Non-stationary Douglas–Rachford and alternating direction method of multipliers: adaptive step-sizes and convergence

Dirk A. Lorenz\textsuperscript{1} · Quoc Tran-Dinh\textsuperscript{2}

Received: 27 September 2018 / Published online: 8 May 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract
We revisit the classical Douglas–Rachford (DR) method for finding a zero of the sum of two maximal monotone operators. Since the practical performance of the DR method crucially depends on the step-sizes, we aim at developing an adaptive step-size rule. To that end, we take a closer look at a linear case of the problem and use our findings to develop a step-size strategy that eliminates the need for step-size tuning. We analyze a general non-stationary DR scheme and prove its convergence for a convergent sequence of step-sizes with summable increments in the case of maximally monotone operators. This, in turn, proves the convergence of the method with the new adaptive step-size rule. We also derive the related non-stationary alternating direction method of multipliers. We illustrate the efficiency of the proposed methods on several numerical examples.

Keywords Douglas–Rachford method · Alternating direction methods of multipliers · Maximal monotone inclusions · Adaptive step-size · Non-stationary iteration

Mathematics Subject Classification 90C25 · 65K05 · 65J15 · 47H05

This material was based upon work partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The work of Q. Tran-Dinh was partially supported by the NSF Grant, No. DMS-1619884 (2016–2019).

Dirk A. Lorenz
d.lorenz@tu-braunschweig.de
Quoc Tran-Dinh
quoctd@email.unc.edu

\textsuperscript{1} Institute of Analysis and Algebra, TU Braunschweig, 38092 Braunschweig, Germany
\textsuperscript{2} Department of Statistics and Operations Research, University of North Carolina at Chapel Hill (UNC-Chapel Hill), 333 Hanes Hall, CB# 3260, UNC Chapel Hill, NC 27599-3260, USA
1 Introduction

In this paper we consider the Douglas–Rachford (DR) method to solve the problem of finding a zero of the sum of two maximal monotone operators, i.e. solving:

\[ 0 \in (A + B)x, \]

where \( A, B : \mathcal{H} \rightrightarrows \mathcal{H} \) are two (possibly multivalued) maximal monotone operators from a Hilbert space \( \mathcal{H} \) into itself [36].

The DR method originated from [16] and was initially proposed to solve the discretization of stationary and non-stationary heat equations where the involved monotone operators are linear (namely, the discretization of second derivatives in different spatial directions, for example \( A \approx -\partial_x^2 \) and \( B \approx -\partial_y^2 \)). The iteration uses resolvents \( J_A = (I + A)^{-1} \) (\( I \) is the identity map) and \( J_B = (I + B)^{-1} \), and from the original paper [16, Eqs. (7.4), (7.5)] one can extract the iteration

\[ u^{n+1} := J_tB \left( J_tA \left( (I - tB)u^n \right) + tBu^n \right), \]

where \( t > 0 \) is a given step-size. This iteration scheme also makes sense for general maximal monotone operators as soon as \( B \) is single-valued. It has been observed in [30] that the iteration can be rewritten for arbitrary maximally monotone operators by substituting \( u := J_{tB}y \) and using the identity

\[ tBJ_{tB}y = tB(I + tB)^{-1}y = y - (I + tB)^{-1}y = y - J_{tB}y \]

to get

\[ y^{n+1} := y^n + J_{tA} \left( 2J_{tB}y^n - y^n \right) - J_{tB}y^n. \]

Comparing (2) and (4), we see that (4) does not require to evaluate \( Bu \), which avoids assuming that \( B \) is single-valued as in (2). Otherwise, \( u^{n+1} \) is not uniquely defined. While \( \{u^n\} \) in (2) converges to a solution \( x^* \) of (1), \( \{y^n\} \) in (4) is just an intermediate sequence converging to \( y^* \) such that \( u^* = (I + tB)^{-1}y^* \) is a solution of (1). Therefore, (2) gives us a convenient form to study the DR method in the framework of fixed-point theory. Note that the iterations (2) and (4) are equivalent in the stationary case, but they are not equivalent in the non-stationary case, i.e. when the step-size \( t \) varies along the iterations; we will shed more light on this later in Sect. 2.2.

From a practical point of view, the performance of a DR scheme mainly depends on the following two aspects:

1. **Good step-size \( t \):** It seems to be generally acknowledged that the choice of the step-size is crucial for the algorithmic performance of the method but a general rule to choose the step-size seems to be missing [2,17]. So far, convergence theory of DR methods provides some theoretical guidance to select the parameter \( t \) in a given range in order to guarantee convergence of the method. Such a choice is often globally fixed for all iterations, and does not take into account local structures of the underlying operators \( A \) and \( B \). Moreover, the global convergence rate of the DR method is known to be \( O(1/n) \) under only monotonicity assumptions,
but often using an averaging sequence \([14,15,25]\), where \(n\) is the iteration counter. Several experiments have shown that DR methods have better practical rate than its theoretical bound \([34]\) by using the last iterate (i.e. not an averaging sequence). In the special case of convex optimization problems, the Douglas–Rachford method is equivalent to the alternating direction methods of multipliers (ADMM) (see, e.g., \([22]\), a recent paper for a short historical account) and there a several proposals for dynamic step-sizes for ADMM \([26,29,38,42]\) but we are not aware of a method that applies to DR in the case of monotone operators. The recent works \([13,19,31]\) provide explicit choices for constant step-sizes in cases where the monotone operators posses further properties such as Lipschitz continuity, cocoercivity, or strong monotonicity.

2. **Proper metric:** Since the DR method is not invariant as the Newton method, the choice of metric and preconditioning operator seems to be crucial to accelerate its performance. Some researchers have been studying this aspect from different views, see, e.g., \([8,10,13,20,21,24,35]\). Clearly, the choice of a metric and a preconditioner also affects the choice of the step-size.

Note that a metric choice often depends on the variant of methods, while the choice of step-size depends on the problem structures such as the strong monotonicity parameters and the Lipschitz constants \([31]\). In general cases, where \(A\) and \(B\) are only monotone, we only have a general rule to select the parameter \(t\) to obtain its sublinear convergence rate \([15,17,25]\). This step-size depends on the mentioned global parameters only and does not adequately capture the local structure of \(A\) and \(B\) to adaptively update \(t\). For instance, a linesearch procedure to evaluate a local Lipschitz constant for computing step-size in first-order methods can beat the optimal step-size using global Lipschitz constant \([5]\), or a Barzilai–Borwein step-size in gradient descent methods essentially exploits local curvature of the objective function to obtain a good performance.

**Our contribution:** We prove the convergence of a new version of the non-stationary Douglas–Rachford method for the case where both operators are merely maximally monotone. Moreover, we propose a very simple adaptive step-size rule and demonstrate that this rule does improve convergence in practical situations. We also transfer our results to the case of ADMM and obtain a new adaptive rule that outperforms previously known adaptive ADMM methods and also does have a convergence guarantee. Our step-size rule is relatively simple and does not incur significantly computational effort rather than the norm of two vectors. It also has a theoretical convergence guarantee.

**Paper organization:** We begin with an analysis of the case of linear monotone operators in Sect. 2, analyze the convergence of the non-stationary form of the iteration (2), i.e. the form where \(t = t_n\) varies with \(n\), in Sect. 3, and then propose adaptive step-size rules in Sect. 4. Section 5 extends the analysis to non-stationary ADMM. Finally, Sect. 6 provides several numerical experiments for the DR scheme and ADMM using our new step-size rules.
1.1 State of the art

There are several results on the convergence of the iteration (4). The seminal paper [30] showed that, for any positive step-size $t$, the iteration map in (4) is firmly non-expansive, that the sequence $\{y^n\}$ weakly converges to a fixed point of the iteration map [30, Prop. 2] and that, $\{u^n = J_t B y^n\}$ weakly converges to a solution of the inclusion (1) as soon as $A$, $B$, and $A+B$ are maximal monotone [30, Theorem 1]. In the case where $B$ is coercive and Lipschitz continuous, linear convergence was also shown in [30, Proposition 4]. These results have been extended in various ways. Let us attempt to summarize some key contributions on the DR method. Eckstein and Bertsekas in [17] showed that the DR scheme can be cast into a special case of the proximal point method [36]. This allows the authors to exploit inexact computation from the proximal point method [36]. They also presented a connection between the DR method and the alternating direction method of multipliers (ADMM). In [40] Svaiter proved a weak convergence of the DR method in Hilbert space without the assumption that $A + B$ is maximal monotone and the proof has been simplified in [39]. Combettes [11] casts the DR method as special case of the averaging operator from a fixed-point framework. Applications of the DR method have been studied in [12]. The convergence rate of the DR method was first studied in [30] for the strongly monotone case, while the sublinear rate was then proved in [25]. A more intensive research on convergence rates of the DR methods can be found in [14,15,31,32]. The DR method has been extended to accelerated variant in [34] but specifying for a special setting. In [28] the authors analyzed a non-stationary DR method derived from (4) in the case of convex minimization problems in the framework of perturbations of non-expansive iterations and showed convergence for convergent step-size sequences with summable errors. Their conditions on the variable step-sizes are equivalent to ours, but their analysis is for a different non-stationary scheme (ours is derived from (2) instead of (4) as in [28]).

