Lecture Notes for the Course Numerical Methods for Partial Differential Equations

Walter Zulehner
Institute for Computational Mathematics
Johannes Kepler University Linz

Winter Semester 2005/06

Contents

Chapter 1

Elliptic Differential Equations

1.1 Boundary Value Problems for Second-order Ordinary Differential Equations

Classical Formulation (an example):

Find a function $u : [0,1] \longrightarrow \mathbb{R}$ such that the differential equation

$$-(a(x)u'(x))' + b(x)u'(x) + c(x)u(x) = f(x) \quad x \in (0,1)$$

or, in short,

$$Lu(x) = f(x) \quad x \in (0,1)$$

with the linear differential operator L, given by

$$Lu(x) = -(a(x)u'(x))' + b(x)u'(x) + c(x)u(x),$$

and the boundary conditions

$$u(0) = g_0, \tag{1.1}$$
$$a(1)u'(1) = g_1 \tag{1.2}$$

are satisfied, for given data a, b, c, f, g_0 and g_1.

The boundary condition (1.1) is called Dirichlet boundary condition (or boundary condition of the first kind), the boundary condition (1.2) is called Neumann boundary condition (or boundary condition of the second kind).

All expression are well-defined, for example, for solutions $u \in C^2(0,1) \cap C^1(0,1] \cap C[0,1)$ and data

$$a \in C^1(0,1) \cap C(0,1], \quad b, c, f \in C(0,1), \quad g_0, g_1 \in \mathbb{R}.$$

The function u is called a classical solution.

Special case (model problem):

For $a(x) \equiv 1$, $b(x) \equiv 0$, $c(x) \equiv 0$ one obtains:

$$\begin{aligned} -u''(x) &= f(x) \quad x \in (0,1), \\ u(0) &= g_0, \\ u'(1) &= g_1. \end{aligned}$$

Variational formulation:

Let $v : [0,1] \longrightarrow \mathbb{R}$ be a so-called test function. Under appropriate differentiability and integrability conditions the following steps can be performed:

- Multiplication with a test function v and integration over the interval:

$$\int_0^1 \left[-(a(x)u'(x))' + b(x)u'(x) + c(x)u(x)\right] v(x)\ dx = \int_0^1 f(x)v(x)\ dx$$

- Integration by parts for the principal part:

$$\begin{aligned} &-a(x)u'(x)v(x)\Big|_0^1 + \int_0^1 a(x)u'(x)v'(x)\ dx \\ &\quad + \int_0^1 \left[b(x)u'(x)v(x) + c(x)u(x)v(x)\right]\ dx = \int_0^1 f(x)v(x)\ dx \end{aligned}$$

- Incorporating the boundary conditions for the unknown function u and using the boundary condition $v(0) = 0$ for the test function v :

$$\begin{aligned} &-g_1 v(1) + \int_0^1 a(x)u'(x)v'(x)\ dx \\ &\quad + \int_0^1 \left[b(x)u'(x)v(x) + c(x)u(x)v(x)\right]\ dx = \int_0^1 f(x)v(x)\ dx \end{aligned}$$

  General strategy for incorporating the boundary conditions: Two types of boundary conditions are distinguished: Essential and natural boundary conditions.

  Essential boundary conditions for the solution u are explicitly prescribed, they cause a corresponding homogeneous boundary condition for the test functions v. Here: Dirichlet boundary conditions are essential boundary conditions: $u(0) = g_0$ and $v(0) = 0$ are explicitly prescribed.

  Natural boundary conditions for the solution u are incorporated in the variational equation. The test functions v inherit no boundary condition on such points. Here: Neumann boundary conditions are natural boundary conditions: $a(1)u'(1) = g_1$ is incorporated in the variational equation.

So we obtain the following variational problem: Find a function $u : [0,1] \longrightarrow \mathbb{R}$ with $u(0) = g_0$, such that

$$\int_0^1 [a(x)u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x)]\ dx = \int_0^1 f(x)v(x)\ dx + g_1 v(1)$$

for all test functions $v : [0,1] \longrightarrow \mathbb{R}$ with $v(0) = 0$.

Function spaces

Derivatives occur only behind an integral sign: Weak derivative

$$\int_0^1 u'(x)\varphi(x)\ dx = -\int_0^1 u(x)\varphi'(x)\ dx \quad \text{for all } \varphi \in C_0^\infty(0,1).$$

The existence of the integral expressions (for bounded measurable coefficients $a, b, c \in L^\infty(0,1)$ and square integrable right hand sides $f \in L^2(0,1)$) are guaranteed if

$$u, u', v, v' \in L^2(0,1).$$

This suggests to use the Sobolev space $H^1(0,1)$ as working space:

$$H^1(0,1) = \{v \in L^2(0,1) : v' \in L^2(0,1)\}.$$

Formulation of the essential boundary conditions:

Trace operator:

Lemma 1.1. *There is a constant $C > 0$ with*

$$|v(0)| \le C\,\|v\|_1 \quad \textit{for all } v \in C^1[0,1].$$

Proof. By integrating the identity

$$v(0) = v(x) - \int_0^x v'(y)\ dy$$

one obtains

$$v(0) = \int_0^1 v(x)\ dx - \int_0^1 \int_0^x v'(y)\ dydx = \int_0^1 v(x)\ dx - \int_0^1 (1-y)v'(y)\ dy.$$

Cauchy's inequality implies

$$\begin{aligned} |v(0)| &\le \left(\int_0^1 v(x)^2\ dx\right)^{1/2} + \left(\int_0^1 (1-y)^2\ dy\right)^{1/2} \left(\int_0^1 v'(y)^2\ dy\right)^{1/2} \\ &= \|v\|_0 + \frac{1}{\sqrt{3}}|v|_1 \le \frac{2}{\sqrt{3}}\,\|v\|_1. \end{aligned}$$

□

Hence the so-called trace operator $\gamma_0 : C^1[0,1] \longrightarrow \mathbb{R}$ with $\gamma_0 v = v(0)$ is linear and continuous (bounded) with respect to the H^1-norm. Since $C^1[0,1]$ is dense in $H^1(0,1)$, there is a unique continuous extension of γ_0 on $H^1(0,1)$. In this sense the expression $v(0)(= \gamma_0 v)$ is well-defined for all $v \in H^1(0,1)$.

With

$$V = H^1(0,1), \quad V_0 = \{v \in V | v(0) = 0\}, \quad V_g = \{v \in V | v(0) = g_0\}$$

we obtain the final formulation of the variational problem: Find $u \in V_g$, such that

$$a(u,v) = \langle F, v \rangle \quad \text{for all } v \in V_0$$

where

$$\begin{aligned} a(w,v) &= \int_0^1 [a(x)u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x)]\ dx, \\ \langle F, v \rangle &= \int_0^1 f(x)v(x)\ dx + g_1 v(1). \end{aligned}$$

All expressions are well-defined for data

$$a, b, c \in L^\infty(0,1),\ f \in L^2(0,1),\ g_0, g_1 \in \mathbb{R}.$$

Solutions u of this variational problem are called weak solutions.

Warning:

$$\text{classical solution} \overset{\text{i. A.}}{\not\Rightarrow} \text{weak solution}$$

A classical (smooth) solution is also a weak solution only if the correct integrability conditions are satisfied.

Next we discuss the opposite question: Is a weak solution also a classical solution?

Let $u \in V_g$ be a (weak) solution of the variational problem

$$\int_0^1 [a(x)u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x)]\ dx = \int_0^1 f(x)v(x)\ dx + g_1 v(1)$$

for all $v \in V_0$. Under the assumption that b, c, f are continuous, a is continuously differentiable and u is twice continuously differentiable, one obtains by integration by parts:

$$\begin{aligned} &a(x)u'(x)v(x)\Big|_0^1 - \int_0^1 (a(x)u'(x))'v(x)\ dx \\ &\quad + \int_0^1 [b(x)u'(x)v(x) + c(x)u(x)v(x)]\ dx = \int_0^1 f(x)v(x)\ dx + g_1 v(1). \end{aligned}$$

So

$$\begin{aligned} &a(1)u'(1)v(1) + \int_0^1 [-(a(x)u'(x))' + b(x)u'(x) + c(x)u(x)]v(x)\ dx \qquad (1.3) \\ &= \int_0^1 f(x)v(x)\ dx + g_1 v(1). \end{aligned}$$

If we choose test functions $v \in C_0^\infty(0,1)$ (i.e.: $v(0) = v(1) = 0$), then it follows

$$\int_0^1 [-(a(x)u'(x))' + b(x)u'(x) + c(x)u(x)]v(x)\ dx = \int_0^1 f(x)v(x)\ dx.$$

Since $C_0^\infty(0,1)$ is dense in $C[0,1]$, it easily follows

$$-(a(x)u'(x))' + b(x)u'(x) + c(x)u(x) = f(x) \quad \text{for all } x \in (0,1).$$

Then (1.3) implies

$$a(1)u'(1)v(1) = g_1 v(1) \quad \text{for all } v \in V_0.$$

Since $v(1)$ can be chosen arbitrarily, we obtain the (natural) boundary condition:

$$a(1)u'(1) = g_1.$$

Warning:

$$\text{weak solution} \overset{\text{i. A.}}{\not\Rightarrow} \text{classical solution}$$

A weak solution is a classical solution, only if the correct differentiability conditions are satisfied.

Example: Let $\bar{x} \in (0,1)$. Assume that a is continuously differentiable on $[0,\bar{x}]$ and on $[\bar{x},1]$ and u is twice continuously differentiable on $[0,\bar{x}]$ and on $[\bar{x},1]$. Then one obtains by integration by parts on each of the two sub-intervals:

$$\begin{aligned}
\int_0^1 a(x)u'(x)v'(x)\ dx &= \int_0^{\bar{x}} a(x)u'(x)v'(x)\ dx + \int_{\bar{x}}^1 a(x)u'(x)v'(x)\ dx \\
&= a(x)u'(x)v(x)\Big|_0^{\bar{x}} - \int_0^{\bar{x}} (a(x)u'(x))'\, v(x)\ dx \\
&\quad + a(x)u'(x)v(x)\Big|_{\bar{x}}^1 - \int_{\bar{x}}^1 (a(x)u'(x))'\, v(x)\ dx \\
&= a(1)u'(1)v(1) + [a(\bar{x}+)u'(\bar{x}+) - a(\bar{x}-)u'(\bar{x}-)]v(\bar{x}) \\
&\quad - \int_0^{\bar{x}} (a(x)u'(x))'\, v(x)\ dx - \int_{\bar{x}}^1 (a(x)u'(x))'\, v(x)\ dx
\end{aligned}$$

Then, from the variational equation it easily follows that u satisfies the differential equation on the sub-intervals $(0,\bar{x})$ and $(\bar{x},1)$, and one obtains the additional interface condition:

$$a(\bar{x}+)u'(\bar{x}+) = a(\bar{x}-)u'(\bar{x}-).$$

In order to have $u \in H^1(0,1)$ it is necessary that:

$$u(\bar{x}+) = u(\bar{x}-).$$

So an additional condition follows from the variational formulation, which shows that, in general, the solution is not necessarily continuously differentiable at the point $\bar{x}$.

1.2 The Lax-Milgram Theorem

Properties of the variational problem:

- Linearity: $\langle F, . \rangle$ is linear, $a(.,.)$ is bilinear.
- V-boundedness (continuity) of F:

  $$|\langle F, v\rangle| = \left| \int_0^1 f(x)v(x)\ dx + g_1\, v(1) \right| \leq \|f\|_0 \|v\|_0 + g_1 C \|v\|_1 \leq (\|f\|_0 + g_1 C)\, \|v\|_1.$$

  Observe, that Lemma 1.1 is also valid in in $H^1(0,1)$ (Proof by a closure argument).
- V-boundedness (continuity) of a:

  $$\begin{aligned} |a(w,v)| &= |\int_0^1 [a(x)w'(x)v'(x) + b(x)w'(x)v(x) + c(x)w(x)v(x)]\ dx| \\ &\leq \|a\|_{L^\infty} \|w'\|_0 \|v'\|_0 + \|b\|_{L^\infty} \|w'\|_0 \|v\|_0 + \|c\|_{L^\infty} \|w\|_0 \|v\|_0 \\ &\leq (\|a\|_{L^\infty} + \|b\|_{L^\infty} + \|c\|_{L^\infty})\, \|w\|_1 \|v\|_1 \\ &= \mu_2\, \|w\|_1 \|v\|_1. \end{aligned}$$

- V_0-ellipticity of a:

  Assumptions:

  1. $a(x) \geq a_0 > 0$ for almost all $x \in (0,1)$,
  2. $b(x) \equiv 0$,
  3. $c(x) \geq 0$ for almost all $x \in (0,1)$.

  We have:

  Lemma 1.2 (Friedrichs inequality). *There is a constant $C > 0$ with*

  $$\int_0^1 v(x)^2\ dx \leq c_F^2 \int_0^1 v'(x)^2\ dx \quad \textit{for all } v \in V_0 = \{v \in H^1(0,1) : v(0) = 0\}$$

  Proof. Let $v \in C^1[0,1]$ with $v(0) = 0$. From

  $$v(x) = \int_0^x v'(y)\ dy$$

  one obtains

  $$|v(x)|^2 \leq x \int_0^x |v'(y)|^2\ dy \leq \int_0^1 |v'(y)|^2.$$

  by using the Cauchy inequality. By integrating it follows that

  $$\int_0^1 |v(x)|^2 \leq \int_0^1 |v'(y)|^2\ dy.$$

$C^1[0,1] \cap V_0$ is dense in V_0, all involved (semi-) norms are continuous on V_0. Hence the inequality also holds on the closure of $C^1[0,1] \cap V_0$, which is V_0. □

Remark: Friedrichs inequality is not satisfied in $V = H^1(0,1)$.

Hence

$$\|v\|_0 \leq c_F \, |v|_1,$$

and, therefore,

$$|v|_1^2 \leq \|v\|_1^2 = \|v\|_0^2 + |v|_1^2 \leq (1 + c_F^2)|v|_1^2.$$

So $|v|_1$ and $\|v\|_1$ are equivalent norms on V_0. It follows:

$$a(v,v) \geq a_0 \, |v|_1^2 \geq \frac{a_0}{1 + c_F^2} \|v\|_1^2 = \mu_1 \, \|v\|_1^2.$$

Homogenization:

Find $g \in V = H^1(0,1)$ with $\gamma_0 g = g_0$. For example: $g(x) \equiv g_0$. Then we have

$$V_g = g + V_0 \quad \text{with } g \in V.$$

Using the ansatz $u = g + w$ we obtain a variational problem for $w \in V_0$: Find $w \in V_0$, such that

$$a(w,v) = \langle F, v \rangle - a(g,v) \equiv \langle \hat{F}, v \rangle \quad \text{for all } v \in V_0. \tag{1.4}$$

$\hat{F}$ is linear and bounded:

$$|\langle \hat{F}, v \rangle| \leq |\langle F, v \rangle| + |a(g,v)| \leq (\|F\| + \mu_2 \|g\|) \, \|v\|$$

We assume that V_0 is always a closed subspace of the Hilbert space V. This guarantees that V_0 is also a Hilbert space.

This discussion shows that it suffices to consider the homogenized variational problem. Therefore, we will often discuss in the following only the case

$$g = 0, \quad V = V_g = V_0$$

without loss of generality.

