J. Oper. Res. Soc. China (2015) 3:139–162
DOI 10.1007/s40305-015-0078-y

An Alternating Direction Approximate Newton Algorithm for Ill-Conditioned Inverse Problems with Application to Parallel MRI

William Hager[1] · **Cuong Ngo**[1] · **Maryam Yashtini**[2] · **Hong-Chao Zhang**[3]

Received: 4 September 2014 / Revised: 24 February 2015 / Accepted: 3 April 2015 /
Published online: 12 May 2015

Abstract An alternating direction approximate Newton (ADAN) method is developed for solving inverse problems of the form $\min\{\phi(Bu) + (1/2)\|Au - f\|_2^2\}$, where ϕ is convex and possibly nonsmooth, and A and B are matrices. Problems of this form arise in image reconstruction where A is the matrix describing the imaging device, f is the measured data, ϕ is a regularization term, and B is a derivative operator. The proposed algorithm is designed to handle applications where A is a large dense, ill-conditioned matrix. The algorithm is based on the alternating direction method of multipliers (ADMM) and an approximation to Newton's method in which a term in Newton's

This research was partly supported by National Science Foundation (Nos. 1115568 and 1016204) and by Office of Naval Research Grants (Nos. N00014-11-1-0068 and N00014-15-1-2048).

✉ Hong-Chao Zhang
hozhang@math.lsu.edu
https://www.math.lsu.edu/~hozhang/

William Hager
hager@ufl.edu
http://people.clas.ufl.edu/hager/

Cuong Ngo
ngocuong@ufl.edu
http://people.clas.ufl.edu/ngocuong/

Maryam Yashtini
myashtini3@math.gatech.edu
http://people.math.gatech.edu/~myashtini3/

1 Department of Mathematics, University of Florida, PO Box 118105, Gainesville, FL 32611-8105, USA

2 School of Mathematics, Georgia Institute of Technology, 686 Cherry Street, Atlanta, GA 30332-0160, USA

3 Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803-4918, USA

Hessian is replaced by a Barzilai–Borwein (BB) approximation. It is shown that ADAN converges to a solution of the inverse problem. Numerical results are provided using test problems from parallel magnetic resonance imaging. ADAN was faster than a proximal ADMM scheme that does not employ a BB Hessian approximation, while it was more stable and much simpler than the related Bregman operator splitting algorithm with variable stepsize algorithm which also employs a BB-based Hessian approximation.

Keywords Convex optimization · Total variation regularization · Nonsmooth optimization · Global convergence · Parallel MRI

Mathematics Subject Classification 90C25 · 65K05 · 68U10 · 65J22

1 Introduction

We consider inverse problems that can be expressed in the form

$$\min_{u\in\mathbb{C}^N} \phi(Bu) + \frac{1}{2}\|Au - f\|^2, \tag{1.1}$$

where $\phi\colon \mathbb{C}^m \to (-\infty, \infty)$ is convex, $B \in \mathbb{C}^{m\times N}$, $A \in \mathbb{C}^{M\times N}$, and $f \in \mathbb{C}^M$. Due to its many applications, especially in the image reconstruction field, a variety of numerical algorithms have been proposed for solving (1.1) including [3–5,7,9,13–16,22,24].

We introduce a new variable w to obtain the split formulation of (1.1):

$$\min_{u,w} \phi(w) + \frac{1}{2}\|Au - f\|^2 \quad \text{s.t.} \quad w = Bu, \quad u \in \mathbb{C}^N, \quad w \in \mathbb{C}^m. \tag{1.2}$$

The alternating direction method of multipliers (ADMM) [12] is among the most extensively used techniques for solving (1.2). The augmented Lagrangian associated with (1.2) is

$$\mathcal{L}^\rho(u,\, w,\, b) = \phi(w) + \frac{1}{2}\|Au - f\|^2 + \mathrm{Re}\,\langle b,\, Bu - w\rangle + \frac{\rho}{2}\|Bu - w\|^2, \tag{1.3}$$

where $\rho > 0$ is the penalty parameter, $b \in \mathbb{C}^m$ is a Lagrange multiplier associated with the constraint $Bu = w$, $\langle\cdot,\,\cdot\rangle$ is the Euclidean inner product, and "Re" stands for "real part." In ADMM, each iteration minimizes over u holding w fixed, minimizes over w holding u fixed, and updates an estimate for the multiplier b. More precisely, if b^k is the current approximation to the multiplier, then ADMM [11,12] applied to the split formulation (1.2) is given by the iteration

$$u^{k+1} = \operatorname*{argmin}_u \mathcal{L}^\rho\left(u,\, w^k,\, b^k\right),$$

$$w^{k+1} = \operatorname*{argmin}_w \mathcal{L}^\rho \left(u^{k+1},\ w,\ b^k\right),$$
$$b^{k+1} = b^k + \rho\left(Bu^{k+1} - w^{k+1}\right).$$

After completing the square, this can be written as follows:

$$u^{k+1} = \operatorname*{argmin}_u \Psi(u), \quad \Psi(u) = \frac{1}{2}\|Au - f\|^2 + \frac{\rho}{2}\left\|Bu - w^k + \rho^{-1}b^k\right\|^2, \tag{1.4}$$

$$w^{k+1} = \operatorname*{argmin}_w \left\{\phi(w) + \frac{\rho}{2}\left\|Bu^{k+1} - w + \rho^{-1}b^k\right\|^2\right\}, \tag{1.5}$$

$$b^{k+1} = b^k + \rho\left(Bu^{k+1} - w^{k+1}\right). \tag{1.6}$$

In parallel magnetic resonance imaging (PMRI), the time-consuming part of the ADMM iteration is the update (1.4), since A is a large dense, ill-conditioned matrix.

The linearized proximal method of multipliers is one strategy for speeding up the iteration (1.4) in PMRI. In this approach, a carefully chosen proximal term is introduced into the updates (see [6,8,21,26]). For any Hermitian matrix $Q \in \mathbb{C}^{N\times N}$, we define

$$\|x\|_Q^2 = \langle x,\ Qx\rangle.$$

If Q is a positive definite matrix, then $\|\cdot\|_Q$ is a norm. The proximal version of (1.4) is

$$u^{k+1} = \operatorname*{argmin}_u \left\{\|Au - f\|^2 + \rho\left\|Bu - w^k + \rho^{-1}b^k\right\|^2 + \left\|u - u^k\right\|_Q\right\}. \tag{1.7}$$

As suggested in [26, Eq.(5.12)], the choice $Q = \delta I - A^*A$ will cancel the $\|Au\|^2$ term in this iteration. This yields a partial linearization of the iteration since the quadratic term $\|Au\|^2$ is eliminated, while $\|Bu\|^2$ is retained. In the context of PMRI, $\|Bu\|^2$ is more tractable than $\|Au\|^2$ since B is usually diagonalized by a Fourier transform. We refer to the algorithm based on the updates (1.5), (1.6), and (1.7) as Bregman operator splitting (BOS). BOS is convergent when Q is positive definite according to Theorem 4.2 in [26] or Theorem 5.6 in [21]. If Q is positive definite, then δ should be greater than or equal to the largest eigenvalue of A^*A.

In the BOS with variable stepsize (BOSVS) algorithm, the variable stepsize version of BOS developed in [9,13], and in the approximate Newton algorithm developed in this paper, we replace δ by δ_k, where $\delta_k I$ is a Barzilai–Borwein (BB) [2] approximation to A^*A. As seen in the numerical experiments of Raydan and Svaiter [19], the BB approximation often yields surprisingly fast convergence on ill-conditioned problems. Typically, δ_k is strictly smaller than the largest eigenvalue of A^*A, and $Q = \delta_k I - A^*A$ is indefinite. In this setting where the proximal term is indefinite, a completely new convergence analysis is needed. Moreover, safeguards are needed in the algorithm itself. In BOSVS, one safe guard is a line search in which the initial BB

approximation to A^*A is adjusted. In the alternating direction approximate Newton (ADAN) algorithm developed in this paper, we retain the initial BB approximation, but take a predetermined step along the approximate Newton direction in each iteration. We prove convergence of the resulting algorithm, and we compare its performance to that of both BOS and BOSVS using PMRI image reconstruction problems.

The paper is organized as follows. In Sect. 2, we present the new algorithm, the ADAN method. Section 3 establishes the convergence of ADAN. Section 4 provides numerical experiments based on PMRI.

1.1 Notation

For any matrix M, $\mathcal{N}(M)$ is the null space of M. The superscript $^{\mathsf{T}}$ denotes transpose, while superscript * denotes conjugate transpose with the following exception: superscript * is attached to an optimal solution of (1.1), or in PMRI, to the reference image that we try to reconstruct. For $x \in \mathbb{C}^N$ and $y \in \mathbb{C}^N$, $\langle x, y\rangle = x^*y$ is the standard Euclidean inner product, and (x, y) stacks the vectors x and y vertically. If M_1 and M_2 are Hermitian matrices, we write $M_1 \succeq M_2$ if $M_1 - M_2$ is positive semidefinite. The norm $\|\cdot\|$ is the Euclidean norm given by $\|x\| = \sqrt{\langle x, x\rangle}$. We let Re stand for "real part." For a matrix X, $\|X\|$ is the induced matrix norm, which is the largest singular value. For a differentiable function $F\colon \mathbb{C}^N \to \mathbb{R}$, $\nabla F(x)$ is the gradient of F at x, a column vector. More generally, $\partial F(x)$ denotes the subdifferential set at x.

