Termination criteria for inexact fixed point methods

Philipp Birken[1]

October 1, 2013

[1] *Institute of Mathematics, University of Kassel, Heinrich-Plett-Str. 40, D-34132 Kassel, Germany*
Department of Mathematics/Computer Science, University of Osnabrück, Albrechtstr. 28a, 49076 Osnabrück, Germany
email: birken@mathematik.uni-kassel.de

Abstract

We analyze inexact fixed point iterations where the generating function contains an inexact solve of an equation system. Important examples are the Picard iteration and partitioned fluid structure interaction. The iteration is modelled as a perturbed fixed point iteration and existing analysis is extended to the nested case $\mathbf{x} = \mathbf{F}(\mathbf{S}(\mathbf{x}))$. When applied to inexact fixed iterations, this allows to prove that the iteration converges irrespective of how accurate the equation systems are solved, provided that a specific relative termination criterion is employed. As demonstrated by numerical examples, standard relative and absolute criteria cause convergence to a different solution.

Keywords: Fixed point iteration, Picard iteration, Transmission Problem, Dirichlet-Neumann iteration, Termination criteria

1 Introduction

We consider fixed point iterations where the evaluation of the right hand side is perturbed. The motivation for this is that sometimes, the evaluation of the right hand side corresponds to solving a linear or nonlinear equation system. Two prominent and important examples are the Picard iteration and fluid-structure interaction. For the Picard iteration, which is a common tool in the context of the incompressible Navier-Stokes equations, the evaluation of the right hand side corresponds to solving a linear system. This is typically done by an iterative solver. Strategies for choosing a termination criterion for this are empirically discussed for example in [4, 8].

In Fluid-Structure interaction, a standard approach to solve these coupled problems are partitioned coupling schemes, where existing solvers for the subproblems are reused. Commonly in the form of a Dirichlet-Neumann iteration, this consists of subsequently solving the fluid and the structure problem with appropriate boundary conditions and reasonable tolerances. It is common to formulate the coupling condition at the interface in the form of a fixed point equation.

Recently, it was suggested to use a time adaptive implicit time integration scheme for fluid structure interaction [2], where the time step is chosen based on an error tolerance. As is common in this setting, the tolerances for the solvers for the appearing nonlinear equation systems are chosen such that the iteration error does not interfere with the error from the time integration scheme [1], but nevertheless as large as possible to avoid unnecessary computations. Often, the equation systems are solved using so called inexact Newton's methods, where the appearing linear systems are solved using an iterative solver, where again a tolerance is prescribed. A strategy that retains quadratic convergence of the scheme while solving the linear systems quite coarsely was suggested in [3]. There, the point is that the tolerance for the iterative linear solver has to converge to zero fast enough as the Newton scheme progresses.

In this article, the goal is to analyze what we call inexact fixed point schemes and to suggest reasonable strategies to prescribe tolerances of subsolvers. First, we present the existing analysis on perturbations and apply it to the Picard iteration. Then, we extend the analysis to the case of nested fixed point equations and apply that to the Dirichlet-Neumann iteration. In section three, numerical results are presented.

2 Inexact Fixed Point Methods

2.1 Direct perturbation

Consider the fixed point equation

$$\mathbf{x} = \mathbf{f}(\mathbf{x}) \tag{1}$$

with $\mathbf{x} \in \Omega \subset \mathbb{R}^n$, Ω closed and where we assume that $\mathbf{f} : \Omega \to \Omega$ is Lipschitz continuous with Lipschitz constant $L < 1$. This implies by the Banach fixed point theorem that (1) has a unique solution $\mathbf{x}^*$.

Furthermore, we consider the perturbed fixed point iteration

$$\mathbf{x}^{k+1} = \mathbf{f}(\mathbf{x}^k) + \epsilon, \tag{2}$$

where ϵ is a perturbation that could originate from an iterative solver and for simplicities sake we denote the norm of ϵ by ϵ as well. We furthermore

assume for simplicities sake that $\mathbf{f} + \epsilon$ is also a self-map on Ω. Thus this iteration obtains a solution $\mathbf{x}_\epsilon$ of the perturbed fixed point equation

$$\mathbf{x} = \mathbf{f}(\mathbf{x}) + \epsilon. \tag{3}$$

The question is now: How far is the solution $\mathbf{x}_\epsilon$ of that equation away from $\mathbf{x}^*$? The answer is giving by the following theorem, see for example [6].

Theorem 1 *For the solutions* $\mathbf{x}_\epsilon$ *and* $\mathbf{x}^*$ *of problems* (3) *and* (1) *we have:*

$$\|\mathbf{x}_\epsilon - \mathbf{x}^*\| \leq \epsilon \frac{1}{1-L}. \tag{4}$$

This means that the error is of the order ϵ as is to be expected, but interestingly, it becomes larger, the closer the Lipschitz constant of $\mathbf{f}$ is to one or otherwise put, the less contractive the function is. This implies that in these cases, the error will be much larger than ϵ and thus a much smaller tolerance would have to be supplied to acchieve the desired error. Note that the important Lipschitz constant is the local one in the solution.

If we instead consider a sequence of perturbations ϵ_k, respectively a nonconstant perturbation, and thus the iteration

$$\mathbf{x}^{k+1} = \mathbf{f}(\mathbf{x}^k) + \epsilon_k, \tag{5}$$

the first question is when we obtain convergence of this sequence to $\mathbf{x}^*$. The answer is given by the next theorem, also from [6]:

Theorem 2 *The iteration* (5) *converges to the solution of the unperturbed problem* (1) *if and only if* $\lim_{k\to\infty} \epsilon_k = 0$.

