star)\mathrm{P}_{T_{x^\star}} - H_k, \\ &M_{k,1} = \big[(1+b)W(G_k - G), -bW(G_k - G)\big], \\ &M_{k,2} = \big[((a_k - b_k) - (a - b))W + (b_k - b)WG_k, -((a_k - b_k) - (a - b))W - (b_k - b)WG_k\big]. \end{aligned}$$

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

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

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

(B.6)
$$\begin{aligned} &\|y_{a,k} - x^\star\| = O(\|d_k\|),\, \|y_{b,k} - x^\star\| = O(\|d_k\|),\, \|r_{k+1}\| = O(\|d_k\|), \\ &(\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}} - \mathrm{P}_{T_{x^\star}})(\nabla F(y_{b,k}) - \nabla F(x_{k+1})) = o(\|d_k\|). \end{aligned}$$

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

Proof. We have

(B.8)
$$\begin{aligned} \|y_{a,k} - x^\star\| &= \|(1 + a_k)r_k - a_k r_{k-1}\| \le (1 + a_k)\|r_k\| + a_k\|r_{k-1}\| \\ &\le (1 + a_k)(\|r_k\| + \|r_{k-1}\|) \le \sqrt{2}(1 + a_k)\|d_k\|, \end{aligned}$$

whence we get the first and second estimates. In turn, we obtain

$$\begin{aligned}\|r_{k+1}\| &= \|\mathrm{prox}_{\gamma_k R}(y_{a,k} - \gamma_k \nabla F(y_{b,k})) - \mathrm{prox}_{\gamma_k R}(x^\star - \gamma_k \nabla F(x^\star))\| \\ &\leq \|(y_{a,k} - x^\star) - \gamma_k(\nabla F(y_{b,k}) - \nabla F(x^\star))\| \\ &\leq (1+a_k)\|r_k\| + a_k\|r_{k-1}\| + (1+b_k)\tfrac{\gamma_k}{\beta}\|r_k\| + \tfrac{b_k\gamma_k}{\beta}\|r_{k-1}\| \\ &\leq \big((1+a_k) + (1+b_k)\tfrac{\gamma_k}{\beta}\big)\sqrt{2}\|d_k\|,\end{aligned} \tag{B.9}$$

where we used non-expansiveness of the proximity operator and assumption (**H.2**). This yields the third estimate. Combining Lemma 26, assumption (**H.2**), (B.8) and (B.9), we get

$$\begin{aligned}(\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}} - \mathrm{P}_{T_{x^\star}})\big(\nabla F(y_{b,k}) - \nabla F(x_{k+1})\big) &= o(\|\nabla F(y_{b,k}) - \nabla F(x_{k+1})\|) \\ &= o(\|y_{b,k} - x^\star\|) + o(\|r_{k+1}\|) = o(\|d_k\|).\end{aligned}$$

For (B.7), recall the function Ψ in the proof of Lemma 13(i). First, we have

$$\begin{aligned}\lim_{k\to\infty}\|W(U_k - U)r_{k+1}\|/\|r_{k+1}\| &= \lim_{k\to\infty}\|W(\gamma_k-\gamma)\nabla^2_{\mathcal{M}_{x^\star}}\Psi(x^\star)\mathrm{P}_{T_{x^\star}}r_{k+1}\|/\|r_{k+1}\| \\ &\leq \lim_{k\to\infty}|\gamma_k-\gamma|\|W\|\|\nabla^2_{\mathcal{M}_{x^\star}}\Psi(x^\star)\mathrm{P}_{T_{x^\star}}\| = 0,\end{aligned}$$

which entails $\|W(U_k - U)r_{k+1}\| = o(\|r_{k+1}\|) = o(\|d_k\|)$. Again, since $\gamma_k \to \gamma$,

$$\begin{aligned}\lim_{k\to\infty}\|M_{k,1}d_k\|/\|d_k\| &\leq \lim_{k\to\infty}(1+b)\|W\|\|G_k - G\|(\|r_k\| + \|r_{k-1}\|)/\|d_k\| \\ &\leq \lim_{k\to\infty}(1+b)\|W\||\gamma_k-\gamma|\|\mathrm{P}_{T_{x^\star}}\nabla^2 F(x^\star)\mathrm{P}_{T_{x^\star}}\|\sqrt{2}\|d_k\|/\|d_k\| = 0.\end{aligned}$$