Formulation of operator equation

Because of the continuity of the bilinear form a the linear operator $A : V \longrightarrow V^*$, given by

$$\langle Aw, v \rangle = a(w,v) \quad \text{for all } w, v \in V,$$

is well-defined. Then the variational problem (1.4) can be written as linear operator equation

$$Au = F.$$

Riesz Representation Theorem

Let F be a linear continuous functional on the Hilbert space V. Consider the following variational problem:

Find $u \in V$ with

$$(u, v) = \langle F, v \rangle \quad \text{for all } v \in V. \tag{1.5}$$

This problem is equivalent to the optimization problem:

Find $u \in V$ with

$$J(u) = \min_{v \in V} J(v)$$

where

$$J(v) = \frac{1}{2}(v, v) - \langle F, v \rangle.$$

Proof.

$$\begin{aligned}
& J(u) = \min_{v \in V} J(v) \\
\iff \quad & J(u) \le J(u + tw) \quad \text{for all } w \in V,\ t \in [0, 1] \\
\iff \quad & J(u) \le J(u) + t\left[(u, w) - \langle F, w \rangle\right] + \frac{t^2}{2}(w, w) \quad \text{for all } w \in V,\ t \in [0, 1] \\
\iff \quad & (u, w) - \langle F, w \rangle + \frac{t}{2}(w, w) \ge 0 \quad \text{for all } w \in V,\ t \in [0, 1] \\
\iff \quad & (u, w) - \langle F, w \rangle \ge 0 \quad \text{for all } w \in V
\end{aligned}$$

Choose $w = v$ and $w = -v$ for arbitrary $v \in V$. □

Theorem 1.1 (Riesz Representation Theorem). *There is a unique solution of the variational problem (1.5) and we have $\|u\| = \|F\|$.*

Proof. The functional J is bounded from below:

$$J(v) \ge \frac{1}{2}\|v\|^2 - \|F\|\|v\| \ge -\frac{1}{2}\|F\|^2.$$

Therefore, there is a sequence (u_n) in V with

$$J(u_n) \to \inf_{v \in V} J(v) > -\infty$$

The sequence (u_n) satisfies the Cauchy criterion:

$$\begin{aligned}
\|u_n - u_m\|^2 &= 2\|u_n\|^2 + 2\|u_m\|^2 - \|u_n + u_m\|^2 \\
&= 4J(u_n) + 4J(u_m) - 8J\left(\frac{u_n + u_m}{2}\right) \\
&\le 4J(u_n) + 4J(u_m) - 8\inf_{v \in K} J(v) \to 0
\end{aligned}$$

So (u_n) converge towards a limit value $u \in V$. Because of the continuity of $J(v)$ it follows:

$$J(u) = \lim_{n\to\infty} J(u_n) = \inf_{v\in V} J(v).$$

We have

$$\|u\| = \sup_{v\in V} \frac{(u,v)}{\|v\|} = \sup_{v\in V} \frac{\langle F, v\rangle}{\|v\|} = \|F\|.$$

□

Hence the mapping $\mathcal{J}\colon V^* \longrightarrow V$, given by $\mathcal{J}F = u$, is a isometric isomorphism (Riesz isomorphism).

By using the Riesz Representation Theorem we can reformulate the variational problem Find $u \in V$, such that

$$a(u,v) = \langle F, v\rangle \quad \text{for all } v \in V$$

as a operator equation in the Hilbert space V:

Let $\widetilde{A} : V \longrightarrow V$ be given by

$$(\widetilde{A}w, v) = a(w,v) \quad \text{for all } w, v \in V$$

and $\widetilde{f} \in V$ be given by

$$(\widetilde{f}, v) = \langle F, v\rangle \quad \text{for all } v \in V.$$

Then the variational problem (1.4) can be written as operator equation

$$\widetilde{A}u = \widetilde{f}.$$

It is easy to see that $\widetilde{A} = \mathcal{J}A$ and $\widetilde{f} = \mathcal{J}F$.

Theorem 1.2 (Lax-Milgram Theorem). *Let V be a Hilbert space, $F : V \longrightarrow \mathbb{R}$ a (V-)bounded linear functional ($F \in V^*$) and $a : V \times V \longrightarrow \mathbb{R}$ be a bilinear form with the following properties:*

1. *a is V-elliptic, i.e.: there is a constant $\mu_1 > 0$ with*

$$\mu_1 \|v\|^2 \le a(v,v) \quad \textit{for all } v \in V,$$

2. *a is V-bounded, i.e.: there is a constant $\mu_2 > 0$ with*

$$|a(w,v)| \le \mu_2 \|w\|\|v\| \quad \textit{for all } w, v \in V.$$

Then there exists a unique solution u of the variational problem and we have:

$$\frac{1}{\mu_2}\|F\| \le \|u\| \le \frac{1}{\mu_1}\|F\|.$$

Proof. The linear problem

$$\widetilde{A}u = \widetilde{f}$$

can be written in fixed point form

$$u = u - \tau\,(\widetilde{A}u - \widetilde{f}) \equiv K_\tau u + g_\tau.$$

In the following it will be shown that K_τ is contractive for an appropriate choice of the parameter τ:

$$\begin{aligned}
\|K_\tau v\|^2 &= (K_\tau v, K_\tau v) = ([I - \tau\,\widetilde{A}]v, [I - \tau\,\widetilde{A}]v) \\
&= (v,v) - 2\tau\,(\widetilde{A}v, v) + \tau^2\,(\widetilde{A}v, \widetilde{A}v) \\
&= (v,v) - 2\tau\,a(v,v) + \tau^2\,\|Av\|^2 \\
&\leq (1 - 2\mu_1\tau + \mu_2^2\tau^2)\|v\|^2
\end{aligned}$$

So

$$\|K_\tau v\| \leq q(\tau)\,\|v\| \quad \text{with} \quad q(\tau) = \sqrt{1 - 2\mu_1\tau + \mu_2^2\tau^2}.$$

We have

$$q(\tau) < 1 \Leftrightarrow 0 < \tau < \frac{2\mu_1}{\mu_2^2}.$$

$q(\tau)$ has a minimum for $\tau_{\text{opt}} = \mu_1/\mu_2^2$ and it follows:

$$q_{\text{opt}} = q(\tau_{\text{opt}}) = \sqrt{1 - \left(\frac{\mu_1}{\mu_2}\right)^2}$$

The existence and uniqueness follow from the Banach fixed point theorem.

The estimates follow from:

$$\mu_1\,\|u\|^2 \leq a(u,u) = \langle F, u\rangle \leq \|F\|\|u\|$$

and

$$\|F\| = \sup_{v\neq 0} \frac{\langle F, v\rangle}{\|v\|} = \sup_{v\neq 0} \frac{a(u,v)}{\|v\|} \leq \sup_{v\neq 0} \frac{\mu_2\,\|u\|\|v\|}{\|v\|} = \mu_2\|u\|.$$

□

Consequences:

The Banach Fixed Point Theorem does not only yield the existence of a unique solution but also the construction of a sequence of approximations (u_n), given by

$$u_{n+1} = u_n - \tau(\widetilde{A}u_n - \widetilde{f}),$$

which converges towards the solution. Because of

$$(u_{n+1}, v) = (u_n, v) - \tau[(\widetilde{A}u_n, v) - (\widetilde{f}, v)] = (u_n, v) - \tau[a(u_n, v) - \langle F, v\rangle]$$

for all $v \in V$, one obtains a sequence of variational problems

$$(u_{n+1}, v) = (u_n, v) + \tau[\langle F, v\rangle - a(u_n, v)] \quad \text{for all } v \in V$$

for determining the approximations $u_{n+1} \in V$. From the Banach Fixed Point Theorem the following error estimates can be derived:

- q-linear convergence:
$$\|u_{n+1} - u\| \leq q \, \|u_n - u\|.$$
- r-linear convergence:
$$\|u_n - u\| \leq q^n \, \|u_0 - u\|$$
- constructive a priori estimate:
$$\|u_n - u\| \leq \frac{q^n}{1-q} \, \|u_1 - u_0\|$$
- constructive a posteriori estimate:
$$\|u_n - u\| \leq \frac{q}{1-q} \, \|u_n - u_{n-1}\|$$

with $q = q(\tau)$.

The Lax-Milgram Theorem does not need the symmetry of the bilinear form a. If a is symmetric, the following theorem holds:

Theorem 1.3. *Assume that*

1. *a is symmetric, i.e.*
$$a(w, v) = a(v, w) \quad \textit{for all } w, v \in V,$$
2. *a is non-negative, i.e.*
$$a(v, v) \geq 0 \quad \textit{for all } v \in V.$$

Then $u \in V_g$ is a solution of the variational problem if and only if $u \in V_g$ minimizes the so-called Ritz energy functional J, given by

$$J(v) = \frac{1}{2} \, a(v, v) - \langle F, v\rangle,$$

with respect to the set V_g:

$$J(u) = \min_{w \in V_g} J(w).$$

For the proof, see the discussion of the Riesz Representation Theorem.

Remark: The proof of the last theorem clarifies the role of the test functions v as admissible directions: For $u \in V_g$, we need $v = u + tw$ to stay in the linear manifold $V_g = g + V_0$ for all $t > 0$. So $w \in V_0$.

For symmetric bilinear forms the estimates can be improved: Es gilt:

Theorem 1.4. *Under the assumptions of Theorem 1.2 and the additional condition that the bilinear form a is symmetric we have:*

$$\|K_\tau\| \leq q(\tau) = \max(|1 - \mu_1 \tau|, |1 - \mu_2 \tau|)$$

and

$$q(\tau) < 1 \Leftrightarrow 0 < \tau < \frac{2}{\mu_2}.$$

$q(\tau)$ *has its minimum at* $\tau_{opt} = 2/(\mu_1 + \mu_2)$ *and we have:*

$$q_{opt} = q(\tau_{opt}) = \frac{\mu_2 - \mu_1}{\mu_2 + \mu_1}.$$

Proof. Since K_τ is symmetric, we have:

$$\|K_\tau\| = \sup_{0 \neq v \in V} \frac{|(K_\tau v, v)|}{(v, v)}$$

The estimates easily follow from:

$$(1 - \tau\,\mu_2)\,(v, v) \leq (K_\tau v, v) = (v, v) - \tau\,(\widetilde{A} v, v) \leq (1 - \tau\,\mu_1)\,(v, v)$$

□

1.3 Boundary Value Problems for Second-order Partial Differential Equation

Classical formulation:

Let $\Omega \subset \mathbb{R}^d$ and $\Gamma = \partial\Omega = \Gamma_D \cup \Gamma_N$. Find $u : \overline{\Omega} \longrightarrow \mathbb{R}$, such that the differential equation

$$-\sum_{i,j=1}^{d} \frac{\partial}{\partial x_i} \left(a_{ij}(x)\, \frac{\partial u}{\partial x_j}(x) \right) + \sum_{i=1}^{d} b_i(x) \frac{\partial u}{\partial x_i}(x) + c(x)u(x) \;=\; f(x) \quad x \in \Omega$$

and the boundary conditions

$$\begin{aligned} u(x) &= g_D(x) \quad x \in \Gamma_D \\ \sum_{i=1}^{d} a_{ij}(x) n_i(x) \frac{\partial u}{\partial x_j}(x) &= g_N(x) \quad x \in \Gamma_N \end{aligned}$$

are satisfied. In short:

$$\begin{aligned} -\operatorname{div}(A(x)\operatorname{grad} u(x)) + b(x)\cdot\operatorname{grad} u(x) + c(x)u(x) &= f(x) \quad x\in\Omega, \\ u(x) &= g_D(x) \quad x\in\Gamma_D, \\ A(x)\operatorname{grad} u(x)\cdot n(x) &= g_N(x) \quad x\in\Gamma_N. \end{aligned}$$

with

$$A(x) = (a_{ij}(x))_{i,j=1,\ldots,d}, \quad b(x) = (b_i(x))_{i=1,\ldots,d}.$$

Special case Poisson equation:

$A(x) = I$, $b(x) \equiv 0$, $c(x) \equiv 0$ (Laplace equation: $f(x) \equiv 0$):

$$\begin{aligned} -\Delta u(x) &= f(x) \quad x\in\Omega, \\ u(x) &= g_D(x) \quad x\in\Gamma_D, \\ \frac{\partial u}{\partial n}(x) &= g_N(x) \quad x\in\Gamma_N. \end{aligned}$$

Variational formulation:

Gauss Theorem:

$$\int_\Omega \operatorname{div} w \; dx = \int_\Gamma w\cdot n \; ds$$

in the form

$$\int_\Omega \frac{\partial w}{\partial x_i} \; dx = \int_\Gamma w n_i \; ds$$

yields the identity

$$\int_\Omega \frac{\partial u}{\partial x_i} v \; dx = \int_\Gamma u v n_i \; ds - \int_\Omega u \frac{\partial v}{\partial x_i} \; dx$$

(integration by parts).

Analogous to the one-dimensional problem one obtains the following variational problem: Find $u \in V_g$, such that

$$a(u,v) = \langle F, v\rangle \quad v\in V_0$$

with

$$V = H^1(\Omega), \quad V_0 = H^1_{0,D}(\Omega) = \{v\in V : v = 0 \text{ on } \Gamma_D\}, \quad V_g = \{v\in V : v = g_D \text{ on } \Gamma_D\}$$

and

$$\begin{aligned} a(w,v) &= \int_\Omega \left[\sum_{i,j=1}^d a_{ij}\frac{\partial w}{\partial x_j}\frac{\partial v}{\partial x_i} \; dx + \sum_{i=1}^d b_i \frac{\partial w}{\partial x_i} v + cwv\right] \; dx \\ &= \int_\Omega \left[A\operatorname{grad} w\cdot\operatorname{grad} v + b\cdot\operatorname{grad} w\, v + c\, wv\right] \; dx, \\ \langle F, v\rangle &= \int_\Omega fv \; dx + \int_{\Gamma_N} g_N v \; ds. \end{aligned}$$

Trace Operator:

$$\gamma : C^1(\overline{\Omega}) \longrightarrow C(\Gamma), \quad \|v\|_{L^2(\Gamma)} \leq c\,\|v\|_1$$

Extension:

$$\gamma : H^1(\Omega) \longrightarrow L^2(\Gamma) \qquad \gamma_D : H^1(\Omega) \longrightarrow L^2(\Gamma_D)$$

But: The existence of $g \in V = H^1(\Omega)$ with

$$V_g = g + V_0$$

i.e.: $\gamma_D\, g = g_N$ is non-trivial:

$$\gamma(H^1(\Omega)) = H^{1/2}(\Gamma) \subsetneqq L^2(\Gamma).$$

V_0-ellipticity:

For the case

$$\sum_{i,j=1}^{d} a_{ij}(x)\xi_i\xi_j \geq a_0 \sum_{i=1}^{d} \xi_i^2 \quad \text{for all } \xi \in \mathbb{R}, \text{ for almost all } x \in \Omega$$

with $a_0 > 0$ and

$$b(x) \equiv 0, \quad c(x) \geq 0 \text{ for almost all } x \in \Omega$$

it follows

$$a(v, v) \geq a_0\,|v|_1^2.$$

For $|\Gamma_D| > 0$ the ellipticity follows from the Friedrichs inequality:

$$\|v\|_0^2 \leq c_F^2\,|v|_1^2 \quad \text{for all } v \in V_0.$$

Example: Pure Dirichlet boundary value problem: $\Gamma_D = \Gamma$, $\Gamma_N = \emptyset$, $V_0 = H_0^1(\Omega)$.

For the pure Neumann boundary value problem ($\Gamma_D = \emptyset$, $\Gamma_N = \Gamma$) with $c(x) \equiv 0$ the ellipticity follows from the Poincaré inequality:

$$\|v\|_0^2 \leq c_P^2 \left[\left(\int_\Omega v \; dx \right)^2 + |v|_1^2 \right] \quad \text{for all } v \in H^1(\Omega)$$

in the Hilbert space:

$$V_0 = \{v \in H^1(\Omega) : \int_\Omega v \; dx = 0\}.$$

The condition

$$\langle F, 1 \rangle = 0$$

is necessary and sufficient for the existence of a solution.

1.4 Conforming Finite Element Methods

Let $V_h \subset V$ be finite dimensional, $V_{0h} \subset V_h$ with $V_{0h} \subset V_0$ and $V_{gh} = g_h + V_{0h} \subset V_g$.