2 Proposed Algorithm

The ADMM updated for u given in (1.4) can be expressed as follows:

$$u^{k+1} = \operatorname*{argmin}_u \Psi(u) = u^k - (A^*A + \rho B^*B)^{-1}\nabla\Psi_k, \quad \text{where} \tag{2.1}$$

$$\nabla\Psi_k := A^*\left(Au^k - f\right) + \rho B^*\left(Bu^k - w^k + \rho^{-1}b^k\right). \tag{2.2}$$

Here $(\cdot)^{-1}$ is the generalized inverse, $A^*A + \rho B^*B$ is the Hessian of the objective Ψ, and $\nabla\Psi_k$ is the gradient of Ψ at u^k. The formula for u^{k+1} in (2.1) is exactly the same formula that we would have gotten if we performed a single iteration of Newton's method on the equation $\nabla\Psi(u) = 0$ with starting guess u^k. Since it is not practical to invert $A^*A + \rho B^*B$ in PMRI, due to the large size and dense structure of the matrix, we employ the BB approximation [2]

$$A^*A \approx \delta_k I,$$

where

$$\begin{aligned}\delta_k &= \operatorname{argmin}\left\{\left\|A\left(u^k - u^{k-1}\right) - \delta\left(u^k - u^{k-1}\right)\right\|^2 : \delta \geqslant \delta_{\min}\right\} \\ &= \max\left\{\delta_{\min}, \frac{||A(u^k - u^{k-1})||^2}{||u^k - u^{k-1}||^2}\right\},\end{aligned} \tag{2.3}$$

and $\delta_{\min} > 0$ is a positive lower bound for δ_k. Hence, the Hessian is approximated by $\delta_k I + \rho B^* B$. Often the matrix $B^* B$ in image reconstruction can be diagonalized by a Fourier transform. Since a Fourier transform can be inverted in $O(N \log N)$ flops, the inversion of $\delta_k I + B^* B$ can be accomplished relatively quickly.

After replacing $A^* A$ by $\delta_k I$ in (2.1), the iteration becomes

$$u^{k+1} = u^k + d_k, \quad \text{where} \quad d_k = -\left(\delta_k I + \rho B^* B\right)^{-1} \nabla \Psi_k. \tag{2.4}$$

Note that by substituting $Q = \delta_k I - A^* A$ in (1.7) and solving for the minimizer, we would get exactly the same formula for the minimizer as that given in (2.4). Hence, the approximate Newton formula (2.4) is the same as the proximal update (1.7) with the choice $Q = \delta_k I - A^* A$. On the other hand, by viewing this update in the context of an approximate Newton iteration, there is a natural way to encourage convergence by taking a partial step along the Newton search direction. In particular, we consider the rule

$$u^{k+1} = u^k + \sigma_k d_k,$$

where σ_k is the stepsize in the search direction d_k. We now explain how to choose σ_k to ensure descent.

The inner product between d_k and the objective gradient at u^k is

$$\langle \nabla \Psi_k,\, d_k \rangle = -\langle \nabla \Psi_k,\, \left(\delta_k I + \rho B^* B\right)^{-1} \nabla \Psi_k \rangle = -\left(\delta_k \left\| d_k \right\|^2 + \rho \left\| B d_k \right\|^2\right). \tag{2.5}$$

It follows that d_k is a descent direction. The stepsize σ_k is chosen to ensure an amount of descent similar to what is achieved in an Armijo line search [1]. More precisely, given $\gamma \in (0,\, 1)$, we choose σ_k to be the largest stepsize that satisfies the inequality

$$\frac{\Psi(u^{k+1}) - \Psi(u^k)}{\sigma_k} = \frac{\Psi(u^k + \sigma_k d_k) - \Psi(u^k)}{\sigma_k} \leqslant \gamma \langle \nabla \Psi_k,\, d_k \rangle. \tag{2.6}$$

Since Ψ is a quadratic, the Taylor expansion of $\Psi(u^{k+1})$ around u^k is as follows:

$$\Psi\left(u^{k+1}\right) = \Psi\left(u^k\right) + \langle \nabla \Psi_k,\, u^{k+1} - u^k \rangle + \frac{1}{2} \left\| u^{k+1} - u^k \right\|_H^2, \tag{2.7}$$

where

$$H = A^* A + \rho B^* B. \tag{2.8}$$

We substitute $u^{k+1} - u^k = \sigma_k d_k$ and combine with (2.6) to obtain

$$\sigma_k \left\| d_k \right\|_H^2 \leqslant 2(\gamma - 1) \langle \nabla \Psi_k,\, d_k \rangle. \tag{2.9}$$

We will take σ_k as large as possible while satisfying (2.9). If $\|d_k\|_H \neq 0$, then simply divide by $\|d_k\|_H^2$ to compute the largest σ_k. Based on the following lemma, $\|d_k\|_H = 0$ if and only if u^k minimizes Ψ.

Lemma 2.1 *If $Ad_k = 0 = Bd_k$, then u^k minimizes Ψ.*

Proof If $Ad_k = 0 = Bd_k$, then $\|d^k\|_H = 0$. By a Taylor expansion, as in (2.7), we have

$$\Psi\left(u^k + \sigma d^k\right) = \Psi\left(u^k\right) + \sigma\langle \nabla\Psi_k,\, d_k\rangle + \frac{\sigma^2}{2}\left\|d^k\right\|_H = \Psi\left(u^k\right) + \sigma\langle \nabla\Psi_k,\, d_k\rangle.$$

Clearly, $\langle \nabla\Psi_k,\, d_k\rangle \leqslant 0$ by (2.5). If $\langle \nabla\Psi_k,\, d_k\rangle < 0$, then $\Psi(u^k + \sigma d^k)$ approaches $-\infty$ as σ tends to ∞. This is impossible since $\Psi(u) \geqslant 0$ for all u. Hence, we have

$$\langle \nabla\Psi_k,\, d_k\rangle = \langle \nabla\Psi_k,\, \left(\delta_k I + \rho B^* B\right)^{-1} \nabla\Psi_k\rangle = 0.$$

It follows that $\nabla\Psi_k = 0$, which implies that u^k minimizes the convex function Ψ.

Based on Lemma 2.1, if u^k is not the minimizer of Ψ, then by (2.9), the largest stepsize satisfying (2.6) is given by the following expression:

$$\bar{\sigma}_k = 2(\gamma - 1)\left(\frac{\langle \nabla\Psi_k,\, d_k\rangle}{\|d_k\|_H^2}\right) = 2(1-\gamma)\left(\frac{\delta_k\|d_k\|^2 + \rho\|Bd_k\|^2}{\|Ad_k\|^2 + \rho\|Bd_k\|^2}\right). \tag{2.10}$$

If $\gamma = 1/2$, then $s = \bar{\sigma}_k$ is the exact minimizer of $\Psi(u_k + sd_k)$ over all s. That is, if one solves for the exact minimizer of the quadratic in s, one obtains

$$s = \frac{\delta_k\|d_k\|^2 + \rho\|Bd_k\|^2}{\|Ad_k\|^2 + \rho\|Bd_k\|^2},$$

which is the same as (2.10) for $\gamma = 1/2$. Thus, $\gamma = 1/2$ yields the largest decrease in Φ. However, we need to achieve convergence of the overall algorithm, which also involves updates of w^k and b^k in (1.5) and (1.6). In the convergence analysis, we find that $\gamma > 1/2$ is needed to achieve convergence of the overall algorithm. Hence, we take γ close to 1/2 but not equal to 1/2. Also, the convergence analysis breaks down when the stepsize is too big. Hence, we introduce an upper bound $\sigma_{\max}$ on the stepsize and our safeguarded step is given by

$$\sigma_k = \min\left\{\sigma_{\max},\, \bar{\sigma}_k\right\}. \tag{2.11}$$

The expression $[\Psi(u^{k+1}) - \Psi(u^k)]/\sigma_k$ in (2.6) is a monotone increasing function of σ_k since it is linear in σ_k and the coefficient of the linear term is $\|Ad_k\|^2 + \rho\|Bd_k\|^2$. It follows that the safeguarded step given by (2.11) satisfies

$$\frac{\Psi(u^{k+1}) - \Psi(u^k)}{\sigma_k} \leqslant -\gamma\left(\delta_k\|d_k\|^2 + \rho\|Bd_k\|^2\right),$$

while $\sigma_k = \bar{\sigma}_k$ satisfies this with equality.

The safeguarded approximate Newton algorithm that we analyze in this paper is given Algorithm 1. In Step 1, we first check whether u^k achieves the minimum of Ψ in (1.4). If it is not the minimizer, then we compute the BB parameter δ_k, the search direction d_k, and the safeguarded stepsize σ_k. Step 2 contains some tests that arise in the convergence analysis. If the inequalities in Step 2 are both satisfied, then we need to increase the lower bound for the BB parameter δ_k. If the inequality in Step 3 is satisfied, then we need to decrease the upper bound on the stepsize parameter $\sigma_{\max}$. We take τ close to 1 so that Steps 2 and 3 have minimal impact on the algorithm. Step 4 is the update of u^k and Step 5 is the update of w^k. For total variation (TV) regularization, Step 5 amounts to soft shrinkage [9,23]. Step 6 is the multiplier update.

ADAN is a modification of the BOSVS algorithm of [9]. The initial value for δ_k is computed by the same formula in both ADAN and BOSVS, however, BOSVS employs a line search in which the initial δ_k is increased until a stopping condition is satisfied. ADAN, on the other hand, does not change δ_k; instead, only a partial step is made along the direction d_k, where the stepsize is given by the $\bar{\sigma}_k$ in (2.10). We are able to prove convergence of either algorithm, but the implementation of ADAN is much simpler and requires far few parameters. The experiments of Sect. 4 also reveal some numerical advantages for ADAN when compared to BOSVS.

3 Convergence Analysis

In this section, we establish the convergence of ADAN to a solution of (1.1). We first give the existence and uniqueness result for (1.1):

Algorithm 1: Alternating Direction Approximate Newton Method

Parameter: $0.5 < \gamma < 1 < \tau$, $\rho > 0$, $0 < \delta_{\min} \leqslant \delta_0$, $\sigma_0 = 0$, $\sigma_{\max} = 1$. Initialize $k = 1$.

