Linear Convergence of the ADMM/Douglas Rachford Algorithms for Piecewise Linear-Quadratic Functions and Application to Statistical Imaging

Timo Aspelmeier[*], C. Charitha[†] and D. Russell Luke[‡]

version August 18, 2015

Abstract

We consider the problem of minimizing the sum of a convex, piecewise linear-quadratic function and a convex piecewise linear-quadratic function composed with an injective linear mapping. We show that, for such problems, iterates of the alternating directions method of multipliers converge linearly to fixed points from which the solution to the original problem can be computed. Our proof strategy uses duality and strong metric subregularity of the Douglas-Rachford fixed point mapping. Our analysis does not require strong convexity and yields error bounds to the set of model solutions. We demonstrate an application of this result to exact penalization for signal deconvolution and denoising with multiresolution statistical constraints.

Research of T. Aspelmeier, C. Charitha and D. R. Luke was supported in part by the German Research Foundation grant SFB755-A4.

2010 Mathematics Subject Classification: Primary 49J52, 49M20, 90C26; Secondary 15A29, 47H09, 65K05, 65K10, 94A08.
Keywords: exact penalization, image processing, inverse problems, multiscale analysis

1 Introduction.

The alternating directions method of multipliers method (ADMM) has received a great deal of attention recently for problems involving constraints on the image of the unknowns under some linear mapping or for regularized linear inverse problems. The analysis has focused on either global complexity estimates [29] or sufficient conditions

[*]Institut für Mathematische Stochastics, Universität Göttingen, 37077 Göttingen, Germany
[†]Institut für Numerische und Angewandte Mathematik, Universität Göttingen, 37083 Göttingen, Germany
[‡]Institut für Numerische und Angewandte Mathematik, Universität Göttingen, 37083 Göttingen, Germany

for local linear convergence [22, 39]. The closely related Douglas-Rachford algorithm has also been the focus of recent studies showing global complexity [34, 40] and (local linear) convergence in increasingly complex settings [1, 2, 4–6, 11, 30, 31, 42]. In the convex setting, the convergence studies for both ADMM and Douglas-Rachford share a common thread through the well-known duality between these algorithms [26]. Studies of the ADMM frequently invoke strong convexity. Studies of Douglas-Rachford, on the other hand have, until very recently, been focused on *feasibility* problems and corresponding notions of regularity of intersections. In the present work, we combine an analysis of the ADMM algorithm with facts learned from the local convergence of Douglas-Rachford to provide sufficient conditions for local linear convergence of sequences generated by ADMM without strong convexity. Our theoretical development is specialized to the application of statistical multiscale image denoising/deconvolution following [24].

1.1 Notation and definitions

Though many of the arguments presented here work equally well for infinite dimensional Hilbert spaces, to avoid technicalities, *it will be assumed throughout that U and V are Euclidean spaces.* We denote the extended reals by $(-\infty, +\infty] := \mathbb{R} \cup \{+\infty\}$ and the nonnegative orthant by $\mathbb{R}_+ := \{x \in \mathbb{R} \mid x \geq 0\}$. The closed unit ball centered at the origin is denoted by $\mathbb{B}$. In the usual notation for the natural numbers $\mathbb{N}$ we include 0. The mapping $A : U \to V$ is linear and the functional $J : U \to (-\infty, +\infty]$ is proper (not everywhere $+\infty$ and nowhere $-\infty$), convex and lower semicontinuous (lsc), as is the functional $H : V \to (-\infty, +\infty]$. A proper function $f : U \to (-\infty, +\infty]$ is *strongly convex* if there is a constant $\mu > 0$ such that

$$f\left((1-\tau)x_0 + \tau x_1\right) \leq (1-\tau)f(x_0) + \tau f(x_1) - \tfrac{1}{2}\mu\tau(1-\tau)\|x_0 - x_1\|^2 \tag{1.1}$$

for all x_0 and x_1 and $\tau \in (0, 1)$. We will not assume smoothness of functions and so will require the *subdifferential*. The subdifferential of a function $f : U \to (-\infty, +\infty]$ at a point $\overline{x} \in \operatorname{dom} f$ is defined by

$$\partial f(\overline{x}) := \{v \in U \mid \langle v, x - \overline{x}\rangle \leq f(x) - f(\overline{x}), \text{ for all } x \in U\}. \tag{1.2}$$

When $\overline{x} \notin \operatorname{dom} f$ the subdifferential is defined to be empty. Elements from the subdifferential are called *subgradients*. The subdifferential of a proper, lsc convex function is a *maximally monotone* set-valued mapping [44, Theorem 12.17]. The *Fenchel conjugate* of a function f is denoted by f^* and defined by

$$f^*(y) := \inf\{\langle y,\ x\rangle - f(x)\}.$$

We use the notation $\Phi : U \rightrightarrows V$ to denote a set-valued mapping Φ from U to V. A set valued mapping $\Phi : U \rightrightarrows V$ is said to be *β-inverse strongly monotone* [44, Corollary 12.55] if for all $x, x' \in U$

$$\langle v - v', x - x'\rangle \geq \beta|v - v'|^2, \quad \text{when ever} \quad v \in \Phi(x), v' \in \Phi(x'). \tag{1.3}$$

We denote the *resolvent* of a (set-valued) mapping $\Phi : U \rightrightarrows V$ by $\mathcal{J}_\Phi := (\mathrm{Id} + \Phi)^{-1}$ where Id denotes the identity mapping and the inverse is defined as

$$\Phi^{-1}(y) := \{x \in U \mid y \in \Phi(x)\}. \tag{1.4}$$

The corresponding *reflector* is defined by $R_{\eta\Phi} := 2\mathcal{J}_{\eta\Phi} - \mathrm{Id}$. A set-valued mapping $\Phi : U \rightrightarrows V$ is said to be *Lipschitz continuous* if for all $u, u' \in U$ there exists a $\tau \geq 0$ such that

$$\Phi(u') \subset \Phi(u) + \tau\|u' - u\|\mathbb{B}. \tag{1.5}$$

Φ is said to be a *polyhedral* map (or piecewise polyhedral [44]) if its graph is the union of finitely many sets that are polyhedral convex in $U \times V$ [17]. A key property of set-valued mappings that we will rely on is *metric subregularity* [17, Exercise 3H.4].

Definition 1.1 ((strong) metric subregularity)**.**

(i) The mapping $\Phi : U \rightrightarrows V$ is called metrically subregular at $\overline{x}$ for $\overline{y}$ *if* $(\overline{x}, \overline{y}) \in \operatorname{gph}\Phi$ *and there is a constant $c > 0$ and neighborhoods $\mathcal{O}$ of $\overline{x}$ such that*

$$\operatorname{dist}(x, \Phi^{-1}(\overline{y})) \leq c \operatorname{dist}(\overline{y}, \Phi(x)) \;\forall\; x \in \mathcal{O}. \tag{1.6}$$

(ii) The mapping Φ is called strongly metrically subregular at $\overline{x}$ for $\overline{y}$ *if* $(\overline{x}, \overline{y}) \in \operatorname{gph}\Phi$ *and there is a constant $c > 0$ and neighborhoods $\mathcal{O}$ of $\overline{x}$ such that*

$$\|x - \overline{x}\| \leq c \operatorname{dist}(\overline{y}, \Phi(x)) \;\forall\; x \in \mathcal{O}. \tag{1.7}$$

The constant c measures the stability under perturbations of inclusions $\overline{y} \in \Phi(\overline{x})$.

Proposition 1.2 (polyhedrality implies strong metric subregularity)**.** *Let $W \subset V$ be an affine subspace and $T : W \rightrightarrows W$. If T is polyhedral and $\mathsf{Fix}\, T \cap W$ is an isolated point, $\{\overline{x}\}$, then $\mathrm{Id} - T : W \to (W - \overline{x})$ is strongly metrically subregular, hence metrically subregular, at $\overline{x}$ for 0.*

Proof. If T is polyhedral, so is $\Phi^{-1} := (\mathrm{Id} - T)^{-1}$. Now by [17, Propositions 3I.1 and 3I.2], since Φ^{-1} is polyhedral and $\overline{x}$ is an isolated point of $\Phi^{-1}(0)$, then $\Phi = \mathrm{Id} - T$ is *strongly metrically subregular* at $\overline{x}$ for 0 with constant c on the neighborhood $\mathcal{O}$ of $\overline{x}$ relative to W ((1.7)). □

One prevalent source of polyhedral mappings is the subdifferential of piecewise linear-quadratic functions (see Proposition 2.5 below).

Definition 1.3 (piecewise linear-quadratic functions)**.** *A function $f : \mathbb{R}^n \to [\infty, +\infty]$ is called* piecewise linear-quadratic *if $\operatorname{dom} f$ can be represented as the union of finitely many polyhedral sets, relative to each of which $f(x)$ is given by an expression of the form $\frac{1}{2}\langle x, Ax\rangle + \langle a, x\rangle + \alpha$ for some scalar $\alpha \in \mathbb{R}$ vector $a \in \mathbb{R}^n$, and symmetric matrix $A \in \mathbb{R}^{n\times n}$.*

1.2 Preparatory abstract results

To conclude this section we present general results about types of (firmly) nonexpansive operators that clarify the underlying mechanisms yielding linear convergence of many algorithms. The operative definitions are given here.

Definition 1.4 ((S,ϵ)-(firmly-)nonexpansive mappings)**.** *Let D and S be nonempty subsets of U and let T be a (multi-valued) mapping from D to U.*

(i) T is called (S,ε)-nonexpansive on D *if*

$$\begin{aligned}&\|x_+ - \overline{x}_+\| \le \sqrt{1+\varepsilon}\,\|x-\overline{x}\| \\ &\forall x\in D,\ \forall \overline{x}\in S,\ \forall x_+\in Tx,\ \forall \overline{x}_+\in T\overline{x}.\end{aligned} \tag{1.8}$$

If (1.8) holds with $\epsilon = 0$ then we say that T is S-nonexpansive *on D.*

(ii) T is called (S,ε)-firmly nonexpansive *on D if*

$$\begin{aligned}&\|x_+ - \overline{x}_+\|^2 + \|(x-x_+) - (\overline{x}-\overline{x}_+)\|^2 \le (1+\varepsilon)\,\|x-\overline{x}\|^2 \\ &\forall x\in D,\ \forall \overline{x}\in S,\ \forall x_+\in Tx,\ \forall \overline{x}_+\in T\overline{x}.\end{aligned} \tag{1.9}$$

If (1.9) holds with $\epsilon = 0$ then we say that T is S-firmly nonexpansive *on D. If, in addition, $S = \mathsf{Fix}\, T$, then T is said to be* quasi-firmly nonexpansive*.*

Theorem 1.5 (abstract linear convergence result)**.** *Let $W \subset V$ be an affine subspace and $T : W \rightrightarrows W$ be quasi-firmly nonexpansive on W. Let $\mathsf{Fix}\, T \cap W$ be an isolated point, $\{\overline{x}\}$. If $\mathrm{Id} - T : W \to (W - \overline{x})$ is metrically subregular at $\overline{x}$ for 0, then there is a neighborhood $\mathcal{O} \subset W$ of $\overline{x}$ such that*

$$\mathrm{dist}\,(x_+, \mathsf{Fix}\, T) \le \sqrt{1-\kappa}\,\mathrm{dist}\,(x, \mathsf{Fix}\, T) \quad \forall x_+\in Tx,\ \forall x\in\mathcal{O}, \tag{1.10}$$

where $0 < \kappa = c^{-2}$ for c a constant of metric subregularity of $\mathrm{Id} - T$ at $\overline{x}$ for the neighborhood $\mathcal{O}$. Consequently, the fixed point iteration $x^{k+1} = Tx^k$ converges linearly to $\mathsf{Fix}\, T$ with rate $\sqrt{1-\kappa}$ for all $x^0\in\mathcal{O}$.

