A choice of forcing terms in inexact Newton iterations with application to pseudo-transient continuation for incompressible fluid flow computations

L. Botti[a,*]

[a] *Università degli studi di Bergamo*
via Marconi 4, 24044 Dalmine (BG), Italy

Abstract

A strategy for choosing the forcing terms in inexact Newton iterations is presented. The final goal is to obtain fast steady state marching strategies for the solution of PDEs. The new approach is analysed and tested in the context of inexact Newton methods but is also well suited to be applied in the pseudo-transient continuation framework. To validate the strategy and assess its gains in terms of computational costs we seek approximate solutions of the Incompressible Navier-Stokes equations at high-Reynolds numbers. In particular we consider the well known 2D lid-driven cavity flow and backward-facing step problem focusing on the efficiency of the time marching strategy. Residual history and computation time are monitored and compared with many fixed and adaptive forcing term choices of reference.

Keywords: Discontinuous Galerkin methods, Incompressible Navier-Stokes equations, High-Reynolds numbers, Pseudo-Transient continuation, Steady state time marching
PACS: 02.70.Dh, 47.10.ad, 47.11.Fg

1. Introduction

In this work we devise an effective strategy to define the forcing terms of inexact Newton iterations with application to globalisation strategies like global inexact Newton methods and pseudo-transient continuation algorithms. Globalisation of inexact Newton solvers is necessary to ensure convergence when the initial iterate is far from the solution. The technique is widely em-

*Corresponding author
Email address: lorenzo.botti@unibg.it (L. Botti)

ployed for the computation of steady-state solutions of non-linear differential equations, see *e.g.* [3, 6, 8, 9, 17, 25] for some recently published applications.

As a distinctive feature inexact Newton methods do not require to exactly solve the Newton equations. In iterative Newton or truncated Newton methods iterative solvers are employed to solve linear systems, resulting in a nested inner loop at each outer Newton iteration. Clearly the more effort is spent to solve the system the better its solution is approximated. Nevertheless, if the solution guess is not sufficiently near the exact solution, *oversolving* the Newton equations might result in significant linearisation errors and unsatisfactory convergence rates.

Forcing terms control the expense of the inner loop by prescribing a termination condition for the linear solver. They also influence the local convergence of inexact Newton methods, see Dembo, Eisenstat and Steihaug [11], and the robustness of globalisation strategies. In all the applications in which the cost of the inner iterations is high the choice of forcing term is of primary importance for the efficiency of the whole algorithm.

Many forcing term choices has been proposed in literature to optimize the computational costs involved in inexact Newton methods. Dembo and Steihaug [12] and Brown and Saad [7] devised the first adaptive strategies for decreasing the forcing terms while approaching the exact solution. Later Eisenstat and Walker [14] introduced two effective forcing term choices for inexact Newton methods [13] that are still widely used. Fast local convergence was demonstrated theoretically and the ability to avoid oversolving was analyzed by means of numerical test cases. Recently An, Mo and Liu [2] proposed an approach for choosing the forcing term based on the agreement between the linear and non-linear model at the each Newton iteration. They demonstrated q-superlinear convergence of the resulting Newton strategy and obtained satisfactory numerical results as compared with the choices by Eisenstat and Walker.

Up to the author knowledge, none of the adaptive forcing term choices proposed in the context of inexact Newton methods was systematically applied to pseudo-transient continuation. On one hand, the modification of the Jacobian matrix induced by the discretization of the time derivative reduces the burden of numerically solving the Newton equation while marching with small timesteps. On the other hand, when the timestep is large, the computational cost associated with the inner linear iteration increases to the point that tight termination conditions might be difficult to meet in practice. Several authors favored fixed forcing terms choices, see *e.g.* Tidriri [28], while others opted for a constant work and storage per timestep, *e.g.* Venkatakrishnan and Mavriplis [30] suggested to fix the amount of inner iterations renouncing to control the solution accuracy. Aiming for a prescribed termi-

nation condition but exiting irrespectively of the its fulfillment after a fixed amount of iterations is a common practice, see *e.g.* Gropp, Keyes, McInnes, and Tidriri [18]. An adaptive approach based on the evolution of the Courant number was proposed by Ajmani, Ng and Liou [1] in the context of Computational Fluid Dynamics (CFD) applications. The role of forcing terms in inexact pseudo-transient continuation was analyzed in detail by Kelley and Keyes [21] with specific recommendations for each phase of the global convergence.

In this work a new prediction-correction strategy for computing the forcing terms is proposed. The correction enforces the best possible agreement between the linear and the non-linear model at the current Newton iterate while the prediction modifies the forcing term based on the convergence improvements expected at the next iterate. The material is organized as follows. In Section 2 we briefly formalize inexact Newton methods focusing on computational costs and reporting some theoretical results regarding convergence rates. Section 3 reviews existing forcing term choices and describes in detail the new strategy analyzing the local convergence of the inexact Newton algorithm that results from it. In Section 3.2.4 the forcing term strategies are challenged testing their response to artificially manufactured nonlinear convergence rates typically observed at early stages of a global Newton iteration. In Sections 4 and 5 we consider the new choice in the context of globalization strategies and perform numerical test cases. First we apply global inexact Newton methods to non-linear model problems comparing the new choice with other forcing term choices of reference, see Section 4. Thus, in Section 5, we adapt the new adaptive forcing term choice to pseudo-transient continuation and apply the pseudo-time marching strategy to a discontinuous Galerkin discretization of the incompressible Navier-Stokes equations. We seek steady-state solutions of the 2D lid-driven cavity and backward-facing step problems at high-Reynolds numbers showing how the new forcing term algorithm behaves as compared with other forcing term choices.

2. Inexact Newton methods

Consider the system of (nonlinear) equations

$$F(x) = 0, \tag{1}$$

where $x \in \mathcal{R}^n$ is the global vector of unknown, $F(x) : \mathcal{R}^n \to \mathcal{R}^n$ is the vector of *residuals*, *i.e.* the vector of nonlinear functions of x resulting from the discrete space operators of a finite element discretization.

A classical way of finding $x_* \,|\, F(x_*) = 0$ is Newton's method. Starting from an initial guess x_0 sufficiently close to x_* a sequence of approximations

$\{x_k\}$ is obtained repeatedly solving the linear Newton equation

$$F^{'}(x_k)s_k = -F(x_k), \tag{2}$$

and setting $x_{k+1} = x_k + s_k$. In practice, since equation (2) can be expensive to solve due to the size, the structure and the condition number of the Jacobian matrix $F^{'}(x_k)$, it is convenient to compute approximate solutions. In inexact Newton methods equation (2) is solved inexactly, usually by means of an iterative method, subject to the following condition on the linear system residual

$$\left|\left|F(x_k) + F^{'}(x_k)s_k\right|\right| \leq \eta_k \left|\left|F(x_k)\right|\right|, \tag{3}$$

where $\eta \in [0, 1)$ is the so called *forcing term*. In inexact Newton methods η controls how accurately s_k solves the Newton equation (2). A proper choice of η is of primary importance to ensure computational efficiency and maintain the local convergence properties of Newton's algorithm. As a matter of fact, while η too small negatively impacts computational costs, see Section 2.1, a too loose η value affects Newton method's local convergence rates, see Section 2.2.

2.1. Computational cost and oversolving

When solving equation (2) by means of an iterative solver, condition (3) provides a stopping criteria based on the decrease of the relative residual, *i.e.* $\frac{\left|\left|F(x_k)+F^{'}(x_k)s_k\right|\right|}{||F(x_k)||} \leq \eta_k$. A smaller η_k typically involves additional expense on the iterative solution process that might fail to pay off in terms of global convergence, a phenomenon known as *oversolving*.

Oversolving occurs when stringent requirements on the residual reduction cause a significant disagreement between $F(x)$ and its local linear model, see Eisenstat and Walker [14]. First note that $R(x_k, s_k) = F(x_k) + F^{'}(x_k)s_k$ is the residual of the Newton equation, as well as the local linear model of $F(x)$, *i.e.* $F(x_{k+1}) \approx F(x_k) + F^{'}(x_k)s_k$. Once s_k has been computed, the norm of the linearisation error $E(x_k, s_k)$ can be evaluated as follows

$$||E(x_k, s_k)|| = \left|\left|F(x_{k+1}) - F(x_k) - F^{'}(x_k)s_k\right|\right| = ||F(x_{k+1}) - R(x_k, s_k)||. \tag{4}$$

Following Eisenstat and Walker [13], we define the *predicted reduction*, $pred_k(s_k)$, and the *actual reduction*, $ared_k(s_k)$, at the kth iteration

$$ared_k(s_k) \stackrel{\text{def}}{=} ||F(x_k)|| - ||F(x_{k+1})||, \tag{5}$$

$$pred_k(s_k) \stackrel{\text{def}}{=} ||F(x_k)|| - ||R(x_k, s_k)||. \tag{6}$$

To justify the computational effort required to solve Equation (2), $ared_k(s_k) \simeq pred_k(s_k)$ should be obtained, meaning that the linear and non-linear model for $F(x_k)$ are in good agreement. Indeed the computational expense required to lower $R(x_k, s_k)$ in the inner linear iteration contributes to lower $F(x_{k+1})$ in the outer Newton iteration. The disagreement between $ared_k(s_k)$ and $pred_k(s_k)$ is closely related to the linearisation error $E(x_k, s_k)$, indeed

$$|ared_k(s_k) - pred_k(s_k)| = \big|||F(x_{k+1})|| - ||R(x_k, s_k)||\big| \leq ||E(x_k, s_k)|| \quad (7)$$

Elaborating on the nonlinear residual ratio

$$\begin{aligned} \frac{||F(x_{k+1})||}{||F(x_k)||} &= \frac{\big|\big|F(x_k) + F^{'}(x_k)s_k + E(x_k, s_k)\big|\big|}{||F(x_k)||} \\ &\leq \frac{\big|\big|F(x_k) + F^{'}(x_k)s_k\big|\big| + ||E(x_k, s_k)||}{||F(x_k)||} \quad (8) \\ &\leq \eta_k + \frac{||E(x_k, s_k)||}{||F(x_k)||}, \quad (9) \end{aligned}$$

we easily infer that a tight η_k might be pointless if the last term on the right hand side dominates. In particular the risk of oversolving must be taken into account when the initial guess x_0 is not sufficiently close to the solution x_*, which might induce a significant linearisation error.

2.2. Local convergence rates

Local convergence of inexact Newton methods has been demonstrated by Dembo, Eisenstat and Steihaug [11] under the weak assumption that the forcing sequence $\{\eta_k\}$ is uniformly less than one. We consider the following *standard assumptions* on F.

i Equation (1) has a solution x_*.
ii $F : \mathcal{R}^n \to \mathcal{R}^n$ is continuously differentiable
iii $F^{'}(x_*)$ is nonsingular.

Theorem 2.1 (Dembo et al. [11] Theorem 2.3)**.** *Assume that the standard assumptions hold. Let $\{\eta_k\}$ be such that $\eta_k \leq \eta_{max} < t < 1$ for all k, then there exist $\delta > 0$ such that, for any $x_0 \in N_\delta(x_*) \stackrel{\text{def}}{=} \{x : ||x - x_*|| \leq \delta\}$, the sequence of inexact Newton iterates $\{x_k\}$ converges to x_* in the weighted norm $||\cdot||_* \stackrel{\text{def}}{=} \big|\big|F^{'}(x_*)\cdot\big|\big|$, that is*

$$||x_{k+1} - x_*||_* \leq t\,||x_k - x_*||_* . \quad (10)$$

A similar result was also devised by Kelley [20, Theorem 6.1.4]. Moreover, as remarked by Kelley, q-linear convergence with respect to the norm $||\cdot||_*$ is equivalent to q-linear convergence of the sequence of nonlinear residuals $\{||F(x_n)||\}$. Kelley formalized this result in the following Proposition.

Proposition 1 (Kelley [20] Proposition 6.1.1)**.** *Assume that the standard assumptions hold and let $x_k \to x_*$. Then $||F(x_k)||$ converges q-linearly to 0 if and only if $||x_k - x_*||_*$ does.*

A proof of Proposition 1 is given in Appendix A.

Theorem 2.2 (Dembo et al. [11] Corollary 3.5)**.** *Assume that $F : \mathcal{R}^n \to \mathcal{R}^n$ is continuously differentiable in a neighborhood of $x_* \,|\, F(x_*) = 0$ and $F^{'}(x_*)$ is nonsingular. If the sequence of inexact Newton iterates $\{x_k\}$ converges to x_*, then*

1. *$\{x_k\}$ converges to x_* superlinearly if $\eta_k \to 0$*
2. *the convergence is q-quadratic if $\eta_k = \mathcal{O}\left(||F(x_k)||\right)$ and $F^{'}(x)$ is Lipschitz continuous at x_*.*

Accordingly the forcing term choice might limit the convergence rate of the sequence $\{x_k\}$ if not carefully selected, in particular when the initial guess x_k is close to x_* and the linearisation error is small.

3. Forcing term choice

3.1. Review of existing approaches

In their breakthrough work about forcing term choice in inexact Newton methods Eisenstat an Walker [14] proposed and analyzed two strategies for choosing η. The first couple of alternatives (EW1a and EW1b respectively) reads

$$\eta_{k+1} = \frac{||E(x_k, s_k)||}{||F(x_k)||}, \tag{11}$$

and

$$\eta_{k+1} = \frac{|pred_k(s_k) - ared_k(s_k)|}{||F(x_k)||}. \tag{12}$$

They directly reflect the agreement between $F(x)$ and its local linear model at the previous step. The convergence of the resulting Newton method is q-superlinear and two-step q-quadratic for choices (11) – (12), see [14] for details. The second approach by Eisenstat and Walker (EW2) relates η to the actual residual decrease

$$\eta_{k+1} = \gamma \left(\frac{||F(x_k + s_k)||}{||F(x_k)||} \right)^{\alpha} \tag{13}$$

where $\gamma \in [0,1]$ and $\alpha \in (1,2]$ drive the rate of decrease of the sequence $\{\eta_k\}$. For inexact Newton with choice (13) the convergence is of q-order α if $\gamma < 1$, and of q-order p for ever $p \in [1,\alpha)$ if $\gamma = 1$, see [14] for details. The authors tested several combination of γ and α and also suggested some safeguards that prevents the forcing term from becoming too small too quickly.

Later An, Mo and Liu [2] proposed a different approach (AML) based on agreement between $ared_k(s_k)$ and $pred_k(s_k)$. Defining $t_k = \frac{ared_k(s_k)}{pred_k(s_k)}$ their forcing term choice reads

$$\eta_{k+1} = \begin{cases} 1 - 2p_1, & \text{if } t_k < p_1 \\ \eta_k, & \text{if } p_1 \leq t_k < p_2, \\ 0.8\eta_k, & \text{if } p_2 \leq t_k < p_3, \\ 0.5\eta_k, & \text{if } t_k > p_3, \end{cases} \tag{14}$$

where p_1, p_2, p_3 are user defined parameters such that $0 < p_1 < p_2 < p_3 < 1$ and $p_1 \in (0, 0.5)$. In [2] q-superlinear convergence was proved for inexact Newton with choice (14).

