On the monotone convergence of Jacobi-Newton method for mildly nonlinear systems

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 $u \in \mathbb{R}^N_{>0}$ 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.

In view of the assumption (A_1), the matrix $f'(u) + A$ is an M-matrix for any $u \in \mathbb{R}^N_{>0}$, therefore, it has a positive diagonal (see e.g. 2.4.8 of [17]), and 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 $f_i'(0) = \lim_{u \to 0+} f_i'(u)$.

We note that the analysis presented in the article can be trivially adapted to the case of f_i being convex instead of concave. In addition, some of the assumptions in (A_1) are note essential and can, in principle, be removed. In particular, the analysis can be easily adapted to the case where f is merely piecewise C^1. The assumptions on the domain and the image of f, as well as the assumption $b \in \mathbb{R}^N_{\geq 0}$ can also be relaxed as long as one guaranties existence of the solution to (1).

Before moving any further, let us introduce some essential notations. In order to be able to compare real vectors and matrices we introduce the following component-wise partial order on $\mathbb{R}^N$

and $\mathbb{M}(N)$. For any $u, v \in \mathbb{R}^N$ we say that $u \leq v$ (respectively $u < v$) if $u_i \leq v_i$ (respectively $u_i < v_i$) for all $i \in \{1, \ldots, N\}$. Similarly definition holds for $\mathbb{M}(N)$, in particular, a matrix is said to be nonnegative if it has only nonnegative elements.

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

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

where $\mathbf{V}$ is some given velocity field and λ is a nonnegative scalar diffusion coefficient. 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 [13]). Let us mention that such 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 $L = \Delta t M^{-1} S$ and let D denote the diagonal of L. Setting $A = L - D$ and

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

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 medium equation [24], models of transport in porous media with adsorption (using e.g. the Freundlich isotherm [2]), the Richards' equation [23], [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 Heaviside 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 [14], [15] 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 [17], [11].

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 2.4 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 [16], [17], [21] and [20].

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 [18], [19], or by means of the nested Newton's method [5], [9], [10].

The method of the accelerated monotone iterations, proposed in [19], 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 $\overline{u}$, and where the max operator in (5) acts element wise (recall that F corespondents (1) and involve only the diagonal nonlinearity). One shows that $\underline{u}_n$ (resp. $\overline{u}_n$) is a lower (resp. upper) solution of $F(u) = 0$ for all $n \geq 0$, and that both sequences converge monotonically toward u_s. In addition, one has that $\underline{u}_n \leq u_s \leq \overline{u}_n$, which 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 [5], 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 the 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 [19] and [10].

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 $\beta(u) = u^{1/m}, m > 1$ (this choice corresponds to the porous medium equation [24]), 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 $v = \beta(u)$. 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. Some partial theoretical and numerical results regarding this Jacobi-Newton algorithm were already reported in the proceedings article [3]. Compared to [3], this article delivers the detailed proofs and an extended numerical experiment. 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 as proposed in [7] or [12].

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 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 and globally. 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. We also prove that Newton's method applied to (7) and (8) converges at least as fast as the original method associated with (1). 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. Finally, Section 3 is dedicated to the numerical experiment based on the porous medium and Richards' equations.

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 [17]. 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 condition of the strict inverse isotonicity. For further in-depth discussion of this topic we refer to [22].

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 $u, v \in \mathcal{D}$ 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 [17] (theorem 13.3.4). Note that in contrast with [17], 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 [17].

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. Suppose that $F'(u)$ has a nonnegative inverse for all $u \in \mathcal{D}$, and assume that there exist $u_s \in \mathcal{D}$ satisfying $F(u_s) = 0$ and $u_0 \in \mathcal{D}$ 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 $n \geq 0$. 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, that $F(u_{n+1})$ 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. Hence, u_n 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} belongs to $\mathcal{D}$.*

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 $a \mapsto f_i'(a)^{-1}$ defined over $\mathbb{R}_{>0}$ are continuous, increasing and bounded in the vicinity of zero, they can be extended by continuity to $a = 0$. We then define $g : \mathbb{R}^N \to \mathbb{R}^N$ as a diagonal mapping, whose components g_i are given by

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

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. We will show in Proposition 2.2 below that the inverse of $F_\star'(u), \star = l, r$ is also nonnegative for all $u \in \mathbb{R}^N$. Let us begin with some technical results results, starting with two lemmas from [17]. Let $\rho(M)$ denote the spectral radius of a matrix $M \in \mathbb{M}(N)$, the following results hold.