Similarly, for $M_{k,2}$, since $a_k \to a, b_k \to b$,

$$\lim_{k\to\infty}\|M_{k,2}d_k\|/\|d_k\| \leq \lim_{k\to\infty}(|a_k - a| + |b_k - b|)\|W_k(\mathrm{Id} + G_k)\|\sqrt{2}\|d_k\|/\|d_k\| = 0,$$

where W_k, G_k are bounded. ∎

Proof (Proposition 15). (1.3) and the first-order optimality condition for problem ($\mathcal{P}_{\mathrm{opt}}$) are respectively equivalent to

$$y_{a,k} - x_{k+1} - \gamma_k\big(\nabla F(y_{b,k}) - \nabla F(x_{k+1})\big) \in \gamma_k\partial\Phi(x_{k+1}) \ \text{ and } \ 0 \in \gamma_k\partial\Phi(x^\star).$$

Projecting into $T_{x_{k+1}}$ and $T_{x^\star}$, respectively, and using Fact 7, leads to

$$\begin{aligned}\gamma_k\tau_{k+1}^{-1}\nabla_{\mathcal{M}_{x^\star}}\Phi(x_{k+1}) &= \tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}\big(y_{a,k} - x_{k+1} - \gamma_k\big(\nabla F(y_{b,k}) - \nabla F(x_{k+1})\big)\big) \\ \gamma_k\nabla_{\mathcal{M}_{x^\star}}\Phi(x^\star) &= 0.\end{aligned}$$

Adding both identities, and subtracting $\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}x^\star$ on both sides, we arrive at

$$\begin{aligned}&\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}r_{k+1} + \gamma_k\big(\tau_{k+1}^{-1}\nabla_{\mathcal{M}_{x^\star}}\Phi(x_{k+1}) - \nabla_{\mathcal{M}_{x^\star}}\Phi(x^\star)\big) \\ &= \tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}(y_{a,k} - x^\star) - \gamma_k\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}\big(\nabla F(y_{b,k}) - \nabla F(x_{k+1})\big).\end{aligned} \tag{B.10}$$

In view of Lemma 26, we get

$$\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}r_{k+1} = \mathrm{P}_{T_{x^\star}}r_{k+1} + (\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}} - \mathrm{P}_{T_{x^\star}})r_{k+1} = \mathrm{P}_{T_{x^\star}}r_{k+1} + o(\|r_{k+1}\|).$$

Using [37, Lemma 5.1], we have

$$\begin{aligned}&r_{k+1} = \mathrm{P}_{T_{x^\star}}r_{k+1} + o(\|r_{k+1}\|) \\ \Rightarrow\ &\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}r_{k+1} = r_{k+1} + o(\|r_{k+1}\|) = r_{k+1} + o(\|d_k\|),\end{aligned} \tag{B.11}$$

where we also used (B.6). Similarly

$$\begin{aligned}\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}(y_{a,k} - x^\star) &= \mathrm{P}_{T_{x^\star}}(y_{a,k} - x^\star) + (\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}} - \mathrm{P}_{T_{x^\star}})(y_{a,k} - x^\star) \\ &= \mathrm{P}_{T_{x^\star}}(y_{a,k} - x^\star) + o(\|y_{a,k} - x^\star\|) \\ &= \mathrm{P}_{T_{x^\star}}(y_{a,k} - x^\star) + o(\|d_k\|) \\ &= (1+a_k)\mathrm{P}_{T_{x^\star}}r_k - a_k\mathrm{P}_{T_{x^\star}}r_{k-1} + o(\|d_k\|) \\ &= (1+a_k)r_k - a_k r_{k-1} + o(\|r_k\|) + o(\|r_{k-1}\|) + o(\|d_k\|) \\ &= (y_{a,k} - x^\star) + o(\|d_k\|).\end{aligned} \tag{B.12}$$