3.2. The new choice

In this work we introduce the following new choice for η. Given $\eta_{k=0} \in (0,1)$ and $\alpha \in (1,2]$, compute

$$\eta_{k+1} = \frac{||R(x_k, s_k)||}{||R(x_k, s_k)|| + \alpha(||F(x_k)|| - ||F(x_k + s_k)||)}. \tag{15}$$

The strategy in (15) is here interpreted as prediction-correction strategies for computing η_{k+1} where

- prediction should guarantee satisfactory local convergence of the inexact Newton algorithm, see Section 3.2.2,
- correction attempts to enforce $ared_k(s_{k+1}) \simeq pred_k(s_{k+1})$ so to avoid *oversolving* and obtain the best efficiency from the computational costs viewpoint, see Section 3.2.1.

The user defined parameter α controls the behaviour of the sequence $\{\eta_k\}$ with respect to $\frac{||F(x_k+s_k)||}{||F(x_k)||}$, in particular $\eta_k \to 0$ if $\frac{||F(x_k+s_k)||}{||F(x_k)||} \leq \frac{\alpha-1}{\alpha}$, see Section 3.2.4 for details.

3.2.1. Correction

As a first step, once s_k is obtained according to condition (3), we compute $\bar{\eta}_k$ and c_k such that

$$pred_k(s_k) = ||F(x_k)|| - ||R(x_k, s_k)|| \quad = (1 - \bar{\eta}_k)\,||F(x_k)||\,, \tag{16}$$
$$ared_k(s_k) = ||F(x_k)|| - ||F(x_k + s_k)|| = c_k\bar{\eta}_k\,||F(x_k)||\,. \tag{17}$$

We obtain

$$\bar{\eta}_k = \frac{||R(x_k, s_k)||}{||F(x_k)||}, \tag{18}$$

and

$$c_k = \frac{||F(x_k)|| - ||F(x_k + s_k)||}{\bar{\eta}_k \, ||F(x_k)||} = \frac{||F(x_k)|| - ||F(x_k + s_k)||}{||R(x_k, s_k)||}. \tag{19}$$

Thus we correct $\bar{\eta}_k$ based the optimal condition $ared_k(s_k) = pred_k(s_k)$. To this end we introduce the correction $\hat{\eta}_k$ such that

$$(1 - \hat{\eta}_k) \, ||F(x_k)|| = c_k \hat{\eta}_k \, ||F(x_k)|| \, . \tag{20}$$

We obtain $\hat{\eta}_k = \frac{1}{1+c_k}$, which upon substitution of (19), leads to

$$\hat{\eta}_k = \frac{||R(x_k, s_k)||}{||R(x_k, s_k)|| + ||F(x_k)|| - ||F(x_k + s_k)||}. \tag{21}$$

Note that, according to condition (3), $\eta_k \geq \bar{\eta}_k$ but the two are expected to be comparable since the linear solver terminates as soon as condition (3) is satisfied. Nevertheless, since in the early stages of the Newton iterations one might get $||R(x_k, s_k)|| \ll \eta \, ||F(x_k)||$, we also mention about the possibility to use

$$\hat{c}_k = \frac{||F(x_k)|| - ||F(x_k + s_k)||}{\eta_k \, ||F(x_k)||}. \tag{22}$$

as a practical safeguard to avoid an unintended η decrease. Proceeding as above with $\hat{c}_k$ instead of c_k we get

$$\hat{\eta}_k = \frac{\eta_k \, ||F(x_k)||}{\eta_k \, ||F(x_k)|| + ||F(x_k)|| - ||F(x_k + s_k)||} \tag{23}$$

instead of (21).

3.2.2. Prediction

Once the correction $\hat{\eta}_k$ has been obtained based on the convergence achieved at the current iterate, η_{k+1} is computed based on a prediction of the convergence rate improvements expected at the next iterate. Indeed the correction alone would enforce q-linear convergence of the nonlinear residual impairing the convergence rate of the Inexact Newton Method when x is sufficiently close to x_*, see Theorem 2.2.

The prediction is obtained setting $\eta_{k+1} = \frac{1}{1+\alpha c_k}$, which leads to (15). To further motivate this choice we consider the limit case $\alpha = 2$. For $c \in (0, 0.8]$, leading to a predicted $\eta \in (1, \simeq 0.38)$, we have

$$\left(\frac{1}{1+c}\right)^2 \leq \frac{1}{1+2c} \leq \left(\frac{1}{1+c}\right)^{(1+\sqrt{5})/2},$$

where $1+\sqrt{5}/2$ and 2 correspond to the local convergence of the Secant and the Newton method, respectively, see Figure 1. c values smaller than one are indicative of a significant disagreement between the linear and nonlinear model for $F(x)$. As opposite $c \gg 1$ is obtained in the limit of a vanishing lin-

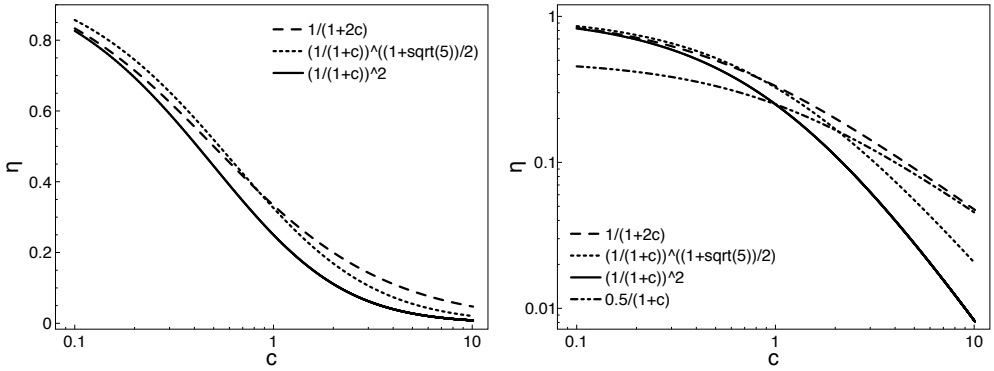

Figure 1: Behaviour of different forcing term prediction strategies for $c \in [0.1, 10]$. *Left*, logarithmic-linear chart to appreciate the differences for a predicted η close to unity. *Right*, logarithmic-logarithmic chart to appreciate the differences for $\eta \to 0$.

earisation error. Since $||E(x_k, s_k)|| \to 0$ implies $||F(x_k + s_k)|| \simeq ||R(x_k, s_k)||$, from (15) we get

$$\begin{aligned} \eta_{k+1} &\simeq \frac{||R(x_k, s_k)||}{||R(x_k, s_k)|| + 2\,||F(x_k)|| - 2\,||R(x_k, s_k)||}, \\ &\simeq \frac{||R(x_k, s_k)||}{2\,||F(x_k)|| - ||R(x_k, s_k)||}, \\ &\simeq \frac{\eta_k}{2 - \eta_k}, \end{aligned} \tag{24}$$

which mimics the strategy to reduce η in the terminal phase of (14), that is $\eta_{k+1} = 0.5\eta_k$. This strategy prevents η to decrease too fast when $x \to x_*$, see Figure 1.

3.2.3. Theoretical results

The following Theorem demonstrates that the strategy in (15) allows to obtain a locally convergent inexact Newton method. The proof is inspired by [2, Theorem 2.1] and requires the following Lemmas formulated by Orthega and Rheinboldt [26].

Let $N_\delta(z) \stackrel{\text{def}}{=} \{x : ||x - z|| \leq \delta\}$.

Lemma 3.1. Orthega and Rheinboldt [26, 2.3.3]. *Assume $F : \mathcal{R}^n \to \mathcal{R}^n$ is continuously differentiable and $x \in \mathcal{R}^n$. If $F^{'}(x)$ is nonsingular, then for any $\epsilon > 0$, there exists $\delta > 0$, such that $F^{'}(y)$ is nonsingular and*

$$\left|\left|F^{'}(y)^{-1} - F^{'}(x)^{-1}\right|\right| < \epsilon,$$

whenever $y \in N_\delta(x)$.

Lemma 3.2. Orthega and Rheinboldt [26, 3.2.10]. *Assume $F : \mathcal{R}^n \to \mathcal{R}^n$ is continuously differentiable. Then for any $z \in \mathcal{R}^n$ and $\epsilon > 0$, there exists $\delta > 0$, such that*

$$\left|\left|F(x) - F(y) - F^{'}(y)(x-y)\right|\right| < \epsilon \left|\left|x - y\right|\right|,$$

whenever $x, y \in N_\delta(z)$.

Theorem 3.1. *Assume that the standard assumptions hold. Let $\eta_0 < 1$ and compute $\eta_{k>0}$ using the forcing term strategy in (15). If x_0 is sufficiently close to x_*, then the sequence $\{x_k\}$ produced by the inexact Newton method converges to x_* q-superlinearly.*

Proof. We first prove that q-linear convergence of sequence $\{x_k\}$ in the weighted norm $\left|\left|\cdot\right|\right|_*$, that is

$$\left|\left|x_{k+1} - x_*\right|\right|_* \leq t \left|\left|x_k - x_*\right|\right|_*. \tag{25}$$

$t \in (0,1)$, implies $\eta_k < 1$ for all k. If $||F(x_{k+1})|| < ||F(x_k)||$ it is readily inferred that (15) leads to $\eta_{k+1} < 1$. Since, under the hypothesis of Proposition 1, $||F(x_k)||$ converges q-linearly to 0 if and only if $\left|\left|x_k - x_*\right|\right|_*$ does, the proof is completed. This means that, by Theorem 2.1, the sequence $\{x_k\}$ converges to x_*.

To prove q-superlinear convergence let $\beta = \left|\left|F^{'}(x_*)^{-1}\right|\right|$ and

$$\bar{\eta}_k = \frac{||R(x_k, s_k)||}{||F(x_k)||} <= \eta_k.$$

By Lemmas 3.1 and 3.2, there exist $\delta > 0$, such that $F'(x)$ is invertible and the inequalities

$$\left|\left|F^{'}(x)^{-1}\right|\right| < 2\beta, \tag{26}$$

and

$$\left|\left|F(x_{k+1}) - F(x_k) - F^{'}(x_k)(s_k)\right|\right| = ||E(x_k, s_k)|| < \frac{(\sqrt{\alpha} - 1)(1 - \bar{\eta}_k)}{\sqrt{\alpha} 2\beta(1 + \bar{\eta}_k)} \left|\left|s_k\right|\right|, \tag{27}$$

hold whenever $x_k, x_{k+1} \in N_\delta(x^*)$. The existence of a positive integer K such that $x_k \in N_\delta(x_*)$ for all $k > K$, follows from q-linear convergence of the sequence $\{x_x\}$. Therefore, for $k > K$, by the inexact Newton condition (3) and (26), we get

$$||s_k|| = \left|\left|F^{'}(x_k)^{-1}\left(F^{'}(x_k)s_k + F(x_k) - F(x_k)\right)\right|\right| \leq 2\beta(1+\bar{\eta}_k)\,||F(x_k)|| \tag{28}$$

Substituting the expression for s_k in (27) we get

$$||E(x_k, s_k)|| < \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}}(1-\bar{\eta}_k)\,||F(x_k)||\,, \tag{29}$$

and, from (8), we infer

$$\begin{aligned}
\frac{||F(x_{k+1})||}{||F(x_k)||} &\leq \bar{\eta}_k + \frac{||E(x_k, s_k)||}{||F(x_k)||} \\
&< \bar{\eta}_k + \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}}(1-\bar{\eta}_k) \\
&= \frac{\bar{\eta}_k}{\sqrt{\alpha}} + \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}} \\
&< 1.
\end{aligned}$$

Using the above result in choice (15) we obtain

$$\begin{aligned}
\eta_{k+1} &= \frac{||R(x_k, s_k)||}{||R(x_k, s_k)|| + \alpha\left(||F(x_k)|| - ||F(x_{k+1})||\right)}\frac{||F(x_k)||}{||F(x_k)||} \\
&= \frac{\bar{\eta}_k}{\bar{\eta}_k + \alpha\left(1 - \frac{||F(x_{k+1})||}{||F(x_k)||}\right)} \qquad (30) \\
&< \frac{\bar{\eta}_k}{\bar{\eta}_k + \alpha\left(1 - \frac{\bar{\eta}_k}{\sqrt{\alpha}} - \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}}\right)} \\
&= \frac{\bar{\eta}_k}{1 + (\sqrt{\alpha}-1)(1-\bar{\eta}_k)} \qquad (31)
\end{aligned}$$

where $(\sqrt{\alpha}-1)(1-\bar{\eta}_k) > 0$. According to (31), for a sufficiently large k, the sequence $\{\eta_k\}$ converges to zero, hence q-superlinear convergence follows thanks to Theorem 2.2.

Note that taking

$$||E(x_k, s_k)|| < \frac{\alpha-1}{\alpha}(1-\bar{\eta}_k)\,||F(x_k)||$$

in place of (29) is sufficient to demonstrate that the sequence $\{\eta_k\}$ is strictly decreasing. Indeed, proceeding as above, we get

$$\begin{aligned}
\eta_{k+1} &= \frac{\bar{\eta}_k}{\bar{\eta}_k + \alpha\left(1 - \frac{||F(x_k+s_k)||}{||F(x_k)||}\right)} \\
&< \frac{\bar{\eta}_k}{\bar{\eta}_k + \alpha\left(1 - \bar{\eta}_k - \frac{\alpha-1}{\alpha}(1-\bar{\eta}_k)\right)} \\
&= \bar{\eta}_k \\
&\leq \eta_k.
\end{aligned}$$

□

3.2.4. Asymptotic forcing terms analysis

The theoretical results of Section 3.2.3 confirms that the new strategy provides satisfactory convergence properties in the terminal phase of the Newton iteration, when x is sufficiently close to x_*. This is no surprise since many of the strategies proposed in literature, and in particular all the strategies considered in Section 3.1, provide similar, or even better, convergence properties. Nevertheless, in view of their application in the context of globalisation strategies, it is of primary importance to analyse how they behave at early stages of the global iteration, when stagnation of the nonlinear residual might be observed. In this phase the forcing term choices differ in how they react to linear model improvement and worsening.

For the sake of comparison we consider a fixed nonlinear residual ratio

$$r = \frac{||F(x_k + s_k)||}{||F(x_k)||}$$

that is we assume that the nonlinear residual converges q-linearly with q-order $r \in (0,1)$ for all k. Imposing $\eta_{k+1} = \eta_k = \bar{\eta}_k$, it is possible to compute the asymptotic forcing term value $\eta_a(r)$ that should be attained for $k \to \infty$. For each strategy the results are as follows.