Lemma 2.1 ([17], 2.4.9) *Let $B, C \in \mathbb{M}(N)$ be such that $|C| \leq B$, then $\rho(C) \leq \rho(B)$.*

Lemma 2.2 ([17], 2.4.17) *Let $A \in \mathbb{M}(N)$ and suppose that $A = B - C$ is a weak regular splitting, that is $B, C \in \mathbb{M}(N)$ satisfy: B^{-1} exists and is nonnegative; $B^{-1}C \geq 0$ and $CB^{-1} \geq 0$. Then $\rho(B^{-1}C) < 1$ if and only if A^{-1} exists and is nonnegative.*

Lemma 2.3 *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,

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

We deduce from Lemmas 2.1 and 2.2 that $\rho(D_2^{-1}B_2) \leq (D_1^{-1}B_1) < 1$, and that $M_2^{-1} \geq 0$. □

The following proposition summarizes the properties of $F_\star'(u), \star = l, r$.

Proposition 2.2 *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; moreover, the right-hand-sides of (16) and (17) are M-matrices, since $A \leq 0$ and g' is nonnegative. Then it follows from Lemma 2.3 that $F_\star'(u)^{-1} \geq 0$ for $\star = l, r$. Again, since $A \leq 0$ and g' is nonnegative, we deduce that $F_\star'(u)$ 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 2.3 (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 Proposition 2.2, the matrix $F_l'(u)$, has a nonnegative inverse and we deduce from Lemma 2.2 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 Proposition 2.2 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 2.3 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 [16] 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 $\xi_s = f(u_s)$.

Now, we are in 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 2.4 (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 $f(\epsilon \mathbf{1}_N) \leq b$, and, therefore, $F_u(\epsilon \mathbf{1}_N) \leq 0$, which makes $u_0 = \epsilon \mathbf{1}_N \in \mathcal{D}$ an appropriate initial guess. Finally, it follows from Proposition 2.1 that F_u is inverse isotone, therefore, $F_u(\epsilon \mathbf{1}_N) \leq 0$ implies that $u_s \geq \epsilon \mathbf{1}_N > 0$, which shows that $u_s \in \mathcal{D}$. $\square$

Proposition 2.5 (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 Propositions 2.2 and 2.3 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 2.5, 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 the degenerate Richards' equation [4], the evolutionary dam problem [8], Stefan or Hele-Shaw problems [24], as well as some classical elliptic or parabolic obstacle problems [6].*

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 one shows that* $I - g(b - Au)'$ *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*

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

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 larger solution updates.

Proposition 2.6 *Let* $u \in \mathbb{R}^N_{\geq 0}$ *be such that* $F_u(u) \leq 0$ *and* $f'(u) < +\infty$. *Let* u_+, u_+^l *and* u_+^r *denote the vector generated by a single iteration of Newton's method applied to* (1), (7) *and* (8) *respectively, starting from the initial guess* u. *Then* $u_+^l \geq u_+$ *and* $u_+^r \geq u_+$.

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_+^l satisfies the equation

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

while u_+ satisfies

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

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 $u_+^l \geq u_+$.

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

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

for some z satisfying $u \leq z \leq u_+^r$. In view of (23), and, observing that $f'(z) \leq f'(u)$, we deduce that $u_+^r \geq u_+$. □

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 $v \in \mathbb{R}^N_{\geq 0}$ 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 2.7 *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. Suppose that $F'(u)$ has a nonnegative inverse for all $u \in \mathcal{D}$, and assume that there exist $u_s \in \mathcal{D}$ satisfying $F(u_s) = 0$ and $u_0 \in \mathcal{D}$ 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

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

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

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

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 $\widehat{u} = u_s$. $\square$

To complete this section we show that the nested Newton's method applied to the problem (7) satisfies the assumptions of Proposition 2.7.

Proposition 2.8 *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_ϵ satisfies

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

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 2.3 that $J_\epsilon^{-1} \geq 0$, and, therefore, $J_\epsilon \leq I$ implies $J_\epsilon^{-1} \geq I$. $\square$

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 2.8 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 2.8, 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). Let l be a real number strictly larger than 1, 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 $w_n - w_{n,\kappa(n)} \leq l^{-n}$.

