



HAL
open science

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

Jingwei Liang, Jalal M. Fadili, Gabriel Peyré

► **To cite this version:**

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

HAL Id: hal-01658850

<https://hal.science/hal-01658850v1>

Submitted on 7 Dec 2017

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

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

3 JINGWEI LIANG[†], JALAL FADILI[†], AND GABRIEL PEYRÉ[‡]

4 **Abstract.** In this paper, we consider a class of Forward–Backward (FB) splitting methods that
5 includes several variants (*e.g.* inertial schemes, FISTA) for minimizing the sum of two proper convex
6 and lower semi-continuous functions, one of which has a Lipschitz continuous gradient, and the other
7 is partly smooth relative to a smooth active manifold \mathcal{M} . We propose a unified framework, under
8 which we show that, this class of FB-type algorithms (i) correctly identifies the active manifold in a
9 finite number of iterations (finite activity identification), and (ii) then enters a local linear convergence
10 regime, which we characterize precisely in terms of the structure of the underlying active manifold.
11 We also establish and explain why FISTA (with convergent sequences) locally oscillates and can
12 be locally slower than FB. These results may have numerous applications including in signal/image
13 processing, sparse recovery and machine learning. Indeed, the obtained results explain the typical
14 behaviour that has been observed numerically for many problems in these fields such as the Lasso,
15 the group Lasso, the fused Lasso and the nuclear norm minimization to name only a few.

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

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

19 **1. Introduction.**

20 **1.1. Non-smooth optimization.** In various fields of science and engineering,
21 such as signal/image processing, inverse problems and machine learning, many prob-
22 lems can be cast as solving a *structured composite non-smooth optimization problem*
23 of the sum of two functions, which usually reads

24 $(\mathcal{P}_{\text{opt}}) \quad \min_{x \in \mathbb{R}^n} \Phi(x) \stackrel{\text{def}}{=} F(x) + R(x),$

25 where

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

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

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

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

32 $(\mathcal{P}_{\text{inc}}) \quad \text{Find } x \in \mathbb{R}^n \text{ such that } 0 \in A(x) + B(x),$

33 where

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

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

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

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

39 $0 \in \partial R(x^*) + \nabla F(x^*),$

40 where ∂R denotes the sub-differential of R at x^* . Clearly, if we let $A = \partial R$ and
41 $B = \nabla F$, then $(\mathcal{P}_{\text{opt}})$ is simply a special case of $(\mathcal{P}_{\text{inc}})$.

*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).

42 In this paper, our main focus is the non-smooth optimization problem (\mathcal{P}_{opt}).
 43 Though some of our results are also valid for the monotone inclusion problem (\mathcal{P}_{inc}),
 44 in particular Algorithm 1 and its global convergence analysis, see Section 2.

45 **1.2. Forward–Backward-type splitting methods.** The Forward–Backward
 46 (FB) splitting method [38] is a powerful tool for solving optimization problems (\mathcal{P}_{opt})
 47 with the additively separable and “smooth + non-smooth” structure. The standard
 48 (non-relaxed) version of FB implements the iterative scheme

$$49 \quad (1.1) \quad x_{k+1} = \text{prox}_{\gamma_k R}(x_k - \gamma_k \nabla F(x_k)), \quad \gamma_k \in [\underline{\epsilon}, 2\beta - \bar{\epsilon}],$$

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

$$51 \quad \text{prox}_{\gamma R}(\cdot) \stackrel{\text{def}}{=} \min_{x \in \mathbb{R}^n} \frac{1}{2} \|x - \cdot\|^2 + \gamma R(x).$$

52 Global convergence of the sequence $(x_k)_{k \in \mathbb{N}}$ generated by the FB method is
 53 well-established in the literature, based on the property that the composed opera-
 54 tor $\text{prox}_{\gamma R}(\text{Id} - \gamma \nabla F)$ is so-called averaged non-expansive [12]. Moreover, sub-linear
 55 $O(1/k)$ convergence rate of the sequence of objective values of FB is also well-known,
 56 e.g. [45, 16, 14].

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

67 On the other hand, in the context of convex optimization, the accelerated FISTA
 68 method was proposed in [14], based on the seminal work [43], which achieves $O(1/k^2)$
 69 convergence rate for the sequence of objective functions. However, while iterates
 70 generated by the FB are convergent, the convergence of FISTA iterates has been an
 71 open problem until recently. This question was first settled in [18], then followed by [9]
 72 in the continuous dynamical system case. More precisely, for $\gamma_k \in]0, \beta]$ and a sequence
 73 of inertial parameter that converges at an appropriate rate (i.e. in Algorithm 1, set
 74 $a_k = b_k = \frac{k-1}{k+q}$, $q > 2$), these authors established (weak in infinite-dimensional Hilbert
 75 spaces) convergence of the iterates sequence while maintaining the $O(1/k^2)$ rate on
 76 the objective values. The rate is actually even $o(1/k^2)$ as proved in [7].

Algorithm 1 A General Inertial Forward–Backward splitting

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

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

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

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

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

- 81 • $a_k = 0$, $b_k = 0$: this is the original FB method [38];
- 82 • $a_k \in [0, \bar{a}]$, $b_k = 0$: this is the case studied in [42] for (\mathcal{P}_{inc}). In the context
 83 of optimization with $R = 0$, one recovers the heavy ball method with friction

84 [49];

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

88 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

89 to the best of our knowledge.

90 **REMARK 1.** *Though Algorithm 1 is stated for the optimization problem $(\mathcal{P}_{\text{opt}})$, it*

91 *readily extends to the monotone inclusion problem $(\mathcal{P}_{\text{inc}})$, for which step (1.3) reads*

$$92 \quad (1.4) \quad x_{k+1} = J_{\gamma_k A}(y_{a,k} - \gamma_k B(y_{b,k})),$$

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

94 For the rest of the paper, we use the terminology *FB-type methods* for any scheme

95 in the form of Algorithm 1 such that the sequence $(x_k)_{k \in \mathbb{N}}$ converges. This will encom-

96 pass the inertial schemes (denoted iFB) that we propose, and the sequence convergent

97 FISTA method [18, 9] that corresponds to the specific choice of inertial sequences

98 $a_k = b_k = \frac{k-1}{k+q}$, $q > 2$. It should be noted, however, that our global convergence

99 analysis to be presented in Section 2 does not cover the case of FISTA, which requires

100 a specific proof strategy as developed in [18, 9].

101 **1.3. Contributions.** The study of (local) linear convergence of FB-type meth-

102 ods in the absence of strong convexity has attracted increasing interest in recent years,

103 see the related work below for details. In general, most of the existing work focuses

104 on some special cases (*e.g.* $R = \|\cdot\|_1$ in $(\mathcal{P}_{\text{opt}})$), and the proofs of the results heavily

105 rely on the specific structure of the function R , which makes them rather difficult to

106 extend to other cases. Therefore, it is important to present a unified analysis frame-

107 work, and possibly with stronger claims. This is one of the main motivations of this

108 work. To be more precise, this paper delivers the following contributions:

109 *A general class of inertial algorithms.* We present a unified iFB splitting class

110 of algorithms for solving $(\mathcal{P}_{\text{opt}})$. It can be viewed as a versatile explicit-implicit

111 discretization of a nonlinear second-order dynamical system with viscous damping,

112 and thus covers existing methods as special cases. We establish global convergence of

113 the iterates, and also stability to errors.

114 *Finite activity identification.* Under the additional assumption that function R

115 is partly smooth at $x^* \in \text{Argmin}(\Phi)$ relative to a C^2 -smooth manifold \mathcal{M}_{x^*} (see

116 Definition 5) and a *non-degeneracy condition* at x^* , we show that any FB-type method

117 to solve $(\mathcal{P}_{\text{opt}})$ has the *finite time activity identification property*. Meaning that, after

118 a finite number of iterations, say K , the iterates $x_k \rightarrow x^*$ built by the FB-type method

119 belong to \mathcal{M}_{x^*} for all $k \geq K$.

120 *Local linear convergence.* Exploiting this identification property, we then show

121 that the FB-type methods, locally along the manifold \mathcal{M}_{x^*} , exhibit a linear conver-

122 gence regime. We characterize this regime and the corresponding rates precisely de-

123 pending on the structure of the active manifold \mathcal{M}_{x^*} . For instance, we provide sharp

124 estimates for the convergence rate. For the sequence convergent FISTA method, we

125 draw two major conclusions:

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

129 This gives an enlightening explanation of the usefulness of the so-called restarting

130 method to locally accelerate the convergence of FISTA used by many authors, for

131 instance in sparse recovery [25, 46, 24]: the algorithm is restarted after a certain

132 number of iterations (set more or less empirically), where the inertial sequence $a_k = b_k$
 133 is reset to 0.

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

139 **1.4. Related work.** Finite support identification and local linear convergence
 140 of FB for solving a special instance of $(\mathcal{P}_{\text{opt}})$ where R is the ℓ_1 -norm is established
 141 in [16, 26]. The same question has been recently addressed for FISTA under some
 142 constraints on the inertial parameter in [54, 32]. [3] proved local linear convergence
 143 of FB to solve $(\mathcal{P}_{\text{opt}})$ for R being a so-called convex decomposable regularizer. Local
 144 linear convergence of FB is studied in [31] for R the nuclear norm and F locally
 145 strongly convex. All these previous functions are subclass of partly smooth functions,
 146 and their results are thus covered by ours under weaker assumptions. The proposed
 147 work is also a deeper and sharper extension of our previous results on FB [37]. Finite
 148 identification of active manifolds associated to partly smooth functions has been shown
 149 in [28, 29, 27] for the (sub)gradient projection method, Newton-like methods, the
 150 proximal point algorithm and the algorithm in [55]. Their work extends that of *e.g.*
 151 [58] on identifiable surfaces (see references therein for related work of Dunn, and Burke
 152 and Moré). However, in all these works, the local linear convergence behaviour was
 153 not addressed.

154 **1.5. Notations.** Throughout the paper, Id denotes the identity operator on \mathbb{R}^n .
 155 For a nonempty convex set $\Omega \subset \mathbb{R}^n$, $\text{ri}(\Omega)$ and $\text{rbd}(\Omega)$ denote its relative interior
 156 and boundary respectively, $\text{aff}(\Omega)$ is its affine hull, and $\text{par}(\Omega) = \mathbb{R}(\Omega - \Omega)$ is the
 157 subspace parallel to it. Denote ι_Ω the indicator function of Ω , σ_Ω its support function
 158 and P_Ω the orthogonal projector onto Ω . For a matrix M , $\ker(M)$ is its null-space.
 159 The subdifferential of a function $R \in \Gamma_0(\mathbb{R}^n)$ is the set-valued operator $\partial R : \mathbb{R}^n \rightrightarrows$
 160 \mathbb{R}^n , $x \mapsto \{u \in \mathbb{R}^n \mid R(z) \geq R(x) + \langle u, z - x \rangle, \forall z \in \mathbb{R}^n\}$.

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

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

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

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

$$177 \quad A^\varepsilon(x) \stackrel{\text{def}}{=} \{v \in \mathbb{R}^n, \langle u - v, y - x \rangle \geq -\varepsilon, \forall y \in \mathbb{R}^n, u \in A(y)\}.$$

178 Denote $\partial^\varepsilon R$ the ε -enlargement of ∂R . We now consider an inexact form of the

179 iFB algorithm where step (1.3) is replaced by finding x_{k+1} such that

$$180 \quad (2.1) \quad y_{a,k} - \gamma_k(\nabla F(y_{b,k}) + \xi_k) - x_{k+1} \in \gamma_k \partial^{\varepsilon_k} R(x_{k+1}),$$

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

185 **THEOREM 3 (Conditional convergence).** *Consider Algorithm 1 with the inexact it-*
186 *eration (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*
187 *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*

$$188 \quad (2.2) \quad \sum_{k \in \mathbb{N}} \max\{a_k, b_k\} \|x_k - x_{k-1}\|^2 < +\infty,$$

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

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

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

$$197 \quad (2.3) \quad a_k = \min\{a, c_{a,k}\}, \quad b_k = \min\{b, c_{b,k}\},$$

198 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
199 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 $c_{b,k}$.

200 One can also devise choices of $(a_k)_{k \in \mathbb{N}}$ and $(b_k)_{k \in \mathbb{N}}$ that are independent of
201 $(x_k)_{k \in \mathbb{N}}$, and still guarantee global convergence. We dub this *unconditional con-*
202 *vergence*. The following result generalizes those in [5, 42, 39].

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

$$206 \quad (2.4) \quad \begin{cases} (1 + a_k) - \frac{\gamma_k}{2\beta}(1 + b_k)^2 > \tau : a_k < \frac{\gamma_k}{2\beta} b_k, \\ (1 - 3a_k) - \frac{\gamma_k}{2\beta}(1 - b_k)^2 > \tau : b_k \leq a_k \text{ or } \frac{\gamma_k}{2\beta} b_k \leq a_k < b_k, \end{cases}$$

207 *and, moreover $\sum_{k \in \mathbb{N}} \varepsilon_k < +\infty$ and $\sum_{k \in \mathbb{N}} \|\xi_k\| < +\infty$. Then $\sum_{k \in \mathbb{N}} \|x_k - x_{k-1}\|^2 <$
208 $+\infty$, and there exists $x^* \in \text{Argmin}(\Phi)$ such that the sequence $(x_k)_{k \in \mathbb{N}}$ converges to x^* .*

209 See Section A for the proof.

210 3. Partial smoothness and finite time activity identification.

