A Nonlinear Discretization Theory

Klaus Böhmer*

Fachbereich Mathematik und Informatik
Universität Marburg
Arbeitsgruppe Numerik
Hans Meerwein Strasse, Lahnberge
D-35032 Marburg
Germany

Robert Schaback

Institut für Numerische und Angewandte Mathematik
Universität Göttingen
Lotzestraße 16-18
D–37073 Göttingen
Germany

Abstract

This paper extends for the first time Schaback's linear discretization theory to nonlinear operator equations, relying heavily on the methods in Böhmer's 2010 book. There is no restriction to elliptic problems or to symmetric numerical methods like Galerkin techniques. Trial spaces can be arbitrary, including spectral and meshless methods, but have to approximate the solution well, and testing can be weak or strong. On the downside, stability is not easy to prove for special applications, and numerical methods have to be formulated as optimization problems. Results of this discretization theory cover error bounds and convergence rates. Some numerical examples are added for illustration.

1. Overview

We start directly with a short and simple version of our discretization theory in Section 2. Since some of its ingredients are somewhat nonstandard, a detailed example follows in Section 3, explaining in particular how strong and weak problems are subsumed. Section 4 contains some results that allow to reduce part of the nonlinear theory to the linear case provided in [1]. Sections 3 and 4 strongly depend upon the nonlinear methods presented in [2]. The most critical ingredient of our theory is proving stability, and we devote Section 5 to its

*Corresponding author
Email addresses: boehmer@Mathematik.Uni-Marburg.de (Klaus Böhmer), schaback@math.uni-goettingen.de (Robert Schaback)

detailed analysis. The main tool are sampling inequalities [3]. We close the paper with some numerical experiments.

2. Nonlinear Discretization Theory

We need four essential ingredients:

1. a *well–posed* nonlinear operator equation,
2. a scale of *trial spaces* that allows to approximate the true solution well,
3. a stable *testing strategy* and
4. a numerical method based on *minimization of residuals*.

Then we shall show how these lead to error bounds and convergence results. But in all possible applications users will have to verify that the four ingredients of the theory are valid. We shall show in Section 4 how this follows by linearization and in Section 6 how to do this for certain examples.

2.1. Well–Posed Problems

We assume a nonlinear operator

$$F \ : \ \mathcal{D}(F) \subset \mathcal{U} \to \mathcal{V} \tag{1}$$

between Banach spaces $\mathcal{U}$ and $\mathcal{V}$, and we want to solve the nonlinear operator equation

$$Fu = f \tag{2}$$

for $u \in \mathcal{D}(F)$ when some $f \in \mathcal{R}(F)$ is given. This combines differential equations and boundary conditions into one single operator, but linear homogeneous boundary conditions should be incorporated into $\mathcal{U}$. Readers should keep in mind that often $\mathcal{U}$ and $\mathcal{V}$ will then be Cartesian products of other Banach spaces. Furthermore, note that physical problems will lead to different operator equations if posed in weak or strong form. This will be illustrated in Section 3.

The nonlinear problem (1), (2) should be *well–posed* in the following sense:

1. There is a locally unique exact solution $u^* \in \mathcal{D}(F)$ with $Fu^* = f$.
2. In a ball around u^* with radius $R > 0$ measured w.r.t. $\| \cdot \|_{\mathcal{U}}$, s.t.

$$K_R(u^*) := \{u \in \mathcal{U} \ : \ \|u - u^*\|_{\mathcal{U}} \leq R\} \subset \mathcal{D}(F),$$

the operator F satisfies the inequalities

$$c_F^{-1}\|u - v\|_{\mathcal{U}} \leq \|Fu - Fv\|_{\mathcal{V}} \leq C_F\|u - v\|_{\mathcal{U}} \text{ for all } u, v \in K_R(u^*) \tag{3}$$

with certain positive constants c_F, C_F, cf. subsection 4.1. For linear F replace $Fu - Fv$ by $F(u - v)$.

If F is Fréchet differentiable in $K_R(u^*)$, the left inequality in (3) implies a boundedly invertible $F'(u)$ for all $u \in K_R(u^*)$, see Theorem 3. In [2] it is shown that a locally unique exact solution $u^* \in \mathcal{D}(F)$ for $Fu^* = f$ usually has the property that $F'(u^*)$ is boundedly invertible.

2.2. Trial Spaces

We assume that there is a scale of linear spaces $\mathcal{U}_r \subseteq \mathcal{U}$ of *trial* functions which allow good approximations $u_r^* \in \mathcal{U}_r$ to u^* in the sense

$$\|u^* - u_r^*\|_{\mathcal{U}} \leq \epsilon(r, u^*) << R. \tag{4}$$

Here, we use a real–valued positive discretization parameter r (for t**R**ial) replacing the standard parameter h. Applications will usually provide a scale of spaces with the properties $\mathcal{U}_{r'} \subset \mathcal{U}_r$ for $r < r'$ and

$$\lim_{r \to 0} \epsilon(r, u^*) = 0. \tag{5}$$

The attainable error of a numerical method for solving an operator equation will basically depend on the attainable error $\epsilon(r, u^*)$ determined by the choice of the *trial space*. If the trial space is chosen badly, there is no hope for a good approximation of the solution. The essential difference between the convergence rates of spectral techniques, meshless methods, the h-FEM or the p-FEM mainly lies in the difference of the chosen trial spaces, not in the difference of their testing strategies that we describe below in general.

To get high convergence rates, for instance with spectral methods, the solution and the trial space must usually have plenty of regularity. We model this by requiring that the exact solution u^* and the trial space $\mathcal{U}_r$ are in a *regularity subspace* $\mathcal{U}_R \subset \mathcal{U}$. The basic assumption on the trial space then is

$$\inf_{u_r \in \mathcal{U}_r} \|u - u_r\|_{\mathcal{U}} \leq \epsilon(r, u)$$

for all $u \in \mathcal{U}_R$ without any connection to any PDE problem. Then (4) follows from the regularity assumption $u^* \in \mathcal{U}_R$.

2.3. Testing

We assume that there is a scale of linear *test mappings*

$$T_s \ : \ \mathcal{D}(T_s) \subseteq \mathcal{V} \to \mathcal{V}_s$$

with values in finite–dimensional spaces $\mathcal{V}_s$. Again, users should think of test spaces $\mathcal{V}_s$ whose dimensions increase when s decreases to 0. We view s as a discretization parameter for te**S**ting, like the standard h. Note that we use s–dependent norms $\|.\|_{\mathcal{V}_s}$ on the finite–dimensional spaces $\mathcal{V}_s$.

In many applications, $\mathcal{V}$ is a space like $L_2(\Omega)$ while T_s performs evaluation of functions on discrete sets X_s. This requires some care with the domain of definition $\mathcal{D}(T_s) \subseteq \mathcal{V}$ of T_s and with boundedness of T_s. We shall avoid to use $\|T_s\|$ as an operator norm in $\mathcal{L}(\mathcal{V}, \mathcal{V}_s)$ as well as using $\|T_s F'(u^*)\|$ as an operator norm in $\mathcal{L}(\mathcal{U}, \mathcal{V}_s)$ or assuming that $T_s F$ is continuous on all of $\mathcal{U}$. Instead, we assume $T_s F$ to be continuous and Fréchet differentiable on the regularity subspace $\mathcal{U}_R$ that contains u^* and the trial spaces $\mathcal{U}_r$.

The basic idea of testing is that the true solution u^* of (1), (2) should theoretically satisfy the discretized *test equations*

$$T_s(Fu) = T_s(f). \tag{6}$$

Note that the test maps T_s are linear, but the test problems (6) are nonlinear due to the nonlinearity of the operator F. The true solution u^* will satisfy *all* test problems, no matter which test maps are used, and testing is at this point completely independent from the trial spaces. It writes down plenty of necessary conditions for the exact solution to satisfy, whatever these conditions are. Clearly, writing down *more* necessary conditions will improve stability, no matter how the latter is actually defined and no matter which numerical method is used to satisfy these conditions approximatively.

Readers should keep in mind that testing in the above setting does not necessarily involve functions, and in particular not "test functions", and not necessarily require numerical integration. While *strong* testing uses evaluations of functions and derivatives and is traditionally called *collocation*, *weak* testing uses test maps working with integrals against test functions. Our framework allows both techniques, and mixtures thereof, and permits all conceivable choices of trial spaces.

2.4. Numerical Solution

We cannot assume that the test equations (6) are solvable on the trial space $\mathcal{U}_r$, since this is not even guaranteed in linear cases [4] where trial spaces and testing are independent. Instead, we assume a numerical process that constructs a function $\hat{u}_r \in \mathcal{U}_r \cap K_R(u^*)$ with

$$\|T_s(f - F\hat{u}_r)\|_{\mathcal{V}_s} \leq 2\|T_s(f - Fu_r^*)\|_{\mathcal{V}_s}, \tag{7}$$

where we used the approximation $u_r^* \in \mathcal{U}_r$ of (4) to the exact solution u^*. This can, for instance, be done by *residual minimization*

$$\hat{u}_r = \arg\min \left\{ \|T_s(f - Fu_r)\|_{\mathcal{V}_s} \ : \ u_r \in \mathcal{U}_r \cap K_R(u^*) \right\}. \tag{8}$$

The nonlinearity of the optimization reflects the nonlinearity of the problem as a whole, and this is a potential challenge. There may be non–convex problems, and there may be plenty of local minima. For instance, the Chladny sound figures problem in section 6.2 has infinitely many analytic solutions anyway, and these will, of course, also occur in the optimization, no matter how it is implemented. But in certain cases, the final problems will be convex, e.g. if a convex term like u^2 is added to a linear elliptic operator.