Starting guess: u^1, w^1 and b^1.

Step 1 If $\nabla \Psi_k$ vanishes in (2.2), then set $u^{k+1} = u^k$, $d_k = 0$, $\delta_k = \delta_{k-1}$, $\sigma_k = \sigma_{k-1}$, and branch to Step 5. Otherwise, set $\sigma_k = \min\{\sigma_{\max}, \bar{\sigma}_k\}$, where $\bar{\sigma}_k$ is defined in (2.10), and

$$\delta_k = \max\left\{\delta_{\min}, \frac{\|A(u^k - u^{k-1})\|^2}{\|u^k - u^{k-1}\|^2}\right\}, \qquad d_k = -(\delta_k I + \rho B^* B)^{-1} \nabla \Psi_k .$$

Step 2 If $\delta_k \sigma_{k-1} > \delta_{k-1} \sigma_k$ and $\delta_k > \max\{\delta_{\min}, \delta_{k-1}\}$, then $\delta_{\min} := \tau \delta_{\min}$.

Step 3 If $\sigma_k < \min\{\sigma_{\max}, \sigma_{k-1}\}$, then $\sigma_{\max} := \tau^{-1} \sigma_{\max}$.

Step 4 Update $u^{k+1} = u^k + \sigma_k d_k$.

Step 5 $w^{k+1} = \arg\min_w \left\{\phi(w) + \frac{\rho}{2}\|w - Bu^{k+1} - \rho^{-1} b^k\|^2\right\}$.

Step 6 $b^{k+1} = b^k + \rho(Bu^{k+1} - w^{k+1})$.

Step 7 If a stopping criterion is satisfied, terminate the algorithm. Otherwise $k = k + 1$ and go to Step 1.

Lemma 3.1 *If $\phi(w)$ tends to infinity as $\|w\|$ tends to infinity, then there exists a solution of (1.1), or equivalently to (1.2). If in addition ϕ is strictly convex and $\mathcal{N}(A) \cap \mathcal{N}(B) = \{0\}$, then the solution of (1.1) is unique.*

Proof Define $\mathcal{S} = \mathcal{N}(A) \cap \mathcal{N}(B)$. Any $u \in \mathbb{C}^N$ can be decomposed as $u = s + t$ where s is the projection of u onto $\mathcal{S}$. As $\|t\|$ tends to ∞, either At or Bt tends to ∞. Let $u_i,\ i = 1,\ 2, \cdots$ denote a minimizing sequence for the objective function $\Phi(u) = \phi(Bu) + (1/2)\|Au - f\|^2$, and decompose $u_i = s_i + t_i$ where s_i is the projection of u_i onto $\mathcal{S}$. If the sequence $\|t_i\|$ is unbounded, then as noted earlier, either $Au_i = At_i$ or $Bu_i = Bt_i$ is unbounded. If $\|Bu_i\|$ is unbounded, then $\Phi(u_i)$ is unbounded by the assumption for ϕ, which contradicts the fact that the u_i form a minimizing sequence. On the other hand, if $\|Bu_i\|$ is bounded and $\|Au_i\|$ is unbounded, then $\|Au_i - f\|$ is unbounded, which implies that $\Phi(u_i)$ is unbounded, again a contradiction. Thus, we conclude that the sequence $\|t_i\|$ is bounded. Let u^* denote a limit for a convergent subsequence of t_i. Since $\Phi(u_i) = \Phi(t_i)$, we deduce that u^* minimizes Φ.

Now suppose in addition that ϕ is strictly convex, $\mathcal{N}(A) \cap \mathcal{N}(B) = \{0\}$, and that (1.1) has two distinct minimizers u_1 and u_2. We show that there is a contradiction unless $u_1 = u_2$. If $q(u) = (1/2)\|Au - f\|^2$ is the quadratic term in (1.1), then for any $\theta \in \mathbb{R}$, we have

$$q\left((1-\theta)u_1 + \theta u_2\right) = (1-\theta)q\left(u_1\right) + \theta q\left(u_2\right) - \frac{\theta(1-\theta)}{2}\left\|A\left(u_1 - u_2\right)\right\|^2. \quad (3.1)$$

Since ϕ is strictly convex, we have

$$\phi\left((1-\theta)Bu_1 + \theta Bu_2\right) \leqslant (1-\theta)\phi\left(Bu_1\right) + \theta\phi\left(Bu_2\right), \quad (3.2)$$

for all $\theta \in [0,\ 1]$. Moreover, the inequality is strict if $\theta \in (0,\ 1)$ and $Bu_1 \neq Bu_2$. We combine (3.1) and (3.2) to obtain

$$\Phi\left((1-\theta)u_1 + \theta u_2\right) \leqslant (1-\theta)\Phi\left(u_1\right) + \theta\Phi\left(u_2\right) - \frac{\theta(1-\theta)}{2}\left\|A\left(u_1 - u_2\right)\right\|^2,$$

for all $\theta \in [0,\ 1]$, where the inequality is strict if $\theta \in (0,\ 1)$ and $Bu_1 \neq Bu_2$. Since u_1 and u_2 are both optimal, we have $\Phi(u_1) = \Phi(u_2)$. Hence, $u = (1-\theta)u_1 + \theta u_2$ with $\theta \in (0,\ 1)$ yields a strictly smaller value for Φ unless $Bu_1 = Bu_2$ and $A(u_1 - u_2) = 0$. This implies that $(u_1 - u_2) \in \mathcal{N}(A) \cap \mathcal{N}(B)$. Since $\mathcal{N}(A) \cap \mathcal{N}(B) = \{0\}$, it follows that $u_1 = u_2$, which completes the proof.

Lemma 3.1 is a strengthening of Lemma 3.1 in [9]. As pointed out in Sect. 4, Lemma 3.1 typically implies the existence of a unique solution to (1.1) in the context of PMRI. Our main convergence result is the following:

Theorem 3.2 *If there exists a solution of (1.1), then the sequence $(u^k,\ w^k,\ b^k)$ generated by ADAN approaches a point $(u^*,\ w^*,\ b^*)$ where the first-order optimality conditions (3.10) are satisfied. Moreover, $(u^*,\ w^*)$ is a solution of (1.2) and u^* is a solution of (1.1).*

The proof of Theorem 3.2 requires several lemmas. In Step 2, $\delta_{\min}$ can grow and in Step 3, $\sigma_{\max}$ can decay, if the stated criteria are satisfied. We first show that these criteria are only satisfied a finite number of times, and hence, $\delta_{\min}$ and $\sigma_{\max}$ converge to positive limits. An upper bound for $\delta_{\min}$ is the following:

Lemma 3.3 *Uniformly in k, we have*

$$\delta_{\min,k} \leqslant \delta_k \leqslant \max\{\bar{\delta},\ \tau\|A\|\},$$

where $\delta_{\min,k}$ is the value of $\delta_{\min}$ at the start of iteration k, and $\bar{\delta} = \delta_{\min,1}$ is the starting $\delta_{\min}$ in ADAN.

Proof Since $\|A(u^k - u^{k-1})\| \leqslant \|A\|\|u^k - u^{k-1}\|$, it follows from Step 1 that

$$\delta_{\min,k} \leqslant \delta_k \leqslant \max\left\{\delta_{\min,k},\ \|A\|\right\}. \tag{3.3}$$

Hence, if $\delta_k > \delta_{\min,k}$, then $\delta_k \leqslant \|A\|$, which implies that $\delta_{\min,k} \leqslant \|A\|$. Consequently, when the second condition in Step 2 is satisfied, the current $\delta_{\min,k} \leqslant \|A\|$, which implies that the new $\delta_{\min,k+1}$ is at most $\tau\|A\|$. If $\delta_{\min,k}$ is never updated in Step 2, then $\delta_{\min,k}$ must equal its starting value $\bar{\delta}$. Hence, in general, we have

$$\delta_{\min,k} \leqslant \max\{\bar{\delta},\ \tau\|A\|\}. \tag{3.4}$$

We combine (3.3) and (3.4), and the fact that $\tau > 1$, to complete the proof.

Next, we show that $\sigma_{\max}$ is uniformly bounded from below:

Lemma 3.4 *Uniformly in k, we have*

$$1 \geqslant \sigma_{\max,k} \geqslant \sigma_k \geqslant \frac{2(1-\gamma)}{\tau} \min\left\{\frac{\bar{\delta}}{\|A\|^2},\ 1\right\}, \tag{3.5}$$

where $\bar{\delta} = \delta_{\min,1}$ is the starting $\delta_{\min}$ in ADAN and $\sigma_{\max,k}$ is the value of $\sigma_{\max}$ at the start of iteration k.

