On global and monotone convergence of the preconditioned Newton's method for some mildly nonlinear systems

Konstantin Brenner

On global and monotone convergence of the preconditioned Newton's method for some mildly nonlinear systems

Konstantin Brenner

1 Introduction

Let β be a diagonal mapping from $\mathbb{R}^N$ to itself and let A be an $N \times N$ real matrix. Given some $r \in \mathbb{R}^N$ we are interested in solving numerically the flowing system of nonlinear equations

$$\beta(u) + Au = r. \tag{1}$$

More precisely, this contribution is concerned with the global convergence analysis of the preconditioned Newton's methods applied to (1).

Nonlinear preconditioning is an increasingly popular technique that may drastically improve the robustness and convergence rate of the linearization schemes such as Newton method. As in the case of linear problems, the nonlinear preconditioning consists in replacing the original system by an equivalent one that can be solved more efficiently. Since more than twenty years the variety of such methods has been proposed including Schwarz-inspired methods ASPIN [4] , MSPIN [13], [9] and RASPEN [6], as well as the nonlinear versions of FETI-DP [10] and BDDC [11].

The nonlinear preconditioning appears to be particularly efficient in the applications to the models of subsurface flow and reactive transport [9], [12], where the failures and lack robustness of the nonlinear solvers is one the major factors limiting the reliability of the simulation codes. In those applications the major benefit seems to result in the form of an extended convergence region, which, in case of time-dependent problems, allows for larger time steps [12].

The present work aims to contribute to the theoretical analysis of the nonlinear preconditioning methods, which remains relatively unexplored. We extend the previous work [2], concerned with the Jacobi-Newton method, and cover the nonlinear counterparts of some popular linear preconditioners based on multi-splitting of the system. Our analysis includes, in particular, the nonlinear preconditioning by block Jacobi or RAS methods. We prove that, under appropriate assumptions discussed

Konstantin Brenner
Université Côte d'Azur, LJAD, CNRS, INRIA, e-mail: konstantin.brenner@univ-cotedazur.fr

below, the one-level RAS-Newton method (or RASPEN [6]) applied to (1) exhibits global and essentially monotone convergence. The analysis of this method is carried out in the framework on nonlinear multi-splitting methods [7], [8], and extends to other methods such as, for example, block Gauss-Seidel.

As an alternative to the nonlinear preconditioning we study a simpler two-step scheme alternating the nonlinear multi-splitting and the standard Newton linearization steps. The two-step multi-splitting/Newton scheme enjoys the same global and monotone convergence properties as the full preconditioned method. We note that in the context of the RAS approach, such scheme, under the name of NKS-RAS method, has been proposed in [5]. It turns out that for single splitting methods, like (block) Jacobi or Gauss-Seidel, the preconditioned Newton's method is equivalent to the former two-step approach.

Our convergence analysis relies Monotone Newton Theorem [1], [15], and requires to major assumptions on the system (1), namely the concavity of the nonlinear map involved in (1) and the assumption that the Jacobian of the system has a nonnegative inverse. More specifically, we will assume the following

(A_1) For each $0 \le i \le N$, the functions $\beta_i \in C^1(\mathbb{R})$ are monotone and concave.
(A_2) For any $u \in \mathbb{R}^N$ the matrix $\beta'(u) + A$ is an M-matrix.

We wish to stress out that the assumptions (A_1) and (A_2) are quite sub-optimal and aim to improve reader's experience at the expense of sharpness. For example, the generalizations of (A_1) can be performed along the following lines. First, one can relax the regularity assumption, clearly piecewise regular functions β_i would do. Secondly, the derivative of β_i need not to be bounded, or alternatively β_i need not to be defined over a whole $\mathbb{R}$, such case has been treated in [2]. As a matter of fact, we believe that the analysis presented here can be extended to β_i being merely maximal monotone and concave (in some appropriate sense). Similarly, the assumption (A_2) can be relaxed by allowing positive off-diagonal elements in the Jacobian, assuming, for example, that A is nonsingular and $A^{-1} \ge 0$.

Before moving any further let us recall some basic properties of the system (1).

Proposition 1 (Existence and uniqueness of solution)

Let $F(u) = \beta(u) + Au$, under the assumptions (A_1) and (A_2) the mapping F^{-1} is well defined on $\mathbb{R}^N$ and is convex.

Next, we state the global version of Monotone Newton Theorem for which we refer to [15].

