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

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

Such mildly nonlinear systems with only a diagonal nonlinearity are commonly found in the context of geophysical flow and transport modeling, where they result from the discretization of nonlinear evolutionary PDEs. They arise, for example, from Richards' and porous medium equations, or, alternatively, from the models of reactive transport involving equilibrium adsorption.

This contribution is concerned with the global convergence analysis of preconditioned Newton's methods applied to (1). Nonlinear preconditioning is an increasingly popular technique that may drastically improve the robustness and convergence rate of linearization schemes such as Newton's 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 variety of nonlinear preconditioning methods has been proposed, including Schwarz-inspired methods ASPIN [4], MSPIN [9], [13] and RASPEN [6], as well as the nonlinear versions of FETI-DP [10] and BDDC [11].

Nonlinear preconditioning appears to be particularly efficient in the application to models of subsurface flow and reactive transport [9], [12], where the failures and lack of robustness of 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].

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

The present work aims to contribute to the theoretical analysis of 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, nonlinear preconditioning by block Jacobi or RAS methods. We prove that, under appropriate assumptions discussed 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 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 has been proposed in [5] under the name of NKS-RAS method. It turns out that for simple 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 on the Monotone Newton Theorem [1, 14]; and requires two 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 $0 \leq i \leq N$, 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 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 be defined over $\mathbb{R}$, 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). We also note that the analysis presented below applies to β convex instead on concave. The explicit concavity assumption is motivated the the applications to porous media flow models that we have in mind. 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 $A^{-1} \geq 0$.

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 the Monotone Newton Theorem, for which we refer to [14].

Theorem 1 (Global monotone Newton theorem)

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

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

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

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 diagonal Jacobi preconditioning [2], the efficiency of the Jacobi-Newton method for systems resulting from the discretization of degenerate PDEs is still controlled by the mesh size, which motivates the use of nonlinear preconditioning of domain decomposition type.

Having in mind the application to the overlapping domain decomposition, 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 (NKS-RAS method from [5]).

2 Nonlinear multi-splitting method

In this section we present the nonlinear preconditioning procedure inspired by the linear multi-splitting methods [7], [15].

Let $(P_i, Q_i)_{i=1,\dots,K}$ be a finite of family matrices such that $A = P_i - Q_i$. We denote $M_i(u) = \beta(u) + P_i u$ and $N_i(u) = Q_i u + r$. If $M_i(u)$ admits an inverse defined on $\mathbb{R}^N$, then one can reformulate the original problem (1) as

$$\mathcal{F}_i(u) := u - M_i^{-1}(N_i(u)) = 0. \tag{2}$$

Let $(E_i)_{i=1,\dots,K}$ be a family of nonnegative diagonal matrices such that $\sum_{i=1}^K E_i = I$. Multiplying (2) by E_i and summing over i we obtain the system

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

Clearly, the solution of the original system satisfies (3). The proposition below states that $\mathcal{F}$ is concave and that $\mathcal{F}'(u)$ has nonnegative inverse for all u, which implies, in particular, that $\mathcal{F}$ is inverse isotone, and, therefore, the solution to (3) is unique.

Proposition 2 *Assume that 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 (3) *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} are 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 for all $u \in \mathbb{R}^N$, the matrix $\mathcal{F}'(u)$ is an M-matrix satisfying $\mathcal{F}'(u)^{-1} \geq I$. We begin with the following spectral bound, which is the founding stone for the analysis of the multi-splitting methods (see [15])

$$\rho\left(\sum E_i M_i'(u)^{-1} Q_i\right) < 1. \tag{4}$$

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 maximum of the vectors $\widetilde{u}_i$; that is $(w)_k = \max_i (\widetilde{u}_i)_k$. Since $u \mapsto M_i'(u)$ is antitone and $M_i'(u)$ has nonnegative inverse, we deduce that $M_i'(\widetilde{u}_i)^{-1} M_i'(w) \leq I$, and, since $M_i'(\widetilde{u}_i)^{-1} Q_i \geq 0$ and $M_i'(w)^{-1} Q_i \geq 0$, 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 (4) that