Proof The lower bound $\sigma_{\max,k} \geqslant \sigma_k$ follows immediately from the formula for σ_k in Step 1, while the upper bound $\sigma_{\max,k} \leqslant 1$ holds since $\sigma_{\max,1} = 1$ and $\sigma_{\max}$ only decreases in ADAN. Let us consider any iteration where $u^{k+1} \neq u^k$. By Lemma 2.1, we cannot have both $Ad_k = 0$ and $Bd_k = 0$. If $Ad_k = 0$, then

$$\bar{\sigma}_k = 2(1-\gamma)\frac{\delta_k\|\delta_k\|^2 + \rho\|Bd_k\|^2}{\rho\|Bd_k\|^2} \geqslant 2(1-\gamma). \tag{3.6}$$

If $Bd_k = 0$, then we have

$$\bar{\sigma}_k = \frac{2(1-\gamma)\delta_k\|d_k\|^2}{\|Ad_k\|^2} \geqslant \frac{2(1-\gamma)\bar{\delta}}{\|A\|^2}, \tag{3.7}$$

since $\delta_k \geqslant \bar{\delta}$ and $\|d_k\|^2/\|Ad_k\|^2 \geqslant 1/\|A\|^2$. If both $Ad_k \neq 0$ and $Bd_k \neq 0$, then we exploit the inequality

$$\frac{a+b}{A+B} \geqslant \min\left\{\frac{a}{A}, \frac{b}{B}\right\}, \quad \text{where} \quad A > 0, \quad B > 0, \quad a \geqslant 0, \quad b \geqslant 0,$$

in (2.10) to obtain

$$\bar{\sigma}_k \geqslant 2(1-\gamma)\min\left\{\frac{\bar{\delta}}{\|A\|^2}, 1\right\}. \tag{3.8}$$

Note that if either (3.6) or (3.7) holds, then (3.8) holds. If $\sigma_k < \min\{\sigma_{\max,k}, \sigma_{k-1}\}$, then $\sigma_{\max,k} > \sigma_k = \bar{\sigma}_k$. We combine this relation with Step 3 and with (3.8) to deduce that in each iteration where $\sigma_k < \min\{\sigma_{\max,k}, \sigma_{k-1}\}$, we have

$$\sigma_{\max,k+1} = \sigma_{\max,k}/\tau > \sigma_k/\tau = \bar{\sigma}_k/\tau \geqslant \frac{2(1-\gamma)}{\tau}\min\left\{\frac{\bar{\delta}}{\|A\|^2}, 1\right\}. \tag{3.9}$$

In each iteration, σ_k is either $\sigma_{\max,k}$ or $\bar{\sigma}_k$. If $\sigma_k = \sigma_{\max,k}$, then by (3.9) and the fact that the right side is independent of k, the lower bound (3.5) holds. If $\sigma_k = \bar{\sigma}_k$, then by (3.8) and the fact that $\tau > 1$, the lower bound (3.5) holds. This completes the proof.

Lemmas 3.3 and 3.4 imply that the conditions in Steps 2 and 3 are only satisfied in a finite number of iterations. In addition, we have the following monotonicity properties:

Corollary 3.5 *For k sufficiently large, we have $\sigma_k \geqslant \sigma_{k-1}$ and $\delta_k/\sigma_k \leqslant \delta_{k-1}/\sigma_{k-1}$.*

Proof As noted before the corollary, in a finite number of iterations, $\delta_{\min,k}$ and $\sigma_{\max,k}$ reach positive limits which we denote $\delta_{\lim}$ and $\sigma_{\lim}$, respectively. That is, for some $K > 0$, $\delta_{\min,k} = \delta_{\lim}$ and $\sigma_{\max,k} = \sigma_{\lim}$ for all $k \geqslant K$. By Step 1, $\sigma_{k-1} \leqslant \sigma_{\lim}$ when $k > K$. By Step 3, $\sigma_k \geqslant \min\{\sigma_{\lim}, \sigma_{k-1}\} = \sigma_{k-1}$ when $k > K$, which proves the first inequality of the corollary. Now, consider the second inequality. If the conditions of Step 2 do not hold, then either $\delta_k/\sigma_k \leqslant \delta_{k-1}/\sigma_{k-1}$ or $\delta_k \leqslant \max\{\delta_{\min}, \delta_{k-1}\}$. If the first condition holds, then we are done. If the second condition holds, then the same analysis used for σ_k reveals that $\delta_k \leqslant \delta_{k-1}$. Together, the relations $\delta_k \leqslant \delta_{k-1}$ and $\sigma_k \geqslant \sigma_{k-1}$ imply that $\delta_k/\sigma_k \leqslant \delta_{k-1}/\sigma_{k-1}$, which completes the proof.

In our convergence analysis of ADAN, we show that the iterates approach a KKT point for (1.2), that is, a point where the following first-order optimality conditions are satisfied:

$$s - b = 0, \quad A^*(Au - f) + B^*b = 0, \quad Bu - w = 0, \quad s \in \partial\phi(w). \tag{3.10}$$

Since ϕ is convex, the KKT conditions (3.10) are necessary and sufficient for optimality.

Besides the matrix H introduced in (2.8), two other matrices appear throughout the convergence analysis:

$$C_k = \sigma_k^{-1}\left(\delta_k I + \rho B^*B\right),$$

and

$$M_k = C_k + H - 2\rho B^* B = \frac{\delta_k}{\sigma_k} I + A^* A + \rho \left(\frac{1}{\sigma_k} - 1 \right) B^* B. \tag{3.11}$$

Both C_k and M_k are uniformly positive definite since $\sigma_k \leqslant 1$ and $\delta_k \geqslant \bar{\delta}$, the positive starting value for $\delta_{\min}$. In fact, we have $M_k \succeq \bar{\delta} I$. M_k is uniformly bounded due to the upper bound for δ_k in Lemma 3.3 and the lower bound for σ_k in Lemma 3.4.

Let us define the error sequences as follows:

$$u_e^k = u^k - \bar{u}, \quad w_e^k = w^k - \bar{w}, \quad b_e^k = b^k - \bar{b}, \quad s_e^k = s^k - \bar{s},$$

where $(\bar{u},\ \bar{w},\ \bar{b},\ \bar{s})$ satisfies the first-order optimality conditions (3.10). The convergence proof has three main parts:

(1) We show that the quantity

$$E_k = \left\| u_e^k \right\|_{M_k}^2 + \rho \left\| w_e^k \right\|^2 + \frac{1}{\rho} \left\| b_e^k \right\|^2,$$

is a monotone decreasing function of k.

(2) We show that

$$\lim_{k\to\infty} \left\| u^{k+1} - u^k \right\| = \lim_{k\to\infty} \left\| w^{k+1} - w^k \right\| = \lim_{k\to\infty} \left\| b^{k+1} - b^k \right\| = 0.$$

(3) By part 1, the sequence $(u^k,\ w^k,\ b^k)$ is uniformly bounded, which implies that a convergent subsequence exists. Using part 2, we will show that the limit is a KKT point. Using part 1 again, we show that the entire sequence approaches the same limit.

We now present each of these parts.

Part 1 Steps 1 and 4 of ADAN can be written as follows:

$$C_k \left(u^{k+1} - u^k \right) = -\nabla \Psi_k, \tag{3.12}$$

where

$$\begin{aligned} \nabla \Psi_k &= A^* \left(A u^k - f \right) + \rho B^* \left(B u^k - w^k + \rho^{-1} b^k \right) \\ &= H u^k - \rho B^* \left(w^k - \rho^{-1} b^k \right) - A^* f. \end{aligned}$$

The first-order optimality condition associated with Step 5 of ADAN is

$$0 = s^{k+1} + \rho \left(w^{k+1} - B u^{k+1} - \rho^{-1} b^k \right), \tag{3.13}$$

for some $s^{k+1} \in \partial \phi(w^{k+1})$. Step 6 can be rearranged as

$$b^{k+1} - b^k = \rho \left(B u^{k+1} - w^{k+1} \right). \tag{3.14}$$

We combine (3.12)–(3.14) with the first-order optimality conditions (3.10) to obtain

$$
\begin{aligned}
&C_k\left(u_e^{k+1}-u_e^k\right)=-Hu_e^k+\rho B^*\left(w_e^k-\rho^{-1}b_e^k\right),\\
&s_e^{k+1}+\rho w_e^{k+1}-\rho\left(Bu_e^{k+1}+\rho^{-1}b_e^k\right)=0,\\
&b_e^{k+1}-b_e^k=\rho\left(Bu_e^{k+1}-w_e^{k+1}\right).
\end{aligned}
$$

We take the inner products between each of these equations and u_e^{k+1}, w_e^{k+1}, and b_e^k, respectively to obtain the following:

$$\left\langle u_e^{k+1},\, C_k\left(u_e^{k+1}-u_e^k\right)\right\rangle+\langle u_e^{k+1},\, Hu_e^k\rangle=\langle Bu_e^{k+1},\, \rho w_e^k-b_e^k\rangle, \tag{3.15}$$

$$\langle s_e^{k+1},\, w_e^{k+1}\rangle+\rho\left\|w_e^{k+1}\right\|^2=\rho\langle Bu_e^{k+1},\, w_e^{k+1}\rangle+\langle b_e^k,\, w_e^{k+1}\rangle, \tag{3.16}$$

$$\frac{1}{\rho}\langle b_e^k,\, b_e^{k+1}-b_e^k\rangle=\langle b_e^k,\, Bu_e^{k+1}-w_e^{k+1}\rangle. \tag{3.17}$$

For any Hermitian matrix M, the following identity holds:

$$2\mathrm{Re}\,\langle a,\, M(a-b)\rangle=\langle a,\, Ma\rangle-\langle b,\, Mb\rangle+\langle a-b,\, M(a-b)\rangle. \tag{3.18}$$

We apply this to twice the real part of (3.17) to get

$$\frac{1}{\rho}\left(\left\|b_e^{k+1}\right\|^2-\left\|b_e^k\right\|^2\right)=\frac{1}{\rho}\left\|b_e^{k+1}-b_e^k\right\|^2+2\mathrm{Re}\,\langle b_e^k,\, Bu_e^{k+1}-w_e^{k+1}\rangle.$$

Since $b_e^{k+1}-b_e^k=\rho(Bu_e^{k+1}-w_e^{k+1})$, this reduces to

$$\frac{1}{\rho}\left(\left\|b_e^{k+1}\right\|^2-\left\|b_e^k\right\|^2\right)=\rho\left\|Bu_e^{k+1}-w_e^{k+1}\right\|^2+2\mathrm{Re}\,\langle b_e^k,\, Bu_e^{k+1}-w_e^{k+1}\rangle. \tag{3.19}$$

The identity (3.18) is also equivalent to

$$2\mathrm{Re}\,\langle a,\, Mb\rangle=\langle a,\, Ma\rangle+\langle b,\, Mb\rangle-\langle a-b,\, M(a-b)\rangle. \tag{3.20}$$