Galerkin method:

Construction of an approximation $u_h \in V_{gh}$, given by the following variational problem: Find $u_h \in V_{gh}$, such that

$$a(u_h, v_h) = \langle F, v_h \rangle \quad \text{for all } v_h \in V_{0h}.$$

Let a symmetric and non-negative:

Ritz method:

Construction of an approximation $u_h \in V_{gh}$, given by the following optimization problem:

$$J(u_h) = \min_{w_h \in V_{gh}} J(w_h).$$

Finite Element Methods: Special construction of V_h.

1.4.1 Finite Element Methods for Boundary Value Problems of Second-order Ordinary Differential Equations

model problem:

Find $u \in V_g = \{v \in V = H^1(\Omega) : v(0) = g_0\}$ with

$$\int_0^1 u'v' \, dx = \int_0^1 fv \, dx + g_1 v(1) \quad \text{for all } v \in V_0 = \{v \in V = H^1(\Omega) : v(0) = g_0\}$$

The Courant Element:

By introducing nodes x_i, $i = 0, 1, \ldots, N_h$, with

$$0 = x_0 < x_1 < \ldots < x_{N_h} = 1$$

one obtains a subdivision $\mathcal{T}_h$ of the interval $\Omega = (0, 1)$ as a set of subintervals (elements) $T_k = (x_{k-1}, x_k)$ for $k = 1, 2, \ldots, N_h$. The mesh size h of the subdivision is given by

$$h = \max_{k=1,\ldots,N_h} h_k \quad \text{with } h_k = |x_k - x_{k-1}|.$$

Let P_k be the set of all polynomials of degree $\leq k$. V_h is the set of all continuous and piecewise linear functions on Ω:

$$V_h = \{v \in C(\overline{\Omega}) : v_{|T} \in P_1 \text{ for all } T \in \mathcal{T}_h\}.$$

We have (conforming FE space):

$$V_h \subset V = H^1(\Omega).$$

Basis (nodal basis) for V_h: Let x_i, $i = 0, 1, \ldots, N_h$ be a node. $\varphi_i \in V_h$ is given by the condition

$$\varphi_i(x_j) = \delta_{ij} \quad i, j = 0, 1, \ldots, N_h.$$

One immediately sees that $\{\varphi_i : i = 0, 1, \ldots, N_h\}$ is a basis of V_h: The functions are linear independent and each function $v_h \in V_h$ can be written in the form

$$v_h(x) = \sum_{i=0}^{N_h} v_i \varphi_i(x)$$

with $v_i = v_h(x_i)$.

Important: basis functions have a local support.

Test functions: $v_h(0) = 0$:

$$V_{0h} = \{v_h \in V_h : v_h(0) = 0\} = \{v_h \in V_h : v_h = \sum_{i=1}^{N_h} v_i \varphi_i\}.$$

Linear manifold for the solution: $v_h(0) = g_0$:

$$V_{gh} = \{v_h \in V_h : v_h(0) = g_0\} = \{v_h \in V_h : v_h = g_0\varphi_0 + \sum_{i=1}^{N} v_i \varphi_i\}.$$

Obviously:

$$V_{0h} \subset V_0, \quad V_{gh} = g_h + V_{0h} \subset V_g \quad \text{with } g_h = g_0\varphi_0.$$

Determination of the approximate solution: ansatz

$$u_h = g_h + \sum_{j=1}^{N_h} u_j \varphi_j$$

with

$$a(g_h + \sum_{j=1}^{N_h} u_j \varphi_j, \sum_{i=1}^{N_h} v_i \varphi_i) = \langle F, \sum_{i=1}^{N_h} v_i \varphi_i \rangle$$

for all $v_i \in \mathbb{R}$, $i = 1, 2, \ldots, N_h$.

Because of the linearity with respect to v it suffices to test only with the basis functions φ_i, $i = 1, 2, \ldots, N_h$:

$$a(g_h + \sum_{j=1}^{N} u_j \varphi_j, \varphi_i) = \langle F, \varphi_i \rangle \quad \text{for all } i = 1, 2, \ldots, N_h.$$

Because of the linearity with respect to u one obtains

$$\sum_{j=1}^{N} a(\varphi_j, \varphi_i) u_j = \langle F, \varphi_i \rangle - a(g_h, \varphi_i) \quad \text{for all } i = 1, 2, \ldots, N_h.$$

Hence

$$K_h \underline{u}_h = \underline{f}_h$$

with

$$K_h = (K_{ij})_{i,j=1,2\ldots,N_h}, \quad K_{ij} = a(\varphi_j, \varphi_i)$$
$$\underline{u}_h = (u_i)_{i=1,2\ldots,N_h}, \qquad \underline{f}_h = (f_i)_{i=1,2\ldots,N_h} \quad f_i = \langle F, \varphi_i \rangle - a(g_h, \varphi_i).$$

K_h is usually called the stiffness matrix, f_h is called the load vector.

Obviously we have the following relation between the bilinear form on $V_{0,h}$ and the stiffness matrix:

$$a(w_h, v_h) = (K_h \underline{w}_h, \underline{v}_h)_{\ell_2} \quad \text{for all } w_h, v_h \in V_{0h}.$$

Here $(.,.)_{\ell_2}$ denotes the Euclidean scalar product. The Euclidean norm is denoted by $\|.\|_{\ell_2}$.

Remark:

1. Sparse stiffness matrix: Most of the entries of the stiffness matrix are 0 due to the local support of the basis functions:

   $$a(\varphi_j, \varphi_i) = \int_\Omega \varphi_j' \varphi_i' \, dx = 0 \quad \text{for } |i - j| > 1$$

   Here in our example we obtain a tridiagonal matrix because of the special numbering of the unknowns.

2. Element stiffness matrices:

   $$\begin{aligned} (K_h \underline{w}_h, \underline{v}_h)_{\ell_2} &= a(w_h, v_h) = \sum_{T \in \mathcal{T}_h} \int_T w_h' v_h' \, dx = \sum_{T \in \mathcal{T}_h} \sum_{i,j} v_i w_j \int_T \varphi_j' \varphi_i' \, dx \\ &= K_h^{(1)} w_1 v_1 + \sum_{k=2}^{N_h} \left(K_h^{(k)} \begin{pmatrix} w_{k-1} \\ w_k \end{pmatrix}, \begin{pmatrix} v_{k-1} \\ v_k \end{pmatrix} \right)_{\ell_2} \end{aligned}$$

   with the element stiffness matrices:

   $$K_h^{(1)} = \int_{T_1} \varphi_1'(x)^2 \, dx, \qquad K_h^{(k)} = \begin{pmatrix} \int_{T_k} \varphi_{k-1}'(x)^2 \, dx & \int_{T_k} \varphi_{k-1}'(x) \varphi_k'(x) \, dx \\ \int_{T_k} \varphi_k'(x) \varphi_{k-1}'(x) \, dx & \int_{T_k} \varphi_k'(x)^2 \, dx \end{pmatrix}$$

   $\varphi_{i|T}$ are called shape functions.

3. Transformation to a reference element: For the Courant Element (but not necessarily for all finite elements) the computation can be performed with the help of a so-called reference element $\hat{T} = (0,1)$, whose nodes are denoted by $\xi_0 = 0$ and $\xi_1 = 1$. Let $F_k : \overline{\hat{T}} \longrightarrow \overline{T_k}$ be a simple bijective map of the reference element $\hat{T}$ onto the element T_k, here we choose the affine map $F_k(\xi) = x_{k-1} + (x_k - x_{k-1})\,\xi$.

   Transformation of basis functions to the reference element:

$$\varphi_{k-1}(F_k(\xi)) = 1 - \xi \equiv \hat{\varphi}_0(\xi), \quad \varphi_k(F_k(\xi)) = \xi \equiv \hat{\varphi}_1(\xi).$$

   Transformation of integrals to the reference element (substitution rule):

$$K_h^{(1)} = \int_{T_1} \varphi_1'(x)^2 \, dx = \int_{\hat{T}} \varphi_1'(F_k(\xi))^2 |F_1'(\xi)| \, d\xi = \int_{\hat{T}} \hat{\varphi}_1'(\xi)^2 \frac{1}{|F_1'(\xi)|} \, d\xi = \frac{1}{h_1}$$

   and analogously

$$\begin{aligned} K_h^{(k)} &= \begin{pmatrix} \int_{T_k} \varphi_{k-1}'(x)^2 \, dx & \int_{T_k} \varphi_{k-1}'(x)\varphi_k'(x) \, dx \\ \int_{T_k} \varphi_k'(x)\varphi_{k-1}'(x) \, dx & \int_{T_k} \varphi_k'(x)^2 \, dx \end{pmatrix} \\ &= \begin{pmatrix} \int_{T_k} \hat{\varphi}_0'(\xi)^2 \frac{1}{|F_k'(\xi)|} \, d\xi & \int_{T_k} \hat{\varphi}_0'(\xi)\hat{\varphi}_1'(\xi) \frac{1}{|F_k'(\xi)|} \, d\xi \\ \int_{T_k} \hat{\varphi}_1'(\xi)\hat{\varphi}_0'(\xi) \frac{1}{|F_k'(\xi)|} \, d\xi & \int_{T_k} \hat{\varphi}_1'(\xi)^2 \frac{1}{|F_k'(\xi)|} \, d\xi \end{pmatrix} = \frac{1}{h_k} \hat{K} \end{aligned}$$

   with

$$\hat{K} = \begin{pmatrix} \int_{\hat{T}} \hat{\varphi}_0'(\xi)^2 & \int_{\hat{T}} \hat{\varphi}_0'(\xi)\hat{\varphi}_1'(\xi) \, d\xi \\ \int_{\hat{T}} \hat{\varphi}_1'(\xi)\hat{\varphi}_0'(\xi) \, d\xi & \int_{\hat{T}} \hat{\varphi}_1'(\xi)^2 \, d\xi \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

4. Element-wise assembling: From the entries of the element (or local) stiffness matrices,

here $K_h^{(k)}$, one easily obtains the (global) stiffness matrix, here:

$$
\begin{aligned}
K_h &= \\
&= \begin{pmatrix}
K_{11}^{(1)} + K_{00}^{(2)} & K_{01}^{(2)} & 0 & \cdots & \cdots & 0 \\
K_{10}^{(2)} & K_{11}^{(2)} + K_{00}^{(3)} & K_{01}^{(3)} & \ddots & & \vdots \\
0 & K_{10}^{(3)} & K_{11}^{(3)} + K_{00}^{(4)} & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & K_{01}^{(N_h-1)} & 0 \\
\vdots & & \ddots & K_{10}^{(N_h-1)} & K_{11}^{(N_h-1)} + K_{00}^{(N_h)} & K_{01}^{(N_h)} \\
0 & \cdots & \cdots & 0 & K_{10}^{(N_h)} & K_{11}^{(N_h)}
\end{pmatrix} \\
&= \begin{pmatrix}
\frac{1}{h_1} + \frac{1}{h_2} & -\frac{1}{h_2} & 0 & \cdots & \cdots & 0 \\
-\frac{1}{h_2} & \frac{1}{h_2} + \frac{1}{h_3} & -\frac{1}{h_3} & \ddots & & \vdots \\
0 & -\frac{1}{h_3} & \frac{1}{h_3} + \frac{1}{h_4} & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & -\frac{1}{h_{N_h-1}} & 0 \\
\vdots & & \ddots & -\frac{1}{h_{N_h-1}} & \frac{1}{h_{N_h-1}} + \frac{1}{h_{N_h}} & -\frac{1}{h_{N_h}} \\
0 & \cdots & \cdots & 0 & -\frac{1}{h_{N_h}} & \frac{1}{h_{N_h}}
\end{pmatrix}.
\end{aligned}
$$

For the special case of an equidistant subdivision on obtains:

$$
K_h = \frac{1}{h} \begin{pmatrix}
2 & -1 & 0 & \cdots & \cdots & 0 \\
-1 & 2 & -1 & \ddots & & \vdots \\
0 & -1 & 2 & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & -1 & 0 \\
\vdots & & \ddots & -1 & 2 & -1 \\
0 & \cdots & \cdots & 0 & -1 & 1
\end{pmatrix}.
$$

5. In order to compute the element stiffness matrices and the resulting assembled matrix K_h two lists are needed: For a given numbering of the nodes and the elements, the first list contains the information from which nodes each individual element is built of:

element index		
1	0	1
2	1	2
$\vdots$	$\vdots$	$\vdots$
N_h	$N_h - 1$	N_h

The second list contains the coordinates of the nodes:

node index	
0	x_0
1	x_1
2	x_2
$\vdots$	$\vdots$
N_h	x_{N_h}

In an analogous way the computation of the right hand side (usually called the load vector):

$$\underline{f}_h = \begin{pmatrix} \langle F, \varphi_1 \rangle \\ \langle F, \varphi_2 \rangle \\ \vdots \\ \langle F, \varphi_{N_h} \rangle \end{pmatrix} + \begin{pmatrix} g_0/h \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

We have

$$\int_\Omega f\varphi_i \, dx = \sum_{T \in \mathcal{T}_h} \int_T f\varphi_i \, dx$$

and

$$\int_T f(x)\varphi_i(x) \, dx = \int_{\hat{T}} f(F_k(\xi))\varphi_i(F_k(\xi))|F_k'(\xi)| \, d\xi$$

These integrals are typically not computed exactly but only approximatively with the help of a so-called quadrature rule, e.g.: the trapezoidal rule:

$$\int_0^1 h(x)dx \approx \frac{1}{2}[h(0) + h(1)]$$

Hence

$$\underline{f}_h = h \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{N_h-1}) \\ f(x_{N_h})/2 \end{pmatrix} + \begin{pmatrix} g_0/h \\ 0 \\ \vdots \\ 0 \\ g_1 \end{pmatrix}.$$

Remark: The FEM (finite element method) can be interpreted as a FDM (finite difference method) for the classical formulation using a central difference quotient

$$u''(x_i) \approx \frac{1}{h^2}[-u_{i-1} + 2u_i - u_{i+1}]$$

for $i = 1, 2, \ldots, N_h - 1$ and a one-sided difference quotient

$$u''(x_i) \approx \frac{2}{h}\left[u_i' - \frac{u_i - u_{i-1}}{h}\right]$$

for $i = N_h$.

1.4.2 Properties of the stiffness matrix K_h

The stiffness matrix is typically a large scale and sparse matrix. For an appropriate numbering of the unknowns it becomes a banded matrix with a small band width compared to the number of unknowns.

Because of the relation

$$a(w_h, v_h) = (K_h \underline{w}_h, \underline{v}_h)_{\ell_2}$$

properties of the bilinear form a carry over to the stiffness matrix K_h, for example:

- a symmetric $\Longrightarrow$ K_h symmetric.
- a elliptic (coercive) $\Longrightarrow$ K_h positive definite.

Eigenvalue estimates:

Let A be symmetric with respect to the inner product $(.,.)$:

$$(Ax, y) = (x, Ay) \quad \text{for all } x, y \in \mathbb{R}^n.$$

For $(.,.) = (.,.)_{\ell_2}$ A is symmetric iff $A^T = A$.

For symmetric matrices it follows that:

$$\sigma(A) \subset \mathbb{R}$$

with $\sigma(A)$ denoting the spectrum of A, i.e., the set of all eigenvalues of A.

For symmetric matrices the following notations are introduced:

1. A is positive semi-definite, in short $A \geq 0$, iff

$$(Ax, x) \geq 0 \quad \text{for all } x \in \mathbb{R}^n.$$

2. A is positive definite, in short $A > 0$, iff

$$(Ax, x) > 0 \quad \text{for all } x \in \mathbb{R}^n \text{ with } x \neq 0.$$

3. $A \geq B$ iff $A - B \geq 0$.
4. $A > B$ iff $A - B > 0$.

Analogously, $A \leq 0$, $A < 0$, $A \leq B$ and $A < B$ are defined.