Proof. Define $\Phi := (\mathrm{Id} - T)$ and note that $\{\overline{x}\} = (\mathrm{Id} - T)^{-1}(0) \iff \{\overline{x}\} = \mathsf{Fix}\, T$, hence

$$\mathrm{dist}\,(x, (\mathrm{Id} - T)^{-1}(0)) = \mathrm{dist}\,(x, \mathsf{Fix}\, T) = \|x-\overline{x}\|.$$

Suppose that Φ is metrically subregular at $\mathsf{Fix}\, T$ for 0. Then by Definition 1.1(i) we have, for all $x\in\mathcal{O}\subset W$ and for all $x^+\in T(x)$,

$$\mathrm{dist}\,(x, (\mathrm{Id} - T)^{-1}(0)) = \|x-\overline{x}\| = \ \le\ c\,\mathrm{dist}\,\left(0, (x-Tx)\right) \le c\|x-x^+\|, \tag{1.11}$$

which is the coercivity condition of [30, Eq.(3.1), Lemma 3.1]. By assumption, T is $(\mathsf{Fix}\, T, 0)$-firmly nonexpansive (i.e., quasi-firmly nonexpansive) on $W\supset\mathcal{O}$ (Definition 1.4 (ii)). The result then follows from [30, Lemma 3.1] with rate $\sqrt{1-\kappa}$ for $\kappa = c^{-2}$. ∎

Corollary 1.6 (Polyhedrality implies linear convergence)**.** *Let $W \subset V$ be an affine subspace and $T : W \rightrightarrows W$ be quasi-firmly nonexpansive on W. Let $\mathsf{Fix}\, T \cap W$ be an isolated point, $\{\overline{x}\}$. If T is polyhedral, then there is a neighborhood $\mathcal{O} \subset W$ of $\overline{x}$ such that*

$$\mathrm{dist}\,(x_+, \mathsf{Fix}\, T) \le \sqrt{1-\kappa}\, \mathrm{dist}\,(x, \mathsf{Fix}\, T) \quad \forall x_+ \in Tx,\ \forall x \in \mathcal{O},$$

where $0 < \kappa = c^{-2}$ for c a constant of metric subregularity of $\mathrm{Id} - T$ at $\overline{x}$ for the neighborhood $\mathcal{O}$. Consequently, the fixed point iteration $x^{k+1} = Tx^k$ converges linearly to $\mathsf{Fix}\, T$ with rate $\sqrt{1-\kappa}$ for all $x^0 \in \mathcal{O}$.

Proof. The result follows immediately from Proposition 1.2 and Theorem 1.5. □

The requirement of uniqueness is common in the inverse problems literature. It is well known, however, that, even if the solution to (2.7) is unique, the set of fixed points of the Douglas Rachford operator T need not be a singleton [36]. Recent work has shown, however, that such uniqueness need only hold on appropriate affine subspaces where the iterates lie [31, 42]. This feature has been exploited in the analysis of the Douglas-Rachford algorithm applied to problems with polyhedral structure [35]. Metric (sub)regularity, on the other hand, is one of the central assumptions of well-posedness of inverse problems [18, 32]. Other useful equivalent characterizations of metric subregularity can be found in [17]. Polyhedrality can be quite easy to verify, as we will see below.

2 Linear Convergence of Douglas-Rachford/ Alternating Directions Method of Multipliers

We consider problems in the following format:

$$\underset{u \in U}{\text{minimize}}\ J(u) + H(Au). \tag{$\mathcal{P}'$}$$

There are many possibilities for solving such problems. We focus our attention on one of the more prevalent methods, the alternating direction method of multipliers (primary sources include [20, 21, 26, 27, 43]). This method is one of many *splitting methods* which are the principle approach to handling the computational burden of large-scale, separable problems [12]. Introducing a new variable $v \in V$, our problem is to solve

$$\underset{(u,v) \in U \times V}{\text{minimize}}\ J(u) + H(v), \qquad \text{subject to } Au = v. \tag{2.1}$$

The augmented Lagrangian $\widetilde{L}$ for (2.1) is given by

$$\widetilde{L}(u, v, b) = J(u) + H(v) + \langle b, Au - v\rangle + \tfrac{\eta}{2}\|Au - v\|^2, \tag{2.2}$$

where $b \in V$, $\eta > 0$ is a fixed penalty parameter. The alternating directions method of multipliers for solving (2.1) is, given (u^k, v^k, b^k), $k \in \mathbb{N}$, compute $(u^{k+1}, v^{k+1}, b^{k+1})$

by

$$\begin{aligned}
u^{k+1} &= \operatorname{argmin}_u \left\{ J(u) + \tfrac{\eta}{2}\|Au - v^k + \eta^{-1}b^k\|^2 \right\}; && (2.3)\\
v^{k+1} &= \operatorname{argmin}_v \left\{ H(v) + \tfrac{\eta}{2}\|Au^{k+1} - v + \eta^{-1}b^k\|^2 \right\}; && (2.4)\\
b^{k+1} &= b^k + \eta(Au^{k+1} - v^{k+1}). && (2.5)
\end{aligned}$$

Using $\frac{\eta}{2}\|Au - v + \eta^{-1}b^k\|^2 - \frac{1}{2\eta}\|b^k\|^2 = \langle b^k, Au - v\rangle + \frac{\eta}{2}\|Au - b^k\|^2$, the algorithm (2.3)-(2.5) can be written equivalently as

Algorithm 2.1 (Alternating Directions Method of Multipliers).
Initialization. *Choose* $\eta > 0$ *and* $(v^0, b^0) \in U \times V \times V$.
General Step $(k = 0, 1, \ldots)$

$$\begin{aligned}
u^{k+1} &= \operatorname{argmin}_u \left\{ J(u) + \langle b^k, Au\rangle + \tfrac{\eta}{2}\|Au - v^k\|_2^2 \right\}; && (2.6a)\\
v^{k+1} &= \operatorname{argmin}_v \left\{ H(v) - \langle b^k, v\rangle + \tfrac{\eta}{2}\|Au^{k+1} - v\|_2^2 \right\}; && (2.6b)\\
b^{k+1} &= b^k + \eta(Au^{k+1} - v^{k+1}). && (2.6c)
\end{aligned}$$

The penalty parameter η need not be a constant, and indeed evidence indicates that the choice of η can greatly impact the complexity of the algorithm, but this is beyond the scope of this investigation, so we have left this parameter constant. It is well known [20,26] that the alternating directions method of multipliers algorithm can be derived from the Douglas Rachford algorithm, and vice verse, and therefore sufficient conditions for convergence of Douglas-Rachford also apply here. Our goal is to determine the rate of convergence of these algorithms so that they may be used as inner routines in an iteratively regularized procedure. Knowing that an algorithm converges linearly, for instance, yields rational stopping criteria with computable estimates for the distance of the current iterate to the solution set.

We present sufficient conditions for *linear* convergence of Algorithm 2.1. The first convergence result was by Lions and Mercier [36], in the dual setting under the assumption of strong convexity and Lipschitz continuity of J. Recent published work in this direction includes [29]. Convergence rate of order $\frac{1}{k}$ was established in [29,46] which is pessimistic compared to our rate $\frac{1}{k^2}$ (for a piecewise linear quadratic map J). A faster convergence rate of order $\frac{1}{k^2}$ due to [46] was under the assumption that J is quadratic. Our strategy is to consider the dual to the alternating directions method of multipliers algorithm, the Douglas-Rachford algorithm, which is more amenable to the tools of abstract fixed point theory presented in Section 1.2. In the first main result, Theorem 2.2, we describe two conditions that guarantee linear convergence of alternating directions method of multipliers . The first of these conditions follows from classical results of Lions and Mercier [36]. The second condition is based on work of more recent vintage [30], and is much more prevalent in applications.

The (Fenchel-Legendre) dual problem corresponding to the problem ($\mathcal{P}'$) is (see, for instance [10])

$$\min_{w\in V} J^*(A^*w) + H^*(-w).$$

Here J^* and H^* are the Fenchel conjugates of J and H respectively. Instead of working with this dual, we work with the following equivalent form with the change of variable

$v = -w$:

$$\min_{v \in V} J^*(-A^* v) + H^*(v). \tag{$\mathcal{D}'$}$$

Under the assumption that the solutions $\overline{u}$ and $\overline{b}$ of the primal and dual problems exist and that the dual gap is zero, the following two inclusions characterize the solutions of the problems ($\mathcal{P}'$) and ($\mathcal{D}'$) respectively:

$$0 \in \partial J(\overline{u}) + \partial (H \circ A)(\overline{u});$$

$$0 \in \partial \left(J^* \circ (-A^*)\right)(\overline{b}) + \partial H^*(\overline{b}).$$

In both cases, one has to solve an inclusion of the form

$$0 \in B + D \tag{2.7}$$

for general set-valued mappings B and D. For any $\eta > 0$, the Douglas Rachford algorithm [19, 36] for solving (2.7) is given by

$$\begin{aligned} b^{k+1} &\in T' b^k \quad (k \in \mathbb{N}) && (2.8)\\ \text{for} \qquad T' &:= \mathcal{J}_{\eta D}\left(\mathcal{J}_{\eta B}(\mathrm{Id} - \eta D) + \eta D\right), && (2.9) \end{aligned}$$

where $\mathcal{J}_{\eta D}$ and $\mathcal{J}_{\eta B}$ are the *resolvents* of ηD and ηB respectively. The connection between the alternating directions method of multipliers algorithm (2.6a)-(2.6c) and the Douglas-Rachford algorithm (2.8) was first discovered by Gabay [26] and is rederived for convenience in Appendix 4.2.

