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 diagonal nonlinearities, semi-global monotone convergence of Newton's method is guarantied provided that the Jacobian of the system is an M-matrix. However, regardless 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 exhibit semi-global convergence.

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 $\mathbb{M}(N)$, $b \in (\mathbb{R}^+)^N$ and the mapping f is defined by

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

with f_i been strictly increasing continuous functions from $\mathbb{R}^+$ to $\mathbb{R}^+$ satisfying $f_i(0) = 0$. More precisely we will assume the following:

(A_1) For each $0 \leq i \leq N$, f_i is strictly increasing, concave and belongs to C^1 on $(0, +\infty)$.

(A_2) The matrix A has zero diagonal elements and non-positive off-diagonal elements.

(A_3) For any $u > 0$ the matrix $f'(u) + A$ is an M-matrix in the sens of the definition below.

Definition 1 *We say that A is an M-matrix if A is invertible, $A^{-1} \geq 0$, and $a_{i,j} \leq 0$ for $i, j = 1, \ldots, N$ with $i \neq j$.*

Note that the assumption (A_3) implies that $f'(u) + A$ has a strictly positive diagonal (see e.g. 2.4.8 of [5]). We also remark that the derivatives of f_i are potentially unbounded at the origin; we will denote $f_i'(0) = \lim_{u \to 0+} f_i'(u)$.

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

$$\partial_t \beta(u) + \operatorname{div}(\mathbf{v}u - \lambda \nabla u) = \gamma(u). \tag{2}$$

Applying the backward Euler scheme and some space discretization method to (2) one typically get 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 $u_h^n, u_h^{n-1} \in \mathbb{R}^N$ 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 contains boundary data.

To fix the ideas let's assume that the Dirichlet boundary conditions are imposed. Several space discretization methods provide (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 is the standard finite volume method with two-point flux approximation and P_1 finite element method with mass lumping under the Delaunay condition on the underlying mesh (see [4]). 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 osculations of the discrete solution.

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

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

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 assumption (A_1) 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 [6], models of transport in porous media with adsorption (using e.g. the Freundlich isotherm [1]), the Richards' equation [2], [3] or the Dupuit-Forchheimer equation [1] (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 where $\beta(u)$ is no longer a function, but rather a monotone graph of the form

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

where $\tilde{f}$ is a function satisfying the assumption (A_1), ζ is a positive real number and $H(u)$ denotes the multivalued Heviside graph. In [3] this type of nonlinearity has been addressed trough the parametrization of f, that is a couple of the functions $\tau \to (\overline{u}(\tau), \overline{v}(\tau))$ with $\overline{v}(\tau) \in f(\overline{u}(\tau))$ for all τ. The problem has been then rewritten in terms of the new variable τ.

Due to its quadratic convergence, Newton's method is a very popular tool that can be to solve the systems (1) numerically; moreover under assumptions $(A_1) - (A_3)$ one can show that Newton's method converges monotonically toward any strictly positive solution $u_\star$ as soon as the initial guess u_0 satisfies $0 < u_0 \leq u_\star$. This *semi-global* convergence result is based on the concavitty of the underlying functional; it is at fact a straightforward adaptation of the convergence results from [5] (see also Proposition 3 below) to the concave setting.

Despite an available convergence result, the numerical evidences presented in [3] suggest that 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 $\beta(u) = u^{\frac{1}{m}}, m \geq 1$ (this choice corresponds to the porous media equation [6]), as demonstrated in the numerical section 3 convergence of Newton's method can be very slow; moreover the number of Newton's iterations required to solve the system grows as m grows. 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 [3]. Compare 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 guarantied.

In this article we reformulate the system (1) in a way that accelerates convergence of Newton's method while preserving concavity of the problem. More precisely we replace the system (1) by a different one having the same solution set but easier to solve using Newton's method. Since the modified system is similar to the one obtained in Jacobi method, we refer to our approach as to Jacobi preconditioned Newton's method.