We have:

$$A \geq 0 \quad \Longleftrightarrow \quad \lambda \geq 0 \text{ for all } \lambda \in \sigma(A) \quad \Longleftrightarrow \quad \lambda_{\min}(A) \geq 0.$$

and

$$A > 0 \quad \Longleftrightarrow \quad \lambda > 0 \text{ for all } \lambda \in \sigma(A) \quad \Longleftrightarrow \quad \lambda_{\min}(A) < 0.$$

We have

$$\begin{aligned} A \geq \alpha\, I &\iff (Ax,x) \geq \alpha\,(x,x) \quad \text{for all } x \in \mathbb{R}^n \\ &\iff \frac{(Ax,x)}{(x,x)} \geq \alpha \quad \text{for all } x \in \mathbb{R}^n \text{ with } x \neq 0 \\ &\iff \inf_{0 \neq x \in \mathbb{R}^n} \frac{(Ax,x)}{(x,x)} \geq \alpha \end{aligned}$$

and

$$\begin{aligned} A \geq \alpha\, I &\iff \lambda - \alpha \geq 0 \quad \text{for all } \lambda \in \sigma(A) \\ &\iff \lambda_{\min}(A) \geq \alpha \end{aligned}$$

for arbitrary $\alpha \in \mathbb{R}$ and, therefore,

$$\lambda_{\min}(A) = \inf_{0 \neq x \in \mathbb{R}^n} \frac{(Ax,x)}{(x,x)}$$

The expression $(Ax,x)/(x,x)$ is called the Rayleigh quotient. Analogously it follows:

$$\lambda_{\max}(A) = \sup_{0 \neq x \in \mathbb{R}^n} \frac{(Ax,x)}{(x,x)}$$

Let A and C be symmetric matrices with respect to the inner product $(.,.)$ and $C > 0$. Then $C^{-1}A$ is symmetric with respect to the inner product $(.,.)_C$, given by

$$(x,y)_C = (Cx,y) \quad \text{for all } x,y \in \mathbb{R}^n,$$

since

$$(C^{-1}Ax,y)_C = (CC^{-1}Ax,y) = (Ax,y) = (x,Ay) = (C^{-1}Cx,Ay) = (Cx,C^{-1}Ay).$$

So it follows

$$\lambda_{\min}(C^{-1}A) = \inf_{0 \neq x \in \mathbb{R}^n} \frac{(C^{-1}Ax,x)_C}{(x,x)_C} = \inf_{0 \neq x \in \mathbb{R}^n} \frac{(Ax,x)}{(Cx,x)}.$$

and

$$\lambda_{\max}(C^{-1}A) = \sup_{0 \neq x \in \mathbb{R}^n} \frac{(C^{-1}Ax,x)_C}{(x,x)_C} = \sup_{0 \neq x \in \mathbb{R}^n} \frac{(Ax,x)}{(Cx,x)}.$$

Eigenvalue estimates for the stiffness matrix K_h:

$$\begin{aligned} \left(K_h^{(k)} \begin{pmatrix} v_{k-1} \\ v_k \end{pmatrix}, \begin{pmatrix} v_{k-1} \\ v_k \end{pmatrix} \right)_{\ell_2} &= \frac{1}{h_k} \left(\hat{K} \begin{pmatrix} v_{k-1} \\ v_k \end{pmatrix}, \begin{pmatrix} v_{k-1} \\ v_k \end{pmatrix} \right) \\ &\leq \frac{1}{h_k}\, \lambda_{\max}(\hat{K})\, \left[v_{k-1}^2 + v_k^2 \right] = \frac{2}{h_k} \left[v_{k-1}^2 + v_k^2 \right] \end{aligned}$$

and, therefore,

$$\begin{aligned}(K_h\underline{v}_h,\underline{v}_h)_{\ell_2}=a(v_h,v_h) &\leq \frac{1}{h_1}v_1^2+\sum_{k=2}^{N_h}\frac{2}{h_k}\left[v_{k-1}^2+v_k^2\right]\\ &\leq \frac{4}{\min_k h_k}\sum_{k=1}^{N_h}v_k^2=\frac{4}{\min_k h_k}\|\underline{v}_h\|_{\ell_2}^2.\end{aligned}$$

i.e.:

$$\lambda_{\max}(K_h)=\sup_{\underline{v}_h\neq 0}\frac{(K_h\underline{v}_h,\underline{v}_h)_{\ell_2}}{(\underline{v}_h,\underline{v}_h)_{\ell_2}}\leq\frac{4}{\min_k h_k}.$$

On the other hand we have (Friedrichs inequality):

$$a(v_h,v_h)\geq\frac{1}{c_F^2}\|v_h\|_0^2.$$

So

$$(K_h\underline{v}_h,\underline{v}_h)_{\ell_2}=a(v_h,v_h)\geq\frac{1}{c_F^2}\|v_h\|_0^2=\frac{1}{c_F^2}(M_h\underline{v}_h,\underline{v}_h)_{\ell_2}$$

with $M_h=(M_{ij})$, given by

$$M_{ij}=\int_\Omega\varphi_j\varphi_i\,dx,$$

the so-called mass matrix. We have

$$\begin{aligned}(M_h\underline{w}_h,\underline{v}_h)_{\ell_2} &= \sum_{T\in\mathcal{T}_h}\int_T w_hv_h\,dx=\sum_{T\in\mathcal{T}_h}\sum_{i,j}v_iw_j\int_T\varphi_j\varphi_i\,dx\\ &= M_h^{(1)}w_1v_1+\sum_{k=2,N_h}\left(M_h^{(k)}\begin{pmatrix}w_{k-1}\\w_k\end{pmatrix},\begin{pmatrix}v_{k-1}\\v_k\end{pmatrix}\right)_{\ell_2}\end{aligned}$$

with the element (or local) mass matrices

$$M_h^{(1)}=\int_{T_1}\varphi_1(x)\varphi_1(x)\,dx,\qquad M_h^{(k)}=\begin{pmatrix}\int_{T_k}\varphi_{k-1}(x)^2\,dx & \int_{T_k}\varphi_{k-1}(x)\varphi_k(x)\,dx\\ \int_{T_k}\varphi_k(x)\varphi_{k-1}(x)\,dx & \int_{T_k}\varphi_k(x)^2\,dx\end{pmatrix}$$

Transformation to a reference element:

$$M_h^{(1)} = \int_{T_1}\varphi_1(x)\varphi_1(x)\,dx=\int_{\hat{T}}\hat{\varphi}_1(\xi)\hat{\varphi}_1(\xi)|F_1'(\xi)|\,d\xi=\frac{h_1}{3}$$

and, analogously,

$$\begin{aligned} M_h^{(k)} &= \begin{pmatrix} \int_{T_k} \varphi_{k-1}(x)^2 \, dx & \int_{T_k} \varphi_{k-1}(x)\varphi_k(x) \, dx \\ \int_{T_k} \varphi_k(x)\varphi_{k-1}(x) \, dx & \int_{T_k} \varphi_k(x)^2 \, dx \end{pmatrix} \\ &= \begin{pmatrix} \int_{T_k} \hat{\varphi}_0(\xi)^2 |F_k'(\xi)| \, d\xi & \int_{T_k} \hat{\varphi}_0(\xi)\hat{\varphi}_1(\xi)|F_k'(\xi)| \, d\xi \\ \int_{T_k} \hat{\varphi}_1(\xi)\hat{\varphi}_0(\xi)|F_k'(\xi)| \, d\xi & \int_{T_k} \hat{\varphi}_1(\xi)^2 |F_k'(\xi)| \, d\xi \end{pmatrix} = h_k \, \hat{M} \end{aligned}$$

with

$$\hat{M} = \begin{pmatrix} \int_{\hat{T}} \hat{\varphi}_0(\xi)^2 \, d\xi & \int_{\hat{T}} \hat{\varphi}_0(\xi)\hat{\varphi}_1(\xi) \, d\xi \\ \int_{\hat{T}} \hat{\varphi}_1(\xi)\hat{\varphi}_0(\xi) \, d\xi & \int_{\hat{T}} \hat{\varphi}_1(\xi)^2 \, dx\xi \end{pmatrix} = \frac{1}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Hence it follows:

$$\begin{aligned} \left(M_h^{(k)} \begin{pmatrix} v_{k-1} \\ v_k \end{pmatrix}, \begin{pmatrix} v_{k-1} \\ v_k \end{pmatrix} \right)_{\ell_2} &= h_k \left(\hat{M} \begin{pmatrix} v_{k-1} \\ v_k \end{pmatrix}, \begin{pmatrix} v_{k-1} \\ v_k \end{pmatrix} \right) \\ &\geq h_k \, \lambda_{\min}(\hat{M}) \, \left[v_{k-1}^2 + v_k^2 \right] = \frac{h_k}{6} \left[v_{k-1}^2 + v_k^2 \right] \end{aligned}$$

and, therefore,

$$\begin{aligned} (M_h \underline{v}_h, \underline{v}_h)_{\ell_2} = \|v_h\|_0^2 &\geq \frac{h_1}{3} v_1^2 + \sum_{k=1}^{N_h} \frac{h_k}{6} \left[v_{k-1}^2 + v_k^2 \right] \\ &\geq \frac{\min_k h_k}{6} \sum_{k=1}^{N_h} v_k^2 = \frac{\min_k h_k}{6} \|\underline{v}_h\|_{\ell_2}^2 \end{aligned}$$

So we have:

$$(K_h \underline{v}_h, \underline{v}_h)_{\ell_2} \geq \frac{\min_k h_k}{6 c_F^2} \|\underline{v}_h\|_{\ell_2}^2,$$

i.e.:

$$\lambda_{\min}(K_h) = \inf_{\underline{v}_h \neq 0} \frac{(K_h \underline{v}_h, \underline{v}_h)_{\ell_2}}{(\underline{v}_h, \underline{v}_h)_{\ell_2}} \geq \frac{\min_k h_k}{6 c_F^2}.$$

For the condition number of K_h one obtains:

$$\kappa(K_h) \leq 24 c_F^2 \, \frac{1}{\min_k h_k^2}$$

Special case equidistant subdivision:

$$\kappa(K_h) \leq 24 c_F^2 \, \frac{1}{h^2} = O(\frac{1}{h^2}).$$

The exponent 2, which corresponds to the order of the differential equation, is sharp.

Remark: In a similar way one can easily show that

$$\kappa(M_h) = O(1)$$

for equidistant subdivisions.

1.4.3 The Discretization Error

Homogenization

Under the assumptions $V_h \subset V$, $V_{0h} \subset V_0$ and $V_{gh} = g_h + V_{0h} \subset V_g$ the variational problem in V as well as the corresponding finite dimensional variational problem can be simultaneously homogenized with the help of g_h. Therefore, without loss of generality, it suffices to consider variational problems with:

$$g_h = g = 0, \quad V = V_g = V_0, \quad V_h = V_{gh} = V_{0h}.$$

The existence and uniqueness of the approximation u_h directly follows under the assumptions of the Lax-Milgram Theorem:

Theorem 1.5. *Let V be a Hilbert space, $F \in V^*$ and $a : V \times V \longrightarrow \mathbb{R}$ be a bilinear form with the following properties:*

1. *a is coercive on V: There is a constant $\mu_1 > 0$ with*

$$\mu_1 \|v\|^2 \leq a(v, v) \quad \text{for all } v \in V,$$

2. *a is bounded on V: There is a constant $\mu_2 > 0$ with*

$$|a(w, v)| \leq \mu_2 \|w\| \|v\| \quad \text{for all } w, v \in V.$$

Moreover, let V_h be a finite dimensional subspace of V. Then there exists a unique solution $u_h \in V_h$ with

$$a(u_h, v_h) = \langle F, v_h \rangle \quad \text{for all } v_h \in V_h.$$

The following theorem (Cea's Theorem) is of fundamental importance for the estimation of the discretization error:

Theorem 1.6 (Cea). *Let V be a Hilbert space, $F \in V^*$ and $a : V \times V \longrightarrow \mathbb{R}$ a bilinear form with the following properties:*

1. *a is coercive on V: There is a constant $\mu_1 > 0$ with*

$$\mu_1 \|v\|^2 \leq a(v, v) \quad \text{for all } v \in V,$$

2. *a is bounded on V: There is a constant $\mu_2 > 0$ with*

$$|a(w, v)| \leq \mu_2 \|w\| \|v\| \quad \text{for all } w, v \in V.$$

Furthermore, let V_h be a finite-dimensional subspace of V. Then we have:

$$\|u - u_h\| \leq \frac{\mu_2}{\mu_1} \inf_{w_h \in V_h} \|u - w_h\|.$$

Proof. By subtracting

$$\begin{aligned} a(u, v_h) &= \langle F, v_h \rangle \quad \text{for all } v_h \in V_h \\ a(u_h, v_h) &= \langle F, v_h \rangle \quad \text{for all } v_h \in V_h \end{aligned}$$

we obtain the so-called Galerkin orthogonality:

$$a(u - u_h, v_h) = 0 \quad \text{for all } v_h \in V_h$$

With $v_h = (u - u_h) - (u - w_h)$ it follows that

$$\mu_1 \|u - u_h\|^2 \leq a(u - u_h, u - u_h) = a(u - u_h, u - w_h) \leq \mu_2 \|u - u_h\| \|u - w_h\|.$$

□

Remark: Let a be additionally symmetric. Then a is a (further) scalar product on V with corresponding norm, given by

$$\|v\|_A^2 = a(v, v),$$

which is equivalent to the original norm $\|\|$:

$$\mu_1 \|v\|^2 \leq \|v\|_A^2 \leq \mu_2 \|v\|^2.$$

The Galerkin orthogonality then becomes the A-orthogonality:

$$u - u_h \perp_A V_h$$

and we have:

$$\|u - u_h\|_A^2 = a(u - u_h, u - u_h) = a(u - u_h, u - w_h) \leq \|u - u_h\|_A \|u - w_h\|_A.$$

Hence

$$\|u - u_h\|_A = \inf_{w_h \in V_h} \|u - w_h\|_A$$

and, therefore,

$$\|u - u_h\| \leq \sqrt{\frac{\mu_2}{\mu_1}} \inf_{w_h \in V_h} \|u - w_h\|$$

Because of

$$\begin{aligned} J(w_h) &= \frac{1}{2} a(w_h, w_h) - \langle F, w_h \rangle = \frac{1}{2} a(w_h, w_h) - a(u, w_h) \\ &= \frac{1}{2} a(w_h - u, w_h - u) - \frac{1}{2} a(u, u) = \frac{1}{2} \|w_h - u\|_A^2 - \frac{1}{2} \|u\|_A^2 \end{aligned}$$

it follows

$$J(u_h) = \inf_{w_h \in V_h} J(w_h).$$

This property is also a direct consequence of Theorem 1.3.

Cea's Theorem states that the discretization error can be estimated by the approximation error.