A specific case is

$$\epsilon_k = cL^k \tag{6}$$

with $c > 0$, which we call (6) the adaptive strategy.

2.2 Application: Picard iteration

As an application of the above theorems, we now analyze the convergence of the Picard iteration. This is often employed in the context of the incompressible Navier-Stokes equation and corresponds to a fixed point iteration for the equation

$$\mathbf{x} = \mathbf{A}^{-1}(\mathbf{x})\mathbf{b},$$

where $\mathbf{A}(\mathbf{x}) \in \mathbb{R}^{n\times n}$ is an approximation of a Jacobian in $\mathbf{x}$ [4]. Thus, the fixed point iteration

$$\mathbf{x}^{k+1} = \mathbf{A}^{-1}(\mathbf{x}^k)\mathbf{b} \tag{7}$$

is implemented by solving

$$\mathbf{A}(\mathbf{x}^k)\mathbf{x}^{k+1} = \mathbf{b} \tag{8}$$

for $\mathbf{x}^{k+1}$ up to a certain tolerance using an iterative scheme. The scheme (7) can be analyzed either as a fixed point scheme, which results in linear convergence provided that the Lipschitz constant L of $\mathbf{A}^{-1}(\mathbf{x})\mathbf{b}$ can be bounded from below away from one or as a method of Newton type where $\mathbf{A}(\mathbf{x})$ is an approximation of the exact Jacobian and we have linear convergence as long as this approximation is good enough.

When solving (8), either the relative termination criterion

$$\|\mathbf{A}(\mathbf{x}^k)\mathbf{x}^{k+1} - \mathbf{b}\| \leq \tau_r \|\mathbf{A}(\mathbf{x}^k)\mathbf{x}^k - \mathbf{b}\|, \tag{9}$$

the relative termination criterion

$$\|\mathbf{A}(\mathbf{x}^k)\mathbf{x}^{k+1} - \mathbf{b}\| \leq \tau_r \|\mathbf{b}\|, \tag{10}$$

or the absolute criterion

$$\|\mathbf{A}(\mathbf{x}^k)\mathbf{x}^{k+1} - \mathbf{b}\| \leq \tau_a \tag{11}$$

are used, where τ_r and τ_a are relative and adaptive tolerances.

To analyze the consequences of choosing one of these using theorems 1 and 2, we need to quantify the perturbation in the form (2). Thus, we define $\mathbf{f}(\mathbf{x}^k) = \mathbf{A}^{-1}(\mathbf{x}^k)\mathbf{b}$ to obtain

$$\mathbf{x}^{k+1} = \mathbf{A}^{-1}(\mathbf{x}^k)\mathbf{b} + \epsilon_k$$

and we can write

$$\epsilon_k = \mathbf{A}^{-1}(\mathbf{x}^k)(\mathbf{A}(\mathbf{x}^k)\mathbf{x}^{k+1} - \mathbf{b}). \tag{12}$$

In the case of the relative termination criterion (9), we can estimate the norm of the right hand side in (12) by

$$\|\epsilon_k\| \leq \|\mathbf{A}^{-1}(\mathbf{x}^k)\| \tau_r \|(\mathbf{A}(\mathbf{x}^k)\mathbf{x}^k - \mathbf{b})\|.$$

We furthermore have

$$\mathbf{A}(\mathbf{x}^k)\mathbf{x}^k - \mathbf{b} = \mathbf{A}(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{A}^{-1}(\mathbf{x}^k)\mathbf{b}) = \mathbf{A}(\mathbf{x}^k)(\mathbf{f}(\mathbf{x}^{k-1}) - \mathbf{f}(\mathbf{x}^k)).$$

Thus,

$$\|(\mathbf{A}(\mathbf{x}^k)\mathbf{x}^k - \mathbf{b})\| \leq \|(\mathbf{A}(\mathbf{x}^k)\| L \|\mathbf{x}^{k-1} - \mathbf{x}^k\| \leq \|(\mathbf{A}(\mathbf{x}^k)\| L^k \|\mathbf{x}^1 - \mathbf{x}^0\|.$$

All in all, we obtain with the condition number $\kappa(\mathbf{A})$

$$\|\epsilon_k\| \leq \underbrace{\tau_r \kappa(\mathbf{A}(\mathbf{x}^k)) L^k \|\mathbf{x}^1 - \mathbf{x}^0\|}_{\to 0} \tag{13}$$

which is a perturbation of the form (6), independent of the choice of τ_r! Thus, this iteration converges to the exact solution independently of how accurate we solve the linear equation systems with the only additional requirement that $\kappa(\mathbf{A}(\mathbf{x}))$ is bounded, which is reasonable in the first place. We now formulate this as a theorem.

Theorem 3 *Let $\mathbf{b} \in \mathbb{R}^n$ and the function $\mathbf{A}(\mathbf{x})$ be given that maps the closed set $\Omega \subset \mathbb{R}^n$ onto quadratic regular matrices. Assume that the function $\mathbf{A}^{-1}(\mathbf{x})\mathbf{b} : \Omega \to \Omega$ is Lipschitz continuous with Lipschitz constant $L < 1$ and correspondingly has a unique fixpoint $\mathbf{x}^*$. Furthermore assume that $\kappa(\mathbf{A}(\mathbf{x}))$ is bounded on Ω and that the inexact fixedpoint iteration defined by (9) converges to a limit $\mathbf{x}_\epsilon$. Then $\mathbf{x}_\epsilon = \mathbf{x}^*$, independent of the choice of τ_r.*

In case of the relative criterion (10), the estimate for the norm of the left hand side in (12)

$$\|\epsilon_k\| \leq \|\mathbf{A}^{-1}(\mathbf{x}^k)\|\tau_r\|\mathbf{b}\|$$

holds which is bounded away from zero provided that $\mathbf{A}^{-1}(\mathbf{x})$ is. Thus, it is not clear if this iteration satisfies theorem 2, but if the inequality is sharp, the iteration will not converge to $\mathbf{x}^*$. Similarly if we use the absolute termination criterion (11), we obtain

$$\|\epsilon_k\| \leq \|\mathbf{A}^{-1}(\mathbf{x}^k)\|\tau_a$$

which is also bounded away from zero.

