Convergence Rates of Forward-Douglas-Rachford Splitting Method

Cesare Molinari, Jingwei Liang, Jalal M. Fadili

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

Convergence Rates of Forward–Douglas–Rachford Splitting Method

Cesare Molinari · Jingwei Liang · Jalal Fadili

Received: date / Accepted: date

Abstract Over the past decades, operator splitting methods have become ubiquitous for non-smooth optimization owing to their simplicity and efficiency. In this paper, we consider the Forward–Douglas–Rachford splitting method, and study both global and local convergence rates of this method. For the global rate, we establish a sublinear convergence rate in terms of a Bregman divergence suitably designed for the objective function. Moreover, when specializing to the Forward–Backward splitting, we prove a stronger convergence rate result for the objective function value. Then locally, based on the assumption that the non-smooth part of the optimization problem is partly smooth, we establish local linear convergence of the method. More precisely, we show that the sequence generated by Forward–Douglas–Rachford first (i) identifies a smooth manifold in a finite number of iteration, and then (ii) enters a local linear convergence regime, which is for instance characterized in terms of the structure of the underlying active smooth manifold. To exemplify the usefulness of the obtained result, we consider several concrete numerical experiments arising from applicative fields including, for instance, signal/image processing, inverse problems and machine learning.

Keywords Forward–Douglas–Rachford · Forward–Backward · Bregman Distance · Partial Smoothness · Finite Identification · Local Linear Convergence

Mathematics Subject Classification (2000) 49J52 · 65K05 · 65K10 · 90C25

Cesare Molinari, Jalal Fadili
Normandie Université, GREYC, ENSICAEN, CNRS, France
E-mail: cecio.molinari@gmail.com; Jalal.Fadili@ensicaen.fr

Jingwei Liang, Corresponding author
DAMTP, University of Cambridge, UK
E-mail: jl993@cam.ac.uk

1 Introduction

Operator splitting methods are iterative schemes to solve inclusion and optimization problems by decoupling the original problem into subproblems, that are easy to solve. These schemes evaluate the individual operators, their resolvents, the linear operators, all separately at various points in the course of iteration, but never the resolvents of sums nor of composition by a linear operator. Since the first operator splitting method developed in the 70's for solving structured monotone inclusion problems, the class of splitting methods has been regularly enriched with increasingly sophisticated algorithms as the structure of problems to handle become more complex. We refer the readers to [1] and references therein for a through account of operator splitting methods.

In this paper, we consider a subspace constrained optimization problem, where the objective function is the sum of a proper convex and lower semicontinuous function and a convex smooth differentiable function with Lipschitz gradient. To efficiently handle the constraint, a provably convergent algorithm is Forward–Douglas–Rachford splitting algorithm (FDR) [2], which is a hybridization of Douglas–Rachford splitting algorithm (DR) [3] and Forward–Backward splitting algorithm (FB) [4]. FDR is also closely related to the generalized Forward–Backward splitting algorithm (GFB) [5,6] and the three-operator splitting method (TOS) [7].

Global sub-linear convergence rate to asymptotic regularity of the sequence generated by FDR (hence all the above-mentioned algorithms) has been recently established in the literature, from the perspective of Krasnosel'skiĭ-Mann fixed-point iteration; see, for instance, [8] and the references therein. This allows to exhibit convergence rates of the distance of 0 to the objective subdifferential evaluated at the iterate. However, very limited results have been reported in the literature on the convergence rate of the objective function value for FDR, except for certain specific cases. For instance, the objective convergence rate of Forward–Backward splitting and its accelerated versions are now well understood [9,10,11,12,13,14]. These results rely essentially on some monotonicity property of a properly designed Lyapunov function. Given that FDR is fixed-point algorithm, it is much more difficult or even impossible to study the convergence rate of the objective function value. Indeed, these algorithms generate several different points along the course of iteration, making it rather challenging to design a proper Lyapunov function (as we shall see for the FDR algorithm in Section 4).

Recently, local linear convergence of operator splitting algorithms for optimization have recently attracted a lot of attention; see [15] for Forward–Backward-type methods, [16] for Douglas–Rachford splitting, and [17] for Primal–Dual splitting algorithms. This work particularly exploits the underlying geometric structure of the optimization problems, achieving a local linear convergence result without assuming conditions like strong convexity, unlike what is proved in [18,8]. In practice, local linear convergence of FDR algorithm is also observed. However, to our knowledge, there is no theoretical explanation available for this local behaviour.

Main Contributions In this paper, we study both the global and local convergence rates of the FDR algorithm. Our main contributions consist of both global and local aspects. First, the global convergence behaviour is studied under a general real Hilbert space setting.

- In Section 4, we first prove the convergence of the newly proposed non-stationary FDR scheme (6). This is achieved by capturing non-stationarity as an error term. The proof exploits a general result on inexact and non-stationary Krasnosel'skiĭ-Mann fixed-point iteration developed in [8].
- We design a Bregman divergence as a meaningful convergence criterion. Under the standard assumptions, we show pointwise and ergodic convergence rates of this criterion (Theorem 4.2). When specializing the result to Forward–Backward splitting, we obtain a stronger claim for the objective convergence rate of the method. The allowed range of step-size for the latter rate to hold is twice larger than the one known in the literature.

For local convergence analysis, we turn to finite-dimension as partial smoothness, which is at the heart of this part is only available in the Euclidean setting.

- Finite Time Activity Identification Under the assumption that the non-smooth component of the optimization problem is partly smooth around a global minimizer relative to its smooth submanifold (see Definition 2.4) and under a non-degeneracy condition (see (31)), we show in Section 5 (Theorem 5.1) that the sequence generated by the non-stationary FDR identifies in finite time the solution submanifold. In plain words, this means that, after a finite number of iterations, the sequence enters the submanifold and never leaves it. We also provide a bound on the number of iterations to achieve identification.
- Local Linear Convergence Exploiting the finite identification property, we then show that the sequence generated by non-stationary FDR converges locally linearly. We characterize the convergence rate precisely based on the properties of the identified partial smoothness submanifolds.
- Three-operator Splitting Given the close relation between the three-operator splitting method and FDR, in Section 5.4, we extend the above local linear convergence result to the case of the three-operator splitting algorithm.

Relation to Prior Work The convergence rate of the objective value for FDR has been studied in [18]. There, the author presented ergodic and pointwise convergence rates on the objective value under different (more or less stringent) assumptions imposed on the non-smooth function in the objective (1). Without any further assumptions other than (**A.1**)-(**A.5**), the author proved a pointwise convergence rate on a criterion associated to the objective value, but in absolute value (see [18, Theorem 3.5]). However, this rate seems quite pessimistic (it suggests that FDR is as slow as sub-gradient descent). Moreover, there is no non-negativity guarantee for such criterion and the obtained rate is thus of a quite limited interest. Improving this rate on the objective value requires quite strong assumptions on the non-smooth component.

As far as local linear convergence of the sequence in absence of strong convexity is concerned, it has received an increasing attention in the past few

years in the context of first-order proximal splitting methods. The key idea here is to exploit the geometry of the underlying objective around its minimizers. This has been done for instance in [19,15,16,17] for the FB scheme, Douglas–Rachford splitting/ADMM and Primal–Dual splitting, under the umbrella of partial smoothness. The error bound property[1], as highlighted in the seminal work of [22,23], is used by several authors to study linear convergence of first-order descent-type algorithms, and in particular FB splitting; see, *e.g.*, [20,24, 25,21]. However, to the best of our knowledge, we are not aware of local linear convergence results for the FDR algorithm.

Paper Organization The rest of the paper is organized as follows. In Section 2, we recall some classical material on convex analysis and operator theory, that are essential to our exposition. We then introduce the notion of partial smoothness. The problem statement and FDR algorithm are presented in Section 3. The global convergence analysis is presented in Section 4, followed by finite identification and local convergence analysis in Section 5. Several numerical experiments are presented in Section 6. Some introductory material on smooth Riemannian manifolds is gathered in the appendix.

2 Preliminaries

Throughout the paper, $\mathcal{H}$ is a Hilbert space equipped with scalar product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. Id denotes the identity operator on $\mathcal{H}$. $\Gamma_0(\mathcal{H})$ denotes the set of proper convex and lower semi-continuous functions on $\mathcal{H}$.

Sets For a non-empty convex set $C \subset \mathcal{H}$, $\mathrm{par}(C) := \mathbb{R}(C - C)$ the smallest subspace parallel to C. Denote ι_C the indicator function of C, $\mathcal{N}_C$ the associated normal cone operator and P_C the orthogonal projection on C. The strong relative interior of C is $\mathrm{sri}(C)$.

Functions Given $R \in \Gamma_0(\mathcal{H})$, its sub-differential is a set-valued operator defined by $\partial R : \mathcal{H} \rightrightarrows \mathcal{H}$, $x \mapsto \big\{v \in \mathcal{H} : R(x') \geq R(x) + \langle v, x' - x \rangle, \forall x' \in \mathcal{H}\big\}$.

Lemma 2.1 (Descent Lemma [26]) *Suppose that $F : \mathcal{H} \to \mathbb{R}$ is convex continuously differentiable and ∇F is $(1/\beta)$-Lipschitz continuous. Then,*

$$F(x) \leq F(y) + \langle \nabla F(y), x - y \rangle + \frac{1}{2\beta}\|x - y\|^2, \ \forall x, y \in \mathcal{H}.$$

Definition 2.1 (Bregman Divergence) Given a function $R \in \Gamma_0(\mathcal{H})$ and two points x, y in its effective domain $\mathrm{dom}(R)$, the Bregman divergence is defined by

$$\mathcal{D}_R^v(y, x) := R(y) - R(x) - \langle v, y - x \rangle,$$

where $v \in \partial R(x)$ is a sub-gradient of R.

[1] For the interplay between the error bound property, the Kurdyka-Łojasiewicz property, and the quadratic growth property; see [20,21].

Notice that the Bregman divergence is not a distance in the usual sense, as it is in general not symmetric[2]. However, it measures the distance of two points in the sense that $\mathcal{D}_R^v(x,x) = 0$ and $\mathcal{D}_R^v(y,x) \geq 0$ for any x, y in $\mathrm{dom}(R)$. Moreover, $\mathcal{D}_R^v(y,x) \geq \mathcal{D}_R^v(w,x)$ for all w in the line segment between x and y.

Operators Given a set-valued mapping $A : \mathcal{H} \rightrightarrows \mathcal{H}$, define its graph as $\mathrm{gph}\,(A) := \{(x,u) \in \mathcal{H} \times \mathcal{H} : \ u \in A(x)\}$, and set of zeros $\mathrm{zer}(A) = \{x \in \mathcal{H} : 0 \in A(x)\}$. Denote $(\mathrm{Id} + A)^{-1}$ the resolvent of A

Definition 2.2 (Cocoercive Operator) Let $\beta > 0$ and $B : \mathcal{H} \to \mathcal{H}$, then B is β-cocoercive, if $\langle B(x_1) - B(x_2),\, x_1 - x_2 \rangle \geq \beta \|B(x_1) - B(x_2)\|^2$, $\forall x_1, x_2 \in \mathcal{H}$.

If an operator is β-cocoercive, then it is β^{-1}-Lipschitz continuous.

Definition 2.3 (Non-expansive Operator) An operator $\mathcal{F} : \mathcal{H} \to \mathcal{H}$ is non-expansive, if $\|\mathcal{F}(x) - \mathcal{F}(y)\| \leq \|x - y\|, \ \ \forall x, y \in \mathcal{H}$. For any $\alpha \in]0,1[$, $\mathcal{F}$ is called α-averaged, if there exists a non-expansive operator $\mathcal{F}'$ such that $\mathcal{F} = \alpha\mathcal{F}' + (1-\alpha)\mathrm{Id}$.

In particular, when $\alpha = \frac{1}{2}$, $\mathcal{F}$ is called *firmly non-expansive*. Several properties of firmly non-expansive operators are collected in the following lemma.

Lemma 2.2 *Let $\mathcal{F} : \mathcal{H} \to \mathcal{H}$, the following statements are equivalent:*
(i) *$\mathcal{F}$ is firmly non-expansive;*
(ii) *$2\mathcal{F} - \mathrm{Id}$ is non-expansive;*
(iii) *$\mathcal{F}$ is the resolvent of a maximal monotone operator $A : \mathcal{H} \rightrightarrows \mathcal{H}$.*

Proof (i)$\Leftrightarrow$(ii) follows [1, Proposition 4.2, Corollary 4.29], and (i)$\Leftrightarrow$(iii) is [1, Corollary 23.8]. □

Lemma 2.3 ([1, Proposition 4.33]) *Let $F : \mathcal{H} \to \mathbb{R}$ be a convex differentiable function, with $\frac{1}{\beta}$-Lipschitz continuous gradient, $\beta \in]0, +\infty[$, then $\mathrm{Id} - \gamma\nabla F$ is $\frac{\gamma}{2\beta}$-averaged for $\gamma \in]0, 2\beta[$.*

The next lemma shows the composition of two averaged operators.

Lemma 2.4 ([27, Theorem 3]) *Let $\mathcal{F}_1, \mathcal{F}_2 : \mathcal{H} \to \mathcal{H}$ be α_1, α_2-averaged respectively, then $\mathcal{F}_1 \circ \mathcal{F}_2$ is α-averaged with $\alpha = \frac{\alpha_1 + \alpha_2 - 2\alpha_1\alpha_2}{1 - \alpha_1\alpha_2} \in]0,1[$.*

Sequence The following lemma is very classical, see *e.g.* [28, Theorem 3.3.1].

Lemma 2.5 *Let the non-negative sequence $\{a_k\}_{k\in\mathbb{N}}$ be non-increasing and summable. Then $a_k = o(k^{-1})$.*

[2] It is symmetric, if and only if R is a non-degenerate convex quadratic form.

Partial Smoothness In this part, let $\mathcal{H} = \mathbb{R}^n$. We briefly introduce the concept of partial smoothness, which was introduced in [29] and lays the foundation of our local convergence analysis.

Let $\mathcal{M}$ be a C^2-smooth manifold of $\mathbb{R}^n$ around a point x. Denote $\mathcal{T}_{\mathcal{M}}(x')$ the tangent space to $\mathcal{M}$ at any point near x in $\mathcal{M}$; See Section 8 for more materials. Below we present the definition of partly smooth functions in $\Gamma_0(\mathbb{R}^n)$ setting.

Definition 2.4 (Partly Smooth Function) Let $R \in \Gamma_0(\mathbb{R}^n)$, and $x \in \mathbb{R}^n$ such that $\partial R(x) \neq \emptyset$. R is then said to be *partly smooth* at x relative to a set $\mathcal{M}$ containing x, if

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

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

Popular examples of partly smooth functions are summarized in Section 6 whose details can be found in [15].

3 Problem and Algorithms

Non-smooth Optimization In this paper, we are interested in the following structured convex optimization problem

$$\min_{x\in\mathcal{H}} \big\{F(x) + R(x) : \quad x \in V\big\}, \tag{1}$$

where the following assumptions are imposed

(**A.1**) R belongs to $\Gamma_0(\mathcal{H})$.
(**A.2**) $F : \mathcal{H} \to \mathbb{R}$ is convex continuously differentiable with ∇F being $(1/\beta)$-Lipschitz continuous.
(**A.3**) The constraint set V is a closed vector subspace of $\mathcal{H}$.
(**A.4**) $\mathrm{Argmin}_V(F + R)$ is non-empty and $0 \in \mathrm{sri}(\mathrm{dom}(R) - V)$.

Typical examples of (1) can be found in the numerical experiment section. These assumptions entail that $F + R + \iota_V \in \Gamma_0(\mathcal{H})$, and moreover, that

$$\mathrm{zer}(\nabla F + \partial R + \mathcal{N}_V) = \mathrm{zer}(\partial(F + R + \iota_V)) = \mathrm{Argmin}_V(F + R) \neq \emptyset,$$

using [1, Theorem 16.37(i)] and Fermat's rule.

Forward–Douglas–Rachford Splitting When $V = \mathcal{H}$, problem (1) can be handled by the classical Forward–Backward splitting method [4], whose iteration, in its relaxed form, reads

$$x_{k+1} = (1 - \lambda_k)x_k + \lambda_k \mathrm{prox}_{\gamma R}\big(x_k - \gamma\nabla F(x_k)\big), \tag{2}$$

where $\gamma \in]0, 2\beta[$ is the step-size and $\lambda_k \in]0, \frac{4\beta-\gamma}{2\beta}[$ is the relaxation parameter. The term $\mathrm{prox}_{\gamma R}$ is called the proximity operator of γR and is defined by

$$\mathrm{prox}_{\gamma R}(x) := \mathrm{argmin}_{u\in\mathcal{H}}\, \gamma R(u) + \tfrac{1}{2}\|u - x\|^2. \tag{3}$$

When V is merely a subspace of $\mathcal{H}$, in principle we still can apply FB splitting method to solve (1). However, even if $\mathrm{prox}_{\gamma R}$ is very easy to compute, the proximity operator of $R + \iota_V$ in general may be rather difficult to calculate. Therefore, new splitting algorithms are needed, and one possible choice is the Forward–Douglas–Rachford splitting method [2] which will be presented shortly. Let us first define P_V as the orthogonal projector onto the subspace V, and the function $G{:=}F \circ \mathrm{P}_V$. Then (1) is, obviously, equivalent to

$$\min_{x\in\mathcal{H}} \big\{\Phi_V(x){:=}G(x) + R(x) + \iota_V(x)\big\}. \tag{4}$$

In turn, owing to assumptions (**A.1**)-(**A.4**), we have

$$\emptyset \neq \mathrm{Argmin}_V(F + R) = \mathrm{Argmin}(\Phi_V) = \mathrm{zer}(\nabla G + \partial R + \mathcal{N}_V).$$

Remark 3.1 From the assumption on F, we have that also G is convex and continuously differentiable with $\nabla G = \mathrm{P}_V \circ \nabla F \circ \mathrm{P}_V$ being $(1/\beta_V)$-Lipschitz continuous (notice that $\beta_V \geq \beta$). The observation of using G instead of F to achieve a better Lipschitz condition was first considered in [7].