Estimation of the approximation error

Let $v \in H^1(0,1)$. Since $H^1(0,1) \subset C[0,1]$ (see the trace operator) the interpolation operator I_h is well-defined: $v_h = I_h v \in V_h$ is that continuous and piecewise linear function, which coincides with v at the nodes:

$$v_h(x_i) = v(x_i) \quad \text{for all } i = 0, 1, \ldots, N_h.$$

The approximation error can be estimated by the interpolation error:

$$\inf_{v_h \in V_h} \|u - v_h\|_1 \leq \|u - I_h u\|_1$$

Lemma 1.3. *For $u \in H^2(0,1)$ it follows: There are constants $C_0, C_1 > 0$ such that*

$$\|v - I_h v\|_0 \leq C_0 \left(\sum_k h_k^4 \, |v|_{2,T_k}^2 \right)^{1/2} \leq C_0 \, h^2 \, |v|_2$$

and

$$|v - I_h v|_1 \leq C_1 \left(\sum_k h_k^2 \, |v|_{2,T_k}^2 \right)^{1/2} \leq C_1 \, h \, |v|_2$$

Proof. Transformation onto the reference element:

$$\begin{aligned}
\int_\Omega |v(x) - I_h v(x)|^2 \, dx &= \sum_{k=1}^{N_h} \int_{T_k} |v(x) - I_h v(x)|^2 \, dx \\
&= \sum_{k=1}^{N_h} h_k \int_{\hat{T}} |(v \circ F_k)(\xi) - I_h v \circ F_k(\xi)|^2 \, d\xi \\
&= \sum_{k=1}^{N_h} h_k \int_{\hat{T}} |(v \circ F_k)(\xi) - \hat{I}(v \circ F_k)(\xi)|^2 \, d\xi.
\end{aligned}$$

On the reference element it follows for $\hat{v}(\xi) = (v \circ F_k)(\xi)$:

$$\begin{aligned}
\hat{v}(\xi) - \hat{I}\hat{v}(\xi) &= \hat{v}(\xi) - [\hat{v}(0) + (v(1) - v(0))\xi] \\
&= \xi \int_0^1 [\hat{v}'(\xi y) - \hat{v}'(y)] \, dy \\
&= \xi \int_0^1 \int_y^{\xi y} \hat{v}''(z) \, dz dy.
\end{aligned}$$

So

$$|\hat{v}(\xi) - \hat{I}\hat{v}(\xi)| \leq C \, \|\hat{v}''\|_0$$

and, therefore,

$$\int_{\hat{T}} |\hat{v}(\xi) - \hat{I}\hat{v}(\xi)|^2 \, d\xi \;\; \leq \;\; C^2 \int_{\hat{T}} |\hat{v}''(\xi)|^2 \, d\xi.$$

By transforming back onto T_k we obtain:

$$\int_{T_k} |v(x) - I_h v(x)|^2 \, dx \;\; \leq \;\; C^2 \, h_k^4 \int_{T_k} |v''(x)|^2 \, dx.$$

Analogously, the second estimate follows using:

$$\begin{aligned} \hat{v}'(\xi) - (\hat{I}\hat{v})'(\xi) &= \hat{v}'(\xi) - (v(1) - v(0)) \\ &= \int_0^1 [\hat{v}'(\xi) - \hat{v}'(y)] \; dy \\ &= \int_0^1 \int_y^\xi \hat{v}''(z) \; dz dy \end{aligned}$$

□

Summarizing, we obtain:

Theorem 1.7. *Let $u \in V \cap H^2(\Omega)$ be the exact solution of the variational problem and let $u_h \in V_h$ be the approximate solution of the FEM with the Courant element. Then we have:*

$$\|u - u_h\|_1 \leq C_1 \left(\sum_k h_k^2 \, |u|_{2,T_k}^2 \right)^{1/2} \leq C_1 \, h \, |u|_2.$$

Remark: Convergence for $u \in H^1(\Omega)$: $H^2(\Omega)$ is dense in $H^1(\Omega)$. Hence

$$\lim_{h \to 0} \|u - u_h\|_1 \leq \frac{\mu_2}{\mu_1} \lim_{h \to 0} \inf_{v_h \in V_h} \|u - v_h\|_1 \to 0.$$

Remark: Estimation in other norms: Example L^2-Norm:

$$\|u - u_h\|_0 \leq C_0 \, h^2 \, |u|_2.$$

by the so-called Aubin-Nitsche duality argument.

a-posteriori error estimators:

For the discretization error we have:

$$a(u - u_h, v) = \langle F, v \rangle - a(u_h, v) \quad \text{for all } v \in V.$$

So, by the Lax-Milgram Theorem:

$$\frac{1}{\mu_2} \sup_{v \in V} \frac{\langle F, v \rangle - a(u_h, v)}{\|v\|} \leq \|u - u_h\| \leq \frac{1}{\mu_1} \sup_{v \in V} \frac{\langle F, v \rangle - a(u_h, v)}{\|v\|}$$

So, the discretization error can be estimated from above and from below by the norm of the residual. Furthermore, we have:

$$\langle F, v\rangle - a(u_h, v) = \langle F, v - v_h\rangle - a(u_h, v - v_h)$$

for all $v_h \in V_h$.

Estimation of the residual for the model problem with $w = v - v_h$:

$$\begin{aligned}
\langle F, w\rangle - a(u_h, w) &= \sum_{k=1}^{N_h} \left[\int_{T_k} fw \; dx - \int_{T_k} u_h' w' \; dx \right] - g_1 w(1) \\
&= \sum_{k=1}^{N_h} \left[\int_{T_k} fw \; dx - u_h' w \Big|_{x_{k-1}}^{x_k} + \int_{T_k} u_h'' w \; dx \right] - g_1 w(1) \\
&= \sum_{k=1}^{N_h} \int_{T_k} (f + u_h'') w \; dx + \sum_{l=1}^{N_h - 1} [u_h'](x_l) w(x_l) + (g_1 - u_h'(x_{N_h}))\, w(1)
\end{aligned}$$

with

$$[u_h'](x_l) = u_h'(x_l+) - u_h'(x_l-).$$

Hence

$$\begin{aligned}
&\langle F, w\rangle - a(u_h, w) \\
&\leq \sum_{k=1}^{N_h} \|f + u_h''\|_{0,T_k} \|w\|_{0,T_k} + \sum_{l=1}^{N_h - 1} |[u_h'](x_l)||w(x_l)| + |g_1 + u_h'(x_{N_h})||w(1)|
\end{aligned}$$

For $v_h = I_h v$, the jump terms vanish (only in 1D). Using the estimation of the interpolation error

$$\|v - v_h\|_{0,T_k}^2 = \int_{T_k} |v - I_h v|^2 \; dx \leq C_I^2 \, h_k^2 \int_{T_k} |v'|^2 \; dx = C_I^2 \, h_k^2 |v|_{1,T_k}^2$$

it follows:

$$\begin{aligned}
&\langle F, w\rangle - a(u_h, w) \\
&\leq C_I \sum_{k=1}^{N_h} h_k \, \|f + u_h''\|_{0,T_k} |v|_{1,T_k} \\
&\leq C_I \left(\sum_{k=1}^{N_h} h_k^2 \, \|f + u_h''\|_{0,T_k}^2 \right)^{1/2} \left(\sum_{k=1}^{N_h} |v|_{1,T_k}^2 \right)^{1/2} = C_I \, \eta \, |v|_1 \leq C_I \, \eta \, \|v\|_1
\end{aligned}$$

In summary, we obtain

$$\|u - u_h\|_1 \leq \frac{C_I}{\mu_1} \, \eta$$

with

$$\eta^2 = \sum_{k=1}^{N_h} \eta_k^2, \quad \eta_k^2 = h_k^2 \, \|f + u_h''\|_{0,T_k}^2.$$

In 1D this coincides with the a-priori estimation.

1.4.4 Finite Element Methods for Boundary Value Problems of Partial Differential Equations

- Subdivisions

  in 2D: triangles, quadrilaterals

  in 3D: tetrahedra, hexahedra, ...

- Basis functions, shape functions:

  $P_k = \{\sum_{|\nu|\leq k} c_\nu\, x^\nu\}$, $Q_k = \{\sum_{\nu_i \leq k} c_\nu\, x^\nu\}$

  in 2D: $k = 1$, P_1 for triangles, basis functions: pyramid functions

  in 2D: $k = 1$, Q_1 (bilinear functions) for quadrilaterals, bilinear transformations onto the reference element = unit square.

- element-by-element assembling: analogous

  here: important in order to efficiently work with unstructured meshes.

- Properties of K_h:

  analogous properties

  BUT: band width grows.

- Discretization error: analogously

  Rule of thumb $O(h^k)$ for P_k, Q_k.

  higher dimensional elements: acute, obtuse angles, regular, quasi-uniform meshes.

  Bramble-Hilbert lemma.

1.5 Iterative Methods for Linear Systems of Equations

1.5.1 The preconditioned Richardson method

Theorem 1.5 also yields an iterative method for determining $u_h \in V_h$:

$$(u_h^{(n+1)}, v_h) = (u_h^{(n)}, v_h) + \tau\,[\langle F, v_h\rangle - a(u_h^{(n)}, v_h)] \quad \text{for all } v_h \in V_h.$$

Here $(.,.)$ denotes the scalar product in V. This scalar product (as any other bilinear form) can be represented by a matrix, say B_h:

$$(w_h, v_h) = (B_h \underline{w}_h, \underline{v}_h)_{\ell_2}.$$

The matrix B_h is symmetric and positive definite. Therefore, we obtain the following iterative method in matrix-vector notation:

$$B_h \underline{u}_h^{(n+1)} = B_h \underline{u}_h^{(n)} + \tau\,(\underline{f}_h - K_h \underline{u}_h^{(n)}),$$

which is a preconditioned Richardson method:

$$\underline{u}_h^{(n+1)} = \underline{u}_h^{(n)} + \tau\, B_h^{-1}(\underline{f}_h - K_h \underline{u}_h^{(n)}).$$

Algorithm:

1. Compute $\underline{r}_h = \underline{f}_h - K_h \underline{u}_h^{(n)}$
2. Solve $B_h \underline{w}_h = \underline{r}_h$
3. Compute $\underline{u}_h^{(n+1)} = \underline{u}_h^{(n)} + \tau\, \underline{w}_h$

Convergence analysis:

The V_h-coercivity of a is equivalent to:

$$\mu_1\,(B_h \underline{v}_h, \underline{v}_h)_{\ell_2} \le (K_h \underline{v}_h, \underline{v}_h)_{\ell_2}$$

The V_h-boundedness of a is equivalent to:

$$(K_h \underline{w}_h, \underline{v}_h)_{\ell_2} \le \mu_2\,(B_h \underline{w}_h, \underline{w}_h)_{\ell_2}^{1/2} (B_h \underline{v}_h, \underline{v}_h)_{\ell_2}^{1/2} \quad \text{for all } \underline{v}_h, \underline{w}_h \in \mathbb{R}^{N_h}$$

From Theorem 1.5 and its consequences it follows immediately that the method is q-linear convergent with respect to the B_h-norm with a convergence factor

$$q = \sqrt{1 - \left(\frac{\mu_1}{\mu_2}\right)^2}$$

with optimal choice of the parameter

$$\tau = \frac{2\mu_1}{\mu_2^2}.$$

Hence, the convergence rate is independent of h!

BUT: The step 2 is too expensive.

Remedy: B_h is replaced by a (symmetric and positive definite) matrix C_h, hence:

$$\underline{u}_h^{(n+1)} = \underline{u}_h^{(n)} + \tau\, C_h^{-1}(\underline{f}_h - K_h \underline{u}_h^{(n)}) \tag{1.6}$$

Algorithm:

1. Compute $\underline{r}_h = \underline{f}_h - K_h \underline{u}_h^{(n)}$
2. Solve $C_h \underline{w}_h = \underline{r}_h$
3. Compute $\underline{u}_h^{(n+1)} = \underline{u}_h^{(n)} + \tau\, \underline{w}_h$

It is assumed that the linear system $C_h \underline{w}_h = \underline{r}_h$ is easy to solve.

One immediately obtains the following convergence result:

Theorem 1.8. *Let C_h be a symmetric and positive definite matrix. Assume that constants ν_1 and ν_2 exist with*

$$\nu_1\,(C_h \underline{v}_h, \underline{v}_h)_{\ell_2} \le (K_h \underline{v}_h, \underline{v}_h)_{\ell_2} \quad \textit{for all } \underline{v}_h \in \mathbb{R}^{N_h}$$

and

$$(K_h \underline{w}_h, \underline{v}_h)_{\ell_2} \le \nu_2\,(C_h \underline{w}_h, \underline{w}_h)_{\ell_2}^{1/2} (C_h \underline{v}_h, \underline{v}_h)_{\ell_2}^{1/2} \quad \textit{for all } \underline{v}_h, \underline{w}_h \in \mathbb{R}^{N_h}$$

Then we have for the iterative method (1.6):

$$\|u_h - u_h^{(n+1)}\|_{C_h} \le q\, \|u_h - u_h^{(n)}\|_{C_h}$$

with

$$q = \sqrt{1 - \left(\frac{\nu_1}{\nu_2}\right)^2}$$

for the choice

$$\tau = \frac{\nu_1}{\nu_2^2}.$$

Discussion of the assumptions:

V_h-ellipticity:

$$\nu_1\,(C_h \underline{v}_h, \underline{v}_h)_{\ell_2} \le (K_h \underline{v}_h, \underline{v}_h)_{\ell_2} = (K_h^{\mathrm{sym}} \underline{v}_h, \underline{v}_h)_{\ell_2} \quad \text{for all } \underline{v}_h \in \mathbb{R}^{N_h} \tag{1.7}$$

i.e.:

$$\nu_1\, C_h \le K_h^{\mathrm{sym}} = \frac{1}{2}(K_h + K_h^T) \quad \text{or} \quad \nu_1 \le \lambda_{\min}(C_h^{-1} K_h^{\mathrm{sym}})$$

V_h-boundedness:

$$\sup_{0 \neq \underline{v}_h \in \mathbb{R}^{N_h}} \frac{(K_h \underline{w}_h, \underline{v}_h)_{\ell_2}}{(C_h \underline{v}_h, \underline{v}_h)_{\ell_2}^{1/2}} \le \nu_2\,(C_h \underline{w}_h, \underline{w}_h)_{\ell_2}^{1/2} \quad \text{for all } \underline{w}_h \in \mathbb{R}^{N_h}$$

or, equivalently:

$$(C_h^{-1} K_h \underline{w}_h, K_h \underline{w}_h)_{\ell_2} \le \nu_2^2\,(C_h \underline{w}_h, \underline{w}_h)_{\ell_2} \quad \text{for all } \underline{w}_h \in \mathbb{R}^{N_h} \tag{1.8}$$

Proof.

$$\begin{aligned}\sup_{0\neq \underline{v}_h\in\mathbb{R}^{N_h}} \frac{(K_h\underline{w}_h,\underline{v}_h)_{\ell_2}}{(C_h\underline{v}_h,\underline{v}_h)_{\ell_2}^{1/2}} &= \sup_{0\neq \underline{v}_h\in\mathbb{R}^{N_h}} \frac{(C_hC_h^{-1}K_h\underline{w}_h,\underline{v}_h)_{\ell_2}}{(C_h\underline{v}_h,\underline{v}_h)_{\ell_2}^{1/2}} = \sup_{0\neq \underline{v}_h\in\mathbb{R}^{N_h}} \frac{(C_h^{-1}K_h\underline{w}_h,\underline{v}_h)_{C_h}}{(\underline{v}_h,\underline{v}_h)_{C_h}^{1/2}} \\ &= \|C_h^{-1}K_h\underline{w}_h\|_{C_h} = (C_hC_h^{-1}K_h\underline{w}_h, C_h^{-1}K_h\underline{w}_h)_{\ell_2}^{1/2} \\ &= (C_h^{-1}K_h\underline{w}_h, K_h\underline{w}_h)_{\ell_2}^{1/2}\end{aligned}$$

□

i.e.:

$$K_h^TC_h^{-1}K_h \leq \nu_2^2\, C_h \quad \text{or} \quad \lambda_{\max}(C_h^{-1}K_h^TC_h^{-1}K_h) \leq \nu_2^2$$

If, in addition, a is symmetric, then K_h is symmetric and the conditions (1.7) and (1.8) simplify to

$$\nu_1\, C_h \leq K_h \leq \nu_2\, C_h$$

and one obtains sharper bounds for the convergence rate:

Theorem 1.9. *Let C_h be a symmetric and positive definite matrix. Assume that constants $\nu_1 > 0$ and ν_2 exist with*

$$\nu_1\, C_h \leq K_h \leq \nu_2\, C_h.$$

Then we have for the iterative method (1.6):

$$\|u_h - u_h^{(n+1)}\|_{C_h} \leq q\, \|u_h - u_h^{(n)}\|_{C_h}$$

with

$$q = \frac{\nu_2/\nu 1 - 1}{\nu_2/\nu_1 + 1} = 1 - \frac{2}{\nu_2/\nu_1 + 1}$$

for the parameter

$$\tau = \frac{2}{\nu_1 + \nu_2}.$$

Remark: For $\nu_1 = \lambda_{\min}(C_h^{-1}K_h)$ and $\nu_2 = \lambda_{\max}(C_h^{-1}K_h)$ one obtains

$$q = \frac{\kappa(C_h^{-1}K_h) - 1}{\kappa(C_h^{-1}K_h) + 1}$$

Hence: preconditioning!