Given b^0 and $v^0 \in Db^0$, following [45], define the new variable $x^0 := b^0 + \eta v^0$ so that $b^0 = \mathcal{J}_{\eta D} x^0$. We thus arrive at an alternative formulation of the Douglas-Rachford algorithm (2.8):

$$\begin{aligned} x^{k+1} &= T x^k \quad (k \in \mathbb{N}) && (2.10)\\ \text{for} \qquad T &:= \tfrac{1}{2}(R_{\eta B} R_{\eta D} + \mathrm{Id}) = \mathcal{J}_{\eta B}(2\mathcal{J}_{\eta D} - \mathrm{Id}) + (\mathrm{Id} - \mathcal{J}_{\eta D}), && (2.11) \end{aligned}$$

where $R_{\eta D}$ and $R_{\eta B}$ are the reflectors of the respective resolvents. This is exactly the form of Douglas-Rachford considered in [36]. Note that for our application $B := \partial\left(J^* \circ (-A^*)\right)$ and $D := \partial H^*$, and so the resolvent mappings are the proximal mappings of the convex functions $(J^* \circ (-A^*))$ and H^* respectively, and hence the resolvent mappings are single-valued [38].

Proposition 2.1. *Let $J : U \to \mathbb{R} \cup \{+\infty\}$ and $H : V \to \mathbb{R}$ be proper, lsc and convex. Let $A : U \to V$ be linear and suppose there exists a solution to $0 \in B + D$, for $B := \partial\left(J^* \circ (-A^*)\right)$ and $D := \partial H^*$. For fixed $\eta > 0$, given any initial points x^0 and $(b^0, v^0) \in \operatorname{gph} D$ such that $x^0 = b^0 + \eta v^0$, the sequences $\left(b^k\right)_{k \in \mathbb{N}}$, $\left(x^k\right)_{k \in \mathbb{N}}$ and $\left(v^k\right)_{k \in \mathbb{N}}$ defined respectively by (2.8), (2.10) and $v^k := \frac{1}{\eta}\left(x^k - b^k\right)$ converge to points $\overline{b} \in \mathsf{Fix}\, T'$, $\overline{x} \in \mathsf{Fix}\, T$ and $\overline{v} \in D\left(\mathsf{Fix}\, T'\right)$. The point $\overline{b} = \mathcal{J}_{\eta D}\overline{x}$ is a solution to ($\mathcal{D}'$), and $\overline{v} = \frac{1}{\eta}\left(\overline{x} - \overline{b}\right) \in D\overline{b}$. If, in addition, A has full column rank, then the sequence $\left(b^k, v^k\right)_{k \in \mathbb{N}}$ corresponds exactly to the sequence of points generated in steps (2.6b) and (2.6c) of Algorithm 2.1 and the sequence $\left(u^{k+1}\right)_{k \in \mathbb{N}}$ generated by (2.6a) converges to $\overline{u}$, a solution to ($\mathcal{P}'$).*

Proof. Following [20, 45], we rewrite the Douglas Rachford iteration 2.8 in two steps: Given $(b^0, v^0) \in \operatorname{gph} D$, for $k \in \mathbb{N}$ do

$$\text{find } (q^{k+1}, s^{k+1}) \in \operatorname{gph}(B) \text{ such that } q^{k+1} + \eta s^{k+1} = b^k - \eta v^k; \tag{2.12a}$$
$$\text{find } (b^{k+1}, v^{k+1}) \in \operatorname{gph}(D) \text{ such that } b^{k+1} + \eta v^{k+1} = q^{k+1} + \eta v^k \quad . \tag{2.12b}$$

The existence and uniqueness in the above steps follows from the representation lemma [20, Corollary 3.6.3]. The mappings B, D are maximal monotone operators as the subdifferentials of proper lsc convex functions. This together with the fact that the solution set of (2.7) is non-empty yields that the sequence $(b^k, v^k)_{k\in\mathbb{N}}$ defined by the algorithm (2.12) converges to some $(\overline{b}, \overline{v})$ such that $\overline{v} \in D\overline{b}$ and $\overline{b}$ solves ($\mathcal{D}'$) [45, Theorem 1]. By the change of variables $x^k = b^k + \eta v^k$, it follows that $x^k \to \overline{x} \in \mathsf{Fix}\, T$ for T given by (2.11).

For these definitions of B and D, the sequence $\left(b^k\right)_{k\in\mathbb{N}}$ generated by $b^k := \mathcal{J}_{\eta D} x^k$ for x^k generated by (2.10) corresponds exactly to the sequence $\left(b^k\right)_{k\in\mathbb{N}}$ generated by (2.8). Moreover, if A is full column rank, then by the discussion in [20] (see Appendix 4.2) both $\left(b^k\right)_{k\in\mathbb{N}}$ and the sequence $\left(v^k\right)_{k\in\mathbb{N}}$ generated by $v^k := \frac{1}{\eta}\left(x^k - b^k\right) \in Db^k$ correspond exactly to the sequences of points b^k and v^k generated by (2.6a)-(2.6c). Consequently, by [20, Proposition 3.42][1] the sequence $\left(u^k\right)_{k\in\mathbb{N}}$ defined by (2.6a) converges to a solution of ($\mathcal{P}'$). □

We now state sufficient conditions guaranteeing linear convergence of the ADMM and the Douglas-Rachford algorithms. The first conditions (i) of Theorem 2.2 are classical. The second conditions are new.

Theorem 2.2 (local linear convergence I)**.** *Let $J : U \to \mathbb{R} \cup \{+\infty\}$ and $H : V \to \mathbb{R}$ be proper, lsc and convex. Suppose there exists a solution to $0 \in B + D$ for $B := \partial\left(J^* \circ (-A^*)\right)$ and $D := \partial H^*$ where $A : U \to V$ is an injective linear mapping. Let $\widehat{x} \in \mathsf{Fix}\, T$ for T defined by (2.11). For fixed $\eta > 0$ and any given triplet of points (b^0, v^0, x^0) satisfying $x^0 := b^0 + \eta v^0$, with $v^0 \in Db^0$, generate the sequence $(v^k, b^k)_{k\in\mathbb{N}}$ by (2.6a)-(2.6c) and the sequence $(x^k)_{k\in\mathbb{N}}$ by (2.10).*

(i) Let $\mathcal{O} \subset U$ be a neighborhood of $\widehat{x}$ on which H is strongly convex with constant μ and ∂H is β-inverse strongly monotone for some $\beta > 0$. Then, for any $(b^0, v^0, x^0) \in \mathcal{O}$ satisfying $x^0 := b^0 + \eta v^0 \in \mathcal{O}$, the sequences $(x^k)_{k\in\mathbb{N}}$ and $(v^k, b^k)_{k\in\mathbb{N}}$ converge linearly to the respective points $\overline{x} \in \mathsf{Fix}\, T$ and $\left(\overline{b}, \overline{v}\right)$ with rate at least $K = (1 - \frac{2\eta\beta\mu^2}{(\mu+\eta)^2})^{\frac{1}{2}} < 1$.

(ii) Suppose that $T : W \to W$ for some affine subspace $W \subset U$ with $\widehat{x} \in W$. On the neighborhood $\mathcal{O}$ of $\widehat{x}$ relative to W, that is $\mathcal{O} \subset W$, suppose there is a constant $\kappa > 0$ such that

$$\|x - x^+\| \geq \sqrt{\kappa}\, \operatorname{dist}\left(x, \mathsf{Fix}\, T\right) \quad \forall x \in \mathcal{O},\ \forall x^+ \in Tx. \tag{2.13}$$

Then the sequences $(x^k)_{k\in\mathbb{N}}$ and $(v^k, b^k)_{k\in\mathbb{N}}$ converge linearly to the respective points $\overline{x} \in \mathsf{Fix}\, T \cap W$ and $\left(\overline{b}, \overline{v}\right)$ with rate bounded above by $\sqrt{1-\kappa}$.

[1] By convergence of $v^k \to \overline{v}$ and $b^k \to \overline{b}$ and the update rule (2.6c), $Au^k \to \overline{v}$, from which the claim follows - see Appendix 4.2.

In either case, the limit point $\overline{b} = \mathcal{J}_{\eta D}\overline{x}$ *is a solution to* ($\mathcal{D}'$), $\overline{v} \in D\overline{b}$ *and the sequence* $\left(u^k\right)_{k\in\mathbb{N}}$ *given by* (2.6a) *of Algorithm 2.1 converges to* $\overline{u}$*, a solution of* ($\mathcal{P}'$).

Proof. The final statement of the theorem and the statements about the sequence $\left(b^k, v^k\right)$ follows from Proposition 2.1 where it is shown that the sequence $(v^k, b^k)_{k\in\mathbb{N}}$ generated by (2.6a)-(2.6c) corresponds to sequences $\left(b^k\right)_{k\in\mathbb{N}}$ and $\left(v^k\right)_{k\in\mathbb{N}}$ generated respectively by (2.8) and $v^k = \frac{1}{\eta}\left(x^k - b^k\right) \in Db^k$ for $\left(x^k\right)_{k\in\mathbb{N}}$ generated by (2.10). The linear convergence of the iterates of Algorithm 2.1 claimed in statements (i) and (ii) follows from the properties of the operators T' and T defined respectively by (2.9) and (2.11).

Part (i). Since H is assumed to be strongly convex with $\mu > 0$ the modulus of convexity on $\mathcal{O}$, ∂H is strongly monotone with modulus of monotonicity μ [3, Example 22.3]. Since ∂H is also maximally monotone, using the identity $\partial H = (\partial H^*)^{-1}$ (see, for example, [41, Corollary 3.49]) we conclude that ∂H^* is Lipschitz continuous with constant $\frac{1}{\mu}$. Moreover, since ∂H is β-inverse strongly monotone on $\mathcal{O}$, we have for any $x, y \in \mathcal{O}$

$$\langle u - v, x - y\rangle \geq \beta|u - v|^2, \quad \text{whenever} \quad u \in \partial H(x), v \in \partial H(y).$$

Hence ∂H^* is strongly monotone with modulus β and Proposition 4 of [36] applies to yield linear convergence of the sequences $\left(x^k\right)$ and $\left(b^k\right)$ to the respective limit points $\overline{x}$ and $\overline{b}$

$$\|x^k - \overline{x}\| \leq LK^k; \quad \|b^k - \overline{b}\| \leq LK^k, \tag{2.14}$$

where L is some constant, $K = (1 - \frac{2\eta\beta}{(1+\eta\xi)^2})^{\frac{1}{2}}$ and $\xi = \frac{1}{\mu}$ is the Lipschitz constant for the set-valued map ∂H^* on $\mathcal{O}$. Now, since $v^k = \frac{1}{\eta}(x^k - b^k)$, we have for $v^k \to \overline{v} := \frac{1}{\eta}(\overline{x} - \overline{b})$ with the same rate as x^k and b^k, modulo a constant:

$$\|v^k - \overline{v}\| \leq \tfrac{1}{\eta}\left(\|x^k - \overline{x}\| + \|\overline{b} - b^k\|\right) \leq \frac{2LK^k}{\eta}. \tag{2.15}$$