The DR method together with its dual variant, ADMM, become extremely popular in recent years due to a wide range of applications in image processing, and machine learning [6,27], which are unnecessary to recall them all here.

In terms of step-size selection for DR schemes as well as for ADMM methods, it seems that there is very little work available from the literature. Some general rules for fixed step-sizes based on further properties of the operators such as strong monotonicity, Lipschitz continuity, and coercivity are given in [19,31], and it is shown that the resulting linear rates are tight. Heuristic rules for fixed step-sizes motivated by quadratic problems are derived in [21]. A self-adaptive step-size for ADMM proposed in [26] seems to be one of the first works in this direction. The recent works [42,43] also proposed an adaptive update rule for step-size in ADMM based on a spectral estimation. Some other papers rely on theoretical analysis to choose optimal step-size such as [18], but it only works in the quadratic case. In [29], the authors proposed a nonincreasing adaptive rule for the penalty parameter in ADMM. Another update rule for ADMM can be found in [38]. While ADMM is a dual variant of the DR scheme, we unfortunately have not seen any work that converts such an adaptive step-size from ADMM to the DR scheme where the more general case of monotone operators can be
Non-stationary Douglas–Rachford and alternating direction…

handled. In addition, the adaptive step-size for the DR scheme by itself seems to not exist in the literature.

1.2 A motivating linear example

While the Douglas–Rachford iteration (weakly) converges for all positive step-sizes $t > 0$, it seems to be folk wisdom, that there is a “sweet spot” for the step-size which leads to fast convergence. We illustrate this effect with a simple linear example. We consider a linear equation

$$0 = Ax + Bx,$$

where $A, B \in \mathbb{R}^{m \times m}$ are two matrices of the size $m \times m$ with $m = 200$. We choose symmetric positive semi-definite matrices with $\text{rank}(A) = \frac{m}{2} + 10$ and $\text{rank} B = \frac{m}{2}$ such that $A + B$ has full rank, and thus the equation $0 = Ax + Bx$ has zero as its unique solution.\(^1\) Since $B$ is single-valued, we directly use the iteration (2).

**Remark 1.1** Note that the shift $\tilde{B}x = Bx - y$ would allow to treat the inhomogeneous equation $(A + B)x = y$. If $x^*$ is a solution of this equation, then one sees that iteration (2) applied to $A + \tilde{B}$ is equivalent to applying the iteration to $A + B$ but for the residual $x - x^*$.

We ran the DR scheme (2) for a given range of different values of $t > 0$, and show the residuals $\| (A + B)x^n \|$ in semi-log-scale on the left of Fig. 1. One observes the following typical behavior for this example:

- A not too small step-size ($t = 0.5$ in this case) leads to good progress in the beginning, but slows down considerably in the end.
- Large step-sizes (larger than 2 in this case) are slower in the beginning and tend to produce non-monotone decrease of the error.
- Somewhere in the middle, there is a step-size which performs much better than the small and large step-sizes.

In this particular example the step-size $t = 1.5$ greatly outperforms the other step-sizes. On the right of Fig. 1 we show the norm of the residual after a fixed number of iterations for varying step-sizes.

One can see that there is indeed a sweet spot for the step-sizes around $t = 1.5$. Note that the value of $t = 1.5$ is by no means universal and this sweet spot of 1.5 varies with the problem size, with the ranks of $A$ and $B$, and even for each particular instance of this linear example.

2 Analysis of the linear monotone inclusion

In order to develop an adaptive step-size for our non-stationary DR method, we first consider the linear problem instance of (1). We consider the original DR scheme (2)

\(^1\) The exact construction of $A$ and $B$ is $A = C^T C$ and $B = D^T D$, where $C \in \mathbb{R}^{(0.5m+10) \times m}$ and $D \in \mathbb{R}^{0.5m \times m}$ are drawn from the standard Gaussian distribution in Matlab.
instead of (4) since (2) generates the sequence \( \{u^n\} \) which converges to a solution of (1), while the sequence \( \{y^n\} \) computed by (4) does not converge to a solution and its limit does depend on the step-size in general.

2.1 The construction of adaptive step-size for single-valued operator \( B \)

When both \( A \) and \( B \) are linear, the DR scheme (2) can be expressed as a fixed-point iteration scheme of the following mapping:

\[ H_t := J_t B \left( J_t A \left( I - tB \right) + tB \right) \]
\[ = \left( I + tB \right)^{-1} \left( I + tA \right)^{-1} \left( I + t^2 AB \right) \]
\[ = \left( I + tA + tB + t^2 AB \right)^{-1} \left( I + t^2 AB \right). \] (6)

Recall that, by Remark 1.1, all of this section also applies not only to problem \((A + B)x = 0\) but also problem \((A + B)x = y\). The notion of a monotone operator has a natural equivalence for matrices, which is, however, not widely used. Hence, we recall that a matrix \( A \in \mathbb{R}^{m \times m} \) is called monotone, if, for all \( x \in \mathbb{R}^m \), it holds that \( \langle x, Ax \rangle \geq 0 \). Note that any symmetric positive semidefinite (spd) matrix is monotone, but a monotone matrix is not necessarily spd. For example, the following matrices are monotone but not spd:

\[ A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad \text{and} \quad A = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \quad \text{with} \ |t| \leq 2. \]

The first matrix is skew symmetric, i.e., \( A^T = -A \) and any such matrix is monotone. Note that even if \( A \) and \( B \) are spd (as in our example in Sect. 1.2), the iteration map \( H_t \) in (6) is not even symmetric. Consequently, the asymptotic convergence rate of the iteration scheme (2) is not governed by the norm of \( H_t \) but by its spectral
radius $\rho(H_t)$, which is the largest magnitude of an eigenvalue of $H_t$ (cf. [23, Theorem 11.2.1]). Moreover, the eigenvalues and eigenvectors of $H_t$ are complex in general.

First, it is clear from the derivation of $H_t$ that the eigenspace of $H_t$ for the eigenvalue $\lambda = 1$ exactly consists of the solutions of $(A + B)x = 0$.

In the following, for any $z \in \mathbb{C}$ (the set of complex numbers) and $r > 0$, we denote by $B_r(z)$ the ball of radius $r$ centered at $z$. We estimate the eigenvalues of $H_t$ that are different from 1.

**Lemma 2.1** Let $A, B \in \mathbb{R}^{n \times n}$ be monotone, and $H_t$ be defined by (6). Let $\lambda \in \mathbb{C}$ be an eigenvalue of $H_t$ with the corresponding eigenvector $z \in \mathbb{C}^n$. Assume that $\lambda \neq 1$ and define $c$ by

$$c := \frac{\Re(\langle Bz, z \rangle)}{t^{-1}z^2 + t\|Bz\|^2}. \quad (7)$$

Then, we have $c \geq 0$ and

$$|\lambda - \frac{1}{2}| \leq \sqrt{\frac{1}{4} - \frac{c}{1 + 2c}} \leq \frac{1}{2},$$

i.e. $\lambda \in B_{\frac{1}{2}}(\frac{1}{2})$, where $\Re(u)$ is the real part of a complex number $u$.

**Proof** Note that for a real, linear, and monotone map $M$, and a complex vector $a = b + ic$, it holds that $\langle Ma, a \rangle = \langle Mb, b \rangle + \langle Mc, c \rangle + i(\langle M^T - M \rangle b, c)$ and thus, $\Re(\langle Ma, a \rangle) \geq 0$. This shows that $c \geq 0$.