211 **3.1. Partial smoothness.** From now on, besides assumption (H.1), we assume
212 that R in $(\mathcal{P}_{\text{opt}})$ is moreover *partly smooth* relative to a smooth manifold. The notion
213 of partial smoothness is first introduced in [35]. This concept, as well as that of identi-
214 fiable surfaces [58], captures the essential features of the geometry of non-smoothness
215 which are along the so-called active/identifiable manifold. For convex functions, a
216 closely related idea is developed in [34]. Loosely speaking, a partly smooth function
217 behaves smoothly as we move on the identifiable submanifold, and sharply if we move
218 normal to the manifold. In fact, the behaviour of the function and of its minimiz-
219 ers depend essentially on its restriction to this manifold, hence offering a powerful
220 framework for algorithmic and sensitivity analysis theory.

221 Let \mathcal{M}_x be a C^2 -smooth embedded submanifold of \mathbb{R}^n around a point x . To
 222 lighten terminology, henceforth we shall state C^2 -manifold instead of C^2 -smooth em-
 223 bedded submanifold of \mathbb{R}^n . The natural embedding of a submanifold \mathcal{M}_x into \mathbb{R}^n
 224 permits to define a Riemannian structure on \mathcal{M}_x , and we simply say \mathcal{M}_x is a Rie-
 225 mannian manifold. $\mathcal{T}_{\mathcal{M}_x}(x')$ denotes the tangent space to \mathcal{M}_x at any point x' near x
 226 in \mathcal{M}_x . More materials on manifolds are given in Section B.1.

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

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

- 231 (i) **Smoothness:** \mathcal{M}_x is a C^2 -manifold around x , R restricted to \mathcal{M}_x is C^2 near x ;
 232 (ii) **Sharpness:** The tangent space $\mathcal{T}_{\mathcal{M}_x}(x)$ coincides with $T_x \stackrel{\text{def}}{=} \text{par}(\partial R(x))^\perp$;
 233 (iii) **Continuity:** The set-valued mapping ∂R is continuous at x relative to \mathcal{M}_x .

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

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

245 [35, Proposition 2.10] allows to prove the following fact.

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

248 We now give expressions of the Riemannian gradient and Hessian (see Section B.1
 249 for definitions) for the case of partly smooth functions relative to a C^2 submanifold.
 250 This is summarized in the following fact which follows by combining (B.2), (B.3),
 251 Definition 5, Fact 6 and [23, Proposition 17] (or [40, Lemma 2.4]).

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

$$253 \quad \nabla_{\mathcal{M}_x} R(x') = P_{T_{x'}}(\partial R(x')),$$

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

$$256 \quad \nabla_{\mathcal{M}_x}^2 G(x')h = P_{T_{x'}} \nabla^2 \tilde{R}(x')h + \mathfrak{W}_{x'}(h, P_{T_{x'}^\perp} \nabla \tilde{R}(x')),$$

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

259 **3.2. Finite time activity identification.** In this section, we state our result
 260 establishing that FB-type methods have the finite activity identification property.

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

$$264 \quad (\text{ND}) \quad -\nabla F(x^*) \in \text{ri}(\partial R(x^*)),$$

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

266 *If moreover,*

- 267 (i) \mathcal{M}_{x^*} is an affine subspace, then $\mathcal{M}_{x^*} = x^* + T_{x^*}$ and $y_{a,k}, y_{b,k} \in \mathcal{M}_{x^*}, \forall k > K$;
 268 (ii) R is locally polyhedral around x^* , then $y_{a,k}, y_{b,k} \in \mathcal{M}_{x^*} = x^* + T_{x^*}$ for all $k > K$,
 269 $\nabla_{\mathcal{M}_{x^*}} R(x_k) = \nabla_{\mathcal{M}_{x^*}} R(x^*)$, and $\nabla_{\mathcal{M}_{x^*}}^2 R(x_k) = 0, \forall k \geq K$.

270 REMARK 9.

- 271 (i) If F is also locally C^2 around x^* , the smooth perturbation rule of partly smooth
 272 functions [35, Corollary 4.7], ensures that $\Phi \in \text{PSF}_{x^*}(\mathcal{M}_{x^*})$.
 273 (ii) The iFB is convergent under the assumptions of Theorem 3 or Theorem 4.
 274 The FISTA method is sequence convergent for $a_k = b_k = \frac{k-1}{k+q}, q > 2$, and
 275 $\gamma_k \equiv \gamma \in]0, \beta]$ [18, 9]. Thus, Theorem 8 holds true for all these instances.
 276 (iii) The non-degeneracy condition (ND) can be viewed as a geometric general-
 277 ization of the strict complementarity of non-linear programming. Building
 278 on the arguments of [29], it is almost a necessary condition for the finite
 279 identification of \mathcal{M}_{x^*} . Relaxing it in general is a challenging problem.
 280 (iv) When R is locally polyhedral around x^* , in addition with the finite identifi-
 281 cation of $\mathcal{M}_{x^*} = x^* + T_{x^*}$, we also have $\nabla_{\mathcal{M}_{x^*}} \Phi(x_k) = \nabla_{\mathcal{M}_{x^*}} \Phi(x^*)$, hence
 282 $\nabla_{\mathcal{M}_{x^*}}^2 \Phi(x_k) = 0$, for k large enough.

283 *Proof.* By assumption, the sequence $(x_k)_{k \in \mathbb{N}}$ created by any FB-type method
 284 converges to some $x^* \in \text{Argmin}(\Phi)$, and the latter is non-empty by assumption (H.3).
 285 Now (1.3) is equivalent to

$$286 \quad y_{a,k} - \gamma_k \nabla F(y_{b,k}) - x_{k+1} \in \gamma_k \partial R(x_{k+1}).$$

287 By (H.2), we get

$$\begin{aligned} & \text{dist}(-\nabla F(x^*), \partial R(x_{k+1})) \\ & \leq \left\| \frac{1}{\gamma_k} (y_{a,k} - x_{k+1}) - \nabla F(y_{b,k}) + \nabla F(x^*) \right\| \\ 288 & \leq \frac{1}{\gamma_k} (a_k \|x_k - x_{k-1}\| + \|x_{k+1} - x_k\|) + \|\nabla F(y_{b,k}) - \nabla F(x^*)\| \\ & \leq \left(\frac{1}{\gamma_k} + \frac{1}{\beta} \right) \|x_k - x_{k-1}\| + \frac{1}{\gamma_k} \|x_{k+1} - x_k\| + \frac{1}{\beta} \|x_k - x^*\|. \end{aligned}$$

289 Since $\liminf \gamma_k = \underline{\epsilon} > 0$ and x_k converges to x^* , we obtain $\text{dist}(-\nabla F(x^*), \partial R(x_k)) \rightarrow$
 290 0. Owing to assumption (H.1), R is subdifferentially continuous at every point in
 291 its domain, and in particular at x^* for $-\nabla F(x^*)$, which in turn entails $R(x_k) \rightarrow$
 292 $R(x^*)$. Altogether, this shows that the conditions of [28, Theorem 5.3] are fulfilled on
 293 $\langle \nabla F(x^*), \cdot \rangle + R$, and the result follows.

- 294 (i) When the active manifold \mathcal{M}_{x^*} is an affine subspace, then $\mathcal{M}_{x^*} = x^* + T_{x^*}$ owing
 295 to the normal sharpness property and the claim follows immediately;
 296 (ii) When R is locally polyhedral around x^* , then \mathcal{M}_{x^*} is an affine subspace and the
 297 identification of $y_{a,k}, y_{b,k}$ follows from (i). For the rest, it is sufficient to observe
 298 that by polyhedrality, for any $x \in \mathcal{M}_{x^*}$ near x^* , $\partial R(x) = \partial R(x^*)$. Therefore,
 299 combining Fact 6 and Fact 7, we get the second conclusion. \square

300 *A bound on the identification iteration.* In Theorem 8, we have not provided an
 301 estimate $K \geq 0$ beyond which finite identification occurs. There is of course a situation
 302 where the answer is trivial, *i.e.* R is the indicator function of an affine subspace.
 303 However, knowing K has practical interest, for instance, if one wants to switch to
 304 higher order acceleration (see Section 4.5). It is then legitimate to wonder whether
 305 such an estimate of K can be given. In the following, we shall give a bound in some
 306 important cases. For the sake of simplicity, we state the result for the case of FB (*i.e.*
 307 $a_k = b_k \equiv 0$ in Algorithm 1). A similar reasoning can be easily generalized to the case
 308 of any converging FB-type method.

309 PROPOSITION 10. *Suppose that the assumptions of Theorem 8 hold. Then the*
 310 *following holds.*

311 (i) *If the iterates are such that $\partial R(x_k) \subset \text{rbd}(\partial R(x^*))$ whenever $x_k \notin \mathcal{M}_{x^*}$, then*

$$312 \quad x_k \in \mathcal{M}_{x^*} \text{ for all } k \geq \frac{\|x_0 - x^*\|^2}{\underline{\epsilon}^2 \text{dist}(-\nabla F(x^*), \text{rbd}(\partial R(x^*)))^2};$$

313 (ii) *If R is separable, i.e. $R(x) = \sum_{i=1}^m \sigma_{C_i}(x_{b_i})$, where $\forall 1 \leq i \leq m, b_i \subset \{1, \dots, n\}$,*
 314 *$\bigcup_{i=1}^m b_i = \{1, \dots, n\}$, and $b_i \cap b_j = \emptyset, \forall i \neq j$, and $\dim(C_i) = |b_i|$, then identifica-*

315 *tion of \mathcal{M}_{x^*} occurs for some k larger than $\frac{\|x_0 - x^*\|^2}{\underline{\epsilon}^2 \sum_{i \in I_{x^*}^c} \text{dist}(-\nabla F(x^*)_{b_i}, \text{rbd}(C_i))^2}$,*

316 *where $I_x \stackrel{\text{def}}{=} \{i : x_{b_i} \neq 0\}$.*

317 *Proof.* (i) By firm non-expansiveness of $\text{prox}_{\gamma_{k-1}R}$, and non-expansiveness of
 318 $\text{Id} - \gamma_{k-1}\nabla F$, we have

$$\begin{aligned} 319 \quad \|x_k - x^*\|^2 &\leq \|(\text{Id} - \gamma_{k-1}\nabla F)(x_{k-1}) - (\text{Id} - \gamma_{k-1}\nabla F)(x^*)\|^2 \\ &\quad - \|x_{k-1} - \gamma_{k-1}\nabla F(x_{k-1}) - x_k + \gamma_{k-1}\nabla F(x^*)\|^2 \\ &\leq \|x_{k-1} - x^*\|^2 - \underline{\epsilon}^2 \|u_k - \nabla F(x^*)\|^2, \end{aligned}$$

320 where we denoted $u_k \stackrel{\text{def}}{=} (x_{k-1} - x_k)/\gamma_{k-1} - \nabla F(x_{k-1})$. By definition, we have
 321 $u_k \in \partial R(x_k)$. Suppose that identification has not occurred at k , i.e. that $x_k \notin$
 322 \mathcal{M}_{x^*} , and hence $u_k \in \partial R(x_k) \subset \text{rbd}(\partial R(x^*))$. Therefore, continuing the above
 323 inequality, we get

$$\begin{aligned} 324 \quad \|x_k - x^*\|^2 &\leq \|x_{k-1} - x^*\|^2 - \underline{\epsilon}^2 \text{dist}(-\nabla F(x^*), \partial R(x_k))^2 \\ 325 \quad &\leq \|x_{k-1} - x^*\|^2 - \underline{\epsilon}^2 \text{dist}(-\nabla F(x^*), \text{rbd}(\partial R(x^*)))^2 \\ 326 \quad &\leq \|x_0 - x^*\|^2 - k\underline{\epsilon}^2 \text{dist}(-\nabla F(x^*), \text{rbd}(\partial R(x^*)))^2, \end{aligned}$$

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

332 (ii) We have $\partial \sigma_{C_i}(x_{b_i}^*) = C_i, \forall i \in I_{x^*}^c$. In turn, by separability, R is partly smooth
 333 at x^* relative to $\mathcal{M}_{x^*} = \bigtimes_{i=1}^m \mathcal{M}_{x_{b_i}^*}$, where $\mathcal{M}_{x_{b_i}^*} = 0$ if $i \in I_{x^*}^c$ and $\mathcal{M}_{x_{b_i}^*} \neq 0$
 334 otherwise. Suppose that at iteration k , $I_{x^*}^c \cap I_{x_k} \neq \emptyset$. Denote $h_{k-1} = x_{k-1} -$
 335 $\gamma_{k-1}\nabla F(x_{k-1})$, and $h^* = x^* - \gamma_{k-1}\nabla F(x^*)$. Thus for any $i \in I_{x^*}^c \cap I_{x_k}$, we have

$$\begin{aligned} 336 \quad x_{k,b_i} - x_{b_i}^* &= h_{k-1,b_i} - P_{\gamma_{k-1}C_i}(h_{k-1,b_i}) \\ 337 \quad &= (h_{k-1,b_i} - h_{b_i}^*) - (P_{\gamma_{k-1}C_i}(h_{k-1,b_i}) - P_{\gamma_{k-1}C_i}(h_{b_i}^*)) \end{aligned}$$