- The EW1a strategy in (11) is not rewritable in terms of r. For the EW1b strategy in (12), which is closely related to (11), we get

$$\begin{aligned}
\eta_{k+1} &= \frac{|pred_k(s_k) - ared_k(s_k)|}{||F(x_k)||} \\
&= \frac{|||F(x_k + s_k)|| - ||R(x_k, s_k))|||}{||F(x_k)||} \\
&= |r - \bar{\eta}_k|
\end{aligned}$$

Imposing $\eta_{k+1} = \eta_k = \bar{\eta}_k$, we infer $\eta_a(r) = r/2$. Convergence towards the asymptotic value is not guaranteed, the sequence $\{\eta_k\}$ oscillate around $\eta_a(r)$.

- The EW2 strategy in (13) reads $\eta_{k+1} = \eta_a(r, \alpha) = r^\alpha < r$. Sudden convergence towards $\eta_a(r, \alpha)$ is observed indeed the asymptotic value is the forcing term at the next iterate.

- For the AML strategy in (14) we get $t = \frac{ared_k(s_k)}{pred_k(s_k)} = \frac{1-r}{1-\bar{\eta}}$ and

$$\eta_a(r, \eta, p1, p2) = \begin{cases} 1 - 2p_1, & \text{if } r > 1 - 2p_1^2, \\ (0, \max(\eta, 1 - 2p_1)] & \text{if } 1 - p_2 < r \leq 1 - 2p_1^2, \\ 0 & \text{if } r \leq 1 - p_2. \end{cases} \tag{32}$$

  Even if an analytic expression for the asymptotic forcing term is not available we obtain $\eta_a(r, \eta, p1, p2) <= r$. The convergence towards the asymptotic value is piecewise linear. Since the asymptotic value is a function of η, the same residual ratio r might be associated to different asymptotes.

- The new strategy in (15) reads

$$\eta_{k+1} \quad = \quad \frac{\bar{\eta}_k}{\bar{\eta}_k + \alpha\,(1 - r)}. \tag{33}$$

  see also (30). Imposing $\eta_{k+1} = \eta_k = \bar{\eta}_k$, we get

$$\eta_a(r, \alpha) = \begin{cases} \alpha\left(r - \frac{\alpha-1}{\alpha}\right), & \text{if } r > \frac{\alpha-1}{\alpha}, \\ 0, & \text{if } r \leq \frac{\alpha-1}{\alpha}. \end{cases} \tag{34}$$

  Note that, since $0 < \frac{\alpha-1}{\alpha} \leq 0.5$, we get $\eta_a(r, \alpha) < r$. The sequence $\{\eta_k\}$ converges smoothly and monotonically to $\eta_a(r, \alpha)$.

Not only the asymptotic forcing term is different for each of the strategies but, most importantly, the way it is approached changes. To outline how the strategies behaves we analyse how they react to artificially manufactured r loads in the following tests.

- Test 1. In order to simulate an improvement in nonlinear convergence, from $r = 0.7$ to $r = 0.65$, we start from $\eta_0 = \eta_a(0.7)$ and analyse the convergence towards $\eta_a(0.65)$ for $k = \{1, 2, ..., 10\}$. To conclude for $k = \{11, 12, ..., 20\}$ we switch back to $r = 0.7$ to simulate a worsening of the nonlinear convergence.

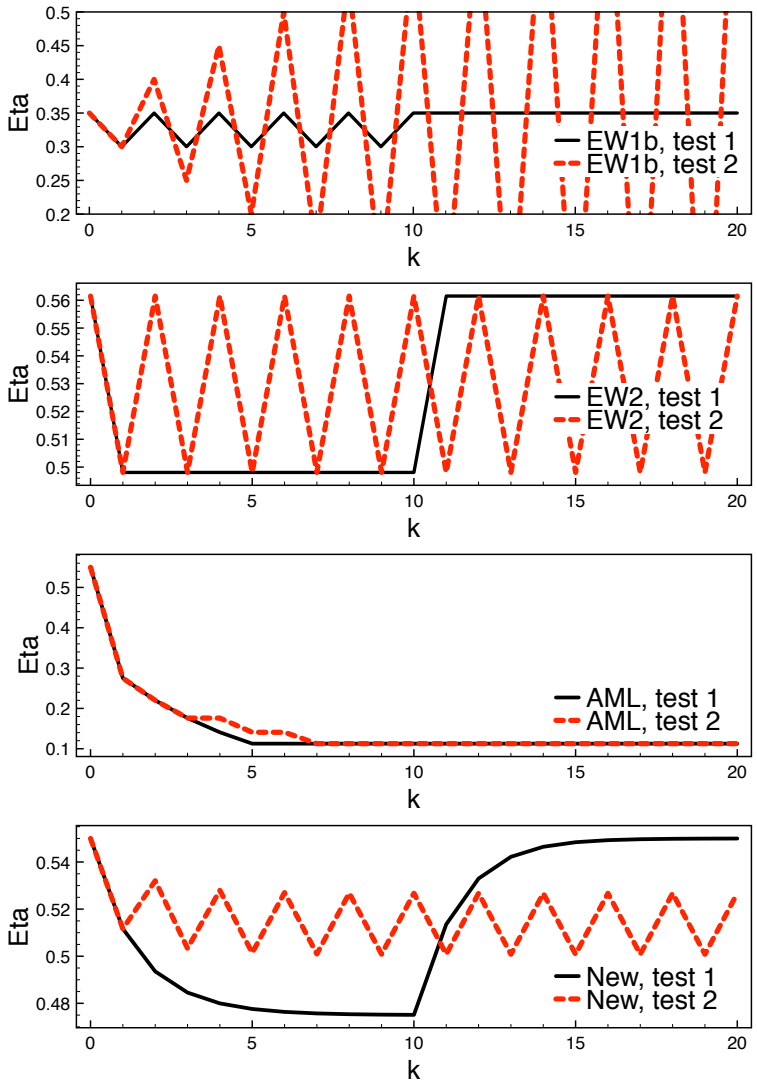

Figure 2: Behaviour of different forcing term prediction strategies (EW1b, EW2, AML, New) under artificially conceived nonlinear residual loads (Test 1 and 2), see text for details.

- Test 2. We consider an oscillatory convergence rate, for $k = \{1, 2, ..., , 20\}$ we set $r = 0.7$ and $r = 0.65$ if k is even and uneven, respectively.

Note that a nonlinear convergence q-factor $r \in [0.65, 0.7]$ might be commonly observed at the early stages of a global strategy.

We considered the following parameters. For EW2 we set $\gamma = 1$ and $\alpha = (1+\sqrt(5))/2$; for New we set $\alpha = 1.5$; for AML we set $p_1 = 0.1$, $p_2 = 0.4$, $p_3 = 0.7$ and we consider $\eta_0 = 0.55$ instead of $\eta_0 = \eta_a(0.7, \eta_0, p_1, p_2)$.

The results are reported in Figure 2. EW2 and New show the better

behavior in Test 1. They converge monotonically towards $\eta_a(0.65)$ and also converge back towards $\eta_a(0.7)$. As a plus New smoothly reacts to sudden changes of r. EW1 oscillates around $\eta_a(0.65)$ without converging to it. AML, after having reached $\eta_a(0.65)$, does not react to the sudden increase of r.

In Test 2 New and EW2 end up oscillating around $(\eta_a(0.65)+\eta_a(0.7))/2$. New oscillates with modest amplitude as compared to the extremes values $\eta_a(0.65)$ and $\eta_a(0.7)$ whereas EW2 oscillates between $\eta_a(0.65)$ and $\eta_a(0.7)$. EW1 shows big oscillations and rapidly goes out of scale while AML converges towards $\eta_a(0.65)$. Similarly to what observed in Test 1 AML displays an asymmetric way to react to an oscillatory load r.

We remark that in these artificial tests we disregarded the practical safeguards proposed by Eisenstat and Walker in [14] as well as the safeguard proposed by An, Mo and Liu in [2]. While the safeguard proposed for EW2 and AML does not influence their behavior in Test 1 and Test 2, the safeguard for EW1 keeps the oscillations under control in Test 2. Nonetheless the amplitude of the oscillations is much wider than $\eta_a(0.7)-\eta_a(0.65)$, which is the amplitude observed for EW2.

4. Global inexact Newton method

Besides the computational costs, the major drawback of Newton's method and inexact Newton methods is that the convergence is only *local*, meaning that the sequence $\{x_k\}$ might not converge to x_* if the initial guess x_0 is not sufficiently close to x_*. Since the local convergence properties are attractive several *globalization* strategies was devised to improve the likelihood of convergence from arbitrary starting points. To prove that the eta choice here proposed are effective in practice we consider numerical test cases based on *globally convergent inexact Newton* methods and compare with *forcing term* choices proposed by in literature.

A general framework for globally convergent inexact newton methods was introduced by Eisenstat and Walker [13]. Globalization can be achieved augmenting the inexact Newton condition (35) with a sufficient decrease condition on $||F(x)||$. In many applications the *forcing term* η_k is specified first, then s_k is computed, inexactly solving equation (2) according to condition (35). At last the residual decrease is enforced, possibly damping s_k and modifying η_k.

We consider the following Inexact Newton Backtracking method (INB) method

Algorithm 4.1 INB

set $x_k \leftarrow x_0$, choose $p \in (0,1)$ and $0 < \theta_{\min} < \theta_{\max} < 1$
for $k = 0$ step 1 until $||F(x_k)||$ is too large **do**
choose an initial $\eta_k \in [0, \eta_{\max})$, $\eta_{\max} < 1$, and determine s_k such that

$$\left|\left|F(x_k) + F^{'}(x_k)s_k\right|\right| \leq \eta_k \, ||F(x_k)|| \tag{35}$$

set $\eta_k^* = \eta_k$
while

$$||F(x_k + s_k)|| \geq [1 - p(1 - \eta_k^*)] \, ||F(x_k)|| \tag{36}$$

do
choose $\theta \in [\theta_{\min}, \theta_{\max}]$
update $s_k \leftarrow \theta s_k$ and $\eta_k^* \leftarrow 1 - \theta(1 - \eta_k^*)$
end while

set $\eta_k \leftarrow \eta_k^*$ (37)

set $x_{k+1} \leftarrow x_k + s_k$
end for

Theorem 6.1 of Eisenstat and Walker [13] states that if $\{x_k\}$ generated by Algorithm INB has a limit point x_* such that $F^{'}(x_*)$ is invertible, then $F(x_*) = 0$ and $x_k \to x_*$. Moreover in this case, for a sufficiently large k, the choice of η_k and s_k is accepted without modification in the backtracking phase. This allows to compare different strategies for computing η_k without any backtracking bias in the terminal phase of the convergence. Nevertheless the forcing term choice might influence the number of backtracking steps when x in far from x_*, see Table 2.

Using definitions (5)-(6) conditions (35)-(36) can be rewritten as

$$pred_k(s_k) \geq (1 - \eta_k) \, ||F(x_k)|| \tag{38}$$

$$ared_k(s_k) \geq p(1 - \eta_k) \, ||F(x_k)|| \tag{39}$$

respectively, see [13] for details, and s_k is accepted if $ared_k(s_k) \geq p \, pred_k(s_k)$. We follow the choice of Eisenstat and Walker [14] and set $p = 10^{-4}$, which seek to minimize the occurrence of backtracking.

In the while loop each θ is computed minimizing over $[\theta_{\min}, \theta_{\max}]$ the quadratic $v(\theta)$ for which $v(0) = g(0)$, $v^{'}(0) = g^{'}(0)$, and $v(1) = g(1)$, where $g(\theta) \stackrel{\text{def}}{=} ||F(x_k + \theta s_k)||_2^2$, as proposed by Eisenstat and Walker [14]. The parameters $\theta_{\min} = 0.1$ and $\theta_{\max} = 0.5$ are used in practice. Convergence is declared when $||F(x_k)|| \leq 10^{-6}$ or $||s_k|| \leq 10^{-12}$. $F^{'}(x)$ is computed analytically for each of the model problems considered. The GMRES algorithm

without restarting [27] is employed for the solution of linear systems which is generally achieved in less than 30 iterations, see Table 2. Additional safeguard conditions might be required depending on the initial η choice to ensure $\eta_k < \eta_{\max}$, the practical limits will be detailed in the next Section.

4.1. Comparison of forcing term choices in Global Inexact Newton

To evaluate the performance of the forcing term choices presented in Section 3 we use them in combination with Algorithm 4.1. We complete the definition of each strategy by setting the relevant parameters and we introduce the following notation for reporting the results.

1. New1.3, New1.5, New2, the new strategy in (15) with $\alpha = 1.3, 1.5, 2$;
2. EW1a, the first strategy given by Eisenstat and Walker, see (11);
3. EW1b, the second variant of the first strategy given by Eisenstat and Walker, see (12);
4. EW2, the second strategy given by Eisenstat and Walker with $\alpha = (1 + \sqrt{5})/2$ and $\gamma = 1$, see (13);
5. AML, the strategy devised by An, Mo and Liu with $p1 = 0.1$, $p2 = 0.4$ and $p3 = 0.7$, see (14);
6. Fixed, we consider the following set of fixed eta values $\{0.5, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$.

We set $\eta_{\max} = 0.99$ and $\eta_0 = 0.9$, while $\eta_{k>0}$ is computed according to the adaptive choices. We verified that further reduction of $\eta_{\max}$ in Algorithm 4.1 has a detrimental effect in most cases. Note that the choice of $\eta_{\max}$ does not affect AML since according to (14) we have $\eta_k <= 0.8$ for all $k > 0$. We implement the safeguards proposed by Eisenstat and Walker [14] to avoid η to become too small too quickly, that is

- EW1a and EW1b: Modify η_{k+1} by $\eta_{k+1} \leftarrow \max\{\eta_{k+1}, \eta_k^{(1+\sqrt{5})/2}\}$ whenever $\eta_k^{(1+\sqrt{5})/2} > 0.1$;
- EW2: Modify η_{k+1} by $\eta_{k+1} \leftarrow \max\{\eta_{k+1}, \gamma\eta_k^{\alpha}\}$ whenever $\gamma\eta_k^{\alpha} > 0.1$.

As suggested by An, Mo and Liu [2] we safeguard their strategy to avoid stagnation of the sequence $\{\eta_k\}$

- AML: Modify η_{k+1} by $0.5\eta_k$, if $t_k, t_{k-1} < p_1$ and $\eta_k, \eta_{k-1} > 0.1$,.

We also introduce a safeguard for the new choice having the goal to avoid an excessive η decrease in the early stages of the Newton iteration, see Section 3.2.1.

- New: Modify η_{k+1} by $\eta_{k+1} \leftarrow \frac{\eta_k ||F(x_k)||}{\eta_k ||F(x_k)|| + \alpha(||F(x_k)|| - ||F(x_k+s_k)||)}$ whenever $k < 4$ and $||R(x_k, s_k)|| < 0.5\eta_k \, ||F(x_k)||$;

For the sake of comparison we also include a computation with the New strategy, $\alpha = 1.3$ and no safeguard, the notation is New1.3ns.