Theorem 1 (Global Monotone Newton Theorem)

Let $F : \mathbb{R}^N \to \mathbb{R}^N$ be continuous, Gâteaux differentiable and concave. Suppose that $F'(u)$ has a nonnegative inverse for all $u \in \mathbb{R}^N$, and assume that $F(u) = 0$ has a solution.Then, for any $u_0 \in \mathbb{R}^N$, the sequence

$$u_{n+1} = u_n - F'(u_n)^{-1} F(u_n), \qquad n \ge 0 \tag{2}$$

satisfies $u_n \le u_{n+1} \le F^{-1}(0)$ and $F(u_n) \le 0$ for all $n \ge 1$. If, in addition, there exists an invertible $P \in \mathbb{M}(N)$ such that $F'(u)^{-1} \ge P \ge 0$ for all $u \in \mathbb{R}^N$, then the sequence u_n converges to $F^{-1}(0)$.

In view of Theorem 1 and Proposition 1 one can deduce that Newton's method applied to the system (1) converges regardless of the initial guess. Unfortunately, depending on the stiffness of the function β this convergence may become arbitrarily slow as pointed out in [3] and [2]. While the lack of robustness with respect to the shape of β can be addressed by the diagonal Jacobi preconditioning [2], for systems resulting from the discretization of the degenerate PDEs the efficiency of the Jacobi-Newton method is still controlled by the mesh size, which motivates the use of the nonlinear preconditioning of the domain decomposition type.

Having in mind the application to the methods involving the RAS approach, we introduce in Section 2 the preconditioning technique based on the nonlinear multi-splitting of (1). We prove that the preconditioned system satisfies Theorem 1 and, therefore, Newton's method is unconditionally convergent. In Section 3 we present the numerical results based on a discretized porous media equation [16] and using some variants of nonlinear RAS method including RASPEN and RAS/Newton two-step methods.

2 Nonlinear multi-splitting method

In this section we present the nonlinear preconditioning procedure inspired by the linear multi-splitting methods [14], [7]. We note that, following [8], the one-level RASPEN method applied to (1) can be expressed in terms of a certain multi-splitting.

Let $(P_i, Q_i)_{i=1,\dots,K}$ be a finite of family matrices such that

$$A = P_i - Q_i.$$

Denoting $M_i(u) = \beta(u) + P_i u$, and assuming that $M_i(u)$ admits an inverse defined on $\mathbb{R}^N$ we can reformulate the original problem as

$$\mathcal{F}_i(u) := u - M_i^{-1}(N_i(u)) = 0 \tag{3}$$

where $N_i(u) = Q_i u + r$. Let $(E_i)_{i=1,\dots,K}$ be a family of nonnegative diagonal matrices such that $\sum_{i=1}^{K} E_i = I$. Multiplying (3) by E_i and summing over i we obtain the system

$$\mathcal{F}(u) := \sum_{i=1}^{K} E_i \mathcal{F}_i(u) = u - \sum_{i=1}^{K} E_i M_i^{-1}(N_i(u)) = 0. \tag{4}$$

Clearly, the solution of the original system satisfies (4). The proposition below states that $\mathcal{F}$ is concave and that $\mathcal{F}'(u)$ has the nonnegative inverse for all u, which implies, in particular, that $\mathcal{F}$ is inverse isotone, and, therefore, the solution to (4) is unique.

Proposition 2 *Assume that the for all $u \in \mathbb{R}^N$ and all i the splitting*

$$F'(u) = M_i'(u) - Q_i$$

is weakly regular and that $M_i'(u)$ is an M-matrix. Then, the mapping $\mathcal{F}(u)$ from (4) *is a concave bijection from $\mathbb{R}^N$ to $\mathbb{R}^N$, and, for all $u \in \mathbb{R}^N$, the matrix $\mathcal{F}'(u)$ is an M-matrix satisfying $\mathcal{F}'(u)^{-1} \geq I$.*

Proof: In view of Proposition 1 the mappings M_i^{-1} is well defined on $\mathbb{R}^N$ and are convex, which implies that $\mathcal{F}$ is concave since $E_i \geq 0$. Let us show that $\mathcal{F}'(u)^{-1} \geq I$ for all $u \in \mathbb{R}^N$. As the matter of fact we will show that the Jacobian is also an M-matrix. We beguine with the the following spectral bound which is the founding stone for the analysis of the multi-splitting methods (see [14])

$$\rho\left(\sum E_i M_i'(u)^{-1} Q_i\right) < 1. \tag{5}$$

Let $\widetilde{u}_i = M_i^{-1}(N_i(u))$, we have

$$\mathcal{F}'(u) = I - \sum_i E_i M_i'(\widetilde{u}_i)^{-1} Q_i.$$

Let $w \in \mathbb{R}^N$ be a component-wise max of the vectors $\widetilde{u}_i$, that is $(w)_k = \max_i (\widetilde{u}_i)_k$. Since M' is antitone and has nonnegative inverse we deduce that $M_i'(\widetilde{u}_i)^{-1} M_i'(w) \leq I$, and, since $M_i'(\widetilde{u}_i)^{-1} Q_i$ and $M_i'(w)^{-1} Q_i$ are nonnegative we obtain