This completes the proof of the first statement. △

Part (ii). Since B and D are maximal monotone operators the reflected resolvents $R_{\eta B}$ and $R_{\eta D}$ are nonexpansive [3, Proposition 23.7]. The composition $R_{\eta B}R_{\eta D}$ is nonexpansive which implies that the T is firmly nonexpansive [3, Proposition 4.2], and hence quasi-firmly nonexpansive on W. Condition (2.13) is the coercivity condition (b) of [30, Lemma 3.1] which guarantees local linear convergence of fixed-point mappings for (S, ϵ)-firmly nonexpansive mappings ($S \subset \mathsf{Fix}\ T \cap W$). Quasi-firmly nonexpansive mappings, under consideration here, are ($\mathsf{Fix}\ T \cap W, 0$)-firmly nonexpansive. Thus, by [30, Lemma 3.1] the sequence $(x^k)_{k\in\mathbb{N}}$ converges linearly on the neighborhood $\mathcal{O}$ with rate $\sqrt{1-\kappa}$. Nonexpansiveness of the resolvent $\mathcal{J}_{\eta D}$ and the relations $b^k = \mathcal{J}_{\eta D}x^k$ and $v^k = \frac{1}{\eta}\left(x^k - b^k\right)$ then complete the proof of the second statement. □

Remark 2.3. *The strong convexity assumption (*(i) *of Theorem 2.2 fails in a wide range of applications, and in particular for feasibility problems (minimizing the sum of* *indicator functions). By Theorem 1.5, case (ii) of Theorem 2.2, in contrast, holds in general for mappings T for which* $\mathrm{Id} - T$ *is* metrically subregular *and the fixed point sets are* isolated points *with respect to an affine subspace to which the iterates are confined. The restriction to the affine subspace W is a natural generalization for the Douglas Rachford algorithm, where the iterates are known to stay confined to affine subspaces orthogonal to the fixed point set [31, 42]. It would be far too restrictive to require that* $\mathsf{Fix}\, T$ *be a singleton on the entire ambient space V rather than with respect to just the affine hull of the iterates. We show that metric subregularity with respect to this affine subspace holds in many applications.*

Remark 2.4. *Proposition 2.1 and Theorem 2.2 and their proofs also hold in infinite dimensional Hilbert spaces. Lemma 3.1 of [30] is stated for Euclidean spaces, but the proof holds also on general Hilbert spaces.*

Proposition 2.5 (polyhedrality of the Douglas-Rachford operator)**.** *Let $J : U \to \mathbb{R} \cup \{+\infty\}$ and $H : V \to \mathbb{R}$ be proper, lsc and convex. Suppose, in addition, that J and H are piecewise linear-quadratic (see Definition 1.3). Define $B := \partial\left(J^* \circ (-A^*)\right)$ and $D := \partial H^*$ where $A : U \to V$ is a linear mapping. With $\eta > 0$ fixed, the operator $T : V \to V$ defined by (2.11) is ployhedral.*

Proof. Since the maps J and H are proper, lsc, convex and piecewise linear-quadratic, by [44, Theorem 11.14] so are the Fenchel conjugates, J^* and H^*. The subdifferentials $B := \partial\left(J^* \circ (-A^*)\right)$ and $D := \partial H^*$ and their resolvents, therefore, are polyhedral mappings [44, Proposition 12.30]. Since the graphs of reflectors $R_{\eta B}$ and $R_{\eta D}$ correspond to the graphs of their respective resolvents $\mathcal{J}_{\eta B}$ and $\mathcal{J}_{\eta D}$ through a linear transformation $R_{\eta B}$ and $R_{\eta D}$ are also polyhedral maps. Note that the resolvent mappings $\mathcal{J}_{\eta B}$ and $\mathcal{J}_{\eta D}$ are the proximal mappings of the convex functions $\left(J^* \circ (-A^*)\right)$ and H^* respectively and, hence, are single-valued [38]. The reflectors $R_{\eta B}$ and $R_{\eta D}$ are then also single-valued and therefore $T = \frac{1}{2}(R_{\eta B} R_{\eta D} + I)$ is polyhedral as the composition of single-valued polyhedral maps. □

Theorem 2.6 (local linear convergence II)**.** *Let $J : U \to \mathbb{R} \cup \{+\infty\}$ and $H : V \to \mathbb{R}$ be proper, lsc, convex, piecewise linear-quadratic functions (see Definition 1.3). Suppose there exists a solution to $0 \in B + D$ for $B := \partial\left(J^* \circ (-A^*)\right)$ and $D := \partial H^*$ where $A : U \to V$ is an injective linear mapping. With $\eta > 0$ fixed, define the operator $T : V \to V$ by (2.11). Suppose $T : W \to W$ for W some affine subspace of V and that* $\mathsf{Fix}\, T \cap W$ *is an isolated point $\{\overline{x}\}$. Then there is a neighborhood $\mathcal{O} \subset W$ of $\overline{x}$ such that, for all starting points (x^0, v^0, b^0) with $x^0 := b^0 + \eta v^0 \in \mathcal{O}$ for $v^0 \in D(b^0)$ so that $\mathcal{J}_{\eta D} x^0 = b^0$, the sequence $(x^k)_{k\in\mathbb{N}}$ generated by (2.10) converges linearly to $\overline{x}$ where $\overline{b} := \mathcal{J}_{\eta D}\overline{x}$ is a solution to $(\mathcal{D}')$. The rate of linear convergence is bounded above by $\sqrt{1-\kappa}$, where $\kappa = c^{-2} > 0$, for c a constant of metric subregularity of $\mathrm{Id} - T$ at $\overline{x}$ for the neighborhood $\mathcal{O}$. Moreover, the sequence $\left(b^k, v^k\right)_{k\in\mathbb{N}}$ generated by Algorithm 2.1 converges linearly to $\left(\overline{b}, \overline{v}\right)$ with $\overline{v} = \frac{1}{\eta}\left(\overline{x} - \overline{b}\right)$, and the sequence $\left(u^k\right)_{k\in\mathbb{N}}$ defined by (2.6a) of Algorithm 2.1 converges to a solution to $(\mathcal{P}')$.*

Proof. By Proposition 2.5 the Douglas-Rachford operator T is polyhedral and thus the first statement follows from Corollary 1.6. The statement about the sequences generated by Algorithm 2.1 follows as in Theorem 2.2. □

3 Error Bounds and Iterative Penalization

3.1 Structured Constraints and penalization

In this section, we discuss an iteratively regularized algorithmic scheme for solving the problems of the form

$$\min\{J(u) \mid u \in U \ \text{ and } f_j(Au) \le \epsilon_j,\ j = 1, 2, \ldots, M\}$$

where $J : U \to (-\infty, +\infty]$ is proper lsc and convex, the mapping $A : U \to V$ is linear, for all j the nonnegative-valued function $f_j : V \to \mathbb{R}_+$ is convex and smooth (at least at points that matter) and $\epsilon_j > 0$. We refer to the inequality constraints as *structured constraints*. It will be convenient to introduce the following notation that will help to reduce clutter. We collect the constraints into a vector-valued function so that we can write the problem as

$$\begin{aligned} &\underset{u\in U}{\text{minimize}} && J(u) \\ &\text{subject to} && F_\epsilon(Au) \le 0 \end{aligned} \tag{$\mathcal{P}$}$$

where

$$F_\epsilon : V \to \mathbb{R}^M := v \mapsto (f_1(v) - \epsilon_1, f_2(v) - \epsilon_2, \ldots, f_M(v) - \epsilon_M)^T. \tag{3.1}$$

Here the vector inequality is understood as holding element-wise.

A common approach to solving problems of the type ($\mathcal{P}$) arising from inverse problems is *implicitly* to apply the structured constraint by adding some (usually smooth) quantification of the constraint violation into the objective function:

$$\underset{u\in U}{\text{minimize}}\ J(u) + \rho\theta(F_\epsilon(Au)), \tag{$\mathcal{P}_\rho$}$$

where $\theta : \mathbb{R}^M \to (-\infty, +\infty]$ is a proper, lsc convex function and $\rho > 0$. Problem ($\mathcal{P}_\rho$) is the specialization of ($\mathcal{P}'$) with $H(Au) = \rho\theta(F_\epsilon(Au))$.

As is often seen in the inverse problems literature, the constraint violation parameter $\epsilon_j = 0$ ($j = 1, \ldots, M$), essentially penalizing divergence from the origin. A prominent instance of this form of regularization is the squared norm: $\theta(v) := \|v\|^2$. There are many efficient methods available for solving ($\mathcal{P}_\rho$). It is clear that for a certain value of ρ the optimal solution to ($\mathcal{P}_\rho$), u_ρ, will satisfy $f_j(Au_\rho) \le \bar{\epsilon}_j(\rho)$ with the *effective* error $\bar{\epsilon}_j(\rho)$ depending on ρ. What is *not* true in general, however, is that the solution to ($\mathcal{P}_\rho$) corresponds to the solution to ($\mathcal{P}$) for the constraint error $\bar{\epsilon}(\rho)$. Moreover, for our intended applications, U is a finite dimensional Euclidean space with dimension n and the dimensionality of the constraints M grows superlinearly as a function of n, so we would like to consolidate the constraints somehow while exploiting the phenomenon that, at the solution to ($\mathcal{P}$) relatively few of the constraints are in fact tight or *active*.

We consider convex penalties that reduce the dimensionality of the constraint structure and have the property that $\theta(F_\epsilon(Au)) = 0$ if and only if $F_\epsilon(Au) \leq 0$. Of particular interest among penalties with this property are *exact penalties*, that is penalties θ with the property that solutions to ($\mathcal{P}_\rho$) correspond to solutions to ($\mathcal{P}$) for all values of ρ beyond a certain threshold $\overline{\rho}$. For more background on exact penalization see, for example, [8, 9, 14, 16, 23, 28, 37]. We point also to Friedlander and Tseng [25] for a connection between exact penalization and what they call *exact regularization* as this fits well with our viewpoint that the structured constraints $F_\epsilon(Au) \leq 0$ constitute a regularization of the *model* with regularization parameter ϵ. This illustrates the distinction between *model-based* regularization, that is, regularization of the constraints motivated by external (eg. statistical) considerations, versus *numerical* regularization motivated solely on the grounds of enabling efficient (approximate) numerical solutions to ($\mathcal{P}$).

