Acceleration of Newton's method using nonlinear Jacobi preconditioning

Konstantin Brenner

Acceleration of Newton's method using nonlinear Jacobi preconditioning

Konstantin Brenner
Université Côte d'Azur, Inria Team Coffee, CNRS, Laboratoire J.A. Dieudonné

Abstract

For mildly nonlinear systems, involving concave or convex diagonal nonlinearities, semi-global monotone convergence of Newton's method is guaranteed provided that the Jacobian of the system has a nonnegative inverse. However, regardless of this convergence result, the efficiency of Newton's method becomes poor for stiff nonlinearities. We propose a nonlinear preconditioning procedure inspired by the Jacobi method and resulting in a new system of equations, which can be solved by Newton's method much more efficiently. The obtained preconditioned method is shown to by globally convergent.

Keywords: Mildly nonlinear systems, Newton's method, Jacobi-Newton method, nonlinear preconditioning, monotone convergence

MSC (2010)**: 58C15, 65H10, 65H20, 65M22**

1 Introduction

Let N be a positive integer we consider the problem of finding $u \in \mathbb{R}^N$ satisfying

$$f(u) + Au = b, \tag{1}$$

where A belongs to the set of real $N \times N$ matrices, denoted in the following by $\mathbb{M}(N)$, and f is a diagonal mapping given by

$$f : u \mapsto \begin{pmatrix} f_1(u_1) \\ \vdots \\ f_N(u_N) \end{pmatrix}.$$

Because of the applications that we have in mind we will assume that f_i are only defined on $\mathbb{R}_{\geq 0}$. More specifically, the analysis presented in this article will be based on the following assumptions:

(A_1) For each $0 \leq i \leq N$, the function f_i is a continuous bijection from $\mathbb{R}_{\geq 0}$ to $\mathbb{R}_{\geq 0}$ belonging to $C^1(0, +\infty)$. The matrix A has zero diagonal and nonpositive off-diagonal elements, and for any $u > 0$ the inverse of $f'(u) + A$ exists and is nonnegative. We assume in addition that $b \in \mathbb{R}^N_{\geq 0}$.

(A_2) For each $0 \leq i \leq N$, the function f_i is concave.

Let us remark that the assumption (A_1) implies that $f'(u) + A$ is an M-matrix and therefore has a positive diagonal (see e.g. 2.4.8 of [9]); it follows in turn that f is increasing and therefore $f(0) = 0$. We also remark that the derivatives of f_i are potentially unbounded at the origin; we will denote $f_i'(0) = \lim_{u \to 0+} f_i'(u)$. Finally, we remark that the analysis presented in the article can be trivially adapted to the case of f_i being convex instead of concave.

The system (1) can be found in the numerical modeling of flow and transport processes. In particular it arises from the discretization of the nonlinear evolutionary PDEs of the form

$$\partial\beta(u) + \operatorname{div}(\mathbf{V}u - \lambda\nabla u) = \gamma(u), \tag{2}$$

where $\mathbf{V}$ is some given velocity field. Applying the backward Euler scheme and some space discretization method to (2) one typically gets the discrete problem of the form

$$\frac{\beta(u_h^n) - \beta(u_h^{n-1})}{\Delta t} + M^{-1} S u_h^n = \gamma(u_h^n) + \sigma_h^n, \tag{3}$$

where u_h^n, u_h^{n-1} are the vectors of the discrete unknowns associated with two sequential time steps, while M and S are respectively the mass and the stiffness matrices, and the vector σ_h^n represent the effect of the boundary conditions.

To fix the ideas let's assume that the Dirichlet boundary conditions are imposed. Several space discretization methods ensure (possibly under some geometrical condition on the mesh) that the matrix $M^{-1}S$ is an M-matrix. In the presence of diffusion (that is $\lambda > 0$), the examples of such monotone discretization schemes would be the standard finite volume method with two-point flux approximation, or P_1 finite element method with mass lumping under the Delaunay condition on the underlying mesh (see [6]). Let us mention that the monotone discretizations are not only beneficial to the nonlinear solver (as it is going to be discussed in this paper), but also allow to preserve the local maximum principle on the discrete level, thus avoiding any spurious oscillations of the discrete solution.

Let D denote the diagonal of $M^{-1}S$ and let $A = \Delta t \left(M^{-1}S - D\right)$. Setting

$$f(u) = \beta(u) + \Delta t(Du - \gamma(u))$$

the system (3) can be written in the form of (1) as

$$f(u_h^n) + A u_h^n = \beta(u_h^{n-1}) + \Delta t \sigma_h^n.$$

Given the assumptions $(A_1) - (A_2)$ on the mapping f, and thus on the nonlinearities $\beta(u)$ and $\gamma(u)$, several physical models are relevant. Such models are for example the porous media equation [10], models of transport in porous media with adsorption (using e.g. the Freundlich isotherm [2]), the Richards' equation [3], [4], or the Dupuit-Forchheimer equation [2] (provided that convection is discretized using an explicit scheme). Let us further remark that the analysis and the algorithms presented in this paper can be extended to the Hele-Shaw or Stefan like problems (see Remark 2.3 below), where $\beta(u)$ is no longer a function, but rather a monotone graph of the form

$$\beta(u) = \zeta H(u) + \tilde{\beta}(u),$$

where $\tilde{\beta}$ is a nondecreasing C^1 concave function, ζ is a nonnegative real number and H denotes the multivalued Heviside graph. In [4] this type of nonlinearity has been addressed through the parametrization of β, that is a couple of the functions $\tau \mapsto (\overline{u}(\tau), \overline{v}(\tau))$ with $\overline{v}(\tau) \in \beta(\overline{u}(\tau))$ for all τ. Then, the problem has been reformulated in terms of this new variable τ.

Due to its quadratic convergence in the vicinity of a solution, Newton's method is a very popular tool for solving systems of algebraic equations, and in particular, those arising from the discretization of the nonlinear PDEs. Let F be some mapping from $\mathbb{R}^N$ to $\mathbb{R}^N$ and assume that F is differentiable in some appropriate sense. Then, starting with some initial guess $u_0 \in \mathbb{R}^N$, Newton's method generates the sequence $(u_n)_n$ defined by

$$u_{n+1} = u_n - F'(u_n)^{-1} F(u_n), \qquad n \geq 0, \tag{4}$$

which hopefully converges to some $u_s \in \mathbb{R}^N$ satisfying $F(u_s) = 0$. Unfortunately the sequence $(u_n)_n$ may not converge, in particular the celebrated Newton-Kantorovich theorem [7], [8] ensures convergence of $(u_n)_n$ only if the initial guess u_0 is sufficiently close to u_s. To overcome this limitation, multiple modifications of a basic Newton's method, involving line search, trust region or homotopy continuation, have been proposed [9], [5].

For system (1) and under Assumption (A_1) there are some variants of Newton's method that ensure convergence of $(u_n)_n$ under some mild if any assumptions on u_0. Those algorithm would typically generate a monotone sequence of lower or upper solutions converging to u_s. Let us briefly review some of those monotone methods. First of all we remark that under assumption (A_2) the sequence generated by (4) will converge monotonically toward any positive solution u_s as soon as the initial guess u_0 satisfies $F(u_0) \leq 0$ (see Proposition 2.3 below), in particular the sequence $(u_n)_n$

satisfies $u_n \leq u_{n+1} \leq u_s$ for all $n \geq 0$. This semi-global convergence result follows from a more general Monotone Newton Theorem (MNT), which were originally introduced by Baluev [1] and is derived from two major assumptions: F is either convex or concave, and $F'(u)^{-1}$ is nonnegative. In this article we use a slightly weaker version of this theorem (Theorem 2.1 below), for more general results we refer to [11], [9], [12] and [13].