339 where we used Moreau identity in the first equality. Since $i \in I_{x_k} \cap I_{x^*}^c$, we
 340 have $h_{k-1,b_i} \notin \gamma_{k-1}C_i$ and $h_{b_i}^* \in \gamma_{k-1}C_i$, or equivalently, that $P_{\gamma_{k-1}C_i}(h_{k-1,b_i}) \in$
 341 $\gamma_{k-1}\text{rbd}(C_i) = \gamma_{k-1}\text{rbd}(\partial \sigma_{C_i}(x_{b_i}^*))$ and $P_{\gamma_{k-1}C_i}(h_{b_i}^*) = h_{b_i}^*$. Combining this with
 342 the fact that the orthogonal projector on $\gamma_{k-1}C_i$ is firmly non-expansive, we get

$$\begin{aligned} 343 \quad \|x_{k,b_i} - x_{b_i}^*\|^2 &\leq \|h_{k-1,b_i} - h_{b_i}^*\|^2 - \|P_{\gamma_{k-1}C_i}(h_{k-1,b_i}) - h_{b_i}^*\|^2 \\ 344 \quad &= \|h_{k-1,b_i} - h_{b_i}^*\|^2 - \|P_{\gamma_{k-1}C_i}(h_{k-1,b_i}) + \gamma_{k-1}\nabla F(x^*)_{b_i}\|^2 \\ 345 \quad &\leq \|h_{k-1,b_i} - h_{b_i}^*\|^2 - \gamma_{k-1}^2 \text{dist}(-\nabla F(x^*)_{b_i}, \text{rbd}(C_i))^2 \\ 346 \quad &\leq \|h_{k-1,b_i} - h_{b_i}^*\|^2 - \underline{\epsilon}^2 \text{dist}(-\nabla F(x^*)_{b_i}, \text{rbd}(C_i))^2. \end{aligned}$$

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

$$\begin{aligned} \|x_k - x^*\|^2 &= \sum_{i \in I_{x^*}^c} \|x_{k,b_i} - x_{b_i}^*\|^2 + \sum_{j \in I_{x^*}} \|x_{k,b_j} - x_{b_j}^*\|^2 \\ &\leq \|h_{k-1} - h^*\|^2 - \underline{\epsilon}^2 \sum_{i \in I_{x^*}^c} \text{dist}(-\nabla F(x^*)_{b_i}, \text{rbd}(C_i))^2 \\ &\leq \|x_{k-1} - x^*\|^2 - \underline{\epsilon}^2 \sum_{i \in I_{x^*}^c} \text{dist}(-\nabla F(x^*)_{b_i}, \text{rbd}(C_i))^2 \\ &\leq \|x_0 - x^*\|^2 - k\underline{\epsilon}^2 \sum_{i \in I_{x^*}^c} \text{dist}(-\nabla F(x^*)_{b_i}, \text{rbd}(C_i))^2, \end{aligned}$$

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 $I_{x^*}^c \cap I_{x_k} \neq \emptyset$ does not exceed the given bound. We then conclude that beyond this bound, there is no i such that $\mathcal{M}_{x_{k,b_i}} \neq \emptyset$ while $\mathcal{M}_{x_{b_i}^*} = \emptyset$. The proof is complete. \square

Note that, as intuitively expected, this bound increases as the non-degeneracy condition (ND) becomes more stringent. However, as it depends on x^* , it is only of theoretical interest. In the separable case, observe that $\sum_{i \in I_{x^*}^c} \text{dist}(-\nabla F(x^*)_{b_i}, \text{rbd}(C_i))^2 = \text{dist}(-\nabla F(x^*), \partial R(x^*))^2$ when σ_{C_i} is differentiable at $x_{b_i}^*$ for all $i \in I_{x^*}$. 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 $x^* \in \text{Argmin}(\Phi)$ (see typically the summability conditions in Theorem 3(i)-(ii)). Then, since $\xi_k \rightarrow 0$, 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 $R + \iota_{\mathcal{M}_{x^*}}$ locally agree nearby x^* . This property is clearly violated if approximate proximal mappings are involved. Here is a simple example.

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

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

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

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

whence the graph of $(\text{Id} + \partial^\varepsilon|\cdot|)^{-1}$, 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. $x_k \notin \mathcal{M}_{x^*}$, for any finite k .

4. Local linear convergence of FB-type methods. We are now in position to present the local linear convergence result for FB-type methods, and all the proofs in this section are collected in Section B. Throughout this section, x^* is a global minimizer of problem $(\mathcal{P}_{\text{opt}})$ to which the sequence $(x_k)_{k \in \mathbb{N}}$ provided by the FB-type method converges. \mathcal{M}_{x^*} is the partial smoothness manifold of R at x^* , and T_{x^*} the corresponding tangent space.

391 *Restricted injectivity.* In addition to **(H.2)**, in the rest of the paper, we also
 392 assume that F is locally C^2 around x^* , and its Hessian fulfills the following *restricted*
 393 *injectivity* condition,

$$394 \text{ (RI)} \quad \ker(\nabla^2 F(x^*)) \cap T_{x^*} = \{0\}.$$

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

$$397 (4.1) \quad \langle h, \nabla^2 F(x)h \rangle > \alpha \|h\|^2, \forall x \in \mathbb{B}_\epsilon(x^*) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : \|x - x^*\| \leq \epsilon\}.$$

398 It turns out that under conditions **(ND)** and **(RI)**, one can show that problem
 399 $(\mathcal{P}_{\text{opt}})$ admits a unique minimizer, and local quadratic growth of Φ if R is moreover
 400 partly smooth. Recall that a function Φ grows quadratically locally around x^* if
 401 $\exists c > 0$ such that $\Phi(x) \geq \Phi(x^*) + c\|x - x^*\|^2$, $\forall x$ near x^* .

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

- 405 (i) x^* is the unique minimizer of $(\mathcal{P}_{\text{opt}})$.
- 406 (ii) If moreover $R \in \text{PSF}_{x^*}(\mathcal{M}_{x^*})$, then Φ has at least a quadratic growth near x^* .

407 **4.1. Locally linearized iteration.** Define the following matrices which are all
 408 symmetric,

$$409 (4.2) \quad H \stackrel{\text{def}}{=} \gamma P_{T_{x^*}} \nabla^2 F(x^*) P_{T_{x^*}}, \quad G \stackrel{\text{def}}{=} \text{Id} - H, \quad U \stackrel{\text{def}}{=} \gamma \nabla_{\mathcal{M}_{x^*}}^2 \Phi(x^*) P_{T_{x^*}} - H,$$

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

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

- 414 (i) **(ND)** holds.
- 415 (ii) \mathcal{M}_{x^*} is an affine subspace.

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

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

419 **LEMMA 14.** *For the matrices in (4.2) and W ,*

- 420 (i) Under **(H.2)** and **(RI)**,
 - 421 (a) H is symmetric positive definite with eigenvalues in $]\gamma\alpha, \frac{\gamma}{\beta}]$.
 - 422 (b) For $\gamma \in [\underline{\epsilon}, 2\beta - \bar{\epsilon}]$, $\underline{\epsilon}$ and $\bar{\epsilon} > 0$, G has eigenvalues in $[-1 + \frac{\bar{\epsilon}}{\beta}, 1 - \alpha\epsilon[$
 423 $]-1, 1[$.
 - 424 (c) For $\gamma \in [\underline{\epsilon}, \beta]$, G is also symmetric positive semi-definite with eigenvalues
 425 in $[0, 1 - \alpha\epsilon[$ $[0, 1[$.
- 426 (ii) If both the assumptions of Lemma 13 and (i) hold, then WG has real eigen-
 427 values lying in $]-1, 1[$. If moreover $\gamma \in [\underline{\epsilon}, \beta]$, then WG has eigenvalues lying
 428 in $[0, 1[$.

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

$$430 (4.3) \quad M \stackrel{\text{def}}{=} \begin{bmatrix} (a-b)W + (1+b)WG & -(a-b)W - bWG \\ \text{Id} & 0 \end{bmatrix}.$$

431 Our interest in the vector d_k is inspired by the convergence rate analysis of the heavy
 432 ball method [50, Section 3.2]. We now show that once the active manifold is identified,
 433 FB-type iteration locally linearizes.

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

437 (4.4) $a_k \rightarrow a \in [0, 1], b_k \rightarrow b \in [0, 1], \gamma_k \rightarrow \gamma \in [\underline{\epsilon}, 2\beta - \bar{\epsilon}],$

438 *then for k large enough, we have*

439 (4.5) $d_{k+1} = Md_k + o(\|d_k\|).$

440 *The $o(\cdot)$ term disappears when R is locally polyhedral around x^* and (γ_k, a_k, b_k) are*
 441 *chosen constant.*

442 REMARK 16.

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

445 (ii) *For the FB method (i.e. $a_k = b_k \equiv 0$), (4.3) can be further simplified, and*
 446 *the corresponding linearized iteration can be given in terms of r_k directly,*

447 (4.6) $r_{k+1} = WGr_k + o(\|r_k\|).$

448 (iii) *Proposition 15 also covers the sequence convergent FISTA method [18, 9],*
 449 *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*
 450 *$a_k \rightarrow a = b = 1$.*

451 **4.2. Spectral properties of M .** Our aim now is to establish local linear con-
 452 vergence of FB-type schemes. For this, given the structure of the locally linearized
 453 iteration (4.5), it is sufficient to strictly upper-bound by 1 the spectral radius of M ,
 454 and conclude using standard arguments. This is what we are about to do.

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

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

463 *Locally polyhedral case.* When R is locally polyhedral around x^* , U vanishes and
 464 $W = \text{Id}$, and M in (4.3) simplifies.

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

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

470 (4.7) $\sigma^2 - ((a - b) + (1 + b)\eta)\sigma + (a - b) + b\eta = 0.$

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

472 (4.8) $(2(b - a) - 1)/(1 + 2b) < \underline{\eta}.$

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

476 *General partly smooth case.* When R is a general partly smooth function, then
 477 U is nontrivial, and the spectral analysis of (4.3) becomes a generalized eigenvalue
 478 problem which is much more complex. Therefore, we assume $b = a$. We have the
 479 following corollary of Proposition 17.

480 COROLLARY 19. Let $b = a$. If $\begin{pmatrix} r_1 \\ r_2 \end{pmatrix}$ be an eigenvector of M corresponding to an
 481 eigenvalue σ , then it must satisfy $r_1 = \sigma r_2$. Moreover r_2 is an eigenvector of G related
 482 to eigenvalue η , where η and σ satisfy the relation

$$483 \quad (4.9) \quad \sigma^2 - (1 + a)\eta\sigma + a\eta = 0,$$

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

$$485 \quad (4.10) \quad -1/(1 + 2a) < \underline{\eta}.$$

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

488 **4.3. Local linear convergence of FB-type methods.** We start with the case
 489 where R is locally polyhedral around x^* .

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

$$495 \quad \|x_k - x^*\| \leq C\rho^{k-K}\|x_K - x^*\|.$$

496 *Proof.* Combining Proposition 15, Proposition 17 and [50, Section 2.1.2, Theo-
 497 rem 1], leads to the claimed result. \square

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

500 Let's turn to the case R is a general partly smooth function, but $b = a \in [0, \bar{a}]$.

501 THEOREM 23. Suppose assumptions (H.1)-(H.3) hold, and the FB-type methods
 502 generate a sequence $x_k \rightarrow x^* \in \text{Argmin}(\Phi)$ such that $R \in \text{PSF}_{x^*}(\mathcal{M}_{x^*})$, F is C^2 near
 503 x^* , and conditions (ND), (RI) are satisfied. If moreover (4.4) holds with $b = a$, and
 504 (4.10) is satisfied, then $(x_k)_{k \in \mathbb{N}}$ converges locally linearly to x^* . More precisely, given
 505 any $\rho \in [\rho(M), 1[$, there exists $K > 0$ and a constant $C > 0$, such that for all $k \geq K$,
 506 there holds

$$507 \quad \|x_k - x^*\| \leq C\rho^{k-K}\|x_K - x^*\|.$$

508 *Proof.* This follows by combining Proposition 15, Corollary 19 and [50, Section
 509 2.1.2, Theorem 1]. \square

510 REMARK 24.

- 511 (i) The limit $b = a$ in (4.4) does not mean that we should set $b_k = a_k, \forall k \in \mathbb{N}$
 512 along the iterations.
- 513 (ii) In contrast to our previous work [37], which addresses the case of FB method,
 514 the rate estimates that we provide here are much sharper in general, and
 515 both estimates only coincide when R is locally polyhedral (see the numerical
 516 experiments for more details). The main reasons underlying this is that, here,
 517 our rate estimate relies on the locally linearized iteration in Proposition 15 and
 518 the spectral properties of M , which takes into account the geometry of the
 519 identified submanifold (its curvature for instance). This is not the case in our
 520 former work.
- 521 (iii) The obtained results can be readily extended to the variable metric FB split-
 522 ting method [21], where a rate under an appropriate metric can be obtained.
 523 However for the sake of brevity, we do not pursue this further.

- (iv) *In our proof of local linear convergence, convexity does play a crucial role. For instance, it was only needed to show that the matrix U is positive semi-definite. This suggests that our local linear convergence claims can be extended to the non-convex case, provided that the Riemannian Hessian of R is assumed positive semi-definite at x^* . In addition, to guarantee finite identification in the non-convex setting, we need global convergence of iFB to a critical point, which can be ensured if for instance Φ satisfies the (non-smooth) Kurdyka-Lojasiewicz inequality [15]. This will be left to a forthcoming paper.*

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

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

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

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

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

4.4. Discussion. We here summarize some main conclusions on the local linear convergence behaviour of FB-types methods. Recall that 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 $\bar{\eta} \geq \underline{\eta} \geq 0$ (see Lemma 14(ii)), and thus condition (4.10) is in force. Moreover $\bar{\eta}$ is also the local convergence rate of the FB method, and $\rho(M)$ depends solely on $\bar{\eta}$ 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 $\bar{\eta}$. One can make the the following observations:

- (i) When $a \in [0, \bar{\eta}]$, we have $\rho(M) \leq \bar{\eta}$. This entails that if iFB is used with such a choice of inertial parameter, it will converge locally lineally faster than FB. For $a \in [\bar{\eta}, 1]$, the situation reverses as $\rho(M) \geq \bar{\eta}$, and iFB becomes slower than FB.
- (ii) In particular, as $a = 1$ for FISTA, we have $\rho(M) = \sqrt{\bar{\eta}} > \bar{\eta}$. 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 $a = \frac{(1-\sqrt{1-\bar{\eta}})^2}{\bar{\eta}}$, 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

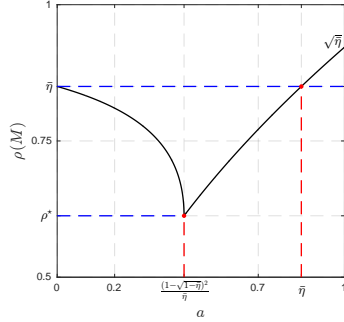


Fig. 1: Let $b = a$, and assume $\underline{\eta}, \bar{\eta}$ are known and also close enough such that the spectral radius $\rho(M)$ is only affected by $\bar{\eta}$, then $\rho(M)$ is a function of a .