Define

$$\mathcal{C} := \{ u \in U \mid F_\epsilon(Au) \leq 0 \}. \tag{3.2}$$

This is a closed convex set since the f_j are lsc and convex. If there exists some $\alpha \in \mathbb{R}$ such that $\mathcal{C} \cap \operatorname{lev}_{\leq\alpha} J$ is nonempty and bounded then ($\mathcal{P}$) has a solution [3, Theorem 11.9]. This will happen, for instance, if $\operatorname{dom}(J) \cap \mathcal{C} \neq \emptyset$ and J is *coercive* [3, Proposition 11.12], that is J satisfies

$$\lim_{\|u\|\to\infty} J(u) = +\infty. \tag{3.3}$$

Such assumptions are naturally satisfied in many applications. Moreover, $\operatorname{lev}_{\leq\overline{\alpha}} J(u)$, the lower level-set of J corresponding to the optimal value $\overline{\alpha}$ in ($\mathcal{P}$), is convex and so the set of optimal solutions to ($\mathcal{P}$) is also convex. Define $J_\rho := J + \rho\theta(F_\epsilon \circ A)$ for the convex, lsc function θ satisfying $\theta(w) \geq 0$ for all w and $\theta(w) = 0$ if and only if $F_\epsilon(w) \leq 0$. Then J_ρ is convex, lsc and corresponds exactly to J on the set $\mathcal{C}$. Otherwise J_ρ increases pointwise to $+\infty$ at points outside $\mathcal{C}$ as $\rho \to \infty$. For $(\rho_k)_{k\in\mathbb{N}}$ with $\rho_k \to \infty$, the sequence of functions (J_{ρ_k}) epi-converges (see [44, Definition 7.1]) to $J + \iota_\mathcal{C}$ as $k \to +\infty$ where $\iota_\mathcal{C}$ is the indicator function of the set $\mathcal{C}$. As we will allow *approximate* solution of problems ($\mathcal{P}_\rho$) it will be helpful to recall the set of γ-minimizers: $\gamma - \operatorname{argmin} J_\rho := \{ u \mid J_\rho(u) \leq \inf J_\rho + \gamma \}$. The relation between the solution sets to ($\mathcal{P}$) and ($\mathcal{P}_\rho$) is detailed in the following, which is a direct application of [44, Theorem 7.33].

Proposition 3.1. *Let $J : U \to (-\infty, +\infty]$, $F_\epsilon : V \to \mathbb{R}^M$ and $\theta : \mathbb{R}^M \to \mathbb{R}$ be proper, lsc and convex, and let $A : U \to V$ be linear. Let J be coercive with $\operatorname{dom} J \cap \mathcal{C} \neq \emptyset$ for $\mathcal{C}$ defined by (3.2). Suppose further that $\theta(w) \geq 0$ and that $\theta(w) = 0$ if and only if $F_\epsilon(w) \leq 0$. Define $J_{\rho_k} := J + \rho_k\theta(F_\epsilon \circ A)$ where $\rho_k \nearrow +\infty$ as $k \nearrow +\infty$. Then $\inf J_{\rho_k} \to \inf J + \iota_\mathcal{C} < +\infty$. Moreover, for any sequence of errors $\gamma_k \searrow 0$ and corresponding points $u^k \in \gamma_k - \operatorname{argmin} J_{\rho_k}$, the sequence $\left(u^k\right)_{k\in\mathbb{N}}$ is bounded, and all its cluster points belong to $\operatorname{argmin}\{J + \iota_\mathcal{C}\}$.*

Proof sketch. The property of the convex penalty θ that $\theta(w) \geq 0$ and $\theta(w) = 0$ if and only if $F_\epsilon(w) \leq 0$ yields epi-convergence of J_{ρ_k} to $J + \iota_{\mathcal{C}}$. Coercivity of J guarantees that J_ρ is level bounded for all values of $\rho > 0$. These two properties, together with lower semicontinuity and the fact that J and J_ρ are proper, are all that is needed to prove the result. $\square$

If the regularization were *exact*, then we would know that for all parameter values ρ large enough, the solutions to ($\mathcal{P}_\rho$) coincide with solutions to ($\mathcal{P}$). We return to this later.

3.2 Solution to the regularized Subproblem and error bounds

We now turn our attention to solution of the problem ($\mathcal{P}_\rho$) for a fixed value of ρ_k. The alternating directions method of multipliers discussed in Section 2 is useful for solving this problem in the sense that it has an error bound under specific assumptions which gives a stopping rule. This is not unique to Algorithm 2.1, but we focus on this method due to its prevalence in practice.

Recall the unregularized problem ($\mathcal{P}$):

$$\begin{array}{ll} \underset{u \in U}{\text{minimize}} & J(u) \\ \text{subject to} & F_\epsilon(Au) \leq 0 \end{array} \tag{$\mathcal{P}$}$$

It will be convenient to rewrite the penalized problem[2] ($\mathcal{P}_\rho$) as

$$\underset{u \in U}{\text{minimize}}\ \frac{1}{\rho} J(u) + \theta(F_\epsilon(Au)). \tag{$\mathcal{P}_\rho$}$$

Consider also the *limiting* problem

$$\underset{u \in U}{\text{minimize}}\ \theta(F_\epsilon(Au)). \tag{$\mathcal{P}_\infty$}$$

We view problem ($\mathcal{P}_\rho$) as the regularized version of ($\mathcal{P}_\infty$) with J as the regularizing functional and $\frac{1}{\rho}$ as the regularization parameter. Denote the solution sets to these problems by

$$\begin{aligned} S &:= \operatorname{argmin}\left\{ J(u) \mid u \in U,\ F_\epsilon(Au) \leq 0 \right\}, \\ S_\rho &:= \operatorname{argmin}\left\{ \frac{1}{\rho} J(u) + \theta\left(F_\epsilon(Au)\right) \mid u \in U \right\}, \\ S_\infty &:= \operatorname{argmin}\left\{ \theta\left(F_\epsilon(Au)\right) \mid u \in U \right\}. \end{aligned}$$

If the penalization θ satisfies $\theta(F_\epsilon(Au)) = 0$ if and only if $F_\epsilon(Au) \leq 0$, then it is immediately clear that S_∞ corresponds to the feasible set of problem ($\mathcal{P}$) hence $S \subset S_\infty$. What is more remarkable is that, if a Lagrange multiplier for ($\mathcal{P}$) exists, then $S_\rho = S$ for all ρ large enough, that is, the penalty θ is *exact*.

[2] Of course, the value of the problem is not the same, but the solutions are.

Theorem 3.2 (Theorem 4.2 of [25]). *Suppose that S is nonempty and compact, and that there exist Lagrange multipliers λ for ($\mathcal{P}$). Let θ in ($\mathcal{P}_\rho$) be convex and satisfy $\theta(F_\epsilon(Au)) = 0$ if and only if $F_\epsilon(Au) \leq 0$. Then the solution set to the penalized problem S_ρ coincides with the solution set to the exact problem, S, for all $\rho > \theta^\circ(\lambda)$ where θ° is the polar function of θ given by $\theta^\circ(\lambda) = \sup_{x \not\leq 0} \frac{\lambda^T x}{\theta(x)}$.*

It is easy to check whether a solution $u_\rho \in S_\rho$ is in fact feasible for ($\mathcal{P}$) (and hence also in S) by simply evaluating the value of $\theta\left(F_q(Au_\rho)\right)$. More generally, one would check whether the first order optimality conditions for ($\mathcal{P}_\infty$) are satisfied at u_ρ, namely

$$0 \overset{?}{\in} \partial\theta\left(F_\epsilon \circ A(\cdot)\right) \text{ at } u_\rho. \tag{3.4}$$

An explicit formula for the subdifferential in (3.4) for image denoising and deconvolution is given in Section 4 as this will be needed for computing Step (2.6b) of Algorithm 2.1.

If, in addition, S_∞ is *weakly sharp*, then one can obtain an upper bound for the distance of solutions to ($\mathcal{P}_\rho$) to *feasible* solutions to ($\mathcal{P}$), even in the absence of Lagrange multipliers for ($\mathcal{P}$). The notion of weak sharpness for convex optimization was introduced by Burke and Ferris [13]. The solution set $\operatorname{argmin}\{f(x) \mid x \in \Omega\}$ for a nonempty closed convex set Ω, is weakly sharp if, for $\overline{p} = \inf_\Omega f$, there exists a positive number α (sharpness constant) such that

$$f(x) \geq \overline{p} + \alpha\, d(x, S_f) \quad \forall x \in \Omega.$$

Similarly, the solution set S_f is weakly sharp of order γ if there exists a positive number α (sharpness constant) such that, for each $x \in \Omega$,

$$f(x) \geq \overline{p} + \alpha\, d(x, S_f)^\gamma \quad \forall x \in \Omega.$$

Assumption 3.3.

(i) The solution set S_∞ of problem ($\mathcal{P}_\infty$) is nonempty.

(ii) $\operatorname{lev}_{\leq\alpha} J$ is bounded for each $\alpha \in \mathbb{R}$ and $\inf_{x\in U} > -\infty$.

(iii) The solution set S_∞ of ($\mathcal{P}_\infty$) is weakly sharp of order $\gamma \geq 1$.

Theorem 3.4. *Suppose Assumption 3.3(i)-(ii) hold.*

(i) For any $\overline{\rho} > 0$, $\bigcup_{\rho \geq \overline{\rho}} S_\rho$ is bounded.

(ii) If, in addition, Assumption 3.3(iii) holds with modulus of sharpness γ, then for any $\overline{\rho} > 0$ there exists $\tau > 0$ such that

$$dist(u_\rho, S_\infty)^{\gamma-1} \leq \frac{\tau}{\rho}, \quad \forall u_\rho \in S_\rho, \quad \rho \geq \overline{\rho}. \tag{3.5}$$

(iii) If, in addition, Assumption 3.3(iii) holds and the penalization θ is exact, then for all ρ large enough, $u_\rho \in S$ and $dist(u_\rho, S_\infty) = dist(u_\rho, S) = 0$.

Proof. (i) and (ii). Under the assumption 3.3, Theorem 5.1 in [25] directly applies to yield the result. △

(iii). If the penalization θ is exact, then $\theta(F_\epsilon(Au)) = 0$ if and only if $F_\epsilon(Au) \le 0$, hence $S = S_\rho$ for all ρ large enough, and S_∞ corresponds exactly to the feasible set in $(\mathcal{P})$. □

Remark 3.5. *The error bound* (3.5) *holds independent of the existence of Lagrange multipliers for* $(\mathcal{P})$*, hence, for exact penalization under Assumption 3.3, Theorem 3.4 yields an upper bound on the distance of solutions to* $(\mathcal{P}_\rho)$ *to feasible points for* $(\mathcal{P})$*.*

While it is nice to know that, with exact penalization, one can achieve an exact correspondence between the original constrained optimization problem and the penalized problem, the whole point of relaxing the constraints is to reduce the computational burden of strictly enforcing the constraints. As is often done in practice, one gradually strengthens the constraints, finding intermediate points that nearly solve the relaxed problem and using these as starting points for solving a more strictly penalized problem. Together with Theorem 3.4, the linear convergence rate established in Theorem 2.2, or alternatively Theorem 2.6, yield estimates on the distance of intermediate points in an iteratively penalized algorithm, not only to the solution set of the relaxed problem, but also to the feasible set of the unrelaxed problem.