In many cases, the residual $f - F\hat{u}_r$ is explicitly available, at least in a fine discretization of graphical accuracy. If users see that it is reasonably small, they often stop asking for a more elaborate error bound, since they know that the given operator equation is solved in a finely discretized form up to a small perturbation in the right–hand side. But even if $T_s(f - F\hat{u}_r)$ vanishes on a perfectly fine discretization, it is mathematically impossible to conclude in general that $f - F\hat{u}_r$ is small as a function.

2.5. Stability

The above argument requires to conclude that a function is small if certain discrete information obtained by a linear test operator is small. This can only work in finite–dimensional subspaces, and even there it needs a thorough analysis.

To make this work, we require a *stability inequality*

$$\|Fu_r - Fv_r\|_{\mathcal{V}} \leq C_S(r,s)\|T_s(Fu_r - Fv_r)\|_{\mathcal{V}_s} \tag{9}$$

for all u_r, $v_r \in \mathcal{U}_r \cap K_R(u^*), r, s \to 0$. Note that the stability inequality relates trial spaces $\mathcal{U}_r$ to test maps T_s and that a stability inequality allows to identify the image Fu_r of a trial function $u_r \in \mathcal{U}_r$ uniquely from its test data T_sFu_r. In some sense, stability means bounded invertibility of the test map T_s on the range of F on the trial functions. Numerical calculations cannot work reasonably without such an assumption. Both (6) and (9) suggest that applications gain more stability by adding more test equations, without sacrificing the solvability of (6) by the exact solution.

In practice, there will be a dependence between the discretization parameters r and s for the trial space and the test strategy, respectively, because a larger trial space will require more test conditions to identify the solution properly. Since the scale of trial spaces should be chosen first, users will choose a test discretization $s = s(r)$ depending on the choice of the trial space with parameter r.

Definition 1. *A choice $s(r)$ of a test discretization for a given trial discretization will be called a* trial/test strategy *in what follows.*

Definition 2. *We call a trial/test strategy $s(r)$* uniformly stable*, if*

$$C_S(r, s(r)) \leq C_T \tag{10}$$

uniformly for all sufficiently small trial parameters r.

By (3) and (9) we have for $F_{r,s} = T_sF\big|_{\mathcal{U}_r}$

$$\begin{aligned} \|u_r - v_r\|_{\mathcal{U}} &\leq c_F\|Fu_r - Fv_r\|_{\mathcal{V}} \\ &\leq c_F C_S(r,s)\|T_s(Fu_r - Fv_r)\|_{\mathcal{V}_s} \\ &= c_F C_S(r,s)\|F_{r,s}u_r - F_{r,s}v_r\|_{\mathcal{V}_r} \end{aligned}$$

for all u_r, $v_r \in \mathcal{U}_r \cap K_R(u^*), r, s \to 0$. Thus well–posedness and our stability inequality (9) imply the classical stability property for $F_{r,s}$ provided that the trial/test strategy $s(r)$ yields $C_S(r, s(r)) \leq C_T$ and thus is uniformly stable in the sense of Definition 2.

2.6. Error Bound

The previous sections on well–posedness, trial spaces, testing, stability, and optimizing solvers allow a surprisingly simple error analysis as follows.

Theorem 1. *Under the assumptions (3),(4), (7), and (9), the approximate solution $\hat{u}_r \in \mathcal{U}_r$ of the nonlinear operator equation (2) has an error bound*

$$\|u^* - \hat{u}_r\|_{\mathcal{U}} \leq \epsilon(r, u^*) + 3c_F C_S(r, s)\delta(r, s, u^*) \tag{11}$$

with

$$\delta(r, s, u^*) := \|T_s(Fu_r^* - Fu^*)\|_{\mathcal{V}_s}. \tag{12}$$

PROOF. We extend the basic argument in [1] to the nonlinear situation and apply all ingredients to get, with (4), (3), (9), (7),

$$\begin{array}{rcl}
\|u^* - \hat{u}_r\|_{\mathcal{U}} & \leq & \|u^* - u_r^*\|_{\mathcal{U}} + \|u_r^* - \hat{u}_r\|_{\mathcal{U}} \\
& \leq & \epsilon(r, u^*) + \|u_r^* - \hat{u}_r\|_{\mathcal{U}} \\
& \leq & \epsilon(r, u^*) + c_F\|Fu_r^* - F\hat{u}_r\|_{\mathcal{V}} \\
& \leq & \epsilon(r, u^*) + c_F C_S(r, s)\|T_s(Fu_r^* - F\hat{u}_r)\|_{\mathcal{V}_s} \\
& \leq & \epsilon(r, u^*) \\
& & + c_F C_S(r, s)\left(\|T_s(Fu_r^* - Fu^*)\|_{\mathcal{V}_s} + \|T_s(f - F\hat{u}_r)\|_{\mathcal{V}_s}\right) \\
& \leq & \epsilon(r, u^*) \\
& & + c_F C_S(r, s)\left(\|T_s(Fu_r^* - Fu^*)\|_{\mathcal{V}_s} + 2\|T_s(f - Fu_r^*)\|_{\mathcal{V}_s}\right) \\
& \leq & \epsilon(r, u^*) + 3c_F C_S(r, s)\|T_s(Fu_r^* - Fu^*)\|_{\mathcal{V}_s} \\
& \leq & \epsilon(r, u^*) + 3c_F C_S(r, s)\delta(r, s, u^*). \quad \square
\end{array}$$

In later applications, we shall use a trial/test strategy $s(r)$ that couples the test discretization parameter s with the trial discretization parameter r.

2.7. Ill–Posed and Inverse Problems

The nonlinear operator F might not have a boundedly invertible linearization, and $Fu = f$ might not have a solution at all. This is a standard situation for ill–posed or inverse problems.

Theorem 2. *Replace the assumptions in Section 2.1 by*

$$\begin{array}{rcl}
\|Fu^* - f\|_{\mathcal{V}} & = & \delta > 0 \\
\|Fu - Fv\|_{\mathcal{V}} & \leq & C_F\|u - v\|_{\mathcal{U}} \text{ for all } u, v \in K_R(u^*),
\end{array}$$

for some $R > 0$ and leave the rest of the assumptions in Theorem 1 unchanged. Then there is an error bound

$$\|f - F\hat{u}_r\|_{\mathcal{V}} \leq \delta + \epsilon(r, u^*)C_F + 3C_S(r, s)(\delta(r, s, u^*) + \delta(r, s, f))$$

using (12) and

$$\delta(r, s, f) := \|T_s(Fu_r^* - f)\|_{\mathcal{V}_s}.$$

PROOF. We proceed like in Theorem 1 to get

$$\begin{aligned}
\|f - F\hat{u}_r\|_{\mathcal{V}} &\leq \|f - Fu_r^*\|_{\mathcal{V}} + \|Fu_r^* - F\hat{u}_r\|_{\mathcal{V}} \\
&\leq \delta + \|Fu_r^* - Fu^*\|_{\mathcal{V}} + \|Fu_r^* - F\hat{u}_r\|_{\mathcal{V}} \\
&\leq \delta + \epsilon(r,u^*)C_F + C_S(r,s)\|T_s(Fu_r^* - F\hat{u}_r)\|_{\mathcal{V}_s} \\
&\leq \delta + \epsilon(r,u^*)C_F + C_S(r,s) \\
&\quad \times (\|T_s(Fu_r^* - f)\|_{\mathcal{V}_s} + \|T_s(f - F\hat{u}_r)\|_{\mathcal{V}_s}) \\
&\leq \delta + \epsilon(r,u^*)C_F \\
&\quad + C_S(r,s)\,(\|T_s(Fu_r^* - f)\|_{\mathcal{V}_s} + 2\|T_s(f - Fu_r^*)\|_{\mathcal{V}_s}) \\
&\leq \delta + \epsilon(r,u^*)C_F + 3C_S(r,s)\|T_s(Fu_r^* - f)\|_{\mathcal{V}_s} \\
&\leq \delta + \epsilon(r,u^*)C_F + 3C_S(r,s)\|T_s(Fu_r^* - Fu^*)\|_{\mathcal{V}_s} \\
&\quad + 3C_S(r,s)\|T_s(Fu^* - f)\|_{\mathcal{V}_s} \\
&\leq \delta + \epsilon(r,u^*)C_F + 3C_S(r,s)(\delta(r,s,u^*) + \delta(r,s,f)). \quad \square
\end{aligned}$$

This means that under uniform stability one can reproduce f approximately with good accuracy.