The iteration of FDR method for solving (4) reads

$$\begin{aligned} u_{k+1} &= \mathrm{prox}_{\gamma R}\big(2x_k - z_k - \gamma\nabla G(x_k)\big),\\ z_{k+1} &= z_k + \lambda_k(u_{k+1} - x_k),\\ x_{k+1} &= \mathrm{P}_V(z_{k+1}), \end{aligned} \tag{5}$$

where γ is step-size and λ_k is relaxation parameter. Recall that, under the conditions that $\gamma \in]0, 2\beta_V[$, $\lambda_k \in]0, \frac{4\beta_V-\gamma}{2\beta_V}[$ and $\sum_{k\in\mathbb{N}} \lambda_k\big(\frac{4\beta_V-\gamma}{2\beta_V} - \lambda_k\big) = +\infty$, the sequences $\{u_k\}_{k\in\mathbb{N}}$, $\{x_k\}_{k\in\mathbb{N}}$ converge to a solution; see [2, Theorem 4.2].

In this paper, we consider a non-stationary version of (5), namely γ may change along the iterations. The method is described below in Algorithm 1.

Algorithm 1: Non-stationary Forward–Douglas–Rachford

Initial: $k = 0, z_0 \in \mathcal{H}, x_0 = \mathrm{P}_V(z_0)$.

repeat

$$\left\lfloor \begin{aligned} u_{k+1} &= \mathrm{prox}_{\gamma_k R}\big(2x_k - z_k - \gamma_k\nabla G(x_k)\big),\ \ \gamma_k \in]0, 2\beta_V[,\\ z_{k+1} &= z_k + \lambda_k(u_{k+1} - x_k),\ \ \lambda_k \in]0, \tfrac{4\beta_V-\gamma}{2\beta_V}[,\\ x_{k+1} &= \mathrm{P}_V(z_{k+1}). \end{aligned} \right. \tag{6}$$

until *convergence*;

Remark 3.2 For global convergence, one can also consider an inexact version of (6) by incorporating additive errors in the computation of u_k and x_k, though we do not elaborate more on this for the sake of local convergence analysis. One can consult [8] for more details on this aspect.

In the next, we suppose the following main assumption on the parameters:

(**A.5**) The sequence of the step-sizes $\{\gamma_k\}_{k\in\mathbb{N}}$ and the one of the relaxation parameters $\{\lambda_k\}_{k\in\mathbb{N}}$ verify:
- $0 < \underline{\gamma} \leq \gamma_k \leq \bar{\gamma} < 2\beta_V$ and $\gamma_k \to \gamma$ for some $\gamma \in [\underline{\gamma}, \bar{\gamma}]$;
- $\lambda_k \in]0, \frac{4\beta_V - \gamma_k}{2\beta_V}[$ such that $\sum_{k\in\mathbb{N}} \lambda_k(\frac{4\beta_V-\gamma_k}{2\beta_V} - \lambda_k) = +\infty$;
- $\sum_{k\in\mathbb{N}} \lambda_k|\gamma_k - \gamma| < +\infty$.

Notice that, for the stationary case (*i.e.* for γ_k constant), assumption (**A.5**) is equivalent to the conditions required in [2, Theorem 4.2] for the convergence of iteration (5). Moreover, to satisfy (**A.5**) in absence of relaxation (*i.e.* when the relaxation parameter is fixed to $\lambda_k \equiv 1$), the sequence of the step-sizes has just to verify $\gamma_k \in]\underline{\gamma}, \bar{\gamma}[$ with $\sum_{k\in\mathbb{N}} |\gamma_k - \gamma| < +\infty$. On the other hand, in general, the summability assumption of $\{\lambda_k|\gamma_k - \gamma|\}_{k\in\mathbb{N}}$ in (**A.5**) is weaker than imposing it without λ_k. Indeed, following the discussion in [30, Remark 5.7], take $q \in]0, 1]$, let $\theta = \frac{4\beta_V - \bar{\gamma}}{4\beta_V} > \frac{1}{2}$ and

$$\lambda_k = \theta - \sqrt{\theta - 1/(2k)} \quad \text{and} \quad |\gamma_k - \gamma| = (\theta + \sqrt{\theta - 1/(2k)})/k^q.$$

Then, it can be verified that

$$\sum_{k\in\mathbb{N}}|\gamma_k - \gamma| = +\infty, \ \sum_{k\in\mathbb{N}}\lambda_k|\gamma_k - \gamma| = \frac{1}{2k^{1+q}} < +\infty \quad \text{and}$$
$$\sum_{k\in\mathbb{N}}\lambda_k\big(\tfrac{4\beta_V-\gamma_k}{2\beta_V} - \lambda_k\big) \geq \sum_{k\in\mathbb{N}}\lambda_k(2\theta - \lambda_k) = \sum_{k\in\mathbb{N}}\frac{1}{2k} = +\infty.$$

As previously mentioned, FDR recovers DR [3] when $F = 0$, and FB [4] when $V = \mathcal{H}$. We briefly introduce below two other closely related operator splitting methods: the generalized Forward–Backward splitting (GFB) [6] and the three-operator splitting (TOS) [7].

Generalized Forward–Backward Splitting Let $m > 0$ be a positive integer. Now for problem (1), let $V = \mathcal{H}$ and suppose we have m non-smooth functionals. The problem then becomes: let $R_i \in \Gamma_0(\mathcal{H})$ for each $i = 1, ..., m$

$$\min_{x\in\mathcal{H}} \big\{F(x) + \textstyle\sum_{i=1}^{m} R_i(x)\big\}, \tag{7}$$

Similar to the situation of FDR algorithm, even if the proximity operator of each R_i can be solved easily, the proximity of the sum of them can be intractable. In [6], the authors propose the GFB algorithm, which achieves the full splitting of the evaluation of the proximity operator of each R_i. Let $(\omega_i)_i \in]0, 1[^m$ such that $\sum_{i=1}^{m} \omega_i = 1$, choose $\gamma \in]0, 2\beta[$ and $\lambda_k \in]0, \frac{4\beta-\gamma}{2\beta}[$:

$$\begin{aligned}
&\text{from } i = 1 \text{ to } m\text{:}\\
&\left\lfloor \begin{aligned} u_{i,k+1} &= \operatorname{prox}_{\frac{\gamma}{\omega_i}R_i}\big(2x_k - z_{i,k} - \gamma\nabla F(x_k)\big)\\ z_{i,k+1} &= z_{i,k} + \lambda_k(u_{i,k+1} - x_k)\end{aligned}\right.\\
&x_{k+1} = \textstyle\sum_{i=1}^{m}\omega_i z_{i,k+1}.
\end{aligned} \tag{8}$$

We refer to [6] for more details of the GFB algorithm. Now define the product space $\boldsymbol{\mathcal{H}} := \mathcal{H} \times ... \times \mathcal{H}$, equipped with proper inner product and norm, the

subspace $\boldsymbol{\mathcal{S}}:=\{\boldsymbol{x}=(x_i)_{i=1,\dots,m}\in\boldsymbol{\mathcal{H}}: x_1=\dots=x_m\}\subset\boldsymbol{\mathcal{H}}$ and let the weights be $\omega_i=\frac{1}{m}, i=1,\dots,m$. Then it can be shown that GFB algorithm is equivalent to applying FDR to the following problem:

$$\min_{\boldsymbol{x}\in\boldsymbol{\mathcal{H}}} F\big(\tfrac{1}{m}\textstyle\sum_{i=1}^m x_i\big)+\sum_{i=1}^m R_i(x_i)+\iota_{\boldsymbol{\mathcal{S}}}(\boldsymbol{x}).$$

We refer to [2,5] for more connections between FDR and GFB.

Three-Operator Splitting Let $m=2$ in problem (7), then it becomes

$$\min_{x\in\mathcal{H}} F(x)+R_1(x)+R_2(x). \tag{9}$$

Notice that (9) can be handled by GFB as it is only a special case of (7). In [7] the author proposed a splitting scheme which resembles FDR yet different: given $\gamma\in]0,2\beta[$ and $\lambda_k\in]0,\frac{4\beta-\gamma}{2\beta}[$, the iteration of TOS reads as follows:

$$\begin{aligned} u_{k+1}&=\operatorname{prox}_{\gamma R_1}\big(2x_k-z_k-\gamma\nabla F(x_k)\big)\\ z_{k+1}&=z_k+\lambda_k(u_{k+1}-x_k),\\ x_{k+1}&=\operatorname{prox}_{\gamma R_2}(z_{k+1}).\end{aligned} \tag{10}$$

It can be observed that the projection operator P_V of FDR is replaced by the proximity operator $\operatorname{prox}_{\gamma R_2}$. Though the difference is only for the update of x_{k+1}, their fixed-point operators are quite different; see in Section 5.4.

4 Global Convergence

In this section, we deliver the global convergence analysis of the non-stationary FDR (6) in a general real Hilbert space setting, including convergence rate.

4.1 Global Convergence of the Non-Stationary FDR

Define the reflection operators of γR and ι_V respectively as $\mathcal{R}_{\gamma R}:=2\operatorname{prox}_{\gamma R}-\mathrm{Id}$ and $\mathcal{R}_V:=2\mathrm{P}_V-\mathrm{Id}$. Moreover, define the following operators:

$$\mathscr{F}_\gamma:=\tfrac{1}{2}(\mathrm{Id}+\mathcal{R}_{\gamma R}\circ\mathcal{R}_V)(\mathrm{Id}-\gamma\nabla G)\quad\text{and}\quad\mathscr{F}_{\gamma,\lambda_k}:=(1-\lambda_k)\mathrm{Id}+\lambda_k\mathscr{F}_\gamma. \tag{11}$$

Then the (stationary) FDR iteration (5) can be written into a fixed-point iteration in terms of z_k [2, Theorem 4.2], namely

$$z_{k+1}=\mathscr{F}_{\gamma,\lambda_k}(z_k). \tag{12}$$

The next lemma shows the property of the fixed-point operator of FDR.

Lemma 4.1 *For the FDR algorithm* (6)*, let* $\gamma\in]0,2\beta_V[$ *and* $\lambda_k\in]0,\frac{4\beta_V-\gamma}{2\beta_V}[$. *Then, we have that* $\mathscr{F}_\gamma$ *is* $\frac{2\beta_V}{4\beta_V-\gamma}$*-averaged and* $\mathscr{F}_{\gamma,\lambda_k}$ *is* $\frac{2\beta_V\lambda_k}{4\beta_V-\gamma}$*-averaged.*

Proof The property of $\mathscr{F}_\gamma$ is a combination of Lemma 2.4 and [2, Proposition 4.1]. For $\mathscr{F}_{\gamma,\lambda_k}$, it is sufficient to apply the definition of averaged operators. □

Owing to [2, Theorem 4.2], under $\sum_{k\in\mathbb{N}} \lambda_k(\frac{4\beta_V-\gamma}{2\beta_V}-\lambda_k)=+\infty$ and conditions (A.1)-(A.4), $\{z_k\}_{k\in\mathbb{N}}$ converges weakly to some $z^\star \in \mathrm{fix}(\mathscr{F}_\gamma)$, and $\{x_k\}_{k\in\mathbb{N}}$ converges weakly to $x^\star{:=}\mathrm{P}_V(z^\star)$, where $\mathrm{P}_V(z^\star)\in \mathrm{Argmin}_V(F+R)$. On the other hand, the non-stationary FDR iteration (6) can be written as

$$z_{k+1}=\mathscr{F}_{\gamma_k,\lambda_k}(z_k)=\big((1-\lambda_k)z_k+\lambda_k\mathscr{F}_\gamma(z_k)\big)+\lambda_k\big(\mathscr{F}_{\gamma_k}(z_k)-\mathscr{F}_\gamma(z_k)\big). \tag{13}$$

We are now ready to state our result on global convergence of Algorithm 1.

Theorem 4.1 *Consider the non-stationary FDR iteration* (6)*. Suppose that Assumptions* (A.1)-(A.5) *hold. Then,* $\sum_{k\in\mathbb{N}}\|z_k-z_{k-1}\|^2<+\infty$*. Moreover,* $\{z_k\}_{k\in\mathbb{N}}$ *converges weakly to a point* $z^\star\in\mathrm{fix}(\mathscr{F}_\gamma)$*, and* $\{x_k\}_{k\in\mathbb{N}}$ *converges weakly to* $x^\star{:=}\mathrm{P}_V(z^\star)\in\mathrm{Argmin}_V(F+R)$*. If, in addition, either* $\inf_{k\in\mathbb{N}}\lambda_k>0$ *or* $\mathcal{H}$ *is finite-dimensional, then* $\{u_k\}_{k\in\mathbb{N}}$ *converges weakly to* $x^\star$*.*

The main idea of the proof of the theorem (see below) is to treat the non-stationarity as a perturbation error of the stationary iteration.

Remark 4.1

- As mentioned in the introduction, Theorem 4.1 remains true if the iteration is carried out inexactly, *i.e.* if $\mathscr{F}_{\gamma_k}(z_k)$ is computed approximately, provided that the errors are summable; see [8, Section 6] for more details.
- With more assumptions on how fast $\{\gamma_k\}_{k\in\mathbb{N}}$ converges to γ, we can also derive the convergence rate of the residuals $\{\|z_k-z_{k-1}\|\}_{k\in\mathbb{N}}$. However, as we will study in Section 5 local linear convergence behaviour of $\{z_k\}_{k\in\mathbb{N}}$, we shall forgo the discussion here. Interested readers can consult [8] for more details about the rate of residuals.

Proof According to [8, Theorem 4], the following conditions are needed to ensure the convergence of the non-stationary iteration:

(1) The set of fixed point of $\mathrm{fix}(\mathscr{F}_\gamma)$ is non-empty;
(2) $\forall k\in\mathbb{N}$, $\mathscr{F}_{\gamma_k}$ is 1-Lipschitz, *i.e.* non-expansive;
(3) $\lambda_k\in]0,\frac{4\beta_V-\gamma_k}{2\beta_V}[$ such that $\inf_{k\in\mathbb{N}}\lambda_k(\frac{4\beta_V-\gamma_k}{2\beta_V}-\lambda_k)>0$;
(4) $\forall\rho\in[0,+\infty[$ and $\Delta_{k,\rho}:=\sup_{\|z\|\leq\rho}\|\mathscr{R}_{\gamma_k}(z)-\mathscr{R}_\gamma(z)\|$ with $\mathscr{R}_{\gamma_k},\mathscr{R}_\gamma$ being some non-expansive operators, there holds $\sum_k\lambda_k\Delta_{k,\rho}<+\infty$.