Example: For the one-dimensional model problem, it was shown that

$$c_1\, h\, I \leq K_h \leq c_2\, h^{-1}\, I,$$

so $\nu_2/\nu_1 = O(h^{-2})$.

Remark: From the error estimates of the Theorems 1.8 and 1.9 it immediately follows that:

$$\|\underline{u}_h^{(n)} - \underline{u}_h\|_{C_h} \leq q^n \, \|\underline{u}_h^{(0)} - \underline{u}_h\|_{C_h}$$

Therefore, the initial error is reduced by a prescribed factor $\varepsilon > 0$, if $q^n \leq \varepsilon$, hence, about

$$n = \frac{-\ln \varepsilon}{-\ln q}$$

iterations are necessary. If $\nu_2/\nu_1 \gg 1$ it follows under the assumptions of Theorem 1.8:

$$n = \frac{-\ln \varepsilon}{-\ln q} = \frac{-\ln \varepsilon}{-\ln \sqrt{1-(\nu_1/\nu_2)^2}} \approx (-2 \ln \varepsilon) \left(\frac{\nu_2}{\nu_1}\right)^2,$$

under the assumptions of Theorem 1.9:

$$n = \frac{-\ln \varepsilon}{-\ln q} = \frac{-\ln \varepsilon}{-\ln(1-2/(\nu_2/\nu_1+1))} \approx \frac{1}{2}(-\ln \varepsilon) \left(\frac{\nu_2}{\nu_1}+1\right).$$

1.5.2 Preconditioning

Let $u_h^{(n)}$ be an approximation of the solution of the variational problem

$$a(u_h, v_h) = \langle F, v_h \rangle \quad \text{for all } v_h \in V_h.$$

Then the exact solution is given by

$$u_h = u_h^{(n)} + w_h,$$

if the correction w_h is the solution of the residual equation

$$a(w_h, v_h) = \langle F, v_h \rangle - a(u_h^{(n)}, v_h) \quad \text{for all } v_h \in V_h$$

Problem: Too expensive

Remedy: Approximate solution of the residual equation on a subspace or on several subspaces:

subspace correction:

Let $W_h \subset V_h$. Consider the following variational problem: Find $w_h \in W_h$ such that

$$a(w_h, v_h) = \langle F, v_h \rangle - a(u_h^{(n)}, v_h) \quad \text{for all } v_h \in W_h.$$

Typically corrections are calculated not only with respect to one subspace but with respect to several subspaces $V_{h,s} \subset V_h$, $s = 1, \ldots, p$ with

$$\sum_s V_{h,s} = V_h.$$

Example: Let $\{\varphi_i : i = 1, 2, \ldots, N_h\}$ be a basis of V_h. For the choice $V_{h,i} = \text{span}(\varphi_i)$, one obtains the following subspace correction equations for the corrections $w_{h,i} = w_i\varphi_i$:

$$a(\varphi_i, \varphi_i)w_i = \langle F, \varphi_i\rangle - a(u_h^{(n)}, \varphi_i)$$

hence

$$K_{ii}w_i = r_i.$$

with

$$\underline{r}_h = (r_i) = \underline{f}_h - K_h\underline{u}^{(n)}.$$

In the following two possibilities are described how to construct a new approximate solution from these corrections:

Additive Schwarz methods

One computes the corrections $w_{h,s} \in V_{h,s}$ for all subspaces always starting with the old approximate solution $u_h^{(n)}$ and sum up these corrections:

$$u_h^{(n+1)} = u_h^{(n)} + \tau \sum_s w_{h,s}.$$

Example:

$$u_h^{(n+1)} = u_h^{(n)} + \tau \sum_i w_i\varphi_i$$

or in matrix-vector notation:

$$\underline{u}_h^{(n+1)} = \underline{u}_h^{(n)} + \tau\, \underline{w}_h$$

with

$$C_h\underline{w}_h = \underline{f}_h - K_h\underline{u}_h^{(n)},$$

where $C_h = D_h = \text{diag}(K_h)$. This corresponds to one step of the Jacobi method, i.e.: the preconditioned Richardson method with preconditioner $C_h = D_h = \text{diag}(K_h)$.

Multiplicative Schwarz methods

After each correction one immediately updates the approximate solution, which is then considered as old approximate solution for the next subspace correction:

$$u_h^{(n+s/p)} = u_h^{(n+(s-1)/p)} + w_{h,s}$$

with

$$a(w_{h,s}, v_h) = \langle F, v_h\rangle - a(u_h^{(n+(s-1)/p)}, v_h) \quad \text{for all } v_h \in V_{h,s}$$

for $s = 1, 2, \ldots, p$.

Example:

$$u_h^{(n+i/N_h)} = u_h^{(n+(i-1)/N_h)} + w_i\varphi_i$$

with

$$a(\varphi_i, \varphi_i)w_i = \langle F, \varphi_i\rangle - a(u_h^{(n+(i-1)/N_h)}, \varphi_i)$$

Hence

$$u^{(n+1)} = u^{(n)} + \sum_{i=1}^{N_h} w_i\varphi_i$$

with

$$\begin{aligned} a(\varphi_i, \varphi_i)w_i &= \langle F, \varphi_i\rangle - a(u_h^{(n+(i-1)/N_h)}, \varphi_i) = \langle F, \varphi_i\rangle - a(u_h^{(n)} + \sum_{j=1}^{i-1} w_j\varphi_j, \varphi_i) \\ &= \langle F, \varphi_i\rangle - a(u_h^{(n)}, \varphi_i) - \sum_{j=1}^{i-1} a(\varphi_j, \varphi_i)w_j, \end{aligned}$$

i.e.

$$\underline{u}_h^{(n+1)} = \underline{u}_h^{(n)} + \underline{w}_h,$$

where $\underline{w}_h = (w_i)$ is the solution of the linear system

$$\sum_{j=1}^{i} K_{ij}w_j = r_i \quad \text{for all } i = 1, 2, \ldots, N_h.$$

Hence

$$C_h\underline{w}_h = \underline{f}_h - K_h\underline{u}_h^{(n)}$$

with

$$C_h = \begin{pmatrix} K_{11} & 0 & \cdots & 0 \\ K_{21} & K_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ K_{N_h,1} & \cdots & \cdots & K_{N_h,N_h} \end{pmatrix}$$

This corresponds to one step of the Gauß-Seidel method.

Remark: For the model problem it can be shown that

$$c_1\, h^2\, D_h \le K_h \le c_2\, D_h$$

This implies for the Jacobi method: $\nu_2/\nu_1 = O(h^{-2})$, which is no essential improvement compared to the choice $C_h = I$. The same is true for the Gauß-Seidel method.

Example: Multilevel preconditioner: Starting with an initial mesh of mesh size h_1 a sequence of subdivisions $\mathcal{T}_l$, $l = 1, 2, \ldots, L$ with mesh sizes $h_l = h_{l-1}/2$ is constructed by successively subdividing each sub-interval in two sub-intervals of equal length. The corresponding continuous piecewise linear basis functions are denoted by $\varphi_{l,i}$, $i = 1, 2, \ldots, N_l$, $l = 1, 2, \ldots, L$. This induces a subdivision of V_h into one-dimensional subspaces

$$V_h = \sum_{l=1}^{L} \sum_{i=1}^{N_l} V_{l,i}$$

with

$$V_{l,i} = \operatorname{span}(\varphi_{l,i}).$$

The corresponding additive Schwarz method is called MDS method (multilevel diagonal scaling). This methods corresponds to the Jacobi method, however, not only applied on the finest mesh, but on the whole hierarchy of subdivisions.

Performing one step of the method requires the solution of $N_1 + N_2 + \cdots N_L = (2 - 2^{-(L-1)})N_L = O(N_L)$ one-dimensional problems:

$$a(\varphi_{l,i}, \varphi_{l,i})\, w_{l,i} = \langle F, \varphi_{l,i} \rangle - a(u_L^{(n)}, \varphi_{l,i}).$$

This can be done in $O(N_L)$ operations. For this one uses the fact that basis functions on a mesh can be easily represented by the basis functions of the next finer grid:

$$\varphi_{l-1,i} = \frac{1}{2}\,\varphi_{l,2i-1} + \varphi_{l,2i} + \frac{1}{2}\,\varphi_{l,2i+1}.$$

In summary, only $O(N_L)$ are required for one step of the iteration.

One step of the MDS method can formally be represented as one step of a preconditioned Richardson method with some preconditioner C_h. For the model problem it can be shown that constants $\nu_1 > 0$ and $\nu_2 > 0$ independent of L exist such that

$$\nu_1\, C_h \leq K_h \leq \nu_2\, C_h.$$

So the number of iterations in order to reduce the initial error by a prescribed factor ε is independent of L.

Altogether one obtains a method of optimal complexity $O(N_L)$.

1.5.3 Krylov Subspace Methods

The preconditioned Richardson method can be accelerated by a so-called Krylov subspace method.

First one the case $C_h = I$ is considered. For simplicity we omit the subscript h and the underlining of vectors.

Krylov subspaces:

For the residuals of the Richardson method we have:

$$r^{(m)} = f - Ku^{(m)} = r^{(m-1)} - \tau\, Kr^{(m-1)}$$

Hence

$$r^{(n-1)} \in \operatorname{span}(r^{(0)}, Kr^{(0)}, \ldots, K^{n-1}r^{(0)}) = \mathcal{K}_n(K, r^{(0)})$$

and, therefore,

$$u^{(n)} \in u^{(0)} + \mathcal{K}_n(K, r^{(0)}).$$

The space $\mathcal{K}_n(K, r^{(0)})$ is called the Krylov subspace.

Obviously, we also have

$$u^{(j)} \in u^{(0)} + \mathcal{K}_n(K, r^{(0)}) \quad \text{for all } j = 0, 1, \ldots, n.$$

and

$$\sum_{j=0}^{n} \omega_{nj} u^{(j)} \in u^{(0)} + \mathcal{K}_n(K, r^{(0)})$$

for all coefficients $\omega_{n0}, \omega_{n1}, \ldots, \omega_{nn}$ with $\sum_{j=0}^{n} \omega_{nj} = 1$.

Basic idea of an acceleration technique:

Starting from a sequence $(u^{(n)})$ (here produced by the Richardson method) a one tries to construct a new sequence $(v^{(n)})$ with

$$v^{(n)} = \sum_{j=0}^{n} \omega_{nj} u^{(j)}$$

and proper coefficients $\omega_{n0}, \omega_{n1}, \ldots, \omega_{nn}$ with $\sum_{j=0}^{n} \omega_{nj} = 1$, which converges faster.

The discussion from above shows that we are looking for a sequence $(v^{(n)})$ with

$$v^{(n)} \in u^{(0)} + \mathcal{K}_n(K, r^{(0)})$$

Instead of determining coefficients ω_{nj} we could directly search for iterates $v^{(n)}$ with

$$v^{(n)} \in u^{(0)} + \mathcal{K}_n(K, r^{(0)})$$

which satisfy some selection rule like:

$$\|f - Kv^{(n)}\|_{\ell_2} = \min_{v \in u^{(0)} + \mathcal{K}_n(K, r^{(0)})} \|f - Kv\|_{\ell_2}$$

Another selection rule is

$$(f - Kv^{(n)}, v)_{\ell_2} = 0 \quad \text{for all } v \in \mathcal{K}_n(K, r^{(0)}).$$

For a symmetric and positive definite matrix K this can also be described as the following selection rule:

$$J(v^{(n)}) = \min_{v \in u^{(0)} + \mathcal{K}_n(K, r^{(0)})} J(v).$$

The main task now is to find an efficient way to compute this sequence $(v^{(n)})$.

Since in this discussion the original sequence $(u^{(n)})$ completely disappeared, we will from now on use $(u^{(n)})$ to denote the new (accelerated) sequence of iterates instead of $(v^{(n)})$.

The CG method

Let K be symmetric and positive definite. Hence, u solves the system

$$Ku = f,$$

if and only if

$$J(u) = \min_v J(v) \quad \text{with} \quad J(v) = \frac{1}{2}(Kv, v)_{\ell_2} - (f, v)_{\ell_2}.$$

The direction of the steepest descent with respect to the functional J is given by the negative gradient of J. Here we have:

$$-\operatorname{grad} J(v) = f - Kv$$

This motivates the so-called gradient method:
Initial settings $r^{(0)} = f - Ku^{(0)}$. For $n = 0, 1, 2, \ldots$:

$$\begin{aligned} p^{(n)} &= r^{(n)}, \\ u^{(n+1)} &= u^{(n)} + \alpha^{(n)} p^{(n)} \quad \text{with} \quad \alpha^{(n)} = \frac{(r^{(n)}, p^{(n)})_{\ell_2}}{(Kp^{(n)}, p^{(n)})_{\ell_2}}, \\ r^{(n+1)} &= r^{(n)} - \alpha^{(n)} K p^{(n)}. \end{aligned}$$

The choice of $\alpha^{(n)}$ guarantees (for each search direction $p^{(n)}$ not only for $p^{(n)} = r^{(n)}$) that

$$(r^{(n+1)}, p^{(n)})_{\ell_2} = 0. \tag{1.9}$$

Hence it follows:

$$J(u^{(n+1)}) = \min_{v \in u^{(n)} + \operatorname{span}(p^{(n)})} J(v).$$

The search direction of the gradient method are the residuals. Successive search directions (residuals) are (ℓ_2-)orthogonal.

The CG method uses this search direction only in the first step:

initial settings $r^{(0)} = f - Ku^{(0)}$. For $n = 0, 1, 2, \ldots$:

$$\begin{aligned}
p^{(n)} &= \begin{cases} r^{(0)} & \text{for} \quad n = 0, \\ r^{(n)} + \beta^{(n-1)} p^{(n-1)} \quad \text{with} \quad \beta^{(n-1)} = -\dfrac{(r^{(n)}, Kp^{(n-1)})_{\ell_2}}{(Kp^{(n-1)}, p^{(n-1)})_{\ell_2}} & \text{for} \quad n \geq 1, \end{cases} \\
u^{(n+1)} &= u^{(n)} + \alpha^{(n)} p^{(n)} \quad \text{with} \quad \alpha^{(n)} = \frac{(r^{(n)}, p^{(n)})_{\ell_2}}{(Kp^{(n)}, p^{(n)})_{\ell_2}}, \\
r^{(n+1)} &= r^{(n)} - \alpha^{(n)} Kp^{(n)}.
\end{aligned}$$

The choice of $\beta^{(n-1)}$ guarantees that successive search directions are conjugate, i.e. K-orthogonal:

$$\begin{aligned}
(Kp^{(n-1)}, p^{(n)})_{\ell_2} &= (Kp^{(n-1)}, r^{(n)} + \beta^{(n-1)} p^{(n-1)})_{\ell_2} \\
&= (Kp^{(n-1)}, r^{(n)})_{\ell_2} + \beta^{(n-1)} \, (Kp^{(n-1)}, p^{(n-1)})_{\ell_2} = 0, \qquad (1.10)
\end{aligned}$$

and that successive residuals are (ℓ_2-)orthogonal:

$$\begin{aligned}
(r^{(n+1)}, r^{(n)})_{\ell_2} &= (r^{(n+1)}, p^{(n)} - \beta^{(n-1)} p^{(n-1)})_{\ell_2} \\
&= -\beta^{(n-1)} (r^{(n+1)}, p^{(n-1)})_{\ell_2} = -\beta^{(n-1)} (r^{(n)} - \alpha^{(n)} Kp^{(n)}, p^{(n-1)})_{\ell_2} \\
&= 0. \qquad (1.11)
\end{aligned}$$

The conditions (1.9), (1.10) and (1.11) are not only valid for successive indices, but we have more generally:

Lemma 1.4. *If $r^{(n-1)} \neq 0$ then:*

1. $p^{(n-1)} \neq 0$
2. $\mathcal{K}_n(K, r^{(0)}) = \operatorname{span}(r^{(0)}, r^{(1)}, \ldots, r^{(n-1)}) = \operatorname{span}(p^{(0)}, p^{(1)}, \ldots, p^{(n-1)})$.
3. $(Kp^{(n)}, p^{(j)})_{\ell_2} = 0$ *for all* $j = 0, 1, \ldots, n-1$.
4. $(r^{(n)}, p^{(j)})_{\ell_2} = 0$ *for all* $j = 0, 1, \ldots, n-1$.
5. $(r^{(n)}, r^{(j)})_{\ell_2} = 0$ *for all* $j = 0, 1, \ldots, n-1$.
6. $u^{(n)} \in u^{(0)} + \mathcal{K}_n(K, r^{(0)})$ *and*

$$J(u^{(n)}) = \min_{v \in u^{(0)} + \mathcal{K}_n(K, r^{(0)})} J(v)$$

Proof. Induction with respect to to n.