In (3.15), we apply (3.20) to $\langle u_e^{k+1},\, Hu_e^k\rangle$ and (3.18) to $\langle u_e^{k+1},\, C_k(u_e^{k+1}-u_e^k)\rangle$ to obtain

$$\left\|u_e^{k+1}\right\|_{C_k+H}^2+\left\|u_e^{k+1}-u_e^k\right\|_{N_k}^2-\left\|u_e^k\right\|_{N_k}^2=2\mathrm{Re}\,\langle Bu_e^{k+1},\, \rho w_e^k-b_e^k\rangle, \tag{3.21}$$

where $N_k=C_k-H$. We add the right sides of (3.19), (3.21), and twice the right side of (3.16) to obtain

$$\rho\left[2\mathrm{Re}\,\langle Bu_e^{k+1},\, w_e^k\rangle+\left\|Bu_e^{k+1}\right\|^2+\left\|w_e^{k+1}\right\|^2\right]. \tag{3.22}$$

Using (3.20), we have

$$2\mathrm{Re}\,\langle Bu_e^{k+1},\ w_e^k\rangle = \left\|Bu_e^{k+1}\right\|^2 + \left\|w_e^k\right\|^2 - \left\|Bu_e^{k+1} - w_e^k\right\|^2.$$

Hence, the right side expression (3.22) can be expressed as follows:

$$\rho\left[2\left\|Bu_e^{k+1}\right\|^2 + \left\|w_e^{k+1}\right\|^2 + \left\|w_e^k\right\|^2 - \left\|Bu_e^{k+1} - w_e^k\right\|^2\right]. \tag{3.23}$$

Now add (3.19), (3.21), and twice (3.16) and take into account the right side shown in (3.23) to get

$$\begin{aligned}&\left\|u_e^{k+1}\right\|_{M_k}^2 + \left\|u_e^{k+1} - u_e^k\right\|_{N_k}^2 + \rho\left\|w_e^{k+1}\right\|^2 + \tfrac{1}{\rho}\left\|b_e^{k+1}\right\|^2 + 2\mathrm{Re}\,\langle s_e^{k+1},\ w_e^{k+1}\rangle\\ &\quad = \left\|u_e^k\right\|_{N_k}^2 + \rho\left\|w_e^k\right\|^2 + \tfrac{1}{\rho}\left\|b_e^k\right\|^2 - \rho\left\|Bu_e^{k+1} - w_e^k\right\|^2,\end{aligned} \tag{3.24}$$

where M_k was defined in (3.11).

Since $u^{k+1} = u^k + \sigma_k d_k$, it follows that $u_e^{k+1} - u_e^k = \sigma_k d_k$. Hence, we have

$$\begin{aligned}\left\|u_e^{k+1} - u_e^k\right\|_{N_k}^2 &= \sigma_k^2\left\|d^k\right\|_{N_k}^2 = \sigma_k^2\langle d_k,\ C_k d_k\rangle - \sigma_k^2\langle d_k,\ Hd_k\rangle\\ &= \sigma_k\left(\delta_k\|d_k\|^2 + \rho\|Bd_k\|^2\right) - {\sigma_k}^2\left(\|Ad_k\|^2 + \rho\|Bd_k\|^2\right).\end{aligned}$$

Combine this with (2.9) to obtain

$$\left\|u_e^{k+1} - u_e^k\right\|_{N_k}^2 \geqslant \sigma_k(2\gamma - 1)\left(\delta_k\|d_k\|^2 + \rho\|Bd_k\|^2\right) \geqslant 0, \tag{3.25}$$

since $\gamma > 1/2$. By Lemma 3.3 in [9], the monotonicity of the subdifferential of ϕ implies that

$$\mathrm{Re}\,\langle s_e^{k+1},\ w_e^{k+1}\rangle = \mathrm{Re}\,\langle s^{k+1} - \bar{s},\ w^{k+1} - \bar{w}\rangle \geqslant 0. \tag{3.26}$$

We drop the nonnegative terms $\|u_e^{k+1} - u_e^k\|_{N_k}^2$, $\|Bu_e^{k+1} - w_e^k\|^2$, and $\mathrm{Re}\,\langle s_e^{k+1},\ w_e^{k+1}\rangle$ in (3.24) to obtain an inequality:

$$\left\|u_e^{k+1}\right\|_{M_k}^2 + \rho\left\|w_e^{k+1}\right\|^2 + \frac{1}{\rho}\left\|b_e^{k+1}\right\|^2 \leqslant \left\|u_e^k\right\|_{N_k}^2 + \rho\left\|w_e^k\right\|^2 + \frac{1}{\rho}\left\|b_e^{k^2}\right\|. \tag{3.27}$$

Notice that

$$M_k = C_k + H - 2\rho B^*B = C_k + A^*A - \rho B^*B \succeq C_k - A^*A - \rho B^*B = N_k.$$

It follows from (3.27) that

$$\left\|u_e^{k+1}\right\|_{M_k}^2 + \rho\left\|w_e^{k+1}\right\|^2 + \frac{1}{\rho}\left\|b_e^{k+1}\right\|^2 \leqslant \left\|u_e^k\right\|_{M_k}^2 + \rho\left\|w_e^k\right\|^2 + \frac{1}{\rho}\left\|b_e^k\right\|^2. \tag{3.28}$$

By Corollary 3.5, $1/\sigma_k \leqslant 1/\sigma_{k-1}$ and $\delta_k/\sigma_k \leqslant \delta_{k-1}/\sigma_{k-1}$ for k sufficiently large. This implies that M_k satisfies $M_k \succeq M_{k+1}$ for k sufficiently large. Hence, (3.28) implies that for k sufficiently large,

$$E_{k+1} \leqslant E_k, \quad \text{where} \quad E_k = \left\|u_e^k\right\|_{M_k}^2 + \rho \left\|w_e^k\right\|^2 + \frac{1}{\rho}\left\|b_e^k\right\|^2. \tag{3.29}$$

This completes the proof of part 1.

Part 2 by (3.28) the nonnegative quantity E_k approaches a limit E_∞ as k tends to ∞. We utilize the inequalities $M_k \succeq M_{k+1}$ and $M_k \succeq N_k$ in (3.24) to obtain

$$E_{k+1} + \left\|u_e^{k+1} - u_e^k\right\|_{N_k}^2 + 2\mathrm{Re}\,\langle s_e^{k+1},\, w_e^{k+1}\rangle + \rho \left\|Bu_e^{k+1} - w_e^k\right\|^2 \leqslant E_k.$$

By (3.25) and (3.26), all the terms on the left side of this inequality are nonnegative. Since E_k approaches a limit, we deduce that

$$\lim_{k\to\infty} \left\|u^{k+1} - u^k\right\|_{N_k}^2 = \lim_{k\to\infty} \left\|Bu^{k+1} - w^k\right\| = \lim_{k\to\infty} \mathrm{Re}\langle s_e^k,\, w_e^k\rangle = 0. \tag{3.30}$$

We drop the B term in (3.25) and use the relation $d^k = (u^{k+1} - u^k)/\sigma_k$ to obtain

$$\left\|u_e^{k+1} - u_e^k\right\|_{N_k}^2 \geqslant \left(\frac{(2\gamma - 1)\delta_k}{\sigma_k}\right) \left\|u^{k+1} - u^k\right\|^2. \tag{3.31}$$

By (3.30), the left side $\|u_e^{k+1} - u_e^k\|_{N_k}^2$ tends to 0 as k tends to ∞. Since $\sigma_k \leqslant 1$, $\gamma > 1/2$, and $\delta_k \geqslant \bar{\delta}$, it follows from (3.31) that

$$\lim_{k\to\infty} \left\|u^{k+1} - u^k\right\| = 0. \tag{3.32}$$

By the triangle inequality, we have

$$\begin{aligned} \left\|w^{k+1} - w^k\right\| &\leqslant \left\|w^{k+1} - Bu^{k+2}\right\| + \left\|B\left(u^{k+2} - u^{k+1}\right)\right\| + \left\|Bu^{k+1} - w^k\right\| \\ &= \left\|w_e^{k+1} - Bu_e^{k+2}\right\| + \left\|B\left(u^{k+2} - u^{k+1}\right)\right\| + \left\|Bu_e^{k+1} - w_e^k\right\|. \end{aligned} \tag{3.33}$$

By (3.32), the middle term on the right side of (3.33) tends to 0; and by (3.30), the first and last terms tend to zero. Hence, we conclude that

$$\lim_{k\to\infty} \left\|w^{k+1} - w^k\right\| = 0. \tag{3.34}$$

By (3.14), we have

$$\left\|b^{k+1} - b^k\right\| = \rho \left\|Bu^{k+1} - w^{k+1}\right\| \leqslant \rho \left\|Bu^{k+1} - w^k\right\| + \rho \left\|w^{k+1} - w^k\right\|.$$

By (3.30) and (3.34), both terms on the right side tend to zero, and consequently,

$$\lim_{k\to\infty} \left\| b^{k+1} - b^k \right\| = 0 = \lim_{k\to\infty} \left\| Bu^k - w^k \right\|. \tag{3.35}$$

This completes the proof of part 2.