We can see from (6) that any pair $(\lambda, z)$ of eigenvalue and eigenvector of $H_t$ fulfills

$$z + t^2ABz = \lambda \left(z + tAz + tBz + t^2ABz\right).$$

Now, if we denote $u := Bz$, then this expression becomes

$$z + t^2Au = \lambda z + \lambda tzA + \lambda tu + \lambda t^2Au,$$

which, by rearranging, leads to

$$-(\lambda - 1)z - \lambda tu = tA(\lambda z + (\lambda - 1)tu).$$

Hence, by monotonicity of $tA$, we can derive from the above relation that

$$0 \leq \Re(\langle \lambda z + (\lambda - 1)tu, -(\lambda - 1)z - \lambda tu \rangle)$$

$$= -\Re(\lambda(\lambda - 1)) \|z\|^2 - (|\lambda|^2 + |\lambda - 1|^2) t \Re(\langle u, z \rangle) - \Re((\lambda - 1)\bar{\lambda}) t^2 \|u\|^2.$$

This leads to

$$\left(|\lambda|^2 + |\lambda - 1|^2\right) \Re(\langle u, z \rangle) \leq \frac{\Re(\lambda - |\lambda|^2)}{t} \|z\|^2 + \Re(\left(\bar{\lambda} - |\lambda|^2\right)) t \|u\|^2.$$
Denoting $\lambda := x + iy \in \mathbb{C}$, the last expression reads as
\[
\left(x^2 + (x - 1)^2 + 2y^2\right) \Re(\langle u, z \rangle) \leq \left(x - x^2 - y^2\right) \left(\frac{\|z\|^2}{t} + t\|u\|^2\right).
\]
Recalling the definition of $c$ in (7), we get
\[
\left(x^2 + (x - 1)^2 + 2y^2\right) c \leq x - x^2 - y^2.
\]
This is equivalent to
\[
0 \leq x - x^2 - y^2 - cx^2 - c(x - 1)^2 - 2cy^2 = (1 + 2c) \left(x - x^2 - y^2\right) - c,
\]
which, in turn, is equivalent to $x^2 - x + y^2 \leq -\frac{c}{1 + 2c}$. Adding $\frac{1}{4}$ to both sides, it leads to $(x - \frac{1}{2})^2 + y^2 \leq \frac{1}{4} - \frac{c}{1 + 2c}$, which shows the desired estimate. $\square$

In general, the eigenvalues of $H_t$ depend on $t$ in a complicated way. For $t = 0$, we have $H_0 = I$ and hence, all eigenvalues are equal to one. For growing $t > 0$, some eigenvalues move into the interior of the circle $B_{1/2}(1/2)$ and for $t \to \infty$, it seems that all the eigenvalues tend to converge to the boundary of such a circle, see Fig. 2 for an illustration of the eigenvalues distribution.

**Remark 2.2** It appears that Lemma 2.1 is related to Proposition 4.10 of [3] and also to the fact that the iteration mapping $H_t$ is (in the general nonlinear case) known to be not only non-expansive, but **firmly non-expansive** (cf. [17, Lemma 1] and [17, Figure 1]). In general, firm non-expansiveness allows over-relaxation of the method, and indeed, one can also easily see this in the linear case as well: If $\lambda$ is an eigenvalue of $H_t$, then it lies in $B_{1/2}(1/2)$ (when it is not equal to one) and the corresponding eigenvalue $\lambda_\rho$ of the relaxed iteration map
\[
H_\rho = (1 - \rho)I + \rho H_t
\]
is \( \lambda_\rho = 1 - \rho + \rho \lambda \) and lies in \( \mathbb{B}_{\rho/2}(1 - \frac{\rho}{2}) \). Therefore, for \( 0 \leq \rho \leq 2 \) all eigenvalues different from one of the relaxed iteration

\[
  u^{n+1} = (1 - \rho)u^n + \rho H_t u^n
\]
lie in a circle of radius \( \rho/2 \) centered at \( 1 - \rho/2 \), and hence, the iteration is still non-expansive. It is known that relaxation can speed up convergence, but we will not investigate this in this paper.

Lemma 2.1 tells us a little more than that all the eigenvalues of the iteration map \( H_t \) lie in a circle centered at \( \frac{1}{2} \) of radius \( \frac{1}{2} \). Especially, all the eigenvalues except for \( \lambda = 1 \) have magnitude strictly smaller than one if \( \text{Re}(\langle Bz, z \rangle) > 0 \) for all the corresponding eigenvectors \( z \). This implies that the iteration map \( H_t \) is indeed asymptotically contracting outside the set of solutions \( \{ x^* \in \mathcal{H} \mid (A + B)x^* = 0 \} \) of (1). This proves that the stationary iteration \( u^{n+1} = H_t u^n \) converges to a zero point of the map \( A + B \) at a linear rate. Note that this does not imply the convergence in the non-stationary case.

To optimize the convergence speed, we aim at minimizing the spectral radius of \( H_t \), which is the magnitude of the largest eigenvalue of \( H_t \) and there seems to be little hope to explicitly minimize this quantity.

Here is a heuristic argument based on Lemma 2.1, which we will use to derive an adaptive step-size rule: Note that \( c \mapsto \frac{c}{1 + 2c} \) is increasing and hence, to minimize the upper bound on \( \lambda \) (more precisely: the distance of \( \lambda \) to \( \frac{1}{2} \)) we want to make \( c \) from (7) as large as possible. This is achieved by minimizing the denominator of \( c \) over \( t \) which happens for

\[
  t = \frac{\|z\|}{\|Bz\|}.
\]

This gives \( c = \text{Re}(\langle Bz, z \rangle)/(2\|z\|\|Bz\|) \) and note that \( 0 \leq c \leq \frac{1}{2} \) (which implies \( 0 \leq \frac{c}{1 + 2c} \leq \frac{1}{4} \)). This motivates an adaptive choice for the step-size \( t_n \) as

\[
  t_n := \frac{\|u^n\|}{\|Bu^n\|},
\]
in the Douglas–Rachford iteration scheme (2).

Remark 2.3 One can use the above derivation to deduce that \( t = 1/\|B\| \) is a good constant step-size. In fact, this is also the step-size that gives the best linear rate derived in [30, Proposition 4], which is minimized when \( t = 1/M \) where \( M \) is the Lipschitz constant of \( B \). However, this choice does not perform well in practice in our experiments.

Since little is known about the non-stationary Douglas–Rachford iteration in general (besides the result from [28] on convergent step-sizes with summable errors), we turn to an investigation of this method in Sect. 3. Before we do so, we generalize the heuristic step-size to the case of multivalued \( B \).
2.2 The construction of adaptive step-size for non-single-valued $B$

In the case of multi-valued $B$, one needs to apply the iteration (4) instead of (2). To motivate an adaptive choice for the step-size in this case, we again consider situation of linear operators.

In the linear case, the iteration (4) is given by the iteration matrix

$$ F_t = J_t A (2J_t B - I) - J_t B + I. $$

Comparing this with the iteration map $H_t$ from (6) (corresponding to (2)) one notes that

$$ F_t = (I + tB) H_t (I + tB)^{-1}, $$

i.e. the matrices $F_t$ and $H_t$ are similar and hence, have the same eigenvalues. Moreover, if $z$ is an eigenvector of $H_t$ with the eigenvalue $\lambda$, then $(I + tB)z$ is an eigenvector of $F_t$ for the same eigenvalue $\lambda$. However, in the case of the iteration (4) we do not assume that $B$ is single-valued, and thus the adaptive step-size using the quotient $\|u\|/\|Bu\|$ cannot be used. However, again due to (3), we can rewrite this quotient without applying $B$ and get, with $J_t B y = u$, that

$$ \frac{\|u\|}{\|Bu\|} = \frac{\|J_t B y\|}{\|y - J_t B y\|} = t \frac{\|J_t B y\|}{\|y - J_t B y\|}. \quad (9) $$

Note that the two iteration schemes (2) and (4) are not equivalent in the non-stationary and non-linear case. Indeed, let us consider $y^n$ such that $u^n := J_{tn-1} B y^n$. By induction, we have $u^{n+1} = J_{tn} B y^{n+1}$. Substituting $u^{n+1}$ into (2), we obtain

$$ y^{n+1} = J_{tn} A (u^n - t_n Bu^n) + t_n Bu^n. \quad (10) $$

From (3) we have

$$ Bu^n = B J_{tn-1} B y^n = \frac{1}{t_{n-1}} (y^n - J_{tn-1} B y^n). $$

Substituting $u^n = J_{tn-1} B y^n$ and $Bu^n$ into (10), we obtain

$$ y^{n+1} = J_{tn} A \left( J_{tn-1} B y^n - \frac{J_{tn-1}}{t_{n-1}} (y^n - J_{tn-1} B y^n) \right) + \frac{J_{tn}}{t_{n-1}} (y^n - J_{tn-1} B y^n) $$

$$ = \frac{1}{t_{n-1}} y^n + J_{tn} A \left( (1 + \frac{J_{tn}}{t_{n-1}}) J_{tn-1} B y^n - \frac{J_{tn}}{t_{n-1}} y^n \right) - \frac{J_{tn}}{t_{n-1}} J_{tn-1} B y^n. $$

Updating $t_n$ by (9) would then give

$$ t_n := \kappa_n t_{n-1}, \quad \text{where} \quad \kappa_n := \frac{\|J_{tn-1} B y^n\|}{\|y^n - J_{tn-1} B y^n\|}. $$

If we would have derived a non-stationary scheme from (4), it would have been

$$ y^{n+1} := y^n + J_{tn} A \left( 2J_{tn} B y^n - y^n \right) - J_{tn} B y^n \quad (11) $$
which is notably different from our scheme and has been studied in [28]. In summary, we can write an alternative DR scheme for solving (1) as

\[
\begin{align*}
\{ u^n \} &:= J_{t_{n-1}} y^n, \\
\kappa_n &: = \frac{\| u^n \|}{\| u^n - y^n \|}, \\
\eta_n &: = \kappa_n t_{n-1}, \\
\nu^n &: = J_{t_n} \left( (1 + \kappa_n)u^n - \kappa_n y^n \right), \\
y_{n+1} &: = \nu^n + \kappa_n (y_n - u^n). \\
\end{align*}
\]

(12)

This scheme essentially has the same per-iteration complexity as in the standard DR method since the computation of \( \kappa_n \) does not significantly increase the cost. Compare (12) and (11), we can see that (12) is different from (11) due to the relaxation parameter \( \kappa_n \) instead of 1. To the best of our knowledge, the scheme (12) is new.