$n = 1$ trivial.

$n \to n + 1$:

Assume $r^{(n)} \neq 0$.

$r^{(n)} = r^{(n-1)} - \alpha^{(n-1)} K p^{(n-1)} \in \mathcal{K}_{n+1}(K, r^{(0)})$, hence

$$\text{span}(r^{(0)}, r^{(1)}, \ldots, r^{(n)}) \subset \mathcal{K}_{n+1}(K, r^{(0)}).$$

Because of $(r^{(n)}, p^{(j)})_{\ell_2} = 0$ for all $j = 0, 1, \ldots, n-1$ it follows that $r^{(n)} \perp \mathcal{K}_n(K, r^{(0)})$ and, therefore,

$$\text{span}(r^{(0)}, r^{(1)}, \ldots, r^{(n)}) = \mathcal{K}_{n+1}(K, r^{(0)}).$$

Because of $p^{(n)} = r^{(n)} - \beta^{(n-1)} p^{(n-1)}$ it immediately follows

$$\text{span}(p^{(0)}, p^{(1)}, \ldots, p^{(n)}) = \mathcal{K}_{n+1}(K, r^{(0)})$$

and, therefore, $p^{(n)} \neq 0$. This implies statements 1 and 2.

For $j = n$ it was already shown that $(r^{(n+1)}, p^{(j)})_{\ell_2} = 0$. For $j = 0, 1, \ldots, n-1$ it follows that

$$(r^{(n+1)}, p^{(j)})_{\ell_2} = (r^{(n)} - \alpha^{(n)} K p^{(n)}, p^{(j)})_{\ell_2} = (r^{(n)}, p^{(j)})_{\ell_2} - \alpha^{(n)} (K p^{(n)}, p^{(j)})_{\ell_2} = 0.$$

So: $r^{(n+1)} \perp \mathcal{K}_{n+1}(K, r^{(0)})$. This implies the statements 4, 5 and 6.

For $j = n$ it was already shown that $(K p^{(n+1)}, p^{(j)}) = 0$. For $j = 0, 1, \ldots, n-1$ it follows that

$$(K p^{(n+1)}, p^{(j)})_{\ell_2} = (p^{(n+1)}, K p^{(j)})_{\ell_2} = (r^{(n+1)}, K p^{(j)})_{\ell_2} + \beta^{(n)} (p^{(n)}, K p^{(j)})_{\ell_2} = 0.$$

This implies statement 3. □

Convergence analysis

From

$$J(v) = \frac{1}{2} \|v - u\|_K^2 - \frac{1}{2} \|u\|_K^2$$

it follows that

$$\|u^{(n)} - u\|_K \quad = \quad \min_{v \in u^{(0)} + \mathcal{K}_n(K, r^{(0)})} \|v - u\|_K$$

We have:

$$\begin{aligned} v - u &= u^{(0)} + c_1 r^{(0)} + c_2 K r^{(0)} + \ldots + c_n K^{n-1} r^{(0)} - u \\ &= (u^{(0)} - u) + c_1 K (u - u^{(0)}) + c_2 K^2 (u - u^{(0)}) + \ldots + c_n K^n (u - u^{(0)}) \\ &= [I - c_1 K^1 - c_2 K^2 - \ldots - c_n K^n](u^{(0)} - u) \\ &= p(K)(u^{(0)} - u) \end{aligned}$$

where p is a polynomial of degree $\leq n$ with $p(0) = 1$.

Let e_i, $i = 1, 2, \ldots, n$, be a complete system of eigenvectors of K with eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. With

$$u^{(0)} - u = \sum_i \alpha_i e_i$$

it follows that

$$p(K)(u^{(0)} - u) = \sum_i \alpha_i p(\lambda_i) e_i$$

and, therefore,

$$\begin{aligned} \|p(K)(u^{(0)} - u)\|_K^2 &= (p(K)(u^{(0)} - u), p(K)(u^{(0)} - u))_K = \sum_i \alpha_i^2 \lambda_i p(\lambda_i)^2 \\ &\leq \max_i p(\lambda_i)^2 \sum_i \lambda_i \alpha_i^2 = \left[\max_i |p(\lambda_i)|\right]^2 \|u^{(0)} - u\|_K^2. \end{aligned}$$

Hence

$$\|u^{(n)} - u\|_K \leq \left[\min_{\substack{p \text{ polynomial} \\ \deg p \leq n,\ p(0)=1}} \max_i |p(\lambda_i)|\right] \|u^{(0)} - u\|_K.$$

Theorem 1.10. *We have:*

$$\|u^{(n)} - u\|_K \leq \frac{2q^n}{1 + q^{2n}} \|u^{(0)} - u\|_K \leq 2q^n \|u^{(0)} - u\|_K$$

with

$$q = \frac{\sqrt{\kappa(K)} - 1}{\sqrt{\kappa(K)} + 1}.$$

Proof.

$$\min_{\substack{p \text{ polynomial} \\ \deg p \leq n,\ p(0)=1}} \max_i |p(\lambda_i)| \leq \min_{\substack{p \text{ polynomial} \\ \deg p \leq n,\ p(0)=1}} \max_{\lambda \in [\lambda_1, \lambda_n]} |p(\lambda)|$$

Let T_n be the n-th Chebychev polynomial:

$$T_n(x) = \frac{1}{2}[(x + \sqrt{x^2 - 1})^n + (x - \sqrt{x^2 - 1})^n]$$

For

$$p(\lambda) = \frac{T_n((\lambda_n + \lambda_1 - 2\lambda)/(\lambda_n - \lambda_1))}{T_n((\lambda_n + \lambda_1)/(\lambda_n - \lambda_1))}$$

it follows that

$$\max_{\lambda \in [a,b]} |(p(\lambda)| = \frac{1}{T_n((\lambda_n + \lambda_1)/(\lambda_n - \lambda_1))} = \frac{1}{T_n((\kappa + 1)/(\kappa - 1))}$$

with $\kappa = \lambda_n/\lambda_1$.

$$\frac{\kappa+1}{\kappa-1} + \sqrt{\left(\frac{\kappa+1}{\kappa-1}\right)^2 - 1} = \frac{\kappa+1+2\sqrt{\kappa}}{\kappa-1} = \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}$$

and

$$\frac{\kappa+1}{\kappa-1} - \sqrt{\left(\frac{\kappa+1}{\kappa-1}\right)^2 - 1} = \frac{\kappa+1-2\sqrt{\kappa}}{\kappa-1} = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}.$$

Hence

$$\frac{1}{T_n((\kappa+1)/(\kappa-1))} = \frac{2}{q^n + q^{-n}} = \frac{2q^n}{1+q^{2n}} \leq 2q^n$$

with

$$q = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}.$$

□

Remark:

1. Because of

   $$(r^{(n)}, p^{(n)})_{\ell_2} = (r^{(n)}, r^{(n)} + \beta^{(n-1}p^{(n-1)})_{\ell_2} = (r^{(n)}, r^{(n)})_{\ell_2}$$

   we also have the representation

   $$\alpha^{(n)} = \frac{(r^{(n)}, r^{(n)})_{\ell_2}}{(Kp^{(n)}, p^{(n)})_{\ell_2}}.$$

2. Because of

   $$(r^{(n)}, Kp^{(n-1)})_{\ell_2} = -\frac{1}{\alpha^{(n-1)}}(r^{(n)}, r^{(n)} - r^{(n-1)})_{\ell_2} = -\frac{1}{\alpha^{(n-1)}}(r^{(n)}, r^{(n)})_{\ell_2}$$

   we also have the representation

   $$\beta^{(n-1)} = \frac{(r^{(n)}, r^{(n)})_{\ell_2}}{(r^{(n-1)}, r^{(n-1)})_{\ell_2}}.$$

PCG-method

By the substitutions $(w,v)_{\ell_2} \to (w,v)_C$, $K \to C^{-1}K$ and $f \to C^{-1}f$ one obtains:

Initial settings $s^{(0)} = C^{-1}(f - Ku^{(0)})$. For $n = 0, 1, 2, \ldots$:

$$\begin{aligned}
p^{(n)} &= \begin{cases} s^{(0)} & \text{for} \quad n = 0, \\ s^{(n)} + \beta^{(n-1)}p^{(n-1)} \quad \text{with} \quad \beta^{(n-1)} = \dfrac{(s^{(n)}, s^{(n)})_C}{(s^{(n-1)}, s^{(n-1)})_C} & \text{for} \quad n \geq 1, \end{cases} \\
u^{(n+1)} &= u^{(n)} + \alpha^{(n)}p^{(n)} \quad \text{with} \quad \alpha^{(n)} = \frac{(s^{(n)}, s^{(n)})_C}{(C^{-1}Kp^{(n)}, p^{(n)})_C}, \\
s^{(n+1)} &= s^{(n)} - \alpha^{(n)}C^{-1}Kp^{(n)}.
\end{aligned}$$

Hence:

Initial settings $r^{(0)} = f - Ku^{(0)}$. For $n = 0, 1, 2, \ldots$:

$$\begin{aligned}
s^{(n)} &= C^{-1}r^{(n)} \\
p^{(n)} &= \begin{cases} s^{(0)} & \text{for} \quad n = 0, \\ s^{(n)} + \beta^{(n-1)}p^{(n-1)} \quad \text{with} \quad \beta^{(n-1)} = \dfrac{(r^{(n)}, s^{(n)})_{\ell_2}}{(r^{(n-1)}, s^{(n-1)})_{\ell_2}} & \text{for} \quad n \geq 1, \end{cases} \\
q^{(n)} &= Kp^{(n)} \\
u^{(n+1)} &= u^{(n)} + \alpha^{(n)}p^{(n)} \quad \text{with} \quad \alpha^{(n)} = \frac{(r^{(n)}, s^{(n)})_{\ell_2}}{(q^{(n)}, p^{(n)})_{\ell_2}}, \\
r^{(n+1)} &= r^{(n)} - \alpha^{(n)}q^{(n)}.
\end{aligned}$$

Symmetry:

$$(C^{-1}Kv, w)_C = (Kv, w)_{\ell_2} = (v, Kw)_{\ell_2} = (Cv, C^{-1}Kw)_{\ell_2} = (v, C^{-1}Kw)_C.$$

Energy functional:

$$\frac{1}{2}(C^{-1}Kv, v)_C - (C^{-1}f, v)_C = \frac{1}{2}(CC^{-1}Kv, v)_{\ell_2} - (CC^{-1}f, v)_{\ell_2} = \frac{1}{2}(Kv, v)_{\ell_2} - (f, v)_{\ell_2}.$$

Therefore, one immediately obtains the analogous convergence properties with the replacement $\kappa(K) \to \kappa(C^{-1}K)$.

GMRES

If K is symmetric and positive definite, the CG method produces a ℓ_2-orthogonal basis of the Krylov subspace $\mathcal{K}_n(K, r^{(0)})$, namely the residuals: $r^{(0)}, r^{(1)}, \ldots, r^{(n-1)}$.

A general method for constructing a ℓ_2-orthogonal basis of the Krylov subspaces $\mathcal{K}_n(K, r^{(0)})$ is the so-called Arnoldi method:

$$\begin{aligned}
v^{(1)} &= r^{(0)}/\|r^{(0)}\|_{\ell_2}, \\
w^{(i)} &= Kv^{(i)}, \\
w_\perp^{(i)} &= w^{(i)} - \sum_{j=1}^{i} h_{ji}\, v^{(j)} \quad \text{with} \quad h_{ji} = (w^{(i)}, v^{(j)}), \\
v^{(i+1)} &= w_\perp^{(i)}/h_{i+1,i} \quad \text{with} \quad h_{i+1,i} = \|w_\perp^{(i)}\|_{\ell_2}.
\end{aligned}$$

Consider now the problem of finding an approximate solution

$$u^{(n)} = u^{(0)} + \sum_{i=1}^{n} y_i v^{(i)} \in u^{(0)} + \mathcal{K}_n(K, r^{(0)})$$

which minimizes the ℓ_2-norm of the residual:

$$\|f - Ku^{(n)}\|_{\ell_2} = \min_{v \in u^{(0)} + \mathcal{K}_n(K, r^{(0)})} \|f - Kv\|_{\ell_2}.$$

We have:

$$h_{i+1,i}\, v^{(i+1)} = Kv^{(i)} - \sum_{j=1}^{i} h_{ji} v^{(j)},$$

so

$$Kv^{(i)} = \sum_{j=1}^{i+1} h_{ji}\, v^{(j)}.$$

Hence

$$\begin{aligned} f - Ku^{(n)} &= r^{(0)} - \sum_{i=1}^{n} y_i \sum_{j=1}^{i+1} h_{ji}\, v^{(j)} = \|r^{(0)}\|_{\ell_2}\, v^{(1)} - \sum_{i=1}^{n} y_i \sum_{j=1}^{i+1} h_{ji}\, v^{(j)} \\ &= \left[\|r^{(0)}\|_{\ell_2} - \sum_{i=1}^{n} h_{1i} y_i\right] v^{(1)} - \sum_{j=2}^{n+1} \left[\sum_{i=j-1}^{n} h_{ji} y_i\right] v^{(j)}. \end{aligned}$$

Therefore

$$\|f - Ku^{(n)}\| = \|\|r^{(0)}\| e_1 - H_n y^{(n)}\|$$

with

$$H_n = \begin{pmatrix} h_{11} & h_{12} & \cdots & \cdots & h_{1n} \\ h_{21} & h_{22} & \ddots & & \vdots \\ 0 & h_{22} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & h_{n-1,n} \\ 0 & \cdots & 0 & h_{n,n-1} & h_{nn} \\ \hline 0 & \cdots & \cdots & 0 & h_{n+1,n} \end{pmatrix} \qquad e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ \vdots \\ 0 \\ \hline 0 \end{pmatrix} \qquad y^{(n)} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{pmatrix}.$$

By a sequence of simple orthogonal matrices (Givens rotations) $H^{(n)}$ can be transformed to a triangular matrix:

$$J_n \cdots J_2 J_1 H_n = \begin{pmatrix} R_n \\ 0 \end{pmatrix}.$$

This implies

$$\left\|\|r^{(0)}\|_{\ell_2}\, e_1 - H_n y^{(n)}\right\|_{\ell_2} = \left\|\|r^{(0)}\|_{\ell_2}\, J_n \cdots J_2 J_1 e_1 - \begin{pmatrix} R_n y^{(n)} \\ 0 \end{pmatrix}\right\|_{\ell_2}.$$

With the notation

$$s^{(n)} = \|r^{(0)}\|_{\ell_2}\, J_n \cdots J_2 J_1 e_1 = \begin{pmatrix} \tilde{s}^{(n)} \\ \hline \tilde{s}_{n+1} \end{pmatrix}$$

one finally obtains:

$$\begin{aligned}
\min_{v\in\mathcal{K}_n(K,r^{(0)})} \|f-Kv\|_{\ell_2}^2 &= \min_{y\in\mathbb{R}^n} \left\| \|r^{(0)}\|_{\ell_2}\, e_1 - H_n y^{(n)} \right\|_{\ell_2}^2 \\
&= \min_{y\in\mathbb{R}^n} \left\| \|r^{(0)}\|_{\ell_2}\, J_n\cdots J_2 J_1 e_1 - \begin{pmatrix} R_n y^{(n)} \\ 0 \end{pmatrix} \right\|_{\ell_2}^2 \\
&= \min_{y} \|\tilde{s}^{(n)} - R_n y^{(n)}\|_{\ell_2}^2 + |\tilde{s}_{n+1}|^2 \\
&= |\tilde{s}_{n+1}|^2
\end{aligned}$$

So, the minimal value of $\|f-Kv\|_{\ell_2}^2$ can be computed by using H_n, the Givens rotations J_1, J_2, ..., J_n and $s^{(n)}$.