Part 3 since $M_k \succeq \bar{\delta} I$, it follows that

$$E_k = \left\| u_e^k \right\|_{M_k}^2 + \rho \left\| w_e^k \right\|^2 + \frac{1}{\rho} \left\| b_e^k \right\|^2 \geqslant \bar{\delta} \left\| u_e^k \right\|^2 + \rho \left\| w_e^k \right\|^2 + \frac{1}{\rho} \left\| b_e^k \right\|^2.$$

Since E_k is monotone decreasing and approaching a limit, it follows that the sequence $(u^k,\ w^k,\ b^k)$ is uniformly bounded. Hence, a convergent subsequence $(u^{k_l},\ w^{k_l},\ b^{k_l})$, $l \geqslant 1$, exists that approaches a limit $(u^\infty,\ w^\infty,\ b^\infty)$. By Theorem 23.4 in [20], $\partial\phi(w^\infty)$ is bounded. Let $s^k \in \partial\phi(w^k)$. By Corollary 24.5.1 of [20], for any $\lambda > 0$, there exists $\mu > 0$ with the following property: if $\|w - w^\infty\| \leqslant \mu$, then for each $s \in \partial\phi(w)$, the distance from s to $\partial\phi(w^\infty)$ is less than or equal to λ. Now choose L large enough that $\|w^{k_l} - w^\infty\| \leqslant \mu$ for all $l \geqslant L$. Corollary 24.5.1 shows that any sequence of subdifferentials $s^{k_l} \in \partial\phi(w^{k_l})$ is bounded uniformly for $l \geqslant L$. Hence, there exists a convergent subsequence, also denoted as s^{k_l} for simplicity, which converges to a limit s^∞. By Theorem 24.4 in [20], we have $s^\infty \in \partial\phi(w^\infty)$.

The left side of (3.12) tends to zero as k tends to ∞, since δ_k is bounded from above by Lemma 3.3, σ_k is bound from below by Lemma 3.4, and $\|u^{k+1} - u^k\|$ tends to zero by (3.32). Also, the $Bu^k - w^k$ term in (3.12) tends to 0 by (3.35). Hence, for the convergent subsequence, (3.12) yields the following relation in the limit:

$$A^*\left(Au^\infty - f\right) + B^* b^\infty = 0. \tag{3.36}$$

Similarly, (3.13) and (3.35) in the limit are

$$s^\infty - b^\infty = 0 \quad \text{and} \quad Bu^\infty - w^\infty = 0. \tag{3.37}$$

By (3.36) and (3.37), the limit $(u^\infty,\ w^\infty,\ b^\infty)$ satisfies the first-order optimality conditions (3.10) for (1.2).

The proof of the theorem starts with an arbitrary extreme point $(\bar{u},\ \bar{w},\ \bar{b})$. Let us now consider the specific extreme point $\bar{u} = u^\infty$, $\bar{w} = w^\infty$, and $\bar{b} = b^\infty$ that is the limit of a convergent subsequence $(u^{k_l},\ w^{k_l},\ b^{k_l})$, $l \geqslant 1$. As noted above (3.11), M_k is uniformly bounded. Since the sequence $(u_e^{k_l},\ w_e^{k_l},\ b_e^{k_l})$ converges to 0, it follows that E_{k_l} tends to 0. Since $E_{k+1} \leqslant E_k$ for each k sufficiently large, it follows that the E_k tends to zero. Since M_k is also uniformly positive definite, we conclude that $(u_e^k,\ w_e^k,\ b_e^k)$ tends to 0, which implies that $(u^k,\ w^k,\ b^k)$ converges to $(\bar{u},\ \bar{w},\ \bar{b})$. Since ϕ is convex, $(u^\infty,\ w^\infty)$ is a solution of (1.2) and u^∞ is a solution of (1.1).

4 Application to Parallel MRI

PMRI records the signal response from coils pointed at a target from different directions. The data corresponds to part of the signal's Fourier transform (k-space data). By only recording some components of the Fourier transform, acquisition time is reduced. By recording multiple data sets in parallel, temporal/spatial resolution is increased, and motion-related artifacts are suppressed [17,18,25]. The missing Fourier components lead to aliasing artifacts in images, which are removed through the image reconstruction process [4,10,17,18,25]. One way to reconstruct the image employs sensitivity encoding which uses knowledge of the coil sensitivities to separate aliased pixels; this approach can be modeled as the minimization (1.1). The complexity of solving (1.1) in PMRI is due not only to the nonsmoothness of the regularization term $\phi(Bu)$, but to the ill-conditioning, huge size, and dense structure of A.

4.1 Model

In a K channel coil array, the undersampled k-space data f_i acquired from the ith channel, $i = 1, \cdots, K$, is related to the original full field of view (FOV) image $u^* \in \mathbb{C}^N$ through the ith channel sensitivity $s_i \in \mathbb{C}^N$ by

$$P\mathcal{F}\left(s_i \odot u^*\right) = f_i + n_i, \quad i = 1, \cdots, K, \tag{4.1}$$

where

- $u^* = (u_1^*, \cdots, u_N^*)^\mathsf{T} \in \mathbb{C}^N$, u_i^* is the intensity of the ith pixel in the image and N is the number of pixels in the image;
- $\mathcal{F} \in \mathbb{C}^{N\times N}$ is a 2D discrete Fourier transform;
- P is the undersampling matrix used by all channels, it is the identity matrix with rows removed corresponding to the components of the Fourier transform that are discarded;
- the symbol $\odot$ is the Hadamard (or componentwise) product between two vectors;
- n_i is complex-valued white Gaussian noise with zero mean and standard deviation s for both real and imaginary parts. We used $s = 0.7 \times 10^{-3}$ in our numerical experiments.

We define

$$A := \begin{bmatrix} P\mathcal{F}S_1 \\ \vdots \\ P\mathcal{F}S_K \end{bmatrix} \quad \text{and} \quad f := \begin{bmatrix} f_1 \\ \vdots \\ f_K \end{bmatrix}, \tag{4.2}$$

where $S_i \in \mathbb{C}^{N\times N}$ is the diagonal matrix defined by $S_i = \operatorname{diag}(s_i)$. This defines the quadratic term in (1.1). For the regularization term ϕ, we employ total variation regularization: $\phi\colon \mathbb{C}^{2\times N} \to \mathbb{R}$ and $\phi(\nabla u) = \alpha \sum_{j=1}^N \|(\nabla u)_j\|$ where $(\nabla u)_j \in \mathbb{C}^2$ is the vector of finite differences along the coordinate directions of the image $u \in \mathbb{C}^N$ at the jth pixel. The parameter α is the weight associated with the regularization term. We solve (1.1) to recover u^* using several different data sets.

For total variation regularization, $Bu = \nabla u$ and $\mathcal{N}(B)$ is the set of constant images; that is, $\mathcal{N}(B)$ consists of all multiples of 1, the vector whose entries are all one. In PMRI all the low frequency components of the Fourier transform are normally retained; in particular, the component of the Fourier transform associated with the zero frequency is retained. Since $S_i 1 = s_i$, the component of the Fourier transform $\mathcal{F} S_i 1$ associated with the zero frequency is the sum of the components of s_i. In practice, this sum never vanishes, which implies that $\mathcal{N}(A) \cap \mathcal{N}(B) = \{0\}$. Since $\|\cdot\|$ is strictly convex and $\mathcal{N}(A) \cap \mathcal{N}(B) = \{0\}$, it follows from Lemma 3.1 that there exists a unique solution to (1.1) in PMRI.

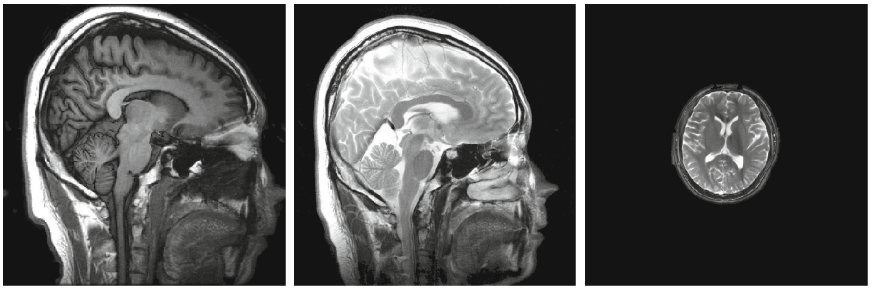
4.2 Data Sets

The three datasets used in our experiments are based on the following two dimensional PMRI images.

Data 1 the *k*-space data corresponds to a sagittal brain image shown in Fig. 1(a) which is fully acquired with an eight-channel head coil. By full acquisition we mean that each receiver coil obtains the complete *k*-space data and hence a high resolution image. The image was acquired on a 3T GE system (GE Healthcare, Waukesha, Wisconsin). The data acquisition parameters were the following: FOV 220 mm^2, size $512 \times 512 \times 8$, repetition time (TR) 3 060 ms, echo time (TE) 126 ms, slice thickness 5 mm, and flip angle 90°. The phase encoding direction was anterior–posterior. This data set was downsampled to size $256 \times 256 \times 8$ to make the dimension of this test set smaller than and different from that of the other test sets.

Data 2 the *k*-space data is associated with a sagittal brain image shown in Fig. 1(b) which is fully acquired with an eight-channel head coil. The image was acquired on a 3T Phillips scanner (Phillips, Best, Netherlands). The acquisition parameters were the following: FOV 205 mm^2, matrix $500 \times 512 \times 8$, slice thickness 5 mm, TR 3 000 ms, TE 85 ms, and flip angle 90°.

Data 3 the *k*-space data is associated with an axial brain image shown in Fig. 1(b) which is fully acquired with an eight-channel head coil. The image was acquired on a 1.5T Siemens Symphony system (Siemens Medical Solutions, Erlangen, Germany).



(a) Sagittal Brain Image **(b)** Sagittal Brain Image **(c)** Axial Brain Image

Fig. 1 (a) Data 1; (b) Data 2 and (c) Data 3

The acquisition parameters were the following: FOV 220 mm^2, matrix $256 \times 512 \times 8$, slice thickness 5 mm, TR 53.5 ms, TE 3.4 ms, and flip angle 75°.

For all three data sets, the ground truth or reference image is given by

$$u_j^* = \left(\sum_{i=1}^{K} |u_{ij}|^2 \right)^{1/2},$$

where K is the number of channels and u_{ij} is the jth component of the image as seen on the ith channel. In the optimization problem (1.1), the matrix A is as given in (4.2), while the vector f is obtained by solving for f_i in (4.1).