572 slower, see Figure 2 and [54] for a FISTA applied to the LASSO problem. In fact,
 573 the iFB scheme shares this property as well when the inertial parameters are large.
 574 Such oscillatory behaviour is due to the fact that, for those inertial parameters, the
 575 eigenvalue σ_{\max} such that $|\sigma_{\max}| = \rho(M)$ is complex. It can then be shown that the
 576 oscillation period of $\|x_k - x^*\|$ is exactly $\frac{\pi}{\theta}$, where θ is the argument of σ_{\max} .

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

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

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

580 **4.5. Acceleration.** The finite time activity identification property (Theorem 8)
 581 implies that, the globally convex but non-smooth problem eventually becomes locally
 582 C^2 -smooth, but possibly non-convex, constrained on the activity manifold. This opens
 583 the door to acceleration, and even finite termination, exploiting the structure of the
 584 objective and that of the identified manifold. There are several ways to achieve this
 585 goal as we explain hereafter.

586 *Optimal first-order method.* In this case, the idea is to keep the scheme imple-
 587 mented in Algorithm 1, and to refine the parameters to minimize the local convergence
 588 rate established in Section 4. Indeed, as shown in Figure 1 and the discussion that
 589 follows, there is a proper choice of the inertial parameters a and b that minimizes
 590 $\rho(M)$. More precisely, choose $\gamma \in]0, \beta]$, then $\bar{\eta} = 1 - \alpha\gamma \geq \underline{\eta} \geq 1 - \gamma/\beta \geq 0$, and $\rho(M)$
 591 depends only on $\bar{\eta}$, a and b . Then with fixed γ (hence $\bar{\eta}$), $\rho(M)$ attains its minimal
 592 value for a and b satisfying

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

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

$$(4.12) \quad \rho^* = 1 - \sqrt{1 - \bar{\eta}} = 1 - \sqrt{\gamma\alpha},$$

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

$$601 \quad \underline{\rho}^* = \frac{1 - \sqrt{\alpha\beta}}{1 + \sqrt{\alpha\beta}} \quad \text{for} \quad \gamma = \frac{4\beta}{(1 + \sqrt{\alpha\beta})^2}, \quad a = \left(\frac{1 - \sqrt{\alpha\beta}}{1 + \sqrt{\alpha\beta}} \right)^2 \quad \text{and} \quad b = 0.$$

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

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

613 **5. Numerical experiments.** In this section, we illustrate the obtained results
 614 by some popular examples originating from linear inverse problems in signal processing
 615 and machine learning. We consider the linear model $y = Lx_{\text{ob}} + w$, where $y \in \mathbb{R}^m$,
 616 $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is some linear operator, and $w \in \mathbb{R}^m$ stands for noise. Solving such a
 617 linear inverse problem can be cast as the optimization problem

$$618 \quad (\mathcal{P}_\lambda) \quad \min_{x \in \mathbb{R}^n} \frac{1}{2} \|y - Lx\|^2 + \lambda R(x),$$

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

620 We use three functions R : the ℓ_1 -norm ($R(x) = \|x\|_1 \stackrel{\text{def}}{=} \sum_{i=1}^n |x_i|$), the $\ell_{1,2}$ -norm
 621 ($R(x) = \|x\|_{1,2} \stackrel{\text{def}}{=} \sum_{b \in \mathcal{B}} \|x_b\|$, for a uniform disjoint partition of $\{1, \dots, n\}$ in blocks
 622 \mathcal{B}), and the nuclear norm ($R(x) = \|x\|_* \stackrel{\text{def}}{=} \|\sigma(x)\|_1$, where $\sigma(x) \in (\mathbb{R}_+ \setminus \{0\})^r$ is the
 623 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
 624 are partly smooth relative to subspaces [57] (ℓ_1 is polyhedral), and the nuclear norm
 625 is partly smooth relative to the constant rank- r manifold [22].

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

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

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

634 *Parameter settings.* We choose $\gamma_k \equiv \beta$ for FISTA. For FB/iFB methods, two
 635 choices of γ_k are considered: $\gamma_k \equiv \beta$ and $\gamma_k \equiv 1.5\beta$. The inertial parameter of iFB
 636 and FISTA are:

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

640 The convergence profiles of $\|x_k - x^*\|$ are shown in Figure 2. As demonstrated by
 641 all the plots, identification and local linear convergence occurs after finite time. The
 642 solid lines (denoted as ‘‘P’’) represent the observed profiles, while dashed ones (denoted
 643 as ‘‘T’’) stand for the theoretically predicted ones. The positions of the *green* points
 644 (or the starting points of the dashed lines) stand for the iteration at which \mathcal{M}_{x^*} has
 645 been identified.

646 *Tightness of predicted rates.* For the ℓ_1 -norm, our predicted rates coincide exactly
 647 with the observed ones (same slopes for the dashed and solid lines). This is due to
 648 the fact that they are all polyhedral and F is quadratic. Note that for FISTA, which
 649 is non-monotone, the prediction coincides with the envelope of the oscillations. For

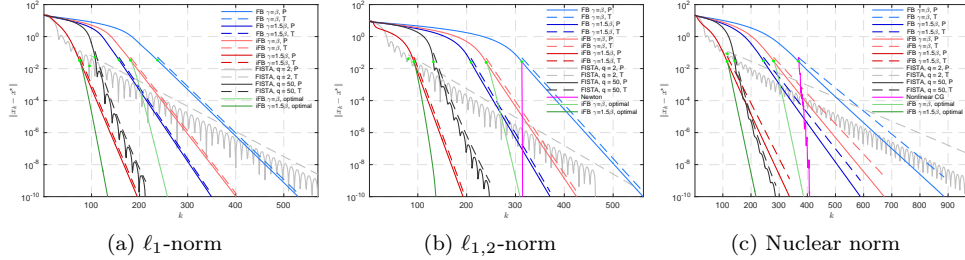


Fig. 2: Local linear convergence and comparison of the FB-type methods (FB, iFB and FISTA) in terms of $\|x_k - x^*\|$. See text for description.

650 the $\ell_{1,2}$ -norm, though it is not polyhedral, our predicted rates still are very tight, due
 651 to the fact that the Riemannian Hessian is taken into account. For the nuclear norm,
 652 whose active manifold is not anymore a subspace, our estimation becomes slightly
 653 less sharp compared to the other examples, though barely visible on the plots. Our
 654 predicted rates for FB are much sharper than in our previous work [37].

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

- 657 (i) Comparison of FB/iFB and FISTA under $\gamma_k \equiv \beta$:
- 658 • Globally, FISTA $q = 50$ is the fastest while $q = 2$ is the slowest. FB and
 659 iFB are in between them with iFB being faster.
 - 660 • For the finite identification, however, FISTA $q = 2$ in general shows the
 661 fastest identification, and FB is the slowest.
 - 662 • Locally, similar to the global convergence, FISTA $q = 50$ has the fastest
 663 rate and $q = 2$ is the slowest. Again, FB and iFB are between them with
 664 iFB being faster than FB.
- 665 (ii) $\gamma_k \equiv \beta$ vs $\gamma_k \equiv 1.5\beta$:
- 666 • For FB, larger γ_k leads to faster global convergence and activity identifi-
 667 cation. However this does not mean that the bigger the better locally. As
 668 we discussed in Section 4.5, the best choice to get the optimal local linear
 669 rate is $2\beta/(1 + \alpha\beta)$.
 - 670 • iFB is faster than FB under the same choice of γ_k . FISTA $q = 50$ is no
 671 longer the fastest one, while it is outperformed by iFB $\gamma_k \equiv 1.5\beta$ for the
 672 first 2 examples.

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

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

685 *Acceleration.* For the ℓ_1 -norm which is polyhedral, we applied the *first-order ac-*
 686 *celeration* described in (4.11) for $\gamma_k \equiv \beta$ and $\gamma_k \equiv 1.5\beta$ respectively (Figure 2(a)).
 687 In fact, acceleration is not even needed in this case and one can access a closed-form

688 solution of x^* once identification occurs. This can be easily achieved by projection the
 689 first-order minimality condition on $\mathcal{M}_{x^*} = x^* + T_{x^*}$, which boils down to solving an
 690 overdetermined linear system which has a unique solution under the restricted injec-
 691 tivity condition (RI). For the $\ell_{1,2}$ -norm, we applied the Riemannian Newton method
 692 which converges quadratically, leading to a dramatic acceleration as can be seen in
 693 Figure 2(b). For the nuclear norm, a non-linear conjugate gradient method is ap-
 694 plied, leading again to a much faster (super-linear) local convergence. To summarize,
 695 in practice, the *inertial+higher-order method* hybrid strategy is an ideal choice for
 696 solving (\mathcal{P}_{opt}).

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

700 Appendix A. Proofs of Section 2.

701 Throughout this section, \mathcal{H} denotes a real Hilbert space. Let $A : \mathcal{H} \rightrightarrows \mathcal{H}$ be a
 702 set-valued operator. The graph of A is the set $\text{gph } A = \{(x, y) \in \mathcal{H} \times \mathcal{H} | y \in A(x)\}$,
 703 and its zeros set is $\text{zer } A = \{x \in \mathcal{H} | 0 \in A(x)\}$. Recall that a set-valued operator
 704 $A : \mathcal{H} \rightrightarrows \mathcal{H}$ is monotone if

$$705 \quad (\text{A.1}) \quad (\forall (x, v) \in \text{gph } A), (\forall (y, u) \in \text{gph } A), \langle x - y, v - u \rangle \geq 0.$$

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

$$708 \quad (\text{A.2}) \quad (\forall x, y \in \mathcal{H}), \beta \|Bx - By\|^2 \leq \langle Bx - By, x - y \rangle.$$

709 *Proof (Theorem 3).* Define the following quantities

$$710 \quad (\text{A.3}) \quad \varphi_k = \frac{1}{2} \|x_k - x^*\|^2, \Delta_k = \frac{1}{2} \|x_k - x_{k-1}\|^2, E_{b,k} = \frac{1}{2} \|y_{b,k} - x_{k+1}\|^2.$$

711 Let $x^* \in \text{zer}(A + B)$, i.e. a solution (\mathcal{P}_{inc}), which exists thanks to (H.6). Recall from
 712 (1.4) and (2.1) that

$$713 \quad -B(x^*) \in A(x^*) \quad \text{and} \quad y_{a,k} - \gamma_k B(y_{b,k}) - \gamma_k \xi_k - x_{k+1} \in \gamma_k A^{\varepsilon_k}(x_{k+1}).$$

714 Thus, we get

$$715 \quad \langle y_{a,k} - x_{k+1} - \gamma_k (B(y_{b,k}) - B(x^*)) - \gamma_k \xi_k, x_{k+1} - x^* \rangle \geq -\gamma_k \varepsilon_k.$$

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

$$717 \quad (\text{A.4}) \quad \begin{aligned} \varphi_k - \varphi_{k+1} &= \frac{1}{2} \langle x_k - x^* + x_{k+1} - x^*, x_k - x_{k+1} \rangle \\ &= \Delta_{k+1} + \langle y_{a,k} - x_{k+1}, x_{k+1} - x^* \rangle - a_k \langle x_k - x_{k-1}, x_{k+1} - x^* \rangle \\ &\geq \Delta_{k+1} + \gamma_k \langle B(y_{b,k}) - B(x^*) + \xi_k, x_{k+1} - x^* \rangle \\ &\quad - a_k \langle x_k - x_{k-1}, x_{k+1} - x^* \rangle - \gamma_k \varepsilon_k. \end{aligned}$$

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

$$719 \quad (\text{A.5}) \quad \langle x_k - x_{k-1}, x_{k+1} - x^* \rangle = \langle x_k - x_{k-1}, x_{k+1} - x_k \rangle + (\Delta_k + \varphi_k - \varphi_{k-1}),$$

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

$$722 \quad (\text{A.6}) \quad \begin{aligned} \varphi_{k+1} - \varphi_k - a_k(\varphi_k - \varphi_{k-1}) &\leq -\Delta_{k+1} - \gamma_k \langle B(y_{b,k}) - B(x^*) + \xi_k, x_{k+1} - x^* \rangle \\ &\quad + a_k \langle x_k - x_{k-1}, x_{k+1} - x_k \rangle + a_k \Delta_k + \gamma_k \varepsilon_k. \end{aligned}$$

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

$$724 \quad (\text{A.7}) \quad \begin{aligned} &\langle B(y_{b,k}) - B(x^*), x_{k+1} - x^* \rangle \\ &\geq \beta \|B(y_{b,k}) - B(x^*)\|^2 + \langle B(y_{b,k}) - B(x^*), x_{k+1} - y_{b,k} \rangle = -\frac{1}{2\beta} E_{b,k}. \end{aligned}$$

