Abstract

We report on methods for computing enclosures of solutions of second-order nonlinear elliptic boundary value problems, simultaneously proving the existence of a solution in the enclosing set. The old-fashioned ‘monotonicity methods’ are well suited for this task, but only for a restricted class of problems. Therefore, we propose a new approach which is based on a suitable fixed-point formulation of the problem and uses, as an essential ingredient, norm bounds for the inverse of the linearization of the given problem at some approximate solution which is computed numerically. These norm bounds are obtained via eigenvalue enclosures. We also give a brief description of an alternative method proposed by M.T. Nakao. © 2001 Elsevier Science Inc. All rights reserved.

Keywords: Elliptic boundary value problem; Nonlinear; Enclosure; Existence

1. Introduction

This paper will primarily be concerned with nonlinear elliptic boundary value problems of the form

\[-\Delta u + F(x, u, \nabla u) = 0 \quad \text{on} \; \Omega,\]
\[u = 0 \quad \text{on} \; \partial \Omega,\]  

where \(\Omega \subset \mathbb{R}^n\) is a bounded domain with Lipschitz-continuous boundary \(\partial \Omega\), and \(F\) is a given nonlinearity on \(\overline{\Omega} \times \mathbb{R} \times \mathbb{R}^n\) with values \(F(x, y, z) \in \mathbb{R}\); \(F\) and its derivatives \(\partial F/\partial y\) and \(\partial F/\partial z = (\partial F/\partial z_1, \ldots, \partial F/\partial z_n)\) are assumed to be continuous. Further regularity assumptions will follow below in appropriate places.

E-mail address: michael.plum@mathematik.uni-karlsruhe.de (M. Plum).
We are interested in computer-assisted methods for proving, under appropriate conditions, the existence of a solution to problem (1) within a ‘close’ and explicitly given neighborhood of some approximate solution. In slightly different words, the methods provide safe and verified error bounds for approximate solutions. The conditions, under which such an existence and enclosure result can be stated, shall moreover be testable in an automatic way on a computer.

When talking about solutions of problem (1), we have essentially two solution concepts in mind:

(i) Strong solutions \( u \in H^2_\Omega \): the differential equation in (1) is required to hold almost everywhere in \( \Omega \).

(ii) Weak solutions \( u \in H^1_\Omega \): the term \( \Delta u \) in (1) is then to be understood in the distributional sense; i.e., (1) is required to hold in the weak sense:

\[
\int_{\Omega} [\nabla u \cdot \nabla \varphi + F(\cdot, u, \nabla u) \varphi] \, dx = 0 \quad \text{for all } \varphi \in H^1_\Omega.
\]

These different solution concepts require different general regularity assumptions for problem (1) and different ‘conditions’ (to be tested on a computer), in order to obtain the desired existence and enclosure statement. This will be made more precise in the following sections.

The general conception for obtaining existence (and enclosure) results for nonlinear problems of various kinds is to transform the problem into some suitable fixed-point equation

\[
u = T(u)
\]

and to apply some fixed-point theorem (e.g., Banach’s or Schauder’s theorem). This requires to construct a fixed-point operator \( T \) with certain topological or metric properties (such as continuity, compactness, contractivity), which moreover maps some suitable subset \( U \) of the chosen underlying Banach space into itself:

\[
T(U) \subset U.
\]

Since the fixed-point theorem provides the existence of a solution \( u^* \) to (1) in \( U \), the desired enclosure is directly given by the set \( U \) itself:

\[
u^* \in U.
\]

Because we are looking for explicit enclosures, the set \( U \) must be constructed quantitatively, i.e., it must be described by explicit numbers (such as, e.g., a norm ball with explicitly given midpoint and radius, or a function interval with explicitly given upper and lower bound functions). Consequently, checking the crucial condition (3) requires explicit information also about the operator \( T \), in form of appropriate bounds for its ‘ingredients’.

The differences between existing existence and enclosure methods for problems like (1) are characterized by different choices of fixed-point operators \( T \) and different types of sets \( U \). The ‘classical’ choice, initialized by Collatz already in the 1950s [6] and developed in detail later by Schröder [24,25], uses function intervals \( U \) and a
monotone fixed-point operator $T$. It will briefly be described in Section 2. These monotonicity methods have proved to be successful in many relevant examples (see, e.g., [7]). However, the class of problems which they can be applied to is restricted by some significant inherent requirements, as will be demonstrated in Section 2. This justifies the search for new existence and enclosure methods.

We will present such a method here, which avoids the aforementioned restrictions. It will first be described in an abstract operator theoretical setting in Section 3, and then be applied to more concrete problem classes (such as problem (1)) in the following sections. The method uses a fixed-point formulation (2), where the fixed-point operator $T$ now involves an approximate solution $\omega$ (usually obtained by numerical means) and the inverse of the linearization $L$ of the given problem at $\omega$. The construction of a suitable set $U$ satisfying (3) therefore requires, in particular, to compute an appropriate bound for $L^{-1}$. The main work for obtaining it consists in the computation of bounds for eigenvalues of $L$ or of $L^*L$.

A typical way of computing an approximate solution $\omega$ is to use a Newton iteration in combination with some projective method (e.g., Ritz–Galerkin or collocation, with finite element or finite Fourier series basis functions). To solve the corresponding large linear systems of equations (approximately), the use of appropriate methods from numerical linear algebra is essential. These are also important for computing the eigenvalue bounds needed for bounding $L^{-1}$; here, first a few approximate eigenpairs of a large matrix eigenvalue problem have to be computed, which in a second step are used to generate small matrix eigenvalue problems, the eigenvalues of which are enclosed by methods from verifying numerical linear algebra; see also Section 4.1.1.

A different approach to existence and enclosure results for problems like (1) has been developed by Nakao [13,14]. It avoids the computation of bounds for $L^{-1}$. Instead, it deals with the inverse of some finite-dimensional projection of $L$, bounding the infinite-dimensional remainder by other means. We will give a brief description of Nakao’s method in Section 8.

In [10], an existence and enclosure result for a spatially periodic Navier–Stokes problem is established; here, a bound for $L^{-1}$ is obtained by Fourier analytic methods which exploit the special finite Fourier series form of the approximation $\omega$.

For boundary value problems with ordinary differential equations (which are of course contained in (1), as long as they are of second order), many existence and enclosure methods can be found in the literature. We will not address these approaches in the present paper. Ordinary differential equations will however appear as a special application of our abstract setting in Section 7.

2. A monotonicity approach

The types of operators $T$ and sets $U$ which were first proposed by Collatz [6] and later developed further by Schröder [24,25], Walter [27] and others, are strongly
related to the classical maximum principle for subharmonic functions. They reduce
the inclusion requirement (3) to a set of inequalities which are often much easier to
handle.

For easy explanation, we assume here that $F$ in (1) is linear in $\nabla u$, i.e., that
\[ F(x, y, z) = b(x) \cdot z + f(x, y), \tag{5} \]
with $b, f,$ and $\partial f/\partial y$ being continuous functions.

The basis of the monotonicity approach is to choose the enclosing set $U$ as a
function interval
\[ U = [v, w] := \{ u \in C(\overline{\Omega}): v(x) \leq u(x) \leq w(x) \text{ for } x \in \overline{\Omega} \}, \]
with functions $v, w \in C(\overline{\Omega})$, $v \leq w$ (pointwise), which are chosen fixed resp. have
to be constructed. Choosing some constant $c_0 > 0$ such that
\[ c_0 \geq \max \left\{ \frac{\partial f}{\partial y}(x, y): x \in \overline{\Omega}, \ v(x) \leq y \leq w(x) \right\} \]
one finds that the operator
\[ L_0[u] := -\Delta u + b(x) \cdot \nabla u + c_0 u \]
is inverse-positive (i.e., for smooth $u$, $L_0[u] \geq 0$ on $\Omega$ together with $u \geq 0$ on $\partial \Omega$
implies $u \geq 0$ on $\overline{\Omega}$; see [24]), so that its inverse $L_0^{-1}$ (for Dirichlet boundary con-
ditions) is a positive operator, and that the mapping $u \mapsto c_0 u - f(\cdot, u)$ is monotone
on $U = [v, w]$. Since, under appropriate regularity conditions, $L_0^{-1}$ is defined on (a
superset of) $C(\overline{\Omega})$, the operator
\[ T : C(\overline{\Omega}) \to C(\overline{\Omega}), \ T(u) := L_0^{-1}[c_0 u - f(\cdot, u)] \tag{6} \]
is well defined and monotone on $U$, and problem (1) (with $F$ given by (5)) is equiva-
 lent to the fixed-point equation (2). The monotonicity immediately provides that, for
$U = [v, w]$, the enclosure $Tv \leq Tu \leq Tw$ holds for all $u \in U$, so that the inclusion
condition (3) is equivalent to the inequalities
\[ v \leq T(v), \quad T(w) \leq w. \tag{7} \]
If $v, w$ are smooth, the inverse-positivity of $L_0$ shows that in turn the inequalities
$L_0[v] \leq L_0[Tv], L_0[Tw] \leq L_0[w]$ (together with $v \leq 0 \leq w$ on $\partial \Omega$), i.e., the dif-
ferential inequalities
\[ -\Delta v + F(\cdot, v, \nabla v) \leq 0 \leq -\Delta w + F(\cdot, w, \nabla w) \quad \text{on } \Omega, \]
\[ v \leq 0 \leq w \quad \text{on } \partial \Omega \tag{8} \]
are sufficient for (7) and thus, for (3). Since $T$ in (6), acting in the Banach space
$(C(\overline{\Omega}), \| \cdot \|_\infty)$, is continuous and compact (under suitable regularity conditions),
and $U$ is closed, bounded, and convex, Schauder’s fixed-point theorem therefore shows
that (8) provides the existence of a solution to problem (1) in $U = [v, w]$.

For practical applications, one will—as already suggested in [6]—start from a
(smooth) numerical approximation $\omega$ to (1) and try to find $v, w$ satisfying (8) in the
form
\[ v = \omega - \psi, \quad w = \omega + \psi \]  
\[ (9) \]
with some ‘small’ function \( \psi \geq 0 \). Inserting (9) into (8) one obtains the inequalities
\[ -\Delta \psi + b \cdot \nabla \psi + [f(\cdot, \omega + \psi) - f(\cdot, \omega)] \geq -d[\omega] \quad \text{on } \Omega, \]
\[ -\Delta \psi + b \cdot \nabla \psi + [f(\cdot, \omega) - f(\cdot, \omega - \psi)] \geq d[\omega] \quad \text{on } \Omega, \]
\[ \psi \geq |\omega| \quad \text{on } \partial \Omega, \]
\[ (10) \]
where \( d[\omega] := -\Delta \omega + F(\cdot, \omega, \nabla \omega) \) denotes the defect of the approximation \( \omega \). Since we are looking for a small function, \( \psi \), we can try to solve (10) by the following approach proposed by Schröder: one first computes some nonnegative function \( \psi_0 \) satisfying
\[ L[\psi_0] \geq |d[\omega]| \quad \text{on } \Omega, \quad \psi_0 \geq |\omega| \quad \text{on } \partial \Omega, \]
\[ (11) \]
with \( L \) denoting the linearization of (1), (5) at \( \omega \),
\[ L[u] := -\Delta u + b \cdot \nabla u + cu, \quad c(x) := \frac{\partial f}{\partial y}(x, \omega(x)) \quad (x \in \Omega), \]
\[ (12) \]
and then (hopefully) verifies (10) for \( \psi := (1 + \delta)\psi_0 \), where \( \delta > 0 \) is chosen heuristically, e.g., \( \delta = 0.01 \). Since \( L[\psi] \) and the left-hand sides of the differential inequalities in (10) differ only by a \( o(\psi) \)-term, this way of proceeding will often be successful (e.g., if \( \delta L[\psi_0] \) dominates this \( o(\psi) \)-term).

Along these or similar lines, e.g., by direct solution of (8) via an ansatz for \( v \) and \( w \), enclosure results for many problems with ordinary and partial differential equations were obtained (partly already decades ago!) by Collatz [7], Schröder [24], Walter [27], and others. Rounding errors were avoided by rational arithmetic or they were neglected. Of course, the latter option violates the rigor of the results, but the influence of rounding errors in the numerical treatment of boundary value problems may usually be regarded to be very small. So monotonicity methods have proved to work very efficiently as long as the given problem belongs to a class where they are applicable, and the author wishes to advocate them clearly for these problem classes.

However, the inequalities (10) (together with the requirement for a small function \( \psi \geq 0 \)) contain an inherent assumption which severely restricts the class of problems where monotonicity methods can be applied to: an ‘almost necessary’ condition (at least under practical aspects) for (10) is
\[ L[\psi] > 0 \quad \text{on } \Omega \]
\[ (13) \]
(compare also (11)), since \( L[\psi] \) and the left-hand sides of the differential inequalities in (10) differ, as already mentioned, only by a \( o(\psi) \)-term.

By [24], condition (13) (together with \( \psi \geq 0 \)) shows that \( L \) must be inverse-positive, which in turn implies that all eigenvalues of \( L \) (subject to Dirichlet boundary conditions) must have positive real part. This is a general restriction on principle, which prevents monotonicity methods from succeeding in many interesting examples, as is illustrated by the following problem taking its rise from semiconductor physics:
Fig. 1. Bifurcation diagram for problem (14).

\[ -u'' - (u \cdot \sin x)' + \lambda(u^2 + u - 1) = 0 \quad (0 < x < 2\pi), \]
\[ u'(0) = u'(2\pi) = 0. \]

(The fact that (14) requires Neumann boundary conditions does not affect the arguments.) Applying a Newton-collocation procedure (together with branch-following and branch-switching techniques) we computed the bifurcation diagram of approximate solutions \( \omega \) shown in Fig. 1. An additional branch is formed by the vertical line at \( \lambda = 0 \), where \( u(x) := \mu \cdot \exp(\cos x) \) is a solution for each \( \mu \in \mathbb{R} \).

On all branches (which were drawn after interpolation of a ‘grid’ of computed approximate solutions \( \omega \)), one finds the respective numbers of negative eigenvalues of the linear operator \( L \) defined in (12). Only two branches show a zero, i.e., only on these two branches all (real) eigenvalues of \( L \) (subject to Neumann boundary conditions) are positive, as required by the ‘almost necessary’ conditions of the ‘monotonicity’ approach.