3. Example

As an illustration for the above framework we consider nonlinear elliptic PDEs of second order in weak and strong forms with Dirichlet boundary conditions on a bounded Lipschitz domain $\Omega \subset \mathbb{R}^2$. As a simple working example we pose the problem

$$\begin{aligned}
-\Delta u + g(u) &= f_1 \quad \text{on } \Omega \\
u &= f_2 \quad \text{on } \partial\Omega
\end{aligned} \tag{13}$$

with a nonlinear function g. This can be considered under various regularity assumptions on the nonlinear g, the right–hand sides, the domain Ω, and the solution u^*.

To bring the problem into the operator equation form (1), (2), we have to fix the regularity assumptions and the testing strategy. If we pose the problem in *strong* form, we can define

$$Fu := (Gu := -\Delta u + g(u), u|_{\partial\Omega}),\ u \in \mathcal{U} := H^m(\Omega),\ Fu \in \mathcal{V} := \mathcal{V}_1 \times \mathcal{V}_2$$

with values in $\mathcal{V} := \mathcal{V}_1 \times \mathcal{V}_2 := H^{m-2}(\Omega) \times H^{m-1/2}(\partial\Omega)$ and reformulate the problem as an identity

$$Fu = f = (f_1, f_2) \in \mathcal{V} = \mathcal{V}_1 \times \mathcal{V}_2 \tag{14}$$

of functions.

Strong testing of our example problem (13) in strong form (14) is done by collocation. We fix point sets

$$X_s = \{x_1, \ldots, x_{K(s)}\} \subset \overline{\Omega} \subset \mathbb{R}^n,\ \textit{or } X_s \subset \Omega,\ Y_s = \{y_1, \ldots, y_{K'(s)}\} \subset \partial\Omega,$$

and observe that a strong solution u will satisfy the equations

$$\begin{aligned} T_s^1 : -\Delta u(x_k) + g\,(u)\,(x_k) &= f_1(x_k),\ 1 \le k \le K(s), \\ T_s^2 : u(y_{k'}) &= f_2(y_{k'}),\ 1 \le k' \le K'(s). \end{aligned} \tag{15}$$

which is (6), since the test space is $\mathcal{V}_s = \mathcal{V}_s^1 \times \mathcal{V}_s^2 = \mathbb{R}^{K(s)+K'(s)}$, and the test map T_s just performs discrete function evaluations on the sets X_s and Y_s. This gives a large nonlinear system of equations, and if these are formulated "*entirely in terms of nodes*" by parametrizing the trial space accordingly, we have a *meshless method* in the sense of [5]. But collocation also works for trial spaces with other parametrizations, e.g. those which are used in *spectral* or *pseudospectral* techniques. It should be clear how this generalizes to other operator equations, including systems, and to other boundary conditions. We shall explicitly formulate this for quasilinear and fully nonlinear equations in a forthcoming paper. Note that collocation by point evaluation requires a regularity $u^* \in C^2(\Omega)$ or $u^* \in C^2(\overline{\Omega})$ or $u^* \in H^m(\Omega)$ of at least $m-2 > n/2$ in n dimensions, i.e. $m > 3$ in two dimensions.

There are different ways to bring (13) into the form (1), (2), when focusing on *weak* testing. All cases employ integration by parts of $-\Delta u$ against a test function v, but they differ in the way they handle the boundary integral arising in

$$-\int_\Omega v\,\Delta u = \int_\Omega \nabla u \cdot \nabla v - \int_{\partial\Omega} v\,\frac{\partial u}{\partial n} =: Lu(v).$$

Anyway, the linear operator $-\Delta$ mapping functions to functions now turns into an operator L mapping functions to linear functionals by $u \mapsto Lu$. This will influence the way we shall define the operator F in (1), (2).

The standard weak form uses test functions v that vanish on the boundary, and then $(u, v) := (Lu, v)$ is a symmetric bilinear form that is well–defined and an inner product on the Hilbert space $H_0^1(\Omega)$ which is the $H^1(\Omega)$ closure of smooth functions on $\overline{\Omega}$ vanishing on the boundary. Thus the first equation of (13) is reformulated in weak form as

$$\int_\Omega (\nabla u \cdot \nabla v + g(u)v) = \int_\Omega v\,f_1 \tag{16}$$

for all $v \in H_0^1(\Omega)$.

To use the symmetry of the bilinear form, the boundary conditions are usually made homogeneous by introducing a function $u_0 \in H^1(\Omega)$ that satisfies the Dirichlet boundary conditions exactly. In terms of a new unknown function $w := u - u_0 \in H_0^1(\Omega)$ the identity (16) turns into

$$\int_\Omega (\nabla w \cdot \nabla v + g(w+u_0)v) = \int_\Omega v\,f_1 - \int_\Omega \nabla u_0 \cdot \nabla v =: \lambda(v) \tag{17}$$

for all $v \in H_0^1(\Omega)$.

This defines a (weak) operator, now in the form

$$G = G_w \ : \ \mathcal{U} := H_0^1(\Omega) \to \mathcal{V} := H_0^1(\Omega)^* \tag{18}$$

via $\tilde{g}(w) := g(w + u_0)$ and

$$G_w w = Gw := (v \mapsto \int_\Omega (\nabla w \cdot \nabla v + \tilde{g}(w)v)) \in \mathcal{V}$$

and reduces the differential equation in (1) and (2) into $Gu = \lambda$ and the functional λ defined in (17).

For weak testing, we fix test functions $v_1, \ldots, v_{K(s)}$ from $H_0^1(\Omega)$ instead of the set X_s used in strong testing. The test map T_s^1 on $\mathcal{V}_1 = H_0^1(\Omega)^*$ in (15) acts on elements $\mu \in \mathcal{V}_1$ as

$$\mu \mapsto (\mu(v_1), \ldots, \mu(v_{K(s)}))^T \in \mathcal{V}_s^1 = \mathbb{R}^{K(s)}.$$

The second component T_s^2 of T_s in (15) remains unchanged. The solution $w^* = u^* - u_0 \in H_0^1(\Omega)$ will satisfy (6) in the form

$$\int_\Omega (\nabla w^* \cdot \nabla v_k + \tilde{g}(w^*)v_k) \quad = \quad \int_\Omega f_1 \, v_k - \int_\Omega \nabla u_0 \cdot \nabla v_k, \quad 1 \le k \le K(s).$$

In this context, it must be kept in mind that no numerical analyst can work in spaces like $H_0^1(\Omega)$. Thus, certain manageable subspaces are used that require additional regularity, at least of the approximations to the solution, and that allow numerical integration with sufficiently small errors. Furthermore, the construction of u_0 is open in some cases.

The above treatment of weak problems in (18) makes use of the identity $\mathcal{U}^* = \mathcal{V}$ and can identify the trial space $\mathcal{U}_r$ with the span of the test functions. This technique is standard for finite element spaces and allows a rather simple convergence analysis with a low convergence rate.

But to show that our discretization theory works much more generally, we do not want to confine ourselves to the above special case. One way is to use test functions from $H_0^1(\Omega)$, but to collocate the boundary data strongly, working with a solution space $\mathcal{U} = H^m(\Omega)$, or $\mathcal{U} = C^2(\Omega)$, and avoiding the construction of the additional function u_0 in (17). Then the first equation of (1), (2) is reformulated as

$$\int_\Omega (\nabla u \cdot \nabla v + g(u)v) = \int_\Omega v \, f_1$$

for all $v \in H_0^1(\Omega)$. This defines a map

$$F \ : \ \mathcal{U} := H^m(\Omega) \to \mathcal{V} := H_0^1(\Omega)^* \times H^{1/2}(\partial\Omega)$$

via

$$Fu := (v \mapsto \int_\Omega (\nabla u \cdot \nabla v + g(u)v), u|_{\partial\Omega}) \in \mathcal{V}$$

and poses the equation (1), (2) with $f = (\lambda_1, f_2)$ and the functional

$$\lambda_1(v) := \int_\Omega v\, f_1 \text{ for all } v \in H_0^1(\Omega),$$

where, formally, the function f_1 is allowed to be in $H^{-1}(\Omega)$. Note that, due to strong collocation of the boundary values, this approach needs regularity $m - 1/2 > (n-1)/2$, i.e. $m > 1$ for $n = 2$. This seems to fall behind the standard weak case, but it is unclear how to find a function $u_0 \in H^1(\Omega)$ with the prescribed function values of some $u \in H^1(\Omega)$ on $\partial\Omega$, if the data function f_2 on the boundary is only in $H^{1/2}$ because u is only in $H^1(\Omega)$. The workaround via u_0 is kind of a cheat that conceals additional regularity needed for actual numerical calculations.

