Acceleration of Newton's method using nonlinear Jacobi preconditioning

Konstantin Brenner

Acceleration of Newton's method using nonlinear Jacobi preconditioning

Konstantin Brenner

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.

Key words: Mildly nonlinear systems, Newton's 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 $\mathbb{M}(N)$, $b \in (\mathbb{R}^+)^N$ and the mapping f is defined by

$$f : u \mapsto (f_1(u_1), \dots f_N(u_N))^T$$

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

Université Côte d'Azur, Inria Team Coffee, CNRS, Laboratoire J.A. Dieudonné, Parc Valrose, 06108 Nice cedex 02, France, e-mail: konstantin.brenner@univ-cotedazur.fr

(A_1) For $1 \leq i \leq N$, f_i is strictly increasing, concave and belongs to C^1 on $(0,+\infty)$.
(A_2) For any $u > 0$ the matrix $f'(u)+A$ is an M-matrix in the sense of the definition below.
(A_3) The matrix A has zero diagonal elements.

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 $i \neq j$.

We remark that the derivatives of f_i are potentially unbounded at the origin.

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

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

where $\mathbf{v}$ is some given velocity field. 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 [3]). Let us mention that the monotone discretizations are not only beneficial to the nonlinear solver (as it is going to be discussed in this paper), but also allow to preserve the local maximum principle on the discrete level, thus avoiding any spurious oscillations of the discrete solution. Let D denote the diagonal of $M^{-1}S$ and let $A = \Delta t \left(M^{-1}S - D\right)$. Setting

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

the system (3) can be written as (1).

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 [5], [2] 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 Heaviside graph. In [2] 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 used to solve the systems (1) numerically; moreover under assumptions (A_1) and (A_2) 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 concavity of the underlying functional and the non-negativity of the inverse of it's Jacobian; it is in fact a straightforward adaptation of the convergence results from [4] (see also Proposition 1 below) to the concave setting.

Despite an available convergence result, the numerical evidences presented in [2] 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]). It is demonstrated in the numerical section 3 that the convergence of Newton's method is slow; moreover the number of Newton's iterations required to solve the system grows 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 [2]. 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 monotone 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 is 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

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

Under the assumption (A_1) the function g is increasing and convex, and therefore $F_\star(u)$, $\star = l, r$ remains concave; moreover the derivative of g is finite for all $u \in (\mathbb{R}^+)^N$ and it can be shown 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 convergence result from [4], we prove monotone convergence of Newton's method applied to the problem (1) in its original formulation and applied to the preconditioned problems (4) and (5). Section 3 is deduced to the numerical experiment.

2 Main results

Let us first present the adaptation of the convergence result 13.3.4 from [4] to the case of concave mappings.

Theorem 1 (Convergence of Newton's method)

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 $u \in (\mathbb{R}^+)^N$. 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 $u_\star$.

Let us denote $F_u(u) = f(u) + Au - b$, based on the assumptions (A_1) and (A_2), it can be shown that the solution of (1) exists and is unique; in addition under the same assumptions it follows from Theorem 1 that Newton's method applied to (1) converges monotonically provided that $u_\star > 0$ and $F_u(u_0) \leq 0$. More precisely the following proposition holds.

Proposition (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 $u_\star$. □

We remark that if $f'(0) = +\infty$ the assumption $b > 0$ can not be avoided, therefore Newton's method can not be applied to $F_u(u) = 0$ unless the solution is strictly positive. In contrast the preconditioned methods can be applied without any additional restrictions on f' or on the sign of the solution. Convergence of the preconditioned methods is summarized by the following proposition, which relies on the assumption (A_3) ensuring the concavity of $F_\star$ and the M-matrix property of $F'_\star$, $\star = l, r$.

Proposition (Convergence of the preconditioned methods) The mappings F_l and F_r satisfy the assumptions of Theorem 1 with $u_0 = 0$; moreover for all $u \in (\mathbb{R}^+)^N$ the matrix $F'_\star(u)$, $\star = l, r$ is such that $F'_\star(u) \leq I \leq F'_\star(u)^{-1}$. □

3 Numerical experiment

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

$$\partial_t \beta(u) - \partial^2_{xx} u = 0 \tag{7}$$

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

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

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". For $m = 10$, $q = 10^4$, $T = 1.2\ 10^{-2}$ and $N_T = 100$ the approximate profile of $\beta(u)$ at different time steps is exhibited at the right side of Figure 2.