If the left hand side of (1) is neither convex nor concave, then the original Newton's method can be modified in a way that the monotone convergence is preserved. This can be achieved either by employing a so-called method of the accelerated monotone iterations [15], [16], or by means of a nested Newton's method [17], [19], [20].

The accelerated monotone iterations, presented in [16], make use of both lower and upper solutions of $F(u) = 0$. In particular, given $\underline{u}_0$ and $\overline{u}_0$ that satisfy $F(\underline{u}_0) \leq 0 \leq F(\overline{u}_0)$, it generates a couple of sequences $(\underline{u}_n)_n$ and $(\overline{u}_n)_n$ defined by

$$\left(\max_{\underline{u}_n \leq \xi \leq \overline{u}_n} F'(\xi) \right) (v_{n+1} - v_n) + F(v_n) = 0 \qquad n \geq 0, \tag{5}$$

with v standing either for $\underline{u}$ or $\overline{u}$. One shows that $\underline{u}_n$ (resp. $\overline{u}_n$) is lower (resp. upper) solution of $F(u) = 0$, that both sequences satisfy $\underline{u}_n \leq u_s \leq \overline{u}_n$ for all $n \geq 0$ and converge monotonically toward u_s. The latter inequality obviously provides a useful error estimate. Note that, for each $n \geq 0$, one has to solve two linear systems resulting from (5) with $v = \underline{u}$ and $v = \overline{u}$. However, those systems only differ by their right-hand-sides, and this situation can be efficiently handled by some linear solvers.

The second method originated from [17], is based on some particular splitting $F(u) = F_1(u) - F_2(u)$, where both mappings F_1 and F_2 are either concave or convex. The system $F_1(u) - F_2(u) = 0$ is solved trough a nested iterative linearization process. The outer loop of the method generates the sequence $(u_n)_n$ defined through the following partial linearization scheme

$$F_1'(u_n)(u_{n+1} - u_n) + F_1(u_n) - F_2(u_{n+1}) = 0, \qquad n \geq 0. \tag{6}$$

One shows that the solution to (6) exists and that the sequence $(u_n)_n$ monotonically converges to u_s. For each $n \geq 0$, the nonlinear system (6) can be solved again by Newton's method. This results in an inner loop generating a sequence that monotonically converges to u_{n+1}. Remark that each inner iteration of the algorithm requires solving a linear system, therefore the total count of linear solves is $N_{outer} \times N_{inner}$.

In contrast with the aforementioned methods we do not aim to relax the convexity/concavity assumption in MNT. Instead, our objective is to accelerate the convergence of the algorithm (4). This will be achieved through a nonlinear preconditioning procedure that preserves the structure required by MNT. As a side note, we remark that, in principle, the preconditioning proposed in this article can be combined with the modified Newton's methods from [16] and [20].

To motivate our study, let us remark that, despite the monotone convergence result, the efficiency of Newton's method applied to (1) can be very poor especially for stiff problems with $f'(0) = +\infty$. To give an example let $\gamma(u) = 0$ and $\beta(u) = u^m, m > 1$ (this choice corresponds to the porous media equation [10]), we demonstrate in the numerical section 3 that the convergence of Newton's method can be very slow; moreover the required number of iterations increases with m. The numerical experiment also demonstrates that the efficiency of Newton's method can be greatly improved by a simple change of the variable $u = \beta(v)$. Let us note that for Richards-like parabolic-elliptic problems with $\beta'(u) = 0$ for $u \geq u_s > 0$ the similar change-of-variable trick can be performed using the variable switching technique as suggested in [4]. Compared to the initial formulation of (1) the drawback of the change-of-variable approaches is that the concavity of the problem is lost, and therefore the monotonic convergence is no longer guaranteed.

In this article, we reformulate (1) in a way that preserves the concavity of the system while offering a much faster convergence of the nonlinear solver. Since the modified system is similar to one obtained in the Jacobi method, we refer to our approach as the Jacobi-Newton method, or the Jacobi preconditioned Newton's method. Note that the Jacobi method can be viewed as a domain decomposition method with the minimal subdomain size and the minimal algebraic overlap. In this regard our approach can be related to the nonlinear domain decomposition methods (see e.g. [23] and [22]).

Because the mapping f is diagonal, strictly increasing and continuous it admits an inverse for all $u \geq 0$. Let g be a diagonal mapping which coincides with f^{-1} on $\mathbb{R}^N_{\geq 0}$, we consider the following left and right-preconditioned problems

$$F_l(u) := u - g(b - Au) = 0 \tag{7}$$

and

$$F_r(\xi) := \xi + Ag(\xi) - b = 0, \tag{8}$$

where (8) has been obtained by a change of variable $u = g(\xi)$. Note that for technical reasons we will also extend g to the whole $\mathbb{R}^N$. This will be done in a way that ensures that g is convex and continuously differentiable on $\mathbb{R}^N$. We then show that $F_\star(u), \star = l, r$ remains concave, that $F'_\star(u)$ exists and has a nonnegative inverse for all $u \in \mathbb{R}^N$. Therefore Newton's iterates corresponding to (7) and (8) converge monotonically. The numerical experiment presented in Section 3 shows that the performance of the preconditioned methods is superior compare to the original formulation of (1), or alternatively to the change-of-variable approaches.

The remainder of the article is organized as follows. In Section 2 we prove the existence and uniqueness of the solution to (1), we present the monotone convergence result for Newton's method applied to the systems (1), (7) and (8); in particular we show that Newton's method applied to (7) and (8) converges independently of the initial guess. In addition in Section 2.1 we deal with the fact that in practice the function g is not evaluated exactly, and we show that a two-level nested Newton's method applied to (7) still exhibits the global convergence.

2 Convergence analysis

In this section we analyze the convergence of Newton's method applied to the problems (1), (7) and (8). To begin with, we present a version of the Monotone Newton Theorem and establish the existence and uniqueness of the solution of (1). Although those two results are quite standard, the proof will be presented for the reader's convenience. Then, the monotone convergence of Newton's method is established for systems (1), (7) and (8). Finally in the subsection 2.1 we investigate the convergence of the preconditioned methods when g is calculated only approximatively.

The analysis presented in this section uses the notions of concavity and inverse isotonicity, so let us recall those definitions. For a more detailed discussion we refer to [9]. Let $\mathcal{D}$ be an open convex subset of $\mathbb{R}^N$ and let $F : \mathcal{D} \to \mathbb{R}^N$ be Gâteaux differentiable. We say that F is concave if

$$F(u) - F(v) \leq F'(v)(u - v) \tag{9}$$

for any $u, v \in \mathcal{D}$, and we say that F is inverse isotone if

$$F(u) \geq F(v) \quad \Rightarrow \quad u \geq v \tag{10}$$

for any $u, v \in \mathcal{D}$; in addition an inverse isotone mapping F is strictly inverse isotone if (10) holds with strict inequalities. Let us remark that inverse isotonicity implies that the equation $F(u) = 0$ has at most one solution. We state below a simple sufficient for the strict inverse isotonicity. For further in-depth discussion of this topic we refer to [14].

Proposition 2.1 (Inverse isotonicity) *Let $\mathcal{D}$ be an open convex subset of $\mathbb{R}^N$ and let $F : \mathcal{D} \to \mathbb{R}^N$, suppose that for any $u, v \in D$ there exists a nonsingular matrix $J(u, v)$ such that $J(u, v)^{-1} \geq 0$ and*

$$F(u) - F(v) \leq J(u, v)(u - v). \tag{11}$$

Then F is strictly inverse isotone.

Let F be a Gâteaux differentiable concave mapping, such that $F'(u)^{-1}$ exists and is nonnegative, then, in view of (9), F satisfies the assumptions of Proposition 2.1 with $J(u, v) = F'(u)$. Similar result holds for a convex mapping with $J(u, v) = F'(v)$. On the other hand, thanks to the mean value theorem, Proposition 2.1 holds for the nonlinear mappings in the left-hand-side of (1), (7) or (8) without convexity/concavity assumption.