The mapping f is diagonal, strictly increasing and continuous and therefore admits an inverse denoted $g = f^{-1}$. We consider the following left-preconditioned and right-preconditioned problems

$$F_l(u) := u - g(b - Au) = 0 \tag{4}$$

or

$$F_r(u) := u + Ag(u) - b = 0. \tag{5}$$

We will show that $F_\star(u)$, $\star = l, r$ remains concave, that $F'_\star(u)$ exists and is an M-matrix for all $u \in (\mathbb{R}^+)^N$. This implies monotone convergence of Newton's method applied to (4) and (5) for any initial guess u_0 satisfying $F_\star(u_0) \leq 0$. The numerical experiment shows (see Section 3) that performance of the preconditioned methods turns out to be superior compare to the original formulation of (1), or alternatively to the change-of-variable approaches.

The reminder of the article is organized as follows. In Section 2, starting with some classical results from [5], we prove existence and uniquness of the solution to (1) and we present the monotone convergence result for Newton's method applied to the problem (1) in its original formulation and applied to the preconditioned problems (4) and (5). In addition in Section 2.3 we deal with the fact that in practice the function g can not be evaluated exactly show that, and we show that a two-level nested Newton's method applied to (4) still exhibits semi-global convergence. Finally, Section 3 is deduced to the numerical experiment.

2 Convergence analysis

2.1 Convergence of Newton's method under partial ordering

In this section we present the adaptation of the convergence result 13.3.4 from [5] to the concave setting. We also prove existence and uniqueness of the solution to (1). Let us first provide some simple observation regarding the M-matrices.

Lemma 1 *Let $A, A' \in \mathbb{M}(n)$ with A been an M-matrix, and A' having non-positive off-diagonal elements and satisfying $A' \geq A$. Then A' is an M-matrix and $(A')^{-1} \leq A^{-1}$.*

Proof: Let D and D' denote the diagonal of A and A' respectively and $B = D - A \geq 0$, $B' = D - A' \geq 0$. Since, $A' \geq A$ we deduce that $D' \geq D$ and $B' \leq B$. The matrix A' can be expressed as follows

$$A' = D' + (D' - D) - (B - (B - B')).$$

Obviously $0 \leq B - B' \leq B$ and the result follows from 2.4.10 of [5].□

Two following lemmas are given without the proof.

Lemma 2 *Let A be an M-matrix and A' be some matrix satisfying $A' \leq A$, then $A^{-1}A' \leq I$ and $A'A^{-1} \leq I$.*

Lemma 3 *Let $A, D \in \mathbb{M}(n)$ with A been an M-matrix and D been strictly positive diagonal matrix. Then the matrices DA and AD are are again the M-matrices.*

The following proposition is necessary to prove uniqueness of the solution to (1).

Proposition 1 (Inverse isotone and uniqueness) *Let F be a continuous G-differentiable concave mapping from $(\mathbb{R}^+)^N$ to $\mathbb{R}^N$ and let $F'(u)$ be an M-matrix for all $u \in (\mathbb{R}^+)^N$. Then F is inverse isotone, that is $F(u) \geq F(v) \Rightarrow u \geq v$ for all $u, v \in (\mathbb{R}^+)^N$, which implies in particular that the solution of $F(u) = 0$ is unique.*

Proof: Since F is concave, we have

$$0 \leq F(u) - F(v) \leq F'(u)(u - v)$$

which provides the result since $F'(u)^{-1} \geq 0$. Let $F(u) = F(v) = 0$, we have

$$F'(v)(u - v) \leq F(u) - F(v) \leq F'(u)(u - v)$$

which implies uniqueness. □

Proposition 2 (Existence and uniqueness of the solution) *The solution to* (1) *exists and is unique.*

Proof: Obviously the problems (1), (4) and (5) have the same solution set. In view of Lemma 4 the mappings F_l and F_r satisfy the assumption of Proposition 1, and therefore (1) has at most one solution.