Numerical results that confirm theorem 3 and demonstrate that the other two iterations behave like being of the form 1 can be found in section 3.1.2. We would like to point out that the criterion (9) is sometimes suggested in the literature on the Picard iteration, e.g. [5, 4], but that an absolute termination criterion is suggested in [8]. There it is suggested to just "gain one digit", meaning to use a tolerance of 0.1.

2.3 Perturbed nested fixed point iteration

Now consider two functions $\mathbf{F}$ and $\mathbf{S}$ and the fixed point equation

$$\mathbf{x} = \mathbf{S}(\mathbf{F}(\mathbf{x})) \tag{14}$$

again with solution $\mathbf{x}^*$. We now consider an iteration where both the evaluation of $\mathbf{F}$ and of $\mathbf{S}$ are perturbed, namely $\mathbf{S}$ is perturbed by δ_k and $\mathbf{F}$ by ϵ_k:

$$\mathbf{x}^{k+1} = \mathbf{S}(\mathbf{F}(\mathbf{x}^k) + \epsilon_k) + \delta_k. \tag{15}$$

Again, assume that this has a unique solution $\mathbf{x}_\epsilon$. Then, we obtain the following theorem.

Theorem 4 *Let* $\mathbf{F}$ *and* $\mathbf{S}$ *be Lipschitz continuous with Lipschitz constants* L_F *and* L_S*, respectively. Assume that* $L_F L_S < 1$*. Then we have, if* $\epsilon_k = \delta_k = \epsilon$ *for all* k*, that*

$$\|\mathbf{x}_\epsilon - \mathbf{x}^*\| \leq \epsilon \frac{1 + L_S}{1 - L_S L_F}. \tag{16}$$

In the case $\epsilon_k = \epsilon$ *and* $\delta_k = \delta$*, we obtain*

$$\|\mathbf{x}_\epsilon - \mathbf{x}^*\| \leq \frac{\epsilon L_S + \delta}{1 - L_S L_F}. \tag{17}$$

Finally, $\mathbf{x}_\epsilon = \mathbf{x}^*$ *if and only if both* δ_k *and* ϵ_k *converge to zero.*

Proof: The proof is technically identical to the one of theorem 1. We have due to the Lipschitz continuity

$$\begin{aligned}
\|\mathbf{x}^{k+1} - \mathbf{x}^*\| = \|\mathbf{S}(\mathbf{F}(\mathbf{x}^k) + \epsilon_k) + \delta_k - \mathbf{x}^*\| &= \|\mathbf{S}(\mathbf{F}(\mathbf{x}^k) + \epsilon_k) + \delta_k - \mathbf{S}(\mathbf{F}(\mathbf{x}^*))\| \\
&\leq L_S \|\mathbf{F}(\mathbf{x}^k) - \mathbf{F}(\mathbf{x}^*) + \epsilon_k\| + \delta_k \leq L_S L_F \|\mathbf{x}^k - \mathbf{x}^*\| + L_S \epsilon_k + \delta_k \\
&\leq (L_S L_F)^2 \|\mathbf{x}^{k-1} - \mathbf{x}^*\| + L_S^2 L_F \epsilon_{k-1} + L_S L_F \delta_{k-1} + L_S \epsilon_k + \delta_k \\
&\leq (L_S L_F)^{k+1} \|\mathbf{x}^0 - \mathbf{x}^*\| + \left(\sum_{j=0}^{k} L_S^{j+1} L_F^j \epsilon_{k-j} \right) + \left(\sum_{j=0}^{k} L_S^j L_F^j \delta_{k-j} \right)
\end{aligned}$$

and thus in the limit $\mathbf{x}^{k+1} \to \mathbf{x}^*$,

$$\|\mathbf{x}_\epsilon - \mathbf{x}^*\| \leq L_S \lim_{k \to \infty} \sum_{j=0}^{k} (L_S L_F)^j \epsilon_{k-j} + \lim_{k \to \infty} \sum_{j=0}^{k} (L_S L_F)^j \delta_{k-j} \tag{18}$$

For a constant perturbation overall, e.g. $\epsilon_k = \delta_k = \epsilon$ for all k, we obtain in the limit

$$\|\mathbf{x}_\epsilon - \mathbf{x}^*\| \leq \epsilon(1 + L_S) \lim_{k \to \infty} \sum_{j=0}^{k} (L_S L_F)^j = \epsilon \frac{1 + L_S}{1 - L_S L_F},$$

which proves the inequality (16). If we have constant but separate perturbations ϵ and δ of $\mathbf{S}$ and $\mathbf{F}$, we obtain (17) from

$$\|\mathbf{x}_\epsilon - \mathbf{x}^*\| \leq \epsilon L_S \lim_{k \to \infty} \sum_{j=0}^{k} (L_S L_F)^j + \delta \lim_{k \to \infty} \sum_{j=0}^{k} (L_S L_F)^j = \frac{\epsilon L_S + \delta}{1 - L_S L_F}.$$

In the general case, due to positivity, the right hand side of (18) is zero if and only if both ϵ_k and δ_k are such that for $\phi_k = \epsilon_k$ or $\phi_k = \delta_k$,

$$\lim_{k \to \infty} \sum_{j=0}^{k} (L_S L_F)^j \phi_{k-j} = 0.$$

By an identical proof to theorem 2, this is the case if and only if both ϵ_k and δ_k converge to zero.