Now, let us present a simplified version of the Monotone Newton Theorem from [9] (theorem 13.3.4). Note that in contrast with [9], the monotone convergence result presented below deals with concave mappings. The proof of Theorem 2.1 below is almost identical to the proof given in [9].

Theorem 2.1 (Monotone Newton Theorem) *Let $\mathcal{D}$ be an open convex subset of $\mathbb{R}^N$ and let $F : \mathcal{D} \to \mathbb{R}^N$ be continuous, Gâteaux differentiable and concave, and suppose that $F'(u)$ has a nonnegative inverse for all $u \in \mathcal{D}$. Assume in addition that there exists $u_s \in \mathcal{D}$ satisfying $F(u_s) = 0$ and $u_0 \in D$ such that $F(u_0) \leq 0$. Then the sequence*

$$u_{n+1} = u_n - F'(u_n)^{-1}F(u_n), \qquad n \geq 0 \tag{12}$$

is well defined, satisfies $u_n \leq u_{n+1} \leq u_s$ and $F(u_n) \leq 0$ for all $n \geq 0$, and is convergent. If in addition there exists an invertible $P \in \mathbb{M}(N)$ such that $F'(u_n)^{-1} \geq P \geq 0$ for all $n \geq 0$, then the sequence u_n converges to u_s.

Proof: Assume that $F(u_n) \leq 0$ for some $n \geq 0$ (e.g. $n = 0$), this implies, in view of Proposition 2.1, that $u_n \leq u_s$. Since $F'(u_n)^{-1}$ is nonnegative we deduce from (12) that $u_{n+1} \geq u_n$. On the other hand we have that

$$F(u_n) - F(u_s) \geq F'(u_n)(u_n - u_s)$$

which implied in view of (12) that

$$u_{n+1} = u_n - F'(u_n)^{-1}\left(F(u_n) - F(u_s)\right) \leq u_s.$$

This shows that u_{n+1} satisfy $u_n \leq u_{n+1} \leq u_s$, and in particular $F(u_{n+1})$ is well defined. Using concavity of F and (12) we have that

$$F(u_{n+1}) - F(u_n) \leq F'(u_n)(u_{n+1} - u_n) = -F(u_n), \tag{13}$$

and, thus $F(u_{n+1}) \leq 0$. We have shown that the sequence $(u_n)_n$ remains in $\mathcal{D}$, is nondecreasing and bounded from above, and hence converges to some $\widehat{u} \in \mathcal{D}$. Let us prove that $\widehat{u} = u_s$, since $F'(u)^{-1} \geq P$ we deduce that $u_{n+1} - u_n \geq -PF(u_n) \geq 0$, implying that $\lim_{n\to\infty} F(u_n) = 0$ since P is nonsingular. From continuity and inverse isotonicity of F we deduce that $\widehat{u} = u_s$. □

Remark 2.1 *Assume that F is such that $u - F'(u)^{-1}F(u) \in \mathcal{D}$ for all $u \in \mathcal{D}$, this is true for example if $\mathcal{D} = \mathbb{R}^N$. Then, the algorithm* (12) *is convergent for any initial guess; in particular the sequence $(u_n)_n$ is monotone starting from $n = 1$. To see that, we remark that the estimate $F(u_{n+1}) \leq 0$ in the proof of Theorem 2.1 resulting from* (13) *does not depend on the sign of $F(u_n)$. In fact* (13) *is valid as soon as u_{n+1} is in $\mathcal{D}$.*

Recall that the mappings F_l and F_r introduced in (7) and (8) rely on the function g which has not yet been compliantly defined. The function g coincides on $\mathbb{R}^N_{\geq 0}$ with the inverse of f, let us define it on the whole $\mathbb{R}^N$. Because the functions $u \mapsto f_i'(u)^{-1}$ are continuous, increasing and bounded in the vicinity of zero they can be extended by continuity to $u = 0$. We then define $g : \mathbb{R}^N \to \mathbb{R}^N$ as a diagonal mapping, whose components g_i are given by

$$g_i(u) = \begin{cases} f_i^{-1}(u), & u \geq 0, \\ f_i'(0)^{-1}u, & u < 0. \end{cases} \tag{14}$$

Since $f_i(0) = 0$, the functions g_i are continuous on $\mathbb{R}$; moreover, they are continuously differentiable, and, since f_i' are decreasing, we deduce that the functions g_i are convex. Clearly the mappings F_l and F_r defined by (7) and (8) are concave. Let us show that the inverse of $F_\star'(u), \star = l, r$ is nonnegative for all $u \in \mathbb{R}^N$. In fact the following proposition holds.

Lemma 2.1 *Assume that (A_1) is satisfied, then the matrix $F_\star'(u), \star = l, r$ is an M-matrix satisfying $F_\star'(u) \leq I \leq F_\star'(u)^{-1}$ for all $u \in \mathbb{R}^N$; moreover $F_\star, \star = l, r$ is strictly inverse isotone.*

Proof: Let us first remark that in view of (14) we only need to prove the statement for $u \in \mathbb{R}^N_{\geq 0}$. Let us denote

$$F_u(u) = f(u) + Au - b, \tag{15}$$

and $w = b - Au \geq 0$. Let $\varepsilon \in \mathbb{R}^N_{>0}$, since g' is increasing and $A \leq 0$ we have that

$$F_l'(u) = I + g'(w)A \geq I + g'(w + \varepsilon)A = f'(g(w + \varepsilon))^{-1}F_u'(g(w + \varepsilon)) \tag{16}$$

and

$$F_r'(u) = I + Ag'(u) \geq I + Ag'(u+\varepsilon) = F_u'(g(u+\varepsilon))f'(g(u+\varepsilon))^{-1}. \tag{17}$$

We remark that the use of ε in the inequalities above is motivated by the fact that F_u' is only defined on $\mathbb{R}^N_{>0}$ and not on $\mathbb{R}^N_{\geq 0}$. Since $f'(g(\cdot))$ is a positive and diagonal matrix, we deduce from (A_1) that the right-hand-sides of (16) and (17) have a positive inverse. In addition, since $A \leq 0$ and g' is nonnegative, we deduce from Lemma 2.2 below that $F_\star'(u)^{-1} \geq 0$ for $\star = l, r$; moreover $F_\star'(u)$ is an M-matrix and satisfies $F_\star'(u) \leq I$, which implies in turn that $F_\star'(u)^{-1} \geq I$.

Finally, because g is diagonal and has continuously differentiable components one deduce from the mean value theorem that

$$F_\star(u) - F_\star(v) = F_\star'(z)(u-v), \qquad \star = l, r,$$

with some $z \in \mathbb{R}^N_{\geq 0}$. In view of Proposition 2.1 this implies that $F_\star$ is strictly inverse isotone.