Moreover owing to Lemma 27 and (B.6),

$$\tag{B.13} \begin{aligned} \tau^{-1}\nabla_{\mathcal{M}_{x^\star}}\Phi(x_{k+1}) - \nabla_{\mathcal{M}_{x^\star}}\Phi(x^\star) &= \nabla^2_{\mathcal{M}_{x^\star}}\Phi(x^\star)\mathrm{P}_{T_{x^\star}} r_{k+1} + o(\|r_{k+1}\|) \\ &= \nabla^2_{\mathcal{M}_{x^\star}}\Phi(x^\star)\mathrm{P}_{T_{x^\star}} r_{k+1} + o(\|d_k\|). \end{aligned}$$

Therefore, inserting (B.11), (B.12) and (B.13) into (B.10), we obtain

$$\tag{B.14} \begin{aligned} &(\mathrm{Id} + \gamma_k \nabla^2_{\mathcal{M}_{x^\star}}\Phi(x^\star)\mathrm{P}_{T_{x^\star}})r_{k+1} \\ &= (y_{a,k} - x^\star) - \gamma_k \tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}\big(\nabla F(y_{b,k}) - \nabla F(x_{k+1})\big) + o(\|d_k\|). \end{aligned}$$

Owing to (B.6) and local C^2-smoothness of F, we have

$$\tag{B.15} \begin{aligned} &\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}\big(\nabla F(y_{b,k}) - \nabla F(x_{k+1})\big) \\ &= \mathrm{P}_{T_{x^\star}}\big(\nabla F(y_{b,k}) - \nabla F(x_{k+1})\big) + o(\|d_k\|) \\ &= \mathrm{P}_{T_{x^\star}}\big(\nabla F(y_{b,k}) - \nabla F(x^\star)\big) - \mathrm{P}_{T_{x^\star}}\big(\nabla F(x_{k+1}) - \nabla F(x^\star)\big) + o(\|d_k\|) \\ &= \mathrm{P}_{T_{x^\star}}\nabla^2 F(x^\star)\mathrm{P}_{T_{x^\star}}(y_{b,k} - x^\star) - \mathrm{P}_{T_{x^\star}}\nabla^2 F(x^\star)\mathrm{P}_{T_{x^\star}}(x_{k+1} - x^\star) + o(\|d_k\|). \end{aligned}$$

Injecting (B.15) into (B.14), we get

$$\tag{B.16} \begin{aligned} &(\mathrm{Id} + \gamma_k \nabla^2_{\mathcal{M}_{x^\star}}\Phi(x^\star)\mathrm{P}_{T_{x^\star}} - \gamma_k \mathrm{P}_{T_{x^\star}}\nabla^2 F(x^\star)\mathrm{P}_{T_{x^\star}})r_{k+1} \\ &= (\mathrm{Id} + U_k)r_{k+1} = (y_{a,k} - x^\star) - H_k(y_{b,k} - x^\star) + o(\|d_k\|), \end{aligned}$$

which can be further written as,

$$\begin{aligned} (\mathrm{Id} + U_k)r_{k+1} &= (\mathrm{Id} + U)r_{k+1} + (U_k - U)r_{k+1} = (y_{a,k} - x^\star) - H_k(y_{b,k} - x^\star) + o(\|d_k\|) \\ &= \big((1+a_k)r_k - a_k r_{k-1}\big) - H_k\big((1+b_k)r_k - b_k r_{k-1}\big) + o(\|d_k\|) \\ &= \big((1+a_k)r_k - (1+b_k)H_k r_k\big) - \big(a_k r_{k-1} - b_k H_k r_{k-1}\big) + o(\|d_k\|) \\ &= \big((a_k - b_k)\mathrm{Id} + (1+b_k)G_k\big)r_k - \big((a_k - b_k)\mathrm{Id} + b_k G_k\big)r_{k-1} + o(\|d_k\|) \\ &= \big[(a_k - b_k)\mathrm{Id} + (1+b_k)G_k \quad -\big((a_k - b_k)\mathrm{Id} + b_k G_k\big)\big]\, d_k + o(\|d_k\|). \end{aligned}$$

Inverting $\mathrm{Id} + U$ (which is possible thanks to Lemma 13), we obtain

$$\begin{aligned} &r_{k+1} + W(U_k - U)r_{k+1} \\ &= \big[(a_k - b_k)W + (1+b_k)WG_k \quad -(a_k - b_k)W - b_k W G_k\big]\, d_k + o(\|d_k\|). \end{aligned}$$