Remark 2.4 *The result similar to Proposition 2.8 can be established for the right-preconditioned method. In that case one has to require that w_ϵ satisfies*

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

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

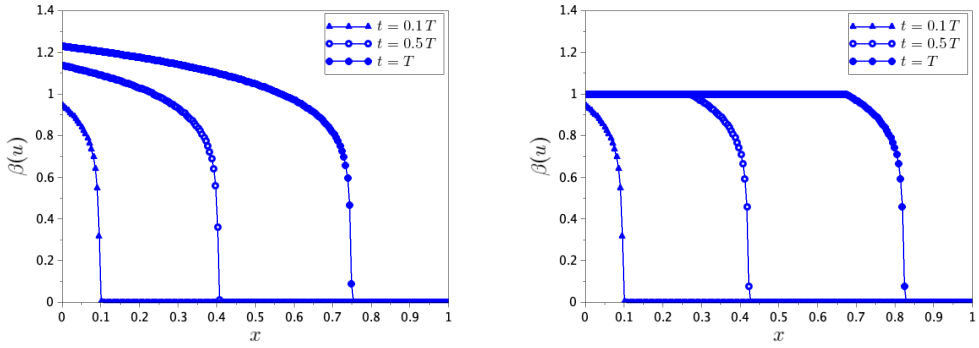


Figure 1: Approximate solution at different times for PME (left) and RE (right) with $m = 8$ and $d = 1$.

3 Numerical experiment

In this section we evaluate the performance of the Jacobi-Newton method applied to the systems resulting from the finite volume discretization of the following evolutionary PDE

$$\partial_t \beta(u) - \Delta u = r. \tag{33}$$

Let $m > 1$, the function β is given either by $\beta(u) = u^{1/m}$, $m > 1$, which corresponds to the porous medium equation [24], or by $\beta(u) = \min(u^{1/m}, 1)$, in which case (33) is parabolic-elliptic and can be interpreted as a zero-gravity Richards' equation using Kirchhoff generalized pressure formulation [4]. We remark that for $m = 2$, one can interpret (33) as the Dupuit-Forchheimer equation [2], [10]. The equation (33) is considered in the space-time domain $\Omega \times (0, T)$, where Ω is a d-dimensional cube, $\Omega = (0, 1)^d$ with $d = 1, 2$. We impose the no-flux Neumann boundary conditions on $\partial\Omega$ and the constant initial condition $u(\mathbf{x}, 0) = u_{ini} > 0$. The flow is driven by the singular right-hand-side term $r = q\delta_{\mathbf{x}_\epsilon}$ with $q > 0$, and $\mathbf{x}_\epsilon \in \Omega$ being some point in the vicinity of the origin. The value of u_{ini} is going to be chosen close to zero, which will lead to the solution exhibiting a sharp front. Figure 1 reports the approximate profile of $\beta(u)$ at times steps computed for $m = 8, q = 8$ and $T = 0.1$.

Let us present the standard implicit finite volume discretization of (33), for more details we refer to [13] and [4]. Let N_x be a positive integer and let $h = 1/N_x$, the space discretization relies on the regular grid composed of $N = N_x^d$ isometric $d-$cubes; more precisely we introduce the family $(K_i)_{i\in\{1,\dots,N\}}$ of $d-$dimentional cubes of measure h^d such that $\overline{\Omega} = \bigcup_{i\in\{1,\dots,N\}} \overline{K_i}$. We denote by $\mathcal{N}_i$ the set of cells neighboring to a given cell K_i, that is

$$\mathcal{N}_i = \{j \in \{1, \dots, N\} | \text{meas}_{d-1}\left(\overline{K_i} \cap \overline{K_j}\right) \neq 0\},$$

where $\text{meas}_d(\cdot)$ stands the $d-$dimentional measure. Without loss of generality we can assume that $K_1 = (0, h)^d$. In addition, we assume that the right-hand-side in (33) is chosen such that $\mathbf{x}_\epsilon \in K_1$; in particular, for any $i \in \{1, \dots, N\}$, we have that $\displaystyle\int_{K_i} r\mathrm{d}\mathbf{x} = q\delta_{i,1}$, where $\delta_{i,1}$ stands for the Kronecker symbol.

Let N_T be a positive integer, the time integration step Δt is defined by $\Delta t = T/N_T$. We denote by u_i^n the discrete solution value associated with the cell K_i and the time step n, and we set the initial condition $u_i^0 = u_{ini}$, $i \in \{1, \dots, N\}$. The finite volume scheme reads as follows: For each $n \in \{1, \dots, N_T\}$ find $(u_i^n)_{i\in\{1,\dots,N\}}$ such that for all $i \in \{1, \dots, N\}$

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

Let L denote the matrix associated with the discrete diffusion operator in the left-hand-side of (34), let D denote the diagonal of L, and let b_n denote the right-hand-side of (34). Imposing (34)

for all $i \in \{1, \ldots, N\}$ we obtain the following system of algebraic equations

$$\beta(u^n) + Lu^n = b_n. \tag{35}$$

Alternatively we can express (35) as

$$f(u^n) + Au^n = b_n, \tag{36}$$

where $f(u) = \beta(u) + Du$ and $A = L - D$. It is easy to show that f and A satisfy the assumptions $(A_1) - (A_2)$ except for the regularity condition on f. The latter assumption is violated in the case $\beta(u) = \min(u^{1/m}, 1)$.