Proposition 2.2 (Existence and uniqueness of the solution) *Assume that* (A_1) *is satisfied, then the solution to* (1) *exists and is unique.*

Proof: Let us consider the mappings $G_l : u \mapsto g(b - Au)$, and let us show that G_l is a contraction. Since $G_l' \geq 0$, we have that $F_l'(u) = I - G_l'(u)$ is a weak regular splitting of $F_l'(u)$ for all u. In view of Lemma 2.1, the matrix $F_l'(u)$, has a nonnegative inverse and we deduce from 2.4.17 of [9] that $\rho(G_l'(u)) < 1$ for all u. This shows that G_l is contractive on $\mathbb{R}^N$ (with respect to some appropriate norm), and, thus G_l has a unique fixed point u_s; moreover the sequence generated by

$$u_{n+1} = G_l(u_n) \tag{18}$$

converges u_s for any u_0. Since $F_l(0) \leq 0$, it follows from Lemma 2.1 that $u_s \geq 0$, and because the restriction of g on $\mathbb{R}^N_{\geq 0}$ is a bijection we deduce that u_s is the unique solution of (1). □

Let us remark that the proof of Proposition 2.2 does not rely on the concavity of $F_\star$. In addition, since (18) can be expressed as

$$u_{n+1} = u_n - F_l(u_n) \tag{19}$$

we observe that the iterative Jacobi process (18) converges component-wise monotonically. We also note that, since $F_\star'(u) \leq I$, we can interpret (19) as a modified Newton method, where $F_l'(u_n)^{-1}$ has been replaced by I, which is a nonnegative subinverse of $F_l'(u)$. We refer to [11] for the analysis of other Newton-like methods of this kind. Let us also note that the stationary iterations

$$\xi_{n+1} = b - Ag(\xi_n)$$

corresponding to the system (8) converge to $\xi_s = f(u_s)$.

We now in the position to prove that Newton's method applied both to the original problem formulation (1) and the preconditioned problems (7) and (8) converges monotonically. Remark however that, since f' may be unbounded at the origin, the mapping F_u' from (15) is only well defined on $\mathcal{D} = \mathbb{R}^N_{>0}$. On the other hand the initial guess u_0 required by Theorem 2.1 needs to satisfy $u_0 \leq u_s$, while u_s does not have to be strictly positive. Therefore, unless some additional hypotheses are made, Newton's method may be inapplicable to the original formulation (1).

Proposition 2.3 (Convergence of the original method) *Assume that* $b > 0$*, then there exists an initial guess* $u_0 > 0$ *such that Newton's method applied to* (1) *converges monotonically.*

Proof: Let F_u be given by (15) and let $\mathcal{D} = \mathbb{R}^N_{>0}$, in view of the assumption (A_1) the mapping F_u defined on $\mathcal{D}$ is continuously differentiable and concave, in addition $F_u'(u)$ has a nonnegative inverse for all $u \in \mathcal{D}$. It remains to show that $u_s \in \mathcal{D}$ and that there exists $u_0 \in \mathcal{D}$ satisfying $F_u(u_0) \leq 0$.

Let $\mathbf{1}_N$ denote the element of $\mathbb{R}^N$ with all unit components, from continuity of f and the fact that $f(0) = 0$ we deduce that there exists $\epsilon > 0$ such that $f(\mathbf{1}_N) \leq b$, and therefore $F_u(\epsilon \mathbf{1}_N) \leq 0$. This implies that $u_0 = \mathbf{1}_N > 0$ is an appropriate initial guess. □

Proposition 2.4 (Convergence of the preconditioned methods) *Newton's method applied to* (7) *and* (8) *converges for any initial guess. In particular the sequence of Newton iterates* $(u_n)_n$ *converges monotonically starting from* $n = 1$.

Proof: It follows form Lemma 2.1 and Proposition 2.2 that $F_\star, \star = l, r$ satisfies the assumptions of Theorem 2.1 with $\mathcal{D} = \mathbb{R}^N$ and $u_0 = 0$. The global convergence follows from Remark 2.1. $\square$

Remark 2.2 *In order to fit the problems* (7) *and* (8) *into the framework of Theorem 2.1 we have defined the mapping g is an extension of f^{-1} to the whole $\mathbb{R}^N$. This is however a rather theoretical construction, since in view of Proposition 2.4 the iterates starting from any $u_0 \geq 0$ such that $F_\star(u_0) \leq 0$, $\star = l, r$ (e.g. $u_0 = 0$) will remain in $\mathbb{R}^N_{\geq 0}$.*

Remark 2.3 *Let us note that the convergence analysis presented above applies to some mildly nonlinear systems that can not be written in the form of* (1). *Let us consider the system*

$$F(\tau) := \overline{v}(\tau) + L\overline{u}(\tau) - b = 0 \tag{20}$$

where $L \in \mathbb{M}(N)$, while $\overline{v}$ and $\overline{u}$ are the diagonal mappings from $\mathbb{R}^N$ to $\mathbb{R}^N$ that are nondecreasing, but not necessarily strictly increasing. The system (20) *typically results from the discretization of some constraint PDEs. Examples of problems leading to* (20) *include degenerate Richards' equation [4], the evolutionary dam problem [21], Stefan or Hele-Shaw problems [10], as well as some classical elliptic or parabolic obstacle problems [18].*

Let D be the diagonal of L and $A = D - L$, then, denoting $\psi(u) = \overline{v}(\tau) + D\overline{u}(\tau)$, we can express the system (20) *as*

$$\psi(\tau) + A\overline{u}(\tau) - b = 0.$$

Assume that ψ is strictly increasing, then, using a new variable $\xi = \psi(\tau)$, and denoting $g(\xi) = \overline{u}(\psi^{-1}(\xi))$, we obtain the system

$$\xi + Ag(\xi) = b \tag{21}$$

similar to (8). *Now, if $\overline{u}$ is merely nondecreasing, then g is not bijective and* (21) *can not be cast into* (1). *Nevertheless preconditioned system similar to* (7) *can be obtained in the following form: Find ξ such that*

$$\xi = b - Au \qquad with \qquad u - g(b - Au) = 0. \tag{22}$$

It is easy to show that ξ is a solution of (22) *if and only if it solves* (21).

Now, let us show that the systems (21) *and* (22) *can be fitted into the framework of the Monotone Newton Theorem. Assume that F form* (20) *is defined on $\mathbb{R}^N$ and that $F'(\tau)$ is an M-matrix for all τ, then on shows that the $I - g(b - Au)'$ and $I + Ag(\xi)'$ are also M-matrices whose inverses are bounded from below by I. Assume in addition that $F(\tau) = 0$ has a solution, then in order to apply Theorem 2.1 it remains to show that g_i are concave for all $i \in \{1, \dots N\}$. In order to do so let us assume that $\overline{v}$ is concave and $\overline{u}$ is convex. Denoting $\zeta = \psi_i^{-1}(\xi)$, we have*

$$g_i'(\xi) = \overline{u}_i'(\zeta)\psi_i'(\zeta)^{-1} = \frac{\overline{u}_i'(\zeta)}{\overline{v}_i'(\zeta) + D_i\overline{u}_i'(\zeta)}.$$

Since the function

$$\gamma(p, q) = \frac{p}{q + D_i p}, \qquad q, p \geq 0$$

is nonincreasing with respect to q and nondecreasing with respect to p we deduce that

$$g'' = \left(\frac{\partial \gamma}{\partial p}\overline{u}_i'' + \frac{\partial \gamma}{\partial q}\overline{v}_i''\right)(\psi_i^{-1})'$$

is nonnegative.

The numerical experiment presented in Section 3 provides the evidences that the preconditioning substantially improves the convergence of Newton's method. To support this observation theoretically we present the following proposition stating that the preconditioned methods lead to a larger solution updates.