For completeness, we also want to point out how to fit a variation of the Meshless Local Petrov–Galerkin method of Atluri and collaborators [6] into this framework. There, degrees of freedom for testing are not only introduced by test functions, but also by allowing plenty of small local domains Ω_h of integration on which the integration by parts is performed. Thus (13) turns into

$$\int_{\Omega_h} (\nabla u \cdot \nabla v + g(u)v) + \int_{\partial\Omega_h} v\, \frac{\partial u}{\partial n} = \int_{\Omega_h} v\, f_1$$

and one can even use $v = 1$ to arrive at test equations

$$\int_{\Omega_h} g(u) + \int_{\partial\Omega_h} \frac{\partial u}{\partial n} = \int_{\Omega_h} f_1$$

that have to be written down for many local domains Ω_h. Dirichlet boundary conditions can be added as before by strong collocation. The map F is defined as in the previous case, but its first component will map to a functional on the span of characteristic functions on subdomains.

The huge variety of strong and weak formulations and their possible mixtures indicates that it must be a major problem to prove fairly general stability inequalities. We shall come back to this problem in Section 5. Note also that variations of weak formulations modify the operator of the general equation (1), (2), i.e. they essentially change the problem itself.

For all of these techniques, one can replace $u \mapsto -\Delta u + g(u)$ by a general nonlinear second order elliptic operator G with an again *elliptic* Fréchet derivative $G'(u)$. However, usually the simple choice of $\mathcal{U} = H^m(\Omega)$ has to be carefully monitored. In fact, for a nonlinear G the $G(u)$ is only defined in appropriate subsets of $\mathcal{U} = H^m(\Omega)$. This problem has to be discussed for every new problem. Analogously, one can handle the boundary conditions, but sometimes they can nicely be shifted into the choice of the trial space. In a forthcoming paper we shall consider quasilinear and fully nonlinear equations.

4. Linearization

We now describe how the a–priori properties of the previous sections for a nonlinear F can be deduced from its linearized problem operator, $F'(u^*)$. This

allows us to go partially back to the linear case treated in [1].

4.1. Well–Posedness

To derive the well–posedness in the sense of the previous section from linearization, we invoke standard perturbation arguments to yield

Theorem 3. *Let F be Fréchet–differentiable in each point $u \in K_{R'}(u^*)$ for some $u^* \in \mathcal{U}$ and let the Fréchet derivatives $F'(u)$ at u be bounded and Lipschitz continuous, i.e.*

$$\|F'(v) - F'(u)\| := \|F'(v) - F'(u)\|_{\mathcal{L}(\mathcal{U},\mathcal{V})} \leq C''\|u - v\|_{\mathcal{U}} \text{ for all } u, v \in K_{R'}(u^*). \tag{19}$$

Finally, let $F'(u^*)$ have a bounded inverse. Then the assumption (3) of Section 2.1 holds for F in some ball of positive radius at most $R \leq R'$ around u^*. Furthermore, all Fréchet derivatives are uniformly bounded and have uniformly bounded inverses in $K_R(u^*)$.

Proof. The inequality (19) implies

$$\|F'(u)\| \leq \|F'(u^*)\| + C''R' \text{ for all } u \in K_{R'}(u^*) \tag{20}$$

and, combined with the Taylor formula (here $\overline{uv}$ indicates the closed linear segment connecting the end points u, v)

$$\|Fv - Fu - F'(u)(v - u)\| \leq \|v - u\| \sup_{x \in \overline{uv}} \|F'(x) - F'(u)\|, \tag{21}$$

it yields, by (19), (20), (21), the inequality

$$\begin{aligned} \|Fu - Fv\|_{\mathcal{V}} &\leq \|Fv - Fu - F'(u)(v - u)\|_{\mathcal{V}} + \|F'(u)(v - u)\|_{\mathcal{V}} \\ &\leq C''\|u - v\|_{\mathcal{U}}^2 + (\|F'(u^*)\| + C''R')\,\|u - v\|_{\mathcal{U}} \\ &\leq C''R'\,(2 + \|F'(u^*)\|)\,\|u - v\|_{\mathcal{U}} \end{aligned}$$

for all u, $v \in K_{R'}(u^*)$, hence the right hand side of inequality (3).

For the left inequality in (3), we assume the radius $R \leq R'$ of $K_R(u^*)$ to be small enough to satisfy

$$R\|(F'(u^*))^{-1}\|C'' \leq \frac{1}{4}. \tag{22}$$

For all $u \in K_R(u^*)$, $v \in \mathcal{U}$ we get

$$\begin{aligned} \|v\|_{\mathcal{U}} &\leq \|(F'(u^*))^{-1}\|\|F'(u^*)v\|_{\mathcal{V}} \\ &= \|(F'(u^*))^{-1}\|\,(\|F'(u^*)v - F'(u)v\|_{\mathcal{V}} + \|F'(u)v\|_{\mathcal{V}}) \\ &\leq \|(F'(u^*))^{-1}\|\,(C''\|u - u^*\|_{\mathcal{U}}\|v\|_{\mathcal{U}} + \|F'(u)v\|_{\mathcal{V}}) \\ &\leq \frac{1}{2}\|v\|_{\mathcal{U}} + \|(F'(u^*))^{-1}\|\|F'(u)v\|_{\mathcal{V}} \end{aligned}$$

by (22), even with $\leq 1/2$ instead of $\leq 1/4$, and via (19). This implies

$$\|v\|_{\mathcal{U}} \leq 2\|(F'(u^*))^{-1}\|\|F'(u)v\|_{\mathcal{V}}$$

for all $v \in \mathcal{U}$. In particular, $F'(u)v = 0$ implies $v = 0$, so $F'(u)$ is injective and thus $(F'(u))^{-1} : \mathcal{R}(F'(u)) \to \mathcal{U}$ exists and satisfies

$$\|(F'(u))^{-1}\| \quad \leq \quad 2\|(F'(u^*))^{-1}\| \tag{23}$$

for all $u \in K_R(u^*)$, proving uniform bounded invertibility of all local Fréchet derivatives. Then we get

$$\begin{array}{rcl} \|v-u\|_{\mathcal{U}} & \leq & \|(F'(u))^{-1}\|\|F'(u)(v-u)\|_{\mathcal{V}} \\ & \leq & \|(F'(u))^{-1}\|\|F'(u)(v-u) - Fv + Fu\|_{\mathcal{V}} \\ & & +\|(F'(u))^{-1}\|\|Fu - Fv\|_{\mathcal{V}} \\ & \leq & \|(F'(u))^{-1}\|C''\|u-v\|_{\mathcal{U}}^2 + \|(F'(u))^{-1}\|\|Fu - Fv\|_{\mathcal{V}} \\ & \leq & \dfrac{1}{2}\|u-v\|_{\mathcal{U}} + \|(F'(u))^{-1}\|\|Fu - Fv\|_{\mathcal{V}} \end{array}$$

using (19), (21), (22), now with $\leq 1/4$, and finally by (23)

$$\begin{array}{rcl} \|v-u\|_{\mathcal{U}} & \leq & 2\|(F'(u))^{-1}\|\|Fu - Fv\|_{\mathcal{V}} \\ & \leq & 4\|(F'(u^*))^{-1}\|\|Fu - Fv\|_{\mathcal{V}} \end{array} \tag{24}$$

for all $u, v \in K_R(u^*)$, hence the left hand side of inequality (3) with $c_F = 4\|(F'(u^*))^{-1}\|$. □

4.2. Stability

To derive stability inequalities of test maps in the sense of the previous section from linearization, we continue using Theorem 3 with some additional assumptions concerning testing. The basic idea is to repeat the proof of Theorem 3 for the maps $G_s := T_s \circ F$.

Theorem 4. *Assume the hypotheses of Theorem 3 and the existence of Lipschitz continuous Fréchet derivatives of $G_s := T_sF$ like (19), i.e.*

$$\|T_sF'(v) - T_sF'(u)\|_{\mathcal{L}(\mathcal{U},\mathcal{V}_s)} \leq C''(s)\|u-v\|_{\mathcal{U}} \text{ for all } u, v \in K_{R'}(u^*) \cap \mathcal{U}_R \tag{25}$$

with some constant $C''(s)$. Assume further that $G_s'(u^*)$ has a bounded inverse and the linearized problem at u^* has a stability inequality with a constant $C_S(r,s)$. Then on a ball $K_R(u^*) \cap \mathcal{U}_R$ with a radius R satisfying

$$RC_S(r,s)C''(s) \leq \frac{1}{4} \tag{26}$$

the nonlinear problem satisfies a stability inequality with constant $4c_FC_S(r,s)$.