Because β' (and thus f') is singular at the origin, the performance of Newton's method applied to (36) may be very poor and deteriorates as m increases. This singularity can be removed by some *ad hoc* change of the variable (see (40) and (41) below) leading to radical improvement in the performance of the resulting Newton's method. However the drawback of the change-of-variable approaches is that, in general, the concavity of the discretized problem is lost, and, therefore, the monotonic convergence is no longer guaranteed. Alternatively, the Jacobi-Newton method allows to remove the singularity in (36), while preserving the monotone convergence. In the numerical experiment we will evaluate the efficiency of Newton's method (NM) applied to left and right-preconditioned problems

$$F_l^n(u) := u - g(b_n - Au) = 0 \tag{37}$$

and

$$F_r^n(u) := u + Ag(u) - b_n = 0. \tag{38}$$

Those Jacobi-Newton methods are compared to three more traditional approaches specified below.
$u-$**formulation:** NM applied to (36) in the original form

$$F_u^n(u) := \beta(u) + Lu - b_n = 0. \tag{39}$$

In view of Proposition 2.4, this method is monotonically convergent provided that the initial guess is a lower solution of (39).
$v-$**formulation:** The problem (36) is reformulated with respect to the variable v with $u = \beta^{-1}(v)$ and NM is applied to

$$F_v^n(v) := v + L\beta^{-1}(v) - b_n = 0. \tag{40}$$

This approach requires β to be invertible and therefore can not be applied in the parabolic-elliptic case $\beta(u) = \min(u^{1/m}, 1)$.
$\tau-$**formulation:** Following [4] we introduce the function pair $\tau \mapsto (\overline{u}(\tau), \overline{v}(\tau))$ satisfying

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

for all τ. Then, NM is applied to

$$F_\tau^n(\tau) := \overline{v}(\tau) + L\overline{u}(\tau) - b_n = 0. \tag{41}$$

Following [4], we define $\overline{u}(\tau)$ and $\overline{v}(\tau))$ based on the condition

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

This leads to the following explicit expressions:

$$\overline{v}(\tau) = \begin{cases} \tau & \text{if } \tau < \widetilde{\tau}, \\ (\tau - \widetilde{\tau} + \widetilde{\tau}^m)^{1/m} & \text{if } \tau \geq \widetilde{\tau}, \end{cases} \tag{42}$$

and

$$\overline{u}(\tau) = \begin{cases} \tau^m & \text{if } \tau < \widetilde{\tau}, \\ \tau - \widetilde{\tau} + \widetilde{\tau}^m & \text{if } \tau \geq \widetilde{\tau}, \end{cases} \tag{43}$$

with $\widetilde{\tau} = m^{-1/(m-1)}$. At each time step n and for each of the formulations (37)-(41) 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.

For a given time step n the initial guess is given either by the initial data (for $n = 1$) or by the approximate solution obtained at the previous time step (for $n \geq 2$). In the latter case the initial guess will eventually differ from one formulation to another. Let us recall that, in view of Proposition 2.5, the Jacobi-Newton methods (37) and (38) converge regardless of the initial guess. In contrast, the original formulation (39) requires the initial guess to be a lower solution (39). Luckily, the approximate solution obtained at the time step $n-1$ is a lower solution for the problem at time step n. This observation is the object of the following remark.

Remark 3.1 *The solution of* (33) *(under given initial and boundary conditions, as well as the source term) satisfies $\partial_t u \geq 0$. This property is reproduced at the discrete level by the approximate solution resulting from the formulations* (37), (38), *and* (39)*; more precisely, let u_ϵ^n denote the approximate solution of $F_\star^n(u) = 0$ for some $\star = u, r, l$, then*

$$u_\epsilon^n \geq u_\epsilon^{n-1}, \qquad F_u^n(u_\epsilon^n) \leq 0 \qquad \text{and} \qquad F_u^n(u_\epsilon^{n-1}) \leq 0 \tag{44}$$

for any $n \geq 1$. The last statement in (44) *means that $u_0^n = u_\epsilon^{n-1}$ constitutes an appropriate initial guess.*

Let us prove (44) *by induction. We only consider the case $\star = u$, the proof for the 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 2.4, that the sequence of Newton's iterates is 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 (44) *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 (44) *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 2.4.*