Note that this implies that the sequence ϵ_k perturbing the inner function $\mathbf{F}$ is less important by a factor of the Lipschitz constant L_S of the outer function. Thus, a possible strategy is to define

$$\epsilon_k = \delta_k / L_S, \tag{19}$$

meaning that we solve the fluid part less accurate by a factor of L_S. Unfortunately, L_S has to be known for this.

2.4 Application: Dirichlet-Neumann coupling for Transmission problem

As an application of the theory from section 2.3, we consider a problem that is a basic building block in fluid structure interaction, namely the transmission problem, where the Laplace equation with right hand side $f(x, y)$ on a domain Ω is cut into two domains $\Omega = \Omega_1 \cup \Omega_2$ using transmission conditions at the interface $\Gamma = \Omega_1 \cap \Omega_2$:

$$\begin{aligned} \Delta u_i(x,y) &= f(x,y),\ (x,y) \in \Omega_i \subset \mathbb{R}^2,\ i = 1,2 \\ u_i(x,y) &= 0,\ (x,y) \in \partial\Omega_i\ \partial\Omega_1 \cap \Omega_2 \\ u_1(x,y) &= u_2(x,y),\ (x,y) \in \Gamma \\ \partial u_1(x,y) \cdot \mathbf{n} &= \partial u_2(x,y) \cdot \mathbf{n},\ (x,y) \in \Gamma \end{aligned} \tag{20}$$

We now employ a standard Dirichlet-Neumann iteration to solve it. Using any linear discretization, this corresponds to alternately solving the problems

$$\mathbf{A}\mathbf{u}_1^{k+1} = \mathbf{b}_1(\mathbf{u}_2^k) \tag{21}$$

and

$$\mathbf{B}\mathbf{u}_2^{k+1} = \mathbf{b}_2(\mathbf{u}_1^{k+1}) \tag{22}$$

were problem (21) corresponds to a discretization of the transmission problem (20) on Ω_1 only with Dirichlet data on Γ given by $\mathbf{u}_2^k$ on the coupling interface and problem (22) corresponds to a discretization of (20) on Ω_2 only with Neumann data on Γ given by the discrete normal derivative of $\mathbf{u}_1$ on Γ. It can be shown that convergence of the approximate solutions on the whole domain is equivalent to the convergence of the solution on the interface only [7].

By considering (21)-(22) as one iteration, we obtain a fixed point formulation

$$\mathbf{u}_\Gamma = \mathbf{S}(\mathbf{F}(\mathbf{u}_\Gamma))$$

where $\mathbf{u}_\Gamma$ is $\mathbf{u}_2$ on the interface, $\mathbf{F} = \mathbf{D}_{\mathbf{n}_\Gamma}\mathbf{A}^{-1}\mathbf{b}_1(\mathbf{u}_\Gamma)$ and $\mathbf{S} = \mathbf{P}_\Gamma\mathbf{B}^{-1}\mathbf{b}_2(\mathbf{u}_1)$. Hereby $\mathbf{D}_{\mathbf{n}_\Gamma}$ is the matrix that computes the discrete normal derivatives on

Γ and $\mathbf{P}_\Gamma$ is a projection of the space that $\mathbf{u}_2$ is in onto the space of discrete unknowns on Γ.

In practice, the linear equation systems are solved iteratively, typically using the conjugate gradient method (CG) up to a relative tolerance of τ. Thus, we obtain a perturbed nested fixed point iteration of the form (15) and the question is now again if we can quantify this perturbation. We have

$$\mathbf{u}_{1_\epsilon}^{k+1} = \mathbf{A}^{-1}\mathbf{b}_1(\mathbf{u}_\Gamma^k) + \epsilon_k \tag{23}$$

and

$$\mathbf{u}_{2_\epsilon}^{k+1} = \mathbf{B}^{-1}\mathbf{b}_2(\mathbf{u}_{1_\epsilon}^{k+1}) + \delta_k. \tag{24}$$

For the iteration (23) we obtain

$$\|\epsilon_k\| = \|\mathbf{u}_{1_\epsilon}^{k+1} - \mathbf{A}^{-1}\mathbf{b}_1(\mathbf{u}_\Gamma^k)\| \leq \|\mathbf{A}^{-1}\|\|\mathbf{A}\mathbf{u}_{1_\epsilon}^{k+1} - \mathbf{b}_1(\mathbf{u}_\Gamma^k)\|.$$