Since the New strategy does not resort to η_k, the backtracked forcing term η_k^* in algorithm INB has no direct influence on η_{k+1}. Note that, even if the safeguard relies on η_k, backtracking is never triggered during the first iterations. As opposite the other strategies and safeguards might be influenced by the choice to omit line (37). We verified that in practice discarding η_k^* has a detrimental effect in case of EW1a and EW1b, doest not influence EW2 and New, while AML might perform slightly better. In what follows only the AML results have been obtained obtained without the update, in agreement with the INB algorithm proposed by An, Mo and Liu in [2]. Note that all choices are influenced by the backtracked value of s_k, in particular $F(x_k+s_k)$, $R(x_k, s_k)$ and $E(x_k, s_k)$ need to be recomputed after having updated s_k.

To compare the forcing term choices here considered we apply the INB Algorithm 4.1 for the solution of the following test problems proposed in literature, each with its own standard initial guess x_s.

Problem 4.1 (Three Diagonal system of Rosenbrock [2])**.**

$$\begin{cases} f_1(x) &= -4c(x_2 - x_1^2)x_1 - 2(1 - x_1), \\ f_i(x) &= 2c(x_i - x_{i-1}^2) - 4c(x_{i+1} - x_i^2)x_i - 2(1 - x_i), \quad i = 2, 3, ..., n-1, \\ f_n(x) &= 2c(x_n - x_{n-1}^2), \quad c = 2 \end{cases}$$

with $x_s = (1.2, 1.2, ..., 1.2)^T$.

Problem 4.2 (Three Diagonal system of Li [22], TD Li)**.**

$$\begin{cases} f_1(x) &= 4(x_1 - x_2^2), \\ f_i(x) &= 8x_i(x_i^2 - x_{i-1}) - 2(1 - x_i) + 4(x_i - x_{i+1}^2), \quad i = 2, 3, ..., n-1, \\ f_n(x) &= 8x_n(x_n^2 - x_{n-1}) - 2(1 - x_n), \end{cases}$$

with $x_s = (12, 12, ..., 12)^T$.

Problem 4.3 (Five Diagonal system of Li [22], FD Li)**.**

$$\begin{cases} f_1(x) &= 4(x_1 - x_2^2) + x_2 - x_3^2, \\ f_2(x) &= 8x_2(x_2^2 - x_1) - 2(1 - x_2) + 4(x_2 - x_3^2) + x_3 - x_4^2, \\ f_i(x) &= 8x_i(x_i^2 - x_{i-1}) - 2(1 - x_i) + 4(x_i - x_{i+1}^2) + \\ &\quad + x_{i-1}^2 - x_{i-2} + x_{i+1} - x_{i+2}^2, \quad i = 3, 4, ..., n-2, \\ f_{n-1}(x) &= 8x_{n-1}(x_{n-1}^2 - x_{n-2}) - 2(1 - x_{n-1}) + 4(x_{n-1} - x_n^2) + \\ &\quad + x_{n-2}^2 - x_{n-3}, \\ f_n(x) &= 8x_n(x_n^2 - x_{n-1}) - 2(1 - x_n) + x_{n-1}^2 - x_{n-2}, \end{cases}$$

with $x_s = (-2, -2, ..., -2)^T$.

Problem 4.4 (Seven Diagonal system of Li [22], SD Li)**.**

$$\begin{cases} f_1(x) & = 4(x_1 - x_2^2) + x_2 - x_3^2 + x_3 - x_4^2, \\ f_2(x) & = 8x_2(x_2^2 - x_1) - 2(1 - x_2) + 4(x_2 - x_3^2) \\ & \quad + x_1^2 + x_3 - x_4^2 + x_4 - x_5^2, \\ f_3(x) & = 8x_3(x_3^2 - x_2) - 2(1 - x_3) + 4(x_3 - x_4^2) \\ & \quad + x_2^2 - x_1 + x_4 - x_5^2 + x_1^2 + x_5 - x_6^2, \\ f_i(x) & = 8x_i(x_i^2 - x_{i-1}) - 2(1 - x_i) + 4(x_i - x_{i+1}^2) + \\ & \quad + x_{i-1}^2 - x_{i-2} + x_{i+1} - x_{i+2}^2 + \\ & \quad + x_{i-2}^2 + x_{i+2} - x_{i-3} - x_{i+3}^2, \quad i = 4, 5, ..., n-3, \\ f_{n-2}(x) & = 8x_{n-2}(x_{n-2}^2 - x_{n-3}) - 2(1 - x_{n-2}) + 4(x_{n-2} - x_{n-1}^2) + \\ & \quad + x_{n-3}^2 - x_{n-4} + x_{n-1} - x_n^2 + x_{n-4}^2 + x_n - x_{n-5}, \\ f_{n-1}(x) & = 8x_{n-1}(x_{n-1}^2 - x_{n-2}) - 2(1 - x_{n-1}) + 4(x_{n-1} - x_n^2) + \\ & \quad + x_{n-2}^2 - x_{n-3} + x_n + x_{n-3}^2 - x_{n-4}, \\ f_n(x) & = 8x_n(x_n^2 - x_{n-1}) - 2(1 - x_n) + x_{n-1}^2 - x_{n-2} + \\ & \quad + x_{n-2}^2 - x_{n-3}, \end{cases}$$

with $x_s = (-3, -3, ..., -3)^T$.

Problem 4.5 (Three Diagonal system of Broyden [23], TD Br)**.**

$$\begin{cases} f_1(x) & = x_1(0.5x_1 - 3) + 2x_2 - 1, \\ f_i(x) & = x_i(0.5x_i - 3) + x_{i-1} + 2x_{i+1} - 1, \quad i = 2, 3, ..., n-1, \\ f_n(x) & = x_n(0.5x_n - 3) - 1 + x_{n-1}, \end{cases}$$

with $x_s = (-1, -1, ..., -1)^T$.

Problem 4.6 (Three Diagonal Trigonometric Exponential system [29], TD TrEx)**.**

$$\begin{cases} f_1(x) & = 3x_1^3 + 2x_2 - 5 + sin(x_1 - x_2)sin(x_1 + x_2), \\ f_i(x) & = 3x_i^3 + 2x_{i+1} - 5 + sin(x_i - x_{i+1})sin(x_i + x_{i+1}) + \\ & \quad + 4x_i - x_{i-1}exp(x_{i-1} - x_i) - 3, \quad i = 2, 3, ..., n-1, \\ f_n(x) & = 4x_n - x_{n-1}exp(x_{n-1} - x_n) - 3, \end{cases}$$

with $x_s = (0, 0, ..., 0)^T$.

Each of the six problems is tested for $n = 5000$. The number of Newton iterations (Nit), the total number of GMRES iterations (Git), the average number of GMRES iterations per Newton step (Git/Nit), and the number of backtracking iterations (BTit) are reported in Tables 1 and 2 for all fixed and adaptive forcing term strategies, respectively.

Eta		TD Li	TD Ros	TD TrEx	TD Broy	FD Li	SD Li
0.5	Nit	204	19	18	15	27	28
	Git	1592	62	20	29	72	67*
	Git/Nit	7.8	3.7	1.1	1.9	2.7	2.4
	BTit	2153	0	1	0	5	6
10^{-1}	Nit	44	9	9	7	15	19
	Git	330	53	18*	25*	67*	77
	Git/Nit	7.5	5.9	2	3.6	4.5	4.1
	BTit	118	0	2	0	5	19
10^{-2}	Nit	21	6	7	5	16	13
	Git	148	45*	18*	27	115	80
	Git/Nit	7.04	7.5	2.6	5.4	7.2	6.2
	BTit	17	0	2	0	14	2
10^{-3}	Nit	15	5	6	4	14	17
	Git	110*	45*	28	28	141	330
	Git/Nit	7.3	9	4.7	7	10.1	19.4
	BTit	1	0	2	0	6	22
10^{-4}	Nit	19	5	6	4	16	17
	Git	226	62	38	38	289	447
	Git/Nit	11.9	12.4	6.33	9.5	18.1	26.3
	BTit	11	0	2	0	25	22

Table 1: Iterations count for the INB algorithm with different fixed forcing term choices. Git: GMRES iterations, Nit: Newton iterations, BTit: Backtracking iterations. * indicates the lowest number of GMRES iterations among the fixed η choices considered.

The fixed η results allows to appreciate the influence of the forcing term on the iteration counts and to identify the optimal η for each model problem. The best η is chosen as the one that provides the smaller Git count, here considered as an indication of the total computation time. A great variability is observed, from a stringent 10^{-3} for the Three Diagonal Rosenbrock system up to 0.5 for the Seven Diagonal case. For the Five and Seven Diagonal systems a significant Git counts increase is observed moving towards stringent forcing terms. For the Three Diagonal Li system a clear Git minimum is present at 10^{-3} with a significant Git increase is observed departing from this value. The Li system appears quite pathological being the sole case in which less stringent forcing terms increase the number of backtracking iterations. Indeed this tridiagonal system has the farthermost initial guess with respect to the exact solution, which increases the possibility to encounter some local minima.

Eta		TD Li	TD Ros	TD TrEx	TD Broy	FD Li	SD Li
New1.3	Nit	15	9	8	7	15	18
	Git	72*	45*	18*	28	62*	67*
	Git/Nit	4.8	5	2.3	4	4.1	3.7
	BTit	1	0	1	0	5	7
New1.5	Nit	15	8	8	7	15	17
	Git	73*	49	18*	28	64*	59*
	Git/Nit	4.9	6.1	2.3	4	4.3	3.5
	BTit	1	0	1	0	5	7
New2	Nit	15	7	8	7	15	15
	Git	80*	48	18*	34	74	65*
	Git/Nit	5.3	6.6	2.3	4.9	4.9	4.3
	BTit	1	0	1	0	5	8
New1.3ns	Nit	12	5	6	6	18	15
	Git	109*	38*	17*	26	70	72
	Git/Nit	9.1	7.6	2.8	4.3	3.9	4.8
	BTit	0	0	1	0	6	1
EW1a	Nit	91	12	8	9	23	18
	Git	256	53	17*	35	65*	53*
	Git/Nit	2.8	4.4	2.1	3.9	2.8	2.9
	BTit	78	0	1	0	5	2
EW1b	Nit	89	11	11	9	22	21
	Git	268	44*	21	44	58*	46*
	Git/Nit	3	4	1.9	4.9	2.6	2.2
	BTit	75	0	1	0	5	2
EW2	Nit	75	6	7	5	15	14
	Git	246	39*	21	31	61*	65*
	Git/Nit	3.3	6.5	3	6.2	4.1	4.6
	BTit	71	0	1	0	5	2
AML	Nit	17	8	8	7	18	14
	Git	129	48	16*	28	63*	65*
	Git/Nit	7.6	6	2	4	3.5	4.6
	BTit	5	0	1	0	5	2

Table 2: Iterations count for the INB algorithm with different adaptive forcing term choices, see text for details. Git: GMRES iterations, Nit: Newton iterations, BTit: Backtracking iterations. * indicates a number of GMRES iterations equal lower than the best result obtained with a fixed forcing term, see Table 1.

The New strategy provide the smallest number of GMRES iterations if we consider all the model problems ($\text{Git}_{sum} = 292, 291, 319$ for $\alpha = 1.3, 1.5, 2$,

respectively), smaller than the sum of GMRES iterations obtained considering the best fixed η for each problem ($\text{Git}_{sum} = 332$). Also New1.3ns with $\alpha = 1.3$ and no safeguard performs relatively well with $\text{Git}_{sum} = 332$. EW1a and EW1b highly reduce the average number of GMRES iteration per Newton step (Git/Nit), minimizing the risk of oversolving. This behaviour causes a relatively high number of GMRES iterations on the Three Diagonal Li case (TD Li), as opposite very good results are obtained for the Five and Seven Diagonal Li systems.

It is very hard (and out of the scope of this comparison) to indicate which is the best performing forcing term strategy based on the limited number of model problems here considered. Some might be more suited than others depending on the practical application at hand. Nevertheless the results in Table 2 confirms that the new strategy provides good local convergence properties and controls the expense of the inner iterations. Interestingly the iteration count is not highly sensitive to the choice of the parameter α. The results are satisfactory in each of the model problems, even if the INB algorithm requires different forcing term choices to perform at best.

5. Pseudo-transient continuation

In this section we consider globalization by means of *pseudo-transient continuation* methods which have proved well suited for CFD applications. In order to obtain a globally convergent Newton algorithm for the solution of equation (1) pseudo-transient continuation (Ψ_{tc}) exploits the following system of (nonlinear) ODEs

$$\frac{\partial x}{\partial t} + F(x) = 0, \tag{40}$$

which is simply the unsteady analogue of problem (1). Equation (40), supplemented with a suitable initial condition $x(0) = x_0$, can be numerically integrated to steady state repeatedly solving

$$\left(\frac{1}{\delta_k} M + F^{'}(x_k)\right) s_k = -F(x_k), \tag{41}$$

and setting $x_{k+1} = x_k + s_k$. In (41) $F^{'}(x_k)$ is the Jacobian matrix, M is the Mass matrix, *e.g.* a block diagonal or a diagonal matrix, depending on the finite element discretization, and δ_k is the variable timestep to be adapted during the time marching strategy. Small timesteps are commonly employed in the initial phase, which is dominated by the incorrectness of the initial guess, while the timesteps are increased as $F(x_k)$ approaches 0.

As in the case of inexact Newton methods it is convenient to solve Equation (41) by means of an iterative solver, subject to the following condition on the relative residual decrease

$$\left|\left| F(x_k) + \left(\frac{1}{\delta_k} M + F^{'}(x_k) \right) s_k \right|\right| \leq \eta_k \left|| F(x_k) \right||, \tag{42}$$

where $R_{\Psi_{tc}}(x_k, s_k, \delta_k) = F(x_k) + \left(\frac{1}{\delta_k} M + F^{'}(x_k) \right) s_k$ is the linear system residual. Accordingly the forcing term η_k has a strong influence on the number of iteration required to satisfy condition (42) and on the computational costs of the whole algorithm.

We refer the reader to Kelley and Keyes [21] for a detailed derivation of the time marching strategy and to Section 5.1 for an overview of the algorithm. Here we remark that Equation (41) can be interpreted an inexact Newton's method for $F(x) = 0$ where the exact Jacobian $F^{'}(x)$ has been modified according to the following considerations

i) Introducing

$$G(x) \stackrel{\text{def}}{=} M \frac{x - x_k}{\delta_k} + F(x), \tag{43}$$

equation (41) can be rewritten as $x_{k+1} = x_k - G^{'}(x_k)^{-1} F(x_k)$, which is the first Newton iterate for $G(x) = 0$. A small δ_k in equation (43) guarantees that x_{k+1} is sought sufficiently close to the current solution x_k. The timestep controls the solution increment so that the dynamics of the physical transient are tracked sufficiently well and the sequence $\{x_k\}$ is globally convergent.

ii) The sole difference between the inexact Newton equation (2) and (41) is the presence of the term $\frac{M}{\delta_k}$. If the timestep choice is such that $\delta_k \to \infty$ as $F(x) \to 0$, x_k is sufficiently close to x_* and η_k is small enough, pseudo-transient continuation maintains the local convergence properties of the Newton's method, see Kelley and Keyes [21].

iii) From the algebraic viewpoint the term $\frac{M}{\delta_k}$ increases the diagonal dominance of the system matrix simplifying the solution Equation (41) by means of an iterative solver. Since equation (2) might be very expensive to solve up to the prescribed accuracy, the timestep choice is of primary importance to control the computational costs.