Using the estimates (B.7), we get

$$\begin{aligned} d_{k+1} &= \begin{bmatrix} (a_k - b_k)W + (1+b_k)WG_k & -(a_k - b_k)W - b_k W G_k \\ \mathrm{Id} & 0 \end{bmatrix} d_k + o(\|d_k\|) \\ &= \Big(M + \begin{bmatrix} M_{k,1} \\ 0 \end{bmatrix} + \begin{bmatrix} M_{k,2} \\ 0 \end{bmatrix}\Big) d_k + o(\|d_k\|) = M d_k + o(\|d_k\|). \end{aligned}$$

□

Proof (Proposition 17).

(i) We have

$$M\begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} (a-b)r_1 + (1+b)Gr_1 - (a-b)r_2 - bGr_2 \\ r_1 \end{pmatrix} = \sigma \begin{pmatrix} r_1 \\ r_2 \end{pmatrix},$$

and thus $r_1 = \sigma r_2$. Inserting this in the first identity, we obtain

$$\begin{aligned} \sigma^2 r_2 &= (a-b)\sigma r_2 + (1+b)\sigma G r_2 - (a-b)r_2 - bGr_2 \\ \Leftrightarrow Gr_2 &= \Big(\big((a-b)(1-\sigma) + \sigma^2\big)/\big((1+b)\sigma - b\big)\Big) r_2 = \eta r_2 \\ \Rightarrow 0 &= \sigma^2 - \big((a-b) + (1+b)\eta\big)\sigma + (a-b) + b\eta. \end{aligned}$$

(ii) For this quadratic equation of σ, the two roots are

$$\tag{B.17} \sigma_1 = \big((a-b) + (1+b)\eta + \sqrt{\Delta_\sigma}\big)/2, \ \ \sigma_2 = \big((a-b) + (1+b)\eta - \sqrt{\Delta_\sigma}\big)/2.$$

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)
$$a_1 = (1-\eta)b - \eta, \;\; a_3 = (1-\eta)b - (1+\eta)/2$$
$$a_2 = (1-\eta)b + (1-\sqrt{1-\eta})^2 \begin{cases} \Delta_\sigma \leq 0 : a \in [a_2, (1-\eta)b + (1+\sqrt{1-\eta})^2], \\ \Delta_\sigma \geq 0 : a \leq a_2. \end{cases}$$

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 η.

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

Subcase $a \in [a_2, 1[$**:** $\sigma_{1,2}$ are complex, hence

$$|\sigma|^2 = \big(((a-b)+(1+b)\eta)^2 - \Delta_\sigma\big)/4 = a - b + b\eta. \tag{B.19}$$

As $a_2 \leq 1 \Leftrightarrow b \leq \frac{1-(1-\sqrt{1-\eta})^2}{1-\eta}$, 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 σ_2 has the bigger absolute value, then

(B.20)
$$|\sigma_2| < 1 \Leftrightarrow -\big((a-b)+(1+b)\eta\big) + \sqrt{\Delta_\sigma} < 2 \Leftrightarrow \frac{2(b-a)-1}{1+2b} < \eta,$$

which means $|\sigma_2| \leq 1$ for $a \in [a_3, a_2]$, and $|\sigma_2| \geq 1$ for $a \in [0, a_3]$. Moreover, $a_3 \leq 0$ for $b \in [0, \frac{1+\eta}{2(1-\eta)}]$, meaning that if $\eta \geq \frac{1}{3}$, $|\sigma_2| \leq 1$ for $a \in [0, a_2]$.