3 Convergence of the non-stationary DR method

In this section, we prove weak convergence of the new non-stationary scheme (12). We follow the approach in [39,40] and restate the DR iteration as follows: Given \((u^0, b^0)\) such that \(b^0 \in B(u^0)\) and a sequence \(\{t_n\}_{n \geq 0}\), at each iteration \(n \geq 0\), we iterate

\[
\begin{align*}
\{ a^n \} & \in A(v^n), \quad v^n + t_n a^n = u^{n-1} - t_n b^{n-1}, \\
\{ b^n \} & \in B(u^n), \quad u^n + t_n b^n = v^n + t_n b^{n-1}.
\end{align*}
\]

(13)

Note that, in the case of single-valued \(B\), this iteration reduces to

\[
u^n = J_{t_n} \left( J_{t_n} \left( u^{n-1} - t_n B u^{n-1} \right) + t_n B u^{n-1} \right),
\]

and this scheme can, as shown in Sect. 2.2, be transformed into the non-stationary iteration scheme (12).

Below are some consequences which we will need in our analysis:

\[
\begin{align*}
u^{n-1} - u^n &= t_n \left( a^n + b^n \right), \\
t_n \left( b^{n-1} - b^n \right) &= u^n - v^n, \\
u^n - v^n + t_n \left( a^n + b^n \right) &= t_n \left( a^n + b^{n-1} \right) = u^{n-1} - v^n.
\end{align*}
\]

(14)

(15)

(16)

Before proving our convergence result, we state the following lemma.

**Lemma 3.1** Let \(\{\alpha_n\}, \{\beta_n\}, \) and \(\{\omega_n\}\) be three nonnegative sequences, and \(\{\tau_n\}\) be a bounded sequence such that for \(n \geq 0\):

\[
0 < \tau \leq \tau_n \leq \bar{\tau}, \quad |\tau_n - \tau_{n+1}| \leq \omega_n, \quad \text{and} \quad \sum_{n=0}^{\infty} \omega_n < \infty.
\]
If \( \alpha_{n-1} + \tau_n \beta_{n-1} \geq \alpha_n + \tau_n \beta_n \), then \( \{\alpha_n\} \) and \( \{\beta_n\} \) are bounded.

**Proof** If \( \tau_{n+1} \leq \tau_n \), then

\[
\alpha_{n-1} + \tau_n \beta_{n-1} \geq \alpha_n + \tau_n \beta_n \geq \alpha_n + \tau_{n+1} \beta_n.
\]

If \( \tau_{n+1} \geq \tau_n \), then \( \frac{\tau_n}{\tau_{n+1}} \leq 1 \) and

\[
\alpha_{n-1} + \tau_n \beta_{n-1} \geq \alpha_n + \tau_n \beta_n \geq \frac{\tau_n}{\tau_{n+1}} \alpha_n + \tau_n \beta_n = \frac{\tau_n}{\tau_{n+1}} (\alpha_n + \tau_{n+1} \beta_n).
\]

By the assumption that \( \frac{\tau_n}{\tau_{n+1}} \geq 1 - \frac{\omega_n}{\tau} \) and, without loss of generality, we assume that the latter term is positive (which is fulfilled for \( n \) large enough, because \( \omega_n \to 0 \)). Thus, in both cases, we can show that

\[
\alpha_{n-1} + \tau_n \beta_{n-1} \geq \left(1 - \frac{\omega_n}{\tau}\right) (\alpha_n + \tau_{n+1} \beta_n).
\]

Recursively, we get

\[
\alpha_0 + \tau_1 \beta_0 \geq \prod_{l=1}^n \left(1 - \frac{\omega_l}{\tau}\right) (\alpha_n + \tau_{n+1} \beta_n).
\]

Under the assumption \( \sum_{n=0}^{\infty} \omega_n < +\infty \), we have \( \prod_{l=1}^n \left(1 - \frac{\omega_l}{\tau}\right) \geq M \) for some \( M > 0 \) and all \( n \geq 1 \). Then, we have \( \alpha_n + \tau_{n+1} \beta_n \leq \frac{1}{M} (\alpha_0 + \tau_1 \beta_0) \). This shows that \( \{\alpha_n + \tau_{n+1} \beta_n\} \) is bounded. Since \( \{\alpha_n\}, \{\beta_n\}, \) and \( \{\tau_n\} \) are all nonnegative, it implies that \( \{\alpha_n\} \) and \( \{\beta_n\} \) are bounded. \( \square \)

**Theorem 3.2** (Convergence of non-stationary DR) Let \( A \) and \( B \) be maximally monotone and \( \{t_n\} \) be a positive sequence such that

\[
0 < \underline{t} \leq t_n \leq \bar{t} \quad \text{and} \quad \sum_{n=0}^{\infty} |t_n - t_{n+1}| < \infty,
\]

where \( 0 < \underline{t} \leq \bar{t} \leq +\infty \) are given. Then, the sequence \( \{ (u^n, b^n) \} \) generated by the iteration scheme (13) weakly converges to some \( (u^*, b^*) \) in the extended solution set \( S(A, B) = \{(z, w) | w \in B(z), -w \in A(z)\} \) of (1), so in particular, \( 0 \in (A+B)(u^*) \).

**Proof** The proof of this theorem follows from the proof of [40, Theorem 1]. First, we observe that, for any \( (u, b) \in S(A, B) \), we have

\[
\left( u^{n-1} - u^n, u^n - u \right) = t_n \left( a^n + b^n, u^n - u \right)
\]

by (14)

\[
= t_n \left[ (a^n + b, u^n - u) + (b^n - b, u^n - u) \right]
\]

\[
\geq t_n \left( a^n + b, u^n - u \right)
\]

\[
= t_n \left[ (a^n + b, u^n - v^n) + (a^n + b, v^n - u) \right]
\]

\[
\geq t_n \left( a^n + b, u^n - v^n \right).
\]

\( B \) is monotone\( A \) is monotone
From this and (15) it follows that
\[
\left\langle u^{n-1} - u^n, u^n - u \right\rangle + t_n^2 \left\langle b^{n-1} - b^n, b^n - b \right\rangle \geq t_n \left\langle a^n + b, u^n - v^n \right\rangle \\
+ t_n \left\langle u^n - v^n, b^n - b \right\rangle = t_n \left\langle a^n - v^n, a^n + b^n \right\rangle.
\]

Moreover, by (14) and (15) it holds that
\[
\|u^{n-1} - u^n\|^2 + t_n^2 \|b^{n-1} - b^n\|^2 = t_n^2 \|a^n + b^n\|^2 + \|u^n - v^n\|^2,
\]
and thus
\[
\|u^{n-1} - u\|^2 + t_n^2 \|b^{n-1} - b\|^2 = \|u^{n-1} - u^n + u^n - u\|^2 + t_n^2 \|b^{n-1} - b^n + b^n - b\|^2 \\
= \|u^{n-1} - u^n\|^2 + 2 \left\langle u^{n-1} - u^n, u^n - u \right\rangle + \|u^n - u\|^2 \\
+ t_n^2 \|b^{n-1} - b^n\|^2 + 2 \left\langle b^{n-1} - b^n, b^n - b \right\rangle + \|b^n - b\|^2 \\
\geq t_n^2 \|a^n + b^n\|^2 + \|u^n - v^n\|^2 + 2 t_n \left\langle u^n - v^n, a^n + b^n \right\rangle \\
+ \|u^n - u\|^2 + t_n^2 \|b^n - b\|^2 \\
= \|u^n - u\|^2 + t_n^2 \|b^n - b\|^2 + \|u^n - v^n + t_n (b^n + a^n)\|^2.
\]

We see from (17) that
\[
\|u^{n-1} - u\|^2 + t_n^2 \|b^{n-1} - b\|^2 \geq \|u^n - u\|^2 + t_n^2 \|b^n - b\|^2,
\]
and using Lemma 3.1 with \(\alpha_n = \|u^n - u\|^2\), \(\tau_n = t_n^2\) and \(\beta_n = \|b^n - b\|^2\), we can conclude that both sequences \(\{\|u^n - u\|\}\) and \(\{\|b^n - b\|\}\) are bounded.