Again, the second factor is what is tested in the termination criterion of CG. In the case of the relative criterion (9), here stated as

$$\|\mathbf{A}\mathbf{u}_{1_\epsilon}^{k+1} - \mathbf{b}_1(\mathbf{u}_\Gamma^k)\| \leq \tau_r\|\mathbf{A}\mathbf{u}_1^k - \mathbf{b}_1(\mathbf{u}_\Gamma^k)\|,$$

we obtain

$$\|\epsilon_k\| \leq \|\mathbf{A}^{-1}\|\tau_r\|\mathbf{A}\mathbf{u}_{1_\epsilon}^k - \mathbf{b}_1(\mathbf{u}_\Gamma^k)\| \leq \kappa(\mathbf{A})\tau_r\|\mathbf{u}_1^k - \mathbf{A}^{-1}\mathbf{b}_1(\mathbf{u}_\Gamma^k)\|$$

$$= \kappa(\mathbf{A})\tau_r\|\mathbf{u}_1^k - \mathbf{u}_1^{k+1}\|$$

Now the point is that since the problem is linear, we can write down a linear mapping that maps $\mathbf{u}_1^k$ onto $\mathbf{u}_1^{k+1}$ for arbitrary k. Let this have Lipschitz constant L_1, then we have

$$\|\epsilon_k\| \leq \tau_r\kappa(\mathbf{A})L_1^k\|\mathbf{u}_1^0 - \mathbf{u}_1^1\|.$$

Thus the perturbation has limit zero if $L_1 < 1$. This is the case if and only if the sequence $(\mathbf{u}_1^k)_k$ is convergent, which is in fact the case provided that $f(x, y)$ is sufficiently harmless, as can be seen from the literature on domain decomposition methods, e.g. [7, ch. 4].

As for the Picard iteration, if we choose the absolute termination criterion (11) or the relative one based on the right hand side (10), we obtain a bound of the form

$$\|\epsilon_k\| \leq \|\mathbf{A}^{-1}\|\tau_r\|\mathbf{b}(\mathbf{u}_\Gamma^k)\|,$$

respectively

$$\|\epsilon_k\| \leq \|\mathbf{A}^{-1}\|\tau_a.$$

Again, unless these inequalities are sharp, we cannot make a statement on the limit of ϵ_k.

In the second case, meaning the iteration with Neumann data (24), we obtain

$$\|\delta_k\| = \|\mathbf{u}_{2_\epsilon}^{k+1} - \mathbf{B}^{-1}\mathbf{b}_2(\mathbf{u}_{1_\epsilon}^{k+1})\| \leq \|\mathbf{B}^{-1}\|\|\mathbf{B}\mathbf{u}_{2_\epsilon}^{k+1} - \mathbf{b}_2(\mathbf{u}_{1_\epsilon}^{k+1})\|$$

and analogous arguments produce the same results for δ_k. Thus, we have that when using the relative criterion we obtain convergence to the exact solution for any τ.

2.5 A note on the termination criterion and convergence speed

It is important to note that under the assumptions, all sequences considered, wether perturbed or not, are converging and therefore, the iteration will terminate when using the standard criterion

$$\|\mathbf{x}^{k+1} - \mathbf{x}^k\| \leq \tau.$$

However, as just shown, the perturbed iteration converges to an approximation of the unperturbed fixed point and thus, the algorithm can terminate when we are in fact not τ-close to the solution.

A further difference between the different iterations that should be stressed is that the iterations perturbed by a constant is a fixed point iteration, wheras the schemes with a variable perturbation are in fact, not. Thus, the convergence speed, which is otherwise linear with constant L, is not clear and numerical evidence suggests that it is in fact slower than for the other iteration.

Thus, we could argue to employ the schemes with constant perturbation, measure the Lipschitz constant numerically after a few iterations and then adjust the tolerance based on theorem 1 or 4. Unfortunately, it is not clear what the ϵ from these theorems is, respectively, it is based on quantities that are hard to measure like $\|\mathbf{A}^{-1}\|$. Thus, we cannot guarantee a certain tolerance in this way.

Finally, it is important to note that this analysis is mostly relevant to the time independent case. Otherwise, when considering this inside an implicit time integration scheme, additional requirements on the solutions appear, namely that the solutions in the subdomains have a certain accuracy.

3 Numerical Results

For all numerical experiments, the fixed point iteration is terminated when the norm $\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2$ is smaller than 10^{-14}. Furthermore, with the exception of the results on the Picard iteration, all computations were performed in MATLAB.

L \ ϵ	1e-1	1e-2	1e-3
0.009868	1.010e-1	1.010e-2	1.010e-3
0.101239	1.090e-1	1.089e-2	1.089e-3
0.899524	2.016e-1	1.827e-2	1.813e-3
0.996035	2.290e-1	1.981e-2	1.961e-3

Table 1: $|x_\epsilon - x^*|$ (left) and $\frac{\epsilon}{1-L}$ (right) for different values of ϵ and L for the solution of the scalar nonlinear equation (25)

3.1 Direct Perturbation

3.1.1 Testcase: Scalar nonlinear system

As a first example, we employ the nonlinear scalar equation

$$x = e^{\gamma x}/4 \tag{25}$$

with $x \in [0, 1]$ and $\gamma < 1$ given. Thus, the Lipschitz constant L on $[0, 1]$ is equal to $\gamma e^\gamma/4 < 1$. We solve this equation for $\gamma = 0.3, 1.145, 1.2$.

Employing the fixed point method with constant perturbation, we provide the values of $|x_\epsilon - x^*|$ in table 1. The difference in solutions is proportional to ϵ, as suggested by theorem 1. However, the dependence on the Lipschitz constant is very weak and the error does not become worse when it approaches one. This is because the problem is nonlinear and thus, the Lipschitz constant is domain dependent. The local Lipschitz constant near the solution is actually well smaller than one, which reminds us that for nonlinear problems, the Lipschitz constant does not always describe a problem well.

Furthermore, we tested the adaptive strategy and there $|x_\epsilon - x^*|$ tends to machine accuracy, as predicted by the theory.