The Selective Evolution Relaxation (SER) strategy introduced by [24], see also Section 5.2, is a standard and effective strategy to control timesteps in pseudo-transient continuation algorithms. In [21] Kelley and Keyes thoroughly analyzed the convergence rates of pseudo-transient continuation coupled with SER based time stepping. In what follows we briefly summarize

their findings regarding the three distinct phases taking place during the numerical integration towards the steady state.

i) Initial phase. x is far from the steady state solution and a small δ is required to guarantee global convergence. The convergence rates are not relevant but, nevertheless, an accurate time integration should be performed according to stability considerations.
ii) Midrange phase. Here pseudo-transient continuation should produce an accurate x and a large δ. During this phase x_k converges at most q-linearly to x_*.
iii) Terminal phase, δ is large and x is near the steady state solution. Here the local convergence properties of Newton's method can be fruitfully exploited.

Kelley and Keyes [21] also analyzed how the choice of η in (42) influences the convergence rates in each phase. Their estimates suggest that specific care must be devoted to the choice of η. While in the initial and midrange phase it is of primary importance to avoid oversolving, in order to reduce the computational expense, in the terminal phase, Newton like convergence should be guaranteed.

5.1. Inexact SER pseudo-transient continuation with backtracking

We employ the following pseudo-transient continuation algorithm

Algorithm 5.1 Ψ_{tc}

1: set $x_k \leftarrow x_0$, $\delta_k \leftarrow \delta_0$, $\eta_k \leftarrow \eta_0 \in (0,1)$, $f_{\min} = ||F(x_0)||$
2: **for** $k = 0$ step 1 until $||F(x_k)||$ is too large **do**
3: possibly find s_k such that

$$\left|\left|\left(\frac{M}{\delta_k} + F^{'}(x)\right) s_k + F(x)\right|\right| \leq \eta_k \, ||F(x)|| \tag{44}$$

4: compute δ_{k+1}
5: **if** $||F(x + s_k)|| < 1.2 \, f_{\min}$ **then**
6: set $x_{k+1} \leftarrow x_k + s_k$
7: compute η_{k+1}
8: set $f_{\min} \leftarrow ||F(x_k + s_k)||$
9: **end if**
10: **end for**

Algorithm 5.1 fits in the general framework analyzed by Kelley and Keyes [21] but was modified to include the safeguard condition at line 5. A 20%

residual increase is considered as an indication of the inability to follow the dynamics of the physical transient due to excessively large timesteps, the solution update is discarded to avoid the break down of the algorithm and the timestep is updated (read reduced, as will be clear in what follows). This very rough form of backtracking is usually invoked during the very first iterations due to the inconsistency of the initial solution x_0. As a matter of fact the rapid settlement of boundary conditions usually leads to a strong residual decrease involving an uncontrolled early timestep increase.

In practice line (44) implies the solution of (41) by means of an iterative solver. If the solver fails to converge but the safeguard condition at line 5 is satisfied, it is effective to update the solution without recomputing s_k. The solver and/or the preconditioner options are possibly tuned for the next iteration.

Algorithm 5.1 requires ad-hoc strategies to update the timestep and the forcing term at each iteration. The update procedures for δ and η are described in detail in Section 5.2 and Section 5.3, respectively. In particular the new forcing term choice proposed in Section 5.3 is the adaptation of strategy (15) to the Ψ_{tc} framework.

5.2. Timestep choice

As for the timestep update the well known Successive Evolution Relaxation (SER) method [24] is here modified to take into account the backtracking introduced in Algorithm 5.1. The timestep update is as follows

Algorithm 5.2 B-SER (Backtracking-SER)

1: **if** $||F(x_k + s_k)|| > 1.2\ ||F(x_k)||$ **then**
2: $\quad \delta_{k+1} \leftarrow \frac{4}{5}\delta_k$
3: **else**
4: $\quad \delta_{k+1} \leftarrow \delta_k \frac{||F(x_k)||}{||F(x_k+s_k)||}$
5: **end if**

If the residual stays within the safeguard limit the timestep increases in inverse proportion to the residual reduction (as in the standard SER strategy). As opposite, based on the observation that the solution increment s_k will be discarded in the event of backtracking, we also disregard the tentative residual $F(x + s_k)$ and simply reduce the timestep by 20%. This fixed relaxation procedure is based on the observation that backtracking is often an indication of an uncontrolled residual increase which risks to impact the timestep too severely. While in SER $\delta_k = \delta_0 \frac{||F(x_0)||}{||F(x_k)||}$, in B-SER we lose this property as a consequence of backtracking.

5.3. Algorithm for forcing term choice

In this Section we devise an algorithm for choosing forcing terms in inexact pseudo-transient continuation based on the strategy in (15). As a first step we remark that, in the context of pseudo-transient continuation, conditions (16)-(17) can be rewritten as follows

$$pred_k(s_k) = ||F(x_k)|| - ||R_{\Psi_{tc}}(x_k, s_k, \delta_k)|| = (1 - \bar{\eta}) \, ||F(x_k)|| \,, \tag{45}$$

$$ared_k(s_k) = ||F(x_k)|| - ||F(x_k + s_k)|| \quad = c_k \bar{\eta}_k \, ||F(x_k)|| \,. \tag{46}$$

Here the achievement of $ared_k(s_k) \simeq pred_k(s_k)$, that is q-linear convergence with q-factor η_k for the nonlinear residuals, is even more attractive. Indeed, as demonstrated by Kelley and Keyes in [21], q-linear convergence for the sequence $\{x_k\}$ is the best possible result in all but the terminal phase, see Section 5.

Proceeding as in Section 3.2.1 we get once again $\hat{\eta}_k = \frac{1}{1+c_k}$, where $c_k = \frac{||F(x_k)|| - ||F(x_k+s_k)||}{\left|\left|R_{\Psi_{tc}}(x_k,s_k,\delta_k)\right|\right|}$. Thus setting $\alpha = 2$, see Section 3.2.2, we predict $\eta_{k+1} = \frac{1}{1+2c_k}$. As compared to global inexact Newton, in Ψ_{tc} additional care is required to safely deal with residual increase, commonly observed and tolerated when x is far from x_*. We propose the following algorithm

Algorithm 5.3 Forcing term choice (Variable Eta)

if $k < 10$ **then**
  set $c_k \leftarrow \frac{1-\eta_{\max}}{2\eta_{\max}}$
  set $\eta_{k+1} \leftarrow \eta_{\max}$ {safeguard}
else
  compute $\bar{c}_k = \frac{||F(x_k)|| - ||F(x_k+s_k)||}{\left|\left|R_{\Psi_{tc}}(x_k,s_k,\delta_k)\right|\right|}$
  if $\bar{c}_k \geq c_{k-1}$ **then**
    set $c_k = 0.5c_{k-1} + 0.5\bar{c}_k$ {deferred correction}
  else
    set $c_k = 0.75c_{k-1} + 0.25\bar{c}_k$ {deferred correction}
  end if
  set $\eta_{k+1} \leftarrow \min(\frac{1}{1+2c_k}, \eta_{\max})$ {prediction}
end if

To avoid an excessive forcing term decrease in the initial phase of the global convergence we rely on the maximum admissible forcing term value $\eta_{\max}$. As a consequence of small timesteps and loose values of $\eta_{\max}$ just a few linear solver iterations are usually required in the first ten Ψ_{tc} iterations. In order to obtain a smoother behavior of the sequence η_k and possibly account

for residual increase the correction $\bar{c}_k$ is deferred based on the previous iterate. Negative $\bar{c}_k$ values are admitted whenever $||F(x_k)|| < ||F(x_k + s_k)|| < 1.2\,||F(x_k)||$, while the practical safeguard on the prediction ensures that $\eta_k <= \eta_{\max}$. The deferred correction also avoids a sudden increase of η_{k+1} in case of linear convergence failure, that is the inability to solve the modified Newton equation using the prescribed η.

Linear convergence failure is common practical issue. As remarked in Section 5, the ease of solving (44) as compared to (35) when the timestep is small, is a big advantage of Ψ_{tc}. As opposite, in the *terminal* phase, the amount of work required to satisfy condition (44) increases dramatically, being comparable to the expense required to satisfy condition (35) in inexact Newton methods. As a result, in the late midrange and terminal phase of the convergence, the number of iterations performed by the linear solver typically increases and the solver might fail to converge within the (user defined) maximum number of linear iterations. As detailed in Section 5.1 the occurrence of linear convergence failure is acknowledged without discarding the solution increment in order to optimize the computational effort.

5.4. Inexact pseudo-transient continuation for incompressible fluid flow problems

In order to demonstrate the effectiveness of the proposed forcing term strategy we apply the inexact pseudo-transient continuation Algorithm 5.1 to find steady state numerical solutions of convection dominated incompressible fluid flow problems. In particular we deal with the well known lid-driven cavity and backward-facing step problems at high-Reynolds numbers in two space dimension.

In the context of steady-state incompressible flow computations inexactly solving the Newton equations can be considered as an effective means of introducing artificial compressibility at the discrete level and alleviate the expense of solving saddle point problems. Indeed, since at each pseudo transient continuation iteration the continuity equation is approximatively solved, the incompressibility constraint is not fulfilled and only the final (steady state) solution is truly divergence free. Overall the computational burden associated to the solution process is strongly reduced without impacting the final accuracy.

To spatially discretize the Incompressible Navier-Stokes (INS) equations we rely on the dG formulation proposed by Bassi et al. [4]. The method relies on a local artificial compressibility perturbation of the equations to compute the inviscid numerical fluxes and provides a convergence rate of h^{k+1} and h^k for the velocity and pressure error in L^2 norm, respectively, when using a polynomial expansion of degree k. The accuracy of the scheme was

checked comparing with accurate reference computations of the lid-driven cavity problem available in literature, see [4] for details.

To solve the modified Newton equation (41) resulting from the application of the discrete dG space operators we use an ILU preconditioned GMRES(i) solver [27] restarted after i iterations and we set the maximum number of linear iteration is set to $2i$. If the solver fails to converge within this limit convergence failure is detected and we set $i \mathrel{+}= 20$, starting from $i = 120$ at $k = 0$. This practice allows to deal with stringent termination conditions, permitting to reach optimal convergence rates in the terminal phase.

The choice of the preconditioner side deserves specific attention. Let's consider the Newton equation $F^{'}(x)s = -F(x)$, where the termination condition for the linear solver reads

$$\left|\left|F^{'}(x)s + F(x)\right|\right| = ||R(x)|| \leq \eta \, ||F(x)|| \,, \tag{47}$$

see Section 2 for details. While a left preconditioned GMRES solver minimizes $P_l^{-1}R(x)$, where P_l^{-1} is the left preconditioner matrix, a right preconditioner minimizes the linear system residual $R(x) = F(x) + F^{'}(x)P_r^{-1}P_r s$. Left preconditioning is usually cheaper but is clearly not compatible with condition (47). The forcing term choice here proposed requires a relative residual based stopping criteria which in turn requires right preconditioning. In order to fairly compare with fixed forcing terms strategy, lid-driven computations are performed considering both left and right preconditioning options.

Global convergence of the sequence x_k is detected if $||x_{k+1} - x_k|| < 10^{-11}$ or $F(x_k) < 10^{-11}$. These tight tolerances allows to analyze all the convergence phases, in particular the terminal one could not have been observed with looser conditions. All the simulations but the higher-order computation of Section 5.6 are performed in serial. All the runs are performed and profiled on a Intel Xeon workstation.

5.4.1. Lid-driven cavity

Many publications have demonstrated the possibility to obtain accurate solutions of the lid-driven cavity problem at high-Reynolds numbers, see *e.g.* [4], [5], [16] and [31]. This model problem is widely employed to test the approximation capabilities of numerical schemes due to the simplicity of the computational domain and the complexity of the flow structures. The achievement of steady state solutions of lid-driven cavity is challenging from the marching strategy viewpoint as the counter-rotating vortices localized at the corners of the cavity and the main vortex itself need time to develop and reach a stable equilibrium. Clearly the higher the Reynolds number the

higher the flow complexity and the difficulty to reach a steady state solution. We remark that the 2D lid-driven cavity problem was also included in the test suite of Eisenstat and Walker, see [14].

For all the computations we employ second degree polynomial expansions for both the velocity and the pressure unknown over a fine 100^2 quadrilateral elements grid. The computational domain is the unit square.

To demonstrate the effectiveness of our forcing term choice coupled with SER based pseudo-time stepping we compare it with many fixed forcing term computations and we evaluate the residual decrease versus the number of Newton iteration and the simulation time. Thus, in order to compare with the adaptive forcing term strategies proposed in literature, we consider the forcing term choices presented in Section 3 and use them in combination with Algorithm 5.1.

Comparison with fixed forcing term choices

For the sake of comparison we consider the lid-driven cavity problem at three Reynolds numbers $\{10^4, 2\cdot 10^4, 3\cdot 10^4\}$, and in addition to the variable η strategy here proposed, we simulate 4 to 5 fixed η choices for each Reynolds number. Besides freezing η and possibly changing the preconditioner side, the B-SER algorithm for time step choice, see Section 5.2, the Ψ_{tc} marching strategy, see Section 5.1, and the linear solver options are kept unchanged. In all the computations consider fluid at rest as initial guess and start with a pseudo timestep $\delta_0 = 0.1$. As for the variable η strategy we set $\eta_{\max} = 0.9$.

The η, δ and $||F(x)||$ values at each Ψ_{tc} iteration, and the number of Krylov Spaces of the restarted GMRES solver (that is the number i of iteration before restart), are represented in Figure 3. Each increase in the number of Krylov spaces is an indication of linear convergence failure, and it is possible to appreciate that higher-Reynolds number computations require an higher number of Krylov Spaces, more linear iterations, and also more outer Ψ_{tc} iterations. In all the computations the final timestep grows above 10^{10} and the final residual is below 10^{-11}, while the forcing term is below 0.1. It is possible to appreciate that almost each iteration of the terminal phase is associated with linear convergence failure and triggers an increase of the number of Krylov Spaces. As a consequence smaller values of η would not have improved convergence rates.