Proof. The maps $G_s'(u) = T_sF'(u)$ restricted to $\mathcal{U}_r$ are linear maps between finite–dimensional linear spaces, thus continuous. The proof structure of Theorem 3 just needs existence of the Fréchet derivatives and (19), but no other explicit quantitative form of Fréchet differentiability. Thus we can use it with (25) and $C''(s)$. Stability of the linearized problem means

$$\|F'(u^*)v_r\|_{\mathcal{V}} \leq C_S(r,s)\|T_sF'(u^*)v_r\|_{\mathcal{V}_s} \text{ for all } v_r \in \mathcal{U}_r, \tag{27}$$

and this is the boundedness of the inverse of $G'_s(u^*)$ on $\mathcal{U}_r$. We can now follow the proof of Theorem 3 verbatim, and the crucial condition for the radius R is now to be posed like in (22) and using (27) as

$$R\|(G'_s(u^*))^{-1}\|C''(s) \leq RC_S(r,s)C''(s) \leq \frac{1}{4}.$$

Thus by the proof of Theorem 3 with F replaced by $G'_s = (T_s \circ F)'$ we obtain (29) with

$$c(r,s) \leq 4C_S(r,s). \tag{28}$$

The inequalities (27) and (28), combined with (24), yield the updated inequality

$$\|u_r - v_r\|_{\mathcal{U}} \leq c(r,s)\|T_sFu_r - T_sFv_r\|_{\mathcal{V}_s} \tag{29}$$

on a neighborhood of u_r^* in $\mathcal{U}_r \subset \mathcal{U}_R \subset \mathcal{U}$.
The stability of the nonlinear case now follows with (29) and with the right hand side of inequality (3), which is proved in Theorem 3 for F, via

$$\begin{aligned} \|Fu_r - Fv_r\|_{\mathcal{V}} &\leq C_F\|u_r - v_r\|_{\mathcal{U}} \\ &\leq C_F c(r,s)\|T_sFu_r - T_sFv_r\|_{\mathcal{V}_s} \\ &\leq 4C_FC_S(r,s)\|T_sFu_r - T_sFv_r\|_{\mathcal{V}_s} \end{aligned}$$

holding on a neighborhood of both u^* and u_r^*. □

We then have to make r small enough to ensure that the neighborhoods of u^* for Theorem 3 and the neighborhood of u_r^* at the end of the last proof have a nonempty intersection that is a neighborhood of both u^* and u_r^*.
The upshot is that the constant in the stability inequality just takes a factor of $4c_F$ when going from the linear to the nonlinear case, but only on a ball with a radius R that may dramatically depend on r and s.

Corollary 1. *If the trial/test strategy $s(r)$ is uniformly stable in the sense of Definition 2 for the linear case, Theorem 4 yields uniform stability also in the nonlinar case, cf (26), but on a strongly discretization–dependent neighborhood. If $C''(r,s(r))$ is bounded uniformly, the uniform stability holds in the nonlinear case for a fixed radius R. This is correct e.g. for Lipschitz continuous F' and bounded T_s.*

4.3. Numerical Solution

To allow numerical solutions via linearization, we could iterate by residual minimization

$$u_{j+1} := \arg\min\{\|T_s(f - F'(u_j)(u_r - u_j))\|_{\mathcal{V}_s} \ : \ u_r \in \mathcal{U}_r \cap K_R(u^*)\}$$

starting from some $u_0 \in \mathcal{U}_r$ sufficiently close to u_r^*. But we can simply invoke any method for minimizing the residual in (8) and leave linearization to the optimizer. We shall demonstrate this in section 6 for two examples.

5. Proving Stability

We now come back to the stability problem and want to formulate a convenient framework (see [3]) to prove stability inequalities. We outline it here for convenience and for application in examples like those in sections 3 and 6. By Theorems 3 and 4 we can restrict the discussion to linear operators $F = L$.

Consider test maps $T_s : \mathcal{V} \to \mathcal{V}_s$ like in sections 2 and 3. In many cases, testing can be analyzed independently of the underlying linear operator equation $L\,u = f$ by applying the operator L to the trial spaces $\mathcal{U}_r$ to get subspaces

$$\mathcal{W}_r \subset \mathcal{V} \text{ defined as } \mathcal{W}_r := L\,\mathcal{U}_r$$

on which the test maps T_s act. Thus we assume a scale of subspaces $\mathcal{W}_r \subset \mathcal{V}$ carrying hidden information on the linear operator and the trial space. Note that these functions are *not* acting as test functions in the usual sense. They usually are just the images of the *trial* functions under the operator, and consequently their span uses the t**R**ial scale parameter r. Note that due to $\mathcal{W}_r = L\mathcal{U}_r$ we have $\mathcal{W}_r \subset \mathcal{W}_{r'}$ for $r > r'$.

Usually we will employ values of $s = s(r) < r$ with $\dim \mathcal{V}_s > \dim \mathcal{W}_r$. This prerequisite of a useful trial/test strategy $s(r)$ fits well with the residual minimization in 4.3, and we can expect that excessive testing will improve stability. Then we go for a stability inequality, like (9) in the linearized form

$$\|w_r\|_{\mathcal{V}} \le C_S(r,s)\|T_s w_r\|_{\mathcal{V}_s} \text{ for all } w_r \in \mathcal{W}_r, \text{ for all } r, s \to 0, \tag{30}$$

which formally is not related to operator equations anymore.

These stability inequalities can be obtained by combining *sampling and inverse inequalities*, as we shall outline now. A typical sampling inequality is

$$\|w\|_{\mathcal{V}} \le C(\epsilon(s)\|w\|_{\mathcal{W}} + \|T_s w\|_{\mathcal{V}_s}) \forall w \in \mathcal{W}, \tag{31}$$

with a (smooth) subspace $\mathcal{W}$ of $\mathcal{V}$, and $\Omega \subset \mathbb{R}^n$. A typical case is $\mathcal{W} = H^m(\Omega)$ $m - 2 > n/2$. For differential operators G of order 2, the last inequality allows the point evaluation of $(Gu)(x_k)$ for $x_k \in \Omega$. Often in (31) the norm $\|w\|_{\mathcal{W}}$ is replaced by the corresponding semi norm $|w|_{\mathcal{W}}$.

The inequality (31) should be interpreted as follows. The subspace $\mathcal{W}$ has additional smoothness and allows to bound the weaker norm $\|w\|_{\mathcal{V}}$ by the stronger norm $\|w\|_{\mathcal{W}}$ with a small factor $\epsilon(s)$ provided that the discrete test data $T_s w$ are small. The basic principle is that a function is small in a weak norm, if it has a bound in a strong norm and takes very small values at plenty of points well–distributed in the domain. Some typical examples are in [3], and in Section 6.

Such a sampling inequality yields

$$\|w_r\|_{\mathcal{V}} \le C(\epsilon(s)\|w_r\|_{\mathcal{W}} + \|T_s w_r\|_{\mathcal{V}_s}) \ \forall w_r \in \mathcal{W}_r.$$

The second part of the right–hand side is what we want, provided that the assumption $\mathcal{W}_r \subset \mathcal{W}$ is correct, which we assume from now on. In particular,

for smooth trial spaces, like for instance, "kernel–based" and "spectral" spaces we usually will get $\mathcal{W}_r \subset \mathcal{W}$ without any problems.

To eliminate the first part, one can use an *inverse inequality*

$$\|w_r\|_{\mathcal{W}} \leq D(r)\|w_r\|_{\mathcal{V}} \text{ for all } w_r \in \mathcal{W}_r \tag{32}$$

with a constant $D(r)$ which normally increases when r decreases. Together we have

$$\|w_r\|_{\mathcal{V}} \leq C(\epsilon(s)D(r)\|w_r\|_{\mathcal{V}} + \|T_s w_r\|_{\mathcal{V}_s}) \text{ for all } w_r \in \mathcal{W}_r,$$

and if one can guarantee

$$C\epsilon(s)D(r) \leq \frac{1}{2} \tag{33}$$

by a suitable choice of r and s, then

$$\|w_r\|_{\mathcal{V}} \leq 2C\|T_s w_r\|_{\mathcal{V}_s} \text{ for all } w_r \in \mathcal{W}_r \tag{34}$$

is a stability inequality of the form (30), which leads to uniform stability (10).

To get (33) one has to pick a suitably fine test discretization (i.e. a suitably small s) to provide a stable testing of the finite–dimensional space $\mathcal{V}_r$ with a possibly rather large $D(r)$. In ideal cases, one has

$$\begin{array}{rcl} \epsilon(s) & \leq & C_1 s^{\beta} \\ D(r) & \leq & C_2 r^{-\beta} \end{array} \tag{35}$$

with the same positive exponent β, and then uniform stability is guaranteed if the quotient s/r is sufficiently small.

Theorem 5. *Under the conditions (31), (32), (35) for a sufficiently small quotient s/r we obtain the stability inequality (34) and uniform stability.*

Note that this framework leaves open how to establish the ingredients (31) and (32) to prove stability inequalities (30), and how to guarantee the sufficient condition (33) for uniform stability. But this is a purely theoretical issue. In practice, users will first specify their trial space, thus fixing r and determining implicitly the achievable error. Then they will pick a test strategy that is fine enough to guarantee that the numerical subproblems do not suffer from rank loss or instability. If instabilities occur, testing must be finer. If the error is too large, the trial space must be enlarged, possibly requiring a finer testing as well. We give two specific examples in the next section.

6. Examples

Here, we present two numerical cases illustrating our nonlinear discretization theory. We focus on calculations first, and then explain how to prove uniform stability in these cases.