Proposition 2.5 *Let $u \in \mathbb{R}^N_{\geq 0}$ be such that $F_u(u) \leq 0$ and $f'(u) < +\infty$. Let $u_{\rm up}$, $u^l_{\rm up}$ and $u^r_{\rm up}$ denote the update generated by Newton's method applied to* (1), (7) *and* (8) *respectively, starting from the initial guess u. Then, $u^l_{\rm up} \geq u_{\rm up}$ and $u^r_{\rm up} \geq u_{\rm up}$.*

Proof: Let us first consider the system (7), and let us denote $w = b - Au$. Since $F_u(u) \leq 0$ we deduce that $f(u) \leq w$ and, thanks to the mean value theorem, we have that

$$F_l(u) = u - g(w) = g(f(u)) - g(w) = f'(g(z))^{-1} F_u(u)$$

for some z satisfying $f(u) \leq z \leq w$. One the other hand,

$$F_l'(u) = I + g'(w)A = I + f'(g(w))^{-1} A.$$

Therefore $u_{\rm up}^l$ satisfies the equation

$$f'(g(z)) \left(I + f'(g(w))^{-1} A\right) \left(u_{\rm up}^l - u\right) = -F_u(u),$$

while $u_{\rm up}$ satisfies

$$f'(u) \left(I + f'(u)^{-1} A\right) (u_{\rm up} - u) = -F_u(u). \tag{23}$$

Since f' is nonincreasing, A is nonpositive and $u \leq g(z) \leq g(w)$ we have that

$$f'(g(z)) \left(I + f'(g(w))^{-1} A\right) \leq f'(u) \left(I + f'(u)^{-1} A\right).$$

In view of (A_1) both sides of the above inequality are the M-matrices, therefore, we deduce that $u_{\rm up}^l \geq u$.

Now, we consider the system (8) and we denote $\xi = f(u)$ and $\xi_{\rm up} = f(u_{\rm up}^r)$. Writing down a single step of Newton's method, and using again the mean value theorem, we have

$$-F_u(u) = (I + Ag'(\xi)))(\xi_{\rm up} - \xi) = (I + Af'(u)^{-1}))f'(z)(u_{\rm up}^r - u)$$

for some z satisfying $u \leq z \leq u_{\rm up}$. In view of (23) and observing that $f'(z) \leq f'(u)$ and deduce that $u_{\rm up}^r \geq u_{\rm up}$. □

2.1 Convergence of the inexact methods

The application of Newton's method to the preconditioned problems (7) and (8) requires evaluation of the function g, which in general can not be done exactly. In order to compute $g(v)$ for some $v \in \mathbb{R}_{\geq 0}^N$ one has to solve a set of scalar nonlinear equations of the form $f(w) = v$. This can be achieved by any appropriate iterative method, such as bisection, *regula falsi* or Newton's method again. The fact that in practice the function g is evaluated only approximatively gives rise to the following sequence of the inexact iterations

$$u_{n+1} = u_n - J_{n,\epsilon}^{-1} F_{n,\epsilon}, \qquad n \geq 0. \tag{24}$$

Here $F_{n,\epsilon}$ and $J_{n,\epsilon}$ denote some approximations of $F(u_n)$ and $F'(u_n)$ respectively. Let us give the conditions under which the inexact method (24) converges to u_s.

Proposition 2.6 *Let $\mathcal{D}$ be an open convex subset of $\mathbb{R}^N$ and let $F : \mathcal{D} \to \mathbb{R}^N$ be continuous, Gâteaux differentiable and concave, and suppose that $F'(u)$ have a nonnegative inverse for all $u \in \mathcal{D}$. Assume in addition that there exists $u_s \in \mathcal{D}$ satisfying $F(u_s) = 0$ and $u_0 \in D$ such that $F(u_0) \leq 0$. Let $(u_n)_n$ be a sequence constructed by the following algorithm: For all $n \geq 0$*

1. *Choose $F_{n,\epsilon}$ such that*
$$F(u_n) \leq F_{n,\epsilon} \leq 0. \tag{25}$$
2. *Choose $J_{n,\epsilon}$ such that*
$$J_{n,\epsilon}^{-1} \geq 0 \qquad \text{and} \qquad J_{n,\epsilon} \geq F'(u_n). \tag{26}$$
3. *Use (24) to compute u_{n+1}.*

Then, the sequence $(u_n)_n$ is well defined for all $n \geq 1$ and satisfy $u_n \leq u_{n+1} \leq u_s$ and $F(u_n) \leq 0$ for all $n \geq 0$.

If in addition

$$\text{there exists an invertible } P \in \mathbb{M}(N) \text{ such that } J_{n,\epsilon}^{-1} \geq P \geq 0 \text{ for all } n \tag{27}$$

and

$$\text{there exists a sequence } (\sigma_n)_n \geq 0 \text{ such that } \lim_{n\to\infty} \sigma_n = 0 \text{ and}$$
$$-\sigma_n \leq F(u_n) - F_{n,\epsilon}, \tag{28}$$

then u_n converges to u_s.

Proof: Let $F(u_n) \leq 0$ for some $n \geq 0$ (e.g. for $n = 0$), since $F_{n,\epsilon} \leq 0$ we deduce from (24) that $u_{n+1} \geq u_n$. Let us show that $u_{n+1} \leq u_s$. From (24), (25), (26) and using concavity of F we deduce that

$$u_{n+1} \leq u_n - F'(u_n)^{-1}F(u_n) = u_n - F'(u_n)^{-1}(F(u_n) - F(u_n)) \leq u_s.$$

This implies in particular that $u_{n+1} \in \mathcal{D}$. It follows from concavity of F that

$$F(u_{n+1}) - F(u_n) \leq F'(u_n)(u_{n+1} - u_n),$$

and using (24) we obtain

$$F(u_{n+1}) - F(u_n) \leq -F'(u_n)J_{n,\epsilon}^{-1}F_{n,\epsilon}$$

or

$$F(u_{n+1}) \leq \left(I - F'(u_n)J_{n,\epsilon}^{-1}\right) F_{n,\epsilon} + F(u_n) - F_{n,\epsilon}.$$

Since $J_{n,\epsilon}^{-1} \geq 0$ we deduce from (26) that

$$I - F'(u_n)J_{n,\epsilon}^{-1} \geq 0,$$

and, in view of (25), we deduce that $F(u_{n+1}) \leq 0$.

The sequence $(u_n)_n$ is nondecreasing and bounded from above; therefore $(u_n)_n$ converges to some $\widehat{u}$. Now, assume that (27) and (28) are satisfied. Combining (27) and (24) we find that

$$0 \geq -P \lim_{n\to\infty} F_{n,\epsilon} \geq 0,$$

which implies that $\lim_{n\to\infty} F_{n,\epsilon} = 0$. In turn, the condition (28) and the continuity of F yield $F(\widehat{u}) = 0$, and in view of Proposition 2.1, we deduce that $\widehat{u} = u_s$. □

To complete this section we show that the nested Newton's method applied to the problem (7) satisfies the assumptions of Proposition 2.6. We begin with a following technical lemma.

Lemma 2.2 *Let $M_1, M_2 \in \mathbb{M}(N)$ with M_1 being an M-matrix, and M_2 having nonpositive off-diagonal elements and satisfying $M_2 \geq M_1$, then $M_2^{-1} \geq 0$.*

Proof: Let D_1 and D_2 denote the diagonal of M_1 and M_2 respectively and $B_1 = D_1 - M_1 \geq 0$, $B_2 = D_2 - M_2 \geq 0$. Since, $M_2 \geq M_1$ we deduce that $D_2 \geq D_1 \geq 0$ and $B_1 \geq B_2 \geq 0$, and therefore

$$0 \leq D_2^{-1}B_2 \leq D_1^{-1}B_1.$$

Denoting by $\rho(M)$ denote the spectral radius of a matrix M, we deduce from 2.4.8 and 2.4.17 of [9] that $\rho(D_2^{-1}B_2) \leq (D_1^{-1}B_1) < 1$, and that $M_2^{-1} \geq 0$. □

The following proposition draws the connection between Proposition 2.6 and the approximate evaluation of the function g.