3.1.2 Testcase: Picard iteration

We now consider the Picard iteration (7). The equation system considered arises from the discretization of the incompressible Navier-Stokes equations on the unit square with homogeneous Dirichlet boundary conditions and a viscosity of $\nu = 1/1000$. The grid is cartesian with $128 \times 128 = 16384$ cells. For the computations, the code MooNMD by John et. al. [5] was used. The Finite Element discretization employs P_1/Q_2 elements, resulting in 181250 unknowns overall, thereof 66049 for the velocities (including the Dirichlet nodes) and 49151 for the pressure.

The right hand side is chosen that the solution is given by $(u_1, u_2) = (d\psi/dy, -d\psi/dx)$ with $\psi = x^2(1-x)^2y^2(1-y)^2$, resulting in

τ_r	Fixp. iter	GMRES it.	$\|res\|$
1e-01	12	42	1.793e-15
1e-02	9	48	8.559e-15
1e-03	9	58	2.318e-15
1e-04	8	57	2.121e-15
1e-05	7	59	5.299e-15
1e-06	7	64	5.266e-15
1e-07	7	67	5.265e-15

Table 2: $\|res\|$, total number of GMRES iterations and fixpoint iterations for different values of τ_r when using termination criterion (9)

τ_a	Fixp. iter	GMRES it.	$\|res\|$
1e-01	1	0	8.212092e-03
1e-02	1	0	8.212092e-03
1e-03	1	4	6.356705e-04
1e-04	1	6	2.131814e-05
1e-05	1	7	4.018437e-06
1e-06	2	11	3.016563e-07
1e-07	2	14	6.249272e-08

Table 3: $\|res\|$, total number of GMRES iterations and fixpoint iterations for different values of τ_a when using termination criterion (11)

$$u_1(x,y) = x^2(1-x)^2[2y(1-y)^2 - 2y^2(1-y)],$$
$$u_2(x,y) = [2x^2(1-x) - 2x(1-x)^2]y^2(1-y)^2,$$
$$p(x,y) = x^3 + y^3 - 1/2 \in L_0^2.$$

To solve the linear systems (8), GMRES is employed and terminated based on a a relative tolerance τ_r.

In table 2, $\|res\|$, as well as the total number of inner GMRES iterations and the number of Picard iterations needed to reach machine accuracy are shown for different values of τ_r, where the termination criterion (9) was used. As predicted by the theory, all schemes converge to the exact solution. Furthermore, it takes slightly more fixed point iterations to reach machine accuracy if the linear systems are solved very inaccurately. Nevertheless, the most efficient scheme is the one with $\tau_r = 1e-1$.

In table 3, we show the same quantities, but for the termination criterion (11). As can be seen, the Picard iteration does not converge to the exact solution, and how close we get is proportional to τ_a. This suggests that here,

the upper bound on the perturbation is accurate, thus having a situation as in theorem 1.

3.2 Nested fixed point iteration

3.2.1 Testcase: Linear Equation System with Matrix Product

We now consider the linear problem

$$(\mathbf{I} - \mathbf{AB})\mathbf{x} = \mathbf{b} \Leftrightarrow \mathbf{x} = \mathbf{ABx} + \mathbf{b} \tag{26}$$

with

$$\mathbf{A} = \begin{pmatrix} \alpha & 0 \\ 0.001 & 0.001 \end{pmatrix}, \mathbf{B} = \begin{pmatrix} \beta & 0 \\ 0.001 & 0.001 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Thus, $\mathbf{S}(\mathbf{x}) = \mathbf{Ax} + \mathbf{b}$ with $L_S = \|\mathbf{A}\|_2 \approx \alpha$ and $\mathbf{F}(\mathbf{x}) = \mathbf{Bx}$ with $L_F = \|\mathbf{B}\|_2 \approx \beta$.

ϵ	α \ β	0.1	0.9	0.99
1e-1	0.1	1.111e-1 1.058e-1	1.209e-1 1.111e-1	1.221e-1 1.117e-1
	0.9	2.087e-1 1.638e-1	1.000e-0 7.107e-1	1.743e-0 1.235e-0
	0.99	2.209e-1 1.715e-1	1.826e-1 1.293e-1	1.000e+1 7.071e-0
1e-2	0.1	1.111e-2 1.058e-2	1.209e-2 1.111e-2	1.221e-2 1.117e-2
	0.9	2.088e-2 1.638e-2	1.000e-1 7.107e-2	1.743e-1 1.235e-1
	0.99	2.209e-2 1.715e-2	1.826e-1 1.293e-1	1.000e-0 7.071e-1
1e-3	0.1	1.111e-3 1.058e-3	1.209e-3 1.111e-3	1.221e-3 1.117e-3
	0.9	2.088e-3 1.638e-3	1.000e-2 7.107e-3	1.743e-2 1.235e-3
	0.99	2.209e-3 1.715e-3	1.826e-2 1.293e-3	1.000e-1 7.071e-2

Table 4: Estimate (17) and $\|\mathbf{x}_\epsilon - \mathbf{x}^*\|_2$ for different values of ϵ, L_S and L_F for equation (26)

As a perturbation, we use a constant vector with eucledian norm $\epsilon = \delta$. The difference $\|\mathbf{x}_\epsilon - \mathbf{x}^*\|_2$ can be seen in table 4. As initial guess, the zero vector was used. The results demonstrate that (17) is a very good estimate of the true error and that the errors are perfectly proportional to ϵ. We furthermore tested the adaptive strategy and that iteration indeed converges to $\mathbf{x}^*$.