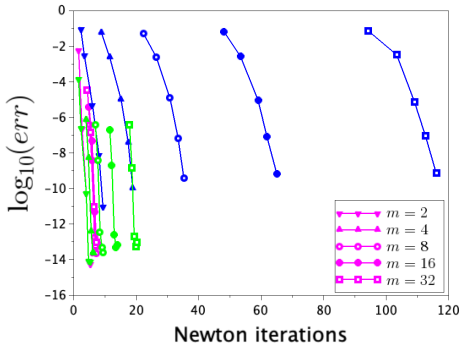
We present below the results of the numerical experiment. For a given value of m, the tolerance ϵ and a specific solution method $\star$ let us denote by $\left(u_{m,\epsilon}^{n,\star}\right)_{n\in\{1,\dots,N_T\}}$ the approximate solution of (35). The methodology of the numerical experiment is similar 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 $\left(u_{m,ref}^n\right)_{n\in\{1,\dots,N_T\}}$. Then, for each solution method (37)-(41) and for the tolerance values $\epsilon \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-6}, 10^{-8}\}$, we perform the computations measuring the total number of Newton's iteration, the required computational 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

$$err_{m,\epsilon}^\star = \frac{\|u_{m,\epsilon}^{\cdot,\star} - u_{m,ref}^{\cdot}\|_{L^\infty(0,T;L^1(0,1))}}{\|u_{m,ref}^{\cdot}\|_{L^\infty(0,T;L^1(0,1))}}. \tag{45}$$

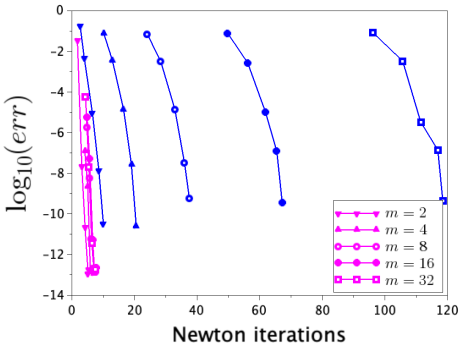
The test case is configurated as follows: in order to allow for the use of $u-$formulation we chose a positive initial condition with $\beta(u_{ini}) = 10^{-10}$, while m takes values in the set $\{2, 4, 8, 16, 32\}$. Let $V_T = 0.8$ be the total injected volume, we set $q = V_T/T$, with $T = 0.1$ for $d = 1$ and $T = 1$ for $d = 2$. This particular parameter set aims insuring that u takes values above one.

3.1 Performance in terms of Newton's iterations

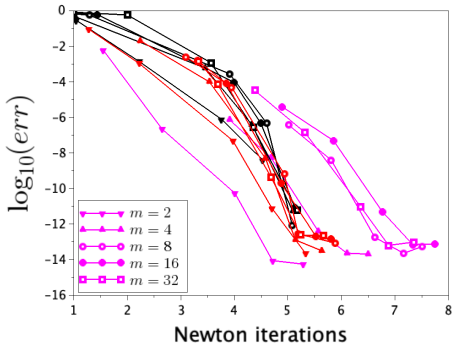
The first set of tests is performed using the fixed mesh size parameters: $N = 200$ for $d = 1$ and and $N = 40^2$ for $d = 2$. The total number of time steps is given by $N_T = 100$. We report on Figures 2 and 3 the performance of the formulations (37)-(41) for $d = 1$ and $d = 2$ respectively; more specifically, those figures exhibit the accuracy of the approximate solution measured in terms



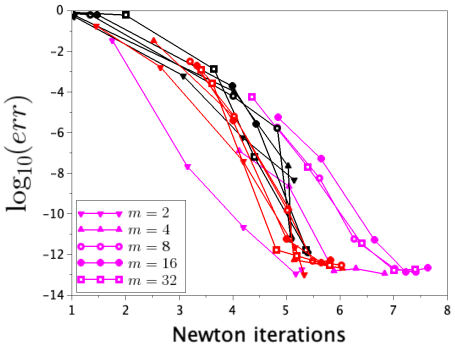
(a) $err^{u,\star}_{m,\epsilon}$ for $u-$formulation (blue), $\tau-$formulation (magenta) and $v-$formulation (green) for PME.



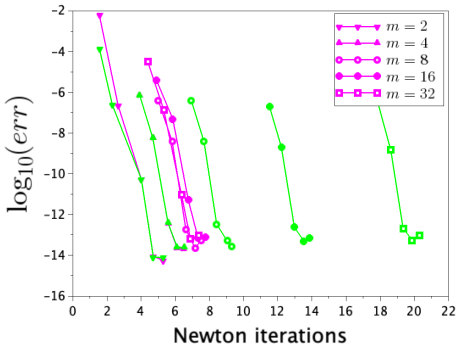
(b) $err^{u,\star}_{m,\epsilon}$ for $u-$formulation (blue) and $\tau-$formulation (magenta) for RE.



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



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



(e) $err^{\star}_{m,\epsilon}$ for $\tau-$formulation (magenta) and $v-$formulation (green) for PME.