Proposition 2.7 *Let $u \in \mathbb{R}^N_{\geq 0}$ be such that $F_l(u) \leq 0$. Let w denote the unique solution of $f(w) = b - Au$ and let w_ϵ satisfy*

$$u \leq w_\epsilon \leq w.$$

Let

$$F_\epsilon = u - w_\epsilon \qquad \text{and} \qquad J_\epsilon = I + f'(w_\epsilon)^{-1}A,$$

then

$$F_l(u) \leq F_\epsilon \leq 0, \tag{29}$$

and

$$J_\epsilon \geq F_l'(u) \qquad \text{and} \qquad J_\epsilon^{-1} \geq I. \tag{30}$$

Proof: Note that

$$F_l(u) = u - w \tag{31}$$

and

$$F_\epsilon = u - w_\epsilon \leq 0. \tag{32}$$

Subtracting (32) from (31) we find

$$F_l(u) - F_\epsilon = w_\epsilon - w \leq 0,$$

which, combined with (32), implies (29).

Since $w_\epsilon \leq w$ and since f is diagonal and concave we have that

$$f'(w)^{-1}A \leq f'(w_\epsilon)^{-1}A \leq 0,$$

which implies that $J_\epsilon \geq F_l'(u)$. In turn it follows from Lemma 2.2 that $J_\epsilon^{-1} \geq 0$, and therefore $J_\epsilon \leq I$ implies $J_\epsilon^{-1} \geq I$. □

Let us show that, for a given u_n satisfying $F_l(u_n) \leq 0$, the computation of an approximation $w_{n,\epsilon}$ of w_n satisfying Proposition 2.7 can be achieved by Newton's method. Let $r_n = b - Au_n$, since $F_l(u_n) \leq 0$ and in view of (31) we have that $u_n \leq w_n$ implying that $f(u_n) \leq f(w_n) = r_n$. Let $w_{n,0} = u_n$, we define the sequence $(w_{n,k})_k$ by

$$w_{n,k+1} = w_{n,k} - f'(w_{n,k})^{-1}(f(w_{n,k}) - r_n), \qquad k \geq 0.$$

Using similar arguments as in the proof of Theorem 2.1 one shows that $u_n \leq w_{n,k} \leq w_{n,k+1} \leq w_n$ for all $k \geq 0$ and that the sequence $(w_{n,k})_k$ converges toward w_n. In view of Proposition 2.7 we have that for any $k \geq 0$ the quantities

$$F_{n,\epsilon} = u_n - w_{n,k} \qquad \text{and} \qquad J_{n,\epsilon} = I + (f'(w_{n,k}))^{-1}A$$

satisfy (25), (26) and (27). The extraction of the sequence $w_{n,\epsilon}$ providing (28) can be done by setting $w_{n,\epsilon} = w_{n,\kappa(n)}$ where $\kappa(n)$ is the smallest integer satisfying $w_n - w_{n,\kappa(n)} \leq 10^{-n}$.

Remark 2.4 *The result similar to Proposition 2.7 can be established for the right-preconditioned method. In that case one has to require that w_ϵ satisfies*

$$w_\epsilon \leq w \qquad \textit{and} \qquad u \leq b - Aw_\epsilon.$$

However, in contrast with the left preconditioned method, it is unclear how such approximated values w_ϵ can be constructed in practice.

3 Numerical experiment

Let us consider the porous medium equation (see [10])

$$\partial_t \beta(u) - \partial_{xx} u = 0 \tag{33}$$

on $(0,1) \times (0,T)$. The nonlinearity in the accumulation term is given by $\beta(u) = u^{1/m}$ with $m > 1$. We consider the Neumann boundary conditions

$$\partial_x u(0,t) = -q, \quad \partial_x u(1,t) = 0 \text{ for all } t \in (0,T)$$

with $q > 0$, and the constant initial condition $u(x,0) = u_{ini} > 0$. The value of u_0 is going to be chosen close to zero leading to a traveling wave like solution. Figure 1 exhibit the approximate profile of $\beta(u)$ at different time steps computed for $m = 10, q = 10^4$ and the time step $\Delta t = 1.2\,10^{-4}$. Equation (33) is discretized using the standard finite volume method with the time integration performed by the backward Euler scheme.

Let N be a positive integer, let $h = 1/N$ and let $x_i = i/N$ for $i \in \{0, \ldots, N\}$; the set of the finite volumes $(K_i)_{i\in\{1,\ldots,N\}}$ is defined by $K_i = (x_{i-1}, x_i)$. Let N_T be a positive integer and let

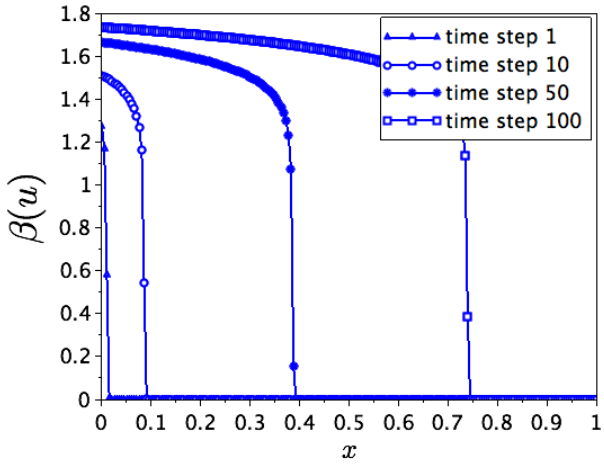


Figure 1: Approximate solution at different time steps

$\Delta t = T/N_T$ define the time step length. Integrating (33) over $K_i \times ((n-1)\Delta t, n\Delta t)$ and using the standard implicit in time finite difference approximation of $\partial_x u$ we obtain the following equation

$$\beta(u_i^n) + \frac{\Delta t}{h^2} \sum_{j \in \mathcal{N}_i} (u_i^n - u_j^n) = \beta(u_i^{n-1}) + \Delta t q \delta_{i,1}, \tag{34}$$

where $\delta_{i,1}$ stands for the Kronecker symbol and where $\mathcal{N}_i$ denotes the set of cells neighboring with K_i

$$\mathcal{N}_i = \begin{cases} \{2\} & i = 1, \\ \{i-1, i+1\} & 1 < i < N, \\ \{N-1\} & i = N. \end{cases}$$

Let L denote the tridiagonal matrix associated to the discretization of the diffusion operator in the left-hand-side of (34) and let, for a given $n \geq 0$, b_n denote the right-hand-side of (34). Imposing (34) for all $i \in \{1, \ldots, N\}$ leads to the following system of algebraic equations

$$\left(\beta(u) + \frac{\Delta t}{h^2} u\right) + \left(L - \frac{\Delta t}{h^2} I\right) u = b_n, \tag{35}$$

which has to be solved for each $n \in \{1, \ldots, N_T\}$. It is easy to verify that $b_n \geq 0$ and that $f(u) = \beta(u) + \dfrac{\Delta t}{h^2} u$ and $A = L - \dfrac{\Delta t}{h^2} I$ satisfy the assumptions $(A_1) - (A_2)$.

The objective of the numerical experiment is to evaluate the efficiency of Newton's method (NM) applied to left and right-preconditioned problems

$$F_l^n(u) := u - g(b_n - Au) = 0 \tag{36}$$

and

$$F_r^n(u) := u + Ag(u) - b_n = 0. \tag{37}$$

Those Jacobi-Newton methods are compared, in terms of the performance, with three more traditional approaches specified below.

$u-$**formulation:** NM applied to (35) in the original form

$$F_u^n(u) := \beta(u) + Lu - b_n = 0. \tag{38}$$

In view of Proposition 2.3 this method is monotonically convergent provided that the initial guess satisfies $F(u_0) \leq 0$.