725 Denote $\mu_k = 1 - \frac{\gamma_k}{2\beta} \in [\frac{\bar{\epsilon}}{2\beta}, 1 - \frac{\bar{\epsilon}}{2\beta}]$, $\nu_k = a_k - \frac{\gamma_k b_k}{2\beta}$ and $v_k = x_{k+1} - x_k - \frac{\nu_k}{\mu_k}(x_k - x_{k-1})$.
 726 Substituting (A.7) back into (A.6), and since $E_{b,k} = \Delta_{k+1} + b_k^2 \Delta_k + b_k(x_k - x_{k+1}, x_k -$
 727 $x_{k-1})$, we get

(A.8)

$$\begin{aligned} & \varphi_{k+1} - \varphi_k - a_k(\varphi_k - \varphi_{k-1}) \\ & \leq -\Delta_{k+1} + \frac{\gamma_k}{2\beta} E_{b,k} + a_k(x_k - x_{k-1}, x_{k+1} - x_k) + a_k \Delta_k + \gamma_k \varepsilon_k - \gamma_k \langle \xi_k, x_{k+1} - x^* \rangle \\ 728 & = -\frac{\mu_k}{2} \|v_k\|^2 + (a_k + \frac{\nu_k^2}{\mu_k} + \frac{\gamma_k b_k^2}{2\beta}) \Delta_k + \gamma_k (\varepsilon_k + \sqrt{2} \|\xi_k\| \sqrt{\varphi_{k+1}}) \\ & \leq -\frac{\mu_k}{2} \|v_k\|^2 + (\frac{2a_k}{\mu_k} + \frac{\gamma_k b_k}{2\beta}) \Delta_k + \gamma_k (\varepsilon_k + \sqrt{2} \|\xi_k\| \sqrt{\varphi_{k+1}}) \\ & \leq -\frac{\mu_k}{2} \|v_k\|^2 + (\frac{4\beta}{\bar{\epsilon}} a_k + (1 - \frac{\bar{\epsilon}}{2\beta}) b_k) \Delta_k + \bar{\gamma} (\varepsilon_k + \sqrt{2} \|\xi_k\| \sqrt{\varphi_{k+1}}). \end{aligned}$$

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

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

$$\begin{aligned} 731 \text{ (A.9)} & \leq \prod_{j=1}^k a_j \theta_1 + \sum_{j=1}^k (\prod_{l=j}^k a_{l-j}) (\delta_j + \bar{\gamma} \varepsilon_j + \sqrt{2\bar{\gamma}} \|\xi_j\| \sqrt{\varphi_{j+1}}) \\ & \leq \bar{a}^k \varphi_1 + \sum_{j=1}^k \bar{a}^{k-j} (\delta_j + \bar{\gamma} \varepsilon_j + \sqrt{2\bar{\gamma}} \|\xi_j\| \sqrt{\varphi_{j+1}}). \end{aligned}$$

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

$$\begin{aligned} 733 & \sum_{m=1}^k \theta_{m+1} = \varphi_{k+1} - \varphi_1 \\ 734 & \leq \frac{1}{1-\bar{a}} \varphi_1 + \sum_{m=1}^k \sum_{j=1}^m \bar{a}^{k-j} (\delta_j + \bar{\gamma} \varepsilon_j + \sqrt{2\bar{\gamma}} \|\xi_j\| \sqrt{\varphi_{j+1}}) \\ 735 & \leq \frac{1}{1-\bar{a}} \varphi_1 + \sum_{m=1}^k (\sum_{j=1}^{k-m} \bar{a}^j) (\delta_m + \bar{\gamma} \varepsilon_m + \sqrt{2\bar{\gamma}} \|\xi_m\| \sqrt{\varphi_{m+1}}) \\ 736 & \leq \frac{1}{1-\bar{a}} (\varphi_1 + \sum_{m=1}^k (\delta_m + \bar{\gamma} \varepsilon_m + \sqrt{2\bar{\gamma}} \|\xi_m\| \sqrt{\varphi_{m+1}})), \end{aligned}$$

737 which entails

(A.10)

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

741 where $c = \varphi_1 + \frac{1}{1-\bar{a}} (\varphi_1 + \sum_{m \in \mathbb{N}} \delta_m + \bar{\gamma} \sum_{m \in \mathbb{N}} \varepsilon_m) \geq 0$. By assumption
 742 on the sequences $(\varepsilon_m)_{m \in \mathbb{N}}$ and $(\delta_m)_{m \in \mathbb{N}}$, c is bounded. Using the fact that
 743 $(\|\xi_m\|)_{m \in \mathbb{N}}$ is summable, it can be easily shown, e.g. [6, Lemma A.9], that
 744 since $(\varphi_k)_{k \in \mathbb{N}}$ satisfies (A.10), it also obeys $\varphi_k \leq \sqrt{c} + \sum_{j \in \mathbb{N}} \|\xi_j\| < +\infty$.

745 Denote $t = \sqrt{c} + \sum_{j \in \mathbb{N}} \|\xi_j\|$. Then, (A.9) becomes

$$\begin{aligned} 746 & \theta_{k+1} \leq -\frac{\mu_k}{2} \|v_k\|^2 + \bar{a} \theta_k + \delta_k + \bar{\gamma} \varepsilon_k + \sqrt{2t\bar{\gamma}} \|\xi_k\| \\ 747 \text{ (A.11)} & \leq -\frac{\mu_k}{2} \|v_k\|^2 + a_k [\theta_k]_+ + \delta_k + \bar{\gamma} \varepsilon_k + \sqrt{2t\bar{\gamma}} \|\xi_k\| \end{aligned}$$

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

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

751 where $e_k = \delta_k + \bar{\gamma} \varepsilon_k + \sqrt{2\bar{\gamma}} \sqrt{t} \|\xi_k\|$ is a summable sequence by assumption.
 752 Therefore, using that $\bar{a} < 1$ and applying [20, Lemma 3.1(iv)], it follows that
 753 $[\theta_k]_+$ is summable. In turn,

$$754 \varphi_{k+1} - \sum_{j=1}^{k+1} [\theta_j]_+ \leq \varphi_{k+1} - \theta_{k+1} - \sum_{j=1}^k [\theta_j]_+ = \varphi_k - \sum_{j=1}^k [\theta_j]_+.$$

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

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

$$759 \quad \begin{aligned} \varphi_{k+1} &\leq \varphi_k + \delta_k + \bar{\gamma}\varepsilon_k + \sqrt{2\bar{\gamma}}\|\xi_k\|\sqrt{\varphi_k} \\ &\leq \varphi_1 + \sum_{j \in \mathbb{N}} \delta_j + \bar{\gamma} \sum_{j \in \mathbb{N}} \varepsilon_j + \sqrt{2\bar{\gamma}} \sum_{j=1}^k \|\xi_j\|\sqrt{\varphi_{j+1}}. \end{aligned}$$

760 Again, by virtue of [6, Lemma A.9] and the summability of the sequences
761 $(\delta_j)_{k \in \mathbb{N}}$, $(\varepsilon_j)_{k \in \mathbb{N}}$ and $(\|\xi_j\|)_{k \in \mathbb{N}}$, we have $\varphi_k \leq t = \sqrt{\varphi_1 + \sum_{j \in \mathbb{N}} (\delta_j + \bar{\gamma}\varepsilon_j + \|\xi_j\|)} < +\infty$. Consequently, we have

$$763 \quad \varphi_{k+1} \leq \varphi_k + \delta_k + \bar{\gamma}\varepsilon_k + \sqrt{2t\bar{\gamma}}\|\xi_k\|.$$

764 We then conclude that the sequence $(x_k)_{k \in \mathbb{N}}$ is quasi-Fejér monotone (of type
765 III) relative to $\text{zer}(A + B)$ [20, Definition 1.1(3)], and thus φ_k is convergent
766 [20, Proposition 3.6].

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

769 By assumption (2.2), $a_k(x_k - x_{k-1}) \rightarrow 0$ and $b_k(x_k - x_{k-1}) \rightarrow 0$, and thus

$$770 \quad (\text{A.12}) \quad \frac{\nu_k}{\mu_k}(x_k - x_{k-1}) \rightarrow 0,$$

771 since $\mu_k \geq \frac{\bar{\varepsilon}}{2\beta} > 0$. Moreover, from (A.11), we obtain

$$772 \quad \sum_{k \in \mathbb{N}} \|v_k\|^2 \leq \frac{4\beta}{\varepsilon} (\bar{a}\varphi_0 + \sum_{k \in \mathbb{N}} (\bar{a}[\theta_k]_+ + e_k)) < +\infty.$$

773 Consequently, $v_k \rightarrow 0$. Combining this with (A.12), we get that $x_{k+1} - x_k \rightarrow 0$. In
774 turn, $y_{a,k} - x_{k+1} \rightarrow 0$ and $y_{b,k} - x_{k+1} \rightarrow 0$. Let \bar{x} be a weak cluster point of $(x_k)_{k \in \mathbb{N}}$,

775 and let us fix a subsequence, say $x_{k_j} \rightharpoonup \bar{x}$. Denote $u_{k_j} \stackrel{\text{def}}{=} \frac{y_{a,k_j} - x_{k_j+1}}{\gamma_{k_j}} - B(y_{b,k_j}) - \xi_{k_j}$.

776 Since B is cocoercive and $y_{b,k_j} \rightharpoonup \bar{x}$, we have $B(y_{b,k_j}) \rightarrow B(\bar{x})$. In turn, $u_{k_j} \rightarrow$
777 $-B(\bar{x})$ since $\gamma_k \geq \underline{\varepsilon} > 0$ and $\xi_k \rightarrow 0$. Since $(x_{k_j+1}, u_{k_j}) \in \text{gph } A^{\varepsilon_{k_j}}$, and the graph
778 of the enlargement of A is weakly-strongly sequentially closed in $\mathbb{R}_+ \times \mathcal{H} \times \mathcal{H}$ [53,
779 Proposition 3.4(b)], we get that $-B(\bar{x}) \in A(\bar{x})$, *i.e.* \bar{x} is a solution of $(\mathcal{P}_{\text{inc}})$. Opial's
780 theorem [47] concludes the proof. \square

781 *Proof (Theorem 4).* In view of the imposed assumptions, we deduce from Theo-
782 rem 3 that $(x_k)_{k \in \mathbb{N}}$ is bounded, and thus $c = \sup_{k \in \mathbb{N}} \|x_k - x^*\| < +\infty$. From (A.8),
783 we apply Young's inequality to get

$$784 \quad \begin{aligned} &\varphi_{k+1} - \varphi_k - a_k(\varphi_k - \varphi_{k-1}) \\ &\leq \left(\frac{\gamma_k}{2\beta} - 1\right)\Delta_{k+1} + |a_k - \frac{\gamma_k b_k}{2\beta}|(\Delta_{k+1} + \Delta_k) + \left(\frac{\gamma_k}{2\beta} b_k^2 + a_k\right)\Delta_k + \gamma_k(\varepsilon_k + c\|\xi_k\|) \\ &= s_k \Delta_{k+1} + t_k \Delta_k + \bar{\gamma}(\varepsilon_k + c\|\xi_k\|), \end{aligned}$$

785 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
786 and γ_k are non-decreasing so that s_k , t_k are also non-decreasing. Denote $\phi_k = \varphi_k -$
787 $a_k \varphi_{k-1} + t_k \Delta_k$ and $\delta_k = \bar{\gamma}(\varepsilon_k + c\|\xi_k\|)$,

$$788 \quad (\text{A.13}) \quad \begin{aligned} \phi_{k+1} - \phi_k &\leq (\varphi_{k+1} - \varphi_k) - a_k(\varphi_k - \varphi_{k-1}) + t_{k+1}\Delta_{k+1} - t_k\Delta_k \\ &\leq s_k \Delta_{k+1} + t_k \Delta_k + t_{k+1}\Delta_{k+1} - t_k \Delta_k + \delta_k \\ &= (s_k + t_{k+1})\Delta_{k+1} + \delta_k. \end{aligned}$$

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

$$791 \quad (\text{A.14}) \quad \begin{aligned} \phi_{k+1} - \phi_k &\leq (s_{k+1} + t_{k+1})\Delta_{k+1} \\ &= ((3a_{k+1} - 1) + \frac{\gamma_{k+1}}{2\beta}(1 - b_{k+1})^2)\Delta_{k+1} + \delta_k \leq -\tau\Delta_{k+1} + \delta_k. \end{aligned}$$

792 (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
793 (A.13) we have,

$$794 \begin{aligned} \phi_{k+1} - \phi_k &\leq (s_{k+1} + t_{k+1})\Delta_{k+1} + \delta_k \\ &\leq \left(\frac{\gamma_{k+1}}{2\beta} - 1 + 2|a_{k+1} - \frac{\gamma_{k+1}}{2\beta}b_{k+1}| + \frac{\gamma_{k+1}}{2\beta}b_{k+1}^2 + a_{k+1}\right)\Delta_{k+1} + \delta_k. \end{aligned}$$

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

797 (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),
(A.15)

$$798 \phi_{k+1} - \phi_k \leq \left((3a_{k+1} - 1) + \frac{\gamma_{k+1}}{2\beta}(1 - b_{k+1})^2\right)\Delta_{k+1} + \delta_k \leq -\tau\Delta_{k+1} + \delta_k.$$

799 (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),
(A.16)

$$800 \phi_{k+1} - \phi_k \leq \left(-(1 + a_{k+1}) + \frac{\gamma_{k+1}}{2\beta}(1 + b_{k+1})^2\right)\Delta_{k+1} + \delta_k \leq -\tau\Delta_{k+1} + \delta_k.$$

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

$$802 \sum_{j=1}^k \Delta_{j+1} \leq \frac{1}{\tau}(\phi_1 - \phi_{k+1}) + \sum_{j=1}^k \delta_j \leq \frac{1}{\tau}(\phi_1 + \bar{a}\varphi_k) + \sum_{j=1}^k \delta_j < +\infty.$$

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

$$805 \sum_{j=1}^k \Delta_{j+1} \leq \frac{1}{\tau}(\phi_1 - \phi_{k+1}) \leq \frac{1}{\tau}(\phi_1 + \bar{a}\varphi_k) \leq \frac{1}{\tau}(\bar{a}^k\varphi_1 + \frac{\phi_1}{1-\bar{a}}) < +\infty.$$

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

808 Appendix B. Proofs of Section 4.

809 **B.1. Riemannian Geometry.** Let \mathcal{M} be a C^2 -smooth embedded submanifold
810 of \mathbb{R}^n around a point x . With some abuse of terminology, we shall state C^2 -manifold
811 instead of C^2 -smooth embedded submanifold of \mathbb{R}^n . The natural embedding of a
812 submanifold \mathcal{M} into \mathbb{R}^n permits to define a Riemannian structure and to introduce
813 geodesics on \mathcal{M} , and we simply say \mathcal{M} is a Riemannian manifold. Denote respectively
814 $\mathcal{T}_{\mathcal{M}}(x)$ and $\mathcal{N}_{\mathcal{M}}(x)$ the tangent and normal space of \mathcal{M} at point near x in \mathcal{M} .