Owing to Lemma 4.1, given $\gamma_k\in[0,2\beta_V]$, we have that $\mathscr{F}_{\gamma_k}$ is α_k-averaged with $\alpha_k=\frac{2\beta_V}{4\beta_V-\gamma_k}$. This means that there exists a non-expansive operator $\mathscr{R}_{\gamma_k}$ such that $\mathscr{F}_{\gamma_k}=\alpha_k\mathscr{R}_{\gamma_k}+(1-\alpha_k)\mathrm{Id}$. Similarly, for $\gamma\in[0,2\beta_V]$, we have that $\mathscr{F}_\gamma$ is α-averaged with $\alpha=\frac{2\beta_V}{4\beta_V-\gamma}$ and so that there exists a non-expansive operator $\mathscr{R}_\gamma$ such that $\mathscr{F}_\gamma=\alpha\mathscr{R}_\gamma+(1-\alpha)\mathrm{Id}$. Provided z_k, define the error term $e_k=(\mathscr{F}_{\gamma_k}-\mathscr{F}_\gamma)(z_k)$. Then iteration (13) can be written as

$$z_{k+1}=(1-\lambda)z_k+\lambda_k\mathscr{F}_{\gamma_k}(z_k)=(1-\lambda)z_k+\lambda_k\big(\mathscr{F}_\gamma(z_k)+e_k\big). \tag{14}$$

From the assumptions (A.1)-(A.5), we can derive the following results:

- We have $\operatorname{Argmin}_V(F+R) = \operatorname{zer}(\nabla G + \partial R + \mathcal{N}_V) = \mathrm{P}_V(\operatorname{fix}(\mathscr{F}_\gamma))$ from the discussion of Assumptions (**A.1**)-(**A.4**). It then follows that $\operatorname{fix}(\mathscr{F}_\gamma) \neq \emptyset$.
- Owing to Lemma 4.1, we have $\mathscr{F}_{\gamma_k,\lambda_k}$ is $(\alpha_k\lambda_k)$-averaged non-expansive.
- Owing to the averageness of $\mathscr{F}_\gamma$ and $\mathscr{F}_{\gamma_k}$, we have

$$\mathscr{R}_\gamma = \mathrm{Id} + \tfrac{1}{\alpha}(\mathscr{F}_\gamma - \mathrm{Id}) \text{ and } \mathscr{R}_{\gamma_k} = \mathrm{Id} + \tfrac{1}{\alpha_k}(\mathscr{F}_{\gamma_k} - \mathrm{Id}).$$

Let $\rho > 0$ be a positive number. Then, $\forall z \in \mathcal{H}$ such that $\|z\| \leq \rho$,

$$\begin{aligned} \|\mathscr{R}_{\gamma_k}(z) - \mathscr{R}_\gamma(z)\| &= \|\tfrac{1}{\alpha_k}(\mathscr{F}_{\gamma_k} - \mathrm{Id})(z) - \tfrac{1}{\alpha}(\mathscr{F}_\gamma - \mathrm{Id})(z)\| \\ &\leq \tfrac{|\gamma_k-\gamma|}{2\beta_V}\big(2\rho + \|\mathscr{F}_\gamma(0)\|\big) + \tfrac{1}{\alpha_k}\|\mathscr{F}_{\gamma_k}(z) - \mathscr{F}_\gamma(z)\|. \end{aligned} \tag{15}$$

Given $\gamma \in]0, 2\beta_V[$, define the two operators $\mathscr{F}_{1,\gamma} = \frac{1}{2}(\mathrm{Id} + \mathcal{R}_{\gamma R} \circ \mathcal{R}_V)$ and $\mathscr{F}_{2,\gamma} = \mathrm{Id} - \gamma\nabla G$. Then $\mathscr{F}_{1,\gamma}$ is firmly expansive (Lemma 2.2) and $\mathscr{F}_{2,\gamma}$ is $\frac{\gamma}{2\beta_V}$-averaged (Lemma 2.3). Now we have

$$\|\mathscr{F}_{\gamma_k}(z) - \mathscr{F}_\gamma(z)\| \leq \|\mathscr{F}_{2,\gamma_k}(z) - \mathscr{F}_{2,\gamma}(z)\| + \|\mathscr{F}_{1,\gamma_k}\mathscr{F}_{2,\gamma}(z) - \mathscr{F}_{1,\gamma}\mathscr{F}_{2,\gamma}(z)\|. \tag{16}$$

For the first term of (16),

$$\begin{aligned} \|\mathscr{F}_{2,\gamma_k}(z) - \mathscr{F}_{2,\gamma}(z)\| &= |\gamma_k - \gamma|\|\nabla G(z)\| \\ \text{(Triangle inequality and } \nabla G \text{ is } \beta_V^{-1}\text{-Lip.)} &\leq |\gamma_k - \gamma|(\beta_V^{-1}\rho + \|\nabla G(0)\|), \end{aligned} \tag{17}$$

where $\nabla G(0)$ is obviously bounded. Now for the second term of (16), denote $z^V = \mathrm{P}_V(z)$ and $z^{V^\perp} = z - z^V$, it can be derived that

$$v = \mathscr{F}_{1,\gamma}\mathscr{F}_{2,\gamma}(z) \Longleftrightarrow v = z^{V^\perp} + \operatorname{prox}_{\gamma R}(z^V - z^{V^\perp} - \gamma\nabla G(z^V)).$$

Denote $y = z^V - z^{V^\perp} - \gamma\nabla G(z^V)$. Then we have

$$\mathscr{F}_{1,\gamma_k}\mathscr{F}_{2,\gamma}(z) - \mathscr{F}_{1,\gamma}\mathscr{F}_{2,\gamma}(z) = \operatorname{prox}_{\gamma_k R}(y) - \operatorname{prox}_{\gamma R}(y).$$

Denote $w_k = \operatorname{prox}_{\gamma_k R}(y)$ and $w = \operatorname{prox}_{\gamma R}(y)$. Using the resolvent equation [31] and firm non-expansiveness of the proximity operator yields

$$\begin{aligned} \|w_k - w\| &= \|\operatorname{prox}_{\gamma_k R}(\tfrac{\gamma_k}{\gamma}y + (1 - \tfrac{\gamma_k}{\gamma})w) - \operatorname{prox}_{\gamma_k R}(y)\| \\ &\leq \|(1 - \tfrac{\gamma_k}{\gamma})(y - w)\| = \tfrac{|\gamma_k-\gamma|}{\gamma}\|y - w\| \\ &\leq \tfrac{|\gamma_k-\gamma|}{\gamma}\|(\mathrm{Id} - \operatorname{prox}_{\gamma R})y\| \leq \tfrac{|\gamma_k-\gamma|}{\gamma}(\|y\| + \|\operatorname{prox}_{\gamma R}(0)\|). \end{aligned} \tag{18}$$

Using the triangle inequality and non-expansiveness of $\beta_V\nabla G$, we obtain

$$\begin{aligned} \|y\| &\leq \|z^V - z^{V^\perp}\| + \gamma\|\nabla G(z^V)\| \leq \rho + \gamma\|\nabla G(z^V) - \nabla G(0)\| + \gamma\|\nabla G(0)\| \\ &\leq \rho + \gamma\beta_V^{-1}\|z\| + \gamma\|\nabla G(0)\| \leq \rho + \bar{\gamma}\beta_V^{-1}\rho + \bar{\gamma}\|\nabla G(0)\|. \end{aligned} \tag{19}$$

Define $\Delta_{k,\rho} := \sup_{\|z\|\leq\rho} \|\mathscr{R}_{\gamma_k}(z) - \mathscr{R}_{\gamma}(z)\|$. Then, putting together (15), (17), (18) and (19), we get that $\forall \rho \in [0, +\infty[$

$$\begin{aligned}\sum_{k\in\mathbb{N}}\lambda_k\alpha_k\Delta_{k,\rho} &= \sum_{k\in\mathbb{N}}\lambda_k\alpha_k \sup_{\|z\|\leq\rho} \|\mathscr{R}_{\gamma_k}(z) - \mathscr{R}_{\gamma}(z)\| \\ &\leq C\sum_{k\in\mathbb{N}} \lambda_k|\gamma_k - \gamma| < +\infty,\end{aligned}$$

where $C = \frac{2\rho+\|\mathscr{F}_\gamma(0)\|}{4\beta_V-\bar{\gamma}} + \frac{\rho}{\beta_V}(1+\frac{\beta_V}{\underline{\gamma}}+\frac{\bar{\gamma}}{\underline{\gamma}}) + (1+\frac{\bar{\gamma}}{\underline{\gamma}})\|\nabla G(0)\| + \frac{1}{\underline{\gamma}}\|\mathrm{prox}_{\gamma R}(0)\|$ is finite valued.

To this point, we verified that all the conditions of [8, Theorem 4] are met for the non-stationary FDR. Weak convergence of the sequence $\{z_k\}_{k\in\mathbb{N}}$ then follows. In turn, since P_V is linear, weak convergence of $\{x_k\}_{k\in\mathbb{N}}$ is also obtained.

For the sequence $\{u_k\}_{k\in\mathbb{N}}$, observe from the second equation in (6) that $u_{k+1} = (z_{k+1} - z_k)/\lambda_k + x_k$, hence $\|u_{k+1} - x_k\| \leq \|z_{k+1} - z_k\|/\lambda_k$. It follows from $\|z_{k+1} - z_k\| \to 0$ and the condition $\inf_{k\in\mathbb{N}} \lambda_k > 0$ that $u_{k+1} - x_k$ converges strongly to 0. We thus obtain weak convergence of u_k. If $\mathcal{H}$ is finite-dimensional, using (30) and the same argument as for inequality (18), we get $\|u_{k+1} - x^\star\| \leq \frac{|\gamma_k-\gamma|}{\underline{\gamma}}((2+\bar{\gamma}\beta_V)\|x_k - x^\star\| + \|z_k - z^\star\| + \|\mathrm{prox}_{\gamma R}(0)\|) \to 0$ which concludes the proof. □

4.2 Convergence Rate of the Bregman Divergence

In this part, we discuss the convergence rate of a specifically designed Bregman divergence associated to the objective value. As we have seen from the FDR iteration (5), there are three different points z_k and u_k, x_k generated along the iteration, which makes very difficult to establish a convergence rate on the objective value directly, unless the constraint subspace V is the whole space. For instance, in [18] the author obtained an $o(1/\sqrt{k})$ convergence rate on $(R(u_k) + G(x_k)) - (R(x^\star) + G(x^\star))$, which in general is *not* a non-negative quantity. Moreover, the functions R and G in the criterion are not evaluated at the same point. So the latter convergence rate is not only pessimistic (when specialized to $V = \mathcal{H}$ it gives a convergence rate as slow as subgradient descent), but is also of a limited interest given the lack of non-negativity. Our result in this part successfully avoids such drawbacks.

As in Theorem 4.1, let $z^\star \in \mathrm{fix}(\mathscr{F}_\gamma)$ and $x^\star := \mathrm{P}_V(z^\star) \in \mathrm{Argmin}(\Phi_V)$. Thus (**A.4**) and Fermat's rule allow to deduce that there exists a normal vector $v^\star \in V^\perp = \mathcal{N}_V(x^\star)$ such that $v^\star \in \nabla G(x^\star) + \partial R(x^\star)$. Now denote $\Phi := R + G$. Recalling Definition 2.1, for $y \in \mathbb{R}^n$, define the following Bregman divergence to the solution $x^\star$

$$\mathcal{D}_\Phi^{v^\star}(y) := \mathcal{D}_\Phi^{v^\star}(y, x^\star) = \Phi(y) - \Phi(x^\star) - \langle v^\star,\, y - x^\star\rangle = \Phi(y) - \Phi(x^\star) - \langle v^\star,\, y^{V^\perp}\rangle, \tag{20}$$

where $y^{V^\perp} := \mathrm{P}_{V^\perp}(y)$ is the projection of y onto $V^\perp$. In the last equality, we used the trivial fact that $\langle v^\star,\, x^\star\rangle = 0$.

The motivation of choosing the above function to quantify the convergence rate of FDR algorithm is due to the fact that it measures both the discrepancy of the objective to the optimal value and violation of the constraint on V.

Lemma 4.2 hereafter will provide us a key estimate on $\mathcal{D}_{\Phi}^{v^\star}(u_k)$ which will be used to derive the convergence rate of $\{\mathcal{D}_{\Phi}^{v^\star}(u_k)\}_{k\in\mathbb{N}}$. Denote $z_k^{V^\perp}:=\mathrm{P}_{V^\perp}(z_k)$ the projection of z_k onto $V^\perp$, $\phi_k:=\frac{1}{2\gamma_k}(\|z_k^{V^\perp}+\gamma_k v^\star\|^2+\|x_k-x^\star\|^2)$ and two auxiliary quantities $\xi_k:=\frac{|\gamma-\beta_V|}{2\underline{\gamma}\beta_V}\|z_k-z_{k-1}\|^2$, $\zeta_k:=\frac{|\gamma_k-\gamma_{k-1}|}{2\underline{\gamma}^2}\|z_k-x^\star\|^2$.

Lemma 4.2 *Considering the non-stationary FDR iteration in* (6)*. Suppose that Assumptions* (**A.1**)-(**A.5**) *hold with* $\lambda_k\equiv 1$*. Then,*

(i) *We have that* $\mathcal{D}_{\Phi}^{v^\star}(y)\geq 0$ *for every* y *in* $\mathcal{H}$*. Moreover, if* y *is a solution then* $\mathcal{D}_{\Phi}^{v^\star}(y)=0$ *(in particular,* $\mathcal{D}_{\Phi}^{v^\star}(x^\star)=0$*). On the other hand, if* y *is feasible (*$y\in V$*) and* $\mathcal{D}_{\Phi}^{v^\star}(y)=0$*, then* y *is solution.*

(ii) *For the sequence* $\{u_k\}_{k\in\mathbb{N}}$*, if* $v^\star$ *is bounded we have*

$$\mathcal{D}_{\Phi}^{v^\star}(u_{k+1})+\phi_{k+1}\leq\phi_k+\tfrac{\gamma_{k+1}-\gamma_k}{2}\|v^\star\|^2+\xi_{k+1}+\zeta_{k+1}<+\infty. \tag{21}$$

Remark 4.2 If we restrict $\gamma_k\in]0,\beta_V]$, then the term ξ_k in (21) can be discarded. If we assume $\{\gamma_k\}_{k\in\mathbb{N}}$ is monotonic, then the term ζ_k also disappears.

Proof The non-negativity of $\mathcal{D}_{\Phi}^{v^\star}(u_k)$ is rather obvious, as Φ is convex. Therefore, next we focus on the second claim. Define $y^{V^\perp}:=\mathrm{P}_{V^\perp}(y)$, $u_k^{V^\perp}:=\mathrm{P}_{V^\perp}(u_k)$, $z_k^{V^\perp}:=\mathrm{P}_{V^\perp}(z_k)$ the projections of y,u_k,z_k onto $V^\perp$ respectively.

The update of u_k in (6) and definition of proximity operator imply that

$$(2x_k-z_k-u_{k+1})/\gamma_k-\nabla G(x_k)\in\partial R(u_{k+1}).$$

For the convexity of R, we obtain that, for every $y\in\mathcal{H}$,

$$\begin{aligned}R(y)&\geq R(u_{k+1})+\langle(2x_k-z_k-u_{k+1})/\gamma_k-\nabla G(x_k),\,y-u_{k+1}\rangle\\&=R(u_{k+1})+\tfrac{1}{\gamma_k}\langle 2x_k-z_k-u_{k+1},\,y-u_{k+1}\rangle-\langle\nabla G(x_k),\,y-u_{k+1}\rangle.\end{aligned} \tag{22}$$

Notice that $u_{k+1}=x_k+z_{k+1}-z_k$. Then, the first inner product of the last line of (22) can be re-written as

$$\begin{aligned}&\langle 2x_k-z_k-u_{k+1},\,y-u_{k+1}\rangle\\&=\langle x_k-z_k,\,y-x_k\rangle+\langle y-z_k,\,z_k-z_{k+1}\rangle+\|z_{k+1}-z_k\|^2\\&=-\langle z_k^{V^\perp},\,y\rangle+\tfrac{1}{2}\big(\|z_{k+1}-z_k\|^2+\|z_{k+1}-y\|^2-\|z_k-y\|^2\big),\end{aligned} \tag{23}$$

where $2\langle c_2-c_1,\,c_1-c_3\rangle=\|c_2-c_3\|^2-\|c_1-c_2\|^2-\|c_1-c_3\|^2$ is applied to $\langle y-z_k,\,z_k-z_{k+1}\rangle$. Combining (23) with (22),

$$\begin{aligned}R(u_{k+1})-R(y)\leq\langle\nabla G(x_k),\,y-u_{k+1}\rangle+\tfrac{1}{\gamma_k}\langle z_k^{V^\perp},\,y\rangle\\+\tfrac{1}{2\gamma_k}\big(\|z_k-y\|^2-\|z_{k+1}-y\|^2-\|z_{k+1}-z_k\|^2\big).\end{aligned} \tag{24}$$

Since G is convex, given any x_k and $y \in \mathcal{H}$, we have

$$G(x_k) - G(y) \leq \langle \nabla G(x_k),\, x_k - y \rangle. \tag{25}$$

Recall $\Phi = R + G$. Summing up (24) and (25) and rearranging the terms, then

$$\begin{aligned}
&\big(R(u_{k+1}) + G(x_k)\big) - \Phi(y) + \tfrac{1}{2\gamma_k}\big(\|z_{k+1} - y\|^2 - \|z_k - y\|^2\big) - \tfrac{1}{\gamma_k}\langle z_k^{V^\perp},\, y\rangle \\
&\leq -\tfrac{1}{2\gamma_k}\|z_{k+1} - z_k\|^2 + \langle \nabla G(x_k),\, x_k - u_{k+1}\rangle.
\end{aligned}$$

Since G has Lipschitz continuous gradient, applying Lemma 2.1 yields

$$G(u_{k+1}) - G(x_k) \leq \langle \nabla G(x_k),\, u_{k+1} - x_k\rangle + \tfrac{1}{2\beta_V}\|u_{k+1} - x_k\|^2.$$

Sum up the above two inequalities and recall $\xi_{k+1} := \frac{|\gamma - \beta_V|}{2\underline{\gamma}\beta_V}\|z_{k+1} - z_k\|^2$, then

$$\begin{aligned}
&\Phi(u_{k+1}) - \Phi(y) + \tfrac{1}{2\gamma_k}\big(\|z_{k+1} - y\|^2 - \|z_k - y\|^2\big) - \tfrac{1}{\gamma_k}\langle z_k^{V^\perp},\, y\rangle \\
&\leq -\tfrac{1}{2\gamma_k}\|z_{k+1} - z_k\|^2 + \tfrac{1}{2\beta_V}\|u_{k+1} - x_k\|^2 \\
&= \tfrac{\gamma_k - \beta_V}{2\gamma_k\beta_V}\|z_{k+1} - z_k\|^2 \leq \tfrac{|\gamma_k - \beta_V|}{2\gamma_k\beta_V}\|z_{k+1} - z_k\|^2 \leq \xi_{k+1}.
\end{aligned} \tag{26}$$