$$M_i'(\widetilde{u}_i)^{-1} Q_i = M_i'(w)^{-1} Q_i + \left(M_i'(\widetilde{u}_i)^{-1} M_i'(w) - I\right) M_i'(w)^{-1} Q_i \leq M_i'(w)^{-1} Q_i.$$

and

$$0 \leq \sum_i E_i M_i'(\widetilde{u}_i)^{-1} Q_i \leq \sum_i E_i M_i'(w)^{-1} Q_i$$

It follows from (5) that

$$\rho(\sum_i E_i M_i'(\widetilde{u}_i)^{-1} Q_i) \leq \rho\left(\sum_i E_i M_i'(w)^{-1} Q_i\right) < 1,$$

which implies in turn that $\mathcal{F}'(u)^{-1} \geq 0$. Clearly the off-diagonal part of $\mathcal{F}'(u)$ is nonpositive, implying that it is M-matrix; moreover, since, $\mathcal{F}'(u) \leq I$, we deduce that $\mathcal{F}'(u)^{-1} \geq I$. □

Based on Proposition 2 one shows that the mapping $\mathcal{F}$ satisfies the assumptions of Theorem 1. In addition we consider the following multi-splitting/Newton two-step scheme: Given $u_0 \in \mathbb{R}^N$ compute for all $n \geq 0$

$$\widetilde{u}_n = \sum_i E_i M_i^{-1}(N_i(u_n)) \tag{6}$$

and

$$u_{n+1} = \widetilde{u}_n - F'(\widetilde{u}_n)^{-1} F(\widetilde{u}_n). \tag{7}$$

We note that (6) can be interpreted as a step of a quasi-Newton method applied to (4), where the matrix $\mathcal{F}(u)^{-1}$ has been replaced by it's subinverse I. It can be shown

that the (6)-(7) leads again to a globally convergent scheme. Remarkably enough, in the case of a simple splitting the two-step scheme is equivalent to the preconditioned Newton's method.

Proposition 3 *Let $A = P - Q$ be some splitting such that the inverse of $M = \beta(u) + Pu$ is well defined and $M'(u)$ is non-singular for all $u \in \mathbb{R}^N$. Then, the two-step scheme is equivalent to the preconditioned Newton's method.*

Proof: Let $\widetilde{u} = M^{-1}(N(u))$, we remark that $\mathcal{F}(u) = u - \widetilde{u}$ and

$$\mathcal{F}'(u) = I - M'(\widetilde{u})^{-1} = M'(\widetilde{u})^{-1} F'(\widetilde{u}).$$

Therefore, the update generated by the preconditioned Newton's method starting from u is given by

$$\delta_{MS-N}(u) = M'(\widetilde{u})^{-1} F'(\widetilde{u}) \, (\widetilde{u} - u) \, . \tag{8}$$

Now, let us consider the update generated by the two-step method (6)-(7), we have

$$\delta_{MS/N}(u) = \widetilde{u} - u - F'(\widetilde{u})^{-1} F(\widetilde{u}).$$

We remark that

$$F(\widetilde{u}) = M(\widetilde{u}) - N(\widetilde{u}) = N(u) - N(\widetilde{u}),$$

and using linearity of N, we deduce that $F(\widetilde{u}) = Q(u - \widetilde{u})$. Therefore,

$$\delta u_{MS/N} = \left(I + F'(\widetilde{u})^{-1} Q \right) (\widetilde{u} - u) = F'(\widetilde{u})^{-1} M'(\widetilde{u})^{-1} (\widetilde{u} - u),$$

which provides $\delta_{MS/N}(u) = \delta_{MS-N}(u)$. □

3 Numerical experiment

We now proceed with the numerical experiment that illustrates the performance of block Jacobi-Newton, RASPEN and RAS/Newton methods applied to the system resulting from the discretization of a degenerate parabolic equation. In particular, the numerical experiment shows that, in contrast to the Jacobi-Newton method, the performance of the former methods is virtually independent of the mesh size.