However, the existence and enclosure method which we are going to propose in the following sections was successful on all branches, except in direct neighborhoods of turning points and bifurcation points, where the inverse of \( L \) does not exist or has at least a very large norm. Nevertheless, in Section 6, we will discuss extensions of our method which are able to provide existence and enclosure results also in neighborhoods of (and even in) such singular points.
3. A new method—abstract formulation

The drawbacks of the classical monotonicity methods mentioned just before justify the development of new methods avoiding the restriction on the eigenvalues of the linearization \( L \). Here, we propose such a method first on a more abstract level, which will later be made more concrete in several different ways.

Let \( X, Y, Z \) denote three Banach spaces, \( X \subset Y \), and let \( L_0 \in \mathcal{B}(X, Z) \) (the space of bounded linear operators from \( X \) to \( Z \)). Moreover, let \( \mathcal{F} : Y \to Z \) denote a Fréchet-differentiable operator.

**Remark 3.1.** The Fréchet differentiability of \( \mathcal{F} \) can be relaxed: \( \mathcal{F} \) need only be continuous on \( Y \), and Fréchet differentiable at the approximation \( \omega \in X \) introduced further below. This is sometimes helpful for applications.

We consider the problem

\[
\begin{align*}
    u & \in X, \quad L_0[u] + \mathcal{F}(u) = 0, \\
    (15)
\end{align*}
\]

again aiming at existence and enclosure statements. The enclosing set \( U \) (see (4)) will now be a norm ball with known center and radius (the latter being moreover small). We make the following abstract regularity assumptions (in particular the first one can be weakened):

(A) The embedding \( E_Y^X : X \to Y \) is compact.

(B) For some \( \sigma \in \mathcal{B}(Y, Z) \), \( L_0 + \sigma E_Y^X : X \to Z \) is one-to-one and onto.

As a consequence we obtain, for every \( \rho \in \mathcal{B}(Y, Z) \), the implication

\[
\begin{align*}
    (L_0 + \rho E_Y^X)^{-1} & \in \mathcal{B}(Z, X). \\
    (16)
\end{align*}
\]

(For proof, observe first that assumption (B) and the open mapping theorem show that \( (L_0 + \sigma E_Y^X)^{-1} : Z \to X \) exists and is bounded. Consequently, for given \( r \in Z \), the equation \( (L_0 + \rho E_Y^X)[u] = r \) can be rewritten as \( u = Ku + s \), where \( K := (L_0 + \sigma E_Y^X)^{-1}(\sigma - \rho)E_Y^X : X \to X \) is compact due to assumption (A), and \( s := (L_0 + \sigma E_Y^X)^{-1}[r] \in X \). Fredholm’s alternative therefore provides a unique solution \( u \), since the homogeneous equation \( (r = 0) \) has only the trivial solution because \( L_0 + \rho E_Y^X \) is one-to-one. Thus, \( (L_0 + \rho E_Y^X)^{-1} : Z \to X \) exists, and its boundedness follows again from the open mapping theorem.)

Now, let \( \omega \in X \) denote some approximate solution to problem \((15)\), and denote by

\[
\begin{align*}
    d & := L_0[\omega] + \mathcal{F}(\omega) \in Z. \\
    (17)
\end{align*}
\]

its defect (residual). Simple calculations show that the following equation for the error \( v = u - \omega \) is equivalent to \((15)\):

\[
\begin{align*}
    v & \in X, \quad L_0[v] + \mathcal{F}(\omega + v) - \mathcal{F}(\omega) = -d. \\
    (18)
\end{align*}
\]
Now, with $F'_0(Y, Z)$ denoting the Fréchet derivative of $F$ at $\omega$, let

$g(v) := F(\omega + v) - F(\omega) - F'(\omega)[v]$ for $v \in Y$, \hfill (19)

$L := L_0 + F'(\omega)E_X^Y$. \hfill (20)

Assuming that $L : X \to Z$ is one-to-one, so that it is also onto and $L^{-1}$ is bounded according to (16), we can therefore rewrite (18) as

$v \in X, \quad v = -L^{-1}[d + g(E_X^Y v)] =: T(v)$ \hfill (21)

and apply Schauder’s fixed-point theorem: since $L^{-1}$ is bounded, $g$ is continuous, and $E_X^Y$ is compact, we conclude that $T : X \to X$ is continuous and compact. We are therefore left to find a closed, bounded, and convex set $V \subset X$ such that $T(V) \subset V$. Here, we aim in particular at a norm ball $V$ centered at $\omega$ with radius $\alpha$ (to be constructed). For this purpose, suppose that constants $\delta, C,$ and $K$, as well as some monotonically nondecreasing function $G : [0, \infty) \to [0, \infty)$ are known which satisfy:

\begin{align*}
\|L_0[\omega] + F'(\omega)\|_Z &\leq \delta, \quad \text{(22)} \\
\|u\|_Y &\leq C\|u\|_X \quad \text{for all } u \in X, \quad \text{(23)} \\
\|u\|_X &\leq K\|L[u]\|_Z \quad \text{for all } u \in X, \quad \text{(24)} \\
\|F(\omega + u) - F(\omega) - F'(\omega)[u]\|_Z &\leq G(\|u\|_Y) \quad \text{for all } u \in Y, \quad \text{(25)} \\
G(t) &= o(t) \quad \text{for } t \to 0+, \quad \text{(26)}
\end{align*}

( regards that (25) and (26) are consistent due to the Fréchet differentiability of $F$ at $\omega$). Using (17), (19), and (22)–(25), we obtain from (21) that

$\|T(v)\|_X \leq K[\delta + G(C\|v\|_X)]$ \quad for each $v \in X$, \hfill (27)

so that the norm ball $V = \{v \in X : \|v\|_X \leq \alpha\}$ is mapped into itself by $T$ if $K[\delta + G(C\alpha)] \leq \alpha,$ i.e., if

$\delta \leq \frac{\alpha}{K} - G(C\alpha). \hfill (27)$

We have therefore proved the following:

**Existence and Enclosure Theorem.** If (27) holds for some $\alpha \geq 0,$ there exists a solution $u \in X$ to problem (15) satisfying

$\|u - \omega\|_X \leq \alpha$ \hfill (28)

(i.e., $u$ is contained in the $\|\cdot\|_X$-ball centered at $\omega$ with radius $\alpha$), and in particular, using (23) again,

$\|u - \omega\|_Y \leq C\alpha.$ \hfill (29)
Remark 3.2. (a) An important observation is that, due to (26), the crucial condition (27) is indeed satisfied for some small $\alpha$ if the constant $\delta$ is sufficiently small, which means according to (22) that the approximate solution $\omega$ of problem (15) must be computed with sufficient accuracy, and (27) tells how accurate the computation has to be. This meets the general philosophy of computer-assisted proofs: The 'hard work' of the proof is left to the computer!

(b) If one dispenses with an error bound in the norm $\| \cdot \|_X$ and is content with an error bound in $\| \cdot \|_Y$, one may replace (23) and (24) by the single condition that some $K$ is known which satisfies

$$
\| u \|_Y \leq K \| L[u] \|_Z \quad \text{for all } u \in X.
$$

Then, $K$ and $C$ in (27) have to be replaced by $K$ and 1, respectively, and the assertion is that a solution $u \in X$ exists such that $\| u - \omega \|_Y \leq \alpha$. For proof, apply the 'old' result with $\| \cdot \|_X = \max(\| v \|_Y, \varepsilon \| v \|_X)$ in place of $\| \cdot \|_X$, where $\varepsilon > 0$ is sufficiently small.

We are left to describe how $\delta$, $C$, $K$, and $G$ satisfying (22)–(26) can be computed. We will do so in the following sections in our different realizations of the abstract operator setting.

We remark that the requirement of $L$ being one-to-one (which was stated after (20)) is contained in (24). Observe moreover that (24) does not require any restriction on the eigenvalues of the linearization $L$ except that they must be bounded away from zero. This makes the concept presented here much more general than the monotonicity approach described in Section 2.

4. Strong solutions

Here we describe the application of our abstract results to the elliptic boundary value problem (1), choosing

$$
L_0 := -\Delta, \quad \mathcal{F}(u)(x) := F(x, u(x), \nabla u(x)),
$$

$$
X := H^1_0(\Omega) := \text{closure}_{H^1(\Omega)} \{ v \in C^2(\Omega) : v|_{\partial \Omega} = 0 \}, \quad Z := L^2(\Omega),
$$

that is, we are aiming at existence and enclosure results for strong solutions of problem (1) now. For the choice of the Banach space $Y$ we have to ensure that $E^1_Y$ is compact and that $\mathcal{F} : Y \to Z$ is Fréchet differentiable; these two requirements point into opposite directions concerning the 'strength' of the norm in $Y$. A good ‘balance’ is achieved for the following choices (we restrict ourselves to dimensions $n \leq 3$ here):

(a) If $n = 1$ (so that $\Omega$ is a bounded open real interval and $\Delta u = u''$), the choice $Y := C^1(\overline{\Omega})$ (endowed with its Banach space norm $\| u \| := \| u \|_\infty + \gamma \| u' \|_\infty$, where $\gamma > 0$ denotes a fixed scaling parameter) is appropriate. The compactness of $E^1_Y$ is provided by the Sobolev–Kondrachev embedding theorem [1], and the Fréchet
differentiability of \( \mathcal{F} \) (even as a mapping from \( C_1(\Omega) \) to \( C(\overline{\Omega}) \)) follows from the
continuity of \( F, \partial F/\partial y, \) and \( \partial F/\partial z \).

(b) If \( n \leq 3 \) and \( F \) does not depend on \( z \), one can choose \( Y := C(\Omega) \) (with norm \( \| \cdot \|_\infty \)). The arguments are as before: the Sobolev–Kondrachev theorem yields the compactness of \( E_X^Y \), and the Fréchet differentiability of \( \mathcal{F} \) (even as a mapping from \( C(\Omega) \) to \( C(\overline{\Omega}) \)) follows from the continuity of \( F \) and \( \partial F/\partial y \).

(c) If \( n \in [2, 3] \), the choice \( Y := H_{1,p}(\Omega) \) (for some \( p \in (1, \infty) \)) can be successful. The compactness of the embedding \( E_X^Y \) is provided by the Sobolev–Kondrachev theorem if \( p \cdot (n-2) < 2n \). It is furthermore convenient to choose \( p > n \) in order to have a continuous embedding \( H_{1,p}(\Omega) \to C(\overline{\Omega}) \). The Fréchet differentiability resp. continuity of \( \mathcal{F} : H_{1,p}(\Omega) \to L_2(\Omega) \) now requires (besides the continuity of \( F, \partial F/\partial y, \) and \( \partial F/\partial z \)) the following growth restriction for \( F \) with respect to \( z \): for each \( y_0 \geq 0 \), there exists a constant \( C_0 = C_0(y_0) > 0 \) such that

\[
|F(x, y, z)| \leq C_0(1 + |z|^{p/2})
\]

for all \( x \in \Omega, \ y \in \mathbb{R}, \ |y| \leq y_0, \ z \in \mathbb{R}^n \). \( (33) \)

The special choice \( p := 4 \) is treated further below. See in particular Section 4.2.

We wish to remark that also for dimensions \( n \geq 4 \), the choices \( Y := H_{1,p}(\Omega) \) or (if \( F \) does not depend on \( z \)) \( Y := L_p(\Omega) \) may be successful if \( p \in (1, \infty) \) is chosen suitably. These spaces \( Y \), however, require now always growth restriction for \( F \), not only with respect to \( z \) (as in (33)) but also with respect to \( y \).

The regularity assumption (B) (see Section 3) requires here that the Poisson equation

\[
-\Delta u = r \quad \text{on } \Omega, \quad u = 0 \quad \text{on } \partial \Omega
\]

has a unique solution \( u \in H^2_0(\Omega) \) for each \( r \in L_2(\Omega) \). (It does not make a difference here if the operator \( \sigma \) from the abstract formulation is taken into account or put equal to zero.) In fact, this is the usual \( H_2 \)-regularity condition on the domain \( \Omega \) resp. on its boundary \( \partial \Omega \). It is satisfied, e.g., for \( C_2 \)-smoothly bounded domains \( \Omega \) \( [8] \) or for convex polygonal domains \( \Omega \subset \mathbb{R}^2 \) \( [9] \); it excludes, e.g., domains with reentrant corners.

For the computation of an approximate solution \( \omega \), any numerical method providing an approximation in \( H^2_0(\Omega) \) is suitable. In particular, \( \omega \) has to satisfy the Dirichlet boundary conditions exactly. (However, a generalization of our method is able to handle approximations \( \omega \) satisfying also the boundary conditions only approximately.) Using finite elements for computing \( \omega \), one must choose \( C_1 \)-elements (e.g., triangular Argyris- or Bell-elements, or rectangular Bogner–Schmidt–Fox-elements), in order to meet the desired \( H_2 \)-property of \( \omega \). This is certainly a disadvantage, but on the other hand a rather natural condition when one looks for strong solutions.

We now comment on the computation of the terms \( \delta, \ C, \ K, \) and \( G \) satisfying (22)–(26). Condition (22) reads, in the present context,

\[
\| -\Delta \omega + F(\cdot, \omega, \nabla \omega) \|_{L_2} \leq \delta, \quad (35)
\]
that is, a bound for an integral is required. Depending on \( F \) and on the concrete representation of \( \omega \), such a bound can be computed either by explicit integration using a computer algebra package (e.g., if \( F \) is a polynomial function and \( \omega \) is piecewise polynomial), or by use of a quadrature formula and a bound for its remainder term [26]; the latter can often be obtained by automatic differentiation techniques. In any case, interval arithmetic [11] has to be used in all numerical evaluations here (but not during the computation of \( \omega \) described above!), in order to take rounding errors into account.

For the computation of \( C, K, \) and \( G \), we restrict ourselves here, for easier presentation, to the case \( n \in \{2, 3\} \) (for \( n = 1 \), see [15]). As indicated in (c) above, we now choose

\[
Y := H_{1,4}(\Omega).
\]

Since \( n \leq 3 \), the embedding \( H_{1,4}(\Omega) \hookrightarrow C(\overline{\Omega}) \) is continuous due to Sobolev’s embedding theorem [1]. Therefore, the norm

\[
\|u\|_Y := \max\{\|u\|_\infty, \gamma \|\nabla u\|_{L_2}\},
\]

with \( \gamma > 0 \) denoting a scaling parameter specified later (in (44)), is equivalent to the usual \( H_{1,4} \)-norm and can therefore be chosen as norm in \( Y \).