Note that we applied again the equivalence $u_{k+1} = x_k + z_{k+1} - z_k$. Furthermore, define $\zeta_{k+1}^y := \frac{|\gamma_{k+1} - \gamma_k|}{2\underline{\gamma}^2}\|z_{k+1} - y\|^2$. Then, from (26), we have

$$\begin{aligned}
&\Phi(u_{k+1}) + \tfrac{1}{2\gamma_{k+1}}\|z_{k+1} - y\|^2 \\
&= \Phi(u_{k+1}) + \tfrac{1}{2\gamma_k}\|z_{k+1} - y\|^2 + \big(\tfrac{1}{2\gamma_{k+1}} - \tfrac{1}{2\gamma_k}\big)\|z_{k+1} - y\|^2 \\
&\leq \Phi(y) + \tfrac{1}{\gamma_k}\langle z_k^{V^\perp},\, y^{V^\perp}\rangle + \tfrac{1}{2\gamma_k}\|z_k - y\|^2 + \xi_{k+1} + \zeta_{k+1}^y.
\end{aligned} \tag{27}$$

Recall that $x_k \in V$. Hence, $\mathrm{P}_{V^\perp}(x_k) = 0$. Then, using (27), we have the following estimate for the Bregman divergence (defined in (20)):

$$\begin{aligned}
&\mathcal{D}_\Phi^{v^\star}(u_{k+1}) - \mathcal{D}_\Phi^{v^\star}(y) = \Phi(u_{k+1}) - \Phi(x^\star) - \langle v^\star,\, u_{k+1}^{V^\perp} - y^{V^\perp}\rangle \\
&\leq \tfrac{1}{\gamma_k}\langle z_k^{V^\perp},\, y^{V^\perp}\rangle - \langle v^\star,\, u_{k+1}^{V^\perp} - y^{V^\perp}\rangle + \tfrac{1}{2\gamma_k}\|z_k - y\|^2 - \tfrac{1}{2\gamma_{k+1}}\|z_{k+1} - y\|^2 + \xi_{k+1} + \zeta_{k+1}^y \\
&= \tfrac{1}{2\gamma_k}\big(\|y^{V^\perp} + \gamma_k v^\star\|^2 - 2\|\gamma_k v^\star\|^2 + \|z_k^{V^\perp} + \gamma_k v^\star\|^2 + \|z_k^V - y^V\|^2\big) + \xi_{k+1} + \zeta_{k+1}^y \\
&\quad + \tfrac{1}{2\gamma_{k+1}}\big(-\|z_{k+1}^{V^\perp} + \gamma_{k+1}v^\star\|^2 + \|z_{k+1}^{V^\perp}\|^2 + \|\gamma_{k+1}v^\star\|^2 - \|z_{k+1}^{V^\perp} - y^{V^\perp}\|^2 - \|z_{k+1}^V - y^V\|^2\big),
\end{aligned}$$

where $y^V := \mathrm{P}_V(y)$, $z_k^V := \mathrm{P}_V(z_k)$ are the projections of y, z_k onto V respectively. From the above inequality, we deduce the following result

$$\begin{aligned}
&\mathcal{D}_\Phi^{v^\star}(u_{k+1}) - \mathcal{D}_\Phi^{v^\star}(y) + \phi_{k+1} - \phi_k - (\xi_{k+1} + \zeta_{k+1}^y) \\
&\leq \tfrac{1}{2\gamma_k}\big(\|y^{V^\perp} - \gamma_k v^\star\|^2 - 2\|\gamma_k v^\star\|^2\big) + \tfrac{1}{2\gamma_{k+1}}\big(\|z_{k+1}^{V^\perp}\|^2 + \|\gamma_{k+1}v^\star\|^2 - \|z_{k+1}^{V^\perp} - y^{V^\perp}\|^2\big) \\
&= \tfrac{1}{2\gamma_k}\big(\|y^{V^\perp}\|^2 - 2\gamma_k\langle y^{V^\perp},\, v^\star\rangle - \|\gamma_k v^\star\|^2\big) + \tfrac{1}{2\gamma_{k+1}}\big(\|\gamma_{k+1}v^\star\|^2 - \|y^{V^\perp}\|^2 + 2\langle z_{k+1}^{V^\perp},\, y^{V^\perp}\rangle\big) \\
&= \tfrac{\gamma_{k+1} - \gamma_k}{2\gamma_k\gamma_{k+1}}\|y^{V^\perp}\|^2 + \tfrac{\gamma_{k+1} - \gamma_k}{2}\|v^\star\|^2 + \tfrac{1}{\gamma_{k+1}}\langle z_{k+1}^{V^\perp} - \gamma_{k+1}v^\star,\, y^{V^\perp}\rangle.
\end{aligned}$$

In particular, taking $y = x^\star \in V$ in the last inequality and using the fact that $\mathrm{P}_{V^\perp}(x^\star) = 0$, we obtain the desired result. □

With the above property of $\mathcal{D}_\Phi^{v^\star}(u_k)$, we are able to present the main result on the convergence rate of the Bregman divergence.

Theorem 4.2 *Consider the non-stationary FDR iteration* (6)*. Suppose that Assumptions* (A.1)-(A.5) *hold with* $\lambda_k \equiv 1$*. If moreover* $v^\star$ *is bounded, then for any* $k \geq 0$*,*

$$\inf_{0\leq i\leq k} \mathcal{D}_\Phi^{v^\star}(u_i) = o\big(\tfrac{1}{k+1}\big) \;\; \text{and} \;\; D(\bar{u}_k) = O\big(\tfrac{1}{k+1}\big), \;\; \text{where } \bar{u}_k = \frac{1}{k+1}\sum_{i=0}^{k} u_i.$$

Remark 4.3

- A typical situation that ensures the boundedness of $v^\star$ is when $\partial R(x^\star)$ is bounded. Such requirement can be removed if we choose more carefully the element $v^\star$. For instance, one can easily show from Theorem 4.1 that the subgradient $v_k{:=}(x_k - z_k)/\gamma_k = -\mathrm{P}_{V^\perp}(z_k)/\gamma_k$ converges weakly to $v^\star{:=}(x^\star - z^\star)/\gamma \in V^\perp \cap (\nabla G(x^\star) + \partial R(x^\star))$.
- The main difficulty in establishing the convergence rate directly on $\mathcal{D}_\Phi^{v^\star}(u_k)$ (rather that on the best iterate) is that, for $V \subsetneq \mathcal{H}$, we have no theoretical guarantee that $\mathcal{D}_\Phi^{v^\star}(u_k)$ is decreasing, *i.e.* no descent property on $\mathcal{D}_\Phi^{v^\star}(u_k)$.

Proof Define $\theta_k{:=}\min_{0\leq i\leq k} \mathcal{D}_\Phi^{v^\star}(u_i) \leq \mathcal{D}_\Phi^{v^\star}(u_k)$. Summing inequality (21) up to some $k \in \mathbb{N}$ yields

$$(k+1)\theta_k \leq \sum_{i=0}^{k}\mathcal{D}_\Phi^{v^\star}(u_i) \leq \phi_0 + \tfrac{\gamma_\infty - \gamma_0}{2}\|v^\star\|^2 + \sum_{k\in\mathbb{N}}\xi_k + \sum_{k\in\mathbb{N}}\zeta_k.$$

Since $v^\star$ is bounded, so is ϕ_0. Then, owing to Theorem 4.1, we have

$$\sum_{k\in\mathbb{N}}\xi_k = \tfrac{|\gamma-\beta_V|}{2\underline{\gamma}\beta_V}\sum_{k\in\mathbb{N}}\|z_k - z_{k-1}\|^2 < +\infty.$$

Lastly, as $\{z_k\}_{k\in\mathbb{N}}$ is bounded, so is $\{\|z_k - x^\star\|\}_{k\in\mathbb{N}}$. Recall that, by assumptions, $\{\gamma_k\}_{k\in\mathbb{N}}$ converges to some $\gamma \in]0, 2\beta_V[$ with $\{|\gamma_k - \gamma|\}_{k\in\mathbb{N}}$ being summable. Then

$$\begin{aligned}\sum_{k\in\mathbb{N}}\zeta_k &\leq \tfrac{1}{2\underline{\gamma}^2}\sup_{k\in\mathbb{N}}\|z_k - x^\star\|^2 \sum_{k\in\mathbb{N}}|\gamma_{k+1} - \gamma_k| \\ &\leq \tfrac{1}{2\underline{\gamma}^2}\sup_{k\in\mathbb{N}}\|z_k - x^\star\|^2 \sum_{k\in\mathbb{N}}(|\gamma_{k+1} - \gamma| + |\gamma_k - \gamma|) < +\infty.\end{aligned}$$

Summing up the above results, we have that $(k+1)\theta_k \leq C < +\infty$ holds for all $k \in \mathbb{N}$, which means $\theta_k = O(1/(k+1))$. Now, owing to the definition of θ_k,

$$\sum_{k\in\mathbb{N}}\theta_k \leq \sum_{k\in\mathbb{N}}\mathcal{D}_\Phi^{v^\star}(u_k) \leq \phi_0 + \tfrac{\gamma_\infty - \gamma_0}{2}\|v^\star\|^2 + \sum_{k\in\mathbb{N}}(\xi_k + \zeta_k) < +\infty.$$

Moreover, it is immediate that, for every $k \geq 1$,

$$\theta_k = \min(\mathcal{D}_\Phi^{v^\star}(u_k), \theta_{k-1}) \leq \theta_{k-1},$$

that is, the sequence $\{\theta_k\}_{k\in\mathbb{N}}$ is non-increasing. Invoking Lemma 2.5 on $\{\theta_k\}_{k\in\mathbb{N}}$ concludes the proof.

For the ergodic rate, we start again from (21) and apply Jensen's inequality to $\mathcal{D}_\Phi^{v^\star}$ which is a convex function, and get

$$(k+1)\mathcal{D}_\Phi^{v^\star}(\bar{u}_k) \leq \sum_{i=0}^{k}\mathcal{D}_\Phi^{v^\star}(u_i),$$

where the right-hand side is bounded by arguing as above. □

4.3 Application to Forward–Backward Splitting

Assume now that $V = \mathcal{H}$, in which case problem (4) simplifies to

$$\min_{x\in\mathcal{H}} \big\{\Phi(x) := F(x) + R(x)\big\}.$$

In this case, the FDR iteration (6) is nothing but the FB splitting scheme (2). The non-relaxed and non-stationary version of it reads as

$$x_{k+1} = \mathrm{prox}_{\gamma_k R}\big(x_k - \gamma_k \nabla F(x_k)\big). \tag{28}$$

We get $\mathcal{D}_\Phi^{v^\star}(y) = \Phi(y) - \Phi(x^\star)$ by specializing the Bregman divergence (20) to Φ, which is simply the objective value error. We have the following result.

Corollary 4.1 *Consider the Forward–Backward iteration* (28)*. Suppose that conditions* (A.1)-(A.5) *hold with* $V = \mathcal{H}$ *and* $\lambda_k \equiv 1$*. Then*

$$\Phi(x_k) - \Phi(x^\star) = o(1/k).$$

Remark 4.4

- The $o(1/k)$ convergence rate for the large choice $\gamma_k \in]0, 2\beta[$ appears to be new for the FB splitting algorithm. The rate $O(1/k)$ is known in the literature for several choices of the step-size; see e.g., [12, Theorem 3.1] for $\gamma_k \in]0, \beta]$ or with backtracking, and [11, Proposition 2] for $\gamma_k \in]0, 2\beta[$.
- For the global convergence of the sequence $\{x_k\}_{k\in\mathbb{N}}$ generated by the non-stationary FB iteration, neither convergence of γ_k to γ nor summability of $\{|\gamma_k - \gamma|\}_{k\in\mathbb{N}}$ is required. See [32, Theorem 3.4].

Proof First, weak convergence of the non-stationary FB iteration follows from Theorem 4.1. On the one hand, specializing (21) to the case of FB, we get

$$\begin{aligned}\Phi(x_{k+1}) - \Phi(x^\star) \leq \tfrac{1}{2\gamma_k}\|x_k - x^\star\|^2 - \tfrac{1}{2\gamma_{k+1}}\|x_{k+1} - x^\star\|^2 \\ + \tfrac{|\gamma-\beta|}{2\underline{\gamma}\beta}\|x_k - x_{k-1}\|^2 + \tfrac{|\gamma_{k+1}-\gamma_k|}{2\underline{\gamma}^2}\|x_{k+1} - x^\star\|^2,\end{aligned} \tag{29}$$

which means that

$$\begin{aligned}\textstyle\sum_{k\in\mathbb{N}}(\Phi(x_k) - \Phi(x^\star)) \leq \tfrac{1}{2\gamma_0}\|x_0 - x^\star\|^2 + \tfrac{|\gamma-\beta|}{2\underline{\gamma}\beta}\sum_{k\in\mathbb{N}}\|x_k - x_{k-1}\|^2 \\ + \tfrac{1}{\underline{\gamma}^2}\sup_{k\in\mathbb{N}}\|x_k - x^\star\|^2 \textstyle\sum_{k\in\mathbb{N}}|\gamma_k - \gamma| < +\infty.\end{aligned}$$

On the other hand, owing to inequality (26) in the proof of Lemma 4.2, $\forall y \in \mathcal{H}$,

$$\Phi(x_{k+1}) + \tfrac{1}{2\gamma_k}\|x_{k+1} - y\|^2 \leq \Phi(y) + \tfrac{1}{2\gamma_k}\|x_k - y\|^2 + \big(\tfrac{1}{2\beta} - \tfrac{1}{2\gamma_k}\big)\|x_{k+1} - x_k\|^2.$$

Choosing $y = x_k$, we obtain

$$\begin{aligned}\big(\Phi(x_{k+1}) - \Phi(x^\star)\big) - \big(\Phi(x_k) - \Phi(x^\star)\big) &\leq \big(\tfrac{1}{2\beta} - \tfrac{1}{\gamma_k}\big)\|x_{k+1} - x_k\|^2 \\ &\leq -\delta\|x_{k+1} - x_k\|^2,\end{aligned}$$

where $\delta = \frac{1}{\overline{\gamma}} - \frac{1}{2\beta} > 0$ since $\overline{\gamma} < 2\beta$. This implies that the sequence $\{\Phi(x_k) - \Phi(x^\star)\}_{k\in\mathbb{N}}$ is positive and non-increasing. Summing up both sides of the above inequality and applying Lemma 2.5 leads to the claimed result. □

5 Local Linear Convergence

From now on, we turn to the local convergence analysis of FDR. Given that partial smoothness is so far available only in finite dimension, in this section, we consider a finite-dimensional setting, *i.e.* $\mathcal{H} = \mathbb{R}^n$. In the sequel, we denote $z^\star \in \mathrm{fix}(\mathscr{F}_\gamma)$ a fixed point of iteration (6) and $x^\star = \mathrm{P}_V(z^\star) \in \mathrm{Argmin}(\Phi_V)$ a global minimizer of problem (4). For simplicity, we also fix $\lambda_k \equiv 1$.

5.1 Finite Activity Identification

We start with the finite activity identification, which means that in a finite number of iterations the iterates identify the manifold in which the solution $x^\star$ lives. Under the condition of Theorem 4.1, we know that $\gamma_k \to \gamma$, $z_k \to z^\star$ and $u_k, x_k \to x^\star$. Moreover, we have the following optimality conditions

$$(x^\star - z^\star)/\gamma \in \nabla G(x^\star) + \partial R(x^\star) \;\text{ and }\; (z^\star - x^\star)/\gamma \in V^\perp,\; x^\star \in V. \tag{30}$$

The condition needed for identification result is built upon these monotone inclusions. Since x_k is the projection of z_k onto V, we have $x_k \in V$ for all $k \geq 0$. Therefore, we only need to discuss the identification property of u_k.

Theorem 5.1 *For the non-stationary FDR* (6)*. Suppose that Assumptions* (**A.1**)-(**A.5**) *hold, so that* $(u_k, x_k, z_k) \to (x^\star, x^\star, z^\star)$ *where* $z^\star \in \mathrm{fix}(\mathscr{F}_\gamma)$ *and* $x^\star = \mathrm{P}_V(z^\star) \in \mathrm{Argmin}(\Phi_V)$*. Moreover, suppose that* $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}^R_{x^\star})$ *and that the following non-degeneracy condition holds*

$$(x^\star - z^\star)/\gamma - \nabla G(x^\star) \in \mathrm{ri}\big(\partial R(x^\star)\big). \tag{31}$$

Then,

(i) *there exists* $K \in \mathbb{N}$ *such that, for all* $k \geq K$*, we have* $u_k \in \mathcal{M}^R_{x^\star}$*.*

(ii) *Moreover, for every* $k \geq K$*,*

(a) *if* $\mathcal{M}^R_{x^\star} = x^\star + T^R_{x^\star}$*, then* $T^R_{u_k} = T^R_{x^\star}$*.*

(b) *If* R *is locally polyhedral around* $x^\star$*, then* $x_k \in \mathcal{M}^R_{x^\star} = x^\star + T^R_{x^\star}$*,* $T^R_{u_k} = T^R_{x^\star}$*,* $\nabla_{\mathcal{M}^R_{x^\star}} R(u_k) = \nabla_{\mathcal{M}^R_{x^\star}} R(x^\star)$*, and* $\nabla^2_{\mathcal{M}^R_{x^\star}} R(u_k) = 0$*.*

Remark 5.1 As we mentioned before, for global convergence, approximation errors can be allowed, *i.e.* $\mathrm{prox}_{\gamma R}$ and ∇G can be computed approximately. However, for the finite activity, we have no identification guarantees for (u_k, x_k) if such an approximation is allowed. For example, if we have $x_k = \mathrm{P}_V(z_k) + \varepsilon_k$ where $\varepsilon_k \in \mathbb{R}^n$ is the error of approximating $\mathrm{P}_V(z_k)$. Then, unless $\varepsilon_k \in V$, we can no longer guarantee that $x_k \in V$.