Again from (17) we can deduce using (16) that
\[
\|u^{n-1} - u\|^2 + t_n^2 \|b^{n-1} - b\|^2 \geq \|u^n - u\|^2 + t_n^2 \|b^n - b\|^2 + \|u^{n-1} - v^n\|^2 \\
= \|u^n - u\|^2 + t_n^2 \|b^n - b\|^2 + t_n^2 \|a^n + b^{n-1}\|^2.
\]

The first line gives
\[
\|u^{n-1} - u\|^2 + t_n^2 \|b^{n-1} - b\|^2 \geq \|u^n - u\|^2 + t_{n+1}^2 \|b^n - b\|^2 + \|u^{n-1} - v^n\|^2 \\
+ \left( t_n^2 - t_{n+1}^2 \right) \|b^n - b\|^2.
\]
Summing this inequality from \( n = 1 \) to \( n = N \), we get

\[
\sum_{n=1}^{N} \|u^{n-1} - v^n\|^2 \leq \|u^0 - u\|^2 + t_1^2 \|b^0 - b\|^2 - \left( \|u^N - u\|^2 + t_{N+1}^2 \|b^N - b\|^2 \right) + \sum_{n=1}^{N} \left( t_{n+1}^2 - t_n^2 \right) \|b^n - b\|^2.
\]

Now, since \( \|b^n - b\|^2 \) is bounded and it holds that

\[
\sum_{n=1}^{\infty} |t_n^2 - t_{n+1}^2| = \sum_{n=1}^{\infty} |t_n - t_{n+1}| |t_n + t_{n+1}| \leq 2T \sum_{n=1}^{\infty} |t_n - t_{n+1}| < \infty
\]

by our assumption, we can conclude that

\[
\sum_{n=1}^{\infty} \|u^{n-1} - v^n\|^2 < \infty,
\]

i.e., by (16), we have

\[
\lim_{n \to \infty} (u^{n-1} - v^n) = \lim_{n \to \infty} (a^n + b^{n-1}) = 0.
\]

This expression shows that \( v^n \) and \( a^n \) are also bounded. Due to the boundedness of \( \{(u^n, b^n)\} \), we conclude the existence of weak convergence subsequences \( \{u_{n_l}\}_l \) and \( \{b_{n_l}\}_l \), such that

\[
u^{n_l} \to u^*, \quad b^{n_l} \to b^* ,
\]

and by the above limits, we also have

\[
u^{n_l+1} \to u^*, \quad a^{n_l+1} \to b^* .
\]

From [1, Corollary 3] it follows that \( (u^*, b^*) \in S(A, B) \). This shows that \( \{(u^n, b^n)\} \) has a weak cluster point and that all such points are in \( S(A, B) \).

Note that the condition \( \sum_{n=0}^{\infty} |t_n - t_{n+1}| < \infty \) implies that \( \lim_{n \to \infty} t_n = t^* < \infty \). Now, we deduce from (18) that

\[
\|u^n - u^*\|^2 + (t^*)^2 \|b^n - b^*\|^2 \leq \|u^{n-1} - u^*\|^2 + (t^*)^2 \|b^{n-1} - b^*\|^2 + |t_n^2 - (t^*)^2| \left( \|b^{n-1} - b^*\|^2 - \|b^n - b^*\|^2 \right).
\]

Since \( \|b^n - b^*\|^2 \) is bounded and \( t^n \to t^* \), this shows that the sequence \( \{(u^n, b^n)\} \) is quasi-Fejer convergent (in the sense of [2], Definition 5.32) to the extended solution set \( S(A, B) \) with respect to the distance \( d((u, b), (z, w)) = (\|u - z\|^2 + (t^*)^2 \|b - w\|^2)^{1/2} \). Thus, similar to the proof of [40, Theorem 1], we conclude that the whole sequence \( \{(u^n, b^n)\} \) weakly converges to an element of \( S(A, B) \).
4 An adaptive step-size for DR methods

The step-size $t_n$ suggested by (8) or by (9) is derived from our analysis of a linear case and it does not guarantee the convergence in general. In this section, we suggest modifying this step-size so that we can prove the convergence of the DR scheme. We build our adaptive step-size based on two insights:

- The estimates of the eigenvalues of the DR-iteration in the linear case from Sect. 2.1 motivated the adaptive step-size

$$t_n = \frac{\|u^n\|}{\|Bu^n\|}$$

for single-valued $B$ is single-valued and for the general case, we consider

$$t_n = \frac{\|J_{t_n-1}By^n\|}{\|y^n - J_{t_n-1}By^n\|} t_n^{-1}$$

from Sect. 2.2.

- Theorem 3.2 ensures the convergence of the non-stationary DR-iteration as soon as the step-size sequence is convergent with summable increments. However, the sequences (19) and (20) are not guaranteed to converge (and numerical experiments indicate that, indeed, divergence may occur). Here is a way to adapt the sequence (19) to produce a suitable step-size sequence in the single-valued case:

1. Choose safeguards $0 < t_{\min} < t_{\max} < \infty$, a summable “conservation sequence” $\omega_n \in (0, 1]$ with $\omega_0 = 1$ and start with $t_0 = 0$.

2. Let $\text{proj}_{[\gamma, \rho]}(\cdot)$ be the projection onto a box $[\gamma, \rho]$. We construct $\{t_n\}$ as

$$t_n = (1 - \omega_n)t_{n-1} + \omega_n \text{proj}_{[t_{\min}, t_{\max}]}\left( \frac{\|u^n\|}{\|Bu^n\|} \right).$$

The following lemma ensures that this will lead to a convergent sequence $\{t_n\}$.

**Lemma 4.1** Let $\{\alpha_n\}$ be a bounded sequence, i.e. $\underline{\alpha} \leq \alpha_n \leq \bar{\alpha}$, and $\{\omega_n\} \subset (0, 1]$ such that $\sum_{n=0}^{\infty} \omega_n < \infty$ and $\omega_0 = 1$. Then, the sequence $\{\beta_n\}$ defined by $\beta_0 = 0$ and

$$\beta_n = (1 - \omega_n) \beta_{n-1} + \omega_n \alpha_n,$$

is in $[\underline{\alpha}, \bar{\alpha}]$ and converges to some $\beta^*$ and it holds that $\sum_{n=0}^{\infty} |\beta_{n+1} - \beta_n| < \infty$.

**Proof** Obviously, $\beta_0 = \alpha_0$ and since $\beta_n$ is a convex combination of $\alpha_n$ and $\beta_{n-1}$, one can easily see that $\beta_n$ obeys the same bounds as $\alpha_n$, i.e. $\underline{\alpha} \leq \beta_n \leq \bar{\alpha}$. Moreover, it holds that

$$\beta_n - \beta_{n-1} = \omega_n \alpha_n + (1 - \omega_n) \beta_{n-1} - \beta_{n-1} = \omega_n (\alpha_n - \beta_{n-1}),$$

thus $|\beta_n - \beta_{n-1}| \leq \omega_n (\bar{\alpha} - \underline{\alpha})$ from which the assertion follows, since $\omega_n$ is summable.
Clearly, if we apply Lemma 4.1 to the sequence \( \{t_n\} \) defined by (21), then it converges to some \( t^* \).

We use a similar trick to construct an adaptive step-size based on the choice (20) in the case of multi-valued operators. More precisely, we construct \( \{t_n\} \) as follows:

1. Choose safeguards \( 0 < \kappa_{\min} < \kappa_{\max} < \infty \), a summable “conservation sequence” \( \{\omega_n\} \subset (0, 1) \), and \( t_0 = 1 \).
2. We construct \( \{t_n\} \) as

\[
\kappa_n := \text{proj}_{[\kappa_{\min}, \kappa_{\max}]} \left( \frac{\|J_{t_{n-1}}B y^{n-1}\|}{\|y^{n-1} - J_{t_{n-1}}B y^{n-1}\|} \right),
\]

\[
I_n := v_n t_{n-1}, \quad \text{where} \quad v_n := 1 - \omega_n + \omega_n \kappa_n.
\]

In this case we get that \( t_n = \prod_{k=1}^{n} v_k t_0 \) and since \( |v_n - 1| = \omega_n |\kappa_n - 1| \) and \( \kappa_n \) is bounded, the summability of \( \omega_n \) implies the summability of \( |v_n - 1| \). This implies that \( \prod_{k=1}^{n} v_k \) converges to some positive value and and thus, \( t_n \to t^* > 0 \), too.

The step-size sequence \( \{t_n\} \) constructed by either (21) or (22) fulfills the conditions of Theorem 3.2. Hence, the convergence of the nonstationary DR scheme using this adaptive step-size follows as a direct consequence. We will provide guidelines on how to choose the safeguards and the conservation sequence in practice in Sect. 6.1.