815 *Exponential map.* Geodesics generalize the concept of straight lines in \mathbb{R}^n , pre-
816 serving the zero acceleration characteristic, to manifolds. Roughly speaking, a geodesic
817 is locally the shortest path between two points on \mathcal{M} . We denote by $\mathfrak{g}(t; x, h)$
818 the value at $t \in \mathbb{R}$ of the geodesic starting at $\mathfrak{g}(0; x, h) = x \in \mathcal{M}$ with velocity
819 $\dot{\mathfrak{g}}(t; x, h) = \frac{d\mathfrak{g}}{dt}(t; x, h) = h \in \mathcal{T}_{\mathcal{M}}(x)$ (which is uniquely defined). For every $h \in \mathcal{T}_{\mathcal{M}}(x)$,
820 there exists an interval I around 0 and a unique geodesic $\mathfrak{g}(t; x, h) : I \rightarrow \mathcal{M}$ such that
821 $\mathfrak{g}(0; x, h) = x$ and $\dot{\mathfrak{g}}(0; x, h) = h$. The mapping

$$822 \text{Exp}_x : \mathcal{T}_{\mathcal{M}}(x) \rightarrow \mathcal{M}, \quad h \mapsto \text{Exp}_x(h) = \mathfrak{g}(1; x, h),$$

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

$$825 \text{Exp}_x(h) = z = \mathfrak{g}(1; x, h).$$

826 *Parallel translation.* Given two points $x, z \in \mathcal{M}$, let $\mathcal{T}_{\mathcal{M}}(x), \mathcal{T}_{\mathcal{M}}(z)$ be their cor-
827 responding tangent spaces. Define

$$828 \tau : \mathcal{T}_{\mathcal{M}}(x) \rightarrow \mathcal{T}_{\mathcal{M}}(z),$$

829 the parallel translation along the unique geodesic joining x to z , which is isomorphism
830 and isometry w.r.t. the Riemannian metric.

831 *Riemannian gradient and Hessian.* For a vector $v \in \mathcal{N}_{\mathcal{M}}(x)$, the Weingarten map
 832 of \mathcal{M} at x is the operator $\mathfrak{W}_x(\cdot, v) : \mathcal{T}_{\mathcal{M}}(x) \rightarrow \mathcal{T}_{\mathcal{M}}(x)$ defined by

$$833 \quad \mathfrak{W}_x(\cdot, v) = -\mathbb{P}_{\mathcal{T}_{\mathcal{M}}(x)} dV[h],$$

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

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

$$840 \quad \langle \nabla_{\mathcal{M}}G(z), h \rangle = \left. \frac{d}{dt} G(\mathbb{P}_{\mathcal{M}}(z + th)) \right|_{t=0}, \quad \forall h \in \mathcal{T}_{\mathcal{M}}(z),$$

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

$$843 \quad (\text{B.1}) \quad \langle \nabla_{\mathcal{M}}^2 G(z)h, h \rangle = \left. \frac{d^2}{dt^2} G(\mathbb{P}_{\mathcal{M}}(z + th)) \right|_{t=0}, \quad \forall h \in \mathcal{T}_{\mathcal{M}}(z).$$

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

$$850 \quad (\text{B.2}) \quad \nabla_{\mathcal{M}}G(z) = \mathbb{P}_{\mathcal{T}_{\mathcal{M}}(z)} \nabla \tilde{G}(z),$$

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

$$852 \quad (\text{B.3}) \quad \begin{aligned} \nabla_{\mathcal{M}}^2 G(z)h &= \mathbb{P}_{\mathcal{T}_{\mathcal{M}}(z)} d(\nabla_{\mathcal{M}}G)(z)[h] = \mathbb{P}_{\mathcal{T}_{\mathcal{M}}(z)} d(z \mapsto \mathbb{P}_{\mathcal{T}_{\mathcal{M}}(z)} \nabla_{\mathcal{M}} \tilde{G})[h] \\ &= \mathbb{P}_{\mathcal{T}_{\mathcal{M}}(z)} \nabla^2 \tilde{G}(z)h + \mathfrak{W}_z(h, \mathbb{P}_{\mathcal{N}_{\mathcal{M}}(z)} \nabla \tilde{G}(z)), \end{aligned}$$

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

$$856 \quad \nabla_{\mathcal{M}}^2 G(z) = \mathbb{P}_{\mathcal{T}_{\mathcal{M}}(z)} \nabla^2 \tilde{G}(z) \mathbb{P}_{\mathcal{T}_{\mathcal{M}}(z)}.$$

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

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

859 **LEMMA 26.** *Let $x \in \mathcal{M}$, and x_k a sequence converging to x in \mathcal{M} . Denote $\tau_k :$
 860 $\mathcal{T}_{\mathcal{M}}(x) \rightarrow \mathcal{T}_{\mathcal{M}}(x_k)$ be the parallel translation along the unique geodesic joining x to
 861 x_k . Then, for any bounded vector $u \in \mathbb{R}^n$, we have*

$$862 \quad (\tau_k^{-1} \mathbb{P}_{\mathcal{T}_{\mathcal{M}}(x_k)} - \mathbb{P}_{\mathcal{T}_{\mathcal{M}}(x)})u = o(\|u\|).$$

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

$$864 \quad \tau_k^{-1} = \mathbb{P}_{\mathcal{T}_{\mathcal{M}}(x)} + o(\|x_k - x\|).$$

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

$$866 \quad \begin{aligned} \lim_{k \rightarrow \infty} \frac{\|(\tau_k^{-1} \mathbb{P}_{\mathcal{T}_{\mathcal{M}}(x_k)} - \mathbb{P}_{\mathcal{T}_{\mathcal{M}}(x)})u\|}{\|u\|} &\leq \lim_{k \rightarrow \infty} \frac{\|\mathbb{P}_{\mathcal{T}_{\mathcal{M}}(x)}(\mathbb{P}_{\mathcal{T}_{\mathcal{M}}(x_k)} - \mathbb{P}_{\mathcal{T}_{\mathcal{M}}(x)})\| \|u\|}{\|u\|} + o(\|x_k - x\|) \\ &\leq \lim_{k \rightarrow \infty} \|\mathbb{P}_{\mathcal{T}_{\mathcal{M}}(x_k)} - \mathbb{P}_{\mathcal{T}_{\mathcal{M}}(x)}\| + o(\|x_k - x\|) = 0. \quad \square \end{aligned}$$

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

$$870 \quad \tau^{-1} \nabla_{\mathcal{M}} \Phi(z) = \nabla_{\mathcal{M}} \Phi(x) + \nabla_{\mathcal{M}}^2 \Phi(x) \mathbb{P}_{\mathcal{T}_{\mathcal{M}}(x)}(z - x) + o(\|z - x\|).$$

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

$$873 \quad (\text{B.4}) \quad \tau^{-1}\nabla_{\mathcal{M}}\Phi(z) = \nabla_{\mathcal{M}}\Phi(x) + \nabla_{\mathcal{M}}^2\Phi(x)h + o(\|h\|).$$

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

$$875 \quad \text{P}_{\mathcal{T}_{\mathcal{M}}(x)}(z) = \text{P}_{\mathcal{T}_{\mathcal{M}}(x)}(\text{Exp}_x(h)) = \text{P}_{\mathcal{T}_{\mathcal{M}}(x)}(x) + h + o(\|h\|^2).$$

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

877 B.2. Proofs.

878 *Proof (Proposition 12).*

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

$$881 \quad \Phi(x^* + \delta) - \Phi(x^*) = \frac{1}{2}\langle \delta, \nabla^2 F(x^* + t\delta)\delta \rangle + R(x^* + \delta) - R(x^*) + \langle \nabla F(x^*), \delta \rangle.$$

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

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

$$884 \quad \Phi(x^* + \delta) - \Phi(x^*) \geq \frac{1}{2}\langle \delta, \nabla^2 F(x_t)\delta \rangle > 0.$$

885 (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
886 x^* . Moreover $\partial R(x^*) \neq \emptyset$ ($-\nabla F(x^*)$ is in it), and thus the directional
887 derivative $R'(x^*, \cdot)$ is proper and closed, and it is the support of $\partial R(x^*)$
888 [51, Theorem 8.30]. It then follows from the separation theorem [30,
889 Theorem V.2.2.3] that

$$890 \quad \begin{aligned} & -\nabla F(x^*) \in \text{ri}(\partial R(x^*)) \\ & \Leftrightarrow R'(x^*, \delta) > -\langle \nabla F(x^*), \delta \rangle, \forall \delta \text{ s.t. } R'(x^*, \delta) + R'(x^*, -\delta) > 0. \end{aligned}$$

891 Since (RI) holds and $\nabla^2 F(x)$ depends continuously on $x \in \mathbb{B}_{\epsilon}(x^*)$, (4.1)
892 holds for any such x , and in particular at x_t . Combining with the fact
893 that $\ker(R'(x^*, \cdot)) = T_{x^*}$ [56, Proposition 3(iii) and Lemma 10], we get
894 $-\nabla F(x^*) \in \text{ri}(\partial R(x^*)) \Leftrightarrow R'(x^*, \delta) > -\langle \nabla F(x^*), \delta \rangle, \forall \delta \notin T_{x^*}$

$$\Rightarrow R'(x^*, \delta) > -\langle \nabla F(x^*), \delta \rangle, \forall \delta \in \ker(\nabla^2 F(x_t)) \setminus \{0\}.$$

895 Thus, classical properties of the directional derivative of a convex func-
896 tion yield

$$897 \quad \begin{aligned} & \Phi(x^* + \delta) - \Phi(x^*) \\ & = R(x^* + \delta) - R(x^*) + \langle \nabla F(x^*), \delta \rangle \geq R'(x^*, \delta) + \langle \nabla F(x^*), \delta \rangle > 0. \end{aligned}$$

898 (ii) Let Ψ as defined in the proof of Lemma 13. If $R \in \text{PSF}_{x^*}(\mathcal{M}_{x^*})$, the Rie-
899 mannian Hessian of Φ reads

$$900 \quad \nabla_{\mathcal{M}_{x^*}}^2 \Phi(x^*) = \text{P}_{T_{x^*}} \nabla F(x^*) \text{P}_{T_{x^*}} + \nabla_{\mathcal{M}_{x^*}}^2 \Psi(x^*).$$

901 In view of Lemma 13(i), $\nabla_{\mathcal{M}_{x^*}}^2 \Psi(x^*)$ is positive semi-definite on T_{x^*} . On the
902 other hand, hypothesis (RI) entails positive definiteness of $\text{P}_{T_{x^*}} \nabla F(x^*) \text{P}_{T_{x^*}}$.
903 Altogether, this shows that $\nabla_{\mathcal{M}_{x^*}}^2 \Phi(x^*)$ is positive definite on $T_{x^*} \setminus \{0\}$. Local
904 quadratic growth of Φ near x^* then follows by combining [35, Definition 5.4],
905 [40, Theorem 3.4] and [28, Theorem 6.2]. \square

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

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

$$911 \quad \begin{aligned} & \gamma \nabla_{\mathcal{M}_{x^*}}^2 \Psi(x^*)h = \gamma \text{P}_{T_{x^*}} \nabla^2 \tilde{R}(x^*)h + \gamma \mathfrak{W}_{x^*}(h, \text{P}_{T_{x^*}^{\perp}} \nabla \tilde{\Phi}(x^*)) \\ & = \gamma \nabla_{\mathcal{M}_{x^*}}^2 \Phi(x^*) \text{P}_{T_{x^*}} h - Hh = Uh, \end{aligned}$$

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

$$913 \quad \partial^2 R(x^* | -\nabla F(x^*))h = \begin{cases} \nabla_{\mathcal{M}_{x^*}}^2 \Psi(x^*)h + T_{x^*}^\perp, & h \in T_{x^*}, \\ \emptyset, & h \notin T_{x^*}, \end{cases}$$

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

$$919 \quad 0 \leq \gamma \langle \partial^2 R(x^* | -\nabla F(x^*))h, h \rangle = \gamma \langle \nabla_{\mathcal{M}_{x^*}}^2 \Psi(x^*)h, h \rangle = \langle Uh, h \rangle.$$