Pseudo-transient continuation in combination with Algorithm 5.3 for forcing term choice (variable eta) yields smaller execution times than any fixed η choice here considered, see Figure 4 and Figure 5 for right and left preconditioning options, respectively. The first three charts (top row and bottom row left, one for each Reynolds number) compare the residual history for all

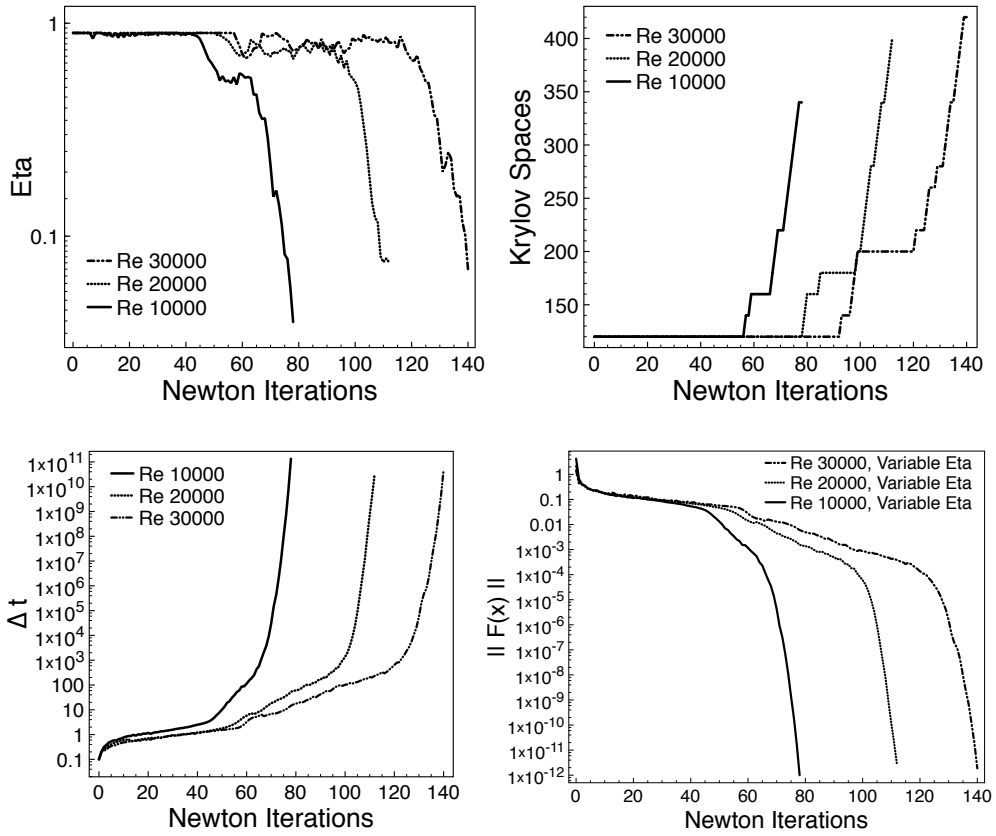

Figure 3: Steady state solution of the lid-driven cavity problem at high Reynolds number (Re) by means of a pseudo-transient continuation algorithm with adaptive forcing term choice (Variable Eta). First row, forcing term history and number of Krylov spaces (ILU precondtioned GMRES linear solver) history. Second row, timestep history and residual history.

the η choices, while the last chart (bottom row right) compares the variable η strategy with the best performing fixed η choice. It is interesting to remark that in case of right preconditioning the best performing fixed forcing term is $\eta = 0.5$ for all the Reynolds numbers here considered, the effectiveness of this choice was also mentioned by Kelley in [20]. The gains of variable η compared to $\eta = 0.5$ ranges from 3% at Reynolds 10^4, up to 17% at Reynolds $3 \cdot 10^4$.

In case of left preconditioning the linear system residual is polluted by the preconditioner, as a result it is harder to pick a value for η. After some attempts we identified an effective η range one order of magnitude smaller than in case of right preconditioning. As opposite to right preconditioning the best performing fixed η depends on the Reynolds number. The gains ranges from 1% at Reynolds 10^4 (fixed $\eta = 0.1$), up to 8% at Reynolds $3 \cdot 10^4$

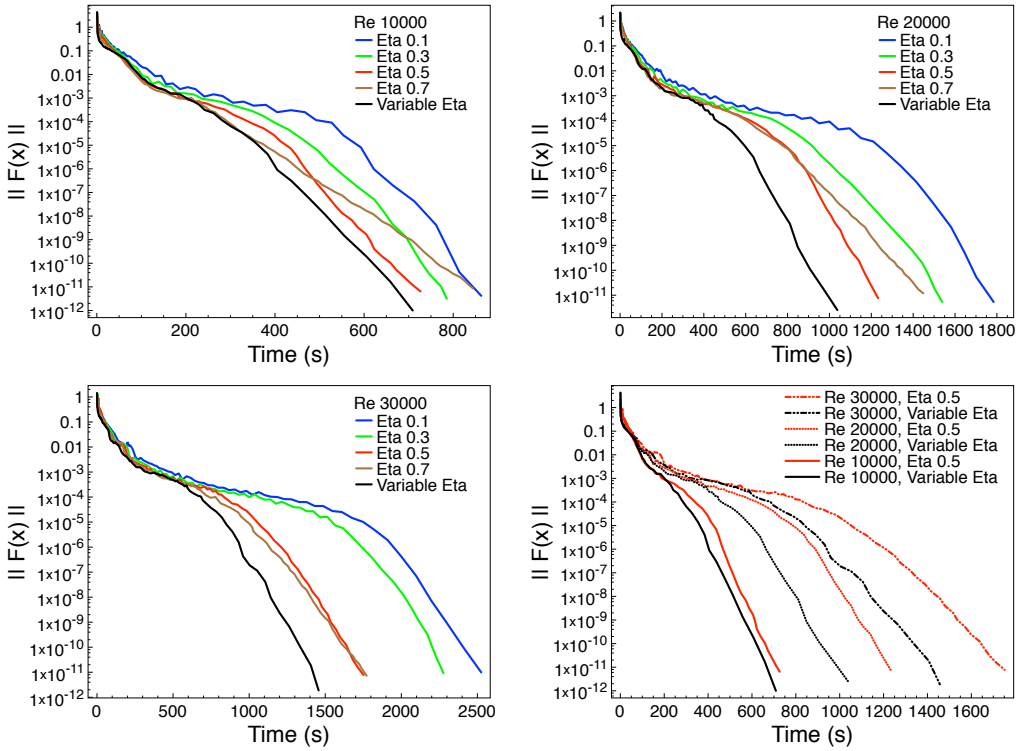

Figure 4: Steady state solution of the lid-driven cavity problem at high Reynolds number by means of a pseudo-transient continuation algorithm with inexact right preconditioned GMRES linear solver. Residual versus computation time for fixed and adaptive forcing terms (Eta) choices. First row, Re=10000 and Re=20000. Second row, Re=30000 and summary of best results.

(fixed $\eta = 0.075$). Even if the benefits seem limited one has to consider that the best fixed η is not known a priori and small deviations from the optimal value might cause significant performance penalty. For example at Reynolds $3 \cdot 10^4$, compared to $\eta = 0.05$ and $\eta = 0.1$ the gains are 18% and 50%, respectively, in favor of the adaptive algorithm, see Figure 5.

In Figure 6 we also compare the effects of η on the number of Ψ_{tc} iterations considering right and left preconditioning. Figure 6 shows that the variable η choice yields a number of global iterations comparable with the tightest tolerance $\eta = 0.1$, meaning that oversolving is avoided without impacting the residual decrease. Furthermore at Reynolds 30000 the adaptive η strategy yields the smallest number of Ψ_{tc} iterations indicating that forcing the inner loop according to the confidence in the linear model for $F(x)$ might also improve the outer loop.

It is possible to appreciate that $\eta = 0.5$, the best fixed η in terms of execution time, leads to a q-linear convergence of the residual in the terminal

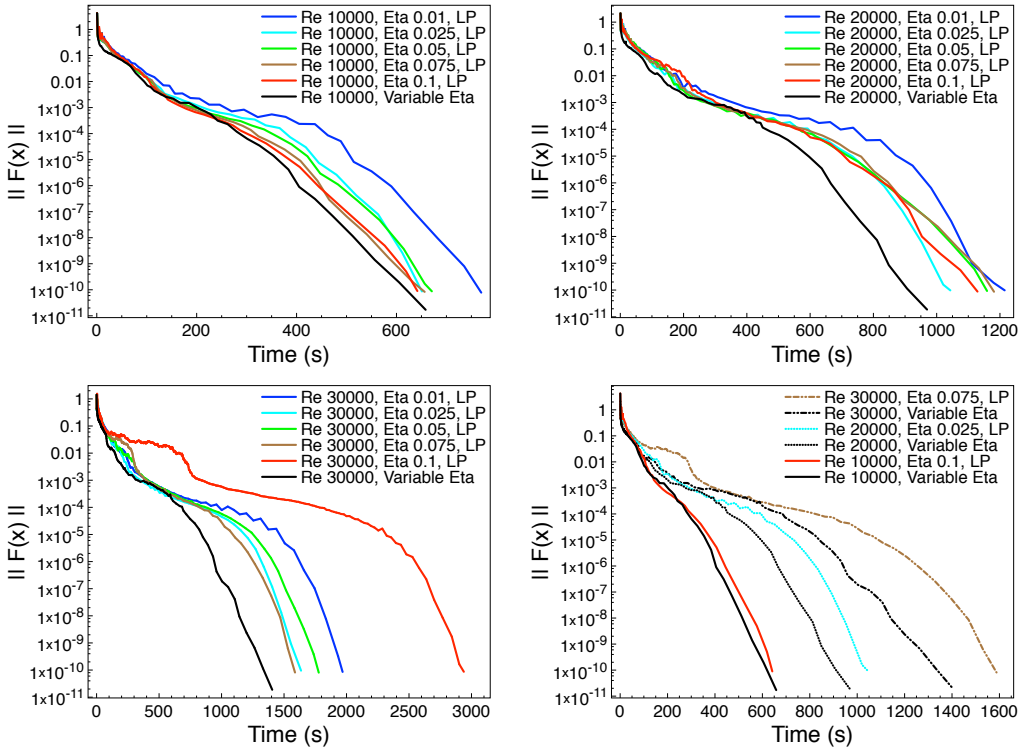

Figure 5: Steady state solution of the lid-driven cavity problem at high Reynolds number by means of a pseudo-transient continuation algorithm with inexact left preconditioned (LP) GMRES linear solver. Residual versus computation time for fixed and adaptive forcing terms (Eta) choices. First row, Re=10000 and Re=20000. Second row, Re=30000 and summary of best results.

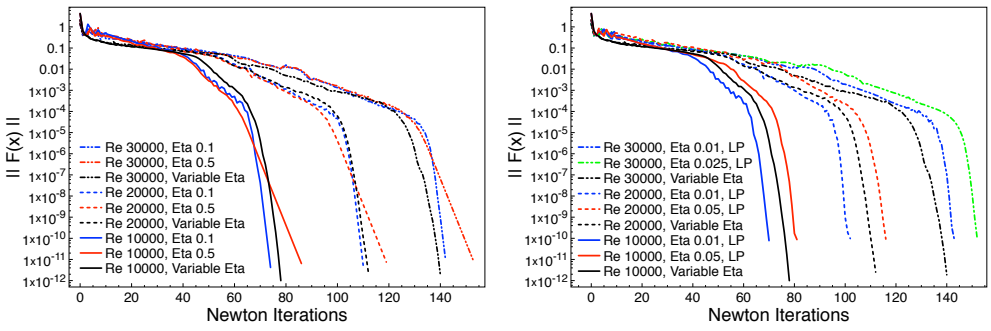

Figure 6: Steady state solution of the lid-driven cavity problem at high Reynolds number by means of a pseudo-transient continuation algorithm with inexact ILU preconditioned GMRES linear solver. Residual history for fixed and adaptive forcing terms (Eta) choices. Left and right, fixed eta computation are performed with right and left preconditioner options, respectively.

phase. Indeed avoiding oversolving in the initial and midrange phases might be more important than obtaining the theoretical convergence rate in the terminal phase. This is confirmed by the execution time associated with $\eta = 0.7$ at the highest Reynolds. Indeed $\eta = 0.7$ is very close to beating the $\eta = 0.5$ run despite the poor terminal phase performance. Clearly, while a fixed η choice must strike a balance between oversolving and optimal convergence the adaptive strategy has the possibility to prevent the former and achieve the latter.

Despite the poor performance in terms of execution times, see Figure 5, choosing a tight fixed $\eta = 0.01$ in combination with left preconditioning seems to reduce the number of Ψ_{tc} iterations at the lower Reynolds numbers, see Figure 6. As opposite, at the highest Reynolds number, the adaptive strategy allows to complete the computation in less outer iterations. Figure 5 allows to appreciate that, due to use of a left preconditioner, the linear solver is not able to drive the global residual norm below 10^{-11}, while the preconditioned residual is reduced up the numerical precision. As a consequence the iteration would stagnate around 10^{-11} and possibly terminate due to satisfaction of the terminating condition on the solution increment, that is $||x_{k+1} - x_k|| <= 10^{-11}$.

It is interesting to remark that tight η values might induce small high-frequency residual oscillations that are taken into account gracefully without requiring backtracking by admitting a 20% residual increase, see Section 5.1. On the other hand the oscillation are almost completely healed by the adaptive η strategy.

Comparison of adaptive forcing term strategies

In this section we consider the application of all the adaptive forcing term strategy of Section 3 (but we avoid EW1a (11) which behaves similarly to to (12) but is less trivial to implement, see also [14]) to pseudo-transient continuation. We complete the definition of each strategy by setting the relevant parameters and we use the almost same notation of Section 4.1 for reporting the results.

1. New, the new strategy in (15) with $\alpha = 2$;
2. EW1, the second variant of the first strategy given by Eisenstat and Walker, see (12);
3. EW2, the second strategy given by Eisenstat and Walker with $\alpha = 2$ and $\gamma = 1$, see (13);
4. AML, the strategy devised by An, Mo and Liu with $p1 = 0.1$, $p2 = 0.4$ and $p3 = 0.7$, see (14);

All the strategies can be easily applied to pseudo-transient continuation considering $R_{\Psi_{tc}}$ in place of R in the definition of $pred_k(s_k)$, see (6). Note that we also consider the New strategy as is, in order to outline the improvements obtained by the deferred correction introduced in Algorithm 5.3 (Variable Eta in the charts).