5 Application to ADMM

It is well-known that the alternating direction method of multipliers (ADMM) for convex optimization with linear constraint can be interpreted as the DR method on its dual problem, see, e.g. [17]. In this section, we apply our adaptive step-size to ADMM to obtain a new variant for solving the following constrained problem:

\[
\min_{u, v} \left\{ \phi(u, v) = \varphi(u) + \psi(v) \mid Du + Ev = c \right\}, \tag{23}
\]

where \( \varphi : \mathcal{H}_u \to \mathbb{R} \cup \{+\infty\} \), \( \psi : \mathcal{H}_v \to \mathbb{R} \cup \{+\infty\} \) are two proper, closed, and convex functions, \( D : \mathcal{H}_u \to \mathcal{H} \) and \( E : \mathcal{H}_v \to \mathcal{H} \) are two given bounded linear operators, and \( c \in \mathcal{H} \).

The dual problem associated with (23) becomes

\[
\min_x \left\{ \varphi^*(D^Tx) + \psi^*(E^Tx) - c^T x \right\}, \tag{24}
\]

where \( \varphi^* \) and \( \psi^* \) are the Fenchel conjugate of \( \varphi \) and \( \psi \), respectively. The optimality condition of (24) becomes

\[
0 \in D\partial \varphi^*(D^Tx) - c + E\partial \psi^*(E^Tx), \tag{25}
\]

which is of the form (1).
In the stationary case, ADMM is equivalent to the DR method applying to the dual problem (25), see, e.g., [17]. However, for the non-stationary DR method, we can derive a different parameter update rule for ADMM. Let us summarize this result into the following theorem for the non-stationary scheme (12). The proof of this theorem is given in Appendix 1.

**Theorem 5.1** Given $0 < t_{\text{min}} < t_{\text{max}} < +\infty$, the ADMM scheme for solving (23) derived from the non-stationary DR method (12) applying to (25) becomes:

\[
\begin{align*}
  u^{n+1} &:= \arg\min_u \left\{ \varphi(u) - \langle Du, w^n \rangle + \frac{t_{n-1}}{2} \| Du + Ev^n - c \|^2 \right\}, \\
  v^{n+1} &:= \arg\min_v \left\{ \psi(v) - \langle Ev, w^n \rangle + \frac{t_{n-1}}{2} \| Du^{n+1} + Ev - c \|^2 \right\}, \\
  w^{n+1} &:= w^n - t_{n-1} \left( Du^{n+1} + Ev^{n+1} - c \right), \\
  t_n &:= (1 - \omega_n) t_{n-1} + \omega_n \text{proj}_{[t_{\text{min}}, t_{\text{max}}]} \left( \frac{\| u^{n+1} \|}{\| Ev^{n+1} \|} \right), \quad \omega_n \in (0, 1).
\end{align*}
\]

Consequently, the sequence $\{w^n\}$ generated by (26) weakly converges to a solution $x^*$ of the dual problem (24).

The ADMM variant (26) is essentially the same as the standard ADMM, but its parameter $t_n$ is adaptively updated. This rule is different from [26,42].

### 6 Numerical experiments

In this section we provide several numerical experiments to illustrate the influence of the step-size and the adaptive choice in practical applications. Although we motivate the adaptive step-size only for linear problems, we will apply it to problems that do not fulfill this assumption since the convergence of the method is ensured by Theorem 3.2 in all cases. We also note that the steps of the non-stationary method may be more costly than the one with constant step-size, if the evaluation of the resolvents is costly and the constant step-size can be leveraged to precompute something. This is the case when $A$ and/or $B$ is linear and the resolvents involve the solution of a linear system for which a matrix factorization can be precomputed. However, there are tricks to overcome this issue, see [7, pp. 28–29], but we will not go in more detail here.

For the Douglas–Rachford method we just provide illustrative examples since we are not aware of any adaptive rule that applies to the Douglas–Rachford method in the general case of monotone operators. For the ADMM there are several other adaptive rules available and we do a comparison in Sect. 6.2.

#### 6.1 Experiments for non-stationary Douglas–Rachford

We provide four numerical examples to illustrate the new adaptive DR scheme (12) on some well-studied problems in the literature. The step-sizes (21) (in the case of single valued $B$) and (22) (in the case of multivalued $B$) come with new parameters: the safeguards $t_{\text{min}}/\text{max}$ and $\kappa_{\text{min}}/\text{max}$ and a “conservation” term $\omega_n$. Since $B$ is single...
valued, in all experiments we always used (21) and we also fixed $t_{\min} = 10^{-4}$, $t_{\max} = 10^4$ and $\omega_n = 2^{-n/100}$ for all experiments.

6.1.1 The linear toy example

We start with the linear toy example from Sect. 1.2. The residual sequence along the iterations is shown on the left of Fig. 3. Additionally, we determined the step-size $t_{\text{opt}}$ that leads to the smallest asymptotic convergence rate, i.e. to the smallest spectral radius of the iteration map $H_{t_{\text{opt}}}$ (in this case $t_{\text{opt}} = 1.367$) and also plot the corresponding residual sequence with this optimal constant step-size in the same figure. The adaptive step-size does indeed improve the convergence considerably both by using small steps in the beginning and automatically tuning to a step-size $t$ that is close to the optimal one (cf. Fig. 1, right). It also outperforms the optimal constant step-size $t_{\text{opt}}$.

6.1.2 LASSO problems

The LASSO problem is the minimization problem

$$\min_x \left[ F(x) = \frac{1}{2} \| K x - b \|^2_2 + \alpha \| x \|_1 \right]$$

and is also known as basis pursuit denoising [41]. We will treat this with the Douglas–Rachford method as follows: We set $F = f + g$ with

$$g(x) = \frac{1}{2} \| K x - b \|^2_2, \quad \quad B = \nabla g(x) = K^T (K x - b)$$

$$f(x) = \alpha \| x \|_1, \quad \quad A = \partial f(x).$$
In this particular example we take $K \in \mathbb{R}^{100 \times 1000}$ with orthonormal rows, and hence, by the matrix inversion lemma, we get

$$(I + tB)^{-1}x = \left(I + tK^T K\right)^{-1} \left(x + tK^T b\right) = \left(I - \frac{t}{t+1}K^T K\right) \left(x + tK^T B\right).$$

The resolvent of $A$ is the so-called soft-thresholding operator:

$$(I + tA)^{-1}x = \max(|x| - t\alpha, 0) \text{sign}(x).$$

Note that $B$ is single-valued and $A$ is a subgradient and hence, the adaptive step-size $t_n$ computed by (21) does apply. Figure 4 shows the result of the Douglas–Rachford iteration with constant and adaptive step-sizes, and also a comparison with the FISTA [4] method. (Note that if $K$ would not have orthonormal rows, one would have to solve a linear system at each Douglas–Rachford step which would make the comparison with FISTA by iteration count unfair.) As shown in this plot, the adaptive step-size again automatically tunes to a step-size close to 10 which, experimentally, seems to be the best constant step-size for this particular instance.

6.1.3 Convex–concave saddle-point problems

Let $X$ and $Y$ be two finite dimensional Hilbert spaces, $K : X \to Y$ be a bounded linear operator and $f : X \to \mathbb{R} \cup \{+\infty\}$ and $g : Y \to \mathbb{R} \cup \{+\infty\}$ be two proper, convex and lower-semicontinuous functionals. The saddle point problem then reads as

$$\min_{x \in X} \max_{y \in Y} \left\{ f(x) + \langle Kx, y \rangle - g(y) \right\}. $$

Saddle points $(x^*, y^*)$ are characterized by the inclusion

$$0 \in \begin{bmatrix} \partial f & K^T \\ -K & \partial g \end{bmatrix} \begin{bmatrix} x^* \\ y^* \end{bmatrix}.$$ 

\(\square\) Springer
To apply the Douglas–Rachford method we split the optimality system as follows. We denote \( z = (x, y) \) and set

\[
A = \begin{bmatrix}
\partial f & 0 \\
0 & \partial g
\end{bmatrix}, \quad B = \begin{bmatrix}
0 & K^T \\
-K & 0
\end{bmatrix},
\]

(cf. \([9,33]\)). The operator \( A \) is maximally monotone as a subgradient and \( B \) is linear and skew-symmetric, hence maximally monotone and even continuous.

One standard problem in this class in the so-called Rudin–Osher–Fatemi model for image denoising \([37]\), also known as total variation denoising. For a given noisy image \( u_0 \in \mathbb{R}^{M \times N} \) one seeks a denoising image \( u \) as the minimizer of

\[
\min_u \left\{ \frac{1}{2} \| u - u_0 \|_2^2 + \lambda \| \nabla u \|_1 \right\},
\]

where \( \nabla u \in \mathbb{R}^{M \times N \times 2} \) denotes the discrete gradient of \( u \) and \( |\nabla u| \) denotes the components-wise magnitude of this gradient. The penalty term \( \| |\nabla u|\|_1 \) is the discretized total variation, and \( \lambda > 0 \) is a regularization parameter. The saddle point form of this minimization problem is

\[
\min_u \max_{|\phi| \leq \lambda} \left\{ \frac{1}{2} \| u - u_0 \|_2^2 + \langle \nabla u, \phi \rangle \right\}.
\]

We test our DR scheme \((12)\) using the adaptive step-size \( t_n \) and compare with two constant step-sizes \( t = 1 \) and \( t = 13 \). The constant step-size \( t = 13 \) seems to be the best among many trial step-sizes after tuning. The convergence behavior of these cases is plotted in Fig. 5 for one particular image called \( \text{auge} \) of the size \( 256 \times 256 \).