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

$$923 \quad \begin{aligned} 0 &\leq t^{-2} \langle v - u, x_t - x^* \rangle = t^{-1} \langle P_{T_{x_t}} v - P_{T_{x^*}} u, h \rangle \\ &\quad (\text{by Fact 7}) = \langle t^{-1} (\nabla_{\mathcal{M}_{x^*}} R(x_t) - \nabla_{\mathcal{M}_{x^*}} R(x^*)), h \rangle \\ &\quad (\text{by (B.2)}) = \langle t^{-1} P_{T_{x^*}} (\nabla \tilde{R}(x^* + tP_{T_{x^*}} h) - \nabla \tilde{R}(x^*)), h \rangle. \end{aligned}$$

924 Since \tilde{R} is C^2 , passing to the limit as $t \rightarrow 0$ leads to the desired result. \square

925 *Proof (Lemma 14).*

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

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

$$930 \quad \|W^{1/2}GW^{1/2}\| \leq \|W^{1/2}\| \|G\| \|W^{1/2}\| < 1,$$

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

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

$$\begin{aligned} 935 \quad & \text{(B.5)} \quad H_k = \gamma_k P_{T_{x^*}} \nabla^2 F(x^*) P_{T_{x^*}}, \quad G_k = \text{Id} - H_k, \quad U_k = \gamma_k \nabla_{\mathcal{M}_{x^*}}^2 \Phi(x^*) P_{T_{x^*}} - H_k, \\ & M_{k,1} = [(1+b)W(G_k - G), -bW(G_k - G)], \\ & M_{k,2} = [((a_k - b_k) - (a - b))W + (b_k - b)WG_k, -((a_k - b_k) - (a - b))W - (b_k - b)WG_k]. \end{aligned}$$

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

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

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

$$941 \quad \text{(B.6)} \quad \begin{aligned} \|y_{a,k} - x^*\| &= O(\|d_k\|), \quad \|y_{b,k} - x^*\| = O(\|d_k\|), \quad \|r_{k+1}\| = O(\|d_k\|), \\ & (\tau_{k+1}^{-1} P_{T_{x_{k+1}}} - P_{T_{x^*}})(\nabla F(y_{b,k}) - \nabla F(x_{k+1})) = o(\|d_k\|). \end{aligned}$$

$$943 \quad \text{(B.7)} \quad \|W(U_k - U)r_{k+1}\| = o(\|d_k\|), \quad \|M_{k,1}d_k\| = o(\|d_k\|) \quad \text{and} \quad \|M_{k,2}d_k\| = o(\|d_k\|).$$

944 *Proof.* We have

$$945 \quad \text{(B.8)} \quad \begin{aligned} \|y_{a,k} - x^*\| &= \|(1 + a_k)r_k - a_k r_{k-1}\| \leq (1 + a_k)\|r_k\| + a_k\|r_{k-1}\| \\ &\leq (1 + a_k)(\|r_k\| + \|r_{k-1}\|) \leq \sqrt{2}(1 + a_k)\|d_k\|, \end{aligned}$$

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

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

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

$$\begin{aligned}
951 \quad (\tau_{k+1}^{-1}P_{T_{x_{k+1}}} - P_{T_{x^*}})(\nabla F(y_{b,k}) - \nabla F(x_{k+1})) &= o(\|\nabla F(y_{b,k}) - \nabla F(x_{k+1})\|) \\
&= o(\|y_{b,k} - x^*\|) + o(\|r_{k+1}\|) = o(\|d_k\|).
\end{aligned}$$

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

$$\begin{aligned}
953 \quad \lim_{k \rightarrow \infty} \|W(U_k - U)r_{k+1}\|/\|r_{k+1}\| &= \lim_{k \rightarrow \infty} \|W(\gamma_k - \gamma)\nabla_{\mathcal{M}_{x^*}}^2 \Psi(x^*)P_{T_{x^*}}r_{k+1}\|/\|r_{k+1}\| \\
&\leq \lim_{k \rightarrow \infty} |\gamma_k - \gamma| \|W\| \|\nabla_{\mathcal{M}_{x^*}}^2 \Psi(x^*)P_{T_{x^*}}\| = 0,
\end{aligned}$$

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

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

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

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

958 where W_k, G_k are bounded. \square

959 *Proof (Proposition 15).* **(1.3)** and the first-order optimality condition for problem
960 $(\mathcal{P}_{\text{opt}})$ are respectively equivalent to

$$961 \quad y_{a,k} - x_{k+1} - \gamma_k(\nabla F(y_{b,k}) - \nabla F(x_{k+1})) \in \gamma_k \partial \Phi(x_{k+1}) \quad \text{and} \quad 0 \in \gamma_k \partial \Phi(x^*).$$

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

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

964 Adding both identities, and subtracting $\tau_{k+1}^{-1}P_{T_{x_{k+1}}}x^*$ on both sides, we arrive at

$$\begin{aligned}
965 \quad (\text{B.10}) \quad \tau_{k+1}^{-1}P_{T_{x_{k+1}}}r_{k+1} + \gamma_k(\tau_{k+1}^{-1}\nabla_{\mathcal{M}_{x^*}}\Phi(x_{k+1}) - \nabla_{\mathcal{M}_{x^*}}\Phi(x^*)) \\
= \tau_{k+1}^{-1}P_{T_{x_{k+1}}}(y_{a,k} - x^*) - \gamma_k\tau_{k+1}^{-1}P_{T_{x_{k+1}}}(\nabla F(y_{b,k}) - \nabla F(x_{k+1})).
\end{aligned}$$

966 In view of Lemma 26, we get

$$967 \quad \tau_{k+1}^{-1}P_{T_{x_{k+1}}}r_{k+1} = P_{T_{x^*}}r_{k+1} + (\tau_{k+1}^{-1}P_{T_{x_{k+1}}} - P_{T_{x^*}})r_{k+1} = P_{T_{x^*}}r_{k+1} + o(\|r_{k+1}\|).$$

968 Using [37, Lemma 5.1], we have

$$\begin{aligned}
969 \quad (\text{B.11}) \quad r_{k+1} &= P_{T_{x^*}}r_{k+1} + o(\|r_{k+1}\|) \\
&\Rightarrow \tau_{k+1}^{-1}P_{T_{x_{k+1}}}r_{k+1} = r_{k+1} + o(\|r_{k+1}\|) = r_{k+1} + o(\|d_k\|),
\end{aligned}$$

970 where we also used **(B.6)**. Similarly

$$\begin{aligned}
971 \quad (\text{B.12}) \quad \tau_{k+1}^{-1}P_{T_{x_{k+1}}}(y_{a,k} - x^*) &= P_{T_{x^*}}(y_{a,k} - x^*) + (\tau_{k+1}^{-1}P_{T_{x_{k+1}}} - P_{T_{x^*}})(y_{a,k} - x^*) \\
&= P_{T_{x^*}}(y_{a,k} - x^*) + o(\|y_{a,k} - x^*\|) \\
&= P_{T_{x^*}}(y_{a,k} - x^*) + o(\|d_k\|) \\
&= (1 + a_k)P_{T_{x^*}}r_k - a_kP_{T_{x^*}}r_{k-1} + o(\|d_k\|) \\
&= (1 + a_k)r_k - a_kr_{k-1} + o(\|r_k\|) + o(\|r_{k-1}\|) + o(\|d_k\|) \\
&= (y_{a,k} - x^*) + o(\|d_k\|).
\end{aligned}$$

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

$$973 \quad (B.13) \quad \begin{aligned} \tau^{-1} \nabla_{\mathcal{M}_{x^*}} \Phi(x_{k+1}) - \nabla_{\mathcal{M}_{x^*}} \Phi(x^*) &= \nabla_{\mathcal{M}_{x^*}}^2 \Phi(x^*) P_{T_{x^*}} r_{k+1} + o(\|r_{k+1}\|) \\ &= \nabla_{\mathcal{M}_{x^*}}^2 \Phi(x^*) P_{T_{x^*}} r_{k+1} + o(\|d_k\|). \end{aligned}$$

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

$$975 \quad (B.14) \quad \begin{aligned} &(\text{Id} + \gamma_k \nabla_{\mathcal{M}_{x^*}}^2 \Phi(x^*) P_{T_{x^*}}) r_{k+1} \\ &= (y_{a,k} - x^*) - \gamma_k \tau_{k+1}^{-1} P_{T_{x_{k+1}}} (\nabla F(y_{b,k}) - \nabla F(x_{k+1})) + o(\|d_k\|). \end{aligned}$$

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

$$977 \quad (B.15) \quad \begin{aligned} &\tau_{k+1}^{-1} P_{T_{x_{k+1}}} (\nabla F(y_{b,k}) - \nabla F(x_{k+1})) \\ &= P_{T_{x^*}} (\nabla F(y_{b,k}) - \nabla F(x_{k+1})) + o(\|d_k\|) \\ &= P_{T_{x^*}} (\nabla F(y_{b,k}) - \nabla F(x^*)) - P_{T_{x^*}} (\nabla F(x_{k+1}) - \nabla F(x^*)) + o(\|d_k\|) \\ &= P_{T_{x^*}} \nabla^2 F(x^*) P_{T_{x^*}} (y_{b,k} - x^*) - P_{T_{x^*}} \nabla^2 F(x^*) P_{T_{x^*}} (x_{k+1} - x^*) + o(\|d_k\|). \end{aligned}$$

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

$$979 \quad (B.16) \quad \begin{aligned} &(\text{Id} + \gamma_k \nabla_{\mathcal{M}_{x^*}}^2 \Phi(x^*) P_{T_{x^*}} - \gamma_k P_{T_{x^*}} \nabla^2 F(x^*) P_{T_{x^*}}) r_{k+1} \\ &= (\text{Id} + U_k) r_{k+1} = (y_{a,k} - x^*) - H_k (y_{b,k} - x^*) + o(\|d_k\|), \end{aligned}$$

980 which can be further written as,

$$\begin{aligned} (\text{Id} + U_k) r_{k+1} &= (\text{Id} + U) r_{k+1} + (U_k - U) r_{k+1} = (y_{a,k} - x^*) - H_k (y_{b,k} - x^*) + o(\|d_k\|) \\ &= ((1 + a_k) r_k - a_k r_{k-1}) - H_k ((1 + b_k) r_k - b_k r_{k-1}) + o(\|d_k\|) \\ 981 \quad &= ((1 + a_k) r_k - (1 + b_k) H_k r_k) - (a_k r_{k-1} - b_k H_k r_{k-1}) + o(\|d_k\|) \\ &= ((a_k - b_k) \text{Id} + (1 + b_k) G_k) r_k - ((a_k - b_k) \text{Id} + b_k G_k) r_{k-1} + o(\|d_k\|) \\ &= [(a_k - b_k) \text{Id} + (1 + b_k) G_k \quad -((a_k - b_k) \text{Id} + b_k G_k)] d_k + o(\|d_k\|). \end{aligned}$$

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

$$983 \quad \begin{aligned} &r_{k+1} + W(U_k - U) r_{k+1} \\ &= [(a_k - b_k) W + (1 + b_k) W G_k \quad -(a_k - b_k) W - b_k W G_k] d_k + o(\|d_k\|). \end{aligned}$$

984 Using the estimates (B.7), we get

$$\begin{aligned} 985 \quad d_{k+1} &= \begin{bmatrix} (a_k - b_k) W + (1 + b_k) W G_k & -(a_k - b_k) W - b_k W G_k \\ \text{Id} & 0 \end{bmatrix} d_k + o(\|d_k\|) \\ &= \left(M + \begin{bmatrix} M_{k,1} \\ 0 \end{bmatrix} + \begin{bmatrix} M_{k,2} \\ 0 \end{bmatrix} \right) d_k + o(\|d_k\|) = M d_k + o(\|d_k\|). \end{aligned} \quad \square$$

986 *Proof (Proposition 17).*

987 (i) We have

$$988 \quad M \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} (a-b)r_1 + (1+b)Gr_1 - (a-b)r_2 - bGr_2 \\ r_1 \end{pmatrix} = \sigma \begin{pmatrix} r_1 \\ r_2 \end{pmatrix},$$

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

$$\begin{aligned} 990 \quad \sigma^2 r_2 &= (a-b)\sigma r_2 + (1+b)\sigma Gr_2 - (a-b)r_2 - bGr_2 \\ &\Leftrightarrow Gr_2 = (((a-b)(1-\sigma) + \sigma^2) / ((1+b)\sigma - b)) r_2 = \eta r_2 \\ &\Rightarrow 0 = \sigma^2 - ((a-b) + (1+b)\eta)\sigma + (a-b) + b\eta. \end{aligned}$$

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

$$992 \quad (B.17) \quad \sigma_1 = ((a-b) + (1+b)\eta + \sqrt{\Delta_\sigma})/2, \quad \sigma_2 = ((a-b) + (1+b)\eta - \sqrt{\Delta_\sigma})/2.$$

where $\Delta_\sigma = ((a-b) + (1+b)\eta)^2 - 4((a-b) + b\eta)$ is the discriminant, which is a quadratic polynomial of three variables. Consider the following three linear functions of a

(B.18)

$$a_1 = (1-\eta)b - \eta, \quad a_3 = (1-\eta)b - (1+\eta)/2$$

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

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

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

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)

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

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

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

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

Summarizing this discussion yields the claimed result. \square

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

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

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

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

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

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

It is straightforward to check that $N_k \stackrel{\text{def}}{=} \begin{bmatrix} (1+a_k)\text{Id} & -a_k\text{Id} \\ \text{Id} & 0_n \end{bmatrix}$ is invertible and admits two eigenvalues $a_k > 0$ and 1 respectively. Iterating the above argument, and owing