### 4.1. Computation of \( C \) and \( K \)

For the computation of \( C \) and \( K \) satisfying (23) and (24) we first need the following results, Theorem 4.1 and Lemma 4.2, which may be regarded as explicit versions of the embeddings \( H_{2}(\Omega) \hookrightarrow C(\overline{\Omega}) \) and \( H_{2}^{2}(\Omega) \hookrightarrow H_{1,4}(\Omega) \).

**Theorem 4.1** (See [16, Corollary 1]). For all \( u \in H_{2}(\Omega) \),

\[
\|u\|_\infty \leq C_0 \|u\|_{L_2} + C_1 \|\nabla u\|_{L_2} + C_2 \|u_{xx}\|_{L_2},
\]

with \( u_{xx} \) denoting the Hessian matrix of \( u \), and with

\[
C_j = \frac{\gamma_j}{\text{vol}(\mathbb{B})} \left[ \max_{x_0 \in \mathbb{B}} \int_{\mathbb{B}} |x - x_0|^{2j} \, dx \right]^{1/2} \quad (j = 0, 1, 2),
\]

where

\[
\gamma_0 = 1, \quad \gamma_1 = 1.1548, \quad \gamma_2 = 0.22361 \quad \text{if } n = 2,
\]

\[
\gamma_0 = 1.0708, \quad \gamma_1 = 1.6549, \quad \gamma_2 = 0.41413 \quad \text{if } n = 3,
\]

and \( \mathbb{B} \subset \mathbb{R}^n \) is a compact and convex set with nonempty interior such that, for each \( x_0 \in \overline{\Omega} \), a congruent image \( \mathbb{B} \) of \( \mathbb{B} \) satisfies \( x_0 \in \mathbb{B} \subset \overline{\Omega} \), i.e., there exist an orthogonal matrix \( T \in \mathbb{R}^{n,n} \) and some \( b \in \mathbb{R}^n \) (both possibly depending on \( x_0 \)) such that

\[
x_0 \in \phi(\mathbb{B}) \subset \overline{\Omega} \quad \text{for } \phi(x) := Tx + b \quad (x \in \mathbb{R}^n).
\]

(In particular, \( \mathbb{B} := \overline{\Omega} \) may be chosen if \( \Omega \) is convex.)
Example 1. If $\Omega$ is a ball of radius $R$, we can choose $\mathcal{Q}$ to be a (closed) ball of radius $R_{\mathcal{Q}}$ and obtain from (38) by straightforward calculations:

\[
C_0 = 0.56419 \cdot \rho^{-1}, \quad C_1 = 0.79789, \quad C_2 = 0.23033 \cdot \rho
\]

if $n = 2$,

\[
C_0 = 0.52319 \cdot \rho^{-3/2}, \quad C_1 = 1.0228 \cdot \rho^{-1/2}, \quad C_2 = 0.37467 \cdot \rho^{1/2}
\]

if $n = 3$.

Example 2. If $\Omega$ is a rectangle with sidelengths $L_1, \ldots, L_n$, we can choose $\mathcal{Q}$ to be a rectangle with sidelengths $l_i \in (0, L_i)$ ($i = 1, \ldots, n$) and obtain from (38):

\[
C_0 = \frac{\gamma_0}{\sqrt{l_1 \cdot l_2(l_3)}}
\]

\[
C_1 = \frac{\gamma_1}{\sqrt{3}} \sqrt{\frac{l_1^2 + l_2^2( + l_3^2)}{l_1 \cdot l_2(l_3)}} \tag{39}
\]

\[
C_2 = \frac{\gamma_2}{3} \cdot \sqrt{\left[\frac{l_1^2 + l_2^2( + l_3^2)}{l_1 \cdot l_2(l_3)}\right]^2 + \frac{4}{3} \left[\frac{l_4^2}{l_1 \cdot l_2(l_3)}\right]^2}
\]

Example 3. To consider finally a nonconvex domain, let $\Omega$ denote the L-shaped domain $(-1, 1)^2 \setminus \{0, 1\}^2 \subset \mathbb{R}^2$. Then $\mathcal{Q}$ may be chosen to be any rectangle with sidelenghts $l_1 \in (0, 1), l_2 \in (0, 2)$, and $C_0, C_1, C_2$ are given by (39).

Lemma 4.2 (See [17, Lemma 4]). For all $u \in H^1_2(\Omega)$,

\[
\|\nabla u\|_{L_4} \leq \|u\|_\infty\|\Delta u\|_{L_2} + 2\|u_{xx}\|_{L_2}. \tag{40}
\]

Observing the right-hand sides of (37) and (40) we now fix the norm in $X = H^1_2(\Omega)$ to

\[
\|u\|_X := \max\left\{C_0\|u\|_{L_2} + C_1\|\nabla u\|_{L_2} + C_2\|u_{xx}\|_{L_2}, \frac{\gamma}{2}(C_0\|u\|_{L_2} + C_1\|\nabla u\|_{L_2} + C_2\|u_{xx}\|_{L_2})^2 + \frac{1}{2\gamma}(\|\Delta u\|_{L_2} + 2\|u_{xx}\|_{L_2})^2\right\} \tag{41}
\]

with $\gamma$ from (36), and with $\gamma > 0$ denoting an additional scaling parameter. Since, due to Theorem 4.1 and Lemma 4.2, the first term in the max in (41) is bigger than $\|u\|_\infty$, and the second is bigger than $\gamma\|\nabla u\|_{L_4}$, we obtain from (36) that $\|u\|_X \leq \|u\|_X$ for all $u \in H^1_2(\Omega)$, i.e., that (23) holds for

$C := 1$. 

For computing some $K$ satisfying (24), suppose that we know constants $K_0, K_1, K_2$, $\kappa$ such that, for all $u \in H^2_T(\Omega)$,

(a) $\|u\|_{L^2} \leq K_0 \|L[u]\|_{L^2},$

(b) $\|\nabla u\|_{L^2} \leq K_1 \|L[u]\|_{L^2},$

(c) $\|u_{tx}\|_{L^2} \leq K_2 \|L[u]\|_{L^2},$

(d) $\|\Delta u\|_{L^2} \leq \kappa \|L[u]\|_{L^2},$

with $L$ denoting the linear operator defined in (20) in the abstract setting; here it reads

$$L[u] = -\Delta u + b \cdot \nabla u + cu,$$

$$b(x) := \frac{\partial F}{\partial z}(x, \omega(x), \nabla \omega(x)), \quad c(x) := \frac{\partial F}{\partial y}(x, \omega(x), \nabla \omega(x)).$$

From (41) and (42) we immediately obtain (24) for

$$K := \max \left\{ C_0 K_0 + C_1 K_1 + C_2 K_2, \right.$$  

$$\gamma \sqrt{\frac{\eta}{2} (C_0 K_0 + C_1 K_1 + C_2 K_2)^2 + \frac{1}{2 \eta} (\kappa + 2 K_2)^2} \bigg\},$$

Fixing now $\eta := (\kappa + 2 K_2)/(C_0 K_0 + C_1 K_1 + C_2 K_2)$ and

$$\gamma := \sqrt{\frac{C_0 K_0 + C_1 K_1 + C_2 K_2}{\kappa + 2 K_2}},$$

we obtain

$$K = C_0 K_0 + C_1 K_1 + C_2 K_2.$$  

We are left to compute the constants in (42). Here, only the computation of $K_0$ will require essential numerical work in the form of eigenvalue bounds to be obtained, while $K_1, K_2$, and $\kappa$ can be calculated rather directly from $K_0$ and the data of the problem. Since also $C_0, C_1, C_2$, and $K$ can (then) be obtained by rather direct formulas ((38) and (45)), the eigenvalue computations form indeed the only time-consuming part in the calculation of $K$ and $C$.

4.1. Computation of $K_0$

Consider the eigenvalue problem for $L^*L$ in weak formulation:

$$u \in H^2_T(\Omega), \quad (L[u], L[\varphi])_{L^2} = \lambda \langle u, \varphi \rangle_{L^2} \quad \text{for all } \varphi \in H^2_T(\Omega).$$

The variational characterization of the smallest eigenvalue $\lambda_1$ of problem (46),

$$\lambda_1 = \min_{u \in H^2_T(\Omega) \setminus \{0\}} \frac{(L[u], L[u])_{L^2}}{(u, u)_{L^2}}.$$
immediately shows that $L$ is one-to-one (as required) if and only if $\lambda_1 > 0$, in which case (42(a)) holds for any

$$K_0 \geq \frac{1}{\sqrt{\lambda_1}}. \quad (47)$$

Thus, we need a positive lower bound for $\lambda_1$ (resp. for the smallest singular value $\sqrt{\lambda_1}$ of $L$). In [23], Rump proposes to use a lower bound for the smallest singular value also in the context of enclosure methods for large sparse systems of equations in $\mathbb{R}^N$. This reflects the similarity of large sparse systems and elliptic differential equations when quantitative questions are under consideration.

There exist efficient computer-assisted methods for computing eigenvalue bounds for self-adjoint problems such as (46). In particular, we mention the (rather simple) Rayleigh–Ritz method providing upper eigenvalue bounds, and the Temple–Lehmann–Goerisch method for obtaining lower bounds. By variational tools and by use of finite-dimensional subspaces (obtained as the linear hull of some approximate eigenfunctions to problem (46), which in turn are computed by numerical linear algebra), these methods reduce the computation of the required bounds to (symmetric) matrix eigenvalue problems, for which eigenvalue bounds can be computed by more direct procedures. However, the Temple–Lehmann–Goerisch method requires knowledge of a certain spectral parameter carrying some rough spectral a priori information about the given problem. This makes it necessary, in many examples, to combine the variational arguments with a homotopy method connecting the given eigenvalue problem to a ‘simple’ one with known eigenvalues. For details, which we are not going to describe here, see [2,22].

In the particular case where the coefficient function $b$ of $L$ (see (43)) satisfies $b = \nabla \phi$ for some Lipschitz-continuous scalar function $\phi$, the following alternative for computing $K_0$ may be used. The operator $L$ is now symmetric on $H_0^2(\Omega)$ with respect to the $L_2$-inner product $\langle \cdot , \cdot \rangle_{\phi^2}$ weighted by $e^{-\phi}$, so that an $\langle \cdot , \cdot \rangle_{\phi^2}$-orthonormal and complete system of eigenfunctions of $L$ exists. Using eigenfunction series expansions one easily derives that, with $\| \cdot \|_{\phi} := \sqrt{\langle \cdot , \cdot \rangle_{\phi^2}}$,

$$\|L[u]\|_{\phi} \geq \sigma \|u\|_{\phi} \quad \text{for all } u \in H_0^2(\Omega),$$

where $\sigma > 0$ is such that

$$\sigma \leq |\lambda| \quad \text{for each eigenvalue } \lambda \text{ of } L. \quad (48)$$

If $\phi, \overline{\phi}$ are constants satisfying $\phi \leq \phi \leq \overline{\phi}$ on $\Omega$, (42(a)) therefore holds for

$$K_0 := \frac{1}{\sigma} \exp \left[ \frac{1}{2} (\overline{\phi} - \phi) \right]. \quad (49)$$

provided that a positive $\sigma$ satisfying (48) can be computed.

So this time we need eigenvalue bounds for $L$ itself, which are often easier to compute than for problem (46), by the variational methods described above. The additional transformation $v = \exp(-\phi/2)u$, which transforms the problem $L[u] = \lambda u$ into
-Δv + \left( \frac{1}{4} |b|^2 - \frac{1}{2} \text{div } b + c \right) v = λv,

further facilitates the application of the variational methods and of the homotopy method. On the other hand, due to the rough estimate 6 and (48) and (49) provide a less accurate constant $K_0$ than the way via problem (46), except when $φ$ is constant, i.e., when $b \equiv 0$.

4.1.2. Calculation of $K_1$, $K_2$ and $κ$

The following three lemmata provide direct formulas for constants $K_1$, $K_2$, and $κ$ satisfying (42(b)–(d)). We assume here that the coefficient functions $b$ and $c$ of $L$ are bounded on $Ω$; due to (43) and to the continuity of $∂F/∂y$ and $∂F/∂z$, this assumption is satisfied for all ‘usual’ approximations $ω$. The proofs of Lemmata 4.3 and 4.4, even of more general versions, can be found in [17, Lemmata 5 and 6]; however, in the simple version presented here, the proof of Lemma 4.4 is trivial.

For Lemma 4.5 we assume in addition that $\partial Ω$ is a piecewise $C_2$-hypersurface, i.e., that some measure-zero-subset $Z ⊆ \partial Ω$ exists such that $\partial Ω \setminus Z$ is a relatively open subset of $\partial Ω$ and a $C_2$-hypersurface of $\mathbb{R}^n$. Consequently, the mean curvature $H$ with respect to the outer unit normal $v$ is defined almost everywhere on $\partial Ω$. The proof of Lemma 4.5 can be found in [16, Lemmata 1–3]; see also [17, Lemma 7].

Lemma 4.3. Let (42(a)) hold with some constant $K_0$, and let $c$ and $b$ denote a lower bound for $c$, and an upper bound for $|b|$, respectively. Then (42(b)) holds with

$$K_1 := \begin{cases} \frac{1}{2 \sqrt{\lambda - b}} & \text{if } c > 0 \text{ and } (2 \sqrt{\lambda - b}) \cdot \sqrt{\lambda} K_0 \geq 1, \\ \frac{1}{2} \bar{b} K_0 + \sqrt{\left( \frac{1}{4} \bar{b}^2 - c \right) \frac{\lambda^2 K_0^2}{4} + K_0} & \text{otherwise}. \end{cases}$$

Lemma 4.4. Let (42(a),(b)) hold with constants $K_0$ and $K_1$, respectively. Then (42(d)) is true for

$$κ := 1 + \|b\|_∞ K_1 + \|c\|_∞ K_0.$$

Lemma 4.5. Let (42(a),(b),(d)) hold with constants $K_0$, $K_1$, and $κ$. Suppose that a Lipschitz-continuous function $g : Ω \rightarrow \mathbb{R}^n$ exists such that $g \cdot v \geq (n - 1) H$ a.e. on $\partial Ω$, and that nonnegative constants $G_0$, $G_1$ are known such that

$$|g| \leq G_0, \quad -\text{div } g + λ_{\max} \left[ J(g) + J(g)^T \right] \leq G_1 \quad \text{a.e. on } Ω,$$

with $J(g)$ denoting the Jacobian matrix of $g$, and $λ_{\max}$ indicating the maximal eigenvalue. Then (42(c)) holds with

$$K_2 := [κ^2 + 2κ G_0 K_1 + G_1 K_1]^{1/2}.$$
Corollary 4.6. If $\Omega$ is convex and (42(d)) holds for some constant $\kappa$, then (42(c)) is true for $K_2 := \kappa$.

Proof. For convex domains, the mean curvature $H$ is nonpositive a.e. on $\partial \Omega$, so that the conditions of Lemma 4.5 hold true with $g \equiv 0$, $G_0 = G_1 = 0$. □

4.2. Construction of $G$, smoothness of $F$

For computing a monotonically nondecreasing function $G : [0, \infty) \to [0, \infty)$ satisfying (25) and (26) (with $\| \cdot \|_T$ given by (36) and (44)) we assume now that $F$ satisfies condition (33) for $p = 4$ (as indicated in (c) at the beginning of Section 4), i.e., $F$ satisfies a quadratic growth condition at $\infty$ with respect to $z$.

Moreover, we assume that $V \omega$ is bounded on $\Omega$, which is satisfied for all usual numerical approximations $\omega \in H^2_0(\Omega)$. ($\omega$ itself is anyway bounded due to the embedding $H^2_0(\Omega) \hookrightarrow C(\overline{\Omega})$.)

First we construct a function $\tilde{G} : [0, \infty) \times [0, \infty) \to [0, \infty)$ with the following properties:

(G1) $|F(x, \omega(x) + y, \nabla \omega(x) + z) - F(x, \omega(x), \nabla \omega(x)) - b(x) 
\cdot z - c(x)y| \leq \tilde{G}(|y|, |z|)$ for $x \in \Omega$, $y \in \mathbb{R}$, $z \in \mathbb{R}^n$;

(G2) $\tilde{G}(\mu t, t) = o(t)$ for $t \to 0$ and each fixed $\mu > 0$;

(G3) $\tilde{G}$ is monotonically nondecreasing in both variables;

(G4) for each fixed $\alpha \geq 0$, $[\tilde{G}(\alpha, t^{1/4})]^2$ is a continuous and concave function of $t$.

Possibly the easiest way to obtain such a function $\tilde{G}$ is to look for it in the form

$$\tilde{G}(s, t) = \sum_{k=1}^{M} g_k(s)t^{\mu_k}$$

with exponents $\mu_k \in [0, 2]$ and monotonically nondecreasing functions $g_k$ satisfying $g_k(s) = o(s^{1-\mu_k})$ (for $s \to 0$), so that properties (G2)–(G4) are fulfilled, and to arrange the integer $M$, the exponents $\mu_k$, and the functions $g_k$ appropriately in order to satisfy property (G1), too. Observe that the exponent restriction $\mu_k \in [0, 2]$, which is necessary for condition (G4), requires the quadratic growth restriction (33) (for $p = 4$), via condition (G1).

The existence of a function $\tilde{G}$ satisfying (G1)–(G4) follows in general (without relying on the success of the ansatz (50)) by the following arguments, which may also be used, alternatively to (50), for constructing $\tilde{G}$ concretely: let $\tilde{G}(s, t)$ denote the supremum of the left-hand side of the inequality in (G1), taken over all $x \in \overline{\Omega}$, $y \in \mathbb{R}$, $z \in \mathbb{R}^n$ such that $|y| \leq s$, $|z| \leq t$. Obviously, $\tilde{G}$ satisfies (G1) and (G3), and also (G2) follows for $\tilde{G}$ due to the boundness of $\omega$ and $V \omega$, and to the uniform
continuity of $\partial F/\partial y$ and $\partial F/\partial z$ on compact sets. A similar argument also provides the continuity of $\tilde{G}$. Now it can be shown that the function

$$G(s, t) := \sup \left\{ \left( \tilde{G}(s, t_1)^2 - t_1^4 + \tilde{G}(t, t_2)^2 - t_2^4 \right)^{1/2} : 0 \leq t_1 < t_2 < \infty, t_1 \leq t \leq t_2 \right\}$$

(observe that the supremum is finite since $\tilde{G}(s, t_2)$ grows at most quadratically with respect to $t_2$ as $t_2 \to \infty$) has all properties (G1)–(G4). In fact, $\tilde{G}$ is the smallest function (in the pointwise sense) which has all these properties. Properties (G1)–(G4) provide the following:

**Lemma 4.7.** Let $g(x, y, z)$ denote the term in modulus on the left-hand side of the inequality in (G1). Then, for each $u \in H_{1,4}(\Omega)$, $g(\cdot, u, \nabla u) \in L_2(\Omega)$ and

$$\|g(\cdot, u, \nabla u)\|_{L_2(\Omega)} \leq \sqrt{\text{vol}(\Omega)} \cdot \tilde{G}(\|u\|_{L_4}, \|\nabla u\|_{L_4}).$$

**Proof.** Let $u \in H_{1,4}(\Omega)$ and $\psi(t) := [\tilde{G}(\|u\|_{L_4}, t^{1/4})]^2$. Due to (G4), $\psi$ is continuous and concave. Moreover, let $v := |\nabla u|^4$. Properties (G1) and (G3) provide, for almost all $x \in \Omega$,

$$|g(x, u(x), \nabla u(x))|^2 \leq [\tilde{G}(\|u(x)\|_{L_4}, |\nabla u(x)|)]^2 \leq [\tilde{G}(\|u\|_{L_4}, v(x)^{1/4})]^2 = \psi(v(x)),$$

and Jensen’s inequality yields, since $v \in L_1(\Omega)$ and $\psi$ is continuous, concave, and nonnegative:

$$\psi \circ v \in L_1(\Omega) \quad \text{and} \quad \frac{1}{\text{vol}(\Omega)} \int_D \psi(v(x)) \, dx \leq \psi \left( \frac{1}{\text{vol}(\Omega)} \int_D v(x) \, dx \right).$$

Consequently, $|g(\cdot, u, \nabla u)|^2 \in L_1(\Omega)$ and

$$\int_D |g(\cdot, u, \nabla u)|^2 \, dx \leq \text{vol}(\Omega) \cdot \psi \left( \frac{1}{\text{vol}(\Omega)} \int_D v(x) \, dx \right) = \text{vol}(\Omega) \left[ \tilde{G}(\|u\|_{L_4}, \|\nabla u\|_{L_4}) \right]^2.$$

By (36) and (G3), $\tilde{G}(\|u\|_{L_4}, \text{vol}(\Omega)^{-1/4} \|\nabla u\|_{L_4}) \leq \tilde{G}(\|u\|_{L_4}, \|\nabla u\|_{L_4})$ for $u \in H_{1,4}(\Omega)$. Thus, Lemma 4.7 and property (G2) show that $\mathcal{F} : H_{1,4}(\Omega) \to L_2(\Omega)$ is Fréchet-differentiable at $u$, with $\mathcal{F}'(u)[u] = b(x) \cdot \nabla u + c(x) u$ (this was already used in (43)), and furthermore, that (25) and (26) are satisfied for

$$G(t) := \sqrt{\text{vol}(\Omega)} \cdot \tilde{G}(t, \text{vol}(\Omega)^{-1/4} \gamma^{-1} t). \quad (51)$$
Moreover, \( \mathcal{F} \) is continuous. (In fact, \( \mathcal{F}(\omega + u) - \mathcal{F}(\omega) - \mathcal{F}'(\omega)[u] \) is continuous due to [17, Lemma 2], which implies the continuity of \( \mathcal{F} \).) According to Remark 3.1, \( \mathcal{F} \) has therefore the required smoothness properties.

### 4.3. Numerical examples

To test the strong solution approach in the frame of our existence and enclosure method, several examples have been treated. Some of the results will be presented here. Our first example is the boundary value problem

\[
-\Delta u = u \left( \lambda - \frac{1}{2} |\nabla u|^2 \right) \quad \text{on} \quad \Omega := (0, 1)^2, \quad u = 0 \quad \text{on} \quad \partial \Omega, \tag{52}
\]

which has an infinite number of possible bifurcation points (from the trivial solution \( u \equiv 0 \)) at the Dirichlet-eigenvalues \( \lambda_{k,l} = (k^2 + l^2)\pi^2 \) of \( -\Delta \). For several values of \( \lambda \), we computed approximate solutions \( \omega \) on the first two nontrivial branches (bifurcating from zero at \( \lambda_{1,1} = 2\pi^2 \) and at \( \lambda_{2,1} = \lambda_{1,2} = 5\pi^2 \)), by a finite element method with bi-quintic rectangular \( (C_1-) \) elements, combined with a Newton-iteration and path-following methods; the numerical process was started, at \( 2\pi^2 \) and at \( 5\pi^2 \), with appropriate multiples of eigenfunctions corresponding to the eigenvalues \( 2\pi^2 \) and \( 5\pi^2 \) of \( -\Delta \). By exploiting symmetries of the (expected) solutions, we could (i) avoid trivial nonuniqueness arising from symmetry transformations, (ii) reduce the numerical computations to the subdomains \( (0, 1/2)^2 \) (first branch) and \( (0, 1/4) \times (0, 1/2) \) (second branch), respectively, (iii) remove several eigenvalues of the corresponding problem (46), which facilitates the eigenvalue enclosure methods described in Section 4.1.1.

With \( 9 \times 9 \) bi-quintic elements for the first, and \( 7 \times 14 \) elements for the second branch, our existence and enclosure method was successful for \( \lambda \) up to \( 3.7\pi^2 \) on the first, and up to \( 7.8\pi^2 \) on the second branch; see Fig. 2. In immediate neighborhoods of the bifurcation points, the computed constants \( K \) (see (24) and (45)) are—as
Table 1

Existence and enclosure results for problem (52)

| $\lambda/k^2$ | $||w||_\infty$ | $\delta$ | $K$ | $\alpha$ |
|---------------|----------------|---------|-----|---------|
| **First branch** |               |         |     |         |
| 2.0001        | 0.02309   | 0.3059E-7 | 2030.0 | 0.6615E-4 |
| 2.001         | 0.07303   | 0.9676E-7 | 201.5  | 0.1963E-4 |
| 2.01          | 0.2309    | 0.3235E-6 | 20.51  | 0.6644E-5 |
| 2.1           | 0.7277    | 0.3475E-5 | 2.441  | 0.8490E-5 |
| 2.5           | 1.611     | 0.5503E-4 | 0.9232 | 0.5095E-4 |
| 3.0           | 2.264     | 0.2774E-3 | 0.9377 | 0.2646E-3 |
| 3.5           | 2.767     | 0.8215E-3 | 1.273  | 0.1204E-2 |
| 3.7           | 2.945     | 0.1159E-2 | 1.500  | 0.3047E-2 |
| 3.8           | 3.030     | 0.1354E-2 | 1.639  | -        |
| **Second branch** |           |         |     |         |
| 5.0001        | 0.01461   | 0.1500E-6 | 2438.0 | -        |
| 5.001         | 0.04619   | 0.4743E-6 | 230.9  | 0.1173E-3 |
| 5.01          | 0.1460    | 0.1525E-5 | 23.14  | 0.3553E-4 |
| 5.1           | 0.4608    | 0.9968E-5 | 2.488  | 0.2486E-4 |
| 5.5           | 1.021     | 0.1130E-5 | 0.6649 | 0.7542E-4 |
| 6.0           | 1.427     | 0.4424E-3 | 0.4628 | 0.2069E-3 |
| 6.5           | 1.728     | 0.1145E-2 | 0.4184 | 0.4923E-3 |
| 7.0           | 1.974     | 0.2406E-2 | 0.4184 | 0.1084E-2 |
| 7.5           | 2.184     | 0.4375E-2 | 0.4406 | 0.2382E-2 |
| 7.8           | 2.297     | 0.5915E-2 | 0.4615 | 0.4959E-2 |
| 7.9           | 2.333     | 0.6484E-2 | 0.4696 | -        |

expected—very large, since the linearization $L$ (see (43)) is almost singular there; this caused failure of our method for $\lambda = 5.0001\pi^2$ on the second branch. Table 1 contains the approximate sizes of $||w||_\infty$, the defects bounds $\delta$ (see (22) and (35)), the constants $K$ (see (24) and (45)), and the error bounds $\alpha$ (see (28) and (29)).

We wish to remark that, on the second branch, the linearized operator $L$ is not inverse-positive, so that monotonicity methods cannot be applied there.

In our second example, we consider the problem

\[
\Delta u + u^2 = \lambda \sin(\pi x_1) \sin(\pi x_2) \quad \text{on } \Omega := (0, 1)^2, \quad u = 0 \quad \text{on } \partial \Omega.
\]

(53)

The results presented here are joint work with McKenna and Breuer and will be reported in detail in a paper which is presently in preparation; see also [3].

In the PDE-community it has been an open question since many years if problem (53) has at least four solutions for sufficiently large values of $\lambda$. Apparently, this question could not be answered by purely analytical means. By our existence and enclosure method, combined with a numerical mountain pass method (see [5]), we could give a positive answer, at least for the particular value $\lambda = 800$. 
The numerical mountain pass method was used to find four essentially different approximate solutions, which were then improved by a Newton iteration, where the linear problems in the Newton steps were treated by a collocation method with \( \sim 16,000 \) trigonometric basis functions. In this way, we arrived at highly accurate approximations \( \omega_i \) (\( i = 1, \ldots, 4 \)) with defect bounds \( \delta_i \) (see (22), (35)) in order of magnitude of 0.001 to 0.01. Since the eigenvalues of \( L_i = -\Delta - 2\omega_i \) turned out to be well separated from zero for all four approximations \( \omega_i \), our existence and enclosure method was successful in proving the existence of four solutions \( u_i \) (\( i = 1, \ldots, 4 \)) of problem (53) such that \( \|u_i - \omega_i\|_\infty \leq \alpha_i \) (\( i = 1, \ldots, 4 \)), with error bounds \( \alpha_1, \ldots, \alpha_4 \) between \( 5 \times 10^{-4} \) and \( 5 \times 10^{-2} \). Since simple computations show \( \|\omega_i - \omega_j\|_\infty > \alpha_i + \alpha_j \) for \( i \neq j \), the four solutions \( u_1, \ldots, u_4 \) are indeed pairwise different.

It should be remarked that the linearization \( L_1 \) in the approximation \( \omega_1 \) has only positive eigenvalues, while \( L_2, L_3, \) and \( L_4 \) have also negative eigenvalues, so that monotonicity methods could not be used for our purpose.

Fig. 3 shows plots of the four solutions. \( u_1 \) and \( u_2 \) are ‘fully’ symmetric (i.e., with respect to reflection at the axes \( x_1 = 1/2, x_2 = 1/2, x_1 = x_2 \), and \( x_1 = 1 - x_2 \)), while \( u_3 \) is symmetric only with respect to \( x_2 = 1/2 \), and \( u_4 \) only with respect to \( x_1 = x_2 \). (Of course, further approximations resp. solutions arise from \( \omega_3 \) and \( \omega_4 \) resp. \( u_3 \) and \( u_4 \) by rotations.)
5. Weak solutions

Besides providing the strong solution approach, our abstract operator setting is also appropriate for yielding existence and enclosure results for weak solutions to problem (1). For this purpose, we choose now

\[ X := H^1_0(\Omega), \quad Z := H_{-1}(\Omega) \]  

(54)

(where \( H^1_0(\Omega) \) is endowed with the norm \( \|u\|_X = \|\nabla u\|_{L^2} \), and \( H_{-1}(\Omega) \) denotes the topological dual space of \( H^1_0(\Omega) \)). The operators \( L_0 \) and \( \mathcal{F} \) are chosen as before in (31) (where for \( u \in H^1_0(\Omega) \), the distribution \( -\Delta u \in H_{-1}(\Omega) \) is defined as usual by \( (-\Delta u)(\varphi) := \int_\Omega \nabla u \cdot \nabla \varphi \, dx \) for all \( \varphi \in H^1_0(\Omega) \)). Here, we assume that \( F \) is independent of \( z \) resp. \( \nabla u \), and moreover that \( F \) has subcritical growth with respect to \( y \), i.e., that some \( C > 0 \) and some \( r \geq 1 \) exist satisfying

\[ (n - 2)r < n + 2, \quad |F(x, y)| \leq C(1 + |y|^r) \]

for all \( x \in \overline{\Omega}, \ y \in \mathbb{R} \).

Then, some \( p > 1 \) exists such that

\[ \frac{n - 2}{2n} < \frac{1}{p} < \frac{n + 2}{2nr}. \]  

(56)

Here, the first inequality ensures that the embedding \( H^1_0(\Omega) \hookrightarrow L_p(\Omega) \) is compact, and the second will be shown in Section 5.3 to provide, together with (55), the required smoothness properties of \( \mathcal{F} : L_p(\Omega) \to H_{-1}(\Omega) \). Consequently, in our abstract setting we can choose

\[ Y := L_p(\Omega). \]  

(57)

The abstract regularity assumption (B) requires here that the Poisson equation (34) has a unique solution \( u \in H^1_0(\Omega) \) for each \( r \in H_{-1}(\Omega) \). By the Riesz representation lemma for bounded linear functionals, this is true for every domain \( \Omega \). We do not need further regularity conditions; in particular, also domains \( \Omega \) with reentrant corners are allowed.

Another advantage of the weak solutions approach is that the approximate solution \( \omega \) need only be in \( H^1_0(\Omega) \) here. Thus, in the finite element context, \( C_0 \)-elements are sufficient!

In the following, we assume that \( \omega \) is bounded on \( \Omega \), which is satisfied for all usual numerical approximations \( \omega \).

Again, we comment now on the computation of the terms \( \delta, C, K \), and \( G \) satisfying (22)–(26).

5.1. Computation of a defect bound \( \delta \)

The defect \( -\Delta \omega + F(\cdot, \omega) \) has to be bounded, according to (22) and (54), in the \( H_{-1} \)-norm, which under some aspects is a bit more involved than the simple \( L_2 \)-bound needed for strong solutions. The direct definition of the dual space norm gives
a supremum over \( \mathcal{H}^1_0(\Omega) \setminus \{0\} \) and is therefore not well suited for the computation of an upper bound.

However, the following simple (but very fruitful) complementary variational characterization of the \( H^{-1} \)-norm removes this difficulty:

**Lemma 5.1.** For all \( w \in H^{-1}(\Omega) \),

\[
\|w\|_{H^{-1}} = \sup_{\varphi \in \mathcal{H}^1_0(\Omega) \setminus \{0\}} \left\{ \|\nabla \varphi\|_{L^2}^{-1} [\text{div } \sigma][\varphi] \right\}
\]

(where, as usual, \( [\text{div } \sigma][\varphi] := - \int_{\Omega} \sigma \cdot \nabla \varphi \, dx \)). Moreover, the minimum in (58) is attained at \( \sigma^* := - \nabla u \), where \( u \in H^1_0(\Omega) \) solves \( -\Delta u = w \), and thus, at the unique solution \( \sigma^* \in L^2(\Omega)^n \) of the equations

\[
\text{div } \sigma = w, \quad \int_{\Omega} \sigma \cdot \tau \, dx = 0
\]

for all \( \tau \in L^2(\Omega)^n \) such that \( \text{div } \tau = 0 \).

**Proof.** For each \( \sigma \in L^2(\Omega)^n \) satisfying \( \text{div } \sigma = w \),

\[
\|w\|_{H^{-1}} = \sup_{\varphi \in \mathcal{H}^1_0(\Omega) \setminus \{0\}} \left\{ \|\nabla \varphi\|_{L^2}^{-1} [\text{div } \sigma][\varphi] \right\}
\]

and equality holds if \( \sigma \) is the gradient of an \( H^1_0 \)-function, and thus, for \( \sigma = \sigma^* \).

Applying Lemma 5.1 to the specific case \( w = -\Delta \omega + F(\cdot, \omega) \) (and transforming \( \sigma = - \nabla \omega + q \) in (58)), we obtain

\[
\| - \Delta \omega + F(\cdot, \omega) \|_{H^{-1}} = \min \left\{ \|\nabla \omega - q\|_{L^2} : q \in L^2(\Omega)^n, \text{div } q = F(\cdot, \omega) \right\},
\]

and moreover, that the minimum in (60) is attained at the unique \( q \in L^2(\Omega)^n \) satisfying \( \text{div } q = F(\cdot, \omega) \) and

\[
\int_{\Omega} q \cdot \tau \, dx = 0 \quad \text{for all } \tau \in L^2(\Omega)^n \text{ such that } \text{div } \tau = 0.
\]

Thus, a defect bound \( \delta \) can be obtained by computing a \( q \in L^2(\Omega)^n \) satisfying the side condition \( \text{div } q = F(\cdot, \omega) \) exactly, but solving Eq. (61) only approximately, and then computing an upper bound \( \delta \) for \( \|\nabla \omega - q\|_{L^2} \) (by verified integration), which is the desired defect bound according to (60). \( q \) can be computed, e.g., in the form

\[
q = \tilde{q} + q_0,
\]

where

\[
\tilde{q} := (\tilde{q}_1, 0, \ldots, 0),
\]

\[
\tilde{q}_1(x) := \int_{s(x_2, \ldots, x_n)}^x F(t, x_2, \ldots, x_n, \omega(t, x_2, \ldots, x_n)) \, dt
\]
(provided that \( \Omega \) allows the definition of such line integrals), and \( q_0 \) satisfies \( \text{div} \ q_0 = 0 \) (exactly) and is an approximate solution of the inhomogeneous problem

\[
\int_{\Omega} q_0 \cdot \tau \ dx = -\int_{\Omega} \tilde{q} \cdot \tau \ dx
\]

for all \( \tau \in L^2(\Omega)^n \) such that \( \text{div} \ \tau = 0 \).

If \( n = 2 \) and \( \Omega \) is simply connected, \( q_0 \) can be put up in the form

\[
q_0 = \frac{\dot{\psi}}{\partial x_1} + \frac{\ddot{\psi}}{\partial x_2},
\]

with \( \psi \in H^1(\Omega) \), and (63) amounts to a Neumann boundary value problem for \( \psi \) (to be solved approximately).

At least from the technical point of view, this approach has the disadvantage that an exact solution of the equation \( \text{div} \ q \cdot \div D \) (!) is needed; e.g., the computation of the line integrals in (62) may be rather tedious. The following modification avoids this difficulty:

Let \( \tilde{q} \in L^2(\Omega)^n \) satisfy both (61) and the side condition \( \text{div} \ q = F(\cdot, \omega) \) only approximately (or, what nearly amounts to the same if \( \omega \) is a ‘good’ approximation to (1), let \( \tilde{q} \) approximate \( \nabla \omega \), such that, as the only ‘sharp’ requirement,

\[
\text{div} \ \tilde{q} \in L^2(\Omega) \quad \text{(i.e.,} \quad \int_{\Omega} \tilde{q} \cdot \nabla \omega \ dx \leq \text{const.} \| \varphi \|_{L^2} \text{ for all } \varphi \in H^1_0(\Omega) \).
\]

Then, using (60) and, in the last equality, Lemma 5.1 again, we obtain

\[
\| -\Delta \omega + F(\cdot, \omega) \|_{H^{-1}} \leq \| \nabla \omega - \tilde{q} \|_{L^2} + \text{min} \left\{ \| \tilde{q} - q \|_{L^2} : q \in L^2(\Omega)^n, \ \text{div} \ q = F(\cdot, \omega) \right\}
\]

\[
= \| \nabla \omega - \tilde{q} \|_{L^2} + \text{min} \left\{ \| \sigma \|_{L^2} : \sigma \in L^2(\Omega)^n, \ \text{div} \ \sigma = \text{div} \ \tilde{q} - F(\cdot, \omega) \right\}
\]

\[
= \| \nabla \omega - \tilde{q} \|_{L^2} + \| \text{div} \ \tilde{q} - F(\cdot, \omega) \|_{H^{-1}}.
\]

Now suppose further that a constant \( \widehat{c} \) is known which satisfies the Poincaré inequality

\[
\| u \|_{L^2} \leq \widehat{c} \| \nabla u \|_{L^2} \quad \text{for all } u \in H^1_0(\Omega).
\]

Such a constant can be computed via a lower bound for the smallest Dirichlet eigenvalue of \( -\Delta \) on \( \Omega \), or by formula (69) of Section 5.2. From (66) we obtain, for each \( v \in L^2(\Omega) \),

\[
\| v \|_{H^{-1}} = \sup_{\varphi \in H^1_0(\Omega) \setminus \{0\}} \left\{ \frac{\| \nabla \varphi \|_{L^2}^{-1} \int_{\Omega} v \varphi \ dx}{ \| v \|_{L^2} \} \right\}
\]

\[
\leq \sup_{\varphi \in H^1_0(\Omega) \setminus \{0\}} \left\{ \frac{\| \nabla \varphi \|_{L^2}^{-1} \| \varphi \|_{L^2}}{ \| v \|_{L^2} } \right\} \cdot \| v \|_{L^2} \leq \widehat{c} \| v \|_{L^2}.
\]

Combination with (65) finally provides (regard (64))
\[ \| -\Delta \omega + F(\cdot, \omega) \|_{H^{-1}} \leq \| \nabla \omega - \tilde{q} \|_{L^2} + \| \text{div} \tilde{q} - F(\cdot, \omega) \|_{L^2}, \] (67)

which is a result of the desired kind since the right-hand side can be bounded by verified integration.

A function \( \tilde{q} \in L^2(\Omega) \) satisfying (64) and solving Eqs. (61) and \( \text{div} \, q = F(\cdot, \omega) \) approximately (resp. approximating \( \nabla \omega \)) can be obtained in different ways. For example, in the finite element context, condition (64) requires (besides smoothness on each element) that the normal component of \( \tilde{q} \) is continuous across element edges; this property is provided, e.g., by Raviart–Thomas elements. So one can compute an approximation to \( \omega \) in an appropriate space of Raviart–Thomas elements (or even of \( H^1(\Omega) \)-elements), to obtain \( \tilde{q} \). In a very elegant way, \( \tilde{q} \) is provided if one uses mixed finite element methods from the beginning, i.e., for problem (1), since these methods yield approximations \( \omega \) to the solution \( u \) and \( \tilde{q} \) to the flux \( \nabla u \) independently; so no further computation is necessary if the mixed method provides \( \omega \) in \( H^1_0(\Omega) \) and \( \tilde{q} \) in a Raviart–Thomas element space (or in another space ensuring (64)).

### 5.2. Computation of \( C \) and \( K \)

To satisfy (23) we need a constant \( C \) such that

\[ \| u \|_{L^2} \leq C \| \nabla u \|_{L^2} \quad \text{for all} \quad u \in H^1_0(\Omega). \] (68)

There exist several approaches to this problem in the literature. A constant \( C \) which is (in general) not optimal but simple is given in [18, Lemma 5.1]:

\[
C = \frac{1}{2\sqrt{2}} \sqrt{\text{vol}(\Omega)^{1/p} \left[ \prod_{j=0}^{p-1} \left( \frac{p}{2} - j \right) \right]^{1/p}} \quad \text{if} \quad n = 2 \quad \text{and} \quad p \geq 2,
\]

\[
C = \frac{n-1}{\sqrt{n(n-2)}} \sqrt{\text{vol}(\Omega)^{\frac{p-1}{2}}} \quad \text{if} \quad n \geq 3,
\] (69)

where \( n \) is the largest integer \( \leq p/2 \). If \( p < 2 \) (and \( n = 2 \)), Hölder's inequality shows that the above formula is still correct when the term in brackets (including the exponent) is replaced by 1.

For the analysis of (24) we first observe that the linearization \( L \) is here (formally) again given by (43), with \( b \equiv 0 \) now, and that (24) now reads

\[ \| \nabla u \|_{L^2} \leq K \| L[u] \|_{H^{-1}} \quad \text{for all} \quad u \in H^1_0(\Omega). \] (70)

By rather straightforward calculations (see [18]) one arrives at the eigenvalue problem

\[ u \in H^1_0(\Omega), \quad L[u] = \lambda (-\Delta u) \] (71)

and obtains that (70) holds for

\[ K := \left[ \inf \left\{ \| \lambda \| : \lambda \text{ eigenvalue of problem (71)} \right\} \right]^{-1}, \]

so that again eigenvalue bounds, which can be obtained by the methods described in Section 4.1.1, are required to compute \( K \). Observe that the eigenvalues of (71)
(counted by multiplicity) accumulate at 1, which may complicate the computation of eigenvalue bounds.

In the particular case \( c > 0 \) (which is, roughly speaking, the ‘maximum principle case’), the Rayleigh quotient

\[
\int_D (|\nabla u|^2 + cu^2) \, dx \bigg/ \int_D |\nabla u|^2 \, dx
\]

of problem (71) is always \( > 1 \), so that all eigenvalues are \( > 1 \), which provides \( K = 1 \) without any further computation. This may be regarded as reflection of the ‘simplicity’ of the ‘maximum principle case’, which in turn is closely related to the cases where the monotonicity methods described in Section 2 are successful.

5.3. Construction of \( G \), smoothness of \( \mathcal{F} \)

A monotonically nondecreasing function \( G \) satisfying (25), i.e.,

\[
\|F(\cdot, \omega + u) - F(\cdot, \omega) - cu\|_{H^{-1}} \leq G(\|u\|_{L_p}) \quad \text{for all } u \in L_p(\Omega),
\]

and condition (26), is obtained in a similar way as in the strong solution approach (see Section 4.2). First we calculate a function \( \tilde{G} : [0, \infty) \to [0, \infty) \) with the following properties:

(G1) \(|F(x, \omega(x) + y) - F(x, \omega(x)) - c(x)y| \leq \tilde{G}(|y|) \) for \( x \in \Omega, \ y \in \mathbb{R} \), (where \( c(x) = \frac{PF}{\omega}(x, \omega(x)) \)),

(G2) \( \tilde{G}(t) = o(t) \) for \( t \to 0 \),

(G3) \( \tilde{G} \) is monotonically nondecreasing,

(G4) \( \tilde{G}(t^{1/r}) \) is a continuous and concave function of \( t \) (where \( r \) is the growth exponent satisfying (55)).

Such a function \( \tilde{G} \) can be obtained in a similar way as described in Section 4.2 for the function \( \bar{G} \) which was under consideration there. Then, using the embedding constant \( C = C(p) \) satisfying (68), we obtain from (G1), for all \( u \in L_p(\Omega), \)

\[
\|F(\cdot, \omega + u) - F(\cdot, \omega) - cu\|_{H^{-1}} \\
\leq \sup_{\varphi \in H^1_0(\Omega) \setminus \{0\}} \left\{ \|\nabla \varphi\|_{L^2}^{-1} \int_\Omega \tilde{G}(|u|) |\varphi| \, dx \right\} \\
\leq \sup_{\varphi \in H^1_0(\Omega) \setminus \{0\}} \left\{ \|\nabla \varphi\|_{L^2}^{-1} \|\varphi\|_{L_{p/(p-r)}} \cdot \|\tilde{G}(|u|)\|_{L_{p/r}} \right\} \\
\leq C(p/(p - r)) \cdot \|\tilde{G}(|u|)\|_{L_{p/r}}.
\]

Furthermore, \( \psi(t) := \tilde{G}(t^{1/r})^{p/r} \) is concave (which follows from (G4) by straightforward calculations), so that Jensen’s inequality provides

\[
\|\tilde{G}(|u|)\|_{L_{p/r}} \leq \text{vol}(\Omega)^{r/p} \cdot \tilde{G}(\text{vol}(\Omega)^{-1/p} \cdot \|u\|_{L_p}).
\]
By (73), (74), and (G2), \( \mathcal{F} \) is Fréchet differentiable at \( \omega \), with \( \mathcal{F}'(\omega)[u] = c \cdot u \), and moreover, (25) (resp. (72)) and (26) hold for
\[
G(t) := C \frac{p}{(p - r)} \operatorname{vol}(\Omega)^{1/p} \cdot \tilde{G}(\operatorname{vol}(\Omega)^{-1/p} \cdot t).
\]
Moreover, \( \mathcal{F} : L_p(\Omega) \to H^{-1}(\Omega) \) is continuous (see [18, Lemma 2.2], and use that, due to (56), the embedding \( L_p(\Omega) \hookrightarrow H^{-1}(\Omega) \) is bounded). Due to Remark 3.1, \( \mathcal{F} \) has therefore the required smoothness properties.

5.4. Numerical examples

As a first test example for our weak solution approach, we looked for a nontrivial solution of Emden's equation
\[
-\Delta u = u^2 \quad \text{on } \Omega := (0, 1)^2, \quad u = 0 \quad \text{on } \partial \Omega.
\] (75)
Again, we used a Newton-iteration in combination with finite element methods to compute an approximation \( \omega \in H^0(\Omega) \). Here, however, we chose quadratic triangular elements, since \( C_0 \)-elements are sufficient for the weak solution approach. (In principle, also linear elements could have been used, which however have turned out to provide too poor accuracy.) Exploiting reflection symmetries of the expected solution, we performed the actual numerical computations only on the triangle \( T := \{(x_1, x_2) \in \mathbb{R}^2 : 1/2 \leq x_2 \leq x_1 \leq 1\} \), with Dirichlet boundary conditions posed at \( x_1 = 1, 1/2 \leq x_2 \leq 1 \).

To obtain a starting approximation for the Newton iteration, we embedded problem (75) into the one-parameter-family \(-\Delta u = u^2 + \lambda \) and followed a path of approximate solutions starting at \( \lambda = 0, u \equiv 0 \), going up to a turning point at \( \lambda \approx 20 \), and returning to \( \lambda = 0 \), which yields an approximation to a nontrivial solution of (75).

Fig. 4 shows a plot of the computed approximate solution \( \omega \) to problem (75). Its maximum is attained in the midpoint \((1/2, 1/2)\) of \( \Omega \) and equals approximately 29.26. For three different mesh sizes (64 uniformly distributed triangular elements corresponding to 136 global variables, 256 elements corresponding to 528 variables, and 1024 elements corresponding to 2080 variables), Table 2 contains the computed defect bounds \( \delta \) (obtained via (60), (62), (63)), the constants \( K \) satisfying (24) resp. (70), and the error bounds \( \alpha \) for \( \|\nabla(u - \omega)\|_{L^2} \) provided by (28). It is remarkable that 64 elements are already sufficient to prove the existence of a nontrivial solution to Emden’s equation, even if the computed error bound \( \alpha \) is not very accurate. It should be noted that the linear operator \( L \) has one negative eigenvalue here, so that monotonicity methods are not applicable.

In the second example, the L-shaped domain \( \Omega := (0, 1)^2 \setminus ([1/2, 1] \times (0, 1/2)) \) is considered; the problem is
\[
-\Delta u - 56u = u^2 + \lambda \quad \text{on } \Omega, \quad u = 0 \quad \text{on } \partial \Omega.
\] (76)
The factor $-56$ is chosen to ‘spoil’ (like any other sufficiently negative factor) the applicability of the monotonicity methods on the expected solution branch passing through $(\lambda = 0, \mu \equiv 0)$.

For several values of $\lambda$, we computed approximate solutions on a uniform mesh of 768 quadratic triangular elements (corresponding to 1488 global variables), and applied our existence and enclosure method; the results are reported in Table 3. Fig. 5 shows a plot of the negative of $\omega$ for $\lambda = 16$; observe that $\omega$ changes sign close to the ‘outer’ part of the boundary $\partial \Omega$.

The rather low precision of the computed error bounds $\alpha$ is partly due to the low-degree-elements, but mainly due to the singularity caused by the reentrant corner at $(1/2, 1/2)$, which has not been treated by mesh refinement techniques here.

### 6. Turning and bifurcation points

One of the assumptions of our existence and enclosure method is that the linearization $L: X \rightarrow Z$ defined in (20) (resp. in (43)) is one-to-one. This condition is in particular contained in assumption (24), which in turn is controlled essentially by the eigenvalue computations mentioned in the previous two sections.

For parameter-dependent problems

$$-\Delta u + F(x, u, \nabla u, \lambda) = 0 \quad \text{on } \Omega, \quad u = 0 \quad \text{on } \partial \Omega$$

(77)

(where now, in addition to our general assumptions on problem (1), $F$ depends on a real parameter $\lambda$, and $\partial F/\partial \lambda$ is continuous), the condition of $L$ being one-to-one
fails close to turning points and to bifurcation points of solution branches \( (u_\lambda) \) of problem (77). This is illustrated, e.g., by the first example (52) in Section 4.3, where our method fails close to the bifurcation point \( (u_0, D; 5\pi^2) \).

Combinations of our abstract approach with regularization techniques, however, provide extensions which are applicable also in neighborhoods of simple turning points and even of simple bifurcation points, provided that the bifurcation is symmetry-breaking.

### 6.1. Simple turning points

Suppose that a series of numerical approximations to problem (77) for several values of \( \lambda \) (or some theoretical considerations) give rise to the conjecture that there is a solution branch \( (u_\lambda) \) containing a turning point (i.e., \( \lambda \) changes nonmonotonously along the branch).

We choose some smooth function \( \Phi : \bar{\Omega} \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R} \) such that the expression

\[ \int_{\Omega} \Phi(x, u(x), \nabla u(x), \lambda) \, dx \quad (78) \]
is well defined (which may require growth restrictions on $\Phi$) and moreover, is (expected to be) monotone along the conjectured solution branch. Such a function $\Phi$ can often be obtained from considerations based on numerical approximations alone, without further knowledge on the expected solution branch. We now add the bordering equation
\[
\int_{Q} \Phi(x, u(x), \nabla u(x), \lambda) \, dx = \mu
\]
(79)
to problem (77), with a new independent parameter $\mu$. Solutions of the augmented problem (77), (79) are now pairs $(u, \lambda)$, depending on $\mu$. Due to our assumption on the monotonicity of expression (78) along the conjectured solution branch, we may expect the augmented problem to be turning-point-free, since (79) leads along the corresponding branch by monotone variation of $\mu$.

To incorporate the augmented problem into our abstract setting (of Section 3), let $X_0, Y_0, Z_0$ denote the old spaces used in Section 4 or in Section 5, and, for fixed $\mu$,
\[
X := X_0 \times \mathbb{R},
\]
\[
Y := Y_0 \times \mathbb{R},
\]
\[
Z := Z_0 \times \mathbb{R},
\]
\[
L_0 \left[ \begin{pmatrix} u \\ \lambda \end{pmatrix} \right] := \begin{pmatrix} -\Delta u \\ 0 \end{pmatrix},
\]
\[
\mathcal{F} \left[ \begin{pmatrix} u \\ \lambda \end{pmatrix} \right] := \begin{pmatrix} \int_{Q} \Phi(x, u(x), \nabla u(x), \lambda) \, dx - \mu \end{pmatrix}.
\]
Provided that the old regularity assumptions (A) and (B) hold true (i.e., if the embedding $X_0 \hookrightarrow Y_0$ is compact, and the Poisson equation (34) has a unique solution in $X_0$, for each $r \in Z_0$, then both regularity assumptions are also satisfied in the new ‘augmented’ setting; for (B), choose $\sigma \left[ \begin{pmatrix} 0 \\ \lambda \end{pmatrix} \right] := \left[ \begin{pmatrix} 0 \\ \lambda \end{pmatrix} \right]$ now.

With $\omega = (\tilde{u}, \tilde{\lambda}) \in X$ denoting an approximate solution to problem (77), (79), the linear operator $L$ defined in (20) now reads
\[
L \left[ \begin{pmatrix} v \\ \sigma \end{pmatrix} \right] := \begin{pmatrix} -\Delta v + b \cdot \nabla v + cv + \sigma \psi \\ \int_{Q} [\varphi(x) \cdot \nabla \psi(x) + \chi(x) \psi(x)] \, dx + \tau \sigma \end{pmatrix}
\]
(80)
for $\left[ \begin{pmatrix} v \\ \sigma \end{pmatrix} \right] \in X$, where
\[
b := \frac{\partial F}{\partial \tilde{u}} (\cdot, \tilde{u}, \nabla \tilde{u}, \tilde{\lambda}), \quad c := \frac{\partial F}{\partial y} (\cdot, \tilde{u}, \nabla \tilde{u}, \tilde{\lambda}),
\]
\[
\psi := \frac{\partial F}{\partial \lambda} (\cdot, \tilde{u}, \nabla \tilde{u}, \tilde{\lambda}), \quad \varphi := \frac{\partial \Phi}{\partial \tilde{u}} (\cdot, \tilde{u}, \nabla \tilde{u}, \tilde{\lambda}),
\]
The computation of the terms satisfying (22)–(26) is similar to the corresponding computations in Section 4 resp. 5. The defect bound $\delta$ requires now, in addition to $\| - \Delta \bar{u} + F(\bar{u}, \bar{v}, \bar{\lambda}) \|_{L_0}$, a bound for $| \int_{\Omega} \Phi(x, \bar{u}(x), \nabla \bar{u}(x), \bar{\lambda}) \ dx - \mu |$, which can again be obtained by verified integration. The bound $C$ for $\| E_X \|_{L_0}$ is derived easily from the bound for $E_X^{0,0}$. The computation of a constant $K$ satisfying (24) is again based on eigenvalue bounds (for an eigenvalue problem now involving the operator $L$ in (80)!), and on the explicit embeddings and the a priori bounds presented in Section 4 resp. 5. The success in computing a 'moderate' constant $K$ (also for $(\bar{u}, \bar{\lambda})$ close to a turning point of the original problem (77)) finally confirms the 'expectations' which were the basis of the construction of the bordering equation (79). A function $G$ satisfying (25), (26) is computed here based on bounds (like in (G1) in Sections 4 and 5) for differences of values (and derivatives with respect to $y, z,$ and $\lambda$) for $F$ and $\Phi$. The details are a bit technical and will be omitted here; see also [19].

A further extension of the method which provides enclosures not only for single solutions $(u, \lambda)$, but also for branches of solutions (see [20]), can serve for proving the presence of the conjectured turning point, and for enclosing it.

As an example, we study the well-known Gelfand equation

$$-\Delta u = \lambda e^u \quad \text{on } \Omega := (0, 1)^2, \quad u = 0 \quad \text{on } \partial \Omega, \quad \lambda = 6.8$$

which has its roots, e.g., in combustion theory. Using a Newton iteration in combination with a finite element method with bi-quintic rectangular elements (which we applied in fact after transforming problem (81) to weaken the corner singularities), we computed the (known) bifurcation diagram shown in Fig. 6. It indicates the presence of a turning point at $\lambda^* \approx 6.808$, and moreover, that the expression $\int_{\Omega} \exp(u(x)) \ dx$ may be expected to be monotone along the solution branch.

The application of our 'direct' method described in Section 4 failed—as expected—close to $\lambda = 6.8$ due to a very large constant $K$. However, adding the bordering equation (see (79))

$$\int_{\Omega} \exp(u(x)) \ dx = \mu$$

(82)

to problem (81) and applying the method described in this section to the augmented problem (81), (82), we obtained existence and enclosure results with 'moderate' constants $K$ along the whole part of the branch we investigated (going up to the turning point and down to $\lambda \approx 4$ again), without any significant behavior close to the turning point; see [19] for more numerical details.

6.2. Simple bifurcation points

Computing enclosures for solutions of problem (77) close to bifurcation points (i.e., isolated intersection points of two (or more) solution branches), or even proving
bifurcation and enclosing bifurcation points, is a more difficult task than the corresponding one for turning points: Obviously, a bifurcation cannot be removed by a simple change of parameters. An even more serious difficulty consists in the fact that arbitrarily small perturbations (of the right-hand side 0 of problem (77), say) are ‘very likely’ to dissolve two crossing branches into two nontouching, veering branches, as a consequence of the Sard–Smale theorem [28, Theorem 4.18]; more precisely, the set of perturbations for which this dissolution takes place is open and dense in the image space (e.g., $L^2_0$). Already the simplest conceivable bifurcation problem $\lambda x = 0$ ($\lambda \in \mathbb{R}$, $x \in \mathbb{R}$) shows this phenomenon. This unstable character of a bifurcation causes fundamental difficulties for any method designed to prove bifurcation by numerical enclosures, because such methods—if they are successful—automatically compute enclosures also for solutions of perturbed problems ‘neighboring’ the given problem.

To avoid this difficulty, we restrict ourselves to a more specific (but nevertheless relevant) type of bifurcation, namely, the symmetry-breaking bifurcation. Here, the solutions $U$ on a ‘basic’ branch belong to a certain symmetry class generated by a symmetry of problem (77); this symmetry is ‘broken’ on another branch bifurcating from the symmetric one.

Here, a symmetry of problem (77) is a bounded linear operator $S : Z_0 \to Z_0$ such that $S(X_0) \subset X_0$ and

$$M(Su, \lambda) = SM(u, \lambda) \quad \text{for all } u \in X_0, \ \lambda \in \mathbb{R},$$

with $M(u, \lambda)$ denoting the left-hand side of the differential equation in (77), and with $X_0, Y_0, Z_0$ again denoting the ‘old’ spaces used in Section 4 resp. 5. The closed linear subspace $\Sigma(S) := \{ u \in Z_0 : u = Su \}$ of $Z_0$ is the space of symmetric functions.

Now suppose again that numerical computations (or other considerations) indicate the presence of a symmetry-breaking bifurcation. Our first goal is to compute
enclosures on the (conjectured) symmetric solution branch (consisting of solutions $U_\lambda \in X_0 \cap \Sigma(S)$, for some fixed symmetry $S$ of problem (77)). For this purpose, we choose the following realization of our abstract setting (for fixed $\lambda$):

$$
X := X_0 \cap \Sigma(S), \quad Y := \text{closure}_{X_0}(X_0 \cap \Sigma(S)),
$$

$$
Z := Z_0 \cap \Sigma(S), \quad L_0[u] := -\Delta u + b_0 \cdot \nabla u + c_0 u,
$$

$$
F(u) := F(\cdot, u, \nabla u, \lambda) - b_0 \cdot \nabla u - c_0 u,
$$

where $b_0 := \partial F/\partial z(\cdot, 0, 0, \lambda)$, $c_0 := \partial F/\partial y(\cdot, 0, 0, \lambda)$. Since $L_0$ is the Fréchet derivative of $M(\cdot, \lambda)$ at 0, condition (83) ensures that $L_0$ commutes with $S$, so that $L_0$ maps in fact $X$ into $Z$, and moreover, that $F$ maps $Y$ continuously into $Z$ (observe that $F|_{X_0} = M(\cdot, \lambda) - L_0$ maps $X_0$ continuously into $Z_0$).

The regularity assumptions (A) and (B) are satisfied if the old assumptions (A) and (B) hold true (i.e., if the embedding $X_0 \hookrightarrow Y_0$ is compact, and the Poisson equation (34) has a unique solution $u \in X_0$, for each $r \in Z_0$). For assumption (B), we choose here $T u \in D$, where $2 \in \mathbb{R}$ is sufficiently large.

Clearly, the setting (84) admits only symmetric solutions to problem (77), so that the conjectured symmetry-breaking branch is no longer present! Even stronger: if $\lambda$ is close to the conjectured bifurcation point and $\omega \in X$ is a (symmetric!) approximate solution of problem (77), then the linearization $L$ in the original setting (31), (32) resp. (31), (54) is not invertible or its inverse has a very large norm, because the linearization in the exact bifurcation point $(u^*, \lambda^*)$ (if this really exists) has a zero eigenvalue. But in a (simple) symmetry-breaking bifurcation, the corresponding eigenfunction is usually not symmetric, so that in the symmetric setting (84), this eigenfunction is no longer present, i.e., the (almost) zero eigenvalue has disappeared!

Thus, for the linearization $L$ in the new setting, we may expect a ‘moderate’ constant $K$ satisfying (24). The computation of $K$, and as well of the other terms satisfying (22), (23), (25), and (26), can be carried out by the methods described in Sections 4 and 5.

In this way, we can compute enclosures for symmetric solutions $U_\lambda$ for several values of $\lambda$ in an interval $I$ which contains the (unknown) parameter value $\lambda^*$ of the conjectured bifurcation point. By the extension of our method described in [20], we can then obtain enclosures for a whole branch $(U_\lambda)_{\lambda \in I}$ of symmetric solutions, simultaneously proving its existence.

The next aim is the computation of enclosures on the conjectured symmetry-breaking branch. Similar to the treatment of turning points, we add a bordering equation

$$
\int_Q \Phi(x, u(x), \lambda) \, dx - \int_Q \Phi(x, U_\lambda(x), \nabla U_\lambda(x), \lambda) \, dx = \mu
$$

(85)

to problem (77), where $\Phi$ is chosen such that the left-hand side of (85) is (as a functional of $(u, \lambda)$) expected to be monotone along the conjectured symmetry-breaking solution branch. Of course, in contrast to turning-point problems, the augmented
problem (77), (85) is not yet regularized for a symmetry-breaking branch since the symmetric branch $U/_{2}$ is still present at $D_0$. For regularization we transform problem (77), (85) as follows: for any $(u, \lambda, \mu)$ satisfying (77), (85), $\mu \neq 0$, let

$$w := \frac{1}{\mu}(u - U/_{\lambda}).$$

(86)

It is then easy to verify that $(w, \lambda, \mu)$ satisfies

$$-\Delta w + \tilde{F}(\cdot, w, \nabla w, \lambda, \mu) = 0 \quad \text{on } \Omega, \quad w = 0 \quad \text{on } \partial \Omega,$$

(87)

where

$$\tilde{F}(x, y, z, \lambda, \mu) := \frac{1}{\mu}[F(x, U/_{\lambda}(x) + \mu y, \nabla U/_{\lambda}(x) + \mu z, \lambda) - F(x, U/_{\lambda}(x), \nabla U/_{\lambda}(x), \lambda)].$$

(88)

Extending $\tilde{F}$ and $\tilde{\Phi}$ continuously to $\mu = 0$ (which involves derivatives of $F$ and $\Phi$), we now look for a solution branch $(w/_{\mu}, \lambda/_{\mu})_{\mu \in J}$ of problem (87) such that $0 \in J$. In case of success, inversion of the transformation (86) provides a solution branch $(u/_{\mu}, \lambda/_{\mu})_{\mu \in J}$ of problem (77), (85), which can be shown to be symmetry-breaking if the eigenfunction ($!$) $w_0$ is not symmetric.

The big advantage of problem (87) (compared with (77), (85)) is that $(w, \lambda, \mu)$ does not solve (87) since, due to (88), $\Phi(x, 0, 0, \lambda, \mu)$ equals 0. Because $w \equiv 0$ corresponds, according to (86), to $u \equiv U/_{\lambda}$, we may therefore regard the symmetric solution branch $(U/_{\lambda}, \lambda/_{\lambda})_{\lambda \in J}$ as being transformed away in problem (87)! Thus, there is a good chance for a regular solution branch $(u/_{\mu}, \lambda/_{\mu})_{\mu \in J}$ of problem (87).

To incorporate problem (87) into our abstract setting, we proceed now similarly as for the treatment of turning-points: with $X_0$, $Y_0$, $Z_0$ denoting the 'old' spaces used in Section 4 resp. 5 we define, for fixed $\mu$,

$$X := X_0 \times \mathbb{R},$$

$$Y := Y_0 \times \mathbb{R},$$

$$Z := Z_0 \times \mathbb{R},$$

$$L_0 \left[ \begin{array}{c} w \\ \lambda \end{array} \right] := \left( \begin{array}{c} -\Delta w \\ 0 \end{array} \right),$$

$$\mathcal{F} \left[ \begin{array}{c} w \\ \lambda \end{array} \right] := \left( \begin{array}{c} \tilde{F}(\cdot, w, \nabla w, \lambda, \mu) \\ \int_{\Omega} \tilde{\Phi}(x, w(x), \nabla w(x), \lambda, \mu) \ dx - 1 \end{array} \right)$$
(where we have to extend \( U \) continuously (e.g., constantly) from \( I \) to \( \mathbb{R} \), in order to make \( F \) really defined on \( Y \)). The further proceeding concerning the computation of the terms satisfying (22)–(26) is again similar to turning-point problems. An additional difficulty consists here in the fact that the branch \( (U_{\lambda})_{\lambda \in I} \) which enters \( Q \) and \( U \) via (88), is known only up to error bounds; nevertheless, this knowledge turns finally out to be sufficient. For more details, see [21].

We applied the methods of this section to the ODE example
\[
-u'' - 65u + u^3 = \lambda x(1-x) \quad \text{on} \ (0,1), \quad u(0) = u(1) = 0. \tag{89}
\]
Our numerical computations (using Newton’s iteration in combination with a collocation procedure with polynomial basis functions) indicate the bifurcation diagram shown in Fig. 7, with two turning points at \( \lambda \approx \pm 731.72 \), and two bifurcation points at \( \lambda \approx \pm 685.30 \). Moreover, the computations suggest that the solutions on the ‘basic’ 3-shaped branch are symmetric with respect to reflection at \( x = 1/2 \) (i.e., with respect to \( (5u)(x) := u(1-x) \)), and that this symmetry is broken on the ‘circular’ branch.

Using the methods of this section (with \( \Phi(x, y, z, \lambda) := x(1-x)y \) in (79) to treat the turning points, and with \( \Phi(x, y, z, \lambda) := \hat{w}(x)y \) in (85), with \( \hat{w} \) denoting a (reflection-antisymmetric) approximation of the eigenfunction corresponding to the zero eigenvalue of the linearization in the conjectured bifurcation point), and applying the ‘direct’ approach presented in Section 4 away from the singular points, we obtained (supported by the extension in [20]) a verification of the whole bifurcation diagram shown in Fig. 7, including a proof of bifurcation.

7. Further applications of the abstract setting

In this section, we will very briefly report on some further applications of our abstract operator theoretical setting presented in Section 3.

(a) Let \( X_0, Y_0, Z_0 \) denote complex Banach spaces such that \( X_0 \subset Y_0 \) and let \( A \in \mathcal{B}(X_0, Z_0), B \in \mathcal{B}(Y_0, Z_0) \). Consider the eigenvalue problem
\[
A[u] = \lambda B[u]. \tag{90}
\]
We wish to enclose eigenpairs \( (u, \lambda) \in (X_0 \setminus \{0\}) \times \mathbb{C} \) of problem (90). Observe that (90) may be a non-self-adjoint eigenvalue problem, so that the variational methods mentioned in Section 4.1.1 are not applicable. Similar to the bordering equation (79), we add here a normalizing equation
\[
\varphi(u) = 1 \tag{91}
\]
to (90). Here, \( \varphi \) is a bounded linear functional on \( Y_0 \) which is chosen such that we may expect \( \varphi(u) \neq 0 \) (and thus, w.l.o.g., \( \varphi(u) = 1 \)) for the eigenfunction \( u \) to be enclosed. This choice of \( \varphi \) can be based, e.g., on a numerical approximation to \( u \).

To incorporate problem (90), (91) into our abstract setting, we choose
\[
X := X_0 \times \mathbb{C}, \quad Y := Y_0 \times \mathbb{C}, \quad Z := Z_0 \times \mathbb{C},
\]
It is easy to check that \( \mathcal{F} \) is Fréchet-differentiable on \( Y \); observe that \( \mathcal{F} \) is really nonlinear! The regularity assumption (A) requires here that the embedding \( X_0 \hookrightarrow Y_0 \) is compact; \( (92) \)

while assumption (B) is satisfied if

the resolvent set of problem (90) is nonempty

(i.e., \( A - \lambda^* B : X_0 \to Z_0 \) is one-to-one and onto, for some \( \lambda^* \in \mathbb{C} \)), since then one can choose

\[
\sigma \left[ \begin{pmatrix} u \\ \lambda \end{pmatrix} \right] := \begin{pmatrix} -\lambda B[u] \\ \varphi(u) - 1 \end{pmatrix}
\]

in assumption (B).

The computation of \( \delta, C, \) and \( G \) satisfying (22), (23), (25), and (26) (for some approximate solution \( \omega = (\tilde{u}, \tilde{\lambda}) \in X \) to problem (90), (91)) depends on the spaces \( X_0, Y_0, Z_0 \) and will not be discussed here. For Hilbert-spaces \( X_0, Y_0, Z_0 \), condition (24) (resp. (30), see Remark 3.2(b)) leads to a self-adjoint eigenvalue problem which can be treated by the variational methods mentioned in Section 4.1.1.
In [12], the approach presented here has successfully been applied to the (non-self-adjoint) Orr–Sommerfeld eigenvalue problem, which is a fourth-order ODE boundary value problem and one of the governing equations of hydrodynamic stability. In fact, an unbounded interval is underlying in [12], so that (92) is unfortunately violated, which requires to apply Banach’s in place of Schauder’s fixed-point theorem, and moreover, to prove that $L$ (defined in (20)) is one-to-one and onto by other arguments than those providing (16).

(b) Consider the boundary value problem with a system of ordinary differential equations

$$\begin{align*}
u' + F(x, u) &= 0 \quad \text{on} \ (a, b), \\
g(u(a), u(b)) &= 0, \\
\end{align*}$$

(93)

where $F : [a, b] \times \mathbb{R}^n \to \mathbb{R}^n$ and $\partial F / \partial y_i$ ($i = 1, \ldots, n$) are continuous, and $g : \mathbb{R}^{2n} \to \mathbb{R}^n$ is continuously differentiable. Putting

$$X := H_1(a, b)^n, \quad Y := C(a, b)^n, \quad Z := L_2(a, b)^n \times \mathbb{R}^n,$$

we find that all smoothness and regularity assumptions required in Section 3 are satisfied (for assumption (B), choose $\sigma[u] := (0, \ldots, 0)$). For a given approximate solution $\phi$ to problem (93), a constant $\delta$ satisfying (22) can be computed by verified integration.

For calculating an embedding constant $C$ (satisfying (23)), see [15]. Condition (24) leads again to a self-adjoint eigenvalue problem. A function $G$ satisfying (25), (26) can usually be calculated rather directly here.

This approach has been investigated in detail and applied to several examples in [4].

(c) Let $\mathcal{F} : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable, and consider the system

$$\mathcal{F}(x) = 0$$

(94)

of nonlinear equations. For the choice,

$$X = Y = Z := \mathbb{R}^n, \quad L_0 = 0,$$

all smoothness and regularity assumptions of Section 3 are satisfied (choose $\sigma := \text{id}_\mathbb{R}^n$ for assumption (B)). For some approximate solution $\omega \in \mathbb{R}^n$ to problem (94), a defect bound $\delta$ (satisfying (22)) can be computed by interval evaluation. Condition (23) obviously holds for $C = 1$. If the chosen norm in $\mathbb{R}^n$ is the Euclidean one, condition (24) requires to compute a positive lower bound for the smallest eigenvalue of $\mathcal{F}'(\omega)^T \mathcal{F}'(\omega)$ resp. the smallest singular value of $\mathcal{F}'(\omega)$. For computing a function $G$ satisfying (25), (26), one may use, e.g., automatic differentiation techniques (to provide a local Lipschitz constant for $\mathcal{F}'$).

This approach has not been tested numerically, and we do not really want to advocate it, since for problems in $\mathbb{R}^n$—in contrast to differential equation prob-
lems—more direct methods (e.g., for computing enclosures for inverse matrices) are available, at least for problems of ‘moderate’ size. Nevertheless, it is remarkable that for ‘large’ systems, also Rump proposes to use the smallest singular value in [23].

8. Nakao’s method

In this final section, we will give a brief description of another method for obtaining existence and enclosure results for problem (1), which has been proposed by Nakao, partly in co-authorship with N. Yamamoto and other co-workers (see, e.g., [13,14]). As the approach presented in Sections 3–6, it avoids the drawbacks of monotonicity methods concerning the range of applicability. Furthermore, it also avoids the computation of a bound for $L^{-1}$ which is needed in our approach. On the other hand, it requires the verified solution of large nonlinear and linear systems in $\mathbb{R}^N$, where the latter moreover have an interval right-hand side (see (95) and (108)). If this task is harder or less hard than the computation of the eigenvalue bounds which we use to bound $L^{-1}$, possibly depends on the concrete situation. A given problem may require a very large dimension $N$, while the eigenvalue computations do not cause any difficulties: Consider, e.g., the case where the coefficient $b$ and $c$ in (43) satisfy $b \equiv 0$ and $c \geq 0$, respectively, so that (48) holds for $\sigma := \lambda_{\min}(-\Delta)$ (which can easily be bounded), and (70) holds for $K = 1$ (without any further computation).

But also the opposite situation may occur: e.g., for the Orr–Sommerfeld problem (see Section 7), the dimension $N$ needed for the approximation is rather moderate, while the eigenvalue computations are very tedious due to a complicated structure of the homotopy used there.

Nakao’s method has been proposed in several variants. Here, we describe one of them which we believe is the most relevant one: The domain $\Omega$ is assumed to be a bounded convex polygonal domain in $\mathbb{R}^2$, and the nonlinear operator $f$ defined by $f(u) := F(\cdot, u, \nabla u)$ is required to map $H^1_0(\Omega)$ continuously into $L^2(\Omega)$, moreover mapping bounded sets into bounded sets. With $S_h$ denoting a suitable finite element subspace of $H^1_0(\Omega)$, the first step in Nakao’s method is to compute a verified solution $uh \in S_h$ of the finite element discretization of problem (1) (which amounts to a nonlinear system in $\mathbb{R}^N$):

\[
\int_{\Omega} \nabla uh \cdot \nabla \phi \, dx + \int_{\Omega} f(uh)\phi \, dx = 0 \quad \text{for all } \phi \in S_h.
\]  

(95)

With $\overline{u} \in H^2(\Omega) \cap H^1_0(\Omega)$ denoting the (unknown) solution of

\[-\Delta \overline{u} + f(uh) = 0 \quad \text{on } \Omega, \quad \overline{u} = 0 \quad \text{on } \partial \Omega,
\]  

(96)

the following problem is equivalent to problem (1), via the transformation $w = u - \overline{u}$:

\[-\Delta w + f(\overline{u} + w) - f(uh) = 0 \quad \text{on } \Omega, \quad w = 0 \quad \text{on } \partial \Omega.
\]  

(97)
Using the (compact) solution operator \((-\Delta)^{-1} : L^2(\Omega) \to H_0^1(\Omega)\), one finds that (97) is equivalent to the fixed-point equation
\[
w = -(-\Delta)^{-1}[f(\pi + w) - f(u_h)] =: \tilde{T}w.
\] (98)

\(\tilde{T} : H_0^1(\Omega) \to H_0^1(\Omega)\) is continuous and compact, so that in principle Schauder’s fixed-point theorem can be applied to problem (98), if some closed, bounded, convex subset of \(H_0^1(\Omega)\) can be found which is mapped into itself by \(\tilde{T}\). However, a detailed analysis shows that such a subset exists (except in trivial cases) only for a severely restricted class of problems (1); the restriction is even stronger than the one which is inherently required by monotonicity methods (see Section 2).

To avoid these difficulties, Nakao proposes a ‘Newton-like’ method to be applied to problem (98): with \(P : H_0^1(\Omega) \to S_h\) denoting the \(H_0^1\)-orthogonal projection onto \(S_h\), and \(f'(u_h)\) the Fréchet-derivative of \(f\) at \(u_h\) (the existence of which is required here), he makes the assumption that the operator
\[
\tilde{L} := P\left[\text{id}_{H_0^1(\Omega)} + (-\Delta)^{-1}f'(u_h)\right]|_{S_h} : S_h \to S_h
\] (99)

which is in fact some finite-dimensional projection of \((-\Delta)^{-1}L\) (with \(L\) denoting ‘our’ operator defined in (20)), is one-to-one and (therefore) onto. This assumption has to be checked by proving that the matrix with elements \(G_{ij} = \langle \tilde{L}\phi_i, \phi_j \rangle_{H_0^1}\), i.e.,
\[
G_{ij} = \int_\Omega \left[ \nabla \phi_i \cdot \nabla \phi_j + \frac{\partial F}{\partial x} (\cdot, u_h, \nabla u_h) \cdot (\nabla \phi_i) \phi_j \\
+ \frac{\partial F}{\partial y} (\cdot, u_h, \nabla u_h) \phi_i \phi_j \right] \, dx
\] (100)

(where \((\phi_1, \ldots, \phi_N)\) denotes a basis of \(S_h\), is regular, which requires the verified solution of some linear system with matrix \((G_{ij})\) (and arbitrary right-hand side). This task is solved later automatically when solving system (108).

Using the regularity of \(\tilde{L}\) one finds that the following fixed-point equation is equivalent to (98):
\[
w = Pw - \tilde{L}^{-1}P(w - \tilde{T}w) + (\text{id}_{H_0^1(\Omega)} - P)\tilde{T}w =: Tw.
\] (101)

Here the first part \(Nw:= Pw - \tilde{L}^{-1}P(w - \tilde{T}w)\) is a ‘Newton-like’ operator for problem (98), and the second part is (hopefully) small if \(\dim S_h\) is sufficiently large, so that the chances for finding a closed, bounded, convex set \(W \subset H_0^1(\Omega)\), which is mapped into itself by \(T\), are much better than for (98). Moreover, \(T\) is continuous and compact, so that indeed Schauder’s fixed-point theorem can be applied.

Nakao looks for \(W\) (satisfying \(TW \subset W\)) in the form \(W = W_h \oplus W_\perp\), where \(W_h \subset S_h\) and \(W_\perp \subset S_h^\perp\) (with \(S_h^\perp\) denoting the \(H_0^1\)-orthogonal complement of \(S_h\)), so that the condition \(T(W) \subset W\) splits into the two conditions
\[
\text{(a)} \quad N(W) \subset W_h, \\
\text{(b)} \quad (\text{id}_{H_0^1(\Omega)} - P)\tilde{T}W \subset W_\perp.
\] (102)
For verifying them, suppose that
\[
W_h = \left\{ \sum_{i=1}^{N} \alpha_i \phi_i : \alpha_i \in [\underline{\alpha}_i, \overline{\alpha}_i] (i = 1, \ldots, N) \right\},
\]
(103)
\[
W_\perp = \{ \phi \in S_h : \| \phi \|_{L^2}^2 \leq \alpha \}
\]
with given intervals \([\underline{\alpha}_i, \overline{\alpha}_i]\) and given \(\alpha > 0\). Since (98) and (99) provide, for \(w \in W\),
\[
Nw = \tilde{L}^{-1} [\tilde{L} P w - P w + \tilde{P} w] = \tilde{L}^{-1} \tilde{P} (-\Delta)^{-1} [f'(u_h) P w - f(\overline{\alpha} + w) + f(u_i)],
\]
condition (102(a)) is equivalent to requiring that, for each \(w \in W\), some \((\alpha_1, \ldots, \alpha_N) \in \prod_{i=1}^{N} [\underline{\alpha}_i, \overline{\alpha}_i]\) exists such that
\[
P(-\Delta)^{-1} [f'(u_h) P w - f(\overline{\alpha} + w) + f(u_i)] = \sum_{i=1}^{N} \alpha_i \tilde{L} \phi_i.
\]
(104)
Since both sides of (104) are in \(S_h\), this equation is equivalent to requiring that the \(H^0\)-inner products of both sides with \(\phi_j\) are the same (for \(j = 1, \ldots, N\)), i.e., to
\[
r_j(w) = \sum_{i=1}^{N} \alpha_i G_{ij} \text{ for } j = 1, \ldots, N,
\]
(105)
where \(G_{ij}\) is given by (100), and
\[
r_j(w) := \int_{\Omega} [f'(u_h) P w - f(\overline{\alpha} + w) + f(u_i)] \phi_j \, dx.
\]
(106)
Therefore, condition (102(a)) is equivalent to the requirement that, for each \(w \in W\), the solution \((\alpha_1, \ldots, \alpha_N)\) of the linear system (105) is unique and contained in \(\prod_{i=1}^{N} [\underline{\alpha}_i, \overline{\alpha}_i]\). Thus, a sufficient condition for (102(a)) is
\[
[\beta_j, \overline{\beta}_j] \subset \underline{[\alpha_i, \overline{\alpha}_i]} \text{ for } i = 1, \ldots, N,
\]
(107)
with \(([\beta_j, \overline{\beta}_j])_{i, \ldots, N}\) denoting the interval solution of the (large!) system
\[
G^T \beta = R(W),
\]
(108)
where \(G = (G_{ij}), \ R(W) = (R_1(W), \ldots, R_N(W)), \) and \(R_j(W)\) is an interval enclosing \([r_j(w) : w \in W]\) \((j = 1, \ldots, N)\).

To compute \(R_j(W)\) via (106), Nakao exploits, in his examples, the special (e.g., polynomial) structure of \(f\), and formulas (103) to bound \(w \in W = W_h \oplus W_\perp\), and moreover, the following arguments to bound the function \(\overline{u}\) (solving (96)) which enters (106): (95) and (96) provide \(P(\overline{u}) = \overline{u}_h\). Consequently, the inequality (which is well known, from finite element theory, to hold for convex domains \(\Omega\))
\[
\| u - Pu \|_{H^0} \leq C_0 \| \Delta u \|_{L^2} \text{ for } u \in H^2(\Omega) \cap H^0(\Omega),
\]
(109)
where \( h \) is the element mesh size and \( C_0 \) is a constant which is known in many cases, immediately provides

\[
\| \overline{u} - u_h \|_{H^1_0} \leq C_0 h \| f(u_h) \|_{L^2}.
\]  

(110)

The Aubin–Nitsche lemma gives the additional inequality

\[
\| \overline{u} - u_h \|_{L^2} \leq C_0^2 h^2 \| f(u_h) \|_{L^2}.
\]  

(111)

It should be remarked that, also for bounding \( w \in W \) via (103), the Aubin–Nitsche lemma can be used to obtain the additional information \( \| \phi \|_{L^2} \leq C_0 h \alpha \) for \( \phi \in W_\perp \).

For checking condition (102(b)), Nakao uses inequality (109) again, which provides, together with (98),

\[
\| (\text{id}_{H^1_0} - P) \overline{w} \|_{H^1_0} \leq C_0 h \| f(\overline{w} + w) - f(u_h) \|_{L^2} \quad \text{for } w \in W,
\]

so that (102(b)) holds if

\[
C_0 h \| f(\overline{w} + w) - f(u_h) \|_{L^2} \leq \alpha \quad \text{for all } w \in W.
\]  

(112)

This condition can be checked again using (103) (plus the Aubin–Nitsche extension) to bound \( w \in W \), and (110), (111) to bound \( \overline{w} \).

Thus, the crucial conditions (102) can be tested via (107), (108), and (112) if the intervals \([\underline{\alpha}_i, \overline{\alpha}_i]\) \((i = 1, \ldots, N)\) and the number \( \alpha > 0 \) are given. For constructing them, Nakao proposes an iterative technique: for some current values of \( \underline{\alpha}_i, \overline{\alpha}_i, \alpha \) (entering the right-hand side of (108) and the left-hand side of (112)), the interval solution \( \beta \) of (108) and the left-hand side of (112) define new values of \( \underline{\alpha}_i, \overline{\alpha}_i, \alpha \), and conditions (107) and (112) simply read

\[
[\underline{\alpha}_i^{\text{new}}, \overline{\alpha}_i^{\text{new}}] \subseteq [\underline{\alpha}_i^{\text{old}}, \overline{\alpha}_i^{\text{old}}], \quad \alpha^{\text{new}} \leq \alpha^{\text{old}}.
\]  

(113)

In the iteration, the new values are subject to an \( \varepsilon \)-inflation (i.e., the new intervals \([\underline{\alpha}_i, \overline{\alpha}_i]\) and the new \( \alpha \) are slightly enlarged), before the next iteration step is performed. This inflation technique facilitates convergence of the iteration in the sense that condition (113) is satisfied after a (usually very moderate) number of iteration steps.

Acknowledgement

The author is grateful to the anonymous referee for his very helpful comments and suggestions.

References