As we can see from this figure that the adaptive step-size has a good performance and is comparable with the best constant step-size in this example \((t = 13)\).

### 6.2 Experiments for ADMM with an adaptive step-size

In this subsection we verify the performance of the our adaptive ADMM variant \((26)\). We follow the comparison from \([43]\) where several adaptive variants of ADMM are
We compare the number of iterations needed for the methods to reach a given tolerance as in [42]. We report mean (± standard deviation) for 50 runs on random instances compared. However, we only compare the methods of ADMM that do not involve relaxation, since we did not consider relaxation in this paper.

In our comparison we compare the ADMM with constant step-size which is fixed ad-hoc, the adaptive rule of He [26] which is based on residual balancing (RB), the adaptive ADMM (AADMM) from [42] and our approach from Theorem 5.1. We used five different test problems from the comparison in [43]: Elastic net regression, LASSO regression, quadratic programming, consensus $\ell_1$-regularized logistic regression, and SVM for classification (see [43, Section 6] for details). We also use the code released online from [42].

Table 1 summarizes the results for average number of iterations for 50 runs on random instances of the same size. Note that both the RB ADMM and the AADMM do guarantee convergence only if the adaptivity is switched off at a certain point while our rule comes with a convergence guarantee. Table 1 shows that our adaptive method consistently performs well.

Figure 6 shows example runs for four of the five problem (the fifth being the SVM classification and is omitted due to space reasons). One observes that residual balancing often fails to make progress towards a favorable step-size and that AADMM sometimes shows large oscillations in the step-sizes. Our method leads to a step-size sequence that stabilizes quickly and leads to good reduction of the residual.

### 7 Conclusion

We have attempted to address one fundamental practical issue in the well-known DR method: step-size selection. This issue has been standing for a long time and has not adequately been well-understood. In this paper, we have proposed an adaptive step-size that is derived from an observation of the linear case. Our non-stationary DR method is new; it is derived from the iteration for single-valued $B$ and differs from the standard non-stationary iteration considered previously, e.g. in [28]. Our step-size remains heuristic in the general case, but we can guarantee a global convergence of the DR method. As a byproduct, we have also derived a new ADMM variant that uses a simple adaptive step-size and has a convergence guarantee. This is practically significant since

---

2 The paper [43] has a convergence guarantee for an adaptive relaxed method, but this does not apply to the methods used in this comparison and is not included since it also involves relaxation.
ADMM has been widely used in many areas in the last two decades. Our finding also opens some future research ideas: Although we gained some insight, the linear case is still not properly understood. Since our heuristic applies to general $A$ and $B$, there...
is a possibility to investigate, which operators should be used as “B” to compute the adaptive step-size. As shown in [33], one can rescale convex–concave saddle point problems to use two different step-sizes for the Douglas–Rachford method, and one may extend our heuristic to this case. Moreover, the convergence speed of the non-stationary method under additional assumptions such as Lipschitz continuity or coercivity could be analyzed. Finally, an adaptive rule for the relaxed DR method would be of interest.

Acknowledgements We thank Zheng Xu and Tom Goldstein for sharing the ADMM code.

Appendix: The proof of Theorem 5.1

Let us assume that we apply (12) to solve the optimality condition (25) of the dual problem (24). From (12), i.e.,

\[ y^{n+1} = J_{t_nA} \left( (1 + \kappa_n)J_{t_{n-1}B}y^n - \kappa_n y^n \right) + \kappa_n \left( y^n - J_{t_{n-1}B}y^n \right), \]

we define \( w^{n+1} := J_{t_{n-1}B}y^n \) and \( z^{n+1} := J_{t_nA}((1 + \kappa_n)w^{n+1} - \kappa_n y^n) \) to obtain

\[
\begin{align*}
  w^{n+1} &= J_{t_{n-1}B}y^n \\
  z^{n+1} &= J_{t_nA}((1 + \kappa_n)w^{n+1} - \kappa_n y^n) \\
  y^{n+1} &= z^{n+1} + \kappa_n \left( y^n - w^{n+1} \right).
\end{align*}
\]

Shifting up this scheme by one index and changing the order, we obtain

\[
\begin{align*}
  z^n &= J_{t_{n-1}A}((1 + \kappa_{n-1})w^n - \kappa_{n-1} y^{n-1}) \\
  y^n &= z^n + \kappa_{n-1} \left( y^{n-1} - w^n \right) \\
  w^{n+1} &= J_{t_{n-1}B}y^n = J_{t_{n-1}B} \left( z^n + \kappa_{n-1} \left( y^{n-1} - w^n \right) \right).
\end{align*}
\]

Let \((1 + \kappa_{n-1})w^n - \kappa_{n-1} y^{n-1} = x^n + w^n\). This gives \( x^n = \kappa_{n-1}(w^n - y^{n-1}) \) and hence, \( z^n + \kappa_{n-1}(y^{n-1} - w^n) = z^n - x^n \) and \( x^{n+1} = \kappa_n(w^{n+1} - y^n) = \kappa_n(w^{n+1} - z^n + x^n) \). Substituting these into the above expression of the DR scheme, we obtain

\[
\begin{align*}
  z^n &= J_{t_{n-1}A}(x^n + w^n) \\
  w^{n+1} &= J_{t_{n-1}B}(z^n - x^n) \\
  x^{n+1} &= \kappa_n(x^n + w^{n+1} - z^n),
\end{align*}
\]

where \( x^n = \kappa_{n-1}(w^n - y^{n-1}) \).

From \( z^n = J_{t_{n-1}A}(w^n + x^n) \), we have \( z^n = (I + t_{n-1}A)^{-1}(w^n + x^n) \) or

\[ 0 \in z^n - w^n - x^n + t_{n-1}\left(D\nabla\phi^* \left(D^Tz^n\right) - c\right). \]
Let $u^{n+1} \in \nabla \varphi^*(D^T z^n)$, which implies $D^T z^n \in \partial \varphi(u^{n+1})$. Hence, we have $z^n - w^n - x^n + t_{n-1}(Du^{n+1} - c) = 0$, therefore $D^T z^n = D^T(w^n + x^n - t_{n-1}(Du^{n+1} - c)) \in \partial \varphi(u^{n+1})$. This condition leads to

$$0 \in D^T \left( t_{n-1} \left( Du^{n+1} - c \right) - x^n - w^n \right) + \partial \varphi \left( u^{n+1} \right).$$

This is the optimality condition of

$$u^{n+1} = \arg\min_u \left\{ \varphi(u) + \frac{t_{n-1}}{2} \| Du - c - t_{n-1} (x^n + w^n) \|^2 \right\}.$$

Similarly, from $u^{n+1} = J_{t_{n-1}B}(z^n - x^n)$, if we define $v^{n+1} \in \nabla \psi^*(E^T w^{n+1})$, then we can also derive that

$$v^{n+1} = \arg\min_v \left\{ \psi(v) + \frac{t_{n-1}}{2} \| Ev + t_{n-1} (x^n - z^n) \|^2 \right\}.$$  

From the line $z^n - w^n - x^n + t_{n-1}(Du^{n+1} - c) = 0$ above, we can write $x^n - z^n = t_{n-1}(Du^{n+1} - c) - w^n$. Substituting this expression into the above step, we obtain

$$v^{n+1} = \arg\min_v \left\{ \psi(v) - \langle w^n, Ev \rangle + \frac{t_{n-1}}{2} \| Ev + Du^{n+1} - c \|^2 \right\}.$$  

This is the second line of (26).

Next, from $w^{n+1} - z^n + x^n + t_{n-1} E v^{n+1} = 0$, we have $w^n = z^n - x^n - t_{n-2} E v^n$. This implies $E v^n = -t_{n-2} (x^n - 1 + w^n - z^n - 1)$. From the last line of (28), we have $x^n = \kappa_{n-1} (x^n - 1 + w^n - z^n - 1)$. Combine these two lines, we get $E v^n = -\frac{1}{\kappa_{n-1} t_{n-2}} x^n = -\frac{1}{t_{n-1}} x^n$ due to the update rule (9): $t_{n-1} = \kappa_{n-1} t_{n-2}$. Substituting $E v^n = -\frac{1}{t_{n-1}} x^n$ into the $u$-subproblem, we obtain

$$u^{n+1} = \arg\min_u \left\{ \varphi(u) - \langle w^n, Du \rangle + \frac{t_{n-1}}{2} \| Du + Ev^n - c \|^2 \right\}.$$  

This is the first line of (26).

Now, since $z^n = w^n - t_{n-1}(Du^{n+1} - c) + x^n$, and $w^{n+1} = z^n - x^n - t_{n-1} E v^{n+1}$, combining these expressions, we obtain $w^{n+1} = w^n - t_{n-1}(Du^{n+1} + Ev^{n+1} - c)$. This is the last line of (26).