The test case considered here is similar to the one presented in [2] to which we refer for more detailed discussion. In brief, we are interested in the algebraic system resulting from the implicit in time discretization of the porous media equation [16]. More specifically, focusing on a single step (of length τ) of the backward Euler time integration scheme we consider the system of the form (1) resulting from the finite difference discretization of the following boundary value problem

$$\begin{cases} \beta(u) - \beta(u_{ini}) = \tau \partial^2_{xx} u & x \in (0,1), \\ \partial_x u(0) = -q, \quad \partial_x u(1) = 0, \end{cases} \tag{9}$$

where $\beta(u) = u^{1/m}$ with $m > 1$. We consider the following set of parameters: $m = 10$, $q = 1$, $\tau = 0.5$, and $\beta(u_{ini}) = 10^{-6}$. The problem (9) is discretized using $N = 100$ or 400 degrees of freedom, and the vector u_{ini} is used as the initial guess by the solution methods considered below.

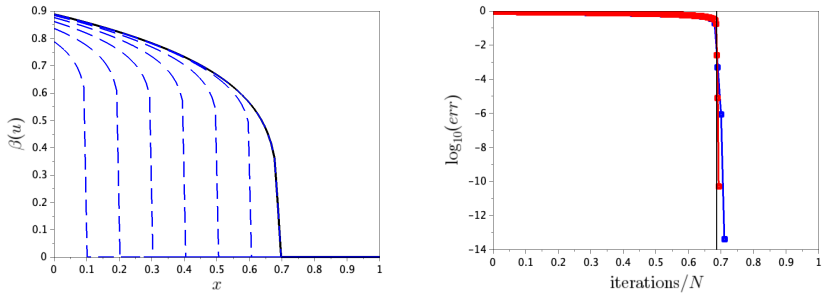

Fig. 1: Left: the solution for $N = 100$ (black) and the iterates u_n of the Jacobi-Newton method for $n = 10, 20, \ldots 70$. Right: convergence history of the Jacobi-Newton method for $N = 100$ (blue) and 400 (red), the iteration count is scaled by the size of the discrete system; the error is measured in l_∞ norm. The vertical black line indicate the location of the solution front.

The solutions of the porous media equation are characterized by the finite speed of propagation of the support. Qualitatively this behavior persists even for strictly positive but small initial data. For the discrete counterpart of the elliptic problem (9) the latter property is reflected on the solver's level. Typically, and unless some Schwarz-type preconditioning is performed, solution fronts resulting from Newton's method can cross at most one degree of freedom at the time. For the Jacobi-Newton method this behavior is illustrated by Figure 1. The left sub-figure exhibits the final position of the solution front and some iterates of the method. The right sub-figure reports the convergence history of the method for two values of the mesh size.

The numerical performance is characterized by two very distinct regimes: a very fast near-solution convergence is preceded by the long period of a slow error decrease. As the matter of fact, the length of the convergence plateau is proportional to the number of the degrees of freedom N_f that has to be crossed by the solution front, and can be expressed as σN_f, where σ is the cost of propagating the front trough one degree of freedom. As shown in [3] and [2], the parameter σ of the standard Newton's method can become arbitrarily large depending on the coefficient m and the initial data In contrast, the Jacobi-Newton method [2] appears to be virtually independent of m and can handle a general nonnegative initial data. Nevertheless, the efficiency of the latter method is still dependent on N_f and thus on the discretization. More precisely, the right side of Figure 1 reflects the convergence of the Jacobi-Newton for $N = 100$ and 400 degrees of freedom. By scaling the iteration count by N we observe that the performance of the method is essentially controlled by N_f. The

vertical black line positioned at N_f/N reflects the final location of the front. Two convergence curves are almost identical, while the total iteration count measured for the Jacobi-Newton method is of 71 for $N = 100$ and 271 for $N = 400$.

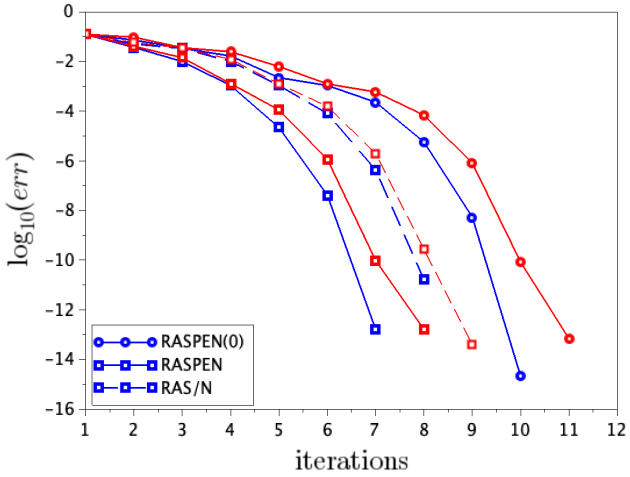

Fig. 2: Convergence history of preconditioned Newton's method for $N = 100$ (blue) and 400 (red); the error is measured in l_∞ norm.