$v-$**formulation:** The problem (35) is reformulated with respect to the variable v with $u = \beta^{-1}(v)$ and NM is applied to

$$F_v^n(v) := v + L\beta^{-1}(v) - b_n = 0. \tag{39}$$

$\tau-$**formulation:** Following [4] we introduce the function pair $\tau \mapsto (\overline{u}(\tau), \overline{v}(\tau))$ such that

$$\overline{v}(\tau) = \beta(\overline{u}(\tau))$$

for all τ and

$$\max\left(\overline{u}'(\tau), \overline{v}'(\tau)\right) = 1.$$

Then NM is applied to

$$F_\tau^n(\tau) := \overline{v}(\tau) + L\overline{u}(\tau) - b_n = 0. \tag{40}$$

At each time step n and for each of the formulations (36)-(40) the sequence of the approximate solutions is computed using Newton's method

$$\xi_{k+1}^n = \xi_k^n - (F_\star^n)'(\xi_k^n)^{-1} F_\star^n(\xi_k), \qquad \star = u, v, \tau, l, r$$

until the stopping criterion

$$\|F_\star^n(\xi_k)\|_\infty < \epsilon$$

is satisfied for some small predefined tolerance parameter ϵ. As the initial guess we use the value of the variable obtained at the previous time step (this value will obviously differ from one formulation to another). This choice of the initial guess is motivated by the following observation.

Remark 3.1 *The solution of* (33) *(under given initial and boundary conditions) satisfies* $\partial_t u \geq 0$. *This property is reproduced at the discrete level by the approximate solution resulting from* $u-$*formulation and the preconditioned methods. For* $\star = u, r, l$, *let us denote by* u_ϵ^n *the approximate solution of* $F_\star^n(u) = 0$, *then one can show that* $F_\star^n(u_\epsilon^{n-1}) \leq 0$, *and therefore* $u_0^n = u_\epsilon^{n-1}$ *provides an appropriate choice of the initial guess. Let us give the proof by induction for the case of* $u-$*formulation. The proof for preconditioned methods is similar, given that for all* $u \in \mathbb{R}^N_{\geq 0}$, *all* n *and* $\star = r, l$ *one has*

$$F_u^n(u) \leq 0 \Leftrightarrow F_\star^n(u) \leq 0.$$

Let $u_\epsilon^0 = u_{ini}\mathbf{1}_N$, *we have* $F_u^1(u_\epsilon^0) < 0$ *providing, in view of Proposition 2.3, that the sequence of Newton's iterates are monotonically increasing, and that* u_ϵ^1 *satisfies* $u_\epsilon^1 \geq u_\epsilon^0$ *and* $F_u^1(u_\epsilon^1) \leq 0$. *Next, we show that if the statement*

$$u_\epsilon^n \geq u_\epsilon^{n-1}, \qquad F_u^n(u_\epsilon^n) \leq 0 \qquad \textit{and} \qquad F_u^n(u_\epsilon^{n-1}) \leq 0 \tag{41}$$

is true for some $n = p \geq 1$, *then it is true for* $n = p + 1$. *To do that we notice that for* $n \geq 1$

$$F_u^{n+1}(u_\epsilon^n) = F_u^n(u_\epsilon^n) - (\beta(u_\epsilon^n) - \beta(u_\epsilon^{n-1})).$$

Therefore, if (41) *is satisfied for some* $n = p \geq 1$, *then* $F_u^{p+1}(u^p) \leq 0$, *which implies* $u_\epsilon^{p+1} \geq u_\epsilon^p$ *and* $F_u^{p+1}(u_\epsilon^{p+1}) \leq 0$ *in view of Proposition 2.3.*

Now, we present the results of the numerical experiment. The test case is configurated as follows: in order to allow for the use of $u-$formulation we chose a positive initial condition $\beta(u_{ini}) = 10^{-10}$, we set $q = 10^4, T = 1.2\,10^{-2}, N_T = 100$ and we let the parameter m to take values in the set $\{4, 8, 16, 32\}$. For a given value of m, the tolerance ϵ and a specific solution method $\star$ we denote by $\left(u_{m,\epsilon}^{n,\star}\right)_{n\in\{1,\ldots,N_T\}}$ and $\left(v_{m,\epsilon}^{n,\star}\right)_{n\in\{1,\ldots,N_T\}}$ the approximate solution of (35).

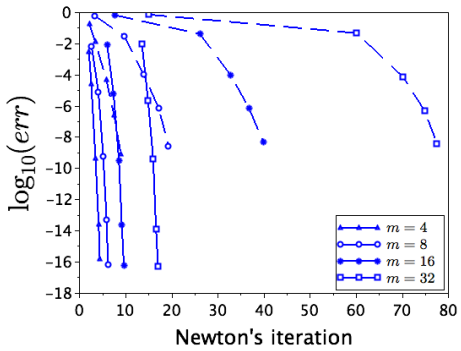
The methodology of the numerical experiment is similar to [4], that is for each value of m we compute, using $\tau-$formulation and the tolerance $\epsilon_{ref} = 10^{-10}$, the reference solution denoted by $\left(u_{m,ref}^{n,\star}\right)_{n\in\{1,\ldots,N_T\}}$ and $\left(v_{m,ref}^{n,\star}\right)_{n\in\{1,\ldots,N_T\}}$. Then, for each solution method (36)-(40) for the tolerance values $\epsilon \in \{10^{-1}, 10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}\}$, we perform the computations measuring the total number of Newton's iteration, required CPU time and the relative deviation from the reference solution. The relative deviation is measured in the discrete $L^\infty(0,T;L^1(0,1))$ norm, and defined by

$$err_{m,\epsilon}^{v,\star} = \frac{\|v_{m,\star}^{\cdot,\star} - v_{m,ref}^{\cdot,\star}\|_{L^\infty(0,T;L^1(0,1))}}{\|v_{m,ref}^{\cdot,\star}\|_{L^\infty(0,T;L^1(0,1))}}.$$

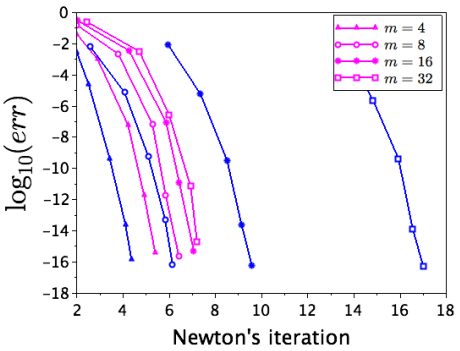
and

$$err^{u,\star}_{m,\epsilon} = \frac{\|u^{\cdot,\star}_{m,\star} - u^{\cdot,\star}_{m,ref}\|_{L^\infty(0,T;L^1(0,1))}}{\|u^{\cdot,\star}_{m,ref}\|_{L^\infty(0,T;L^1(0,1))}}$$

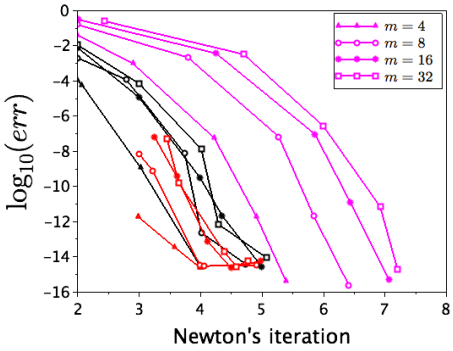
Because the qualitative behaviour of $err^{u,\star}_{m,\epsilon}$ of $err^{v,\star}_{m,\epsilon}$ is similar we will be only reporting the latter error metric.