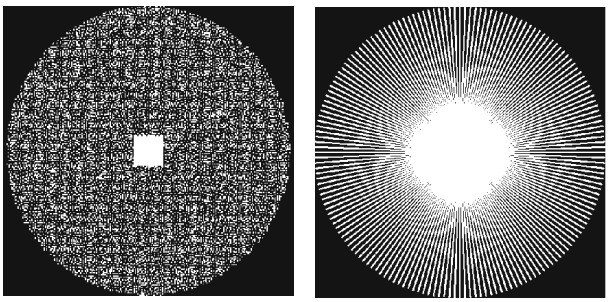
4.3 Observed Data

In (4.1) we use a Poisson random mask (or trajectory) P with a 25% undersampling ratio shown in Fig. 2(a) for Data 1 and Data 2. For Data 3, we use a radial mask (or trajectory) with a 34% undersampling ratio shown in Fig. 2(b). The pixels in the figure are white if the associated Fourier component is recorded and black otherwise. The center of each figure corresponds to the lowest frequency Fourier components.

4.4 Algorithms Comparison

This section compares the performance of the ADAN algorithm to BOS [26] and BOSVS [9] for image reconstruction problems arising in PMRI. For total variation regularization, the computation of the search direction d^k in ADAN can be implemented using a Fourier transform $\mathcal{F}$. That is, as explained in [9,23], there is a diagonal matrix D such that $D = \mathcal{F} B^* B \mathcal{F}^*$. Hence, we have

$$\left(\delta_k I + \rho B^* B\right)^{-1} = \mathcal{F}^* \left(\delta_k I + \rho D\right)^{-1} \mathcal{F} \quad \text{and}$$
$$d^k = \mathcal{F}^* \left(\delta_k I + \rho D\right)^{-1} \mathcal{F} \nabla \Psi_k.$$



(a) Poisson Trajectory **(b)** Radial Trajectory

Fig. 2 (a) Poisson random trajectory with 25% undersampling ratio; (b) radial trajectory with 34% undersampling

The main difference between the BOSVS algorithm and the ADAN algorithm is the computation of u^{k+1}. In ADAN $u^{k+1} = u^k + \sigma_k d^k$, where σ_k is the stepsize. In BOSVS $u^{k+1} = u^k + d^k$, where d^k is initially computed by the same formula (4.3) used in ADAN. If a convergence condition in BOSVS is satisfied, then the update $u^{k+1} = u^k + d^k$ is performed. If the convergence condition is not satisfied, then δ_k is increased and d^k is recomputed. This process of increasing δ_k and recomputing d^k is repeated until a convergence condition is satisfied. The δ_k adjustment process is the following: set $\delta_k = \eta^j \delta_{0,k}$ where $\eta > 1$ and $j \geqslant 0$ is the smallest integer such that $Q_{k+1} \geqslant -C/k^2$ where

$$Q_{k+1} = \xi_k Q_k + \Delta_k, \quad \xi_k = \min\left\{\left(1 - k^{-1}\right)^2, 0.8\right\},$$
$$\Delta_k = \sigma\left(\delta_k \left\|d^k\right\|^2 + \rho \left\|Bd^k + Bu^k - w^k\right\|^2\right) - \left\|Ad^k\right\|^2. \tag{4.3}$$

Here $\delta_{0,k}$ is the same choice for δ_k used in ADAN. Although BOSVS performed well in practice, one needs to choose many parameters in order to achieve this performance. In particular, choices are needed for η, ξ_k, σ, and C. And potentially, one may need to compute several provisional d^k before the subiteration terminates. Also, in BOSVS, the w^k update involves a proximal term with an associated proximal parameter β. For different images, different choices for the parameters may be better than others. In ADAN, these five parameters have been eliminated as well as the recomputation of d^k. ADAN has the same global convergence as that of BOSVS.

4.5 Parameter Settings

Table 1 shows some of the parameter values used in BOS, BOSVS, and ADAN. "NA" means that a parameter is not applicable to an algorithm. The ADAN parameter τ should be larger than 1, while the parameter γ should be larger than 0.5. We choose their values slightly above these lower bounds to guarantee convergence, while not interfering with the performance of the algorithm. On the other hand, the convergence of the BOSVS algorithm is sensitive to the choice of τ. As τ approaches one, the convergence is often faster on average, but highly oscillatory. The choice $\tau = 1.1$ is a compromise between speed and stability. The parameter $\delta_{\min}$ must be positive to ensure that the linear system for the search direction d^k is invertible. $\delta_{\min} = 0.001$ is large enough to ensure invertibility.

The parameters α and ρ are common to all the algorithms, ADAN, BOS, and BOSVS. α is the weight associated with the regularization term, while ρ is the penalty in the augmented Lagrangian (1.3). The parameter α influences the quality of the recon-

Table 1 Parameter values for BOS, BOSVS, and ADAN algorithms

Algorithms	τ	γ	η	σ	C	ξ_k	β
BOS	NA	NA	NA	NA	NA	NA	0
ADAN	1.01	0.5001	NA	NA	NA	NA	NA
BOSVS	1.1	NA	3	0.99	100	(4.3)	0

structed image. As α decreases, more weight is given to the fidelity term $\|Au - f\|^2$ in the objective function, and as α increases, more weight is given to the regularization term ϕ. We adjust α to achieve the smallest error $\|u - u^*\|$ in the reconstructed image u. Table 2 shows how the relative image error $\|u - u^*\|/(\sqrt{N}\|u^*\|)$ depends on α for the three data sets. Based on these results, we took $\alpha = 10^{-5}$.

To obtain a good estimate for the optimal objective in (1.1), we ran ADAN for 100 000 iterations. The optimal objective values for the three data sets were

$$\Phi^* = 0.266\,5 \text{ (Data 1)},$$
$$\Phi^* = 1.052\,5 \text{ (Data 2)},$$
$$\Phi^* = 1.047\,2 \text{ (Data 3)}.$$

The penalty parameter ρ has a significant impact on convergence speed. To determine a good choice for this parameter, we timed how long it took ADAN to reduce the objective error to within 1% of the optimal objective value. The algorithms are coded in MATLAB, version 2011b, and run on a MacbookPro version 10.9.4 with a 2.5 GHz Intel i5 processor. Table 3 shows the number of seconds for each of the three data sets and for ρ between 10^{-2} and 10^{-7}. Based on these results, we took $\rho = 10^{-4}$.

4.6 Experimental Results

This section compares the performance of the existing algorithms BOS and BOSVS with the proposed algorithm ADAN. The initial guess for u^1, b^1, and w^1 was zero for all algorithms. Figure 3 shows the objective value and error as a function of CPU time. Observe that both BOSVS and ADAN converge much more quickly than BOS.

Table 2 Relative image error versus α for the three data sets

Data sets	α				
	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}
1	0.000 278 01	0.000 243 00	0.000 193 17	0.000 255 92	0.000 299 85
2	0.000 085 14	0.000 083 30	0.000 075 15	0.000 076 29	0.000 076 45
3	0.000 071 53	0.000 069 76	0.000 056 43	0.000 052 20	0.000 052 25

Table 3 CPU time in seconds to achieve 1% error versus ρ

Data sets	ρ					
	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}
1	7.968	7.622	8.381	6.531	6.945	13.632
2	50.834	32.876	37.359	38.805	40.659	62.200
3	6.033	6.037	6.065	6.080	6.092	6.565

Hence, there seems to be a significant benefit from using a value for δ_k smaller than the largest eigenvalue of A^*A. Although BOSVS and ADAN are competitive, ADAN is slightly faster and much more stable than BOSVS. The BOSVS objective value can increase or decrease by a factor of 10 in a few iterations.

Table 4 compares the objective value and the numbers of matrix multiplication used by BOS, BOSVS, and ADAN during 100 s CPU time. BOS is able to perform more matrix multiplications during the 100 s, since the algorithm structure is somewhat

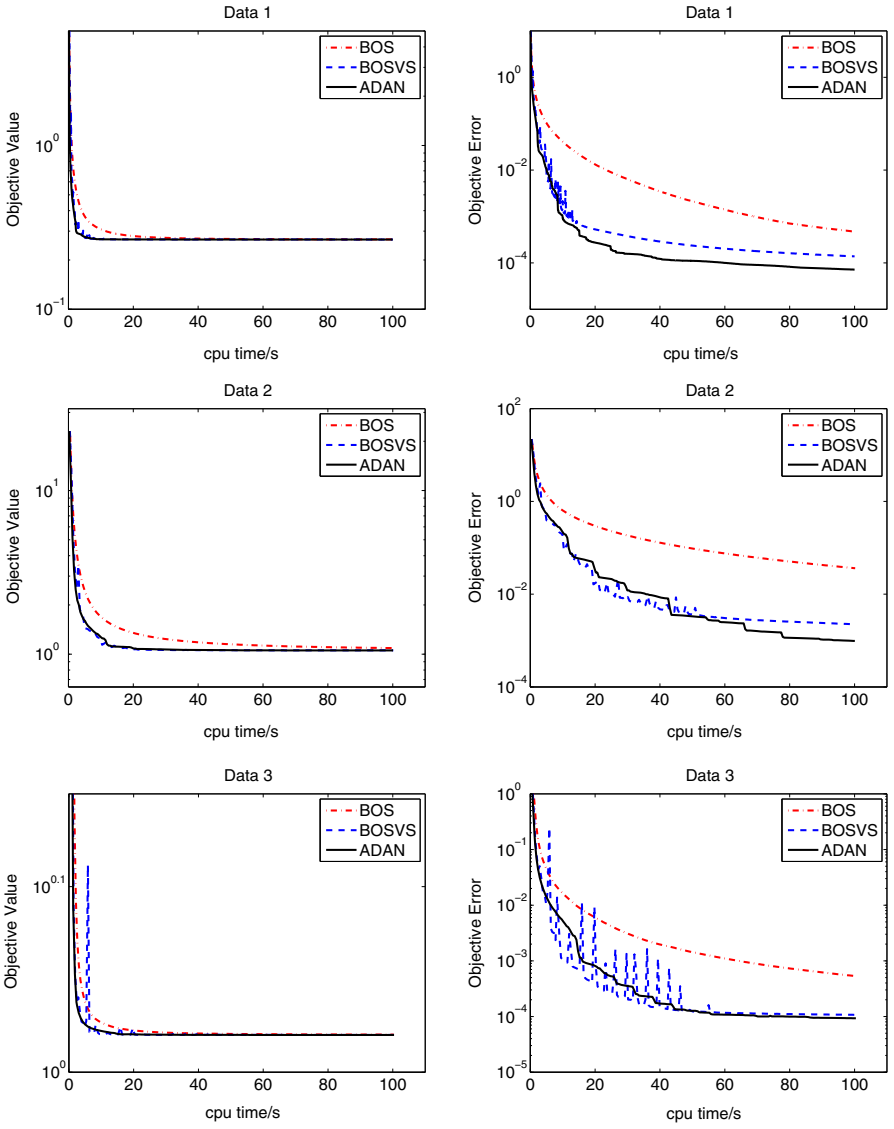


Fig. 3 Comparison of objective values and objective error versus CPU time/s for Data 1, Data 2 and Data 3