3.2.2 Testcase: Scalar nonlinear system

As a second example, we employ the nonlinear scalar problem

$$x = 0.25\gamma_1 e^{\gamma_2 x^2} \tag{27}$$

with $x \in [0, 1]$, $S(x) = 0.25\gamma_1 e^x$ and $L_S = 0.25\gamma_1 e$ and $F(x) = \gamma_2 x^2$ and $L_F = 2\gamma_2$.

The initial guess in the following numerical experiments is $x^0 = 0.5$. In table 5, we show several quantities for different values of ϵ, L_S and L_F where a constant perturbation $\epsilon = \delta$ is employed. First, the estimate (16) using the Lipschitz constants on the interval [0,1], which is referred to as the global estimate. Then the local estimate, which is (16) using the derivatives in the solution, giving an estimate of a local Lipschitz constant. This is reasonable, since all functions are monotonic. Finally, the difference $|x_\epsilon - x^*|$ itself.

As can be seen, we again have the proportionality to ϵ. Furthermore, we see that only when both L_F and L_S are close to one, an influence on the error can be observed, as suggested by theorem 4. Finally, we test the adaptive strategy as an example of perturbations converging to zero and again, we obtain convergence of the new sequence to x^*.

3.2.3 Testcase: Transmission Problem

We now consider the transmission problem (20). Specifically, we use $\Omega_1 = [0, 1] \times [0, 1]$, $\Omega_2 = [1, 2] \times [0, 1]$ and

$$\begin{aligned} f(x, y) = \quad & \sin \pi y^2 (\pi \cos \tfrac{\pi}{2} x^2 - \pi^2 x^2 \sin \tfrac{\pi}{2} x^2) \\ & + \sin \tfrac{\pi}{2} x^2 (2\pi \cos \pi y^2 - 4\pi^2 y^2 \sin \pi y^2). \end{aligned}$$

This was chosen such that the solution is

$$u(x, y) = \sin \pi y^2 \sin \frac{\pi}{2} x^2, \tag{28}$$

which satisfies the boundary conditions.

We discretize this problem using central differences with a constant mesh width of $\Delta x = \Delta y$. As initial guess for the Dirichlet-Neumann procedure, we employ a vector of all zeros. All linear systems are solved using CG.

ϵ	L_S \ L_F	0.01	0.1	0.9	0.99
		1.101e-1	1.111e-1	1.209e-1	1.221e-1
	0.1	1.038e-1	1.038e-1	1.039e-1	1.040e-1
		1.039e-1	1.039e-1	1.042e-1	1.042e-1
		1.917e-1	2.088e-1	1.000e-0	1.743e-0
1e-1	0.9	1.463e-1	1.485e-1	1.724e-1	1.760e-1
		1.350e-1	1.370e-1	1.618e-1	1.658e-1
		2.010e-1	2.209e-1	1.826e-0	1.000e+1
	0.99	1.527e-1	1.555e-1	1.896e-1	1.949e-1
		1.386e-1	1.411e-1	1.746e-1	1.806e-1
		1.101e-2	1.111e-2	1.209e-2	1.221e-2
	0.1	1.038e-2	1.038e-2	1.040e-2	1.040e-2
		1.037e-2	1.037e-2	1.038e-2	1.039e-2
		1.917e-2	2.088e-2	1.000e-1	1.743e-1
1e-2	0.9	1.463e-2	1.485e-2	1.725e-2	1.760e-2
		1.334e-2	1.350e-2	1.525e-2	1.551e-2
		2.010e-2	2.209e-2	1.826e-1	1.000e-0
	0.99	1.527e-2	1.556e-2	1.896e-2	1.949e-2
		1.368e-2	1.388e-2	1.621e-2	1.658e-2
		1.101e-3	1.111e-3	1.209e-3	1.221e-3
	0.1	1.038e-3	1.038e-3	1.039e-3	1.040e-3
		1.037e-3	1.037e-3	1.038e-3	1.038e-3
		1.917e-3	2.088e-3	1.000e-2	1.743e-2
1e-3	0.9	1.463e-3	1.485e-3	1.725e-3	1.760e-3
		1.333e-3	1.348e-3	1.518e-3	1.542e-3
		2.010e-3	2.209e-3	1.826e-2	1.000e-1
	0.99	1.527e-3	1.555e-3	1.896e-3	1.949e-3
		1.366e-3	1.386e-3	1.612e-3	1.646e-3

Table 5: Global estimate, local estimate and $|x_\epsilon - x^*|$ for different values of ϵ, L_S and L_F for equation (27)

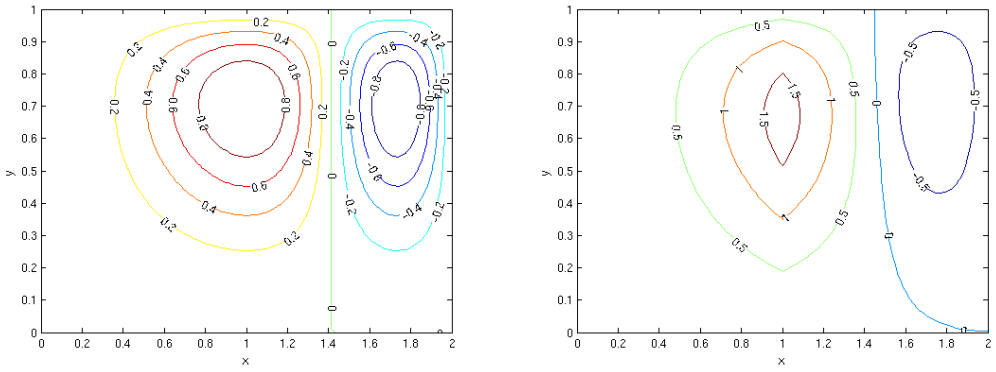

Figure 1: Exact and discrete solution with $\Delta x = 1/40$

τ \ Δx	1/10	1/20	1/40	1/80
1e-1	7.606e-1	4.189e-0	6.519e+149	3.755e+148
1e-2	9.620e-2	2.502e-2	2.621e-1	9.350e-0
1e-3	1.230e-2	2.773e-3	1.192e-1	2.721e-1
1e-4	9.109e-4	1.033e-3	5.545e-3	2.181e-3
1e-5	1.045e-4	7.101e-5	1.571e-5	4.621e-3
1e-6	7.326e-6	8.114e-6	1.011e-4	5.291e-4