Proof From the update of u_{k+1} and the definition of proximity operator, we have $(2x_k - z_k - u_{k+1})/\gamma_k - \nabla G(x_k) \in \partial R(u_{k+1})$. At convergence, we have $(x^\star - z^\star)/\gamma - \nabla G(x^\star) \in \partial R(x^\star)$. Therefore, one can show that

$$\begin{aligned}&\mathrm{dist}\big((x^\star - z^\star)/\gamma - \nabla G(x^\star), \partial R(u_{k+1})\big)\\ &\leq \tfrac{1}{\underline{\gamma}}(2\|x_k - x^\star\| + \|u_{k+1} - x^\star\| + \|z_k - z^\star\|) + \tfrac{|\gamma_k - \gamma|}{\underline{\gamma}^2}\|\mathrm{P}_{V^\perp}(z^\star)\| + \tfrac{1}{\beta_V}\|x_k - x^\star\|.\end{aligned}$$

Theorem 4.1 allows to infer that the right hand side of the inequality converges to 0. In addition, since $R \in \Gamma_0(\mathbb{R}^n)$, R is sub-differentially continuous at every point in its domain [33, Example 13.30], and in particular at $x^\star$. It then follows that $R(u_k) \to R(x^\star)$. Altogether, this shows that the conditions of [34, Theorem 5.3] are fulfilled for R: 1) convergence of sequence; 2) distance $\mathrm{dist}((x^\star - z^\star)/\gamma - \nabla G(x^\star), \partial R(u_{k+1})) \to 0$; 3) convergence of objective function value. The finite identification claim follows.

(a) In this case, $\mathcal{M}^R_{x^\star}$ is an affine subspace, *i.e.* $\mathcal{M}^R_{x^\star} = x^\star + T^R_{x^\star}$. Since R is partly smooth at $x^\star$ relative to $\mathcal{M}^R_{x^\star}$, the sharpness property holds at all nearby points in $\mathcal{M}^R_{x^\star}$ [29, Proposition 2.10]. Thus for k large enough, *i.e.* u_k sufficiently close to $x^\star$ on $\mathcal{M}^R_{x^\star}$, we have $\mathcal{T}_{u_k}(\mathcal{M}^R_{x^\star}) = T^R_{x^\star} = T^R_{u_k}$.

(b) It is immediate to verify that a locally polyhedral function around $x^\star$ is indeed partly smooth relative to the affine subspace $x^\star + T^R_{x^\star}$. Thus, the first claim follows from (ii)(a). For the rest, it is sufficient to observe that by polyhedrality, for any $x \in \mathcal{M}^R_{x^\star}$ near $x^\star$, $\partial R(x) = \partial R(x^\star)$. Therefore, combining local normal sharpness [29, Proposition 2.10] and [15, Lemma 4.3] yields the second conclusion. □

A Bound on the Number of Iterations to Identification In Theorem 5.1, we only assert the existence of some $K \geq 0$ beyond which finite identification occurs. There are situations where a bound of K can be established.

Proposition 5.1 *Suppose that the assumptions of Theorem 5.1 hold. If the iterates are such that* $\partial R(u_k) \subset \mathrm{rbd}(\partial R(x^\star))$ *whenever* $u_k \notin \mathcal{M}_{x^\star}$*, then we have* $u_k \in \mathcal{M}_{x^\star}$ *for some* k *obeying* $k \geq \frac{\|z_0 - z^\star\|^2 + O(\sum_{k\in\mathbb{N}} |\gamma_k - \gamma|)}{\underline{\gamma}^2 \mathrm{dist}(-\nabla G(x^\star), V^\perp + \mathrm{rbd}(\partial R(x^\star)))^2}$.

Remark 5.2 When $V = \mathbb{R}^n$, we recover the result of [15, Proposition 3.6(i)] established for the Forward–Backward splitting method. For $F = 0$, our result also encompasses that of Douglas–Rachford splitting [16, Proposition 5.1].

Proof Recall from the proof of Theorem 4.1 that $\mathscr{F}_\gamma := \mathscr{F}_{1,\gamma} \circ \mathscr{F}_{2,\gamma}$, where $\mathscr{F}_{1,\gamma} = \frac{1}{2}(\mathrm{Id} + \mathcal{R}_{\gamma R} \circ \mathcal{R}_V)$ and $\mathscr{F}_{2,\gamma} = (\mathrm{Id} - \gamma \nabla G)$. From (14), we have $z_{k+1} = \mathscr{F}_\gamma(z_k) + e_k$ where $\{\|e_k\|\}_{k\in\mathbb{N}} = \{|\gamma_k - \gamma|\}_{k\in\mathbb{N}}$ is a summable sequence. Thus arguing as in [35, Theorem 3.1], and using firm non-expansiveness of $\mathscr{F}_{1,\gamma}$ (Lemma 2.2) and non-expansiveness of $\mathscr{F}_{2,\gamma}$ (Lemma 2.3), we get

$$\begin{aligned} \|z_k - z^\star\|^2 &= \|\mathscr{F}_{\gamma_k}(z_{k-1}) - \mathscr{F}_{\gamma_k}(z^\star)\|^2 \leq \|\mathscr{F}_\gamma(z_{k-1}) - \mathscr{F}_\gamma(z^\star)\|^2 + O(\|e_{k-1}\|) \\ &= \|z_{k-1} - z^\star\|^2 - \|g_k + v_{k-1} + \gamma \nabla G(x^\star)\|^2 + O(\|e_{k-1}\|), \end{aligned} \tag{32}$$

where $g_k := 2x_{k-1} - z_{k-1} - u_k - \gamma \nabla G(x_{k-1})$ which verifies $g_k \in \gamma \partial R(u_k)$ and $v_{k-1} := z_{k-1} - x_{k-1} \in V^\perp$. Assume that identification has not occurred yet, *i.e.* $u_k \notin \mathcal{M}_{x^\star}$ which implies $g_k + v_{k-1} \in V^\perp + \partial R(u_k) \subset V^\perp + \mathrm{rbd}(\partial R(x^\star))$. Thus, continuing (32), we get

$$\begin{aligned} &\|z_k - z^\star\|^2 \\ &\leq \|z_{k-1} - z^\star\|^2 - \gamma^2 \mathrm{dist}(-\nabla G(x^\star), V^\perp + \mathrm{rbd}(\partial R(x^\star)))^2 + O(|\gamma_{k-1} - \gamma|) \\ &\leq \|z_0 - z^\star\|^2 - \underline{\gamma}^2 k \mathrm{dist}(-\nabla G(x^\star), V^\perp + \mathrm{rbd}(\partial R(x^\star)))^2 + O(\textstyle\sum_k |\gamma_{k-1} - \gamma|). \end{aligned}$$

Note $\operatorname{dist}(-\nabla G(x^\star), V^\perp + \operatorname{rbd}(\partial R(x^\star))) > 0$ since $-\nabla G(x^\star) \in \operatorname{ri}(V^\perp + \partial R(x^\star))$ by (31). Taking k as the largest integer such that the right hand is positive, we deduce that the number of iterations where identification has not occurred does not exceed the claimed bound. Thus finite identification necessarily occurs at some k larger than this bound. □

5.2 Locally Linearized Iteration

With the finite identification result, in the next we show that the globally nonlinear fixed-point iteration (13) can be locally linearized along the identified manifold $\mathcal{M}^R_{x^\star}$. Define the function $\overline{R}(u) := \gamma R(u) - \langle u,\, x^\star - z^\star - \gamma \nabla G(x^\star)\rangle$. We have the following key property of $\overline{R}$.

Lemma 5.1 *Let $x^\star \in \operatorname{Argmin}(\Phi_V)$, and suppose that $R \in \operatorname{PSF}_{x^\star}(\mathcal{M}^R_{x^\star})$. Then the Riemannian Hessian of $\overline{R}$ at $x^\star$ reads as*

$$H_{\overline{R}} := \mathrm{P}_{T^R_{x^\star}} \nabla^2_{\mathcal{M}^R_{x^\star}} \overline{R}(x^\star) \mathrm{P}_{T^R_{x^\star}}, \tag{33}$$

which is symmetric positive semi-definite under either of the two conditions:
(i) *condition* (31) *holds.*
(ii) *$\mathcal{M}^R_{x^\star}$ is an affine subspace.*
In turn, the matrix $W_{\overline{R}} := (\mathrm{Id} + H_{\overline{R}})^{-1}$ is firmly non-expansive.

Proof See [15, Lemma 4.3] and [1, Corollary 4.3(ii)]. □

From now on, we assume that F (hence G) is locally C^2-smooth around $x^\star$. Define $H_G := \mathrm{P}_V \nabla^2 F(x^\star) \mathrm{P}_V$, $M_{\overline{R}} := \mathrm{P}_{T^R_{x^\star}} W_{\overline{R}} \mathrm{P}_{T^R_{x^\star}}$ and $\mathcal{R}_{M_{\overline{R}}} := 2M_{\overline{R}} - \mathrm{Id}$ and

$$\mathscr{M}_\gamma = \mathrm{Id} + 2M_{\overline{R}}\mathrm{P}_V - M_{\overline{R}} - \mathrm{P}_V - \gamma M_{\overline{R}} H_G = \tfrac{1}{2}\big(\mathcal{R}_{M_{\overline{R}}}\mathcal{R}_V + \mathrm{Id}\big)(\mathrm{Id} - \gamma H_G),$$

and $\mathscr{M}_{\gamma,\lambda} = (1-\lambda)\mathrm{Id} + \lambda \mathscr{M}_\gamma$. We have the following theorem for the linearized fixed-point formulation of (6).

Theorem 5.2 *Consider the non-stationary FDR iteration* (6) *and suppose that* (**A.1**)-(**A.5**) *hold. If moreover, $\lambda_k \to \lambda \in]0, \frac{4\beta_V - \gamma}{2\beta_V}[$ and F is locally C^2 around $x^\star$, then for all k large enough we have*

$$z_{k+1} - z^\star = \mathscr{M}_{\gamma,\lambda}(z_k - z^\star) + \psi_k + \chi_k, \tag{34}$$

where $\psi_k := o(\|z_k - z^\star\|)$ and $\chi_k := O(\lambda_k|\gamma_k - \gamma|)$. Both ψ_k and χ_k vanish when R is locally polyhedral around $x^\star$, F is quadratic and $(\gamma_k, \lambda_k) \in]0, 2\beta_V[\times]0, \frac{4\beta_V - \gamma}{2\beta_V}[$ are chosen constants.

Proof From (6), since V is a subspace, then we have

$$x_k = \mathrm{P}_V(z_k),\ x^\star = \mathrm{P}_V(z^\star) \iff z_k - x_k \in \mathcal{N}_V(x_k),\ z^\star - x^\star \in \mathcal{N}_V(x^\star).$$

Projecting onto V leads to $x_k - x^\star = \mathrm{P}_V(z_k - z^\star)$. Under the assumptions of Theorem 5.1, there exists $K \in \mathbb{N}$ large enough such that for all $k \geq K$,

$u_k \in \mathcal{M}^R_{x^\star}$. Denote $T^R_{u_k}$ and $T^R_{x^\star}$ the tangent spaces corresponding to u_k and $x^\star \in \mathcal{M}^R_{x^\star}$. Denote $\tau_k^R : T^R_{u_k} \to T^R_{x^\star}$ the parallel translation along the unique geodesic on $\mathcal{M}^R_{x^\star}$ joining u_k to $x^\star$. Owing to [19, Lemma 5.1], we have for u_k after identification that $u_k - x^\star = \mathrm{P}_{T^R_{x^\star}}(u_k - x^\star) + o(\|u_k - x^\star\|)$. The update of u_{k+1} in (6) and its convergence are respectively equivalent to

$$\begin{aligned} 2x_k - z_k - u_{k+1} - \gamma_k \nabla G(x_k) &\in \gamma_k \partial R(u_{k+1}) \\ 2x^\star - z^\star - x^\star - \gamma \nabla G(x^\star) &\in \gamma \partial R(x^\star). \end{aligned}$$

Upon projecting onto the corresponding tangent spaces and applying the parallel translation τ_{k+1} from u_{k+1} to $x^\star$, we get

$$\begin{aligned} \gamma_k \tau_{k+1} \nabla_{\mathcal{M}^R_{x^\star}} R(u_{k+1}) &= \mathrm{P}_{T^R_{x^\star}}\big(2x_k - z_k - u_{k+1} - \gamma_k \nabla G(x_k)\big) \\ &\quad + \big(\tau_{k+1}\mathrm{P}_{T^R_{u_{k+1}}} - \mathrm{P}_{T^R_{x^\star}}\big)\big(2x_k - z_k - u_{k+1} - \gamma_k \nabla G(x_k)\big), \\ \gamma \nabla_{\mathcal{M}^R_{x^\star}} R(x^\star) &= \mathrm{P}_{T^R_{x^\star}}\big(2x^\star - z^\star - x^\star - \gamma \nabla G(x^\star)\big). \end{aligned}$$

Subtracting both equations, we obtain

$$\begin{aligned} &\gamma_k \tau_{k+1} \nabla_{\mathcal{M}^R_{x^\star}} R(u_{k+1}) - \gamma \nabla_{\mathcal{M}^R_{x^\star}} R(x^\star) \\ &= \mathrm{P}_{T^R_{x^\star}}\big((2x_k - z_k - u_{k+1} - \gamma_k \nabla G(x_k)) - (2x^\star - z^\star - x^\star - \gamma \nabla G(x^\star))\big) \\ &\quad + \textbf{Term 1} + \textbf{Term 2}, \end{aligned} \tag{35}$$

where we have $\textbf{Term 1} = (\tau_{k+1}\mathrm{P}_{T^R_{u_{k+1}}} - \mathrm{P}_{T^R_{x^\star}})(x^\star - z^\star - \gamma \nabla G(x^\star))$ and $\textbf{Term 2} = (\tau_{k+1}\mathrm{P}_{T^R_{u_{k+1}}} - \mathrm{P}_{T^R_{x^\star}})((2x_k - z_k - u_{k+1} - \gamma_k \nabla G(x_k)) - (2x^\star - z^\star - x^\star - \gamma \nabla G(x^\star)))$. For the term $(\gamma_k - \gamma)\tau_{k+1}\nabla_{\mathcal{M}^R_{x^\star}} R(u_{k+1})$, since the Riemannian gradient $\nabla_{\mathcal{M}^R_{x^\star}} R(u_{k+1})$ is bounded on a bounded set, we have $(\gamma_k - \gamma)\tau_{k+1}\nabla_{\mathcal{M}^R_{x^\star}} R(u_{k+1}) = O(|\gamma_k - \gamma|)$. For **Term 2**, owing to [15, Lemma B.1] and the boundedness of ∇G, we have

$$\begin{aligned} \textbf{Term 2} &= o(\|(2x_k - z_k - u_{k+1} - \gamma_k \nabla G(x_k)) - (2x^\star - z^\star - x^\star - \gamma \nabla G(x^\star))\|) \\ &= o(\|z_k - z^\star\|) + O(|\gamma_k - \gamma|). \end{aligned}$$

Now move **Term 1** to the other side of (35) and combine the definition of $\overline{R}$ and the Riemannian Taylor expansion [15, Lemma B.2], to obtain

$$\begin{aligned} &\gamma \tau_{k+1} \nabla_{\mathcal{M}^R_{x^\star}} R(u_{k+1}) - \gamma \nabla_{\mathcal{M}^R_{x^\star}} R(x^\star) - \big(\tau_{k+1}\mathrm{P}_{T^R_{u_{k+1}}} - \mathrm{P}_{T^R_{x^\star}}\big)(x^\star - z^\star - \gamma \nabla G(x^\star)) \\ &= \mathrm{P}_{T^R_{x^\star}} \nabla^2_{\mathcal{M}^R_{x^\star}} \overline{R}(x^\star) \mathrm{P}_{T^R_{x^\star}}(u_{k+1} - x^\star) + o(\|z_k - x^\star\|). \end{aligned}$$

Owing to [15, Lemma 4.3], that the Riemannian Hessian $\mathrm{P}_{T^R_{x^\star}} \nabla^2_{\mathcal{M}^R_{x^\star}} \overline{R}(x^\star) \mathrm{P}_{T^R_{x^\star}}$ is symmetric positive definite. For the term $\mathrm{P}_{T^R_{x^\star}}(\gamma_k \nabla G(x_k) - \gamma \nabla G(x^\star))$, since we assume that F is locally C^2 around $x^\star$, we can apply the Taylor expansion:

$$\begin{aligned} \gamma_k \nabla G(x_k) - \gamma \nabla G(x^\star) &= \gamma(\nabla G(x_k) - \nabla G(x^\star)) + (\gamma_k - \gamma)\nabla G(x_k) \\ &= \mathrm{P}_V\big(\nabla F(x_k) - \nabla F(x^\star)\big) + O(|\gamma_k - \gamma|) \\ &= \mathrm{P}_V \nabla^2 F \mathrm{P}_V(z_k - z^\star) + o(\|z_k - z^\star\|) + O(|\gamma_k - \gamma|). \end{aligned}$$