Figure 2: Case $d = 1$: relative error $err^{\star}_{m,\epsilon}$ as the function of the average number of Newton's iterations per time step.

of $err^{u,\star}_{m,\epsilon}$ as the function of the average number of NM iterations and the value of the parameter m. The results for the porous medium equation (PME) and the Richards' equation (RE) are shown in the left and the right columns respectively.

First of all, we observe that the performance results corresponding to PME and RE problems are almost identical. In particular, we note that the performance of Newton's method is not affected by the lack of smoothness in β in RE case. Likewise the cases $d = 1$ and $d = 2$ are qualitatively similar. Compared to the case $d = 2$, we observe for $d = 1$ a higher iteration count especially for

$u-$ and $v-$formulations, which is due to the use of a finer mesh. We also wish to comment on the non-monotone convergence curves for $v-$ and $\tau-$formulations reported on Figures 2 and 3 in the range of very small errors (between 10^{-12} and 10^{-14}), which we attribute to the effects of the finite precision arithmetic.

In accordance with the results reported in [4], we observe on Figures 2(a), 2(b), 3(a) and 3(b), that the original formulation (39) based on the variable u performs rather poorly, with Newton's iteration count that increases with m. Yet, we recall that NM applied to (39) enjoys the semi-global monotone convergence. Figures 2(a) and 3(a) (see also 2(e) and 3(e)), report the results obtained with $v-$formulation (40), which is available for the PME. We observe that the simple change of the variable used by this formulation allows for drastic improvement of the performance. Nevertheless, we note that this algorithm is still not robust with respect to m. Compared to the formulations based either on v or u, $\tau-$formulation (41), that switches between those variables, turns out to be very efficient and quite robust.

Figures 2(c), 2(d), 3(c) and 3(d) report the performance of the globally convergent Jacobi-Newton methods (37) and (38) along with the results of $\tau-$formulation used as the reference. We observe that for $m = 2$ both Jacobi-Newton methods are slightly less efficient than $\tau-$formulation. This conclusion also holds for smaller values of the parameter. In contrast, Jacobi-Newton approach becomes advantageous at large values of m. Note that, in accordance with Proposition 2.6 and regardless of the value of m, the Jacobi-Newton method outperforms the original formulation (39) . We also note that the performance of both Jacobi-Newton methods seems to be virtually independent of m.

3.2 Computational overhead due to local problem solution

We have seen that for stiff problems, say with m larger than 4, the preconditioned Newton's methods require fewer iterations than the alternative change-of-variable methods. To assess the overall computational effort required by the preconditioned methods we present the analysis in terms of the CPU time. The Jacobi-Newton method studied in this article naturally results a nested iterative scheme, where each outer iteration involves two sequential steps: 1) solution of the set of scalar nonlinear equations; 2) solution of the linear system of equation. The performance of a given method, thus depends on how effectively each step is executed. In this context, we wish to provide some details on the implementation of the methods. The numerical experiment is performed using Scilab 6.1.1. programming language, with the standard "backslash" direct solver used for the linear systems. On the other hand, each iteration of the Jacobi-Newton algorithm takes to solve a set of equations

$$\beta(w) + d_{ii}w = r_i, \qquad i \in \{1, \dots, N\}, \tag{46}$$

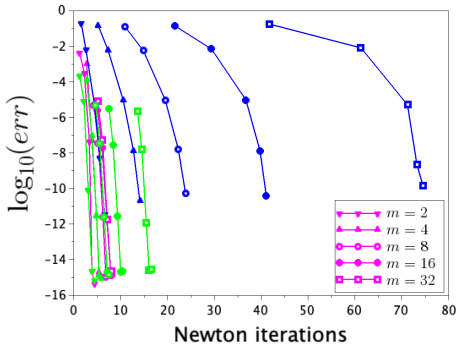
where $d_{ii} > 0$ denoting the diagonal components of the discrete Laplace operator L, and $r_i \geq 0$ is some given right-hand-side. Because the left-hand-side of (46) is concave, Newton's method applied to (46) is monotonically convergent, nevertheless, due to the singularity in β' its performance is rather poor and deteriorates as m grows. To solve (46) more efficiently we set $w = \overline{u}(\tau)$ and reformulate (46) as

$$\overline{v}(\tau) + d_{ii}\overline{u}(\tau) = r_i, \qquad i \in \{1, \dots, N\}, \tag{47}$$

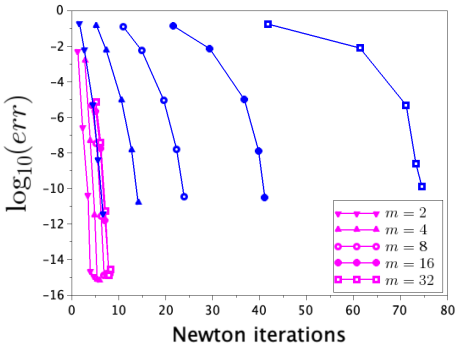
where $\overline{v}$ and $\overline{u}$ are defined by (42) and (43). Equation (47) is then solved by the standard Newton's method with respect to the variable τ. The inner NM iterations are stopped as soon as the maximum norm of the residual drops below 10^{-11}.