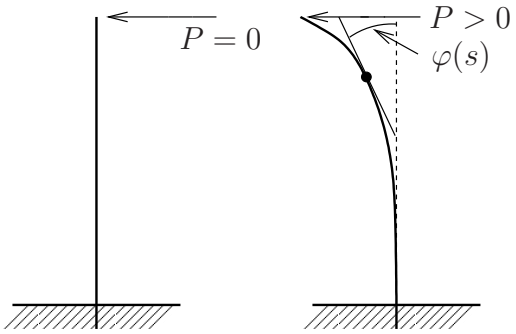

Figure 1: Rod with load perpendicular to the rod

6.1. Bending rod with perpendicular load

This example was carried through with quite some help of Alois Steindl [7].

Let a vertically positioned rod of length L be clamped at the lower end and be free at the upper end, see Figure 1. A load P at the end of the rod, originally perpendicular to the rod, forces it to bend sidewards. The solution φ is written in terms of the angle $\varphi(s)$ at arclength s with respect to the vertical axis. It satisfies the *nonlinear boundary value problem* with the *differential equation*

$$G(\varphi, \lambda) := \frac{d^2\varphi}{ds^2} + \lambda \cos\varphi = 0, \quad \text{for} \quad \lambda := P/\alpha, \tag{36}$$

and the *boundary conditions,* defined by the subspace

$$C_b^2[0, L] := \{\varphi \in C^2[0, L]; \quad \varphi(0) = \frac{d\varphi}{ds}(L) = 0\}, \tag{37}$$

where α is an elasticity parameter.

We transform by $s = t/\sqrt{\lambda}$ to get for $\psi(t) := \varphi(t/\sqrt{\lambda})$ the ODE

$$\psi''(t) + \cos(\psi(t)) = 0$$

with boundary conditions $\psi(0) = 0$ and $\psi'(T) = 0$ where now

$$0 \le t \le T = L\sqrt{\lambda}.$$

The physically interesting solutions lie in the positive quadrant of phase space, and we focus on the solution for $T = 2$ here, which could possibly be obtaind via shooting from a suitable value of $\psi'(0)$, but we want to apply our discretization theory.

The solutions are clearly in C^∞, and to make use of this smoothness, it makes no sense to use a standard h–type finite element discretization. Spectral or p–type FEM techniques are preferable. We simply use polynomials of some degree N as a trial space, and then we can expect that we can approximate the solution with spectral convergence like q^N with some $q < 1$ for $N \to \infty$,

no matter which norm we choose to measure the error. If we have a stable test discretization, we should see this convergence behavior, since our discretization theory implies that the final error is roughly the approximation error, if there is uniform stability. Thus we can expect to get away with moderate values of N, and this should be manageable by a moderate amount of testing.

To set the problem up in MATLAB, we parametrize the trial space via monomials in $[0, 2]$. To incorporate the boundary conditions, we set the lowest coefficient to zero and calculate the highest coefficient to let the derivative vanish at $T = 2$. This means that we only have $N - 1$ variables for degree N. Testing is done on equidistant points of spacing h, leading to test points $x_j = jh,\ 0 \leq j \leq M := 2/h \in \mathbb{N}$. We avoid to linearize the problem and let the optimizing algorithm do the linearization. Therefore we simply invoke the MATLAB routine `lsqnonlin` that minimizes $\|Fa\|_2^2$ for the nonlinear map $F : \mathbb{R}^{N-1} \to \mathbb{R}^{M+1}$ with

$$F_{j+1}a \quad = \quad p_a''(x_j) + \cos(p_a(x_j)),\ 0 \leq j \leq M$$

where p_a is the polynomial of degree N parametrized by a satisfying the boundary conditions. Of course, all of this should be implemented via the Chebyshev basis or Nick Trefethen's `chebfun`[1], but we used the simple `polyval` routine of MATLAB.

In view of our theory, we should make M large enough in order to ensure stability, but it turns out numerically that $M = 20$ suffices for degrees $N \leq 17$. Figure 2 shows the *Root Mean Square Error norm* $\|.\|_{\mathcal{V}_M} = h^{1/2}\|.\|_{\ell_2} = \|.\|_{RMSE}$ of the residual $p_a'' + \sin(p_a)$ for the optimized parameter vector $a \in \mathbb{R}^{N+1}$, evaluated on 10001 points in $[0, 2]$ as a function of N for fixed $h = 0.1$ and $M = 20$. Note that the RMSE norm is the appropriate discrete form of the continuous L_2 norm we shall use in the stability analysis below. The spectral convergence is clearly visible, and plots for smaller h look the same. The routine `lsqnonlin` terminates at 10^{-8} reached for $N = 15$, such that larger $N > 15$ will require a specially tuned–up nonlinear optimizer. The maximal problem size is 14×21 for 10^{-8} accuracy at $N = 15$.

Since we have a nonlinear problem with a trivial solution, the start of the optimization is important. We chose the quadratic polynomial $2x/T - x^2/T^2$ satisfying the boundary conditions to start for $N = 3$, and we used the optimal coefficients of each run as a starting value for the next polynomial with one degree higher. Though the calculations show that stability is not a major problem in this case, we add an argument at the end of this section showing that sufficiently small h ensures uniform stability.

[1] `http://www2.maths.ox.ac.uk/chebfun`

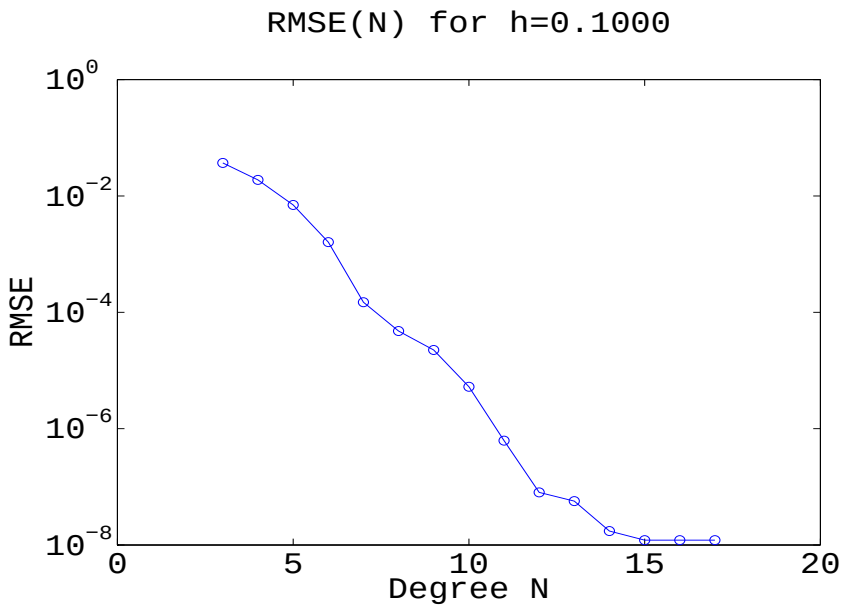

Figure 2: Error for varying trial space degree

6.2. Chladny sound figures

This well–known phenomenon is described by the nonlinear PDE

$$G(u,\lambda) := \Delta u + \lambda \sin u = 0 \text{ in } \Omega = [0,1]\times[0,1] \text{ defined on} \tag{38}$$

$$C_b^2(\Omega) := \{u \in C^2(\Omega); \quad \frac{\partial u}{\partial n} = 0 \text{ on } \partial\Omega\}. \tag{39}$$

where $u = u(x,y)$ is the deviation from the trivial flat position of the plate at the point (x,y). Here $\partial u/\partial n$ is the *normal derivative* into the outer direction of $\partial\Omega$ and $\Delta := \partial^2/\partial x^2 + \partial^2/\partial y^2$ the *Laplacian operator*. For arbitrary λ the trivial flat state $u(x,y) \equiv 0$ represents a solution of ((38)), ((39)). This equation has the awkward property that it has infinitely many trivial solutions for all λ, namely the constant functions $u = k\pi$ for $k \in \mathbb{Z}$.

The eigenvalue problem for the Laplacian, simultaneously $G_u(0,-\lambda)v$, is

$$G_u(0,-\lambda)v = \Delta v - \lambda v = 0.$$

It has the eigenfunctions

$$v_{m,n}(x,y) = \cos(m\pi x)\cos(n\pi y)$$

and corresponding eigenvalues $\lambda_{m,n} := -(m^2+n^2)\pi^2$. A typical case is in Figure 3 which shows $v_{1,3} - v_{3,1}$, one of the many possible linear combinations of eigenfunctions with specific symmetry properties.

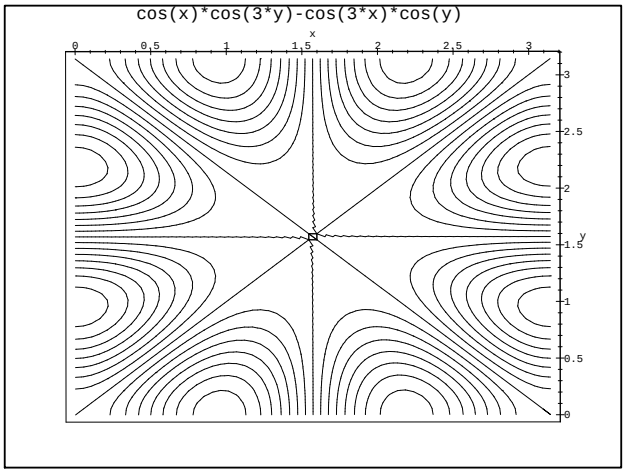

Figure 3: Contour lines of typical eigenfunctions

As in our first example, the solution will be infinitely differentiable, and we should go for a method with spectral convergence. As the above discussion shows, we have the functions $v_{m,n}$, $1 \le m \le M$, $1 \le n \le N$ as ideal candidates for trial functions approximating nontrivial solutions of the nonlinear system. But if the PDE is discretized some way or other, it will still have the unpleasant property that it has infinitely many solutions for each real λ, namely the constants $k\pi$ for all $k \in \mathbb{Z}$, and this will cause trouble for all algorithms minimizing residuals.