Recall that $H_{\overline{R}}:=\mathrm{P}_{T^R_{x^\star}}\nabla^2_{\mathcal{M}^R_{x^\star}}\overline{R}(x^\star)\mathrm{P}_{T^R_{x^\star}}$ and $H_G:=\mathrm{P}_V\nabla^2 F\mathrm{P}_V$. Then, for (35),

$$\begin{aligned}
H_{\overline{R}}(u_{k+1}-x^\star) &= 2\mathrm{P}_{T^R_{x^\star}}(x_k-x^\star)-\mathrm{P}_{T^R_{x^\star}}(z_k-z^\star)-\mathrm{P}_{T^R_{x^\star}}(u_{k+1}-x^\star)\\
&\quad -\gamma H_G(z_k-z^\star)+o(\|z_k-z^\star\|)+O(|\gamma_k-\gamma|)\\
\Longrightarrow (\mathrm{Id}+H_{\overline{R}})\mathrm{P}_{T^R_{x^\star}}(u_{k+1}-x^\star) &= 2\mathrm{P}_{T^R_{x^\star}}(x_k-x^\star)-\mathrm{P}_{T^R_{x^\star}}(z_k-z^\star)\\
&\quad -\gamma H_G(z_k-z^\star)+o(\|z_k-z^\star\|)+O(|\gamma_k-\gamma|)\\
\Longrightarrow \mathrm{P}_{T^R_{x^\star}}(u_{k+1}-x^\star) &= 2M_{\overline{R}}\mathrm{P}_V(z_k-z^\star)-M_{\overline{R}}(z_k-z^\star)-\gamma M_{\overline{R}}H_G(z_k-z^\star)\\
&\quad +o(\|z_k-z^\star\|)+O(|\gamma_k-\gamma|)\\
\Longrightarrow u_{k+1}-x^\star &= 2M_{\overline{R}}\mathrm{P}_V(z_k-z^\star)-M_{\overline{R}}(z_k-z^\star)-\gamma M_{\overline{R}}H_G(z_k-z^\star)\\
&\quad +o(\|z_k-z^\star\|)+O(|\gamma_k-\gamma|),
\end{aligned} \tag{36}$$

where we used several times the relation $u_k-x^\star=\mathrm{P}_{T^R_{x^\star}}(u_k-x^\star)+o(\|u_k-x^\star\|)$. Summing up (36) and $x_k-x^\star=\mathrm{P}_V(z_k-z^\star)$ yields

$$\begin{aligned}
(z_k+u_{k+1}-x_k)-z^\star &= (z_k-z^\star)+(u_{k+1}-x^\star)-(x_k-x^\star)\\
&= \mathscr{M}_\gamma(z_k-z^\star)+o(\|z_k-z^\star\|)+O(|\gamma_k-\gamma|).
\end{aligned}$$

Hence for the non-stationary FDR iteration, we have

$$\begin{aligned}
z_{k+1}-z^\star &= (1-\lambda_k)(z_k-z^\star)+\lambda_k\big((z_k+u_{k+1}-x_k)-(z^\star+x^\star-x^\star)\big)\\
&= (1-\lambda_k)(z_k-z^\star)+\lambda_k\mathscr{M}_\gamma(z_k-z^\star)+o(\|z_k-z^\star\|)+\chi_k\\
&= \mathscr{M}_{\gamma,\lambda}(z_k-z^\star)-(\lambda_k-\lambda)(\mathrm{Id}-\mathscr{M}_\gamma)(z_k-z^\star)+o(\|z_k-z^\star\|)+\chi_k.
\end{aligned}$$

Since $\lim_{k\to+\infty}\frac{\|(\lambda_k-\lambda)(\mathrm{Id}-\mathscr{M}_\gamma)(z_k-z^\star)\|}{\|z_k-z^\star\|}\leq\lim_{k\to+\infty}\frac{|\lambda_k-\lambda|\|\mathrm{Id}-\mathscr{M}_\gamma\|\|z_k-z^\star\|}{\|z_k-z^\star\|}=0$, then we get $z_{k+1}-z^\star=\mathscr{M}_{\gamma,\lambda}(z_k-z^\star)+\psi_k+\chi_k$ and conclude the proof. □

Before presenting the local linear convergence result, we need to study the spectral properties of $\mathscr{M}_{\gamma,\lambda}$, which is presented in the lemma below.

Lemma 5.2 *Given $\gamma\in]0,2\beta_V[$ and $\lambda\in]0,\frac{4\beta_V-\gamma}{2\beta_V}[$, we have that $\mathscr{M}_\gamma$ is $\frac{2\beta_V}{4\beta_V-\gamma}$-averaged and $\mathscr{M}_{\gamma,\lambda}$ is $\frac{2\beta_V\lambda}{4\beta_V-\gamma}$-averaged. Moreover, for all k large enough*

(i) *$\mathscr{M}_{\gamma,\lambda}$ converges to some matrix $\mathscr{M}^\infty_\gamma$ and,*

$$\mathscr{M}^k_{\gamma,\lambda}-\mathscr{M}^\infty_\gamma=(\mathscr{M}_{\gamma,\lambda}-\mathscr{M}^\infty_\gamma)^k \quad \textit{and} \quad \rho(\mathscr{M}_{\gamma,\lambda}-\mathscr{M}^\infty_\gamma)<1.$$

(ii) *Given any $\rho\in]\rho(\mathscr{M}_{\gamma,\lambda}-\mathscr{M}^\infty_\gamma),1[$, $\|\mathscr{M}^k_{\gamma,\lambda}-\mathscr{M}^\infty_\gamma\|=O(\rho^k)$.*

Proof Since $W_{\overline{R}}$ is firmly non-expansive by Lemma 5.1, it follows from [1, Example 4.7] that $M_{\overline{R}}$ is firmly non-expansive and hence $\mathcal{R}_{M_{\overline{R}}}:=2M_{\overline{R}}-\mathrm{Id}$ is non-expansive. Similarly, as P_V is firmly non-expansive, $\mathcal{R}_V:=2\mathrm{P}_V-\mathrm{Id}$ is non-expansive. As a result, $\frac{1}{2}(\mathcal{R}_{M_{\overline{R}}}\mathcal{R}_V+\mathrm{Id})$ is firmly non-expansive [1, Proposition 4.21(i)-(ii)]. Then, given $\gamma\in[0,2\beta_V]$, $\mathrm{Id}-\gamma H_G$ is $\frac{2\beta_V}{4\beta_V-\gamma}$-averaged non-expansive. Therefore, owing to Lemma 4.1, we have the averaged property of $\mathscr{M}_\gamma$ and $\mathscr{M}_{\gamma,\lambda}$. We deduce from [1, Proposition 5.15] that $\mathscr{M}_\gamma$ and $\mathscr{M}_{\gamma,\lambda}$ are

convergent, *i.e.* the limit of $\mathcal{M}_{\gamma,\lambda}^k$ exists as k approaches $+\infty$. It is denoted as $\mathcal{M}_\gamma^\infty$. Moreover, $\mathcal{M}_{\gamma,\lambda}^k - \mathcal{M}_\gamma^\infty = (\mathcal{M}_{\gamma,\lambda} - \mathcal{M}_\gamma^\infty)^k$, $\forall k \in \mathbb{N}$, and $\rho(\mathcal{M}_{\gamma,\lambda} - \mathcal{M}_\gamma^\infty) < 1$ by [36, Theorem 2.12]. The second claim of the lemma is classical using the spectral radius formula; See *e.g.* [36, Theorem 2.12(i)]. □

Owing to Lemma 5.2, we can further simplify the linearized iteration (34).

Corollary 5.1 *Consider the non-stationary FDR iteration* (6) *and suppose that it is run under the assumptions of Theorem 5.2. Then the following holds:*

(i) *Iteration* (34) *is equivalent to*

$$\begin{aligned}&(\mathrm{Id} - \mathcal{M}_\gamma^\infty)(z_{k+1} - z^\star)\\ &= (\mathcal{M}_{\gamma,\lambda} - \mathcal{M}_\gamma^\infty)(\mathrm{Id} - \mathcal{M}_\gamma^\infty)(z_k - z^\star) + (\mathrm{Id} - \mathcal{M}_\gamma^\infty)\psi_k + \chi_k.\end{aligned} \tag{37}$$

(ii) *If moreover R is locally polyhedral around $x^\star$ and F is quadratic, then* $z_{k+1} - z^\star = (\mathcal{M}_{\gamma,\lambda} - \mathcal{M}_\gamma^\infty)(z_k - z^\star)$.

Proof For the first claim. Let $K \in \mathbb{N}$ sufficiently large such that the locally linearized iteration (34) holds, then we have for all $k \geq K$

$$\begin{aligned}z_{k+1} - z^\star &= \mathcal{M}_{\gamma,\lambda}(z_k - z^\star) + \psi_k + \chi_k\\ &= \mathcal{M}_{\gamma,\lambda}\big(\mathcal{M}_{\gamma,\lambda}(z_{k-1} - z^\star) + \psi_{k-1} + \chi_{k-1}\big) + \psi_k + \chi_k\\ &= \mathcal{M}_{\gamma,\lambda}^{k+1-K}(z_K - z^\star) + \textstyle\sum_{j=K}^k \mathcal{M}_{\gamma,\lambda}^{k-j}(\psi_j + \chi_j).\end{aligned} \tag{38}$$

Since $z_k \to z^\star$ and $\mathcal{M}_{\gamma,\lambda}$ is convergent to $\mathcal{M}_\gamma^\infty$ by Lemma 5.2, taking the limit as $k \to +\infty$, we have for all finite $p \geq K$,

$$\lim_{k\to+\infty} \textstyle\sum_{j=p}^k \mathcal{M}_{\gamma,\lambda}^{k-j}(\psi_j + \chi_j) = -\mathcal{M}_\gamma^\infty(z_p - z^\star). \tag{39}$$

Using (39) in (38), we get

$$z_{k+1} - z^\star = (\mathcal{M}_{\gamma,\lambda} - \mathcal{M}_\gamma^\infty)(z_k - z^\star) + (\mathrm{Id} - \mathcal{M}_\gamma^\infty)(\psi_j + \chi_j) + \mathcal{M}_\gamma^\infty(z_{k+1} - z^\star).$$

It is also immediate to see from Lemma 5.2 that $\|\mathrm{Id} - \mathcal{M}_\gamma^\infty\| \leq 1$ and that $(\mathcal{M}_{\gamma,\lambda} - \mathcal{M}_\gamma^\infty)(\mathrm{Id} - \mathcal{M}_\gamma^\infty) = \mathcal{M}_{\gamma,\lambda} - \mathcal{M}_\gamma^\infty$. Rearranging the terms yields the claimed equivalence.

Under polyhedrality and constant parameters, we have from Theorem 5.2 that $o(\|z_k - z^\star\|)$ and $O(\lambda_k|\gamma_k - \gamma|)$ vanish, and the result follows. □

5.3 Local Linear Convergence

We are now in position to claim local linear convergence of the FDR iterates.

Theorem 5.3 *Consider the non-stationary FDR iteration* (6) *and suppose it is run under the conditions of Theorem 5.2. Let be $\rho \in]\rho(\mathcal{M}_{\gamma,\lambda} - \mathcal{M}_\gamma^\infty), 1[$ and $K \in \mathbb{N}$ such that, for all $k \geq K$, $\|\mathcal{M}_{\gamma,\lambda}^k - \mathcal{M}_\gamma^\infty\| = O(\rho^k)$ (see Lemma 5.2). Then the following holds:*

(i) *If there exists* $\eta \in]0, \rho[$ *such that* $\lambda_k|\gamma_k - \gamma| = O(\eta^{k-K})$, *then*

$$\|(\mathrm{Id} - \mathscr{M}_\gamma^\infty)(z_k - z^\star)\| = O(\rho^{k-K}). \tag{40}$$

(ii) *If moreover* R *is locally polyhedral around* $x^\star$, F *is quadratic, and that* $(\gamma_k, \lambda_k) \equiv (\gamma, \lambda) \in]0, 2\beta_V[\times]0, \frac{4\beta_V - \gamma}{2\beta_V}[$, *then we have*

$$\|z_k - z^\star\| \leq \rho^{k-K}\|z_K - z^\star\|. \tag{41}$$

Remark 5.3

- For the first case of Theorem 5.3, if $\mathscr{M}_\gamma^\infty = 0$ then we obtain the convergence rate directly on $\|z_k - z^\star\|$. Moreover, we can further derive the convergence rate of $\|x_k - x^\star\|$ and $\|u_k - x^\star\|$.
- The condition on $\lambda_k|\gamma_k - \gamma|$ in Theorem 5.3(i) implies that $\{\gamma_k\}_{k\in\mathbb{N}}$ should converge fast enough to γ. Otherwise, the local convergence rate would be dominated by that of $\lambda_k|\gamma_k - \gamma|$. Especially, if $\lambda_k|\gamma_k - \gamma|$ converges sublinearly to 0, then the local convergence rate will eventually become sublinear. See Figure 2 in the experiments section for a numerical illustration.
- The above result can be easily extended to the case of GFB method, for the sake of simplicity we shall skip the details here. Nevertheless, numerical illustrations will be provided in Section 6.

Proof For the first claim, let $K \in \mathbb{N}$ be sufficiently large such that (37) holds. We then have from Corollary 5.1(i)

$$\begin{aligned}(\mathrm{Id} - \mathscr{M}_\gamma^\infty)(z_{k+1} - z^\star) &= (\mathscr{M}_{\gamma,\lambda} - \mathscr{M}_\gamma^\infty)^{k+1-K}(\mathrm{Id} - \mathscr{M}_\gamma^\infty)(z_K - z^\star)\\ &\quad + \textstyle\sum_{j=K}^k (\mathscr{M}_{\gamma,\lambda} - \mathscr{M}_\gamma^\infty)^{k-j}\big((\mathrm{Id} - \mathscr{M}_\gamma^\infty)\psi_j + \chi_j\big).\end{aligned}$$

Since $\rho(\mathscr{M}_{\gamma,\lambda} - \mathscr{M}_\gamma^\infty) < 1$ by Lemma 5.2, from the spectral radius formula, we know that for every $\rho \in]\rho(\mathscr{M}_{\gamma,\lambda} - \mathscr{M}_\gamma^\infty), 1[$, there is a constant C such that $\|(\mathscr{M}_{\gamma,\lambda} - \mathscr{M}_\gamma^\infty)^j\| \leq C\rho^j$ holds for all integers j. We thus get

$$\begin{aligned}&\|(\mathrm{Id} - \mathscr{M}_\gamma^\infty)(z_{k+1} - z^\star)\|\\ &\leq C\big(\rho^{k+1-K}\|z_K - z^\star\| + \textstyle\sum_{j=K}^k \rho^{k-j}\chi_j + \sum_{j=K}^k \rho^{k-j}\|(\mathrm{Id} - \mathscr{M}_{\gamma,\lambda})\psi_j\|\big)\\ &= C\big(\rho^{k+1-K}\big(\|z_K - z^\star\| + \rho^{K-1}\textstyle\sum_{j=K}^k \frac{\chi_j}{\rho^j}\big) + \sum_{j=K}^k \rho^{k-j}\|(\mathrm{Id} - \mathscr{M}_{\gamma,\lambda})\psi_j\|\big).\end{aligned}$$

By assumption, $\chi_j = C'\eta^j$ for some constant $C' \geq 0$ and $\eta < \rho$. Then we have

$$\rho^{K-1}\textstyle\sum_{j=K}^k \frac{\chi_j}{\rho^j} \leq C'\rho^{K-1}\sum_{j=K}^\infty (\eta/\rho)^j = \frac{C'\eta^K}{\rho - \eta} < +\infty.$$

Setting $C'' = C(\|z_K - z^\star\| + \frac{C'\eta^K}{\rho-\eta}) < +\infty$, we obtain

$$\|(\mathrm{Id} - \mathscr{M}_\gamma^\infty)(z_{k+1} - z^\star)\| \leq C''\rho^{k+1-K} + C\textstyle\sum_{j=K}^k \rho^{k-j}\|(\mathrm{Id} - \mathscr{M}_\gamma^\infty)\psi_j\|.$$

This, together with the fact that $\|(\mathrm{Id} - \mathscr{M}_\gamma^\infty)\psi_j\| = o(\|(\mathrm{Id} - \mathscr{M}_\gamma^\infty)(z_j - z^\star)\|)$ yields the claimed result. The second claim follows from Corollary 5.1 that $z_k - z^\star = (\mathscr{M}_{\gamma,\lambda} - \mathscr{M}_\gamma^\infty)^{k+1-K}(z_K - z^\star)$ and we conclude the proof. □

5.4 Extension to Three-Operator Splitting

So far, we have presented the global and local convergence analysis of the FDR algorithm. As we recalled in the introduction, FDR is closely related with the three-operator splitting method (TOS) [7]. Therefore, it would be interesting to extend the obtained result to TOS. However, extending the global convergence result to TOS is far from straightforward. Hence, in the following, we mainly focus on the local aspect.

For the sake of notational simplicity, we rewrite problem (9) as

$$\min_{x\in\mathbb{R}^n} \big\{\Psi(x) = F(x) + R(x) + J(x)\big\}, \tag{42}$$

where we suppose the following assumptions:

- (**B.1**) $J, R \in \Gamma_0(\mathbb{R}^n)$.
- (**B.2**) $F : \mathbb{R}^n \to \mathbb{R}$ is convex continuously differentiable with ∇F being $(1/\beta)$-Lipschitz continuous.
- (**B.3**) $\operatorname{Argmin}(\Psi) \neq \emptyset$, *i.e.* the set of minimizers is not empty.