The CPU time assessment is performed over the bidimensional PME case with N taking values in $\{20^2, 40^2, 80^2\}$. Figure 4 reports, for different values of the mesh size parameter and $m \in \{4, 8\}$, the comparison of the Jacobi-Newton methods with the method based on $\tau-$formulation. For the given values of m and ϵ, and the system size N, we denote by $\chi^{\star}_{N,m,\epsilon}$ the computational time required by a particular method $\star$, with $\chi^{\tau}_{N,m,\epsilon}$ denoting the CPU time associated with $\tau-$formulation. We report on Figure 4 the relative deviation from the reference solution as the function of the relative CPU time $\overline{\chi}^{\star}_{N,m,\epsilon} = \chi^{\star}_{N,m,\epsilon}/\chi^{\tau}_{N,m,10^{-10}}$. This scaling aims to bring together the curves corresponding to different system sizes.

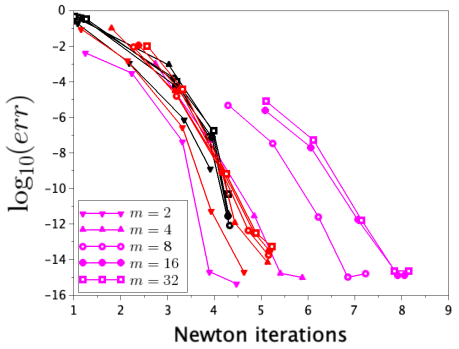
For all except the small problem ($N = 20^2$) we observe that the Jacobi-Newton method performs at least as well as $\tau-$formulation. For $m = 8$ and $N \in \{40^2, 80^2\}$ the preconditioned methods deliver a considerable speedup, which we attribute to a fewer linear problem solves.



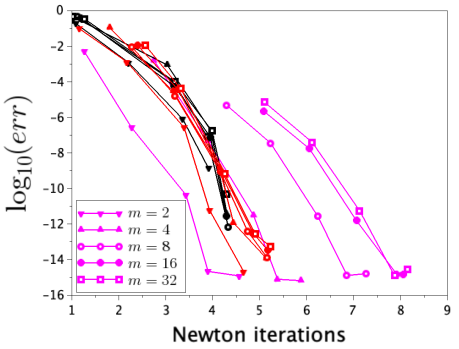
(a) $err^{\star}_{m,\epsilon}$ for $u-$formulation (blue), $\tau-$formulation (magenta) and $v-$formulation (green) for PME.



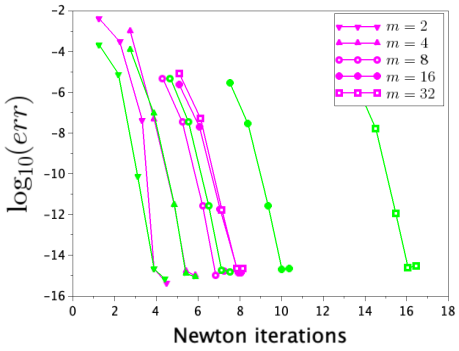
(b) $err^{\star}_{m,\epsilon}$ for $u-$formulation (blue) and $\tau-$formulation (magenta) for RE.



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



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



(e) $err^{\star}_{m,\epsilon}$ for $\tau-$formulation (magenta) and $v-$formulation (green) for PME.

Figure 3: Case $d = 2$: relative error $err^{\star}_{m,\epsilon}$ as the function of the average number of Newton's iterations per time step.

4 Conclusion

For systems involving only diagonal nonlinearities and satisfying the 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