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

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 $n \geq 0$

$$\widetilde{u}_n = \sum_i E_i M_i^{-1}(N_i(u_n)) \tag{5}$$

and

$$u_{n+1} = \widetilde{u}_n - F'(\widetilde{u}_n)^{-1} \left(F(\widetilde{u}_n) - r\right). \tag{6}$$

We note that (5) can be interpreted as a step of a quasi-Newton method applied to (3), where the matrix $\mathcal{F}(u)^{-1}$ has been replaced by its subinverse I. It can be shown that (5)–(6) leads again to a globally convergent scheme. Remarkably enough, in the case of a simple splitting, like (block) Jacobi or Gauss-Seidel, 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 and the preconditioned Newton's generate the same iterates.*

Proof Let $\widetilde{u} = M^{-1}(N(u))$, we remark that $\mathcal{F}(u) = u - \widetilde{u}$ and $\mathcal{F}'(u) = I - M'(\widetilde{u})^{-1}Q = M'(\widetilde{u})^{-1}F'(\widetilde{u})$. Therefore, the update generated by the preconditioned Newton's method starting from u is given by

$$\delta_{\text{prec}}(u) = \left(M'(\widetilde{u})^{-1}F'(\widetilde{u})\right)^{-1}(\widetilde{u} - u)\,. \tag{7}$$

Now, let us consider the update generated by the two-step method (5)-(6). We have $\delta_{\text{two-step}}(u) = \widetilde{u} - u - F'(\widetilde{u})^{-1}F(\widetilde{u})$. 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,

$$\delta u_{\text{two-step}} = \left(I + F'(\widetilde{u})^{-1}Q\right)(\widetilde{u} - u) = F'(\widetilde{u})^{-1}M'(\widetilde{u})(\widetilde{u} - u),$$

which, in view of (7), provides $\delta_{\text{two-step}}(u) = \delta_{\text{prec}}(u)$. □

3 Numerical experiment

We now proceed with the numerical experiment that illustrates the performance of block Jacobi-Newton, RASPEN and the two-step RAS/Newton methods applied to the system resulting from the discretization of a degenerate parabolic equation. The (block) Jacobi-Newton consists of applying Newton's method to the system of the form (2) obtained from a simple splitting $A = P - Q$, where P is a (block) diagonal part of A. On the other hand, RASPEN can be expressed as Newton's method applied to the system (3) resulting from a particular multi-splitting (wee refer to [8] for further details). Using same multi-splitting, the RAS/Newton method is given by (5)–(6).

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

$$\begin{cases} \beta(u) - \beta(u_{ini}) = \tau\partial^2_{xx}u & x \in (0,1), \\ \partial_x u(0) = -q, \quad \partial_x u(1) = 0, \end{cases} \tag{8}$$

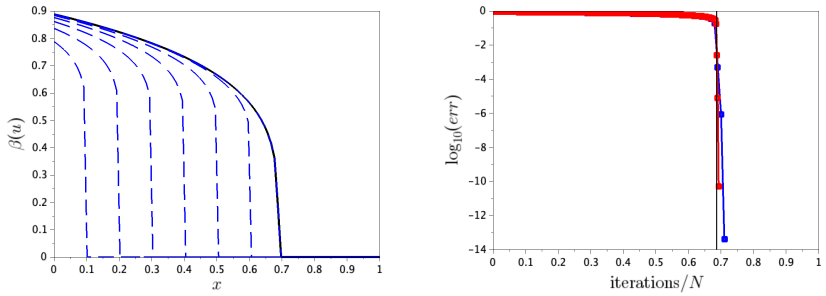

Fig. 1 Left: the solution for $N = 100$ (black) and the iterates $\beta(u_n)$ of the Jacobi-Newton method for $n = 10, 20, \ldots 70$. Right: convergence history of the Jacobi-Newton method for $N = 100$ (blue) and 400 (red), the iteration count is scaled by the size of the discrete system; the error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

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 (8) is discretized using $N = 100$ or 400 degrees of freedom, and the vector u_{ini} is used as the initial guess by the iterative methods under consideration.