To prove existence of the solution we consider the mappings $G_l : u \mapsto g(b - Au)$ and $G_r : u \mapsto b - Ag(u)$. Vector $u_\star$ is the solution to (1) if and only if it is a fixed point of G_l (or equivalently G_r). In view of Lemma 4, the matrix $F'_\star$, $\star = l, r$, has a non-negative inverse and we deduce from 2.4.17 of [5] that $\rho(G'_\star) < 1$, implying that $G_\star$ is a contraction. In particular, for any initial guess u_0 the fixed point iterations

$$u_{n+1} = G_\star(u_n)$$

converges to $u_\star$ solution of (1). □

The following Proposition is the straightforward adaptation of 13.3.4 from [5] to the case of concave mappings.

Theorem 1 (Convergence of Newton's method) *Let F be a mapping satisfying the assumptions of Proposition 1. Assume in addition that there exist $u_\star \in (\mathbb{R}^+)^N$ satisfying $F(u_\star) = 0$ and $u_0 \in (\mathbb{R}^+)^N$ such that $F(u_0) \leq 0$. Then the sequence*

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

is well defined, satisfies

$$u_n \leq u_{n+1} \leq u_\star, \qquad F(u_n) \leq 0$$

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 $u_\star$.

Proof: Assume that $F(u_n) \leq 0$ for some $n \geq 0$ (e.g. $n = 0$), this implies, in view of Proposition 1, that $u_n \leq u_\star$. Using concavity of F and (6) we have that

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

and thus $F(u_{n+1}) \leq 0$ implying, in view of Proposition 1, that $u_{n+1} \leq u_\star$. It also follows from (6) that $u_{n+1} \geq u_n$. The sequence $(u_n)_n$, therefore, is non-decreasing and bounded from above, hence it converges to some $\overline{u}$. Let us prove that $\overline{u} = u_\star$, since $F'(u)^{-1} \geq P$ we deduce that

$$u_{n+1} - u_n \geq -PF(u_n) \geq 0.$$

Passing to the limit, we find in view of continuity of F that $F(\overline{u}) = 0$, and, in view of Proposition 1, $\overline{u} = u_\star$. $\square$

Let us denote

$$F_u(u) = f(u) + Au - b. \tag{7}$$

From Theorem 1 we deduce that Newton's method applied to 1 converges monotonically provided that $u_\star > 0$ and $F_u(u_0) \leq 0$. More precisely the follow Proposition holds.

Proposition 3 (Convergence of the original formulation) *Assume that $b > 0$, then there exists the unique solution $u_\star$ to* (1) *satisfying $u_\star > 0$; moreover there exists u_0 such that $F_u(u_0) \leq 0$ and Newton's iterates* (6) *are well defined and monotonically converge to $u_\star$.*

Proof: Let $\mathbf{1}_N$ denote an element of $\mathbb{R}^N$ with all unit components, since f is continuous and $f(0) = 0$ there exists $\epsilon > 0$ such that $F_u(\varepsilon \mathbf{1}_N) \leq 0$. Let $u_0 = \varepsilon \mathbf{1}_N$ and let $\mathbb{R}^N_{x \geq \delta}$ denote the set $\{x \in \mathbb{R}^N | x \geq \delta \mathbf{1}_N\}$ for any $\delta > 0$. The assumptions of Theorem 1 are satisfied with F_u been G-differentiable on $\mathbb{R}^N_{x>\delta}$ instead of $(\mathbb{R}^+)^N$, which implies that the sequence of Newton's iterates $(u_n)_n$ starting at u_0 converges. In addition from the concavity of f and from Lemma 1 we deduce that $F'_u(u_n)^{-1} \geq F'_u(u_0)^{-1}$ for all $u \geq u_0$ and therefore the sequence $(u_n)_n$ converges toward $u_\star$. $\square$

Let us remark that if $f'(0) = +\infty$ the assumption $b > 0$ can not be avoided, therefore the direct application of Newton's method to (1) is somewhat limited by the data. In contrast the preconditioned methods can be applied without restrictions even if f' is unbounded at the origin.