Equation (7) is discretized using the standard implicit in time finite volume method. Let the positive integers N and N_T denote the number of cells and the total number of time steps, let $h = \dfrac{1}{N}$ be the cell size and $\Delta t = \frac{T}{N_T}$ be the size of the time step. For all cells i and time steps n the discretized version of (7) reads

$$\beta(u_i^n) + \frac{\Delta t}{h^2} \sum_{j \in \mathcal{N}_i} (u_i^n - u_j^n) = \beta(u_i^{n-1}) + \frac{\Delta t}{h} q\ \delta_{i,1}, \tag{8}$$

where $\delta_{i,1}$ stands for the Kronecker symbol and where $\mathcal{N}_i$ denotes the set of the neighbors of the cell i. Let L denote the tridiagonal matrix associated to the discretization of the diffusion operator and D be it's diagonal. We denote by b_n the right-hand-side of (8). The system (8) results in the following problem, which has to be solved for each time step

$$(\beta(u) + Du) + (L - D)\, u = b_n. \tag{9}$$

It is easy to show that $f(u) = \beta(u) + Du$ and $A = L - D$ 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

$$F_l^n(u) := u - g\,(b_n - Au) = 0 \tag{10}$$

and

$$F_r^n(u) := u + Ag(u) - b_n = 0. \tag{11}$$

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

$u-$**formulation:** NM applied to (9) in the original form

$$F_u^n(u) := \beta(u) + Lu - b_n = 0 \tag{12}$$

In view of Proposition 1 this method is monotonically convergent provided that the initial guess satisfy $F(u_0) \leq 0$.

$v-$**formulation:** The problem (9) is reformulated with respect to the variable v with $u = \beta^{-1}(v)$ and NM is applied to

$$F_v^n(v) := v + L\beta^{-1}(v) - b_n = 0 \tag{13}$$

$\tau-$**formulation:** Following [2] we introduce the function pair $\tau \to (\overline{u}(\tau), \overline{v}(\tau))$ such that for all τ it holds $\overline{v}(\tau) = \beta(\overline{u}(\tau))$ and $\max(\overline{u}'(\tau), \overline{v}'(\tau)) = 1$. Then NM is applied to

$$F_\tau^n(\tau) := \overline{v}(\tau) + L\overline{u}(\tau) - b_n = 0. \tag{14}$$

At each time step n and for each of the formulations (10)-(14) the sequence of the approximate solutions $\left(\xi_k^n\right)_k$ (where ξ denotes an appropriate primary variable) is computed using Newton's method until the stopping criterion $\|F_\star^n(\xi_k^n)\|_\infty < \varepsilon$ is satisfied for some small predefined tolerance ε. As the initial guess we use the value of the variable obtained at the previous time step (this value will differ between the formulations). This choice of the initial guess is motivated by the following observation.

Remark 1 Under the given initial and boundary conditions the solution of (7) 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 u^n denote an 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 the appropriate choice of the initial guess.

In the following 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 $\left(u_{m,\varepsilon}^{n,\star}\right)_{n\in\{1,\ldots,N_T\}} \in \mathbb{R}^N$ the approximate solution of (9).

The methodology of the study is similar to [2], that is for each value of m we compute, using $\tau-$formulation and the tolerance $\varepsilon_{ref} = 10^{-10}$, the reference solution denoted by $\left(u_{m,ref}^n\right)_{n\in\{1,\ldots,N_T\}}$. Then, for each solution method (10)-(14) 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 relative deviation from the reference solution measured in terms of the conservative variable $\beta(u_{m,\varepsilon}^\star)$

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

Performance comparison. The first set of tests is performed using the fixed mesh size parameter $N = 100$. In accordance with the results reported in [2], Figure 1 witness the qualitative differences in the performance of u, v and τ-formulations. Com-

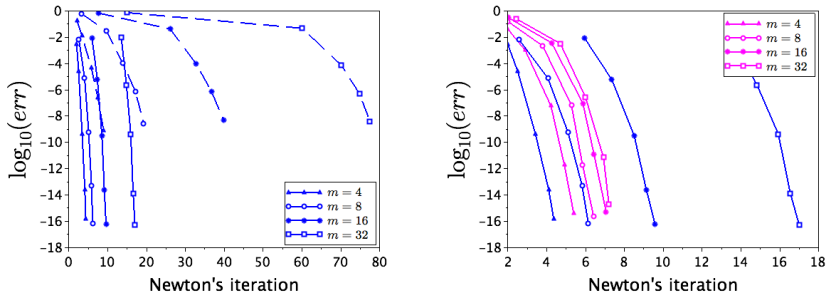

Fig. 1 Relative error $err^{\star}_{m,\varepsilon}$ as the function of the average number of Newton's iterations per time step. Left: for v-formulation (solid blue) and u-formulation (dashed blue). Right: for v-formulation (blue) and τ-formulation (magenta).