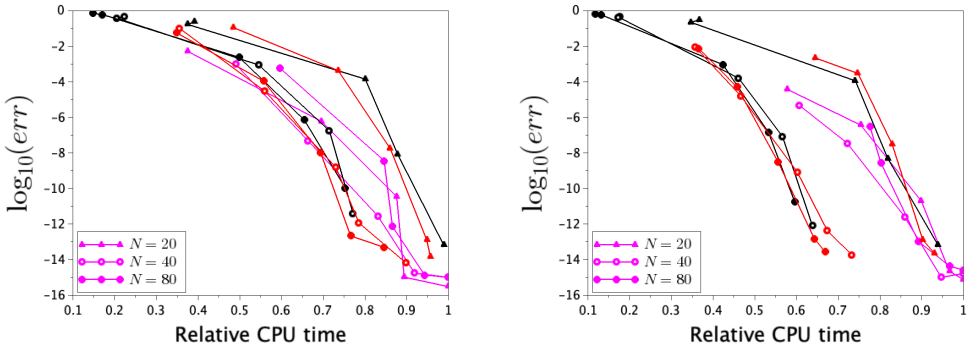


Figure 4: Relative error $err^{\star}_{m,\epsilon}$ as the function of the relative CPU time for $\tau-$formulation (magenta), left preconditioned (black) and right preconditioned (red) Newton's method. Case PME with $d = 2$, $m = 4$ on the left and $m = 8$ on the right.

on the porous medium and Riahcrds equations. 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 robust than the alternative nonmonotone methods based on the change of the variable.

References

[1] A. Baluev. On the abstract theory of chaplygin's method. In *Dokl. Akad. Nauk. SSSR*, volume 83, pages 781–784, 1952.

[2] J. Bear and A. Verruijt. *Modeling groundwater flow and pollution*, volume 2. Springer Science & Business Media, 1987.

[3] K. Brenner. Acceleration of newton's method using nonlinear jacobi preconditioning. In *International Conference on Finite Volumes for Complex Applications*, pages 395–403. Springer, 2020.

[4] K. Brenner and C. Cances. Improving newton's method performance by parametrization: the case of the richards equation. *SIAM Journal on Numerical Analysis*, 55(4):1760–1785, 2017.

[5] L. Brugnano and V. Casulli. Iterative solution of piecewise linear systems and applications to flows in porous media. *SIAM Journal on Scientific Computing*, 31(3):1858–1873, 2009.

[6] L. Brugnano and A. Sestini. Iterative solution of piecewise linear systems for the numerical solution of obstacle problems. *Journal of Numerical Analysis, Industrial and Applied Mathematics*, 6:67–82, 2011.

[7] X.-C. Cai and D. E. Keyes. Nonlinearly preconditioned inexact newton algorithms. *SIAM Journal on Scientific Computing*, 24(1):183–200, 2002.

[8] J. Carrillo Menéndez. On the uniqueness of the solution of the evolution dam problem. 1994.

[9] V. Casulli and P. Zanolli. A nested newton-type algorithm for finite volume methods solving richards' equation in mixed form. *SIAM Journal on Scientific Computing*, 32(4):2255–2273, 2010.

[10] V. Casulli and P. Zanolli. Iterative solutions of mildly nonlinear systems. *Journal of Computational and Applied Mathematics*, 236(16):3937–3947, 2012.

[11] J. E. Dennis Jr and R. B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM, 1996.

[12] 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 Journal on Scientific Computing*, 38(6):A3357–A3380, 2016.

[13] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. *Handbook of numerical analysis*, 7:713–1018, 2000.

[14] L. V. Kantorovich. On newton's method for functional equations. In *Dokl. Akad. Nauk SSSR*, volume 59, pages 1237–1240, 1948.

[15] J. M. Ortega. The newton-kantorovich theorem. *The American Mathematical Monthly*, 75(6):658–660, 1968.

[16] J. M. Ortega and W. C. Rheinboldt. Monotone iterations for nonlinear equations with application to gauss-seidel methods. *SIAM Journal on Numerical Analysis*, 4(2):171–190, 1967.

[17] J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*. SIAM, 2000.

[18] C. Pao. Accelerated monotone iterative methods for finite difference equations of reaction-diffusion. *Numerische Mathematik*, 79(2):261–281, 1998.

[19] C. Pao. Accelerated monotone iterations for numerical solutions of nonlinear elliptic boundary value problems. *Computers & Mathematics with Applications*, 46(10-11):1535–1544, 2003.

[20] F. A. Potra. Newton-like methods with monotone convergence for solving nonlinear operator equations. *Nonlinear Analysis: Theory, Methods & Applications*, 11(6):697–717, 1987.

[21] F. A. Potra and W. C. Rheinboldt. On the monotone convergence of newton's method. *Computing*, 36(1):81–90, 1986.

[22] W. C. Rheinboldt. On m-functions and their application to nonlinear gauss-seidel iterations and to network flows. *Journal of Mathematical Analysis and Applications*, 32(2):274–307, 1970.

[23] C. Van Duyn and L. Peletier. Nonstationary filtration in partially saturated porous media. *Archive for Rational Mechanics and Analysis*, 78(2):173–198, 1982.

[24] J. L. Vázquez. *The porous medium equation: mathematical theory*. Oxford University Press on Demand, 2007.