2.2 Convergence of the exact preconditioned methods

In this section we show that the mappings F_l and F_r satisfy the assumption of Theorem 1.

Lemma 4 *The mappings F_l and F_r are concave; moreover for all $u \in (\mathbb{R}^+)^N$ the matrix $F'_\star(u)$, $\star = l, r$ is an M-matrix satisfying $F'_\star(u) \leq I \leq F'_\star(u)^{-1}$.*

Proof: The functions $f_i(u_i)$ are is strictly increasing and concave. Therefore g is strictly increasing and convex. Since $A \leq 0$ the mapping

$$F_l(u) = u - g(b - Au)$$

is concave.

Let us first remark that since the function $g : u \mapsto (f'(u))^{-1}$ is continuous, increasing and bounded on $(0, +\infty)$ it can be extended by continuity to $u = 0$. This implies in particular that g' is well defined in zero. Next,

$$F'_l(u) = I + g'(b - Au)A,$$

thus, in view of the chain rule $g'(v) = f'(g(v))^{-1}$, we have

$$F'_l(u) = I + f'(g(b - Au))^{-1}A$$

showing that $F_l'(u)$ is an M-matrix in view of Assumption (A_3) and Lemma 3. Obviously $F_l'(u) \leq I$, and it follows from Lemma 1 that $F_l'(u)^{-1} \geq I$. Similarly for the right-preconditioned problem we have

$$F_r'(u) = I + Ag'(u) = I + Af'(g(u))^{-1}$$

providing that $F_r'(u)$ is an M-matrix satisfying $F_r'(u) \leq I \leq F_r'(u)^{-1}$ for all $u \geq 0$. □

Proposition 4 (Convergence of the preconditioned methods) *The mappings F_l and F_r satisfy the assumptions of Theorem 1 with $u_0 = 0$.*

Proof: The result follows form Lemma 4 and Proposition 2. □

2.3 Convergence of the inexact method

The application of Newton's method to the preconditioned problems (4) and (5) requires evaluation of the function g, which in general can not be done exactly. At fact, in order to compute $g(v)$ for some $v \in (\mathbb{R}^+)^N$ one has to solve the set of scalar nonlinear equations $f(w) = v$. This can be achieved by any appropriate method, such as bisection, regula falsi or Newton's method again. The fact that the function g is not evaluated exactly gives rise to the following sequence of inexact iterations

$$u_{n+1} = u_n - J_{n,\epsilon}^{-1} F_{n,\epsilon}. \tag{8}$$

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 (8) converges to $u_\star$.

Proposition 5 *Let F be a mapping satisfying the assumptions of Theorem 1 and u_0 be such that $F(u_0) \leq 0$. Consider the the sequence $(u_n)_n$ constructed by the following algorithm: For all $n \geq 0$*

1. *Choose $F_{n,\epsilon}$ such that*
$$F(u_n) \leq F_{n,\epsilon} \leq 0. \tag{9}$$
2. *Choose an M-matrix $J_{n,\epsilon}$ such that*
$$J_{n,\epsilon} \geq F'(u_n). \tag{10}$$
3. *Use (8) to compute u_{n+1}.*

Then, the sequence $(u_n)_n$ is well defined, in particular $F(u_n) \leq 0$ for all $n \geq 1$; moreover $u_n \leq u_{n+1} \leq u_\star$ for all $n \geq 0$, and therefore the sequence $(u_n)_n$ is convergent. If in addition

$$\text{there exists an invertible } P \in \mathbb{M}(N) \text{ such that } J_{n,\varepsilon}^{-1} \geq P \geq 0 \text{ for all } n \tag{11}$$

and

$$\begin{gathered}\text{there exists a sequence } (\sigma_n)_n \geq 0 \text{ such that } \lim_{n\to\infty} \sigma_n = 0 \text{ and} \\ -\sigma_n \leq F(u_n) - F_{n,\varepsilon},\end{gathered} \tag{12}$$

then u_n converges to $u_\star$.