Correspondingly, the TOS iteration (10) becomes

$$\begin{aligned} u_{k+1} &= \operatorname{prox}_{\gamma R}\big(2x_k - z_k - \gamma\nabla F(x_k)\big)\\ z_{k+1} &= z_k + \lambda_k(u_{k+1} - x_k),\\ x_{k+1} &= \operatorname{prox}_{\gamma J}(z_{k+1}). \end{aligned} \tag{43}$$

We suppose the following assumption on the algorithm parameters:

- (**B.4**) The (constant) step-size verifies $\gamma \in]0, 2\beta[$ and the sequence of relaxation parameters $\{\lambda_k\}_{k\in\mathbb{N}}$ is such that $\sum_{k\in\mathbb{N}} \lambda_k(\frac{4\beta-\gamma}{2\beta} - \lambda_k) = +\infty$.

The fixed-point operator of TOS reads as

$$\mathcal{T}_\gamma = \mathrm{Id} - \operatorname{prox}_{\gamma R} + \operatorname{prox}_{\gamma J}(2\operatorname{prox}_{\gamma R} - \mathrm{Id} - \gamma\nabla F \circ \operatorname{prox}_{\gamma J}), \tag{44}$$

and $\mathcal{T}_{\gamma,\lambda_k} = (1-\lambda_k)\mathrm{Id} + \lambda_k\mathcal{T}_\gamma$. Differently from $\mathcal{F}_\gamma$ (see (11)), $\mathcal{T}_\gamma$ cannot be simplified into a compact form.

Lemma 5.3 ([7, Proposition 2.1 and Theorem 2.1]) *Consider the TOS iteration* (43) *and the fixed-point operator* (44). *Suppose that Assumptions* (**B.1**)-(**B.4**) *hold. Then,*

(i) *the operator* $\mathcal{T}_\gamma$ *is* $\frac{2\beta}{4\beta-\gamma}$*-averaged non-expansive.*

(ii) $\{z_k\}_{k\in\mathbb{N}}$ *converges to some* $z^\star$ *in* $\operatorname{fix}(\mathcal{T}_\gamma)$*; moreover, both* $\{u_k\}_{k\in\mathbb{N}}$ *and* $\{x_k\}_{k\in\mathbb{N}}$ *converge to* $x^\star := \operatorname{prox}_{\gamma J}(z^\star)$*, which is a global minimizer of* Ψ.

Similar to (30), under Lemma 5.3, we have the optimality condition

$$(x^\star - z^\star)/\gamma \in \nabla F(x^\star) + \partial R(x^\star) \;\text{ and }\; (z^\star - x^\star)/\gamma \in \partial J(x^\star).$$

Following Section 5.1-5.3, we present the local convergence of TOS.

Finite Activity Identification We start with the finite identification result, for both u_k, x_k as J is no longer the indicator function of a subspace.

Corollary 5.2 *For the TOS iteration* (43)*. Suppose it is run under the Assumptions* (B.1)-(B.4)*, so that* $(u_k, x_k, z_k) \to (x^\star, x^\star, z^\star)$ *where* $z^\star \in \mathrm{fix}(\mathcal{T}_\gamma)$ *and* $x^\star{:=}\mathrm{prox}_{\gamma J}(z^\star) \in \mathrm{Argmin}(\Psi)$*. Moreover, suppose* $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}^R_{x^\star})$*,* $J \in \mathrm{PSF}_{x^\star}(\mathcal{M}^J_{x^\star})$*, and the following non-degeneracy condition holds*

$$(x^\star - z^\star)/\gamma - \nabla F(x^\star) \in \mathrm{ri}\big(\partial R(x^\star)\big) \;\text{ and }\; (z^\star - x^\star)/\gamma \in \mathrm{ri}\big(\partial J(x^\star)\big). \tag{45}$$

Then, there exists $K \in \mathbb{N}$ *such that* $(u_k, x_k) \in \mathcal{M}^R_{x^\star} \times \mathcal{M}^J_{x^\star}$ *for every* $k \geq K$*.*

Local Linearized Iteration Define $\widetilde{R}(u){:=}\gamma R(u) - \langle u,\, x^\star - z^\star - \gamma\nabla F(x^\star)\rangle$ and $\widetilde{J}(x){:=}\gamma J(x) - \langle x,\, z^\star - x^\star\rangle$. We have the following corollary from Lemma 5.1.

Corollary 5.3 *Suppose that* $J \in \mathrm{PSF}_{x^\star}(\mathcal{M}^J_{x^\star})$ *and* $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}^R_{x^\star})$*. Then their Riemannian Hessians at* $x^\star$ *read*

$$H_{\widetilde{J}}{:=}\mathrm{P}_{T^J_{x^\star}} \nabla^2_{\mathcal{M}^J_{x^\star}} \widetilde{J}(x^\star)\mathrm{P}_{T^J_{x^\star}} \;\text{ and }\; H_{\widetilde{R}}{:=}\mathrm{P}_{T^R_{x^\star}} \nabla^2_{\mathcal{M}^R_{x^\star}} \widetilde{R}(x^\star)\mathrm{P}_{T^R_{x^\star}},$$

which are symmetric positive semi-definite under either of the two conditions:

(i) *condition* (45) *holds.*

(ii) $\mathcal{M}^J_{x^\star}$ *and* $\mathcal{M}^R_{x^\star}$ *are affine subspaces.*

In turn, the matrices $W_{\widetilde{J}}{:=}(\mathrm{Id}{+}H_{\widetilde{J}})^{-1}$ *and* $W_{\widetilde{R}}{:=}(\mathrm{Id}{+}H_{\widetilde{R}})^{-1}$ *are both firmly non-expansive.*

Now assume F is locally C^2-smooth around $x^\star$, and define $H_F{:=}\nabla^2 F(x^\star)$. Define also $M_{\widetilde{J}}{:=}\mathrm{P}_{T^J_{x^\star}} W_{\widetilde{J}} \mathrm{P}_{T^J_{x^\star}}$ and $M_{\widetilde{R}}{:=}\mathrm{P}_{T^R_{x^\star}} W_{\widetilde{R}} \mathrm{P}_{T^R_{x^\star}}$, and the matrices

$$\mathscr{L}_\gamma = \mathrm{Id} + 2M_{\widetilde{R}}M_{\widetilde{J}} - M_{\widetilde{R}} - M_{\widetilde{J}} - \gamma M_{\widetilde{R}} H_F M_{\widetilde{J}} \;\text{ and }\; \mathscr{L}_{\gamma,\lambda} = (1-\lambda)\mathrm{Id} + \lambda\mathscr{L}_\gamma.$$

Lemma 5.4 ([7, Proposition 2.1]) $\mathscr{L}_\gamma$ *is* $\frac{2\beta}{4\beta-\gamma}$*-averaged non-expansive.*

The above lemma entails that $\mathscr{L}_\gamma, \mathscr{L}_{\gamma,\lambda}$ are convergent, hence the spectral properties result in Lemma 5.2 applies to them. Denote $\mathscr{L}^\infty_\gamma{:=}\lim_{k\to+\infty}\mathscr{L}^k_{\gamma,\lambda}$.

Corollary 5.4 *Consider the TOS iteration* (43)*. Suppose it is run under Assumptions* (B.1)-(B.4)*, that* $\lambda_k \to \lambda \in]0, \frac{4\beta-\gamma}{2\beta}[$*, and that* F *is locally* C^2 *around* $x^\star$*. Then we have* $z_{k+1} - z^\star = \mathscr{L}_{\gamma,\lambda}(z_k - z^\star) + o(\|z_k - z^\star\|)$*. If moreover* J, R *are locally polyhedral around* $x^\star$*,* F *is quadratic and* $\lambda_k \equiv \lambda \in]0, \frac{4\beta-\gamma}{2\beta}[$ *is chosen constant, then the term* $o(\|z_k - z^\star\|)$ *vanishes.*

We can also specialize Corollary 5.1 to this context, however we choose to skip it owing to its obviousness.

Local Linear Convergence Finally, we are able to to present the local linear convergence for (43).

Corollary 5.5 *For the TOS iteration* (43)*. Suppose Assumptions* (**B.1**)-(**B.4**) *hold, and that* $\lambda_k \to \lambda \in]0, \frac{4\beta-\gamma}{2\beta}[$*, and that* F *is locally* C^2 *around* $x^\star$*. Then*

(i) *Given any* $\rho \in]\rho(\mathscr{L}_{\gamma,\lambda} - \mathscr{L}_\gamma^\infty), 1[$*, there exists* $K \in \mathbb{N}$ *large enough such that* $\|(\mathrm{Id} - \mathscr{L}_\gamma^\infty)(z_k - z^\star)\| = O(\rho^{k-K})\ \forall k \geq K$*.*

(ii) *If moreover* J, R *are locally polyhedral around* $x^\star$*,* F *is quadratic and* $\lambda_k \equiv \lambda \in]0, \frac{4\beta-\gamma}{2\beta}[$ *is chosen constant, then there exists* $K \in \mathbb{N}$ *such that* $\|z_k - z^\star\| \leq \rho^{k-K}\|z_K - z^\star\|\ \forall k \geq K$*.*

6 Numerical Experiments

In this section, we illustrate our theoretical results on problems arising from statistics, and signal/image processing applications[3].

6.1 Examples of Partly Smooth Functions

Table 1 provides some examples of popular partly smooth functions. More details about them can be found in [15, Section 5] and references therein.

Table 1: Examples of partly smooth functions. For $x \in \mathbb{R}^n$ and some subset of indices $\mathfrak{b} \subset \{1, \ldots, n\}$, $x_\mathfrak{b}$ is the restriction of x to the entries indexed in $\mathfrak{b}$. For ℓ_∞-norm, $I_x = \big\{i : \ |x_i| = \|x\|_\infty\big\}$. D_{DIF} stands for the finite differences operator [37], $I_{D_{\text{DIF}}x} = \{i : (D_{\text{DIF}}x)_i \neq 0\}$. $\mathrm{sign}(x_{I_x})$ is the sign vector of x_{I_x}, and $\mathbb{R}\mathrm{sign}(x_{I_x})$ is the span of $\mathrm{sign}(x_{I_x})$. $\sigma(x)$ denotes the singular values of x.

Function	Expression	Partial smooth manifold
ℓ_1-norm	$\|x\|_1 = \sum_{i=1}^n \|x_i\|$	$\mathcal{M} = \big\{z \in \mathbb{R}^n : \ I_z \subseteq I_x\big\}$, $I_x = \big\{i : \ x_i \neq 0\big\}$
$\ell_{1,2}$-norm	$\sum_{i=1}^m \|x_{\mathfrak{b}_i}\|$	$\mathcal{M} = \big\{z \in \mathbb{R}^n : \ I_z \subseteq I_x\big\}$, $I_x = \big\{i : \ x_{\mathfrak{b}_i} \neq 0\big\}$
ℓ_∞-norm	$\max_{i=\{1,\ldots,n\}} \|x_i\|$	$\mathcal{M} = \big\{z \in \mathbb{R}^n : \ z_{I_x} \in \mathbb{R}\mathrm{sign}(x_{I_x})\big\}$
TV semi-norm	$\|x\|_{\text{TV}} = \|D_{\text{DIF}}x\|_1$	$\mathcal{M} = \big\{z \in \mathbb{R}^n : \ I_{D_{\text{DIF}}z} \subseteq I_{D_{\text{DIF}}x}\big\}$
Nuclear norm	$\|x\|_* = \sum_{i=1}^r \sigma(x)$	$\mathcal{M} = \big\{z \in \mathbb{R}^{n_1 \times n_2} : \ \mathrm{rank}(z) = \mathrm{rank}(x) = r\big\}$

The ℓ_1, ℓ_∞-norms and the anisotropic TV semi-norm are all polyhedral functions, hence the corresponding Riemannian Hessians are simply 0. The $\ell_{1,2}$-norm is not polyhedral yet partly smooth relative to a subspace; the nuclear norm is partly smooth relative to the manifold of fixed-rank matrices, which is no longer a subspace. The Riemannian Hessian of these two functions are non-trivial and can be computed following [38].

[3] MATLAB source for reproducing the numerical result can be found at `https://github.com/jliang993/Rate-FDR`.

6.2 Numerical Experiments

Global Convergence Rate of the Bregman Distance We first demonstrate, numerically, the global $o(1/k)$ convergence rate of the Bregman divergence of Section 4. Towards this goal, we consider the fused LASSO problem [39]

$$\min_{x\in\mathbb{R}^n} \mu_1\|x\|_1 + \mu_2\|D_{\text{DIF}}x\|_1 + \frac{1}{2}\|\mathcal{K}x - f\|^2, \tag{46}$$

where $\mu_1, \mu_2 > 0$ are trade-off weights. Note that all assumptions (**A.1**)-(**A.4**) hold (in particular the set of minimizers is a non-empty compact set by coercivity of $\|\cdot\|_1$). The problem can be solved using the GFB instance of FDR in (8). In the test, we consider $n = 128$ and $\mathcal{K} \in \mathbb{R}^{36\times 128}$ is a random Gaussian matrix. The step-size is chosen as $\gamma_k \equiv \frac{1}{4\|\mathcal{K}\|^2}$ such that we can observe the $o(1/k)$ convergence behaviour for enough number of iterations.

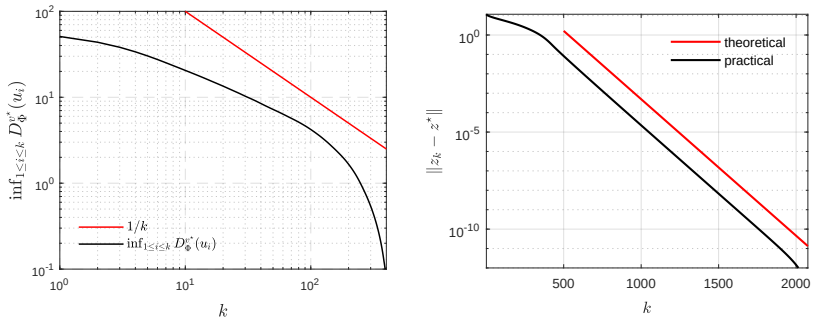

Fig. 1: Results of applying (8) to solve the fused LASSO problem (46). (a): convergence profile of the Bregman distance $\inf_{0\le i\le k} \mathcal{D}_\Phi^{v^\star}(u_i)$. (b): convergence profile of $\|z_k - z^\star\|$.

The convergence profile of $\min_{0\le i\le k} \mathcal{D}_\Phi^{v^\star}(u_i)$ is shown in Figure 1(a). The plot is in log-log scale, where the red line corresponds to the sub-linear $O(1/k)$ rate and the black line is $\min_{0\le i\le k} \mathcal{D}_\Phi^{v^\star}(u_i)$. One can then confirm numerically the prediction of Theorem 4.2.

However, it can be observed that beyond some iteration, *e.g.* 10^2 for the consider example, the convergence rate changes to linear. We argue in the next section that this is likely to be due to finite activity identification since ℓ_1-norm and total variation are partly smooth (in fact even polyhedral) and that, for all k large enough, GFB enters into a local linear convergence regime.

Local Linear Convergence of GFB/FDR Following the above discussion, in Figure 1(b) we present the local linear convergence of FDR in terms of $\|z_k - z^\star\|$ as we are in the scope of Theorem 5.3(ii). We use the same parameters setting as in Figure 1(a). The red line stands for the estimated rate (see Theorem 5.3),

while the black line is numerical observation. The starting point of the red line is the number of iteration where u_k identifies the manifolds. As shown in the figure, we indeed have local linear convergence behaviour of $\|z_k - z^\star\|$. Moreover, since $F = \frac{1}{2}\|\mathcal{K}x - f\|^2$ is quadratic, ℓ_1-norm and total variation are polyhedral, our theoretical rate estimation is tight, *i.e.* the red line has the same slope as the black line.

Non-Stationary FDR We now investigate the convergence behaviour of the non-stationary version of FDR and compare it to the stationary one. We fix $\lambda_k \equiv 1$, *i.e.* the iteration is unrelaxed. The stationary FDR algorithm is run with $\gamma = \beta$. For the non-stationary ones, four choices of γ_k are considered:

$$\text{Case 1: } \gamma_k = (1 + \tfrac{1}{k^{1.1}})\beta, \qquad \text{Case 2: } \gamma_k = (1 + \tfrac{1}{k^2})\beta,$$
$$\text{Case 3: } \gamma_k = (1 + 0.999^k)\beta, \qquad \text{Case 4: } \gamma_k = (1 + 0.5^k)\beta.$$

Obviously, we have $\gamma_k \to \gamma = \beta$ and $\sum_{k\in\mathbb{N}} |\gamma_k - \gamma| < +\infty$ for all cases. Problem (46) is considered. The comparison results are displayed in Figure 2(a).

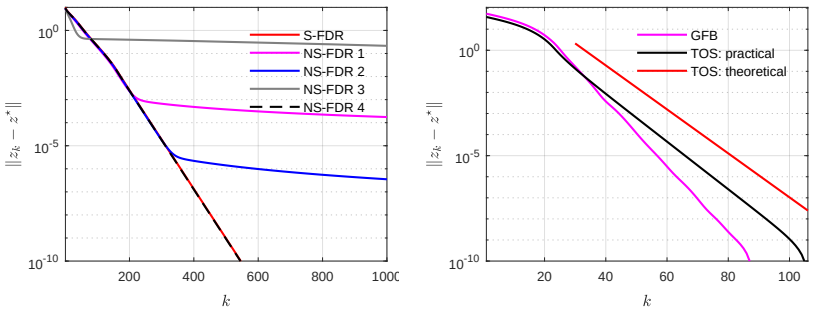

Fig. 2: (a): comparison of $\|z_k - z^\star\|$ between stationary ("S-FDR") and non-stationary FDR ("NS-FDR X", X stands for Case X). (b): comparison of $\|z_k - z^\star\|$ between GFB and TOS for problem (47).

We can make the following observations from the comparison:

- In agreement with our analysis, the local convergence behaviour of the non-stationary iteration is no better than the stationary one. This contrasts with the global behaviour where non-stationarity could be beneficial (see last comment hereafter);
- As argued in Remark 5.3(ii), the convergence rate is eventually controlled by the error $|\gamma_k - \gamma|$, except for "Case 4", Indeed, 0.5 is strictly smaller than the local linear rate of the stationary version (*i.e.* $|\gamma_k - \gamma| = o(\|z_k - z^\star\|)$);
- The non-stationary FDR seems to lead to faster identification, typically for "Case 3". This is the effect of bigger step-size at the early stage.

Local Linear Convergence of GFB/TOS To conclude the numerical experiments, we demonstrate the local convergence behaviour of GFB and TOS algorithms. Consider the non-negative low-rank matrix completion problem

$$\min_{x\in\mathbb{R}^{n\times n}} \mu\|x\|_* + \iota_{\mathbb{R}^{n\times n}_+}(x) + \frac{1}{2}\|\mathcal{K}x - f\|^2, \tag{47}$$

where we recall that $\|\cdot\|_*$ is the nuclear norm (sum of singular values), and $\mathbb{R}^{n\times n}_+$ is the set of matrices with non-negative entries. Again, our main assumptions (**A.1**)-(**A.4**) are verified thanks to continuity, convexity and coercivity. Problem (47) is a special instance of (42) if we let $F = \frac{1}{2}\|\mathcal{K}\cdot - f\|^2$, $R = \mu\|\cdot\|_*$ and $J = \iota_{\mathbb{R}^n_+}(x)$. Hence it can be solved by the TOS scheme (43) and also by the GFB algorithm (8).

In the test, we consider $x \in \mathbb{R}^{50\times 50}$ and $\mathcal{K}$ is the sub-sampling operator (we did not consider larger problem size as computing the theoretical rate is very time and memory consuming). Figure 2(b) shows the convergence profiles of GFB/TOS. Similarly to the observation made in Figure 1(b), both GFB (magenta line) and TOS (black line) converge sub-linearly from the beginning and eventually enter a linear convergence regime. The red line is our theoretical linear rate estimation of TOS. Moreover, for this example, the performances of two algorithms are very close, especially for the global sub-linear regime.

7 Perspectives and Open Problems

In this paper, we address the convergence properties of FDR algorithm from both global and local perspectives. The obtained results allow us to better understand the optimisation problem (1) and FDR algorithm, and moreover lay the foundation for our future research regarding several open problems.

The first open problem is the acceleration of FDR/GFB/TOS, or in general acceleration schemes for non-descent type methods. In recent years, owing to the success of Nesterov's optimal scheme [9] and FISTA [12], inertial technique has been widely adopted to speed up other non-descent type operator splitting methods [40]. However, unlike the results in [9,12], the acceleration effects of inertial technique for these non-descent type methods are rather limited, or even slower than the original method [40, Chapter 4]. As a consequence, a proper acceleration scheme for non-descent methods, including FDR/GFB/TOS, with guaranteed acceleration remains an open problem.

Another direction for acceleration is the incremental version of these algorithms, particularly for GFB as the separable structure of $\sum_i R_i(x)$ in (7) is ideal for designing incremental schemes. Moreover, if F also has finite sum structure, *e.g.* $F(x) = \sum_{i=1}^m f_i(x)$, then similar to [41], we can consider incremental schemes for both smooth and non-smooth components of the problem.

The third perspective would be extending the obtained results to the non-Euclidean setting. More precisely, the proximal mapping of (3) is defined based on the Euclidean distance between u and x. By replacing the Euclidean distance with a Bregman distance, we obtain the Bregman-type splitting algo-

rithms which are much more general. Generalizing the obtained results to Bregman-type splitting setting would be important and challenging.

For the local convergence analysis of FDR algorithm, we have to restrict ourselves to finite dimensional Euclidean space, which is due to the fact that partial smoothness is only available in finite dimension. However, recently it is reported that finite identification also occurs for problems in infinite dimension, such as the off-the-grid compressive sensing [42]. As a result, proper extension of partial smoothness to the infinite dimension is required to explain these phenomena.

8 Conclusions

In this paper, we studied global and local convergence properties of the Forward–Douglas–Rachford method. Globally, we established an $o(1/k)$ convergence rate of the best iterate and $O(1/k)$ ergodic rate in terms of a Bregman divergence criterion designed for the method. We also specialized the result to the case of Forward–Backward splitting method, for which we showed that the objective function of the method converges at an $o(1/k)$ rate. Then, locally, we proved the linear convergence of the sequence when the involved functions are moreover partly smooth. In particular, we demonstrated that the method identifies the active manifolds in finite time and that then it converges locally linearly at a rate that we characterized precisely. We also extended the local linear convergence result to the case of three-operator splitting method. Our numerical experiments supported the theoretical findings.

Acknowledgements CM was supported by CONICYT scholarship CONICYT-PCHA/Doctorado Nacional/2016. JL was supported by the European Research Council (ERC project SIGMA-Vision), and Leverhulme Trust project "Breaking the non-convexity barrier", the EPSRC grant "EP/M00483X/1", EPSRC centre "EP/N014588/1", Cantab Capital Institute for the Mathematics of Information, and Global Alliance project "Statistical and Mathematical Theory of Imaging". JF was partly supported by Institut Universitaire de France. We would like to thank the anonymous reviewers whose comments have greatly improve the quality of this paper.

Appendix: Riemannian Geometry

Let $\mathcal{M}$ be a C^2-smooth embedded submanifold of $\mathbb{R}^n$ around a point x. We denote respectively $\mathcal{T}_{\mathcal{M}}(x)$ and $\mathcal{N}_{\mathcal{M}}(x)$ the tangent and normal space of $\mathcal{M}$ at point near x in $\mathcal{M}$.

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

Parallel Translation Given $x, x' \in \mathcal{M}$, let $\mathcal{T}_{\mathcal{M}}(x), \mathcal{T}_{\mathcal{M}}(x')$ be their corresponding tangent spaces. Define $\tau : \mathcal{T}_{\mathcal{M}}(x) \to \mathcal{T}_{\mathcal{M}}(x')$ the parallel translation along the unique geodesic joining x to x', which is isomorphism and isometry w.r.t. the Riemannian metric.

Riemannian Gradient and Hessian For a vector $v \in \mathcal{N}_{\mathcal{M}}(x)$, the Weingarten map of $\mathcal{M}$ at x is the operator $\mathfrak{W}_x(\cdot, v) : \mathcal{T}_{\mathcal{M}}(x) \to \mathcal{T}_{\mathcal{M}}(x)$ defined by $\mathfrak{W}_x(\cdot, v) = -\mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x)} \mathrm{d}V[h]$ where V is any local extension of v to a normal vector field on $\mathcal{M}$. The definition is independent of the choice of the extension V, and $\mathfrak{W}_x(\cdot, v)$ is a symmetric linear operator which is closely tied to the second fundamental form of $\mathcal{M}$, see [43, Proposition II.2.1].

Let J be a real-valued function which is C^2 along the $\mathcal{M}$ around x. The covariant gradient of J at $x' \in \mathcal{M}$ is the vector $\nabla_{\mathcal{M}} J(x') \in \mathcal{T}_{\mathcal{M}}(x')$ defined by $\langle \nabla_{\mathcal{M}} J(x'), h \rangle = \frac{d}{dt} J\big(\mathrm{P}_{\mathcal{M}}(x' + th)\big)\big|_{t=0}$, $\forall h \in \mathcal{T}_{\mathcal{M}}(x')$, where $\mathrm{P}_{\mathcal{M}}$ is the projection operator onto $\mathcal{M}$. The covariant Hessian of J at x' is the symmetric linear mapping $\nabla^2_{\mathcal{M}} J(x')$ from $\mathcal{T}_{\mathcal{M}}(x')$ to itself which is defined as $\langle \nabla^2_{\mathcal{M}} J(x')h,\, h \rangle = \frac{d^2}{dt^2} J\big(\mathrm{P}_{\mathcal{M}}(x' + th)\big)\big|_{t=0}$, $\forall h \in \mathcal{T}_{\mathcal{M}}(x')$. This definition agrees with the usual definition using geodesics or connections [44]. Now assume that $\mathcal{M}$ is a Riemannian embedded submanifold of $\mathbb{R}^n$, and that a function J has a C^2-smooth restriction on $\mathcal{M}$. This can be characterized by the existence of a C^2-smooth extension (representative) of J, *i.e.* a C^2-smooth function $\widetilde{J}$ on $\mathbb{R}^n$ such that $\widetilde{J}$ agrees with J on $\mathcal{M}$. Thus, the Riemannian gradient $\nabla_{\mathcal{M}} J(x')$ is given by $\nabla_{\mathcal{M}} J(x') = \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')} \nabla \widetilde{J}(x')$ and $\forall h \in \mathcal{T}_{\mathcal{M}}(x')$, the Riemannian Hessian reads $\nabla^2_{\mathcal{M}} J(x')h = \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')} \nabla^2 \widetilde{J}(x')h + \mathfrak{W}_{x'}\big(h, \mathrm{P}_{\mathcal{N}_{\mathcal{M}}(x')} \nabla \widetilde{J}(x')\big)$, where the last equality comes from [45, Theorem 1]. When $\mathcal{M}$ is an affine or linear subspace of $\mathbb{R}^n$, then obviously $\mathcal{M} = x + \mathcal{T}_{\mathcal{M}}(x)$, and $\mathfrak{W}_{x'}(h, \mathrm{P}_{\mathcal{N}_{\mathcal{M}}(x')} \nabla \widetilde{J}(x')) = 0$, and we have $\nabla^2_{\mathcal{M}} J(x') = \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')} \nabla^2 \widetilde{J}(x') \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')}$.

References

1. Bauschke, H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer (2011)
2. Briceño-Arias, L.M.: Forward-douglas–rachford splitting and forward-partial inverse method for solving monotone inclusions. Optimization **64**(5), 1239–1261 (2015)
3. Douglas, J., Rachford, H.H.: On the numerical solution of heat conduction problems in two and three space variables. Transactions of the American mathematical Society **82**(2), 421–439 (1956)
4. Lions, P.L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM Journal on Numerical Analysis **16**(6), 964–979 (1979)
5. Raguet, H.: A note on the forward-douglas–rachford splitting for monotone inclusion and convex optimization. Optimization Letters pp. 1–24 (2018)
6. Raguet, H., Fadili, M.J., Peyré, G.: Generalized forward-backward splitting. SIAM Journal on Imaging Sciences **6**(3), 1199–1226 (2013)
7. Davis, D., Yin, W.: A three-operator splitting scheme and its optimization applications. Set-valued and variational analysis **25**(4), 829–858 (2017)
8. Liang, J., Fadili, J., Peyré, G.: Convergence rates with inexact non-expansive operators. Mathematical Programming: Series A **159**(1-2), 403–434 (2016)
9. Nesterov, Y.: A method for solving the convex programming problem with convergence rate $O(1/k^2)$. Dokl. Akad. Nauk SSSR **269**(3), 543–547 (1983)
10. Nesterov, Y.: Introductory lectures on convex optimization: A basic course, vol. 87. Springer (2004)
11. Bredies, K., Lorenz, D.A.: Linear convergence of iterative soft-thresholding. Journal of Fourier Analysis and Applications **14**(5-6), 813–837 (2008)
12. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences **2**(1), 183–202 (2009)
13. Chambolle, A., Dossal, C.: On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". Journal of Optimization Theory and Applications **166**(3), 968–982 (2015)
14. Attouch, H., Peypouquet, J.: The rate of convergence of Nesterov's accelerated Forward–Backward method is actually faster than $1/k^2$. SIAM Journal on Optimization **26**(3), 1824–1834 (2016)
15. Liang, J., Fadili, J., Peyré, G.: Activity identification and local linear convergence of Forward–Backward-type methods. SIAM Journal on Optimization **27**(1), 408–437 (2017)

16. Liang, J., Fadili, J., Peyré, G.: Local convergence properties of Douglas–Rachford and Alternating Direction Method of Multipliers. Journal of Optimization Theory and Applications **172**(3), 874–913 (2017)
17. Liang, J., Fadili, J., Peyré, G.: Local linear convergence analysis of Primal–Dual splitting methods. arXiv preprint arXiv:1705.01926 (2017)
18. Davis, D.: Convergence rate analysis of the Forward–Douglas–Rachford splitting scheme. SIAM Journal on Optimization **25**(3), 1760–1786 (2015)
19. Liang, J., Fadili, J., Peyré, G.: Local linear convergence of Forward–Backward under partial smoothness. In: Advances in Neural Information Processing Systems, pp. 1970–1978 (2014)
20. Bolte, J., Nguyen, T.P., Peypouquet, J., Suter, B.W.: From error bounds to the complexity of first-order descent methods for convex functions. Mathematical Programming **165**(2), 471–507 (2017). DOI 10.1007/s10107-016-1091-6. URL `https://doi.org/10.1007/s10107-016-1091-6`
21. Drusvyatskiy, D., Lewis, A.: Error bounds, quadratic growth, and linear convergence of proximal methods. Tech. Rep. arXiv:1602.06661 (2016)
22. Luo, Z.Q., Tseng, P.: On the linear convergence of descent methods for convex essentially smooth minimization. SIAM Journal on Control and Optimization **30**(2), 408–425 (1992). DOI 10.1137/0330025. URL `https://doi.org/10.1137/0330025`
23. Luo, Z.Q., Tseng, P.: Error bounds and convergence analysis of feasible descent methods: a general approach. Annals of Operations Research **46**(1), 157–178 (1993). DOI 10.1007/BF02096261. URL `https://doi.org/10.1007/BF02096261`
24. Zhou, Z., So, A.M.C.: A unified approach to error bounds for structured convex optimization problems. Mathematical Programming **165**(2), 689–728 (2017). DOI 10.1007/s10107-016-1100-9. URL `https://doi.org/10.1007/s10107-016-1100-9`
25. Li, G., Pong, T.K.: Calculus of the exponent of Kurdyka–Łojasiewicz inequality and its applications to linear convergence of first-order methods. Foundations of Computational Mathematics (2017). DOI 10.1007/s10208-017-9366-8. URL `https://doi.org/10.1007/s10208-017-9366-8`
26. Bertsekas, D.P.: Nonlinear programming. Athena scientific Belmont (1999)
27. Ogura, N., Yamada, I.: Non-strictly convex minimization over the fixed point set of an asymptotically shrinking nonexpansive mapping. Numerical Functional Analysis and Optimization **23(1-2)**, 113–137 (2002)
28. Knopp, K.: Theory and application of infinite series. Courier Corporation (2013)
29. Lewis, A.S.: Active sets, nonsmoothness, and sensitivity. SIAM Journal on Optimization **13**(3), 702–725 (2003)
30. Combettes, P.L.: Quasi-Fejérian analysis of some optimization algorithms. Studies in Computational Mathematics **8**, 115–152 (2001)
31. Brézis, H.: Opérateurs Maximaux Monotones et Semi-Groupes de Contractions dans les Espaces de Hilbert. North-Holland/Elsevier, New York (1973)
32. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting. Multiscale Modeling & Simulation **4**(4), 1168–1200 (2005)
33. Rockafellar, R.T., Wets, R.: Variational analysis, vol. 317. Springer Verlag (1998)
34. Hare, W.L., Lewis, A.S.: Identifying active constraints via partial smoothness and prox-regularity. Journal of Convex Analysis **11**(2), 251–266 (2004)
35. Combettes, P.L.: Solving monotone inclusions via compositions of nonexpansive averaged operators. Optimization **53**(5-6), 475–504 (2004)
36. Bauschke, H.H., Bello Cruz, J., Nghia, T., Phan, H.M., Wang, X.: Optimal rates of convergence of matrices with applications. Numerical Algorithms (2016). In press (arxiv:1407.0671)
37. Condat, L.: A direct algorithm for 1-d total variation denoising. IEEE Signal Processing Letters **20**(11), 1054–1057 (2013)
38. Vaiter, S., Deledalle, C., Fadili, J.M., Peyré, G., Dossal, C.: The degrees of freedom of partly smooth regularizers. Annals of the Institute of Mathematical Statistics (2015). URL `http://arxiv.org/abs/1404.5557`. To appear
39. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**(1), 91–108 (2005)

40. Liang, J.: Convergence rates of first-order operator splitting methods. Ph.D. thesis, Normandie Université; GREYC CNRS UMR 6072 (2016)
41. Bertsekas, D.P.: Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. Optimization for Machine Learning **2010**(1-38), 3 (2011)
42. Poon, C., Peyré, G.: Multidimensional sparse super-resolution. SIAM Journal on Mathematical Analysis **51**(1), 1–44 (2019)
43. Chavel, I.: Riemannian geometry: a modern introduction, vol. 98. Cambridge University Press (2006)
44. Miller, S.A., Malick, J.: Newton methods for nonsmooth convex minimization: connections among-Lagrangian, Riemannian Newton and SQP methods. Mathematical programming **104**(2-3), 609–633 (2005)
45. Absil, P.A., Mahony, R., Trumpf, J.: An extrinsic look at the Riemannian Hessian. In: Geometric Science of Information, pp. 361–368. Springer (2013)