Anyway, it makes sense to implement a trial space spanned by the $v_{m,n}$ for $0 \le m, n \le N$. Mind that these $v_{m,n}(x, y)$ are the eigenfunctions of the Laplacian and that these trigonometric functions satisfying the boundary conditions have optimal approximation properties. We parametrize the trial functions $u_A(x, y)$ by the $(N + 1) \times (N + 1)$ matrices A and get

$$u_A(x, y) = \sum_{m,n=0}^{N} a_{m,n} v_{m,n}(x, y).$$

These functions are evaluated on a test grid $X_M = X_h \times X_h$ of $(M + 1)^2$ points in $[0, 1]^2$. Note that the action of $A = \Delta$ on such functions is represented by the elementwise (Hadamard or Schur) product of the matrices Λ and A and can be written as $\Lambda. * A$ in MATLAB notation, if Λ is the matrix of eigenvalues. The evaluation on a grid set $X_h \times X_h$ is just matrix multiplication, if we first calculate a matrix Z with entries $\cos(h\pi mj)$, $0 \le m \le N$, $0 \le j \le M$. The residual $\Delta u_A + \lambda \sin(u_A)$ of u_A in MATLAB lingo is one line:

$$Z' * (\Lambda. * A) * Z + \lambda * \sin(Z' * A * Z).$$

The idea is now to solve the equation

$$Fu := \Delta u + \lambda \sin(u) = 0$$

for some fixed λ after discretization, and by minimizing $\|Fu_A\|^2_{2,X_h \times Y_h}$ by `lsqnonlin`.

We need good starting values to avoid that the solver runs into $u_A = 0$. If $\sin(u)$ is linearized for small u, the equation turns into $\Delta u + \lambda u = 0$, and this has solutions $v_{m,n}$ with positive $\lambda = -\lambda_{m,n}$. Thus the starting function should be $v_{m,n}$ while λ is kept fixed somewhat larger than $-\lambda_{m,n}$. If we do this for $m = n = 1$ and $\lambda = 20 = \lceil -\lambda_{1,1} + 0.1 \rceil$ on a test discretization with $M = 50$, i.e. on 51×51 test points for varying trial degree $6 \le N \le 15$, we get Figure 4. Residuals were calculated as Root Mean Square Errors $h\|.\|_{\ell_2} = \|.\|_{RMSE}$ for $h = 0.01$, i.e. on a 101×101 grid. The exact solution has the D_4-symmetry of the square. This implies that the approximation error for the polynomial of next higher even degree is the same as for preceding odd degree. The case for degree 9 is in Figures 5 and 6.

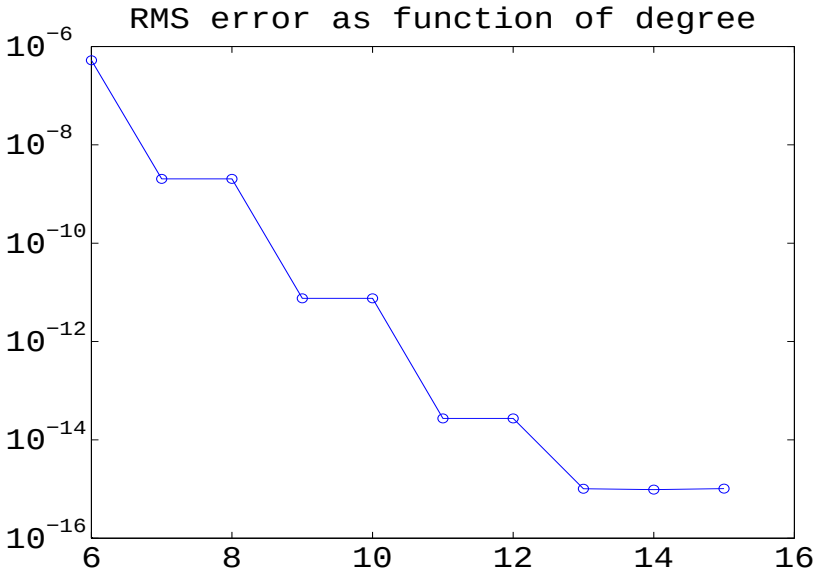

Figure 4: Error for varying trial space degree

6.3. Stability

Both examples can be written as elliptic equations

$$Fu := \Delta u - g\, u = f \in \Omega$$

in linearized form and with homogeneous boundary conditions on $\Omega = [0, T]$ or $\Omega = [0, 1]^2$ that we put into the spaces we work on. We assume f and g

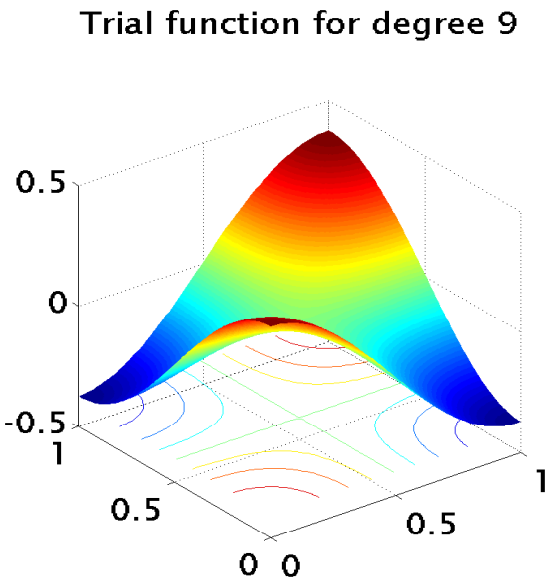

Figure 5: Approximate solution for degree 9

to be smooth, and g should appropriately be nonnegative to ensure ellipticity, according to the boundary conditions. We have enough smoothness to assume that the exact solution u^* is in $H_0^{m+2}(\Omega) := H^{m+2}(\Omega) \cap H_0^1(\Omega)$ for some rather large m while f is at least in $H^m(\Omega)$. Using standard norm notation for Sobolev scales of spaces, and with generic constants that do not depend on the trial space or the test discretization, we have the a priori inequalities (cf. (2.147) and (2.151) in [2])

$$\|u\|_2 \leq C\|Fu\|_0 \text{ for all } u \in H_0^2(\Omega),$$

$$\|u\|_2 \leq C\|\Delta u\|_0 \text{ for all } u \in H_0^2(\Omega),$$

and we have a well–posed problem in the sense of 2.1 using $\mathcal{U} = H_0^2(\Omega)$ and $\mathcal{V} = L^2(\Omega)$.

Our trial spaces $\mathcal{U}_r$ consist of algebraic univariate or trigonometric bivariate polynomials of some degree up to N in each variable satisfying the boundary conditions exactly, and we should rather use $\mathcal{U}_N$ instead of $\mathcal{U}_r$ now. Our exact solutions are so smooth that we can expect errors $\epsilon(N, u^*) \leq CN^{-p}$ for arbitrarily large p even if we measure the error in $\mathcal{U} = H_0^2(\Omega)$.

Our testing maps are based on point evaluations on X_h with either $M+1$ equidistant points in $\Omega = [0, T]$ or $(M+1)^2$ gridded points in $[0, 1]^2$, and we can up to a factor of $\sqrt{2}$ use $h = 1/M$ in what follows. The discrete test spaces $\mathcal{V}_s = \mathcal{V}_M$ are normed with the root–mean–square norm, i.e. $\|.\|_{\mathcal{V}_M} = h^{n/2}\|.\|_{\ell_2} = \|.\|_{RMSE}, n = 1, 2$, up to a constant. Our goal is to find a sufficient

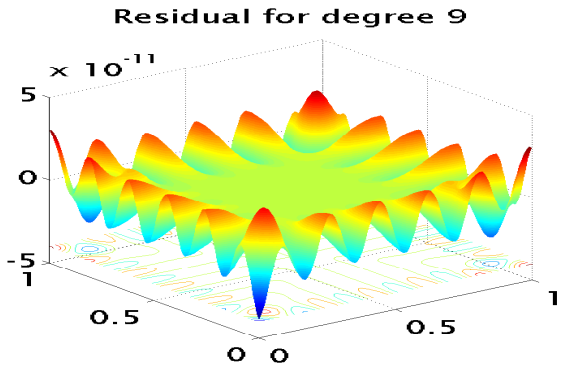

Figure 6: Residual for degree 9

condition on N and M to ensure uniform stability. Clearly, the test maps T_s are defined only on sufficiently smooth functions. The subspace $\tilde{\mathcal{V}} \subseteq L_2(\Omega)$ can be chosen to be $H^2(\Omega)$ in order to make point evaluation possible