Proof: Let $F(u_n) \leq 0$ for some $n \geq 0$ (e,g, for $n = 0$), since $F_{n,\varepsilon} \leq 0$ we deduce from (8) that $u_{n+1} \geq u_n$; in addition

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

and using (8)

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

or

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

In view of Lemma 2

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

and using $F(u_n) \leq F_{n,\epsilon} \leq 0$ we deduce that $F(u_{n+1}) \leq 0$, and hence $u_{n+1} \leq u_\star$. The sequence $(u_n)_n$ is non-decreasing and bounded from above; therefore $(u_n)_n$ is convergent.

If, in addition, $J_{n,\varepsilon} \geq P \geq 0$, then in view of (8)

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

implying that $\lim_{n\to\infty} F_{n,\epsilon} = 0$. Next

$$-\sigma_n \leq F(u_n) - F_{n,\varepsilon} \leq 0$$

implies that $F(\lim_{n\to\infty} u_n) = 0$ in view of continuity of F, and therefore $\lim_{n\to\infty} u_n = u_\star$. □

To complete this section we show that the nested Newton's method applied to the left-preconditioned problem (4) satisfies the assumptions of Proposition 5. We begin with the following proposition showing the connection between Proposition 5 and the approximate evaluation of g.

Proposition 6 *Let u be such that $F_l(u) \leq 0$. Let w denote the unique solution of $f(w) = b - Au$ and let w_ε satisfy*

$$u \leq w_\varepsilon \leq w.$$

Let

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

then

$$F_l(u) \leq F_\varepsilon \leq 0, \tag{13}$$

and J_ε is an M-matrix satisfying

$$F_l'(u) \leq J_\varepsilon \leq I. \tag{14}$$

Proof: Note that

$$F_l(u) = u - w \tag{15}$$

and

$$F_\varepsilon = u - w_\varepsilon \leq 0. \tag{16}$$

Subtracting (16) from (15) we find

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

which, combined with (16), implies (13). Since $w_\varepsilon \leq w$ and since f is concave we have that $f'(w) \leq f'(w_\varepsilon)$, implying that

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

and therefore

$$F_l'(u) \leq J_\varepsilon \leq I.$$

In view of Lemma 1 the latter inequality implies that J_ε is an M-matrix. □

Let us show that, for a given u_n satisfying $F_l(u_n) \leq 0$, the computation of an approximation $w_{n.\varepsilon}$ of w_n satisfying Proposition 6 can be achieved by Newton's method. Let $r_n = b - Au_n$, since $F_l(u_n) \leq 0$ and in view of (15) we have that $u_n \leq w_n$ implying $f(u_n) \leq f(w_n) = r_n$. Let $w_{n,0} = u_n$, we define the sequence $(w_{n,k})_k$ by

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

Using similar arguments as in the proof of Theorem 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 6 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 (9), (10) and (11). The extraction of the sequence $w_{n,\varepsilon}$ providing (12) can be done by setting $w_{n,\epsilon} = w_{n,\kappa(n)}$ where $\kappa(n)$ is a smallest integer satisfying $w_n - w_{n,\kappa(n)} \leq 10^{-n}$.

Remark 1 *The result similar to Proposition 6 can be established for the right-preconditioned method. In that case one has to requiter that w_ϵ satisfies*

$$w_\varepsilon \leq w \quad \textit{and} \quad u \leq b - Aw_\varepsilon$$

However, in contrast with Proposition 6, it is unclear how such quantities w_ϵ can be constructed in practice.

3 Numerical experiment

Let us consider the porous medium equation (see [6])

$$\partial_t \beta(u) - \partial_{xx}^2 u = 0 \tag{17}$$

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

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

with $q > 0$, and the constant initial condition $u(x,0) = u_0 > 0$. The value of u_0 is going to be chosen close to zero leading to "an almost traveling wave solution". Figure 1 exhibit the approximate profile of $\beta(u)$ at different time steps computed for $m = 10$, $q = 10^4$, $T = 1.2\ 10^{-2}$ and $N_T = 100$.