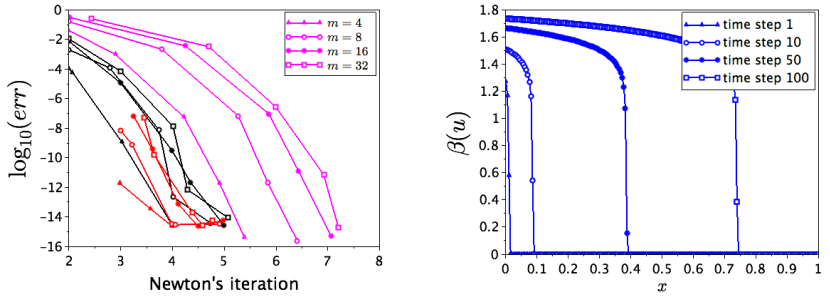

Fig. 2 Left: relative error $err^{\star}_{m,\varepsilon}$ as the function of the average number of Newton's iterations per time step for τ-formulation (magenta), left-preconditioned (black) and right-preconditioned (red) Newton's method (magenta). Right: Approximate solution at different time steps.

pared to the original u-formulation, the formulation using v as the primary variable is few time faster, it also performs slightly better then τ-formulation for the moderate values of m. However, in contrast with τ-formulation, none of the formulations u or v is robust with respect to the variation of m. Finally, Figure 2 shows a relatively similar behavior of τ-formulation and the preconditioned methods, with the latter ones requiring a slightly fewer number of iterations.

Computational overhead due to local problem solution. It can be observed on Figures 1 and 2 that preconditioned Newton's methods require less iterations then the other formulations. However, each iteration of the preconditioned method requires to evaluate the function g, and therefore to solve the set of the scalar nonlinear equations. Those computations, performed again using Newton's method, result in a certain computational overhead which has to be accounted for. To access the overall computational effort required by the preconditioned methods we present the analysis in terms of the CPU time. Figure 3 shows, for different values of the mesh size parameter $N \in \{200, 400, 800, 1200\}$, the comparison of the left (respectively

right) preconditioned NM with the method based on τ-formulation. In can be observed that for the 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.

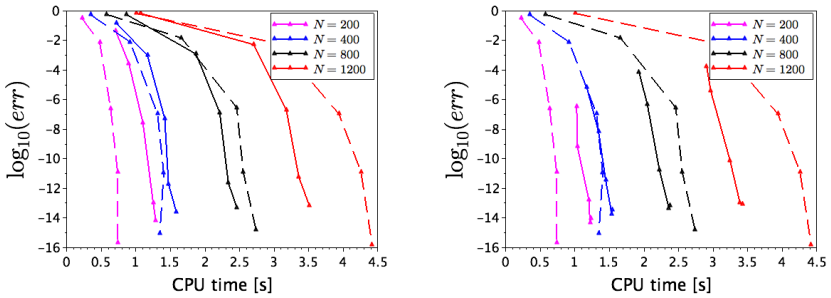

Fig. 3 Relative error $err^{\nu,\star}_{m,\varepsilon}$ as the function of CPU time for different grid sizes. Left: left-preconditioned NM (solid lines) and τ-formulation (dashed lines). Right: right-preconditioned NM (solid lines) and τ-formulation (dashed lines).

References

1. Bear, J., Verruijt, A.: Modeling groundwater flow and pollution. Reidel (1987)
2. Brenner, K., Cancès, C.: Improving newton's method performance by parametrization: the case of richards equation. SIAM Journal on Numerical Analysis (2017)
3. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. Ciarlet, P. G. (ed.) et al., in Handbook of numerical analysis. North-Holland, Amsterdam, pp. 713–1020 (2000)
4. Ortega, J.M., Rheinboldt, W.C.: Iterative Solutions of Nonlinear Equations in Several Variables. Academic Press (1970)
5. Van Duijn, C.J., Peletier, L.A.: Nonstationary filtration in partially saturated porous media. Archive for Rational Mechanics and Analysis **78**(2), 173–198 (1982)
6. Vázquez, J.L.: The Porous Medium Equation - Mathematical theory. The Clarendon Press Oxford University Press (2007)