simpler than either BOSVS or ADAN. On the other hand, the objective error is much larger for BOS after 100 s, when compared to either BOSVS or ADAN (see Fig. 3).

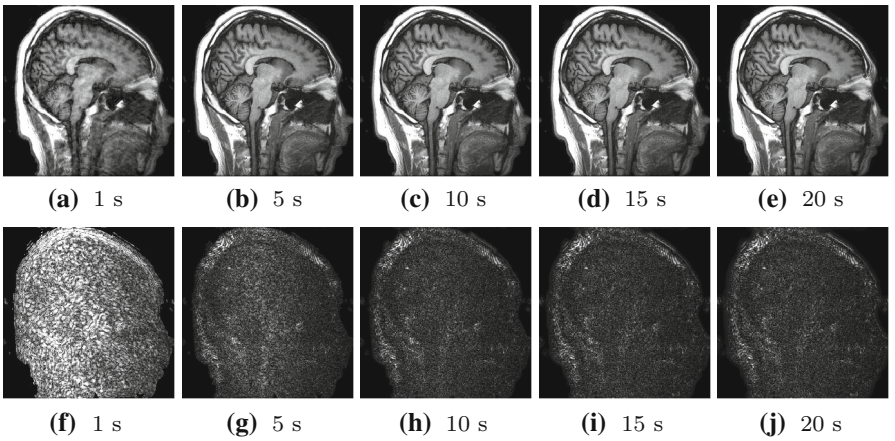
Figure 4 displays the progression of images reconstructed by ADAN verses CPU time. In the error plots, the pixels with the zero absolute error are black, and the pixels with the largest error are white. The reconstructed images for other data sets are of similar quality.

5 Conclusion

The ADAN algorithm was proposed and its global convergence was established for inverse problems of the form (1.1). The algorithm was based on the ADMM [12] and an approximation to Newton's method in which a term in the Hessian is replaced by a BB approximation [2], and a partial step is taken along the approximate Newton search direction. When the algorithm was applied to PMRI reconstruction problems, where A in the fidelity term is a large dense, and ill-conditioned matrix, ADAN was more efficient than an algorithm [6,21,26] based on the proximal ADMM and a fixed positive definite proximal term. In ADAN, the corresponding proximal

Table 4 Objective value (Obj) and the number of matrix multiplications (MM) between A or A' and a vector during 100 s CPU time

Algorithms	Data sets					
	1		2		3	
	Obj	MM	Obj	MM	Obj	MM
BOS	0.267 113	2 031	1.089 897	441	1.047 307	383
BOSVS	0.266 780	1 873	1.055 905	382	1.047 307	352
ADAN	0.266 713	2 014	1.054 646	410	1.047 293	366



(a) 1 s **(b)** 5 s **(c)** 10 s **(d)** 15 s **(e)** 20 s

(f) 1 s **(g)** 5 s **(h)** 10 s **(i)** 15 s **(j)** 20 s

Fig. 4 Data 1: *top row* shows the image reconstruction by ADAN algorithm versus CPU time, while the *bottom row* shows the absolute error for each reconstructed image

term is indefinite; nonetheless, convergence was established and was observed to be relatively fast in the numerical experiments of Sect. 4.

We also compared ADAN to BOSVS [9], a variable stepsize version of BOS which also employs a BB Hessian approximation. ADAN and BOSVS were competitive with each other with regards to speed. However, the algorithms differed significantly in their stability. ADAN tended to converge nearly monotonically, while BOSVS exhibited highly oscillatory convergence; the objective value in BOSVS could increase by more than an order of magnitude within several iterations before dropping a similar amount several iterations later. ADAN is much easier to implement than BOSVS, since the BOSVS line search is replaced by a fixed stepsize in the approximate Newton search direction, and many parameters have been eliminated. In particular, no estimates for Lipschitz constants are required.

Acknowledgments Constructive comments by the reviewers are gratefully acknowledged.

References

[1] Armijo, L.: Minimization of functions having Lipschitz continuous first partial derivatives. Pac. J. Math. **16**, 1–3 (1966)
[2] Barzilai, J., Borwein, J.M.: Two point step size gradient methods. IMA J. Numer. Anal. **8**, 141–148 (1988)
[3] Bioucas-Dias, J., Figueiredo, M., Oliveira, J.P.: Total variation-based image deconvolution: a majorization–minimization approach. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 861–864 (2006)
[4] Block, K., Uecker, M., Frahm, J.: Undersampled radial MRI with multiple coils: iterative image reconstruction using a total variation constraint. Magn. Reson. Med. **57**, 1086–1098 (2007)
[5] Chambolle, A.: An algorithm for total variation minimization and applications. J. Math. Imaging Vis. **20**, 89–97 (2004)
[6] Chambolle, A., Pock, T.: A first-order primal–dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**, 120–145 (2011)
[7] Chan, T.F., Golub, G.H., Mulet, P.: A nonlinear primal–dual method for total variation based image restoration. SIAM J. Optim. **20**, 1964–1977 (1999)
[8] Chen, G., Teboulle, M.: A proximal-based decomposition method for convex minimization problems. Math. Program. **64**, 81–101 (1994)
[9] Chen, Y., Hager, W.W., Yashtini, M., Ye, X., Zhang, H.: Bregman operator splitting with variable stepsize for total variation image reconstruction. Comput. Optim. Appl. **54**, 317–342 (2013)
[10] Doneva, M., Eggers, H., Rahmer, J., Bornert, P., Mertins, A.: Highly undersampled 3D golden ratio radial imaging with iterative reconstruction. In: Proceedings of the International Society for Magnetic Resonance in Medicine, pp. 366 (2008)
[11] Eckstein, J., Bertsekas, D.: On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. **55**, 293–318 (1992)
[12] Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite-element approximations. Comput. Math. Appl. **2**, 17–40 (1976)
[13] Hager, W.W., Yashtini, M., Zhang, H.: An O(1/K) convergence rate for the BOSVS algorithm in total variation regularized least squares. Optimization Online, April 23. http://www.optimizationonline.org/DB_HTML/2015/04/4877.html (2015)
[14] Li, Y., Santosa, F.: An affine scaling algorithm for minimizing total variation in image enhancement. Technical Report TR94-1470. Cornell Theory Center, Cornell University, Ithaca (1994)
[15] Li, Y., Santosa, F.: A computational algorithm for minimizing total variation in image restoration. IEEE Trans. Image Process. **5**, 987–995 (1996)
[16] Osher, S., Burger, M., Goldfarb, D., Xu, J., Yin, W.: An iterative regularization method for total variation-based image restoration. Multiscale Model. Simul. **4**, 460–489 (2005)

[17] Pruessmann, K., Weiger, M., Scheidegger, M., Boesiger, P.: SENSE: sensitivity encoding for fast MRI. Magn. Reson. Med. **42**, 952–962 (1999)
[18] Pruessmann, K., Weiger, M., Bornert, P., Boesiger, P.: Advances in sensitivity encoding with arbitrary *k*-space trajectories. Magn. Reson. Med. **46**, 638–651 (2001)
[19] Raydan, M., Svaiter, B.F.: Relaxed steepest descent and Cauchy–Barzilai–Borwein method. Comput. Optim. Appl. **21**, 155–167 (2002)
[20] Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
[21] Shefi, R., Teboulle, M.: Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. SIAM J. Optim. **24**, 269–297 (2014)
[22] Vogel, C.R.: A multigrid method for total variation-based image denoising. In: Bowers, K., Lund, J. (eds.) Computation and Control IV, Progress in Systems and Control Theory, vol. 20, pp. 323–331. Birkhauser, Boston (1995)
[23] Wang, Y., Yang, J., Yin, W., Zhang, Y.: A new alternating minimization algorithm for total variation image reconstruction. SIAM J. Imaging Sci. **1**, 248–272 (2008)
[24] Yashtini, M., Hager, W.W., Chen, Y., Ye, X.: Partially parallel MR image reconstruction using sensitivity encoding. In: 2012 IEEE International Conference on Image Processing, Orlando, pp. 2077–2080. IEEE, Piscataway, NJ (2012)
[25] Ying, L., Sheng, J.: Joint image reconstruction and sensitivity estimation in SENSE (JSENSE). Magn. Reson. Med. **57**, 1196–1202 (2007)
[26] Zhang, X., Burger, M., Osher, S.: A unified primal–dual algorithm framework based on Bregman iteration. J. Sci. Comput. **46**, 20–46 (2011)