4 Application: image deconvolution and denoising with statistical multiscale analysis

We specialize the above results to the application of optimization with statistical multiscale side constraints. In particular, the problem at hand involves image deconvolution and denoising with statistical multiscale estimation as presented in [24]. We are well aware that there are many ways to model such problems that permit much less computationally intensive numerical solutions than the technique we present here. Our interest in multiresolution deconvolution/denoising model of [24] is two-fold: first, it is one of the few techniques we are aware of that yields quantitative (i.e. statistical) guarantees for the recovered images, and secondly, it is an important instance of convex optimization problems where the number of constraints grows superlinearly as a function of the number of unknowns. Our numerical demonstration addresses the first issue of quantitative image denoising: if the numerics do not permit estimates for the distance to the model solution, then the quantitative assurances of the model are irrelevant. Unlike the numerical approach proposed in [24], the numerical approach we present here permits error bounds to within machine accuracy of our numerical solution to the true model solution.

For our demonstration, we are presented with an image $y \in \mathbb{R}^n$ (Figure 1(a)) generated from a Stimulated Emission Deletion (STED) microscopy experiment conducted at the Laser-Laboratorium Göttingen examining tubulin, represented as the "object" $u \in \mathbb{R}^m$. The imaging model is simple linear convolution, $Au \approx y$ where A

is a convolution matrix with point-spread function shown in Figure 1(e). The measurement y is *noisy* or otherwise inexact, and thus an exact solution $Au = y$ is not desirable. Although the noise in such images is usually modeled by Poisson noise, a Gaussian noise model with constant variance suffices as the photon counts are of the order of 100 per pixel and do not vary significantly across the image. Figure 1 (b) shows a close-up which we used as the noisy data $y \in \mathbb{R}^2$ with $n = 64 \times 64$ data points. We calculate the numerically reconstructed tubulin $\overline{u}$ shown in Figure 1(c) that minimizes the qualitative objective

$$J(u) := \alpha ||u||_2^2, \tag{4.1}$$

subject to structured constraints (see problem ($\mathcal{P}$)) given by a statistical property that is consistent with the noise in the observation y. The numerical "image" generated from the reconstructed tubulin $\overline{u}$ is given by $\overline{v} = A\overline{u}$ and is shown in Figure 1(d).

We emphasize that, since this is experimental data, there is no "truth" for comparison. One can, however, get statistical guarantees on the reliability of the reconstructions. Following the approach proposed in [24] we quantify the difference between an estimate $v = Au$ and the data y via the maximum absolute value of all weighted inner products of the residual function $\triangle(\cdot; y) : \mathbb{R}^n \to \mathbb{R}^n$:

$$f_j(v) := \left|\langle \omega_j, \triangle(v; y)\rangle\right|, \quad j \in \{1, 2, \ldots, M\}. \tag{4.2}$$

The residual function used in [24] $\triangle$ is simply $v - y$. The weights ω_j are normalized window functions of all squares of side lengths 1 and 2 pixels so that the set $\mathcal{I} \subset \{1, 2, \ldots, M\}$ is the index set corresponding to all collections of these square subsets of the image. The statistical multiscale analysis presented in [24] requires that, on each window,

$$\max_{j\in\mathcal{I}}\{f_j(v)\} \leq q. \tag{4.3}$$

The same error q is specified at all scales. Hence F_ϵ in (3.1) specializes to

$$F_q : \mathbb{R}^n \to \mathbb{R}^{M+1} := v \mapsto (f_1(v) - q, f_2(v) - q, \ldots, f_M(v) - q, 0)^T. \tag{4.4}$$

for $f_j : \mathbb{R}^n \to \mathbb{R}$ defined by (4.2) ($j = 1, \ldots, M$) and

$$\theta : \mathbb{R}^{M+1} \to \mathbb{R} : \theta(w) := \max\{w_1, w_2, \ldots, w_{M+1}\}. \tag{4.5}$$

(Here we are expanding the original F_ϵ by the constant function $f_{M+1}(v) := 0$.) The max function is a standard tool in exact penalization methods [9, 16].

Algorithm 4.1 (Sequential ADMM: deconvolution/denoising).
Initialization. *Given an image y, a sequence of error tolerances $(\gamma_k)_{k\in\mathbb{N}}$ with $0 \leq \gamma_k \to 0$ Choose parameters: $\beta > 1$ and the penalty parameter $\eta \in (0,2)$. Initialize $k = 0 = i$, $b^{0,0} = 0, v^0 = y$, $u^{0,0} = A^*y$, and compute $u^{0,1} = \operatorname{argmin}_u \left\{ J(u) + \langle b^{0,0}, Au\rangle + \frac{\eta}{2}\|Au - v^{0,0}\|_2^2 + \frac{1}{2}\|u - u^{0,0}\|_{M_1}^2 \right\}$ for J defined by (4.1).*

For $k = 0, 1, 2, \ldots$

- **While** $\|u^{k,i+1} - \overline{u}^k\| > \gamma_k$
  - *Compute $(v^{k,i+1}, b^{k,i+1})$ via Algorithm 2.1 steps (2.6b)-(2.6c) with $H := \rho_k\theta\left(F_q(\cdot)\right)$ for θ given by (4.5) and F_q given by (4.4).*
  - *Increment $i = i+1$ and calculate $u^{k,i+1}$ via Algorithm 2.1 step (2.6a).*
- **Update/reset:** *Set $u^{k+1,1} := u^{k,i+1}$ and $\rho_{k+1} = \beta\rho_k$. Set $k = k+1$ and $i = 0$. If $\theta\left(F_q\left(u^{k,1}\right)\right) = 0$, set $\gamma_k = 0$.*

For the image size $n = 64 \times 64$ with the window system of squares of lengths 1 and 2, the number of windows is $M = 8065$. The constant α in (4.1) is, strictly speaking, redundant but was introduced as an additional means to balance the contributions of the individual terms to make the most of limited numerical accuracy (double precision). We chose $\alpha = 0.01$. The constant q was taken to be 3σ. Algorithm 4.1 does not specify how the iterates $u^{k,j}$ and $v^{k,j}$ are calculated. We discuss this in the next section.

4.1 Subdifferential Projection

Computation of $u^{k,i+1}$ in Algorithm 4.1 involves minimizing a convex quadratic function without constraints, which is unproblematic. Computation of $v^{k,i+1}$ involves minimizing the sum of a convex, piecewise linear function $\theta(F_q(v))$ and a quadratic function $h(v) := \frac{\eta}{2}\|Au^{k,i+1} - v\|^2$. This can be solved via any number of techniques ranging from first order methods like FISTA [7] to higher-order nonlinear optimization methods like quasi-Newton methods studied in [33]. In order to take advantage of the relative sparsity of the active constraints, we propose the following (exact) algorithm.

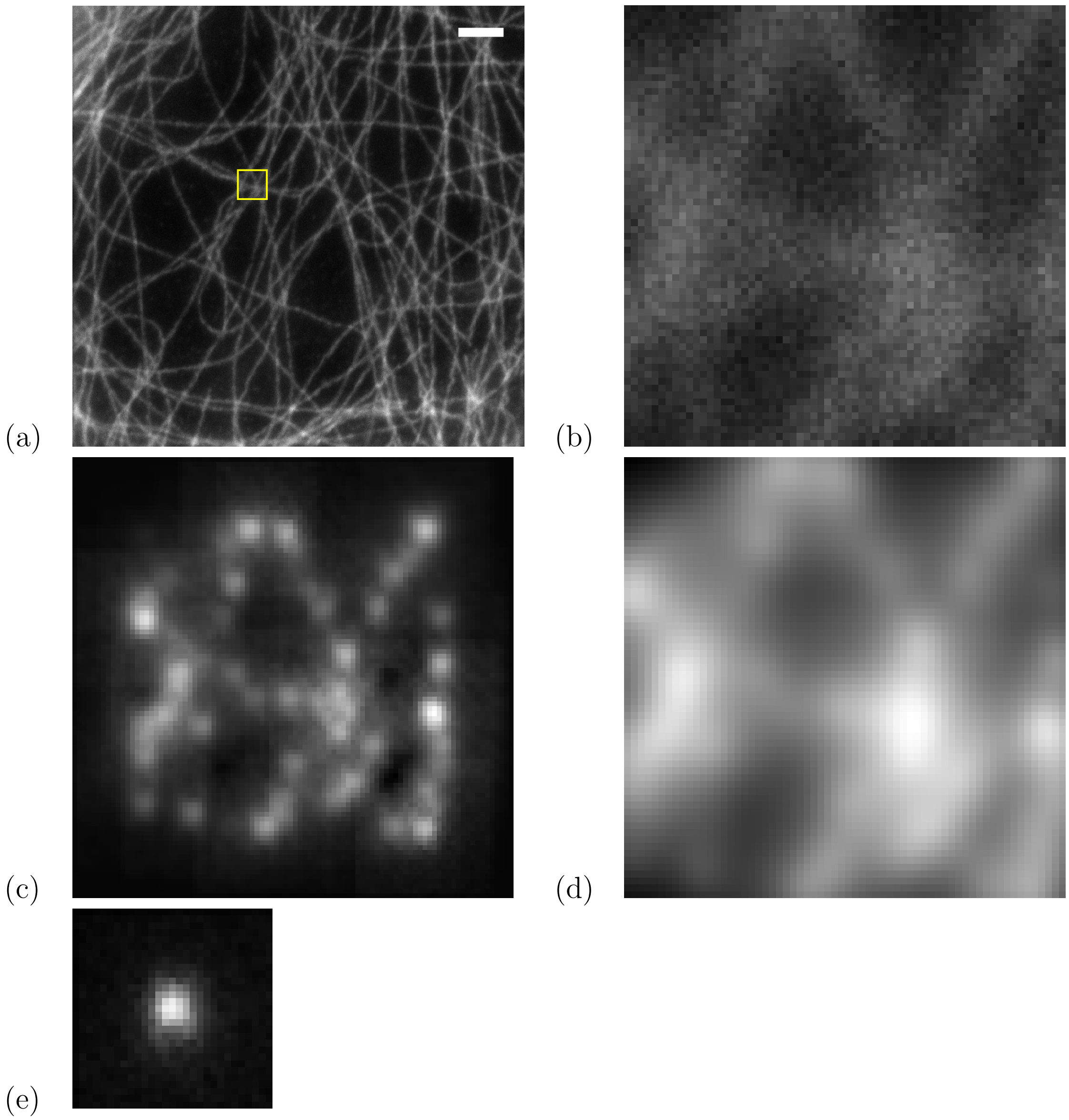

Figure 1: (a) Original data (STED image of Tubulin), (b) an enlargement of the indicated box, (c) the reconstruction for $\rho = 4096$, (d) the reconstruction convolved with the PSF and (e) PSF. Length of scale bar in (a) is 1μm, size of (b), (c) and (d) is $640 \times 640\,\mathrm{nm}^2$, and size of (e) is $290 \times 290\,\mathrm{nm}^2$. Within each window used for the reconstruction, the sum of the pixel values in (d) lie within a confidence interval of 3σ of those in (b).