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

$$a_1 = 0 \Leftrightarrow b = \frac{\eta}{1-\eta} \begin{cases} \leq 1 : \eta \in [0, 0.5], \\ \geq 1 : \eta \in [0.5, 1[. \end{cases}$$

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

$$\sigma_1 < 1 \Leftrightarrow \big((a-b)+(1+b)\eta\big) + \sqrt{\Delta_\sigma} < 2 \Leftrightarrow 0 < 4(1-\eta). \tag{B.21}$$

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 $\nabla_{\mathcal{M}_{x^\star}} \Phi(x_k)$ is locally constant along $\mathcal{M}_{x^\star} = x^\star + T_{x^\star}$ around $x^\star$ (see Remark 9(iii)). Thus, embarking from (B.16) in the proof of Proposition 15, for k large enough, we get

$$x_{k+1} - x^\star = (y_{a,k} - x^\star) - E_k(y_{b,k} - x^\star),$$

where we used the mean-value theorem with $E_k = \gamma_k \int_0^1 \nabla^2 F(x^\star + t(y_{b,k} - x^\star))dt \succeq 0$. Using that E_k is symmetric and $\mathrm{Im}(E_k)^\perp = V$, we have

$$\mathrm{P}_V(x_{k+1} - x^\star) = \mathrm{P}_V(y_{a,k} - x^\star) = (1+a_k)\mathrm{P}_V(x_k - x^\star) - a_k(x_{k-1} - x^\star).$$

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

$$\begin{pmatrix} \mathrm{P}_V(x_{k+1} - x^\star) \\ \mathrm{P}_V(x_k - x^\star) \end{pmatrix} = \begin{bmatrix} (1+a_k)\mathrm{Id} & -a_k\mathrm{Id} \\ \mathrm{Id} & 0 \end{bmatrix} \begin{pmatrix} \mathrm{P}_V(x_k - x^\star) \\ \mathrm{P}_V(x_{k-1} - x^\star) \end{pmatrix}.$$

It is straightforward to check that $N_k \stackrel{\mathrm{def}}{=} \begin{bmatrix} (1+a_k)\mathrm{Id} & -a_k\mathrm{Id} \\ \mathrm{Id} & 0_n \end{bmatrix}$ is invertible and admits two eigenvalues $a_k > 0$ and 1 respectively. Iterating the above argument, and owing to the fact that $x_k, y_{a,k}, y_{b,k} \to x^\star$, we get

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \left(\prod_{j=k}^{\infty} N_j\right) \begin{pmatrix} \mathrm{P}_V(x_k - x^\star) \\ \mathrm{P}_V(x_{k-1} - x^\star) \end{pmatrix},$$

and $\prod_{j=k}^{\infty} N_j$ is invertible. Therefore, we obtain that $x_k - x^\star \in V^\perp$, and in turn, $y_{a,k} - x^\star \in V^\perp$ and $y_{b,k} - x^\star \in V^\perp$, for all large enough k. Observe that $V^\perp \subset T_{x^\star}$, it then follows that

$$x_{k+1} - x^\star = y_{a,k} - x^\star - \mathrm{P}_{V^\perp} E_k \mathrm{P}_{V^\perp}(y_{b,k} - x^\star).$$

By definition, $\mathrm{P}_{V^\perp} E_k \mathrm{P}_{V^\perp}$ 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. ∎

REFERENCES

[1] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*, Princeton University Press, 2009.

[2] P.-A. Absil, R. Mahony, and J. Trumpf, *An extrinsic look at the Riemannian Hessian*, in Geometric Science of Information, Springer, 2013, pp. 361–368.

[3] A. Agarwal, S. Negahban, and M. J. Wainwright, *Fast global convergence of gradient methods for high-dimensional statistical recovery*, The Annals of Statistics, 40 (2012), pp. 2452–2482.

[4] F. Alvarez, *On the minimizing property of a second order dissipative system in Hilbert spaces*, SIAM Journal on Control and Optimization, 38 (2000), pp. 1102–1119.

[5] F. Alvarez and H. Attouch, *An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping*, Set-Valued Analysis, 9 (2001), pp. 3–11.

[6] H. Attouch, Z. Chbani, J. Peypouquet, and P. Redont, *Fast convergence of inertial dynamics and algorithms with asymptotic vanishing damping*, Tech. Report Optimization online 5179, 2015.

[7] H. Attouch and J. Peypouquet, *The rate of convergence of Nesterov's accelerated Forward–Backward method is actually $o(k^{-2})$*, Tech. Report arXiv:1510.08740, 2015.

[8] H. Attouch, J. Peypouquet, and P. Redont, *A dynamical approach to an inertial Forward–Backward algorithm for convex minimization*, SIAM J. Optim., 24 (2014), pp. 232–256.