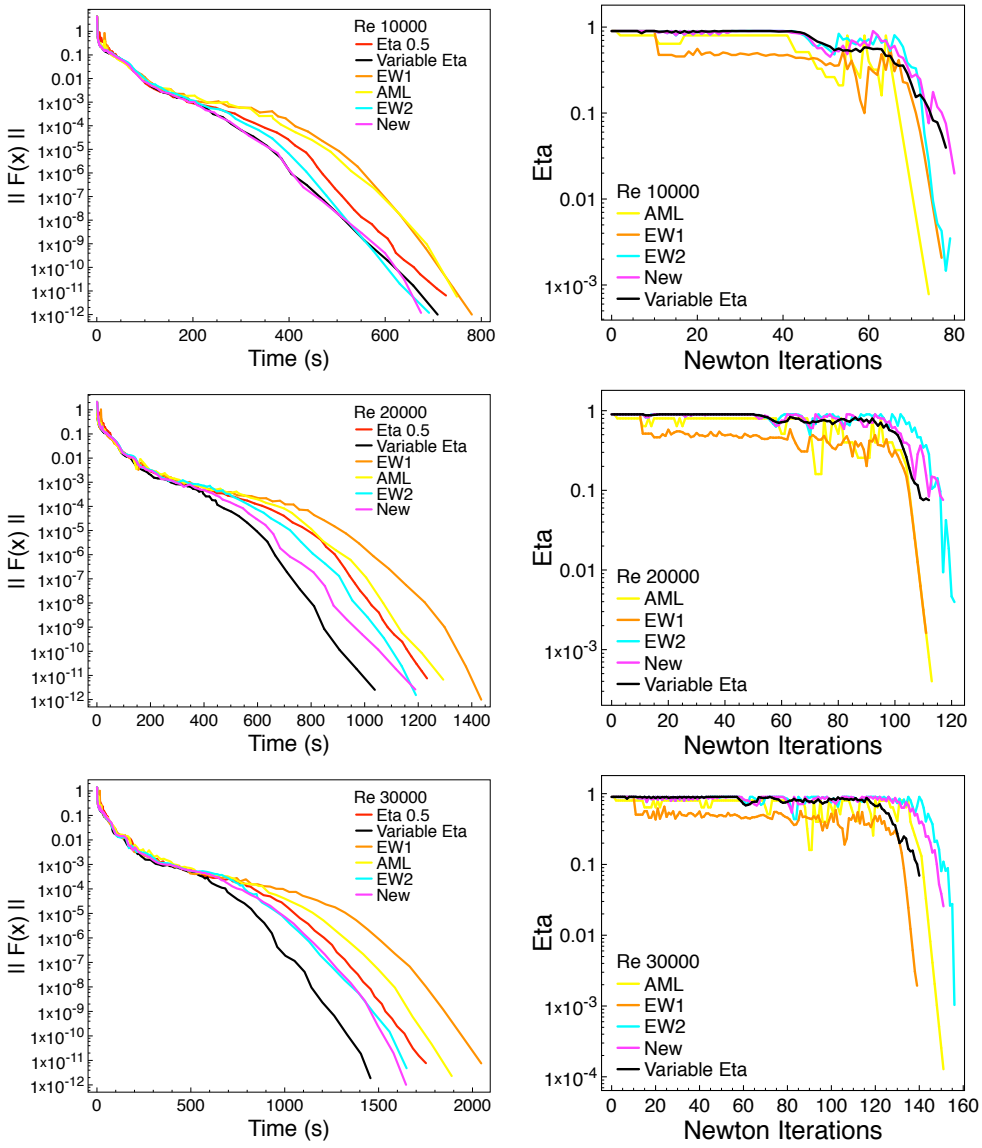

Figure 7: Steady state solution of the lid-driven cavity problem at high Reynolds number by means of a pseudo-transient continuation algorithm with inexact right preconditioned GMRES linear solver. Left, residual versus computation time for different adaptive forcing terms (Eta) choices. Right, η history. First, second and third row, Re=10000, Re=20000 and Re=30000, respectively.

Eta choice	final Krylov space dimension (linear convergence failures)			total Ψ_{tc} iterations		
	Re=10K	Re=20K	Re=30K	Re=10K	Re=20K	Re=30K
AML	360 (12)	440 (16)	520 (20)	74	113	151
EW1	340 (11)	440 (16)	520 (20)	77	111	139
EW2	360 (12)	380 (13)	480 (18)	79	121	156
New	320 (10)	400 (14)	440 (16)	80	117	151
Variable Eta	340 (11)	400 (14)	420 (15)	78	112	140
$\eta = 0.1$	320 (10)	460 (17)	560 (22)	74	110	142

Table 3: Inexact pseudo-transient continuation with SER based timestep choice, see algorithms 5.1 and 5.2, applied to the lid-driven cavity problem. Comparison of different forcing term choices. First column, GMRES search directions at the last Ψ_{tc} iteration and number of linear convergence failures (each linear convergence failure triggers an increase of the number of Krylov spaces, see text for details). Second column, total number of Ψ_{tc} iterations required to achieve a steady state solution.

Since we verified that the practical safeguards of Section 4.1 are mostly detrimental in the context of pseudo-transient continuation, in order to avoid an excessive η decrease at the early stages of the convergence we fix $\eta_k = \eta_{\max}$ for all $k < 10$, see also Algorithm 5.3. We set $\eta_{\max} = 0.9$ for all but AML, where for consistency we impose $\eta_{\max} = 1 - 2p_1 = 0.8$.

In Figure 7 it is possible to appreciate that only EW2, New, and the Variable Eta algorithm designed for pseudo-transient continuation perform better than the optimal fixed forcing term choice $\eta = 0.5$, see also Figure 4. In particular EW2 and New provide comparable execution times thanks to a similar behavior of forcing terms in the initial and midrange phases of the global convergence. As opposite, in the terminal phase, EW2 requires tighter termination conditions with EW1 and AML being even more restrictive. As remarked by Gropp, Keyes, McInnes, and Tidriri [18], the Eisenstat and Walker forcing term choices, although attractive from the convergence properties viewpoint, might be difficult to meet in practice in large, ill-conditioned problems. In particular EW1 achieves a steady state solution with lesser Ψ_{tc} iteration as compared to the others adaptive forcing term choices but requires additional efforts to the linear solver, as can be appreciated in Table 3. Note that a comparable occurrence of linear convergence failure is obtained choosing $\eta = 0.1$.

Overall, the results confirm that New and EW2 allows to optimize the computational cost associated with the inner iteration, possibly increasing the number total number of Ψ_{tc} iterations and, consequently, the number of Jacobian matrix assemblies. This strategy seems to pay off in the context of incompressible fluid flow simulations as the cost of the inner iterations usually

dominates the matrix assembly cost. Nevertheless, when dealing with other CFD applications, and in particular with compressible fluids, the relative weight of inner solves and matrix assemblies on the computation time must be carefully evaluated. If the latter dominates the former EW1 and AML might be more appropriate than New and EW2 and, in general, the benefits of an adaptive forcing term strategy might be less significant.

5.5. Backward-facing step

In addition to the lid-driven cavity flow also the backward-facing step problem has been widely employed as a benchmark for the validation of INS solvers. Albeit the simple geometry, the complex flow features associated with flow separation are fully retained. High Reynolds numbers in two space dimension have been considered by Erturk [15], where accurate numerical solution up to Reynolds 3000 were presented, and Cruchaga [10], who obtained steady state solution up to Reynolds 5500.

The achievement of the steady state solutions of the backward-facing step problem at high Reynolds number is well suited to challenge pseudo-transient continuation combined with the newly introduced adaptive forcing term strategy. Indeed, the impulsive start from flow fluid at rest needs to be dealt with and the pseudo-time integration must be conducted till several recirculating regions develop downstream from the step. The recirculating regions alternates, one at the lower and one at the upper wall of the channel, forcing the bulk of the flow to separate, cross the channel centerline and reattach, see Figure 11.

For all the computations we employ second degree polynomial expansions for both the velocity and the pressure unknown over a uniform 30K quadrilateral elements grid. The computational domains follows the recommendations of Erturk [15], in particular the inlet channel is 20 step heights long while the outflow is located 200 step heights away from the step. We impose a fully developed Dirichlet boundary condition at the inflow and a stress-free boundary condition at the outflow, while a no-slip boundary condition is imposed at the channel walls. The Reynolds number is defined as $Re \stackrel{\text{def}}{=} \frac{\overline{U}\,2h}{\nu}$, where $\overline{U} = \frac{2}{3}U_{\max}$ is the mean inlet velocity (two-thirds of the inlet channel centerline velocity) and h is the inlet channel height (as well as the step height). We simulate Re=2500 and Re=5000 setting the viscosity so to obtain a unit $U_{\max}$.

To demonstrate the effectiveness of our forcing term choice coupled with SER based pseudo-time stepping we compare it with many fixed forcing term computations and we evaluate the residual decrease versus the number of Newton iteration and the simulation time. In addition to the variable η strategy here proposed, we simulate several fixed η choices for each Reynolds

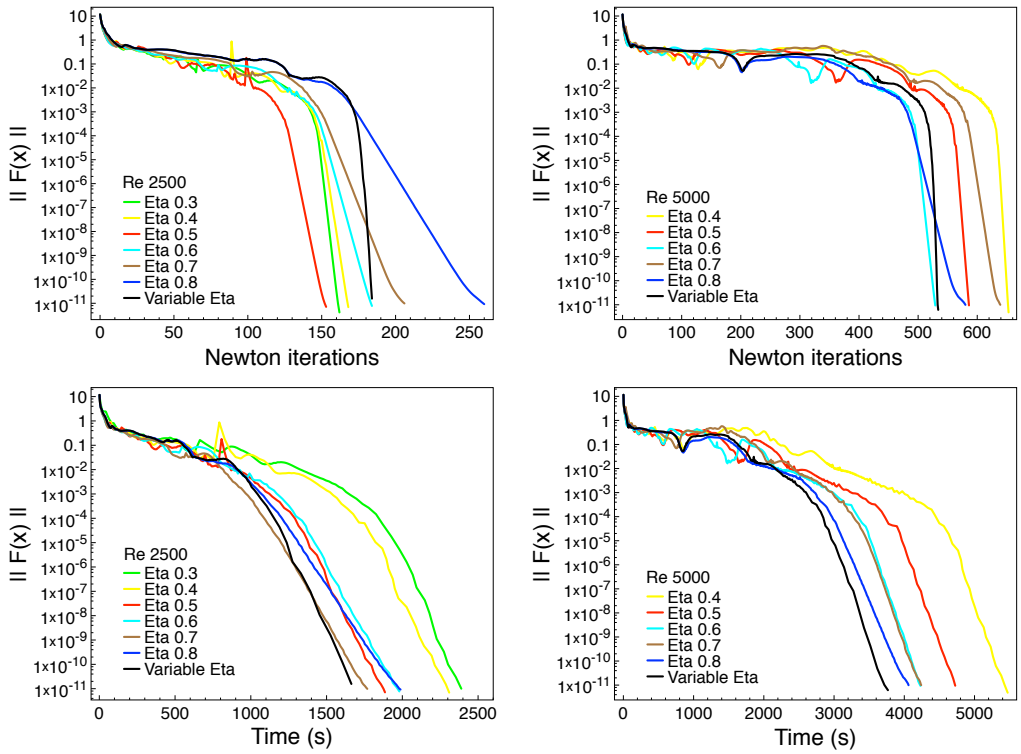

Figure 8: Steady state solution of the backward-facing step problem at high Reynolds number by means of a pseudo-transient continuation algorithm with inexact right preconditioned GMRES linear solver. First row and second, residual history and residual versus computation time for fixed and adaptive forcing terms (Eta) choices.

number. In all the computations we consider fluid at rest as initial guess and we start with a pseudo timestep $\delta_0 = 0.1$. As for the variable η strategy we set $\eta_{\max} = 0.8$.

Once again pseudo-transient continuation in combination with Algorithm 5.3 (Variable Eta) yields smaller execution times than any fixed η choice here considered, see Figure 8. The fastest fixed eta choices are $\eta = 0.7$ and $\eta = 0.8$ at Reynolds 2500 and 5000, respectively. The gains in terms of computation times achieved with the adaptive forcing term strategy are less significant than in the lid-driven cavity case due to the expense of solving the modified Newton equation in the terminal phase of the global convergence. This is confirmed by the fact that the best performance are achieved with looser fixed forcing terms as compared with the lid-driven cavity case (where $\eta = 0.5$ is the gold standard). Nevertheless the reduction in terms of total GMRES iterations (12% at Reynolds 5000), confirms the ability to avoid *oversolving* while maintaining good local convergence properties in the terminal phase,

Eta choice	total GMRES iterations	
	Re=2500	Re=5000
$\eta = 0.3$	12639	
$\eta = 0.4$	12089	23960
$\eta = 0.5$	10233	21216
$\eta = 0.6$	9501	19527
$\eta = 0.7$	8556	17115
$\eta = 0.8$	8659	17677
Variable eta	8015	15539

Table 4: Total number of GMRES iterations required to achieve a steady state solution of the backward-facing step problem. Inexact pseudto-transient continuation with SER based timestep choice, see algorithms 5.1 and 5.2.

see Table 4.

Looking at Figure 8 it is clear that the adaptive forcing term strategy yields smoother residual histories reducing the number of backtracking iterations (identifiable as sudden spikes in the residual history). As for the lid-driven cavity case, the adaptive forcing terms strategy provides a very competitive number of outer iterations, especially at the highest Reynolds numbers. This suggests that forcing the inner loop according to the confidence in the linearisation of $F(x)$ allows to avoid *oversolving* without impacting the residual decrease in the outer loop.

5.6. Higher-order accurate computations

To conclude we also applied the proposed time marching strategy (pseudo-transient continuation with SER-based time stepping and adaptive forcing term choice) for the achievement of higher-order accurate steady state solutions of the lid-driven cavity and backward-facing step problems. The lid-driven cavity computation is performed at Reynolds $5 \cdot 10^4$ employing a sixths polynomial degree dG discretization. The backward-facing step is computed at Reynold 5000 by means of a forth polynomial degree dG discretization. The same 100^2 quadrilateral grid of the unit square and 30K quadrilateral grid are employed for the lid-driven cavity and backward-facing step, respectively. The computations are initialized with fluid at rest.

To solve the modified Newton equation (41) resulting from the application of the discrete dG space operators we use a GMRES(i) solver [27] and we rely on the same solver options employed for serial runs. The computations are performed in parallel dividing the computational domain in eight subdomains. The linear solver is preconditioned with an overlapping Additive Schwartz Method (ASM) setting one level of overlap between the

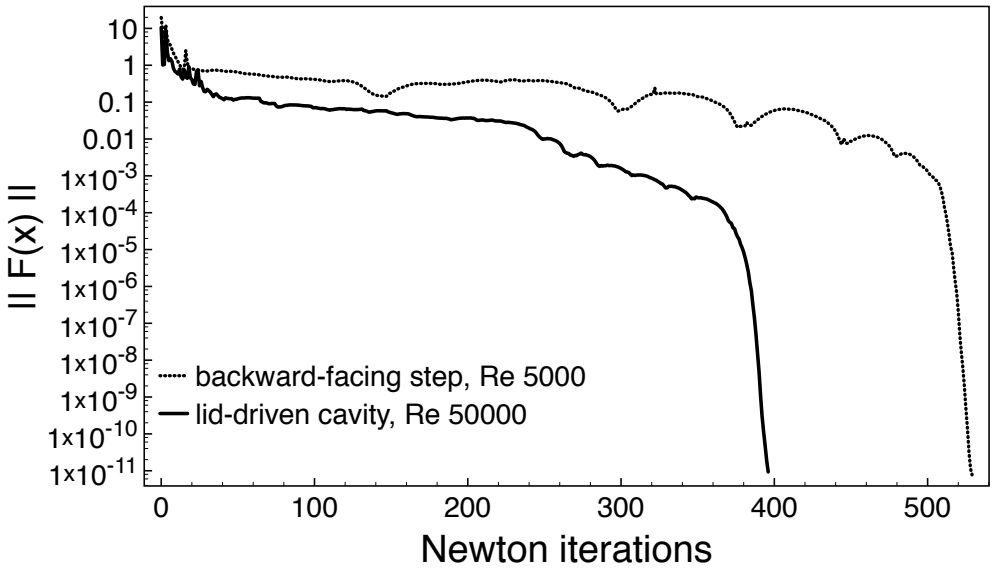

Figure 9: Residual history for the steady state solution of the lid-driven cavity problem (Re 50000, sixth degree dG discretization) and backward-facing step problem (Re=5000, fourth degree dG discretization).