We note that, since the derivative of β is unbounded at the origin, one may consider the change of variable in (8). For example, using $\beta(u)$ as the new unknown will improve the performance of the straightforward Newton's method [3]. Unfortunately, the modified system is no longer concave and, therefore, the monotone convergence is lost; moreover, compared to the splitting-based preconditioning, the convergence of Newton's method applied to the modified system turns out to be slower [2].

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 (8) the latter property is reflected in the performance of Newton's method. Typically, and unless some Schwarz-type preconditioning is performed, the solution fronts resulting from Newton's method can cross at most one degree of freedom at 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 a long period of a slow error decrease. As a 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 general nonnegative initial data. Nevertheless, the efficiency of the latter method is still dependent on N_f and thus on

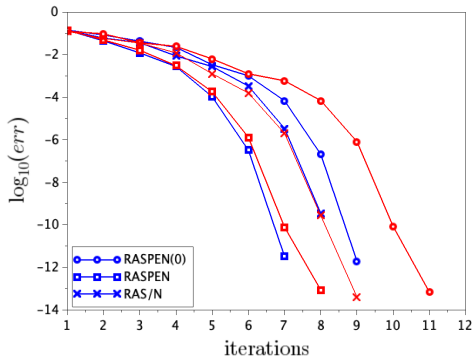

Fig. 2 Convergence history of preconditioned Newton's method for $N = 100$ (blue) and 400 (red); the error is measured in l_∞ norm.

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. The scaled 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 $N = 400$.

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), corresponding to preconditioning based on the block Jacobi method. In this numerical experiment none of the considered methods appear 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 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. On the abstract theory of Chaplygin's method. In: *Dokl. Akad. Nauk. SSSR*, vol. 83, 781–784 (1952).
2. Brenner, K. On the monotone convergence of Jacobi–Newton method for mildly nonlinear systems. *Journal of Computational and Applied Mathematics* **419**, 114719 (2023).
3. Brenner, K. and Cances, C. Improving Newton's method performance by parametrization: the case of the Richards equation. *SIAM Journal on Numerical Analysis* **55**(4), 1760–1785 (2017).
4. Cai, X.-C. and Keyes, D. E. Nonlinearly preconditioned inexact Newton algorithms. *SIAM Journal on Scientific Computing* **24**(1), 183–200 (2002).
5. Cai, X.-C. and Li, X. 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 (2011).
6. Dolean, V., Gander, M. J., Kheriji, W., Kwok, F., and Masson, R. Nonlinear preconditioning: How to use a nonlinear Schwarz method to precondition Newton's method. *SIAM Journal on Scientific Computing* **38**(6), A3357–A3380 (2016).
7. Frommer, A. Parallel nonlinear multisplitting methods. *Numerische Mathematik* **56**, 269–282 (1989).
8. Frommer, A. and Szyld, D. B. An algebraic convergence theory for restricted additive Schwarz methods using weighted max norms. *SIAM journal on numerical analysis* **39**(2), 463–479 (2001).
9. Kern, M., Taakili, A., and Zarrouk, M. Preconditioned iterative method for reactive transport with sorption in porous media. *Mathematical Modelling and Analysis* **25**(4), 546–568 (2020).
10. Klawonn, A., Lanser, M., and Rheinbach, O. Nonlinear feti-dp and bddc methods. *SIAM Journal on Scientific Computing* **36**(2), A737–A765 (2014).
11. Klawonn, A., Lanser, M., and Rheinbach, O. Nonlinear BDDC methods with approximate solvers (2018).
12. Klemetsdal, Ø., Moncorgé, A., Møyner, O., and Lie, K.-A. 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* 1–19 (2021).
13. Liu, L. and Keyes, D. E. Field-split preconditioned inexact Newton algorithms. *SIAM Journal on Scientific Computing* **37**(3), A1388–A1409 (2015).
14. Ortega, J. M. and Rheinboldt, W. C. *Iterative solution of nonlinear equations in several variables*. SIAM (2000).
15. O'Leary, D. P. and White, R. E. Multi-splittings of matrices and parallel solution of linear systems. *SIAM Journal on algebraic discrete methods* **6**(4), 630–640 (1985).
16. Vázquez, J. L. *The porous medium equation: mathematical theory*. Oxford University Press on Demand (2007).