In $H^2(\Omega)$ with $n = \dim \Omega \leq 2$ there is a sampling inequality [8]

$$\|v\|_0 \leq C \left(h^2 \|v\|_2 + h^{n/2} \|v\|_{\ell_2, X_h} \right) \text{ for all } v \in H^2(\Omega)$$

for all finite sets $X_h \subset \Omega$ with *fill distance*

$$h := \sup_{x \in \Omega} \inf_{y \in X_h} \|x - y\|_2 \leq h_0 > 0.$$

We apply this to $v = Fu$ to get

$$\|Fu\|_0 \leq C \left(h^2 \|Fu\|_2 + h^{n/2} \|Fu\|_{\ell_2, X_h} \right).$$

Unfortunately, F does not map polynomials into polynomials, and thus we need a little detour via Δ with

$$\begin{array}{rcl} \|Fu\|_2 & = & \|\Delta u - gu\|_2 \\ & \leq & \|\Delta u\|_2 + \|gu\|_2 \\ & \leq & \|\Delta u\|_2 + C\|u\|_2 \\ & \leq & \|\Delta u\|_2 + C\|\Delta u\|_0 \\ & \leq & C\|\Delta u\|_2 \end{array}$$

and, with the inverse inequality

$$\|p\|_2 \leq C_I(N) \|p\|_0 \text{ for all } p \in \mathcal{U}_N$$

and, by ellipticity,

$$\|p\|_0 \leq \|p\|_2 \leq C\|Fp\|_0$$

also

$$\begin{aligned}\|Fp\|_2 &\leq C\|\Delta p\|_2\\ &\leq CC_I(N)\|\Delta p\|_0\\ &\leq CC_I(N)\left(\|Fp\|_0+\|gp\|_0\right)\\ &\leq CC_I(N)\left(\|Fp\|_0+C\|p\|_0\right)\\ &\leq CC_I(N)\left(\|Fp\|_0+C\|Fp\|_0\right)\\ &\leq CC_I(N)\|Fp\|_0.\end{aligned}$$

Then we combine everything into

$$\begin{aligned}\|Fp\|_0 &\leq C_1\left(h^2\|Fp\|_2+h^{n/2}\|Fp\|_{\ell_2,X_h}\right)\\ &\leq C_1\left(h^2C_2C_I(N)\|Fp\|_0+h^{n/2}\|Fp\|_{\ell_2,X_h}\right)\end{aligned}$$

where we now have named the constants, and we impose the condition

$$h^2C_1C_2C_I(N)<1/2 \tag{40}$$

to get

$$\|Fp\|_0\leq 2h^{n/2}\|Fp\|_{\ell_2,X_h}=2\|Fp\|_{RMSE}.$$

Thus any trial/test strategy $s(r)$ satisfying (40) will lead to

$$C(r,s(r))\leq 2.$$

Using Bernstein–Markov inequalities ([9], p.97) the 1D case with algebraic polynomials has $C_I(N)=N^2(N-1)^2T^2$, while the 2D case with trigonometric polynomials has $C_I(N)=N^2$. Thus the sufficient condition (40) for stability is satisfied for testing on $M+1$ equidistant points for $M=\mathcal{O}(N^2)$ in the 1D case and $M=\mathcal{O}(N)$ in the 2D case.

For error bounds including derivatives, we can assume

$$\begin{aligned}\|u^*-u_r^*\|_\infty &\leq CN^{-p},\\ \|\Delta u^*-\Delta u_r^*\|_\infty &\leq CN^{-p+2},\\ \|Fu^*-Fu_r^*\|_\infty &\leq CN^{-p+2}.\end{aligned}$$

For the error bound (11) we need to evaluate

$$\begin{aligned}\delta^2(r,s,u^*) &= \|T_s(Fu_r^*-Fu^*)\|_{RMSE}^2\\ &= h^d\sum_{x_j\in X_h}(Fu_r^*-Fu^*)^2(x_j)\\ &\leq Ch^d(M+1)^dN^{-2p+4}\\ &\leq CN^{-2p+4}\end{aligned}$$

and since $C(r,s(r))\leq 2$ we get a convergence rate of N^{-p+2} provided that we use enough points for testing.

If the error of $u_r^*-u^*$ converges geometrically like some q^N for some $0<q<1$ (this is observed, but a proof needs analyticity of the solution and a nontrivial application of a Bernstein–type theorem for polynomial approximation), the analogous argument ends up with a convergence like N^2q^N to zero.

To guarantee that uniform stability of the linearized problem carries over to the nonlinear problem via Theorem 4 and Corollary 1, we have to prove that in

$$\|T_s F'(v_r) - T_s F'(u_r)\|_{\mathcal{L}(\mathcal{U},\mathcal{V}_s)} \leq C''(r,s)\|u_r - v_r\|_{\mathcal{U}}$$

for all u_r, v_r in a neighborhood of u_r^* in $\mathcal{U}_r \subset \mathcal{U}$, the constant is uniformly bounded for our trial/test strategy $s(r)$ as given above. We restrict ourselves to the 2D example, because the 1D example is similar and easier.

As a warm–up, let us check (19). This is, for u, v, $w \in H^2$, and by the Cauchy-Schwarz inequality

$$\begin{aligned} \|(F'(u) - F'(v))w\|_{L_2} &= \|\lambda(\sin(u) - \sin(v))w\|_{L_2} \\ &\leq |\lambda|\|u - v\|_{L_2}\|w\|_{L_2} \end{aligned}$$

and thus $C'' = |\lambda|$ can be taken in (19) even if we take only L_2 norms.

The discretized version of this on a finite test set X_s is

$$\begin{aligned} \|(F'(u) - F'(v))w\|_{\ell_2(X_s)} &= \|\lambda(\sin(u) - \sin(v))w\|_{\ell_2(X_s)} \\ &\leq |\lambda|\|u - v\|_{\ell_2(X_s)}\|w\|_{L_\infty} \end{aligned}$$

but this does not help directly since we do not have $\|w\|_{L_2}$ or $\|w\|_{H^2}$ in the right–hand side. But if we use Sobolev embedding in the form

$$\|u\|_{L_\infty} \leq c_e\|u\|_{H^2}$$

we get

$$\|u\|_{RMSE} \leq c_e\|u\|_{H^2} \text{ for all } u \in H^2.$$

Thus

$$\begin{aligned} \|(F'(u) - F'(v))w\|_{RMSE} &\leq |\lambda|\|u - v\|_{RMSE}\|w\|_{L_\infty} \\ &\leq c_e^2|\lambda|\|u - v\|_{H^2}\|w\|_{H^2} \end{aligned}$$

and

$$\|T_s F'(u) - T_s F'(v)\|_{\mathcal{L}(\mathcal{U},\mathcal{V}_s)} \leq c_e^2|\lambda|\|u - v\|_{H^2}$$

and we see that $C''(r,s)$ can be bounded uniformly by $c_e^2|\lambda|$.

On an exact solution u^* with values in $[-\pi/2, \pi/2]$ and for positive λ we have the elliptic operator

$$F'(u^*)w = \Delta w - \lambda\cos(u^*)w$$

which is invertible on L_2, i.e.

$$\|w\|_{L_2} \leq \|w\|_2 \leq C\|F'(u^*)w\|_{L_2}$$

and thus the linearization is continuously invertible.

Thus we have verified all requirements of Section 4 on linearization.

References

[1] R. Schaback, Unsymmetric Meshless Methods for Operator Equations, Numer. Math. 114 (2010) 629–651.

[2] K. Böhmer, Numerical Methods for Nonlinear Elliptic Differential Equations, Oxford University Press, 2010.

[3] C. Rieger, B. Zwicknagl, R. Schaback, Sampling and Stability, in: M. Dæhlen, M. Floater, T. Lyche, J.-L. Merrien, K. Mørken, L. Schumaker (Eds.), Mathematical Methods for Curves and Surfaces, vol. 5862 of *Lecture Notes in Computer Science*, 347–369, 2010.

[4] Y. C. Hon, R. Schaback, On unsymmetric collocation by radial basis functions, Appl. Math. Comput. 119 (2-3) (2001) 177–186, ISSN 0096-3003.

[5] T. Belytschko, Y. Krongauz, D. Organ, M. Fleming, P. Krysl, Meshless methods: an overview and recent developments, Computer Methods in Applied Mechanics and Engineering, special issue 139 (1996) 3–47.

[6] S. N. Atluri, The meshless method (MLPG) for domain and BIE discretizations, Tech Science Press, Encino, CA, 2005.

[7] A. Steindl, TU Wien, private communication, 2010.

[8] W. R. Madych, An estimate for multivariate interpolation. II, J. Approx. Theory 142 (2) (2006) 116–128, ISSN 0021-9045.

[9] R. DeVore, G. Lorentz, Constructive Approximation, vol. 303 of *Grundlehren der mathematischen Wissenschaften*, 1993.