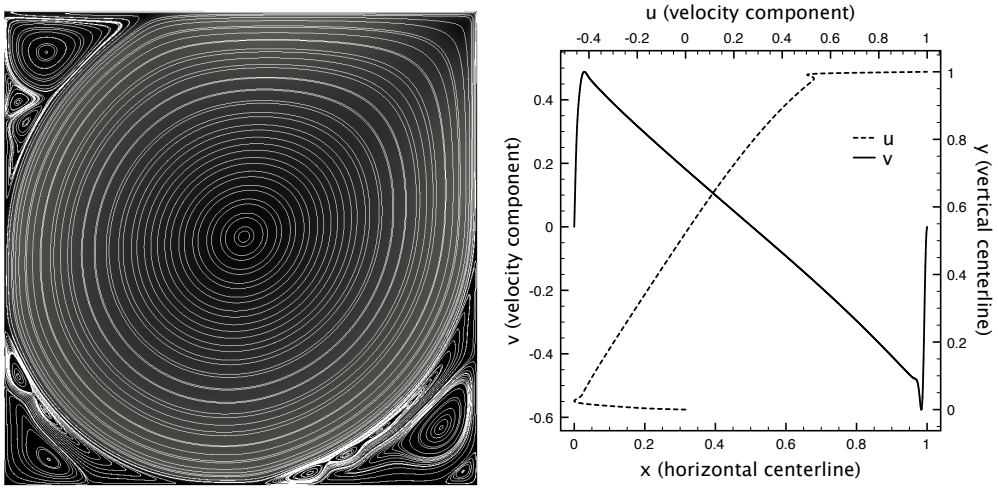

Figure 10: Steady state solution of the lid-driven cavity problem at Re=50000. Sixth polynomial degree dG discretization. Left, streamlines and velocity solution. Right, velocity solutions along the vertical and horizontal centerlines.

subdomains and using an ILU decomposition for each subdomain matrix (point-block ILU).

The residual history is shown in Figure 9. The Ψ_{tc} algorithm continuation algorithm converges in approximatively 400 and 500 iterations for the

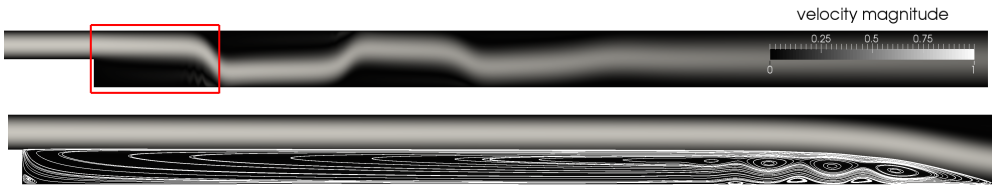

Figure 11: Steady state solution of the backward-facing step problem at Re=5000. Fourth polynomial degree dG discretization. Top, velocity solution in the whole computational domain (the axial coordinate is scaled 1:6.25 for the sake of visualization). Bottom, streamlines in the recirculation region located after the step (outlined by the red box in the top figure).

lid-driven cavity and backward-facing step problems, respectively, with a final residual of 10^{-11}. The streamlines and the velocity solutions across the centerlines of the cavity represented in Figure 10 allows to appreciate the complexity of the flow field and the strong velocity gradients in the boundary layers. Analogously the velocity solution and the streamlines reported in Figure 11 allows to appreciate the succession of recirculating bubbles in the channel and the complexity of the vortex structures in the first recirculating region. While some authors failed to achieve steady state solutions of the lid-driven cavity flow at high-Reynolds, see *e.g.* [16, 19, 31], the pseudo time marching strategy here proposed allows to integrate high-order dG discretizations up to steady state. Moreover the coupling of the time marching strategy with an adaptive forcing term choice lower the computational costs providing a satisfactory smooth residual history.

6. Conclusion

We devised an adaptive strategy for choosing the forcing terms in globalisation strategies. The local convergence properties are demonstrated in the context of inexact Newton and the behavior of the strategy at early stages of the convergence is analyzed with numerical test cases. The results obtained on model nonlinear systems proposed in literature as benchmark problems were encouraging.

The new strategy is adapted to a backtracked pseudo-transient continuation strategy for the computation of steady state solution of the INS equations at high-Reynolds numbers. In pseudo-transient continuation particular care must be devoted to the choice of forcing terms since nonlinear convergence is highly influenced by the timestep choice in all but the terminal phase. As a consequence the forcing term strategy must react as smoothly as possible to linear model improvements to be efficient in practice. Good results are

obtained applying pseudo-transient continuation in combination with selective evolution relaxation time stepping and the newly introduced adaptive forcing terms strategy. In particular, the cost of computing steady state numerical solutions of the lid-driven cavity flow and the backward-facing step at high-Reynolds number is reduced as compared with the best performing fixed forcing term choices. The increased efficiency is related to the decrease of total inner solver iterations which confirms the effectiveness of the proposed approach. The residual history is smooth and also the number of Newton iterations is comparable with those obtained setting tight tolerances on the linear solver convergence.

Appendix A. Proof of Proposition 1

Clearly $F(x_*) = 0$. Set $\beta = \left|\left|F^{'}(x_*)^{-1}\right|\right|$, for any $\sigma \in (0, 1]$ there exist $\delta > 0$ sufficiently small that

$$\left|\left|F(x) - F(x_*) - F^{'}(x_*)(x - x_*)\right|\right| \leq \frac{\sigma}{2\beta} ||x - x_*|| ,$$

whenever $x \in N_\delta(x_*)$. Such a δ exists by Lemma 3.2. If $x \in N_\delta(x_*)$ then

$$\begin{aligned} ||F(x)|| &\geq \left|\left|F^{'}(x_*)(x - x_*)\right|\right| - \left|\left|F(x) - F(x_*) - F^{'}(x_*)(x - x_*)\right|\right| \\ &\geq \frac{1}{||F^{'}(x_*)^{-1}||} ||x - x_*|| - \frac{\sigma}{2\beta} ||x - x_*|| \\ &= \frac{2 - \sigma}{2\beta} ||x - x_*|| \\ &\geq \frac{1}{2\beta} ||x - x_*|| , \end{aligned}$$

so that

$$||x - x_*|| < 2\beta ||F(x)|| . \tag{A.1}$$

Moreover

$$\begin{aligned} \left|\left|F^{'}(x_*)(x - x_*)\right|\right| &\leq ||F(x)|| + \left|\left|-F(x) + F(x_*) + F^{'}(x_*)(x - x_*)\right|\right| \\ &\leq ||F(x)|| + \frac{\sigma}{2\beta} ||x - x_*|| \qquad (A.2) \\ &\leq ||F(x)|| + \frac{\sigma}{2\beta} 2\beta ||F(x)|| \\ &= (1 + \sigma) ||F(x)|| , \end{aligned}$$

where we used (A.1) in (A.2), and

$$\begin{aligned}
||F(x)|| &\leq \left|\left|F^{'}(x_*)(x-x_*)\right|\right| + \left|\left|F(x)-F(x_*)-F^{'}(x_*)(x-x_*)\right|\right| \\
&\leq \left|\left|F^{'}(x_*)(x-x_*)\right|\right| + \frac{\sigma}{2\beta}\,||x-x_*|| \\
&\leq \left|\left|F^{'}(x_*)(x-x_*)\right|\right| + \frac{\sigma}{2\beta}\left|\left|F^{'}(x_*)^{-1}\right|\right|\left|\left|F^{'}(x_*)(x-x_*)\right|\right| \\
&\leq \left|\left|F^{'}(x_*)(x-x_*)\right|\right| + \frac{\sigma}{2}\left|\left|F^{'}(x_*)(x-x_*)\right|\right| \\
&\leq (1+\sigma/2)\left|\left|F^{'}(x_*)(x-x_*)\right|\right|.
\end{aligned}$$

As a result

$$\frac{||F(x)||}{1+\sigma/2} \leq \left|\left|F^{'}(x_*)(x-x_*)\right|\right| \leq (1+\sigma)\,||F(x)||\,, \tag{A.3}$$

which proves Proposition 1 since $0 < \sigma \leq 1$.

[1] Ajmani, K., Ng, W.-F., and Liou, M.-S. Preconditioned conjugate gradient methods for the Navier-Stokes equations. *Journal of Computational Physics 110*, 1 (1994), 68 – 81.

[2] An, H.-B., Mo, Z.-Y., and Liu, X.-P. A choice of forcing terms in inexact Newton method. *Journal of Computational and Applied Mathematics 200*, 1 (2007), 47 – 60.

[3] Bassi, F., Botti, L., Colombo, A., Crivellini, A., Franchina, N., Ghidoni, A., and Rebay, S. Very high-order accurate discontinuous Galerkin computation of transonic turbulent flows on aeronautical configurations. *Notes on Numerical Fluid Mechanics and Multidisciplinary Design 113* (2010), 25–38. cited By (since 1996)16.

[4] Bassi, F., Crivellini, A., Pietro, D. D., and Rebay, S. An artificial compressibility flux for the discontinuous Galerkin solution of the incompressible Navier-Stokes equations. *Journal of Computational Physics 218*, 2 (2006), 794 – 815.

[5] Botti, L., and Pietro, D. A. D. A pressure-correction scheme for convection-dominated incompressible flows with discontinuous velocity and continuous pressure. *Journal of Computational Physics 230*, 3 (2011), 572 – 585.

[6] Brown, J. Efficient nonlinear solvers for nodal high-order finite elements in 3D. *Journal of Scientific Computing 45*, 1-3 (2010), 48–63.

[7] Brown, P., and Saad, Y. Hybrid Krylov methods for nonlinear systems of equations. *SIAM Journal on Scientific and Statistical Computing 11*, 3 (1990), 450–481.

[8] Chisholm, T. T., and Zingg, D. W. A Jacobian-free Newton-Krylov algorithm for compressible turbulent fluid flows. *Journal of Computational Physics 228*, 9 (2009), 3490 – 3507.

[9] Crivellini, A., and Bassi, F. An implicit matrix-free discontinuous Galerkin solver for viscous and turbulent aerodynamic simulations. *Computers & Fluids 50*, 1 (2011), 81 – 93.

[10] Cruchaga, M. A. A study of the backward-facing step problem using a generalized streamline formulation. *Communications in Numerical Methods in Engineering 14*, 8 (1998), 697–708.

[11] Dembo, R., Eisenstat, S., and Steihaug, T. Inexact Newton methods. *SIAM Journal on Numerical Analysis 19*, 2 (1982), 400–408.

[12] Dembo, R., and Steihaug, T. Truncated-Newton algorithms for large-scale unconstrained optimization. *Mathematical Programming 26*, 2 (1983), 190–212.

[13] Eisenstat, S., and Walker, H. Globally convergent inexact Newton methods. *SIAM Journal on Optimization 4*, 2 (1994), 393–422.

[14] Eisenstat, S., and Walker, H. Choosing the forcing terms in an inexact Newton method. *SIAM Journal on Scientific Computing 17*, 1 (1996), 16–32.

[15] Erturk, E. Numerical solutions of 2-D steady incompressible flow over a backward-facing step, part I: High Reynolds number solutions. *Computers & Fluids 37*, 6 (2008), 633 – 655.

[16] Erturk, E., Corke, T. C., and Gökçöl, C. Numerical solutions of 2-D steady incompressible driven cavity flow at high Reynolds numbers. *International Journal for Numerical Methods in Fluids 48*, 7 (2005), 747–774.

[17] Gee, M. W., Kelley, C. T., and Lehoucq, R. B. Pseudo-transient continuation for nonlinear transient elasticity. *International Journal for Numerical Methods in Engineering 78*, 10 (2009), 1209–1219.

[18] Gropp, W. D., Keyes, D. E., McInnes, L., and Tidriri, M. Globalized Newton-Krylov-Schwarz algorithms and software for parallel implicit CFD. *Int. J. High Performance Computing Applications 14* (1998), 102–136.

[19] Hachem, E., Rivaux, B., Kloczko, T., Digonnet, H., and Coupez, T. Stabilized finite element method for incompressible flows with high Reynolds number. *J. Comput. Phys. 229*, 23 (Nov. 2010), 8643–8665.

[20] Kelley, C. *Iterative methods for linear and nonlinear equations*. Society for Industrial and Applied Mathematics, 1995.

[21] Kelley, C., and Keyes, D. Convergence analysis of pseudo-transient continuation. *SIAM Journal on Numerical Analysis 35*, 2 (1998), 508–523.

[22] Li, G. Successive column correction algorithms for solving sparse nonlinear systems of equations. *Math. Program. 43*, 2 (Feb. 1989), 187–207.

[23] Lukšan, L. Inexact trust region method for large sparse systems of nonlinear equations. *J. Optim. Theory Appl. 81*, 3 (June 1994), 569–590.

[24] Mulder, W. A., and Van Leer, B. Experiments with implicit upwind methods for the Euler equations. *Journal of Computational Physics 59*, 2 (1985), 232–246.

[25] Nobile, F., Pozzoli, M., and Vergara, C. Inexact accurate partitioned algorithms for fluid-structure interaction problems with finite elasticity in haemodynamics. *Journal of Computational Physics 273*, 0 (2014), 598 – 617.

[26] Ortega, J. M., and Rheinboldt, W. C. *Iterative Solution of Nonlinear Equations in Several Variables*. Society for Industrial and Applied Mathematics, 2000.

[27] Saad, Y., and Schultz, M. H. GMRES: A Generalized Minimal Residual Algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput. 7*, 3 (July 1986), 856–869.

[28] Tidriri, M. Preconditioning techniques for the Newton-Krylov solution of compressible flows. *Journal of Computational Physics 132*, 1 (1997), 51 – 61.

[29] Toint, P. L. Numerical solution of large sets of algebraic nonlinear equations. *Math. Comput. 46*, 173 (Jan. 1986), 175–189.

[30] Venkatakrishnan, V., and Mavriplis, D. J. Implicit solvers for unstructured meshes. *Journal of Computational Physics 105*, 1 (1993), 83 – 91.

[31] Wahba, E. Steady flow simulations inside a driven cavity up to Reynolds number 35000. *Computers & Fluids 66*, 0 (2012), 85 – 97.