Finally, we derive the update rule for $t_n$. Indeed, note that $y^n = z^n - x^n$, and $z^n - w^n - x^n + t_{n-1}(Du^{n+1} - c) = 0$. These relations show that $y^n = w^n - t_{n-1} (Du^{n+1} - c)$. Moreover, we also have $w^{n+1} = J_{t_{n-1}B}(z^n - x^n) = J_{t_{n-1}B}(y^n)$. In this case, we have $J_{t_{n-1}B}(y^n) = y^n = w^{n+1} - w^n + t_{n-1}(Du^{n+1} - c) = -t_{n-1}(Du^{n+1} + Ev^{n+1} - c) + t_{n-1}(Du^{n+1} - c) = -t_{n-1} E v^{n+1}$. Hence, we can compute $\kappa_n$ as

$$\kappa_n := \frac{\| J_{t_{n-1}B}(y^n) \|}{\| y^n - J_{t_{n-1}B}(y^n) \|} = \frac{\| w^{n+1} \|}{t_{n-1} \| E v^{n+1} \|}.$$
Using the fact that \( t_n := \kappa_n t_{n-1} \), we show that \( t_n := \frac{\|w^{n+1}\|}{\|E v^{n+1}\|} \), which is the last line of (26) after projecting and weighting as in Sect. 4. Since \( \{u^n\} \) is equivalent to the sequence \( \{u^n\} \) in the DR scheme (2) [or equivalently, (12)] applying to the dual optimality condition (25) of the dual problem (24), the last conclusion is a direct consequence of Theorem 3.2.

\[\square\]

References

1. Bauschke, H.H.: A note on the paper by Eckstein and Svaiter on “general projective splitting methods for sums of maximal monotone operators”. SIAM J. Control Optim. 48(4), 2513–2515 (2009)
2. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces, 2nd edn. Springer, Berlin (2017)
3. Bauschke, H.H., Moffat, S.M., Wang, X.: Firmly nonexpansive mappings and maximally monotone operators: correspondence and duality. Set-Valued Var. Anal. 20(1), 131–153 (2012)
4. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. 2(1), 183–202 (2009)
5. Becker, S., Candès, E.J., Grant, M.: Templates for convex cone problems with applications to sparse signal recovery. Math. Program. Comput. 3(3), 165–218 (2011)
6. Becker, S., Combettes, P.L.: An algorithm for splitting parallel sums of linearly composed monotone operators, with applications to signal recovery (2013). arXiv:1305.5828
7. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. 3(1), 1–122 (2011)
8. Bredies, K., Sun, H.P.: Preconditioned Douglas–Rachford algorithms for TV-and TGV-regularized variational imaging problems. J. Math. Imaging Vis. 52(3), 317–344 (2015)
9. Bredies, K., Sun, H.: Preconditioned Douglas–Rachford splitting methods for convex–concave saddle-point problems. SIAM J. Numer. Anal. 53(1), 421–444 (2015)
10. Bredies, K., Sun, H.: Accelerated Douglas–Rachford methods for the solution of convex–concave saddle-point problems (2016). arXiv:1604.06282
11. Combettes, P.L.: Solving monotone inclusions via compositions of nonexpansive averaged operators. Optimization 53(5–6), 475–504 (2004)
12. Combettes, P.L., Pesquet, J.-C.: A Douglas–Rachford splitting approach to nonsmooth convex variational signal recovery. IEEE J. Sel. Top. Signal Process. 1(4), 564–574 (2007)
13. Dao, N.M., Phan, M.H.: Adaptive Douglas–Rachford splitting algorithm for the sum of two operators (2018). arXiv:1809.00761
14. Davis, D.: Convergence rate analysis of the forward-Douglas–Rachford splitting scheme. SIAM J. Optim. 25(3), 1760–1786 (2015)
15. Davis, D., Yin, W.: Convergence rate analysis of several splitting schemes. In: Glowinski, R., Osher, S.J., Yin, W. (eds.) Splitting Methods in Communication, Imaging, Science, and Engineering, pp. 115–163. Springer, Berlin (2016)
16. Douglas Jr., J., Rachford Jr., H.H.: On the numerical solution of heat conduction problems in two and three space variables. Trans. Am. Math. Soc. 82(2), 421–439 (1956)
17. Eckstein, J., Bertsekas, D.P.: On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. 55(1), 293–318 (1992)
18. Ghadimi, E., Teixeira, A., Shames, I., Johansson, M.: Optimal parameter selection for the alternating direction method of multipliers (ADMM): quadratic problems. IEEE Trans. Autom. Control 60(3), 644–658 (2015)
19. Giselsson, P.: Tight global linear convergence rate bounds for Douglas–Rachford splitting. J. Fixed Point Theory Appl. 19(4), 2241–2270 (2017)
20. Giselsson, P., Boyd, S.: Diagonal scaling in Douglas–Rachford splitting and ADMM. In: 2014 IEEE 53rd Annual Conference on Decision and Control (CDC), pp. 5033–5039. IEEE (2014)
21. Giselsson, P., Boyd, S.: Linear convergence and metric selection for Douglas–Rachford splitting and ADMM. IEEE Trans. Autom. Control 62(2), 532–544 (2017)
22. Glowinski, R.: On alternating direction methods of multipliers: a historical perspective. In: Fitzgibbon, W., Kuznetsov, Y., Neittaanmäki, P., Pironneau, O. (eds.) Modeling, Simulation and Optimization for Science and Technology, pp. 59–82. Springer, Dordrecht (2014)

23. Golub, G.H., Van Loan, C.F.: Matrix Computations. Johns Hopkins Studies in the Mathematical Sciences, 4th edn. Johns Hopkins University Press, Baltimore (2013)

24. He, B., Yuan, X.: Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective. SIAM J. Imaging Sci. 5(1), 119–149 (2012)

25. He, B., Yuan, X.: On the $O(1/n)$ convergence rate of the Douglas–Rachford alternating direction method. SIAM J. Numer. Anal. 50(2), 700–709 (2012)

26. He, B.S., Yang, H., Wang, S.L.: Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities. J. Optim. Theory Appl. 106(2), 337–356 (2000)

27. Li, X., Yuan, X.: A proximal strictly contractive Peaceman–Rachford splitting method for convex programming with applications to imaging. SIAM J. Imaging Sci. 8(2), 1332–1365 (2015)

28. Liang, J., Fadili, J., Peyré, G.: Local convergence properties of Douglas–Rachford and alternating direction method of multipliers. J. Optim. Theory Appl. 172(3), 874–913 (2017)

29. Lin, Z., Liu, R., Su, Z.: Linearized alternating direction method with adaptive penalty for low-rank representation. In: Advances in Neural Information Processing Systems, pp. 612–620 (2011)

30. Lions, P.-L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. 16(6), 964–979 (1979)

31. Moursi, W.M., Vandenberghe, L.: Douglas–Rachford splitting for a Lipschitz continuous and a strongly monotone operator (2018). arXiv:1805.09396

32. Nishihara, R., Lessard, L., Recht, B., Packard, A., Jordan, M.: A general analysis of the convergence of ADMM (2015). arXiv:1502.02009

33. O’Connor, D., Vandenberghe, L.: Primal–dual decomposition by operator splitting and applications to image deblurring. SIAM J. Imaging Sci. 7(3), 1724–1754 (2014)

34. Patrinos, P., Stella, L., Bemporad, A.: Douglas–Rachford splitting: complexity estimates and accelerated variants. In: 2014 IEEE 53rd Annual Conference on Decision and Control (CDC), pp. 4234–4239. IEEE (2014)

35. Pock, T., Chambolle, A.: Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 1762–1769. IEEE (2011)

36. Tyrrell Rockafellar, R.: Monotone operators and the proximal point algorithm. SIAM J. Control Optim. 14(5), 877–898 (1976)

37. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Phys. D Nonlinear Phenom. 60(1–4), 259–268 (1992)

38. Song, C., Yoon, S., Pavlovic, V.: Fast ADMM algorithm for distributed optimization with adaptive penalty. In: AAAI, pp. 753–759 (2016)

39. Svaiter, B.F.: A simplified proof of weak convergence in Douglas–Rachford method to a solution of the underlying inclusion problem (2018). arXiv:1809.00967

40. Svaiter, B.F.: On weak convergence of the Douglas–Rachford method. SIAM J. Control Optim. 49(1), 280–287 (2011)

41. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. 58(1), 267–288 (1996)

42. Xu, Z., Figueiredo, M.A.T., Goldstein, T.: Adaptive ADMM with spectral penalty parameter selection (2016). arXiv:1605.07246

43. Xu, Z., Figueiredo, M.A.T., Yuan, X., Studer, C., Goldstein, T.: Adaptive relaxed ADMM: convergence theory and practical implementation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7234–7243. IEEE (2017)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer