Activity Identification and Local Linear Convergence of Douglas–Rachford/ADMM under Partial Smoothness

Jingwei Liang[1,⋆], Jalal Fadili[1], Gabriel Peyré[2], and Russell Luke[3]

[1] GREYC, CNRS, ENSICAEN, Université de Caen, France
{Jingwei.Liang, Jalal.Fadili}@ensicaen.fr
[2] CNRS, Ceremade, Université Paris-Dauphine, France
Gabriel.Peyre@ceremade.dauphine.fr
[3] Institut für Numerische und Angewandte Mathematik,
Universität Göttingen, Germany
r.luke@math.uni-goettingen.de

Abstract. Convex optimization has become ubiquitous in most quantitative disciplines of science, including variational image processing. Proximal splitting algorithms are becoming popular to solve such structured convex optimization problems. Within this class of algorithms, Douglas–Rachford (DR) and ADMM are designed to minimize the sum of two proper lower semi-continuous convex functions whose proximity operators are easy to compute. The goal of this work is to understand the local convergence behaviour of DR (resp. ADMM) when the involved functions (resp. their Legendre-Fenchel conjugates) are moreover partly smooth. More precisely, when both of the two functions (resp. their conjugates) are partly smooth relative to their respective manifolds, we show that DR (resp. ADMM) identifies these manifolds in finite time. Moreover, when these manifolds are affine or linear, we prove that DR/ADMM is locally linearly convergent with a rate in terms of the cosine of the Friedrichs angle between the tangent spaces of the identified manifolds. This is illustrated by several concrete examples and supported by numerical experiments.

Keywords: Douglas–Rachford splitting, ADMM, Partial Smoothness, Finite Activity Identification, Local Linear Convergence

1 Introduction

1.1 Problem formulation

In this work, we consider the problem of solving

$$\min_{x\in\mathbb{R}^n} J(x) + G(x), \tag{1}$$

where both J and G are in $\Gamma_0(\mathbb{R}^n)$, the class of proper, lower semi-continuous (lsc) and convex functions. We assume that $\mathrm{ri}\big(\mathrm{dom}(J)\big) \cap \mathrm{ri}\big(\mathrm{dom}(G)\big) \neq \emptyset$, where $\mathrm{ri}(C)$ is the relative interior of the nonempty convex set C, and $\mathrm{dom}(F)$ is the domain of the function F. We also assume that the set of minimizers is non-empty, and that these two functions are simple, meaning that their respective proximity operators, $\mathrm{prox}_{\gamma J}$ and $\mathrm{prox}_{\gamma G}$, $\gamma > 0$, are easy to compute, either exactly or to a very

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

good approximation. Problem (1) covers a large number of problems including those appearing in variational image processing (see Section 6).

An efficient and provably convergent algorithm to solve this class of problems is the Douglas–Rachford splitting method [16], which reads, in its relaxed form,

$$\begin{cases} v^{k+1} = \mathrm{prox}_{\gamma G}(2x^k - z^k), \\ z^{k+1} = (1-\lambda_k)z^k + \lambda_k\big(z^k + v^{k+1} - x^k\big), \\ x^{k+1} = \mathrm{prox}_{\gamma J} z^{k+1}, \end{cases} \tag{2}$$

for $\gamma > 0$, $\lambda_k \in]0,2]$ with $\sum_{k\in\mathbb{N}} \lambda_k(2-\lambda_k) = +\infty$. The fixed-point operator B_{DR} with respect to z^k takes the form

$$B_{\mathrm{DR}} \overset{\text{def.}}{=} \frac{1}{2}(\mathrm{rprox}_{\gamma G} \circ \mathrm{rprox}_{\gamma J} + \mathrm{Id}),$$
$$\mathrm{rprox}_{\gamma J} \overset{\text{def.}}{=} 2\mathrm{prox}_{\gamma J} - \mathrm{Id}, \quad \mathrm{rprox}_{\gamma G} \overset{\text{def.}}{=} 2\mathrm{prox}_{\gamma G} - \mathrm{Id}.$$

The proximity operator of a proper lsc convex function is defined, for $\gamma > 0$, as

$$\mathrm{prox}_{\gamma J}(z) = \mathrm{argmin}_{x\in\mathbb{R}^n} \frac{1}{2}\|x-z\|^2 + \gamma J(x).$$

Since the set of minimizers of (1) is assumed to be non-empty, so is the $\mathrm{Fix}(B_{\mathrm{DR}})$ since the former is nothing but $\mathrm{prox}_{\gamma J}\big(\mathrm{Fix}(B_{\mathrm{DR}})\big)$. See [3] for a more detailed account on DR in real Hilbert spaces.

Remark 1 *The DR algorithm is not symmetric w.r.t. the order of the functions J and G. Nevertheless, the convergence claims above hold true of course when this order is reversed in* (2)*. In turn, all of our statements throughout also extend to this case with minor adaptations. Note also that the standard DR only accounts for the sum of* 2 *functions. But extension to more than* 2 *functions is straightforward through a product space trick, see Section 5 for details.*

1.2 Contributions

Based on the assumption that both J and G are partly smooth relative to smooth manifolds, we show that DR identifies in finite time these manifolds. In plain words, this means that after a finite number of iterations, the iterates (x^k, v^k) lie respectively in the partial smoothness manifolds associated to J and G respectively. When these manifolds are affine/linear, we establish local linear convergence of DR. We show that the optimal convergence radius is given in terms of the cosine of the Friedrichs angle between the tangent spaces of the manifolds. We generalize these claims to the minimization of the sum of more than two functions. We finally exemplify our results with several experiments on variational signal and image processing.

It is important to note that our results readily apply to the alternating direction method of multipliers (ADMM), since it is well-known that ADMM is the DR method applied to the Fenchel dual problem of (1). More precisely, we only need to assume that the conjugates J^* and G^* are partly smooth. Therefore, to avoid unnecessary lengthy repetitions , we only focus in detail on the primal DR splitting method.

1.3 Relation to prior work

There are problem instances in the literature where DR was proved to converge locally linearly. For instance, in [16, Proposition 4], it was assumed that the "internal" function is strongly convex with a Lipschitz continuous gradient. This local linear convergence result was further investigated in [22, 24] under smoothness and strong convexity assumptions. On the other hand, for the Basis Pursuit (BP) problem, i.e. ℓ_1 minimization with an affine constraint, is considered in [9] and an eventual local linear convergence is shown in the absence of strong convexity. The author in [23] analyzes the local convergence behaviour of ADMM for quadratic or linear programs, and shows local linear convergence if the optimal solution is unique and the strict complementarity holds. This turns out to be a special case of our framework. For the case of two subspaces, linear convergence of DR with the optimal rate being the cosine of the Friedrichs angle between the subspaces is proved in [2]. Our results generalize those of [9, 23, 2] to a much larger class of problems. For the non-convex case, [4] considered DR method for a feasibility problem of a sphere intersecting a line or more generally a proper affine subset. Such feasibility problems with an affine subspace and a super-regular set (in the sense of [14]) with strongly regular intersection was considered in [11], and was generalized later to two (ε, δ)-regular sets with linearly regular intersection [25], see also [18] for an even more general setting. However, even in the convex case, the rate provided in [18] is nowhere near the optimal rate given by the Friedrichs angle.

1.4 Notations

For a nonempty convex set $C \subset \mathbb{R}^n$, $\mathrm{aff}(C)$ is its affine hull, $\mathrm{par}(C)$ is the subspace parallel to it. Denote P_C the orthogonal projector onto C and N_C its normal cone. For $J \in \Gamma_0(\mathbb{R}^n)$, denote ∂J its subdifferential and prox_J its proximity operator. Define the model subspace

$$T_x \stackrel{\text{def.}}{=} \mathrm{par}\big(\partial J(x)\big)^{\perp}.$$

It is obvious that $\mathrm{P}_{T_x}\big(\partial J(x)\big)$ is a singleton, and therefore defined as

$$e_x \stackrel{\text{def.}}{=} \mathrm{P}_{T_x}\big(\partial J(x)\big) = \mathrm{P}_{\mathrm{aff}(\partial J(x))}(0).$$

Suppose $\mathcal{M} \subset \mathbb{R}^n$ is a C^2-manifold around x, denote $\mathcal{T}_{\mathcal{M}}(x)$ the tangent space of $\mathcal{M}$ at $x \in \mathbb{R}^n$.

2 Partly Smooth Functions

2.1 Definition and main properties

Partial smoothness of functions was originally defined in [13], our definition hereafter specializes it to the case of proper lsc convex functions.

Definition 1 (Partly smooth function) *Let $J \in \Gamma_0(\mathbb{R}^n)$, and $x \in \mathbb{R}^n$ such that $\partial J(x) \neq \emptyset$. J is* partly smooth *at x relative to a set $\mathcal{M}$ containing x if*

(1) (Smoothness) *$\mathcal{M}$ is a C^2-manifold around x, $J|_{\mathcal{M}}$ is C^2 near x;*
(2) (Sharpness) *The tangent space $\mathcal{T}_{\mathcal{M}}(x)$ is T_x;*
(3) (Continuity) *The set–valued mapping ∂J is continuous at x relative to $\mathcal{M}$.*

The class of partly smooth functions at x relative to $\mathcal{M}$ is denoted as $\mathrm{PS}_x(\mathcal{M})$. *When $\mathcal{M}$ is an affine manifold, then $\mathcal{M} = x + T_x$, and we denote this subclass as* $\mathrm{PSA}_x(x + T_x)$. *When $\mathcal{M}$ is a linear manifold, then $\mathcal{M} = T_x$, and we denote this subclass as* $\mathrm{PSL}_x(T_x)$.

Capitalizing on the results of [13], it can be shown that, under mild transversality conditions, the set of lsc convex and partly smooth functions is closed under addition and pre-composition by a linear operator. Moreover, absolutely permutation-invariant convex and partly smooth functions of the singular values of a real matrix, i.e. spectral functions, are convex and partly smooth spectral functions of the matrix [7].

Examples of partly smooth functions that have become very popular recently in the signal processing, optimization, statistics and machine learning literature are ℓ_1, $\ell_{1,2}$, ℓ_∞, total variation (TV) and nuclear norm regularizations. In fact, the nuclear norm is partly smooth at a matrix x relative to the manifold $\mathcal{M} = \{x' : \mathrm{rank}(x') = \mathrm{rank}(x)\}$. The first four regularizers are all part of the class $\mathrm{PSL}_x(T_x)$.

We now define a subclass of partly smooth functions where the manifold is affine or linear and the vector e_x is locally constant.

Definition 2 *J belongs to the class* $\mathrm{PSS}_x(x+T_x)$ *(resp.* $\mathrm{PSS}_x(T_x)$*) if and only if* $J \in \mathrm{PSA}_x(x+T_x)$ *(resp. $J \in \mathrm{PSL}_x(T_x)$) and e_x is constant near x, i.e. there exists a neighbourhood U of x such that $\forall x' \in (x + T_x) \cap U$ (resp. $x' \in T_x \cap U$)*

$$e_{x'} = e_x.$$

The class of functions that conform with this definition is that of locally polyhedral functions [21, Section 6.5], which includes for instance the ℓ_1, ℓ_∞ norms and the anisotropic TV semi-norm that are widely used in signal and image processing, computer vision, machine learning and statistics. The indicator function of a polyhedral set is also in $\mathrm{PSS}_x(x + T_x)$ at each x in the relative interior of one of its faces relative to the affine hull of that face, i.e. $x + T_x = \mathrm{aff}(\text{Face of } x)$. Observe that for polyhedral functions, in fact, the subdifferential itself is constant along the partial smoothness subspace.

2.2 Proximity operator

This part shows that the proximity operator of a partly smooth function can be given in an implicit form.

Proposition 1 *Let $p \overset{\text{def.}}{=} \mathrm{prox}_{\gamma J}(x) \in \mathcal{M}$. Assume that $J \in \mathrm{PS}_p(\mathcal{M})$. Then for any point x near p, we have*

$$p = \mathrm{P}_{\mathcal{M}}(x) - \gamma e_p + o\big(\|x - p\|\big).$$

In particular, if $J \in \mathrm{PSA}_p(p + T_p)$ (resp. $J \in \mathrm{PSL}_p(T_p)$), then for any $x \in \mathbb{R}^n$, we have

$$p = \mathrm{P}_{p+T_p}(x) - \gamma e_p \quad (\textit{resp. } p = \mathrm{P}_{T_p}(x) - \gamma e_p).$$

Proof. We start with the following lemma whose proof can be found in [15].

Lemma 1 *Suppose that $J \in \mathrm{PS}_p(\mathcal{M})$. Then any point x near p has a unique projection $\mathrm{P}_{\mathcal{M}}(x)$, $\mathrm{P}_{\mathcal{M}}$ is C^1 around p, and thus*

$$\mathrm{P}_{\mathcal{M}}(x) - p = \mathrm{P}_{T_p}(x - p) + o\big(\|x - p\|\big).$$

Let's now turn to the proof of our proposition. We have the equivalent characterization

$$p = \mathrm{prox}_{\gamma J}(x) \iff x - p \in \gamma \partial J(p). \tag{3}$$

Projecting (3) on T_p and using Lemma 1, we get

$$\mathrm{P}_{T_p}(x-p) = \mathrm{P}_{\mathcal{M}}(x) - p + o\big(\|x-p\|\big) = \gamma e_p,$$

which is the desired result.

When $J \in \mathrm{PSA}_p(p+T_p)$, observe that $\mathrm{P}_{p+T_p}(x) = p + \mathrm{P}_{T_p}(x-p)$ for any $x \in \mathbb{R}^n$. Thus projecting again the monotone inclusion (3) on T_p, we get

$$\mathrm{P}_{T_p}(x-p) = \mathrm{P}_{p+T_p}(x) - p = \gamma e_p,$$

whence the claim follows. The linear case is immediate.

3 Activity Identification with Douglas–Rachford

In this section, we present the finite time activity identification of the DR method.

Theorem 1 (Finite activity identification) *Suppose that the DR scheme* (2) *is used to create a sequence* (z^k, x^k, v^k)*. Then* (z^k, x^k, v^k) *converges to* $(z^\star, x^\star, x^\star)$*, where* $z^\star \in \mathrm{Fix}(B_{\mathrm{DR}})$ *and* $x^\star$ *is a global minimizer of* (1)*. Assume that* $J \in \mathrm{PS}_{x^\star}(\mathcal{M}^J)$ *and* $G \in \mathrm{PS}_{x^\star}(\mathcal{M}^G)$*, and*

$$z^\star \in x^\star + \gamma\big(\mathrm{ri}\big(\partial J(x^\star)\big) \cap \mathrm{ri}\big(-\partial G(x^\star)\big)\big). \tag{4}$$

Then,

(1) *The DR scheme has the finite activity identification property, i.e. for all* k *sufficiently large,* $(x^k, v^k) \in \mathcal{M}^J \times \mathcal{M}^G$*.*
(2) *If* $G \in \mathrm{PSA}_{x^\star}(x^\star + T^G_{x^\star})$ *(resp.* $G \in \mathrm{PSL}_{x^\star}(T^G_{x^\star})$*), then* $v^k \in x^\star + T^G_{x^\star}$ *(resp.* $v^k \in T^G_{x^\star}$*), and in both cases* $T^G_{v^k} = T^G_{x^\star}$ *for all* k *sufficiently large.*
(3) *If* $J \in \mathrm{PSA}_{x^\star}(x^\star + T^J_{x^\star})$ *(resp.* $J \in \mathrm{PSL}_{x^\star}(T^J_{x^\star})$*), then* $x^k \in x^\star + T^J_{x^\star}$ *(resp.* $x^k \in T^J_{x^\star}$*), and in both cases* $T^J_{x^k} = T^J_{x^\star}$ *for all* k *sufficiently large.*

Proof. Standard arguments using that B_{DR} is firmly non-expansive allow to show that the iterates z^k converge globally to a fixed point $z^\star \in \mathrm{Fix}(B_{\mathrm{DR}})$, by interpreting DR as a relaxed Krasnosel'skiĭ-Mann iteration. Moreover, the shadow point $x^\star \overset{\text{def.}}{=} \mathrm{prox}_{\gamma J}(z^\star)$ is a solution of (1), see e.g. [3]. In turn, using non-expansiveness of $\mathrm{prox}_{\gamma J}$, and as we are in finite dimension, we conclude also that the sequence x^k converges to $x^\star$. This entails that v^k converges to $x^\star$ (by non-expansiveness of $\mathrm{prox}_{\gamma G}$).

Now (4) is equivalent to

$$\tfrac{z^\star - x^\star}{\gamma} \in \mathrm{ri}\big(\partial J(x^\star)\big) \quad \text{and} \quad \tfrac{x^\star - z^\star}{\gamma} \in \mathrm{ri}\big(\partial G(x^\star)\big). \tag{5}$$

(1) The update of x^{k+1} and v^{k+1} in (2) is equivalent to the monotone inclusions

$$\tfrac{z^{k+1} - x^{k+1}}{\gamma} \in \partial J(x^{k+1}) \quad \text{and} \quad \tfrac{2x^k - z^k - v^{k+1}}{\gamma} \in \partial G(v^{k+1})\ .$$

It then follows that

$$\operatorname{dist}\big(\tfrac{z^\star - x^\star}{\gamma}, \partial J(x^{k+1})\big) \leqslant \tfrac{1}{\gamma}\big(\|z^{k+1} - z^\star\| + \|x^{k+1} - x^\star\|\big) \to 0$$

and

$$\operatorname{dist}\big(\tfrac{x^\star - z^\star}{\gamma}, \partial G(v^{k+1})\big) \leqslant \tfrac{1}{\gamma}\big(\|z^{k} - z^\star\| + 2\|x^{k} - x^\star\| + \|v^{k+1} - x^\star\|\big) \to 0.$$

By assumption, $J \in \Gamma_0(\mathbb{R}^n)$ and $G \in \Gamma_0(\mathbb{R}^n)$, and thus are sub-differentially continuous at every point in their respective domains [20, Example 13.30], and in particular at $x^\star$. It then follows that $J(x^k) \to J(x^\star)$ and $G(v^k) \to G(x^\star)$. Altogether, this shows that the conditions of [10, Theorem 5.3] are fulfilled for J and G, and the finite identification claim follows.

(2) In this case, we have $v^k \in x^\star + T_{x^\star}^G$ (resp. $v^k \in T_{x^\star}^G$). Since G is partly smooth at $x^\star$ relative to $x^\star + T_{x^\star}^G$ (resp. $T_{x^\star}^G$), the sharpness property holds at all nearby points in $x^\star + T_{x^\star}^G$ (resp. $T_{x^\star}^G$) [13, Proposition 2.10]. Thus for k large enough, i.e. v^k sufficiently close to $x^\star$, we have indeed $\mathcal{T}_{x^\star + T_{x^\star}^G}(v^k) = T_{x^\star}^G = T_{v^k}^G$ as claimed.

(3) Similar to (2).

Remark 2

1. *Condition* (4) *can be interpreted as a non-degeneracy assumption. It can be viewed as a geometric generalization of the strict complementarity of non-linear programming. Such a condition is almost necessary for the finite identification of the partial smoothness active manifolds [8].*
2. *When the minimizer is unique, using the fixed-point set characterization of DR, it can be shown that condition* (4) *is also equivalent to* $z^\star \in \operatorname{ri}\big(\operatorname{Fix}(B_{\mathrm{DR}})\big)$.

4 Local Linear Convergence of Douglas–Rachford

Let us first recall the principal angles and the Friedrichs angle between two subspaces U and V, which are crucial for our quantitative analysis of the convergence rates. Without loss of generality, let $1 \leqslant p \overset{\text{def.}}{=} \dim(U) \leqslant q \overset{\text{def.}}{=} \dim(V) \leqslant n-1$.

Definition 3 (Principal angles) *The principal angles* $\theta_k \in [0, \frac{\pi}{2}]$, $k = 1, \ldots, p$ *between* U *and* V *are defined by, with* $u_0 = v_0 \overset{\text{def.}}{=} 0$

$$\cos\theta_k \overset{\text{def.}}{=} \langle u_k, v_k\rangle = \max\langle u, v\rangle \quad s.t. \quad \begin{array}{l} u \in U, v \in V, \|u\| = 1, \|v\| = 1, \\ \langle u, u_i\rangle = \langle v, v_i\rangle = 0, \ i = 0, \ldots, k-1. \end{array}$$

The principal angles θ_k *are unique with* $0 \leqslant \theta_1 \leqslant \theta_2 \leqslant \ldots \leqslant \theta_p \leqslant \pi/2$.

Definition 4 (Friedrichs angle) *The Friedrichs angle* $\theta_F \in]0, \frac{\pi}{2}]$ *between* U *and* V *is*

$$\cos\theta_F(U,V) \overset{\text{def.}}{=} \max\langle u, v\rangle \ s.t. \ \ u \in U \cap (U \cap V)^\perp, \|u\| = 1, \ v \in V \cap (U \cap V)^\perp, \|v\| = 1.$$

The following relation between the Friedrichs and principal angles is of paramount importance to our analysis, whose proof can be found in [1, Proposition 3.3].

Lemma 2 (Principal angles and Friedrichs angle) *The Friedrichs angle is exactly* θ_{d+1} *where* $d \overset{\text{def.}}{=} \dim(U \cap V)$. *Moreover,* $\theta_F(U,V) > 0$.

Remark 3 *One approach to obtain the principal angles is through the singular value decomposition (SVD). For instance, let $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{n\times q}$ form the orthonormal bases for the subspaces U and V respectively. Let $A\Sigma B^T$ be the SVD of $X^TY \in \mathbb{R}^{p\times q}$, then $\cos\theta_k = \sigma_k,\ k = 1, 2, ..., p$ and σ_k corresponds to the k'th largest singular value in Σ.*

We now turn to local linear convergence properties of DR. Let's denote $S^J_{x^\star} = \big(T^J_{x^\star}\big)^\perp$ and similarly for $S^G_{x^\star}$.

Theorem 2 (Local linear convergence) *Suppose that the DR scheme (2) is used with $\lambda_k \equiv \lambda \in]0, 2[$ to create a sequence (z^k, x^k, v^k) which converges to a pair $(z^\star, x^\star, x^\star)$ such that $J \in \mathrm{PSS}_{x^\star}(T^J_{x^\star})$ and $G \in \mathrm{PSS}_{x^\star}(T^G_{x^\star})$, and (4) holds. Then, there exists $K > 0$ such that for all $k \geqslant K$,*

$$\begin{aligned}\|(z^k - z^\star) - \mathrm{P}_{(T^J_{x^\star}\cap T^G_{x^\star})\oplus(S^J_{x^\star}\cap S^G_{x^\star})}(z^K - z^\star)\| &\leqslant \rho^{k-K}\|(\mathrm{Id} - \mathrm{P}_{(T^J_{x^\star}\cap T^G_{x^\star})\oplus(S^J_{x^\star}\cap S^G_{x^\star})})(z^K - z^\star)\| \\ &\leqslant \rho^{k-K}\|z^K - z^\star\|,\end{aligned} \tag{6}$$

with $\rho = \sqrt{(1-\lambda)^2 + \lambda(2-\lambda)\cos^2\theta_F\big(T^J_{x^\star}, T^G_{x^\star}\big)} \in [0, 1[$, and thus, $z^k - z^\star$ converges locally linearly to $\mathrm{P}_{(T^J_{x^\star}\cap T^G_{x^\star})\oplus(S^J_{x^\star}\cap S^G_{x^\star})}(z^K - z^\star)$ with the optimal rate ρ.

In particular, if $T^J_{x^\star} \cap T^G_{x^\star} = S^J_{x^\star} \cap S^G_{x^\star} = \{0\}$, then z^k converges locally linearly to $z^\star$ with the optimal rate $\sqrt{(1-\lambda)^2 + \lambda(2-\lambda)\cos^2\theta_1\big(T^J_{x^\star}, T^G_{x^\star}\big)} \in [0, 1[$.

This result is only valid for the class PSS. Extending this to general partly smooth functions is left to a forthcoming work.

Remark 4 *It can be observed that the best rate is obtained for $\lambda = 1$. This has been also pointed out in [9] for basis-pursuit. This assertion is however only on the local convergence behaviour and does not mean in general that the DR will be globally faster for $\lambda_k \equiv 1$. Note also that the above result can be straightforwardly generalized to the case of varying λ_k.*

Proof. We give the proof for the affine case, the linear one is similar. To lighten the notation, we will denote $M^\infty = \mathrm{P}_{(T^J_{x^\star}\cap T^G_{x^\star})\oplus(S^J_{x^\star}\cap S^G_{x^\star})}$ (the choice of this notation will be clearer shortly).

Combining Theorem 1(2)-(3), Proposition 1 and the definition of the class $\mathrm{PSS}_x(T_x)$, we get

$$\begin{aligned}x^k &= \mathrm{P}_{T^J_{x^\star}} z^k - \gamma e^J_{x^\star} + \mathrm{P}_{S^J_{x^\star}} x^\star, \\ v^{k+1} &= 2\mathrm{P}_{T^G_{x^\star}} x^k - \mathrm{P}_{T^G_{x^\star}} z^k - \gamma e^G_{x^\star} + \mathrm{P}_{S^G_{x^\star}} x^\star \\ &= 2\mathrm{P}_{T^G_{x^\star}}\mathrm{P}_{T^J_{x^\star}} z^k - \mathrm{P}_{T^G_{x^\star}} z^k - \gamma e^G_{x^\star} - 2\gamma\mathrm{P}_{T^G_{x^\star}} e^J_{x^\star} + 2\mathrm{P}_{T^G_{x^\star}}\mathrm{P}_{S^J_{x^\star}} x^\star + \mathrm{P}_{S^G_{x^\star}} x^\star.\end{aligned}$$

Similarly, we have

$$\begin{aligned}x^\star &= \mathrm{P}_{T^J_{x^\star}} z^\star - \gamma e^J_{x^\star} + \mathrm{P}_{S^J_{x^\star}} x^\star, \\ x^\star &= 2\mathrm{P}_{T^G_{x^\star}}\mathrm{P}_{T^J_{x^\star}} z^\star - \mathrm{P}_{T^G_{x^\star}} z^\star - \gamma e^G_{x^\star} - 2\gamma\mathrm{P}_{T^G_{x^\star}} e^J_{x^\star} + 2\mathrm{P}_{T^G_{x^\star}}\mathrm{P}_{S^J_{x^\star}} x^\star + \mathrm{P}_{S^G_{x^\star}} x^\star.\end{aligned}$$

Combining and rearranging the terms, we get

$$\begin{aligned}(z^k + v^{k+1} - x^k) - z^\star &= (z^k + v^{k+1} - x^k) - (z^\star + x^\star - x^\star) = (z^k - z^\star) + (v^{k+1} - x^\star) - (x^k - x^\star) \\ &= \big(\mathrm{Id} - \mathrm{P}_{T^J_{x^\star}} + 2\mathrm{P}_{T^G_{x^\star}}\mathrm{P}_{T^J_{x^\star}} - \mathrm{P}_{T^G_{x^\star}}\big)(z^k - z^\star) \\ &= (\mathrm{P}_{S^J_{x^\star}} - 2\mathrm{P}_{T^G_{x^\star}}\mathrm{P}_{S^J_{x^\star}} + \mathrm{P}_{T^G_{x^\star}})(z^k - z^\star) = (\mathrm{P}_{S^G_{x^\star}}\mathrm{P}_{S^J_{x^\star}} + \mathrm{P}_{T^G_{x^\star}}\mathrm{P}_{T^J_{x^\star}})(z^k - z^\star),\end{aligned}$$

whence we obtain

$$\begin{aligned}(z^{k+1} - z^\star) - M^\infty(z^K - z^\star) &= M(z^k - z^\star) - M^\infty(z^K - z^\star) \\ &= (M^{k+1-K} - M^\infty)(z^K - z^\star),\end{aligned}$$

where

$$M = (1-\lambda)\mathrm{Id} + \lambda(\mathrm{P}_{S^G_{x^\star}}\mathrm{P}_{S^J_{x^\star}} + \mathrm{P}_{T^G_{x^\star}}\mathrm{P}_{T^J_{x^\star}}).$$

It is immediate to check that M is normal. Moreover, combining [1, Theorem 3.10(ii)] and [2, Proposition 3.6(i)], M is convergent to $\mathrm{P}_{\mathrm{Fix}M} = M^\infty$ if $\lambda \in]0,2[$ (hence the choice of notation above). Thus, combining normality and [1, Theorem 2.16] we get that

$$\|M^{k+1-K} - M^\infty\| = \|M - M^\infty\|^{k+1-K}$$

and $\|M - M^\infty\|$ is the optimal convergence rate of M. Using together Lemma 2 and arguments similar to those of the proof of [2, Theorem 3.10(ii)] (see also [1, Theorem 4.1(ii)]), we get indeed that

$$\|M - M^\infty\| = \rho\ .$$

Finally,

$$\begin{aligned}\|(z^{k+1} - z^\star) - M^\infty(z^K - z^\star)\| &= \|(M^{k+1-K} - M^\infty)(z^K - z^\star)\| \\ &= \|(M^{k+1-K} - M^\infty)(\mathrm{Id} - M^\infty)(z^K - z^\star)\| \\ &\leqslant \|M^{k+1-K} - M^\infty\|\|(\mathrm{Id} - M^\infty)(z^K - z^\star)\| \\ &= \rho^{k+1-K}\|(\mathrm{Id} - M^\infty)(z^K - z^\star)\| \\ &\leqslant \rho^{k+1-K}\|z^K - z^\star\|,\end{aligned}$$

where we used the fact that $M^k M^\infty = M^k \mathrm{P}_{\mathrm{Fix}M} = \mathrm{P}_{\mathrm{Fix}M}$, and $\mathrm{Id} - M^\infty$ is an orthogonal projector, hence non-expansive.

The particular case is immedate. This concludes the proof.

5 Sum of more than two functions

We now want to tackle the problem of solving

$$\min_{x\in\mathbb{R}^n} \textstyle\sum_{i=1}^m J_i(x), \tag{7}$$

where each $J_i \in \Gamma_0(\mathbb{R}^n)$. We assume that all the relative interiors of their domains have a non-empty intersection, that the set of minimizers is non-empty, and that these functions are simple.

In fact, problem (7) can be equivalently reformulated as (1) in a product space, see e.g. [6, 19]. Let $\boldsymbol{\mathcal{H}} = \underbrace{\mathbb{R}^n \times \cdots \times \mathbb{R}^n}_{m \text{ times}}$ endowed with the scalar inner-product and norm

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{\mathcal{H}},\ \langle\!\langle \boldsymbol{x}, \boldsymbol{y} \rangle\!\rangle = \textstyle\sum_{i=1}^m \langle x_i,\, y_i\rangle,\ \|\boldsymbol{x}\| = \sqrt{\sum_{i=1}^m \|x_i\|^2}.$$

Let $\boldsymbol{\mathcal{S}} = \{\boldsymbol{x} = (x_i)_i \in \boldsymbol{\mathcal{H}} :\ x_1 = \cdots = x_m\}$ and its orthogonal complement $\boldsymbol{\mathcal{S}}^\perp = \big\{\boldsymbol{x} = (x_i)_i \in \boldsymbol{\mathcal{H}} : \sum_{i=1}^m x_i = 0\big\}$. Now define the canonical isometry,

$$\boldsymbol{C} : \mathbb{R}^n \to \boldsymbol{\mathcal{S}},\ x \mapsto (x, \cdots, x),$$

then we have $\mathrm{P}_{\boldsymbol{\mathcal{S}}}(\boldsymbol{z}) = \boldsymbol{C}\big(\frac{1}{m}\sum_{i=1}^m z_i\big)$.

Problem (7) is now equivalent to

$$\min_{\boldsymbol{x}\in\boldsymbol{\mathcal{H}}} \boldsymbol{J}(\boldsymbol{x}) + \boldsymbol{G}(\boldsymbol{x}), \quad \text{where} \quad \boldsymbol{J}(\boldsymbol{x}) = \sum_{i=1}^m J_i(x_i) \quad \text{and} \quad \boldsymbol{G}(\boldsymbol{x}) = \iota_{\boldsymbol{\mathcal{S}}}(\boldsymbol{x}). \tag{8}$$

Obviously, $\boldsymbol{J}$ is separable and therefore,

$$\mathrm{prox}_{\gamma\boldsymbol{J}}(\boldsymbol{x}) = \big(\mathrm{prox}_{\gamma J_i}(x_i)\big)_i.$$

Denote $\boldsymbol{T}^{\boldsymbol{J}}_{\boldsymbol{x}^\star} = \times_i T^{J_i}_{x^\star}$, and hence $\boldsymbol{S}^{\boldsymbol{J}}_{\boldsymbol{x}^\star} = (\boldsymbol{T}^{\boldsymbol{J}}_{\boldsymbol{x}^\star})^\perp = \times_i (T^{J_i}_{x^\star})^\perp$, where $\boldsymbol{x}^\star = \boldsymbol{C}(x^\star)$. We have the following result.

Corollary 1 *Suppose that the DR scheme is used to solve* (8) *and creates a sequence* $(\boldsymbol{z}^k, \boldsymbol{x}^k, \boldsymbol{v}^k)$. *Then* $(\boldsymbol{z}^k, \boldsymbol{x}^k, \boldsymbol{v}^k)$ *converges to* $(\boldsymbol{z}^\star, \boldsymbol{x}^\star, \boldsymbol{x}^\star)$, *and* $x^\star$ *is a minimizer of* (7). *Suppose that* $J_i \in \mathrm{PS}_{x^\star}(\mathcal{M}^{J_i})$ *and*

$$\boldsymbol{z}^\star \in \boldsymbol{x}^\star + \gamma\mathrm{ri}\big(\partial\boldsymbol{J}(\boldsymbol{x}^\star)\big) \cap \boldsymbol{\mathcal{S}}^\perp. \tag{9}$$

Then,

(1) *the DR scheme has the finite activity identification property, i.e. for all* k *sufficiently large,* $\boldsymbol{x}^k \in \times_i \mathcal{M}^{J_i}$.
(2) *Assume that* $J_i \in \mathrm{PSS}_{x^\star}(x^\star + T^{J_i}_{x^\star})$ *(or* $J_i \in \mathrm{PSS}_{x^\star}(T^{J_i}_{x^\star})$*) and DR is run with* $\lambda_k \equiv \lambda \in]0, 2[$. *Then, there exists* $K > 0$ *such that for all* $k \geqslant K$,

$$\|(\boldsymbol{z}^k - \boldsymbol{z}^\star) - \mathrm{P}_{(\boldsymbol{T}^{\boldsymbol{J}}_{x^\star}\cap\boldsymbol{\mathcal{S}})\oplus(\boldsymbol{S}^{\boldsymbol{J}}_{x^\star}\cap\boldsymbol{\mathcal{S}}^\perp)}(\boldsymbol{z}^K - \boldsymbol{z}^\star)\| \leqslant \rho^{k-K}\|\boldsymbol{z}^K - \boldsymbol{z}^\star\|,$$

with $\rho = \sqrt{(1-\lambda)^2 + \lambda(2-\lambda)\cos^2\theta_F\big(\boldsymbol{T}^{\boldsymbol{J}}_{x^\star}, \boldsymbol{\mathcal{S}}\big)} \in [0, 1[$, *and thus,* $\boldsymbol{z}^k - \boldsymbol{z}^\star$ *converges locally linearly to* $\mathrm{P}_{(\boldsymbol{T}^{\boldsymbol{J}}_{x^\star}\cap\boldsymbol{\mathcal{S}})\oplus(\boldsymbol{S}^{\boldsymbol{J}}_{x^\star}\cap\boldsymbol{\mathcal{S}}^\perp)}(\boldsymbol{z}^K - \boldsymbol{z}^\star)$ *at the optimal rate* ρ.

Proof.

(1) By the separability rule, $\boldsymbol{J} \in \mathrm{PS}_{\boldsymbol{x}^\star}(\times_i \mathcal{M}^{J_i}_{x^\star})$, see [13, Proposition 4.5]. We also have $\partial\boldsymbol{G}(\boldsymbol{x}^\star) = N_{\boldsymbol{\mathcal{S}}}(\boldsymbol{x}^\star) = \boldsymbol{\mathcal{S}}^\perp$. Thus $\boldsymbol{G} \in \mathrm{PS}_{\boldsymbol{x}^\star}(\boldsymbol{\mathcal{S}})$, i.e. $T^{\boldsymbol{G}}_{\boldsymbol{x}^\star} = \boldsymbol{\mathcal{S}}$. Then (9) is simply a specialization of condition (4) to problem (8). The claim then follows from Theorem 1(1).
(2) This is a direct consequence of Theorem 2.

6 Numerical experiments

Here, we illustrate our theoretical results on several concrete examples. This section is by no means exhaustive, and we only focus on the problems that we consider as representative in variational signal/image processing.

Affinely-constrained Polyhedral Minimization Let us now consider the affine-constrained minimization problem

$$\min_{x\in\mathbb{R}^n} J(x) \quad \text{subject to} \quad y = Ax, \tag{10}$$

where $A \in \mathbb{R}^{m\times n}$, and J is finite-valued polyhedral. We assume that the problem is feasible, i.e. the observation $y \in \mathrm{Im}(A)$. By identifying G with the indicator function of the affine constraint, it is immediate to see that $G = \iota_{\mathrm{Ker}(A)}(\cdot)$, which is polyhedral, hence belongs to PSS, and is simple.

Problem (10) is of important interest in various areas, including signal and image processing to find regularized solutions to linear equations. Typically, J is a regularization term intended to promote solutions conforming to some notion of simplicity/low-dimensional structure. One can think of instance of the active area of compressed sensing (CS) and sparse recovery.

We here solve (10) with J being either ℓ_1, ℓ_∞, and anisotropic TV regularizers. For all these cases, $J \in \Gamma_0(\mathbb{R}^n)$, is simple and $J \in \mathrm{PSS}_{x^\star}(T_{x^\star})$, where $T_{x^\star}$ can be easily computed, see e.g. [21]. In these experiments, A is drawn randomly from the standard Gaussian ensemble, i.e. CS scenario, with the following settings:

(a) ℓ_1-norm: $m = 32$ and $n = 128$, x_0 is 8-sparse;
(b) ℓ_∞-norm: $m = 120$ and $n = 128$, x_0 has 10 saturating entries;
(c) TV semi-norm: $m = 32$ and $n = 128$, (∇x_0) is 8-sparse;

For each setting, the number of measurements is sufficiently large so that one can prove that the minimizer $x^\star$ is unique, and in particular that $\mathrm{Ker}(A) \cap T_{x^\star} = \{0\}$ (with high probability). We also checked that $\mathrm{Im}(A^T) \cap S_{x^\star} = \{0\}$, which in this case is equivalent to uniqueness of the fixed point (see Remark 2(ii)). Thus (4) is obviously fulfilled, and the second part of Theorem 2 applies.

Figure 1(a)-(c) displays the global profile of $\|z^k - z^\star\|$ as a function of k, and the starting point of the solid line is the iteration number at which the partial smooth manifolds (here subspaces) are identified. One can easily see the linear convergence behaviour and that our rate estimate is indeed optimal.

TV based Image Inpainting In this image processing example, we observe $y = Ax_0$, where A is a binary mask operator. We aim at inpainting the missing regions from the observations y. This can be achieved by solving (10) with J the 2D anisotropic TV. The corresponding convergence profile is depicted in Figure 1(d).

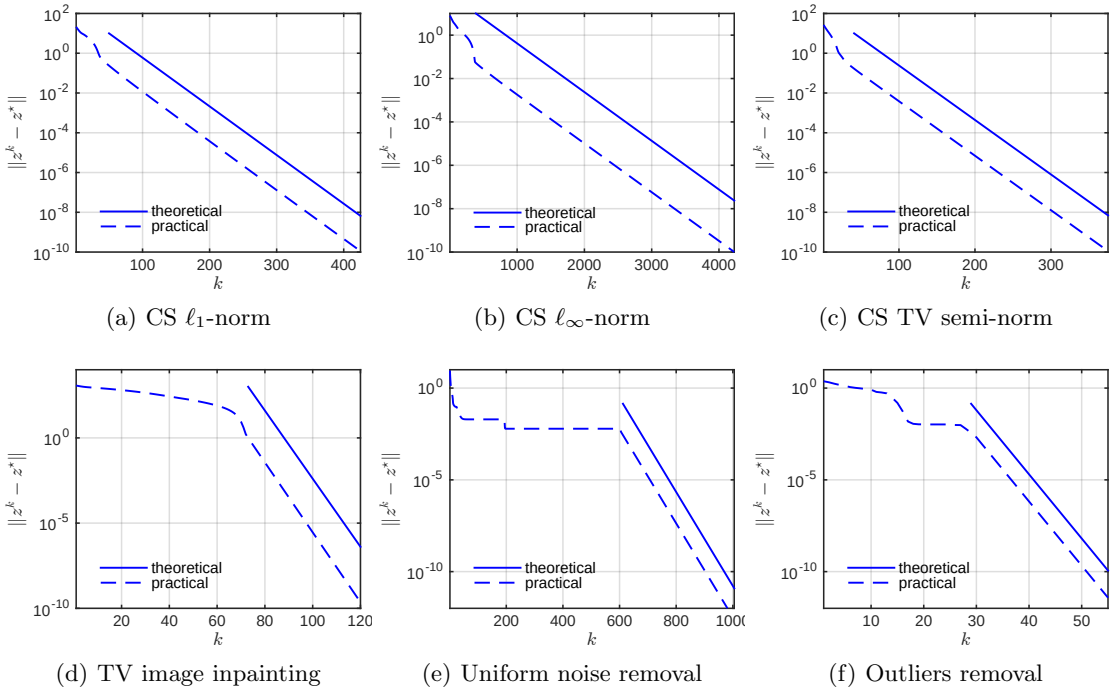
Uniform Noise Removal For this problem, we assume that we observe $y = x_0 + \varepsilon$, where x_0 is a piecewise-smooth vector, and ε is a realization of a random vector whose entries are iid $\sim \mathcal{U}([-a, a])$, $a > 0$. It is then natural to solve the problem

$$\min_{x\in\mathbb{R}^n} \|x\|_{\mathrm{TV}} \quad \text{subject to} \quad \|y - x\|_\infty \leqslant a. \tag{11}$$

G is now identified with the indicator function of the ℓ_∞-ball constraint, which is polyhedral and simple. The local convergence profile is shown in Figure 1(e) where we set $a = 1$ and $n = 100$. Again, the rate estimate is extremely tight.

Outliers Removal Consider solving

$$\min_{x\in\mathbb{R}^n} \|y - x\|_1 + \lambda\|x\|_{\mathrm{TV}}, \tag{12}$$



(a) CS ℓ_1-norm (b) CS ℓ_∞-norm (c) CS TV semi-norm

(d) TV image inpainting (e) Uniform noise removal (f) Outliers removal

Fig. 1. Observed (dashed) and predicted (solid) convergence profiles of DR (2) in terms of $\|z^k - z^\star\|$. (a) CS with ℓ_1. (b) CS with ℓ_∞. (c) CS with TV. (d) TV image inpainting. (e) Uniform noise removal by solving (11). (f) Outliers removal by solving (12). The starting point of the solid line is the iteration at which the manifolds are identified.

where $\lambda > 0$ is the tradeoff parameter. This problem has been proposed by [17] for outliers removal. We take $J = \lambda\|\cdot\|_{\mathrm{TV}}$ and $G = \|y - \cdot\|_1$, which is again simple and polyhedral. For this example we have $n = 100$, and $y - x$ is 10-sparse, the corresponding local convergence profile is depicted in Figure 1(f).

7 Conclusion

In this paper, we first showed that the DR splitting has the finite manifold identification under partial smoothness. When the involved manifolds are affine/linear and the generalized signs are locally constant, we proved local linear convergence of DR and provided a very tight rate estimate as illustrated by several numerical experiments. Our future work will focus on extending the linear convergence result to more general partly smooth functions.

References

1. Bauschke, H.H., Bello Cruz, J.Y., Nghia, T.A., Phan, H.M., Wang, X.: Optimal rates of convergence of matrices with applications. arxiv:1407.0671, (2014).

2. Bauschke, H.H., Bello Cruz, J.Y., Nghia, T.A., Phan, H.M., Wang, X.: The rate of linear convergence of the Douglas–Rachford algorithm for subspaces is the cosine of the friedrichs angle. J. of Approx. Theo., 185:63–79, (2014).
3. Bauschke, H.H., Combettes, P.L.: Convex analysis and monotone operator theory in Hilbert spaces. Springer (2011).
4. Borwein, J.M., Sims, B.: The Douglas–Rachford algorithm in the absence of convexity. (2010).
5. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. of Math. Imag. and Vis., 40(1):120–145, (2011).
6. Combettes, P.L., Pesquet, J.C.: A proximal decomposition method for solving convex variational inverse problems. Inv. Prob., 24(6):065014, (2008).
7. Daniilidis, A., Drusvyatskiy, D., Lewis, A.L.: Orthogonal invariance and identifiability. SIAM Mat. Anal. Appl., (2014).
8. Hare, W.L., Lewis, A.L.: Identifying active manifolds. Alg. Op. Res., 2(2):75–82, (2007).
9. Demanet, L., Zhang, X.: Eventual linear convergence of the Douglas–Rachford iteration for basis pursuit. Math. Prog., (2013).
10. Hare, W.L., Lewis, A.S.: Identifying active constraints via partial smoothness and prox-regularity. J. of Conv. Ana., 11(2):251–266, (2004).
11. Hesse, H., Luke, D.R., Neumann, P.: Alternating Projections and Douglas–Rachford for Sparse Affine Feasibility. IEEE Trans. on Sig. Proc., 62(18):4868–4881, (2014).
12. Hesse, H., Luke, D.R.: Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems. SIAM J. Opt., 23(4):2397–2419, (2013).
13. Lewis, A.S.: Active sets, nonsmoothness, and sensitivity. SIAM J. Opt., 13(3):702–725, (2003).
14. Lewis, A.S., Luke, D.R., Malick, J.: Local linear convergence for alternating and averaged nonconvex projections. Found. Comput. Math., 9(4):485–513, (2009).
15. Liang, J., Fadili, M.J., Peyré, G.: Local linear convergence of Forward–Backward under partial smoothness. NIPS, 1970–1978, (2014).
16. Lions, P.L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. on Num. Ana., 16(6):964–979, (1979).
17. Nikolova, M.: A variational approach to remove outliers and impulse noise. J. of Math. Imag. and Vis., 20(1-2):99–120, (2004).
18. Phan, H.M.: Linear convergence of the Douglas–Rachford method for two closed sets. arXiv:1401.6509v1, (2014).
19. Raguet, H., Fadili, J.M., Peyré, G.: Generalized Forward–Backward splitting. SIAM Im. Sciences, 6(3):1199–1226, (2013).
20. Rockafellar, R.T., Wets, R.: Variational analysis, V317. Springer Verlag, (1998).
21. Vaiter, S., Golbabaee, M., Fadili, M.J., Peyré, G.: Model selection with low complexity priors. Preprint Hal, (2013).
22. D. Davis and W. Yin.: Convergence rates of relaxed Peaceman–Rachford and ADMM under regularity assumptions, arXiv preprint arXiv:1407.5210, 2014.
23. D. Boley.: Local linear convergence of the alternating direction method of multipliers on quadratic or linear programs, SIAM Journal on Optimization, 23(4):2183–2207, 2013.
24. P. Giselsson and S. Boyd.: Metric selection in Douglas–Rachford splitting and ADMM, arXiv preprint arXiv:1410.8479, 2014.
25. R. Hesse and D. R. Luke.: Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems, SIAM Journal on Optimization, 23(4):2397–2419, 2013.