Equation (17) is discretized using the standard finite volume method with the time integration performed by the backward Euler scheme. Let $N \in \mathbb{N}^*$, let $h = \dfrac{1}{N}$ 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 $N_T \in \mathbb{N}^*$ and let $\Delta t = T/N_T$ define the time step length. Integrating (17) 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

$$\beta(u_i^n) + \frac{\Delta t}{h^2} \sum_{j \in \mathcal{N}_i} (u_i^n - u_j^n) = \beta(u_i^{n-1}) + \frac{\Delta t}{h} q\ \delta_{i,1}, \tag{18}$$

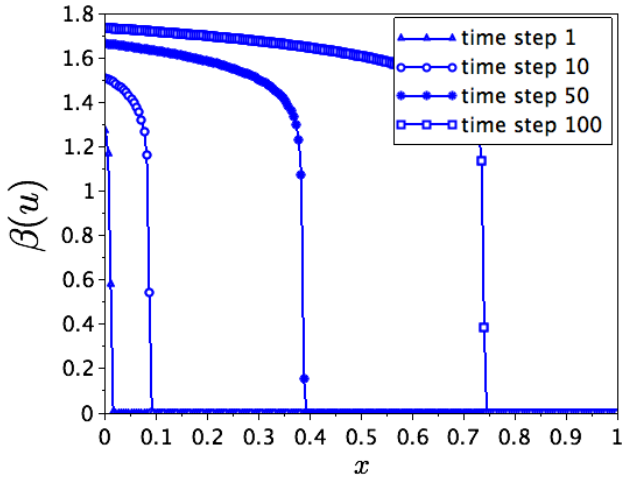


Figure 1: Approximate solution at different time steps

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

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

Let L denote the tridiagonal matrix associated to the discretization of the diffusion operator in the left-hand-side of (18) and let, for a given $n \geq 0$, b_n denote the right-hand-side of (18), the system (18) results in the following problem, which has to be solved for each $n \in \{1, \ldots, N_T\}$

$$\left(\beta(u) + \frac{\Delta t}{h^2} u\right) + \left(L - \frac{\Delta t}{h^2} I\right) u = b_n. \tag{19}$$

It is easy to show that $f(u) = \beta(u) + \dfrac{\Delta t}{h^2} Iu$ and $A = L - \dfrac{\Delta t}{h^2} I$ satisfy the assumptions (A_1)-(A_3).

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

$$F_l^n(u) := u - g\left(b_n - Au\right) = 0 \tag{20}$$

and

$$F_r^n(u) := u + Ag(u) - b_n = 0. \tag{21}$$

Those preconditioned methods are compared, in terms of the performance, with three more standard approaches specified below.

$u-$**formulation:** NM applied to (19) in the original form

$$F_u^n(u) := \beta(u) + Lu - b_n = 0 \tag{22}$$

In view of Proposition 3 this method is monotonically convergent provided that the initial guess satisfy $F(u_0) \leq 0$.

$v-$**formulation:** The problem (19) is reformulated with respect to the variable v with $u = \beta^{-1}(v)$ and NM is applied to

$$F_v^n(v) := v + L\beta^{-1}(v) - b_n = 0 \tag{23}$$

$\tau-$**formulation:** Following [3] we introduce the function pair $\tau \to (\overline{u}(\tau), \overline{v}(\tau))$ 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

$$F_\tau^n(\tau) := \overline{v}(\tau) + L\overline{u}(\tau) - b_n = 0. \tag{24}$$

At each time step n and for each of the formulations (20)-(24) 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

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

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 between the formulations). This choice of the initial guess is motivated by the following observation.