1030 to the fact that $x_k, y_{a,k}, y_{b,k} \rightarrow x^*$, we get

$$1031 \quad \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \left(\prod_{j=k}^{\infty} N_j \right) \begin{pmatrix} P_V(x_k - x^*) \\ P_V(x_{k-1} - x^*) \end{pmatrix},$$

1032 and $\prod_{j=k}^{\infty} N_j$ is invertible. Therefore, we obtain that $x_k - x^* \in V^\perp$, and in turn,
 1033 $y_{a,k} - x^* \in V^\perp$ and $y_{b,k} - x^* \in V^\perp$, for all large enough k . Observe that $V^\perp \subset T_{x^*}$,
 1034 it then follows that

$$1035 \quad x_{k+1} - x^* = y_{a,k} - x^* - P_{V^\perp} E_k P_{V^\perp} (y_{b,k} - x^*).$$

1036 By definition, $P_{V^\perp} E_k P_{V^\perp}$ is symmetric positive definite. Thus, substituting this
 1037 matrix for H_k , and G and M accordingly in Lemma 14 and Corollary 19, and applying
 1038 Theorem 21, leads to the result. \square

1039

REFERENCES

- 1040 [1] P.-A. ABSIL, R. MAHONY, AND R. SEPULCHRE, *Optimization algorithms on matrix manifolds*,
 1041 Princeton University Press, 2009.
- 1042 [2] P.-A. ABSIL, R. MAHONY, AND J. TRUMPF, *An extrinsic look at the Riemannian Hessian*, in
 1043 Geometric Science of Information, Springer, 2013, pp. 361–368.
- 1044 [3] A. AGARWAL, S. NEGAHBAN, AND M. J. WAINWRIGHT, *Fast global convergence of gradient*
 1045 *methods for high-dimensional statistical recovery*, The Annals of Statistics, 40 (2012),
 1046 pp. 2452–2482.
- 1047 [4] F. ALVAREZ, *On the minimizing property of a second order dissipative system in Hilbert spaces*,
 1048 SIAM Journal on Control and Optimization, 38 (2000), pp. 1102–1119.
- 1049 [5] F. ALVAREZ AND H. ATTOUCH, *An inertial proximal method for maximal monotone operators*
 1050 *via discretization of a nonlinear oscillator with damping*, Set-Valued Analysis, 9 (2001),
 1051 pp. 3–11.
- 1052 [6] H. ATTOUCH, Z. CHBANI, J. PEYPOUQUET, AND P. REDONT, *Fast convergence of inertial dy-*
 1053 *namics and algorithms with asymptotic vanishing damping*, Tech. Report Optimization
 1054 online 5179, 2015.
- 1055 [7] H. ATTOUCH AND J. PEYPOUQUET, *The rate of convergence of Nesterov’s accelerated Forward–*
 1056 *Backward method is actually $o(k^{-2})$* , Tech. Report arXiv:1510.08740, 2015.
- 1057 [8] H. ATTOUCH, J. PEYPOUQUET, AND P. REDONT, *A dynamical approach to an inertial*
 1058 *Forward–Backward algorithm for convex minimization*, SIAM J. Optim., 24 (2014),
 1059 pp. 232–256.
- 1060 [9] H. ATTOUCH, J. PEYPOUQUET, AND P. REDONT, *On the fast convergence of an inertial*
 1061 *gradient-like dynamics with vanishing viscosity*, Tech. Report arXiv:1507.04782, 2015.
- 1062 [10] J.-F. AUJOL AND C. DOSSAL, *Stability of over-relaxations for the Forward–Backward algo-*
 1063 *algorithm, application to FISTA*, SIAM Journal on Optimization, 25 (2015), pp. 2408–2433.
- 1064 [11] J. B. BAILLON AND G. HADDAD, *Quelques propriétés des opérateurs angle-bornés et-*
 1065 *rycliquement monotones*, Israel Journal of Mathematics, 26 (1977), pp. 137–150.
- 1066 [12] H. BAUSCHKE AND P. L. COMBETTES, *Convex Analysis and Monotone Operator Theory in*
 1067 *Hilbert Spaces*, Springer, 2011.
- 1068 [13] A. BECK AND M. TEOULLE, *Fast gradient-based algorithms for constrained total variation*
 1069 *image denoising and deblurring problems*, Image Processing, IEEE Transactions on, 18
 1070 (2009), pp. 2419–2434.
- 1071 [14] A. BECK AND M. TEOULLE, *A fast iterative shrinkage-thresholding algorithm for linear in-*
 1072 *verse problems*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 183–202.
- 1073 [15] J. BOLTE, A. DANILIDIS, AND A. LEWIS, *The Łojasiewicz inequality for nonsmooth subanalytic*
 1074 *functions with applications to subgradient dynamical systems*, SIAM J. Optim, 17 (2006),
 1075 pp. 1205–1223.
- 1076 [16] K. BREDIES AND D. A. LORENZ, *Linear convergence of iterative soft-thresholding*, Journal of
 1077 Fourier Analysis and Applications, 14 (2008), pp. 813–837.
- 1078 [17] R. S. BURACHIK AND A. N. IUSEM, *Set-valued Mappings and Enlargements of Monotone Op-*
 1079 *erators*, Optimization and Its Applications, Springer, 2008.
- 1080 [18] A. CHAMBOLLE AND C. DOSSAL, *On the convergence of the iterates of the “fast iterative shrink-*
 1081 *age/thresholding algorithm”*, Journal of Optimization Theory and Applications, 166 (2015),
 1082 pp. 968–982.
- 1083 [19] I. CHAVEL, *Riemannian geometry: a modern introduction*, vol. 98, Cambridge University Press,
 1084 2006.

- 1085 [20] P. L. COMBETTES, *Quasi-Fejérian analysis of some optimization algorithms*, Studies in Com-
 1086 putational Mathematics, 8 (2001), pp. 115–152.
- 1087 [21] P. L. COMBETTES AND B. C. VŰ, *Variable metric Forward–Backward splitting with applica-*
 1088 *tions to monotone inclusions in duality*, Optimization, 63 (2014), pp. 1289–1318.
- 1089 [22] A. DANILIDIS, D. DRUSVYATSKIY, AND A. S. LEWIS, *Orthogonal invariance and identifiability*,
 1090 SIAM Journal on Matrix Analysis and Applications, 35 (2014), pp. 580–598.
- 1091 [23] A. DANILIDIS, W. HARE, AND J. MALICK, *Geometrical interpretation of the predictor-*
 1092 *corrector type algorithms in structured optimization problems*, Optimization: A Journal
 1093 of Mathematical Programming & Operations Research, 55 (2009), pp. 482–503.
- 1094 [24] T. GOLDSTEIN, B. O’DONOGHUE, S. SETZER, AND R. BARANIUK, *Fast alternating direction*
 1095 *optimization methods*, SIAM Journal on Imaging Sciences, 7 (2014), pp. 1588–1623.
- 1096 [25] M. GU, L.-H. LIM, AND C. J. WU, *ParNes: a rapidly convergent algorithm for accurate*
 1097 *recovery of sparse and approximately sparse signals*, Numerical Algorithms, 64 (2012),
 1098 pp. 321–347.
- 1099 [26] E. HALE, W. YIN, AND Y. ZHANG, *Fixed-point continuation for ℓ_1 -minimization: methodology*
 1100 *and convergence*, SIAM Journal on Optimization, 19 (2008), pp. 1107–1130.
- 1101 [27] W. L. HARE, *Identifying active manifolds in regularization problems*, in Fixed-Point Algo-
 1102 rithms for Inverse Problems in Science and Engineering, H. H. Bauschke, R. S., Burachik,
 1103 P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, eds., vol. 49 of Springer Opti-
 1104 mization and Its Applications, Springer, 2011, ch. 13.
- 1105 [28] W. L. HARE AND A. S. LEWIS, *Identifying active constraints via partial smoothness and prox-*
 1106 *regularity*, Journal of Convex Analysis, 11 (2004), pp. 251–266.
- 1107 [29] W. L. HARE AND A. S. LEWIS, *Identifying active manifolds*, Algorithmic Operations Research,
 1108 2 (2007), pp. 75–82.
- 1109 [30] J.-B. HIRIART-URRUTY AND C. LEMARÉCHAL, *Convex Analysis And Minimization Algorithms*,
 1110 vol. I and II, Springer, 2001.
- 1111 [31] K. HOU, Z. ZHOU, A. M.-C. SO, AND Z. LUO, *On the linear convergence of the proximal gra-*
 1112 *dient method for trace norm regularization*, in Advances in Neural Information Processing
 1113 Systems, 2013, pp. 710–718.
- 1114 [32] P. R. JOHNSTONE AND P. MOULIN, *A Lyapunov analysis of FISTA with local linear conver-*
 1115 *gence for sparse optimization*, arXiv preprint arXiv:1502.02281, (2015).
- 1116 [33] J. M. LEE, *Smooth manifolds*, Springer, 2003.
- 1117 [34] C. LEMARÉCHAL, F. OUSTRY, AND C. SAGASTIZÁBAL, *The U-Lagrangian of a convex function*,
 1118 Trans. Amer. Math. Soc., 352 (2000), pp. 711–729.
- 1119 [35] A. S. LEWIS, *Active sets, nonsmoothness, and sensitivity*, SIAM Journal on Optimization, 13
 1120 (2003), pp. 702–725.
- 1121 [36] A. S. LEWIS AND S. ZHANG, *Partial smoothness, tilt stability, and generalized Hessians*, SIAM
 1122 Journal on Optimization, 23 (2013), pp. 74–94.
- 1123 [37] J. LIANG, J. FADILI, AND G. PEYRÉ, *Local linear convergence of Forward–Backward under*
 1124 *partial smoothness*, in Advances in Neural Information Processing Systems, 2014, pp. 1970–
 1125 1978.
- 1126 [38] P. L. LIONS AND B. MERCIER, *Splitting algorithms for the sum of two nonlinear operators*,
 1127 SIAM Journal on Numerical Analysis, 16 (1979), pp. 964–979.
- 1128 [39] D. A. LORENZ AND T. POCK, *An accelerated Forward–Backward algorithm for monotone in-*
 1129 *clusions*, arXiv preprint arXiv:1403.3522, (2014).
- 1130 [40] S. A. MILLER AND J. MALICK, *Newton methods for nonsmooth convex minimization: connec-*
 1131 *tions among-Lagrangian, Riemannian Newton and SQP methods*, Mathematical program-
 1132 ming, 104 (2005), pp. 609–633.
- 1133 [41] B. MORDUKHOVICH, *Sensitivity analysis in nonsmooth optimization*, Theoretical Aspects of In-
 1134 dustrial Design (D. A. Field and V. Komkov, eds.), SIAM Volumes in Applied Mathematics,
 1135 58 (1992), pp. 32–46.
- 1136 [42] A. MOUDAFI AND M. OLINY, *Convergence of a splitting inertial proximal method for monotone*
 1137 *operators*, Journal of Computational and Applied Mathematics, 155 (2003), pp. 447–454.
- 1138 [43] Y. NESTEROV, *A method for solving the convex programming problem with convergence rate*
 1139 *$O(1/k^2)$* , Dokl. Akad. Nauk SSSR, 269 (1983), pp. 543–547.
- 1140 [44] Y. NESTEROV, *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer,
 1141 2004.
- 1142 [45] Y. NESTEROV, *Gradient methods for minimizing composite objective function*, (2007).
- 1143 [46] B. O’DONOGHUE AND E. CANDÉS, *Adaptive restart for accelerated gradient schemes*, Founda-
 1144 tions of computational mathematics, 15 (2015), pp. 715–732.
- 1145 [47] Z. OPIAL, *Weak convergence of the sequence of successive approximations for nonexpansive*
 1146 *mappings*, Bulletin of the American Mathematical Society, 73 (1967), pp. 591–597.

- 1147 [48] R. A. POLIQUIN AND R. T. ROCKAFELLAR, *Tilt stability of a local minimum*, SIAM Journal
1148 on Optimization, 8 (1998), pp. 287–299.
- 1149 [49] B. T. POLYACK, *Some methods of speeding up the convergence of iterative methods*, Zh. Vychisl.
1150 Mat. Mat. Fiz., 4 (1964), pp. 1–17.
- 1151 [50] B. T. POLYAK, *Introduction to optimization*, Optimization Software, 1987.
- 1152 [51] R. T. ROCKAFELLAR AND R. WETS, *Variational analysis*, vol. 317, Springer Verlag, 1998.
- 1153 [52] S. T. SMITH, *Optimization techniques on Riemannian manifolds*, Fields institute communica-
1154 tions, 3 (1994), pp. 113–135.
- 1155 [53] B. F. SVAITER AND R. S. BURACHIK, ε -enlargements of maximal monotone operators in banach
1156 spaces, Set-Valued Anal., 7 (1999), pp. 117–132.
- 1157 [54] S. TAO, D. BOLEY, AND S. ZHANG, *Local linear convergence of ISTA and FISTA on the LASSO*
1158 *problem*, SIAM Journal on Optimization, 26 (2016), pp. 313–336.
- 1159 [55] P. TSENG AND S. YUN, *A coordinate gradient descent method for nonsmooth separable mini-*
1160 *mization*, Math. Prog. (Ser. B), 117 (2009).
- 1161 [56] S. VAITER, M. GOLBABAEE, J. FADILI, AND G. PEYRÉ, *Model selection with low complexity*
1162 *priors*, Information and Inference, (2015), p. iav005.
- 1163 [57] S. VAITER, G. PEYRÉ, AND J. FADILI, *Model consistency of partly smooth regularizers*, arXiv
1164 preprint arXiv:1405.1004, (2014).
- 1165 [58] S. J. WRIGHT, *Identifiable surfaces in constrained optimization*, SIAM Journal on Control and
1166 Optimization, 31 (1993), pp. 1063–1079.