[9] H. Attouch, J. Peypouquet, and P. Redont, *On the fast convergence of an inertial gradient-like dynamics with vanishing viscosity*, Tech. Report arXiv:1507.04782, 2015.

[10] J.-F. Aujol and C. Dossal, *Stability of over-relaxations for the Forward–Backward algorithm, application to FISTA*, SIAM Journal on Optimization, 25 (2015), pp. 2408–2433.

[11] J. B. Baillon and G. Haddad, *Quelques propriétés des opérateurs angle-bornés etn-cycliquement monotones*, Israel Journal of Mathematics, 26 (1977), pp. 137–150.

[12] H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, 2011.

[13] A. Beck and M. Teboulle, *Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems*, Image Processing, IEEE Transactions on, 18 (2009), pp. 2419–2434.

[14] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 183–202.

[15] J. Bolte, A. Daniilidis, and A. Lewis, *The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems*, SIAM J. Optim, 17 (2006), pp. 1205–1223.

[16] K. Bredies and D. A. Lorenz, *Linear convergence of iterative soft-thresholding*, Journal of Fourier Analysis and Applications, 14 (2008), pp. 813–837.

[17] R. S. Burachik and A. N. Iusem, *Set-valued Mappings and Enlargements of Monotone Operators*, Optimization and Its Applications, Springer, 2008.

[18] A. Chambolle and C. Dossal, *On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm"*, Journal of Optimization Theory and Applications, 166 (2015), pp. 968–982.

[19] I. Chavel, *Riemannian geometry: a modern introduction*, vol. 98, Cambridge University Press, 2006.

[20] P. L. Combettes, *Quasi-Fejérian analysis of some optimization algorithms*, Studies in Computational Mathematics, 8 (2001), pp. 115–152.

[21] P. L. Combettes and B. C. Vũ, *Variable metric Forward–Backward splitting with applications to monotone inclusions in duality*, Optimization, 63 (2014), pp. 1289–1318.

[22] A. Daniilidis, D. Drusvyatskiy, and A. S. Lewis, *Orthogonal invariance and identifiability*, SIAM Journal on Matrix Analysis and Applications, 35 (2014), pp. 580–598.

[23] A. Daniilidis, W. Hare, and J. Malick, *Geometrical interpretation of the predictor-corrector type algorithms in structured optimization problems*, Optimization: A Journal of Mathematical Programming & Operations Research, 55 (2009), pp. 482–503.

[24] T. Goldstein, B. O'Donoghue, S. Setzer, and R. Baraniuk, *Fast alternating direction optimization methods*, SIAM Journal on Imaging Sciences, 7 (2014), pp. 1588–1623.

[25] M. Gu, L.-H. Lim, and C. J. Wu, *ParNes: a rapidly convergent algorithm for accurate recovery of sparse and approximately sparse signals*, Numerical Algorithms, 64 (2012), pp. 321–347.

[26] E. Hale, W. Yin, and Y. Zhang, *Fixed-point continuation for ℓ_1-minimization: methodology and convergence*, SIAM Journal on Optimization, 19 (2008), pp. 1107–1130.

[27] W. L. Hare, *Identifying active manifolds in regularization problems*, in Fixed-Point Algorithms for Inverse Problems in Science and Engineering, H. H. Bauschke, R. S., Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, eds., vol. 49 of Springer Optimization and Its Applications, Springer, 2011, ch. 13.

[28] W. L. Hare and A. S. Lewis, *Identifying active constraints via partial smoothness and prox-regularity*, Journal of Convex Analysis, 11 (2004), pp. 251–266.

[29] W. L. Hare and A. S. Lewis, *Identifying active manifolds*, Algorithmic Operations Research, 2 (2007), pp. 75–82.

[30] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis And Minimization Algorithms*, vol. I and II, Springer, 2001.

[31] K. Hou, Z. Zhou, A. M.-C. So, and Z. Luo, *On the linear convergence of the proximal gradient method for trace norm regularization*, in Advances in Neural Information Processing Systems, 2013, pp. 710–718.