Remark 2 *The solution of* (17) *(under the given initial and boundary conditions) satisfies* $\partial_t u \geq 0$. *This property is reproduced by the discrete solution* u^n *resulting from* $u-$*formulation and the preconditioned methods. For* $\star = u, r, l$, *let us denote again by* u^n *the approximate solution of* $F_\star^n(u) = 0$, *then one can show that* $F_\star^n(u^{n-1}) \leq 0$, *and therefore* $u_0^n = u^{n-1}$ *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* $u \geq 0$, *all* n *and* $\star = r, l$ *one has*

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

For $n = 1$ *we have* $F_u^1(u^0) = -b$ *providing, in view of Proposition 3, that the sequence of Newton's iterates is monotonically increasing, and that* u^1 *satisfies* $u^1 \geq u^0$ *and* $F_u^1(u^1) \leq 0$. *Next, we show that the statement*

$$u^n \geq u^{n-1} \qquad F_u^n(u^n) \leq 0 \qquad \textit{and} \qquad F_u^n(u^{n-1}) \leq 0 \tag{25}$$

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

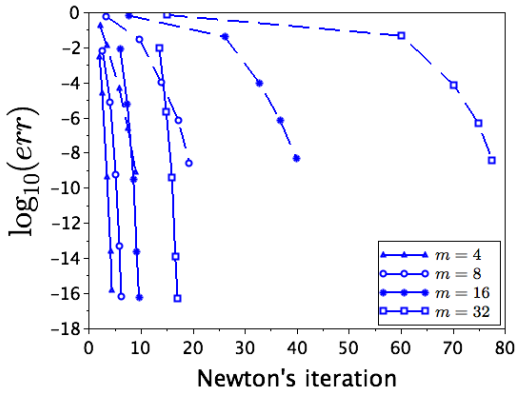
$$F_u^{n+1}(u^n) = F_u^n(u^n) - (\beta(u^n) - \beta(u^{n-1})).$$

Therefore, if (25) *is satisfied for some* $n = m \geq 1$, *then*

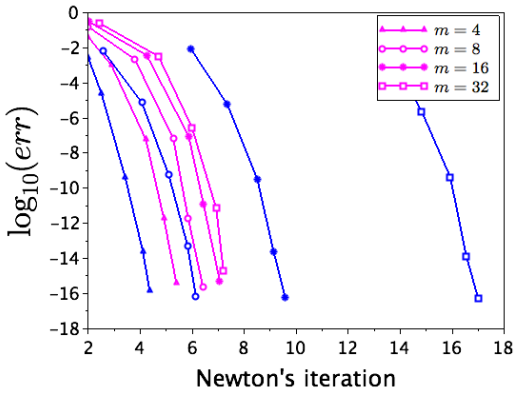
$$F_u^{m+1}(u^m) \leq 0,$$

which implies $u^{m+1} \geq u^m$ *and* $F_u^{m+1}(u^{m+1})$ *in view of Proposition 3.*

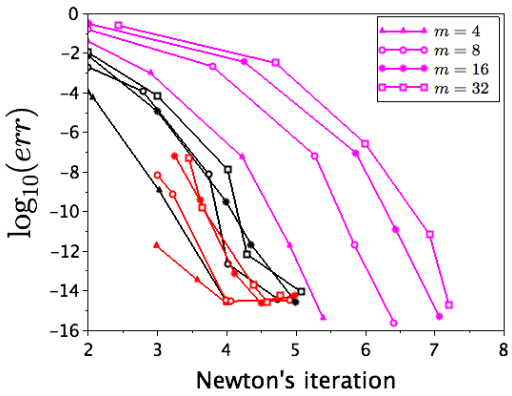
Now, we present the results of the numerical experiment. The test case is configurated as follows: in order to allow for the use of $u-$formulation we chose strictly positive initial condition $\beta(u_0) = 10^{-10}$, we set $q = 10^4$, $T = 1.2\ 10^{-2}$, $N_T = 100$ and we let the parameter m take 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 $\left(u_{m,\varepsilon}^{n,\star}\right)_{n\in\{1,\ldots,N_T\}} \in \mathbb{R}^N$ and $\left(v_{m,\varepsilon}^{n,\star}\right)_{n\in\{1,\ldots,N_T\}} \in \mathbb{R}^N$ the approximate solution of (19).