The dependency of the mesh size can be removed by means of the nonlinear domain decomposition. We report on Figure 2 the convergence history of RASPEN and RAS/Newton (RAS/N) methods for 5 equally sized sub-domain with the relative overlap of 0.1. In addition we consider the case of the minimal algebraic overlap, denoted RASPEN(0), and corresponding to the preconditioning based on the block Jacobi method. In this numerical experiment none of the considered methods appears to exhibit any substantial dependency on the mesh size. Unsurprisingly, the overlap seems to be beneficial for the convergence of both RASPEN and RAS/N. While being slightly less efficient than RASPEN, the two-step RAS/N method still appears as a competitive alternative. The convergence of the nonlinear RAS method, applied as a solver instead of being used as a preconditioner, is not reported here, but roughly speaking the nonlinear RAS method is as inefficient as the linear one.

4 Conclusion

We have analyzed a family of preconditioned Newton's methods based on the nonlinear multi-splitting approach in application to mildly nonlinear systems resulting

from the discretization of some degenerate evolutionary PDEs such as porous media or Richards' equation. Based on the monotone Newton theorem we show that the preconditioned method is globally convergent. The current result extends our previous analysis [2] to the one-level RASPEN method [6]. In addition, for the preconditioning based on a single nonlinear splitting, including the method presented in [2] the preconditioned Newton's method is equivalent to a simpler to implement predictor-corrector scheme. The numerical experiment based on discrete porous media equation shows that the performance of block Jacobi-Newton, RASPEN and RAS/Newton methods is essentially independent of the mesh size.

References

1. Baluev, A. N. (1952). On the abstract theory of Chaplygin's method. In Dokl. Akad. Nauk. SSSR (Vol. 83, pp. 781-784).c
2. Brenner, K. (2023). On the monotone convergence of Jacobi–Newton method for mildly nonlinear systems. Journal of Computational and Applied Mathematics, 419, 114719.
3. Brenner, K., & Cances, C. (2017). Improving Newton's method performance by parametrization: the case of the Richards equation. SIAM Journal on Numerical Analysis, 55(4), 1760-1785.
4. Cai, X. C., & Keyes, D. E. (2002). Nonlinearly preconditioned inexact Newton algorithms. SIAM Journal on Scientific Computing, 24(1), 183-200.
5. Cai, X. C., & Li, X. (2011). Inexact Newton methods with restricted additive Schwarz based nonlinear elimination for problems with high local nonlinearity. Siam journal on scientific computing, 33(2), 746-762.
6. Dolean, V., Gander, M. J., Kheriji, W., Kwok, F., & Masson, R. (2016). Nonlinear preconditioning: How to use a nonlinear Schwarz method to precondition Newton's method. SIAM Journal on Scientific Computing, 38(6), A3357-A3380.
7. Frommer, A. (1989). Parallel nonlinear multisplitting methods. Numerische Mathematik, 56(2), 269-282.
8. Frommer, A., & Szyld, D. B. (2001). An algebraic convergence theory for restricted additive Schwarz methods using weighted max norms. SIAM journal on numerical analysis, 39(2), 463-479.
9. Kern, M., Taakili, A., & Zarrouk, M. M. (2020). Preconditioned iterative method for reactive transport with sorption in porous media. Mathematical Modelling and Analysis, 25(4), 546-568.
10. Klawonn, A., Lanser, M., & Rheinbach, O. (2014). Nonlinear feti-dp and bddc methods. SIAM Journal on Scientific Computing, 36(2), A737-A765.
11. Klawonn, A., Lanser, M., & Rheinbach, O. (2018). Nonlinear BDDC methods with approximate solvers.
12. Klemetsdal, Ø., Moncorgé, A., Møyner, O., & Lie, K. A. (2022). A numerical study of the additive Schwarz preconditioned exact Newton method (ASPEN) as a nonlinear preconditioner for immiscible and compositional porous media flow. Computational Geosciences, 26(4), 1045-1063.
13. L. Liu and D. E. Keyes. Field-split preconditioned inexact Newton algorithms, SIAM J. Sci. Comput. 37 (2015), pp. A1388–A1409.
14. O'Leary, D. P., & White, R. E. (1985). Multi-splittings of matrices and parallel solution of linear systems. SIAM Journal on algebraic discrete methods, 6(4), 630-640.
15. Ortega, J. M., & Rheinboldt, W. C. (2000). Iterative solution of nonlinear equations in several variables. Society for Industrial and Applied Mathematics.
16. Vázquez, J. L. (2007). The porous medium equation: mathematical theory. Oxford University Press on Demand.