Algorithm 4.2 (**Steepest Subdifferential Descent).**
Initialization. *Given b, u, the constant $\eta > 0$ and an initial point v^0, compute the residual $r^0 := b + \eta Au - \eta v^0$ and the projected residual $z^0 := P_{\partial(\theta(F_q(v^0)))}(r^0)$ for $\partial(\theta(F_q(v^0)))$ given by (4.11).*

For $l = 0, 1, 2, \dots$

- **If** $z^l = r^l$
  - *set $\overline{v} = v^l$ and STOP;*
- **else**
  - *set $v^{l+1} = v^l + \lambda_l \left(z^l - r^l\right)$ where $\lambda_l > 0$ is the largest constant λ such that $\theta\left(F_q\left(v^l + \lambda\left(z^l - r^l\right)\right)\right) = f_i\left(v^l + \lambda\left(z^l - r^l\right)\right) - q$ for $i \in I(v^l)$ with*

$$I(v) := \{j \mid f_j(v) - q = \theta(F_q(v))\}\,; \tag{4.6}$$

  - *compute $r^{l+1} := b + \eta Au - \eta v^{l+1}$ and the projected residual*

$$z^{l+1} := P_{\partial(\theta(F_q(v^{l+1})))}(r^{l+1}); \tag{4.7}$$

  - *increment $l = l + 1$.*

Algorithm 4.2 is an *active set method* and the set $I(v)$ defined by (4.6) is the set of *active indexes at* v. Another helpful interpretation is as a steepest subgradient descent method for solving

$$\operatorname{argmin}_v \left\{G(v) := \theta\left(F_q(v)\right) - \langle b, v\rangle + \tfrac{\eta}{2}\|Au - v\|_2^2\right\}. \tag{4.8}$$

The steepest descent step is

$$v^{l+1} = v^l + \lambda_l d^l$$

for $d^l := P_{\partial G(v)}(0) = -r^l + z^l$ with $z^l := P_{\partial\theta\left(F_q\left(v^l\right)\right)}\left(r^l\right)$ and $r^l = b + \eta\left(Au - v^l\right)$. The choice of the step length λ_l ensures that, at each step l, the active set is growing; specifically,

$$I\left(v^l\right) \subset I\left(v^l + \lambda_l d^l\right).$$

At termination, the subdifferential $\partial\theta\left(F_q\left(v^l\right)\right)$ is large enough that it contains the residual r^l. The terminal point of Algorithm 4.2, $\overline{v}$, is a point in (4.8) since it satisfies the first-order optimality conditions:

$$0 = \overline{z} - b - \eta\left(Au - \overline{v}\right) \in \partial\theta\left(F_q(\overline{v})\right) - b - \eta\left(Au - \overline{v}\right) = \partial G(\overline{v}) \tag{4.9}$$

where $\overline{z} = P_{(\partial\theta(F_q(\overline{v})))}\left(b + \eta\left(Au - \overline{v}\right)\right)$. Replacing u and b with $u^{k,i+1}$ and $b^{k,i}$ respectively yields the update for $v^{k,i}$ in Algorithm 4.1.

The expression for the subdiffferential $\partial\theta\left(F_q\right)$ is particularly simple in this case. Note that $I(v) \neq \emptyset$ for all v. Applying the (convex) calculus of subdifferentials to the objective $\theta(F_\epsilon(v))$, as permitted by the regularity of θ and F (see, for instance [15, Section 2.3]), yields

$$\partial\theta\left(F_q(v)\right) = \mathrm{co}\,\left\{\nabla f_j(v) \mid j \in I(v)\right\} \tag{4.10}$$

where co denotes the convex hull of a set of points. This, of course, assumes that f_j is differentiable at v for those $j \in I(v)$. Inspection of (4.2) shows that this is not the case in general, in particular at points v^* where $f_j(v^*) = 0$. However, such points will never be in the active set $I(v^*)$ since $f(v^*) - q < 0 \leq \theta\left(F_q\left(v^*\right)\right)$ for all $q > 0$, so we can safely apply formula (4.10) without further ado. This yields the following specialization for $f_j(v) = |\langle w_j,\ v - y\rangle|$ given by (4.2):

$$\begin{aligned}
\partial\theta\left(F_q\left(v\right)\right) &= \mathrm{co}\,\left\{\nabla f_j(v) \mid j \in I(v)\right\} \qquad (4.11)\\
&= \begin{cases} \mathrm{co}\,\left\{\left\{\mathrm{sign}\left(\langle w_j,\ v-y\rangle\right) w_j \mid j \in I(v)\setminus\{M+1\}\right\},\ 0\right\} & \theta(F_q(v)) \leq 0 \\ \mathrm{co}\,\left\{\mathrm{sign}\left(\langle w_j,\ v-y\rangle\right) w_j \mid j \in I(v)\right\} & \theta(F_q(v)) > 0. \end{cases}
\end{aligned}$$

4.2 Numerical results

In Figure 2 a sample run of the algorithm shows a succession of outer iterations. Within each outer iteration, the inner iteration proceeds with the current value of ρ_k until the step size between successive iterates $u^{k+1,j}$ and $u^{k,j}$ drops below the tolerance $\gamma_k = 10^{-3}$. Then ρ_k is increased by a constant factor. Since, for this model the penalization θ is exact, once the constraints appear to be satisfied (as determined by monitoring the value of $\theta\left(F_q\left(v^k\right)\right)$), the penalty ρ no longer needs to be updated, and the inner loop of the algorithm can be run to the desired accuracy. As indicated in Figure 2, the constraints appear to be satisfied exactly for a value of $\rho_k = 2048$, where the penalty term $\rho_k\theta\left(F_q\left(v^k\right)\right)$ (cyan plot) drops suddenly to 10^{-8}. However, this value of the penalty parameter is apparently just below the critical value, as the algorithm switches back to some constraint violation in preference of reducing the residual error around iteration 24000. Jumps in the constraint penalty term, and hence the step-lengths in the u^k and v^k iterates, reflect the nonsmoothness of the max function. For $\rho_k = 4096$ (beyond iteration 29000) the constraint penalty remains at machine zero for the given residual and step tolerances and the algorithm meets the termination criteria after 31000 iterations.

Acknowledgments

We thank Jennifer Schubert of the Laser-Laboratorium Göttingen for providing us with the STED measurements shown in Fig. 1. Thanks also to Jalal Fadili for fruitful discussions and helpful comments during the preparation of this work.

Appendix

Duality of ADMM and the Douglas-Rachford Algorithm. Consider the sequence $\left(b^k, v^k\right)_{k\in\mathbb{N}}$ of the Douglas Rachford iteration 2.8, for the case $B := \partial(J^* \circ (-A^*))$; $D := \partial H^*$.

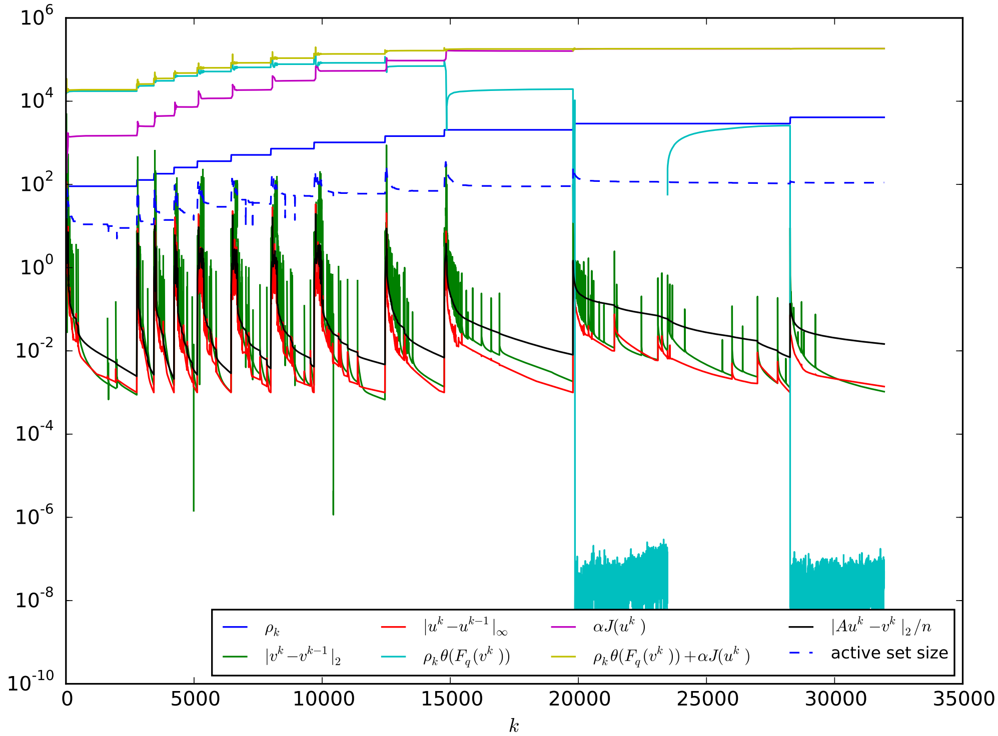

Figure 2: Sample run of the algorithm, starting with $\rho = 1$. Whenever the step size $||u^{k+1} - u^k||$ falls below $\gamma_k := 10^{-3}$, the inner iteration is terminated and ρ_k is increased according to algorithm 4.1 with $\beta = 2$. At $\rho_k = 2048$ the value of the penalty term $\rho_k \theta\left(F_q\left(v^k\right)\right)$ where θ is the max function (cyan plot) drops suddenly to 10^{-8} indicating that the exact constraints ((4.3)) are satisfied to within machine precision. For this value of the penalty parameter, however, the algorithm switches back to some constraint violation in preference of reducing the residual error as can be seen in the jump of the cyan plot around iteration 24000. For $\rho_k = 4096$ (beyond iteration 29000) the constraint penalty remains at machine zero for the given residual and step tolerances and the algorithm meets the termination criteria after 31000 iterations.