(a) $err^{v,\star}_{m,\epsilon}$ for $v-$formulation (solid blue) and $u-$formulation (dashed blue)



(b) $err^{v,\star}_{m,\epsilon}$ for $v-$formulation (solid blue) and $\tau-$formulation (magenta)

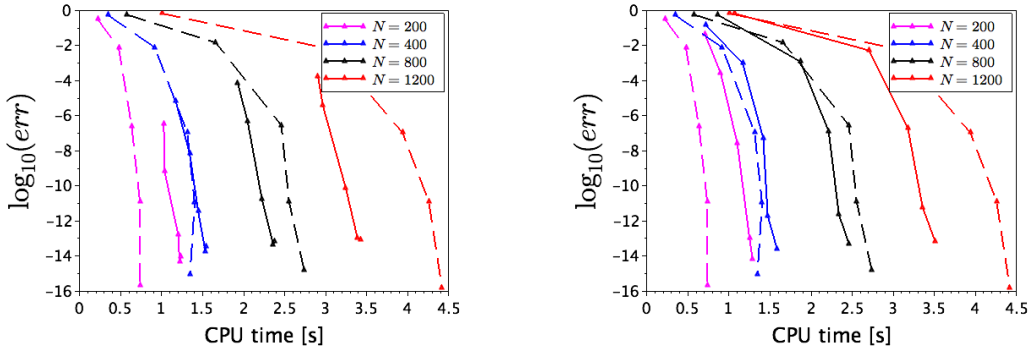


(c) $err^{v,\star}_{m,\epsilon}$ for $\tau-$formulation (magenta), right preconditioned (black) and left preconditioned (red) Newton's method

Figure 2: Relative error $err^{v,\star}_{m,\epsilon}$ as the function of the average number of Newton's iterations per time step

Performance comparison. The first set of tests is performed using the fixed mesh size parameter $N = 100$. In accordance with the results reported in [4], Figures 2(a) and 2(b) witness the qualitative differences in the performance of u and $v-$formulations on one hand, with $v-$formulation being several times faster, and the performance of v and $\tau-$formulations on the other hand, with $\tau-$formulation providing the most significant speedup. It can also be noted that, in contrast with u and $v-$formulations, the sensibility of the $\tau-$formulation to the parameter m is very limited. In turn, Figure 2(c) shows a relatively similar behaviour of $\tau-$formulation and the preconditioned methods, with the latter ones requiring an even smaller number of iterations.

Computational overhead due to local problem solution. As it can be observed on Figure 2(c) the preconditioned Newton's methods require fewer iterations than the method based on $\tau-$formulation. However, each iteration of the Jacobi-Newton method requires to solving a set of the scalar nonlinear equations. Those inner calculations produce a certain computational overhead. To access the overall computational effort required by the preconditioned methods we present the analysis in terms of the CPU time. Figures 3(a) and 3(b) show, for different values of the mesh size parameter $N \in \{200, 400, 800, 1200\}$, the comparison of left (respectively right) preconditioned NM with the method based on $\tau-$formulation. It can be observed that for all except very small prob-



(a) $err^{v,\star}_{m,\epsilon}$ for left preconditioned NM (solid lines) and $\tau-$formulation (dashed lines)

(b) $err^{v,\star}_{m,\epsilon}$ for right preconditioned NM (solid lines) and $\tau-$formulation (dashed lines)

Figure 3: Relative error $err^{v,\star}_{m,\epsilon}$ as the function of CPU time for different grid sizes

lems ($N \lesssim 400$) $\tau-$formulation is outperformed by the preconditioned NM. The observed speedup can be attributed to a smaller number of linear problem solves.

4 Conclusion

For the problems involving only diagonal nonlinearities and satisfying Monotone Newton Theorem, we have proposed a nonlinear preconditioning procedure based on the Jacobi method. This preconditioning is computationally inexpensive and leads to a monotone Newton's method that converges globally and faster than the original one. We believe that this method is particularly efficient for problems involving stiff nonlinearities. This point is illustrated by the numerical experiment based on the porous media equation. We observe that the convergence of the original Newton's method is very slow and deteriorates as the diagonal nonlinearity gets stiffer. In contrast, our newly proposed method exhibits a fast convergence independently of the nonlinear stiffness, which in some sense is absorbed by the preconditioner. The preconditioned method also turns out to be more efficient than the alternative nonmonotone methods based on the change of the variable.

References

[1] A. Baluev. On the method of Chaplygin. Dokl. Akad. Nauk SSSR 83 781-784, 1952.

[2] J. Bear and A. Verruijt. Modeling groundwater flow and pollution. Reidel, 1987.

[3] Duijn, van, C. J., and Peletier, L. A. Nonstationary filtration in partially saturated porous media. Archive for Rational Mechanics and Analysis, 78(2), 173-198, 1982.

[4] K. Brenner, C. Cancès. Improving Newton's method performance by parametrization: the case of Richards equation. SIAM Journal on Numerical Analysis, 2017.

[5] Dennis Jr, John E., and Robert B. Schnabel. Numerical methods for unconstrained optimization and nonlinear equations. Society for Industrial and Applied Mathematics, 1996.

[6] R. Eymard, T. Gallouët, and R. Herbin. Finite Volume Methods, Handbook of Numerical Analysis, volume 7. P.G. Ciarlet and J.L. Lions eds, Elsevier Science B.V., 2000.

[7] L. V. Kantorovich. On Newtons method for functional equations. Dokl. Akad. Nauk SSSR, 59(7):1237–1240, 1948.

[8] J. M. Ortega. The Newton-Kantorovich theorem. Amer. Math. Monthly, 75:658–660, 1968.

[9] J. M. Ortega and W. C. Rheinboldt, Iterative Solutions of Nonlinear Equations in Several Variables, Academic Press, New York, 1970.

[10] J. L. Vázquez. The Porous Medium Equation - Mathematical theory. Oxford Mathematical Monographs. The Clarendon Press Oxford University Press, Oxford, 2007.

[11] J.M. Ortega and W.C. Rheinboldt. Monotone iteration for nonlinear equations with applications to Gauss–Seidel methods. SIAM J. Numer. Analysis, 4 (1967), pp. 171-190

[12] F.A. Potra and W.C. Rheinboldt. On the monotone convergence of Newton's method. Computing, 36 (1986), pp. 81-90

[13] F.A. Potra. Newton-like methods with monotone convergence for solving nonlinear operator equations. Nonlinear Anal., 11 (1987), pp. 697-717.

[14] W. Rheinboldt. On M-functions and their application to nonlinear Gauss-Seidel iterations and to network flows. Journal of Mathematical Analysis and Applications 32 (1970) 274-307

[15] C. V. Pao. Accelerated monotone iterative methods for finite difference equations of reaction-diffusion. Numer. Math. 79, 261–281, 1998

[16] C. V. Pao. Accelerated monotone iterations for numerical solutions of nonlinear elliptic boundary value problems. Comput. Math. Appl., 46, 1535-1544, 2003

[17] L. Brugnano, and V. Casulli. Iterative Solution of Piecewise Linear Systems and Applications to Flows in Porous Media. SIAM J. Sci. Comput., 31(3), 1858–1873, 2009

[18] L. Brugnano, A. Sestini. Iterative solution of piecewise linear systems for the numerical solution of obstacle problems Journal of Numerical Analysis, Industrial and Applied Mathematics, 6 (2012), pp. 67-82

[19] V. Casulli, and P. Zanolli. A Nested Newton-Type Algorithm for Finite Volume Methods Solving Richards' Equation in Mixed Form. SIAM J. Sci. Comput., 32(4), 2255–2273, 2010

[20] V. Casulli, and P. Zanolli. Iterative solutions of mildly nonlinear systems. J. Comput. Appl. Math., 236, 3937–3947, 2012

[21] Carrillo Menéndez, José (1994) On the uniqueness of the solution of the evolution dam problem. Nonlinear Analysis: Theory, Methods & Applications, 22 (5). pp. 573-607.

[22] V. Dolean, M. J. Gander, W. Kheriji, F. Kwok, and R. Masson, Nonlinear preconditioning: how to use a nonlinear Schwarz method to precondition Newton's method, SIAM J. Sci. Comput., 38(6), 2016

[23] X.-C. Cai and D. E. Keyes, Nonlinearly preconditioned inexact Newton algorithms, SIAM J. Sci. Comp., 24(1):183–200, 2002.