Table 6: $\|x_\epsilon - x^*\|_2$ for different values of τ and Δx for the transmission problem (20) with relative termination criterion (10)

The exact solution and the discrete solution with $\Delta x = 1/40$ can be seen in figure 1.

We first look at the convergence properties of the fixed point schemes for different mesh widths and different termination criteria. The difference $\|\mathbf{x}_\epsilon - \mathbf{x}^*\|_2$ for a constant tolerance τ in both CG-subsolvers can be seen in table 6 for the relative termination criterion (10) and in table 7 for the absolute termination criterion (11). As can be seen, the schemes behave essentially as if the perturbation were constant and do not converge to the exact solution. In particular, this implies that the perturbations do not converge to zero. Again, there is a proportionality to τ, though it's not as clear this time. We attribute this to the fact that a relative tolerance in CG is only an upper bound on the perturbation, which can in fact be much smaller than τ if CG oversolves. Furthermore, the perturbed solutions become in general less accurate when the mesh is refined. This can be explained by the dependence on the norms of $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$, which increase with decreasing

τ \ Δx	1/10	1/20	1/40	1/80
1e-1	4.829e-2	4.392e-2	1.998e-2	5.693e-3
1e-2	1.173e-2	6.120e-3	2.828e-3	1.602e-3
1e-3	4.127e-4	2.072e-4	2.118e-4	1.360e-4

Table 7: $\|x_\epsilon - x^*\|_2$ for different values of τ and Δx for the transmission problem (20) with absolute termination criterion (11)

mesh width.

In the case of the termination criterion (9), we recover the exact solution, as predicted by the theory.

We now consider the total number of CG and fixed point iterations when using the termination criterion (9) for different tolerances τ and different mesh widths Δx. As can be seen in table 8, the number of CG iterations increases with decreasing mesh width, which is well known behavior due to the spectrum getting more widely distributed on the real line. Furthermore, the number of fixed point iterations is almost independent of τ. Thus, the most efficient variant is to solve the systems only up to $\tau = 1e - 1$.

4 Summary and Conclusions

We considered perturbed fixed point iterations where the perturbation results from inexact solves of equation systems by iterative solvers. Thereby, we extended a perturbation result for fixed point equations to the case of a nested fixed point equation. Applying these results to the Picard- and the Dirichlet-Neumann iteration for steady states, we showed that these converge to the exact solution indepently of the tolerance in the subsolver, if a specific relative termination criterion is employed. This justifies extremely coarse solves in the inner solvers.

If an absolute or standard relative criterion is used, the theory indicates that we will not converge to the exact solution. Numerical results demonstrate this behavior.

Acknowledgements

Part of this work was funded by the German Research Foundation (DFG) as part of the collaborative research area SFB TRR 30, project C2. Furthermore, I'd like to thank Gunar Matthies for performing the numerical experiments on the Picard iteration in section 3.1.2.

τ \ Δx	1/10		1/20	
	#FP	#GMRES	#FP	#GMRES
1e-1	106	2220	205	8223
1e-2	105	2901	205	11536
1e-3	105	3116	205	12513
1e-4	105	3367	208	13927
1e-5	105	3696	207	15322
1e-6	105	4076	206	16797
τ \ Δx	1/40		1/80	
	#FP	#GMRES	#FP	#GMRES
1e-1	394	28917	380	39582
1e-2	394	43612	760	149699
1e-3	396	48005	793	190175
1e-4	396	52521	848	226154
1e-5	401	59133	759	219592
1e-6	396	64724	767	249260

Table 8: Total CG iterations for the transmission problem (20) for different numbers of τ and Δx when using termination criterion (9)

References

[1] P. Birken, *Numerical Methods for the Unsteady Compressible Navier-Stokes Equations*, Habilitation Thesis, University of Kassel, 2012.

[2] P. Birken, K. J. Quint, S. Hartmann, and A. Meister, *A Time-Adaptive Fluid-Structure Interaction Method for Thermal Coupling*, Comp. Vis. in Science, 13 (2011), pp. 331–340.

[3] S. C. Eisenstat and H. F. Walker, *Choosing the forcing terms in an inexact newton method*, SIAM J. Sci. Comput., 17 (1996), pp. 16–32.

[4] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics*, Oxford University Press, 2005.

[5] V. John and G. Matthies, *MooNMD a program package based on mapped finite element methods*, Comp. Vis. Sci., 170 (2004), pp. 163–170.

[6] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, SIAM, 2nd ed., 2000.

[7] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*, Numerical Mathematics and Scientific Computation, Oxford Science Publications, Oxford, 1999.

[8] S. Turek, *Efficient Solvers for Incompressible Flow Problems: An Algorithmic and Computational Approach*, Springer, Berlin, 1999.