Recalling the two-step implementation (2.12), denote $\bar{p} := b^k - \eta v^k$ and $p' := q^{k+1}$. Then (2.12a) is the proximal step $p' = (I + \eta\partial(J^* \circ (-A^*)))^{-1}\bar{p}$ on the operator $B = \partial(J^* \circ (-A^*))$. If A has full column rank, by [20, Proposition 3.32(iv)], this step can be performed by

$$\begin{aligned} u^{k+1} &= \arg\min_u \{J(u) + \langle \bar{p} + \eta v^k, Au\rangle + \tfrac{\eta}{2}\|Au - v^k\|^2\} \qquad (4.12)\\ p' &= \bar{p} + \eta A u^{k+1}. \qquad (4.13)\end{aligned}$$

Indeed, since A has full rank, $J(u) + \langle \bar{p} + \eta v^k, Au\rangle + \frac{\eta}{2}\|Au - v^k\|^2$ is a proper strongly convex function of u and has a unique minimizer u^{k+1}. From the optimality condition for (4.12),

$$0 \in \partial J(u^{k+1}) + A^*(\bar{p} + \eta A u^{k+1}) = \partial J(u^{k+1}) + A^* p'.$$

Hence, $(u^{k+1}, -A^*p') \in \operatorname{gph} \partial J$ which implies $(-A^*p', u^{k+1}) \in \operatorname{gph} \partial J^*$. This gives

$$\begin{aligned} \Leftrightarrow (p', u^{k+1}) &\in \operatorname{gph} (\partial J^* \circ (-A^*))\\ \Leftrightarrow (p', -Au^{k+1}) &\in \operatorname{gph} (-A \circ \partial J^* \circ (-A^*)) \subseteq \operatorname{gph} \partial\left(J^* \circ (-A^*)\right).\end{aligned}$$

Using (4.13),

$$\begin{aligned} (p', \tfrac{1}{\eta}(\bar{p} - p')) &\in \operatorname{gph} \partial\left(J^* \circ (A^*)\right)\\ \Leftrightarrow p' &= (I + \eta\partial(J^* \circ (A^*)))^{-1}\bar{p}.\end{aligned}$$

Substituting $\bar{p} = b^k - \eta v^k$ in (4.12)-(4.13) yields

$$\begin{aligned} u^{k+1} &= \arg\min_u \{J(u) + \langle b^k - \eta v^k + \eta v^k, Au\rangle + \tfrac{\eta}{2}\|Au - v^k\|^2\}; \qquad (4.14)\\ q^{k+1} &= b^k - \eta v^k + \eta A u^{k+1}. \qquad (4.15)\end{aligned}$$

Similarly, if we denote $\bar{p} := q^{k+1} + \eta v^k (= b^k + \eta A u^{k+1})$ and $p' := b^{k+1}$, (2.12b) is the proximal step $p' = (I + \eta\partial H^*)^{-1}\bar{p}$ on the operator $D = \partial H^*$ which can be performed via

$$\begin{aligned} v^{k+1} &= \arg\min_v \{H(v) - \langle \bar{p} - \eta A u^{k+1}, v\rangle + \tfrac{\eta}{2}\|Au^{k+1} - v\|^2\};\\ p' &= \bar{p} - \eta v^{k+1}.\end{aligned}$$

Substituting $\bar{p} = b^k + \eta A u^{k+1}$,

$$\begin{aligned} v^{k+1} &= \arg\min_v \{H(v) - \langle b^k + \eta A u^{k+1} - \eta A u^{k+1}, v\rangle + \tfrac{\eta}{2}\|Au^{k+1} - v\|^2\} \qquad (4.16)\\ b^{k+1} &= b^k + \eta A u^{k+1} - \eta v^{k+1}. \qquad (4.17)\end{aligned}$$

Now, (4.14)-(4.15) and (4.16)-(4.17) together yield

$$\begin{aligned} u^{k+1} &= \arg\min_u \{J(u) + \langle b^k, Au\rangle + \tfrac{\eta}{2}\|Au - v^k\|^2\}\\ v^{k+1} &= \arg\min_v \{H(v) - \langle b^k, v\rangle + \tfrac{\eta}{2}\|Au^{k+1} - v\|^2\}\\ b^{k+1} &= b^k + \eta(Au^{k+1} - v^{k+1}).\end{aligned}$$

This is the ADMM algorithm (2.6a)-(2.6c) for the primal problem $(\mathcal{P}_\lambda)$. □

References

[1] F. J. Aragón Artacho, J. M. Borwein, and M. Tam. Douglas-Rachford feasibility methods for matrix completion problems. *ANZIAM J.*, 55(4):299–326, 2013.

[2] F. J. Aragón Artacho, J. M. Borwein, and M. Tam. Recent results on Douglas-Rachford methods. *Serdica Math. J.*, 39:313–330, 2013.

[3] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces.* CMS Books in Mathematics. Springer-Verlag, New York, 2011.

[4] H. H. Bauschke, D. Noll, and H. M. Phan. Linear and strong convergence of algorithms involving averaged nonexpansive operators. *J. Math. Analysis and Appl.*, (421):1–20, 2015.

[5] H.H. Bauschke, J.Y. Bello Cruz, T.T.A. Nghia, H.M. Phan, and X. Wang. The rate of linear convergence of the Douglas-Rachford algorithm for subspaces is the cosine of the Friedrichs angle. *J. Approx. Theory*, 185:63–79, 2014.

[6] H.H. Bauschke and D. Noll. On the local convergence of the Douglas-Rachford algorithm. *Archiv der Mathematik*, 102:589–600, 2014.

[7] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[8] D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods.* Academic Press, New York, 1982.

[9] D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, Massachusettes, 2nd edition, 1999.

[10] J. M. Borwein and A. S. Lewis. *Convex analysis and nonlinear optimization : theory and examples.* Springer Verlag, New York, 2nd edition, 2006.

[11] J. M. Borwein and M. Tam. The cyclic Douglas-Rachford method for inconsistent feasibility problems. *J. Nonlinear Convex Anal.*, 16(4):537–584, 2015.

[12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundation and Trends in Machine Learning*, 3(1):1–122, 2011.

[13] J. V. Burke and M. C. Ferris. Weak sharp minima in mathematical programming. *SIAM J. Control. Optim.*, 31:1340–1359, 1993.

[14] J.V. Burke. An exact penalization viewpoint of constrained optimization. *SIAM J. Control Optim.*, 29:968–998, 1991.

[15] F.H. Clarke. *Optimization and Nonsmooth Analysis.* Wiley, New York, 1983. Republished as Vol. 5, Classics in Applied Mathematics, SIAM, 1990.

[16] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust Region Methods*. SIAM Publications, Philadelphia, 2000.

[17] A. L. Dontchev and R. T. Rockafellar. *Implicit functions and solution mapppings*. Srpinger-Verlag, New York, 2009.

[18] A. L. Dontchev and T. Zolezzi. *Well-posed Optimization Problems*. Srpinger-Verlag, New York, 1993.

[19] J. Douglas and H. H. Rachford. On the numerical solution of heat conduction problems in two or three space variables. *Trans. Amer. Math. Soc.*, 82:421–439, 1956.

[20] J. Eckstein. *Splitting Methods for Monotone Operators with Applications to Parallel Optimization*. PhD thesis, MIT, 1989.

[21] J. Eckstein and D. P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Math. Program.*, 55:293–318, 1992.

[22] J. Eckstein and W. Yao. Understanding the convergence of the alternating direction method of multipliers: Theoretical and computational perspectives. *Pacific Journal on Optimization*, to appear.

[23] R. Fletcher. An ℓ_1 penalty method for nonlinear contraints. In P. T. Boggs, R. H. Byrd, and R. B. Schnabel, editors, *Numerical Optimization*, pages 26–40. SIAM, 1985.

[24] K. Frick, P. Marnitz, and A. Munk. Statistical multiresolution dantzig estimation in imaging: Fundamental concepts and algorithmic framework. *Electron. J. Stat.*, 6:231–268, 2012.

[25] M. P. Friedlander and P. Tseng. Exact regularization of convex programs. *SIAM J. Optim.*, 18(4):1326–1350, 2007.

[26] D. Gabay. *Augmented Lagrangian Methods: Applications to the Solution of Boundary- Value Problems*, chapter Applications of the method of multipliers to variational inequalities, pages 299–331. North-Holland,, 1983.

[27] R. Glowinski and A. Marroco. Sur l'approximation, par elements finis d'ordre un, et las resolution, par penalisation-dualitè, d'une classe de problemes de dirichlet non lineares. *Revue Francais d'Automatique, Informatique et Recherche Opérationelle*, 9(R-2):41–76, 1975.

[28] S.-P. Han and O. L. Mangasarian. Exact penalty functions in nonlinear programming. *Math. Program.*, 17:251–269, 1979.

[29] B. He and X. Yuan. On the o(1/n) convergence rate of the Douglas-Rachford alternating direction method. *SIAM J. Numer. Anal.*, 50(2):700–709, 2012.

[30] R. Hesse and D. R. Luke. Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems. *SIAM Journal of Optimization*, 23(4):2397–2419, 2013.

[31] R. Hesse, D. R. Luke, and P. Neumann. Alternating projections and Douglas-Rachford for sparse affine feasibility. *IEEE Transactions on Signal Processing*, 62(18):4868–4881, 2014.

[32] D. Klatte and B. Kummer. Optimization methods and stability of inclusions in Banach spaces. *Math. Program.*, 117(1-2):305–330, 2009.

[33] A. S. Lewis and M. L. Overton. Nonsmooth optimization via quasi-Newton methods. *Mathematical Programming*, 141:135–163, 2013.

[34] G. Li and T. K. Pong. Douglas-Rachford splitting for nonconvex feasibility problems. *Math. Prog. A*, accepted.

[35] J. Liang, G. Peyré, J. Fadili, and D. R. Luke. Activity identification and local linear convergence of Douglas-Rachford under partial smoothness. In *Proceedings of the SSVM 2015*, 2015.

[36] P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.*, 16:964–979, 1979.

[37] O. L. Mangasarian. Sufficiency of exact penalty minimization. *SIAM J. Control Optim.*, 23:30–37, 1985.

[38] J. J. Moreau. Proximité et dualité dans un espace Hilbertian. *Bull. de la Soc. math. de France*, 93(3):273–299, 1965.

[39] R. Nishihara, L. Lessard, B. Recht, A. Packard, and M. Jordan. A general analysis of the convergence of admm. February 2015.

[40] P. Patrinos, L. Stella, and A. Bemporad. Douglas-Rachford splitting: complexity estimates and accelerated variants. 2014.

[41] J.-P. Penot. *Calculus Without Derivatives*. Springer, New York, 2013.

[42] H. Phan. Linear convergence of the Douglas-Rachford method for two closed sets. *Optimization*, accepted.

[43] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.*, 14:87–898, 1976.

[44] R. T. Rockafellar and R. J. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, 1998.

[45] B. F. Svaiter. On weak convergence of the Douglas-Rachford method. *SIAM J. Control. Opt.*, 49(1):280–287, 2011.

[46] B. O'Donoghue T. Goldstein and S. Setzer. Fast alternating direction optimization methods. Technical report, Universität Saarland, 2012.