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



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



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

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

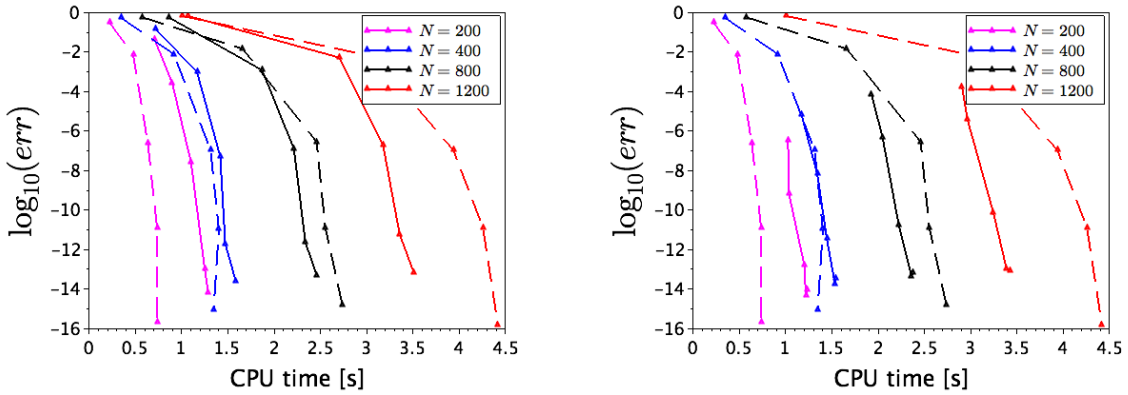
The methodology of the study is similar to [3], that is for each value of m we compute, using $\tau-$formulation and the tolerance $\varepsilon_{ref} = 10^{-10}$, the reference solution denoted by $\left(u^n_{m,ref}\right)_{n\in\{1,\dots,N_T\}}$ and $\left(v^n_{m,ref}\right)_{n\in\{1,\dots,N_T\}}$. Then, for each solution method (20)-(24) and for the tolerance values of $\varepsilon \in \{10^{-1}, 10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}\}$, we perform the computations measuring the total number of Newton's iteration, required CPU time and the deviation from the reference solution. The relative deviation from the reference solution is measured in the discrete $L^\infty(0,T;L^1(0,1))$ norm, and defined by the quantities

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

and

$$err^{v,\star}_{m,\varepsilon} = \frac{\|v^{n,\star}_{m,\varepsilon} - v^n_{m,ref}\|_{L^\infty(0,T;L^1(0,1))}}{\|v^n_{m,ref}\|_{L^\infty(0,T;L^1(0,1))}}.$$

Performance comparison. The first set of tests is performed using the fixed mesh size parameter $N = 100$. In accordance with the results reported in [3], Figures 2a and 2b witness



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

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

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

the qualitative differences in the performance of u and v-formulations on one hand, with v-formulation been few time faster, and the performance of v and τ-formulations on the other hand, with τ-formulation providing the most significant speedup. It can be also noticed that, in contrast with u and v-formulations, the sensibility of the τ-formulation to the parameter m is very limited. In turn, Figure 2c shows a relatively similar behavior of τ-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 2c preconditioned Newton's methods require less iterations then the method based on τ-formulation .However, each iteration of the preconditioned method requires to solve the 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 3a and 3b 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 τ-formulation. In can be observed that for small problems ($N \lesssim 400$) τ-formulation outperforms the preconditioned NM due to the computational overhead related to the latter ones. In turn, for larger problems the preconditioned methods became advantages due to a smaller number of the linear problem solves.

References

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

[2] 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.

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

[4] 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.

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

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