[32] P. R. Johnstone and P. Moulin, *A Lyapunov analysis of FISTA with local linear convergence for sparse optimization*, arXiv preprint arXiv:1502.02281, (2015).

[33] J. M. Lee, *Smooth manifolds*, Springer, 2003.

[34] C. Lemaréchal, F. Oustry, and C. Sagastizábal, *The U-Lagrangian of a convex function*, Trans. Amer. Math. Soc., 352 (2000), pp. 711–729.

[35] A. S. Lewis, *Active sets, nonsmoothness, and sensitivity*, SIAM Journal on Optimization, 13 (2003), pp. 702–725.

[36] A. S. Lewis and S. Zhang, *Partial smoothness, tilt stability, and generalized Hessians*, SIAM Journal on Optimization, 23 (2013), pp. 74–94.

[37] J. Liang, J. Fadili, and G. Peyré, *Local linear convergence of Forward–Backward under partial smoothness*, in Advances in Neural Information Processing Systems, 2014, pp. 1970–1978.

[38] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM Journal on Numerical Analysis, 16 (1979), pp. 964–979.

[39] D. A. Lorenz and T. Pock, *An accelerated Forward–Backward algorithm for monotone inclusions*, arXiv preprint arXiv:1403.3522, (2014).

[40] S. A. Miller and J. Malick, *Newton methods for nonsmooth convex minimization: connections among-Lagrangian, Riemannian Newton and SQP methods*, Mathematical programming, 104 (2005), pp. 609–633.

[41] B. Mordukhovich, *Sensitivity analysis in nonsmooth optimization*, Theoretical Aspects of Industrial Design (D. A. Field and V. Komkov, eds.), SIAM Volumes in Applied Mathematics, 58 (1992), pp. 32–46.

[42] A. Moudafi and M. Oliny, *Convergence of a splitting inertial proximal method for monotone operators*, Journal of Computational and Applied Mathematics, 155 (2003), pp. 447–454.

[43] Y. Nesterov, *A method for solving the convex programming problem with convergence rate $O(1/k^2)$*, Dokl. Akad. Nauk SSSR, 269 (1983), pp. 543–547.

[44] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer, 2004.

[45] Y. Nesterov, *Gradient methods for minimizing composite objective function*, (2007).

[46] B. O'Donoghue and E. Candes, *Adaptive restart for accelerated gradient schemes*, Foundations of computational mathematics, 15 (2015), pp. 715–732.

[47] Z. Opial, *Weak convergence of the sequence of successive approximations for nonexpansive mappings*, Bulletin of the American Mathematical Society, 73 (1967), pp. 591–597.

[48] R. A. Poliquin and R. T. Rockafellar, *Tilt stability of a local minimum*, SIAM Journal on Optimization, 8 (1998), pp. 287–299.
[49] B. T. Polyack, *Some methods of speeding up the convergence of iterative methods*, Zh. Vychisl. Mat. Mat. Fiz., 4 (1964), pp. 1–17.
[50] B. T. Polyak, *Introduction to optimization*, Optimization Software, 1987.
[51] R. T. Rockafellar and R. Wets, *Variational analysis*, vol. 317, Springer Verlag, 1998.
[52] S. T. Smith, *Optimization techniques on Riemannian manifolds*, Fields institute communications, 3 (1994), pp. 113–135.
[53] B. F. Svaiter and R. S. Burachik, *ε-enlargements of maximal monotone operators in banach spaces*, Set-Valued Anal., 7 (1999), pp. 117–132.
[54] S. Tao, D. Boley, and S. Zhang, *Local linear convergence of ISTA and FISTA on the LASSO problem*, SIAM Journal on Optimization, 26 (2016), pp. 313–336.
[55] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Math. Prog. (Ser. B), 117 (2009).
[56] S. Vaiter, M. Golbabaee, J. Fadili, and G. Peyré, *Model selection with low complexity priors*, Information and Inference, (2015), p. iav005.
[57] S. Vaiter, G. Peyré, and J. Fadili, *Model consistency of partly smooth regularizers*, arXiv preprint arXiv:1405.1004, (2014).
[58] S. J. Wright, *Identifiable surfaces in constrained optimization*, SIAM Journal on Control and Optimization, 31 (1993), pp. 1063–1079.