These quantities can be efficiently computed from the previous iteration step: First one calculates the n-th column of H_n:

$$(h_{1n}, h_{2n}, \ldots, h_{nn}, h_{n+1,n})^T$$

The already constructed Givens rotations $J_1, J_2, \ldots, J_{n-1}$ are applied to this column:

$$J_{n-1}\cdots J_2 J_1 (h_{1n}, h_{2n}, \ldots, h_{nn}, h_{n+1,n})^T$$

Subsequently, the Givens rotation J_n is constructed such that the $(n+1)$-th component of this vector vanishes. Therefore, one obtains the n-th column of R_n:

$$(r_{1n}, r_{2n}, \ldots, r_{nn}, 0)^T = J_n J_{n-1}\cdots J_2 J_1 (h_{1n}, h_{2n}, \ldots, h_{nn}, h_{n+1,n})^T$$

and

$$s^{(n)} = J_n s^{(n-1)}.$$

If the minimal value falls below a prescribed bound, the solution of

$$R_n y^{(n)} = \tilde{s}^{(n)}$$

and the approximate solution

$$u^{(n)} = u^{(0)} + \sum_{i=1}^{n} y_i^{(n)} v^{(i)}.$$

are calculated.

Remark: Because of the large amount of storage for GMRES the method is usually restarted after a fixed number m of steps: GMRES(m)

1.6 Boundary Value Problems for Nonlinear Elliptic Differential Equations

Classical formulation:

One-dimensional problems:

$$-\Big(q_1(x,u(x),u'(x))\Big)' + q_0(x,u(x),u'(x)) = f(x) \quad x \in (0,1).$$

Special case linear problems:

$$q_1(x,\xi_0,\xi_1) = a(x)\xi_1, \quad q_0(x,\xi_0,\xi_1) = b(x)\xi_1 + c(x)\xi_0.$$

Higher dimensional problems:

$$-\sum_{i=1}^{d} \frac{\partial}{\partial x_i}\Big(q_i(x,u(x),\operatorname{grad} u(x))\Big) + q_0(x,u(x),\operatorname{grad} u(x)) = f(x) \quad x \in \Omega.$$

Special case linear problems:

$$q_i(x,\xi_0,\xi) = \sum_{j=1}^{d} a_{ij}(x)\xi_j, \quad i = 1,\ldots,n, \qquad q_0(x,\xi_0,\xi) = \sum_{j=1}^{d} b_j(x)\xi_j + c(x)\xi_0$$

with $\xi = (\xi_1,\xi_2,\ldots,\xi_d)^T$.

With

$$q(x,\xi_0,\xi) = (q_i(x,\xi_0,\xi))_{i=1,\ldots,d}$$

a compact representation of the differential equation is obtained:

$$-\operatorname{div}\Big(q(x,u(x),\operatorname{grad} u(x))\Big) + q_0(x,u(x),\operatorname{grad} u(x)) = f(x).$$

For the linear case one obtains

$$q(x,\xi_0,\xi) = A(x)\xi, \qquad q_0(x,\xi_0,\xi) = b(x)\cdot\xi + c(x)\xi_0.$$

Boundary conditions:

1. Dirichlet boundary conditions (boundary condition of the first kind):

$$u(x) = g_D(x) \quad x \in \Gamma_D.$$

2. Neumann boundary conditions (boundary condition of the second kind):

$$q(x,u(x),\operatorname{grad} u(x))\cdot n(x) = g_N(x) \quad x \in \Gamma_N.$$

Variational formulation

Let v be a test function with $v(x) = 0$ for $x \in \Gamma_D$. Then one obtains from the classical formulation:

Find u with $u(x) = g_D(x)$ for $x \in \Gamma_D$ such that

$$\int_\Omega \Big[q(x, u(x), \operatorname{grad} u(x)) \cdot \operatorname{grad} v(x) + q_0(x, u(x), \operatorname{grad} u(x)) v(x) \Big] \, dx$$
$$= \int_\Omega f(x) v(x) \, dx + \int_{\Gamma_N} g_N(x) v(x) \, ds$$

for all v with $v(x) = 0$ for $x \in \Gamma_D$. Or, in short:

$$a(u, v) = \langle F, v \rangle.$$

IMPORTANT: $a(w, v)$ is linear in v, but not necessarily linear in w. By

$$\langle A(w), v \rangle = a(w, v)$$

a nonlinear operator is defined. Then the problem can be written as an operator equation:

$$A(u) = F.$$

Function spaces

Problem: It is not trivial to show the operator is well-defined. Hilbert spaces are not always appropriate, one needs proper Banach spaces, like, e.g., the Sobolev spaces

$$V = W^{k,p}(\Omega) = \{ v \in L^p(\Omega) : D^\alpha v \in L^p(\Omega) \ |\alpha| \leq k \}.$$

Under appropriate assumptions on $a(x, \xi_0, \xi)$ and $a_0(x, \xi_0, \xi)$ it can be shown that $a : V \times V \longrightarrow \mathbb{R}$ and accordingly the nonlinear operators $A : V \longrightarrow V^*$ are well-defined (Nemyzki operators).

Properties of a and A:

As a possible generalization of the ellipticity (coerciveness) of bilinear forms and the corresponding linear operators the following concept is introduced:

Definition 1.1. *Let V be a normed space and V^* its dual space. An operator $A : V \longrightarrow V^*$ is called strongly monotone, if and only if there is a constant $\mu_1 > 0$ such that*

$$\langle A(w) - A(v), w - v \rangle \geq \mu_1 \|w - v\|^2 \quad \textit{for all } w, v \in V.$$

As a possible generalization of the boundedness (continuity) of bilinear forms and the corresponding linear operators the following concept is introduced:

Definition 1.2. *Let V be a normed space and V^* its dual space. An operator $A : V \longrightarrow V^*$ is called Lipschitz continuous, if and only if there is a constant $\mu_2 > 0$ such that*

$$\|A(w) - A(v)\| \leq \mu_2 \|w - v\| \quad \textit{for all } w, v \in V.$$

Examples of strongly monotone and Lipschitz continuous operators: Let

$$q(x, \xi_0, \xi) = \alpha(\|\xi\|_{\ell_2})\, \xi, \quad q_0(x, \xi_0, \xi) \equiv 0.$$

For simplicity only the case of homogenous Dirichlet boundary conditions are discussed:
So we consider boundary value problems of the following form

$$\begin{aligned} -\operatorname{div}\Big(\alpha(\|\operatorname{grad} u(x)\|_{\ell_2}) \operatorname{grad} u(x)\Big) &= f(x) \quad x \in \Omega, \\ u(x) &= 0 \quad x \in \Gamma. \end{aligned}$$

Special case (one-dimensional model problem):

$$\begin{aligned} -\Big(\alpha(|u'(x)|)u'(x)\Big)' &= f(x) \quad x \in \Omega, \\ u(x) &= 0 \quad x \in \Gamma. \end{aligned}$$

We have:

Theorem 1.11. *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain. Assume that the following properties are satisfied for the function $\alpha : [0, \infty) \longrightarrow \mathbb{R}$:*

1. *α is continuous.*
2. *There is a constant $M > 0$ with*

$$|\alpha(t)| \leq M \quad \textit{for all } t \geq 0.$$

Then the operator $A : V \longrightarrow V^*$ is well-defined for $V = H^1(\Omega)$. Additionally, we have

1. *If there is a constant M with*

$$|\alpha(t)t - \alpha(s)s| \leq M|t - s| \quad \textit{for all } s, t \geq 0,$$

   then A is Lipschitz continuous in $H^1(\Omega)$.

2. *If there is a constants $m > 0$ with*

$$\alpha(t)t - \alpha(s)s \geq m(t - s) \quad \textit{for all } t \geq s,$$

   then A is strongly monotone in $H_0^1(\Omega)$.

Proof. $\alpha(\|\operatorname{grad} u(.)\|_{\ell_2})$ is measurable. Since α is bounded we have

$$\begin{aligned}\|\alpha(\|\operatorname{grad} u(.)\|_{\ell_2})\operatorname{grad} u(.)\|_0^2 &= \int_\Omega |\alpha(\|\operatorname{grad} u(x)\|_{\ell_2})|^2 \|\operatorname{grad} u(x)\|_{\ell_2}^2 \, dx \\ &\le M^2 \int_\Omega \|\operatorname{grad} u(x)\|_{\ell_2}^2 \, dx < \infty.\end{aligned}$$

So: $\alpha(\|\operatorname{grad} u(.)\|_{\ell_2})\operatorname{grad} u(.) \in [L^2(\Omega)]^d$, which implies that

$$\int_\Omega \alpha(\|\operatorname{grad} u(x)\|_{\ell_2})\operatorname{grad} u(x) \cdot \operatorname{grad} v(x) \, dx$$

is well-defined.

We have

$$\begin{aligned}&|\langle A(w) - A(v), u\rangle| \\ &= |\int_\Omega \Big(\alpha(\|\operatorname{grad} w(x)\|_{\ell_2})\operatorname{grad} w(x) - \alpha(\|\operatorname{grad} v(x)\|_{\ell_2})\operatorname{grad} v(x)\Big) \cdot \operatorname{grad} u(x) \, dx| \\ &= |\int_\Omega \Big(\alpha(\|\operatorname{grad} w(x)\|_{\ell_2})\big(\operatorname{grad} w(x) - \operatorname{grad} v(x)\big)\Big) \cdot \operatorname{grad} u(x) \, dx \\ &\quad + \int_\Omega \Big(\alpha(\|\operatorname{grad} w(x)\|_{\ell_2}) - \alpha(\|\operatorname{grad} v(x)\|_{\ell_2})\Big)\operatorname{grad} v(x) \cdot \operatorname{grad} u(x) \, dx| \\ &\le M \|\operatorname{grad} w - \operatorname{grad} v\|_0 \|\operatorname{grad} u\|_0 \\ &\quad + \int_\Omega \Big|\alpha(\|\operatorname{grad} w(x)\|_{\ell_2}) - \alpha(\|\operatorname{grad} v(x)\|_{\ell_2})\Big| \|\operatorname{grad} v(x)\|_{\ell_2} \|\operatorname{grad} u(x)\|_{\ell_2} \, dx \\ &= M \|\operatorname{grad} w - \operatorname{grad} v\|_0 \|\operatorname{grad} u\|_0 \\ &\quad + \int_\Omega \Big|\alpha(\|\operatorname{grad} w(x)\|_{\ell_2})\|\operatorname{grad} v(x)\|_{\ell_2} - \alpha(\|\operatorname{grad} v(x)\|_{\ell_2})\|\operatorname{grad} v(x)\|_{\ell_2}\Big| \\ &\qquad \|\operatorname{grad} u(x)\|_{\ell_2} \, dx \\ &= M \|\operatorname{grad} w - \operatorname{grad} v\|_0 \|\operatorname{grad} u\|_0 \\ &\quad + \int_\Omega \Big|\alpha(\|\operatorname{grad} w(x)\|_{\ell_2})\|\operatorname{grad} w(x)\|_{\ell_2} - \alpha(\|\operatorname{grad} v(x)\|_{\ell_2})\|\operatorname{grad} v(x)\|_{\ell_2} \\ &\qquad + \alpha(\|\operatorname{grad} w(x)\|_{\ell_2})\big(\|\operatorname{grad} v(x)\|_{\ell_2} - \|\operatorname{grad} w(x)\|_{\ell_2}\big)\Big| \, \|\operatorname{grad} u(x)\|_{\ell_2} \, dx \\ &\le M \|\operatorname{grad} w - \operatorname{grad} v\|_0 \|\operatorname{grad} u\|_0 \\ &\quad + 2M \int_\Omega \Big|\|\operatorname{grad} v(x)\|_{\ell_2} - \|\operatorname{grad} w(x)\|_{\ell_2}\Big| \, \|\operatorname{grad} u(x)\|_{\ell_2} \, dx \\ &\le M \|\operatorname{grad} w - \operatorname{grad} v\|_0 \|\operatorname{grad} u\|_0 \\ &\quad + 2M \int_\Omega \|\operatorname{grad} v(x) - \operatorname{grad} w(x)\|_{\ell_2} \|\operatorname{grad} u(x)\|_{\ell_2} \, dx \\ &\le 3M \|\operatorname{grad} w - \operatorname{grad} v\|_0 \|\operatorname{grad} u\|_0 = 3M \, |w - v|_1 \, |u|_1 \le 3M \, \|w - v\|_1 \, \|u\|_1.\end{aligned}$$

Hence

$$\|A(w) - A(v)\| = \sup_{u \neq 0} \frac{|\langle A(w) - A(v), u\rangle|}{\|u\|_1} \le 3M \, \|w - v\|_1.$$

With

$$\alpha(t) = m + \alpha_2(t)$$

it follows that

$$\alpha_2(t)t - \alpha_2(s)s \geq 0 \quad \text{for all } s, t \geq 0.$$

So

$$\begin{aligned}
&\langle A(w) - A(v), w - v\rangle \\
&= \int_\Omega \Big(\alpha(\|\operatorname{grad} w(x)\|_{\ell_2})\ \operatorname{grad} w(x) - \alpha(\|\operatorname{grad} v(x)\|_{\ell_2})\ \operatorname{grad} v(x)\Big) \\
&\qquad \cdot\big(\operatorname{grad} w(x) - \operatorname{grad} v(x)\big)\ dx \\
&= m \int_\Omega \|\operatorname{grad} w(x) - \operatorname{grad} v(x)\|_{\ell_2}^2\ dx \\
&\quad + \int_\Omega \Big(\alpha_2(\|\operatorname{grad} w(x)\|_{\ell_2})\ \operatorname{grad} w(x) - \alpha_2(\|\operatorname{grad} v(x)\|_{\ell_2})\ \operatorname{grad} v(x)\Big) \\
&\qquad \cdot\big(\operatorname{grad} w(x) - \operatorname{grad} v(x)\big)\ dx \\
&\geq m \int_\Omega \|\operatorname{grad}(w(x) - v(x))\|_{\ell_2}^2\ dx \\
&\quad + \int_\Omega \Big(\alpha_2(\|\operatorname{grad} w(x)\|_{\ell_2})\left(\|\operatorname{grad} w(x)\|_{\ell_2}^2 - \|\operatorname{grad} w(x)\|_{\ell_2}\|\operatorname{grad} v(x)\|_{\ell_2}\right) \\
&\qquad + \alpha_2(\|\operatorname{grad} v(x)\|_{\ell_2})\left(\|\operatorname{grad} v(x)\|_{\ell_2}^2 - \|\operatorname{grad} w(x)\|_{\ell_2}\|\operatorname{grad} v(x)\|_{\ell_2}\right)\Big)\ dx \\
&= m \int_\Omega \|\operatorname{grad}(w(x) - v(x))\|_{\ell_2}^2\ dx \\
&\quad + \int_\Omega \Big(\alpha_2(\|\operatorname{grad} w(x)\|_{\ell_2})\|\operatorname{grad} w(x)\|_{\ell_2} - \alpha(\|\operatorname{grad} v(x)\|_{\ell_2})\|\operatorname{grad} v(x)\|_{\ell_2}\Big) \\
&\qquad \Big(\|\operatorname{grad} w(x)\|_{\ell_2} - \|\operatorname{grad} v(x)\|_{\ell_2}\Big)\ dx \\
&\geq m \int_\Omega \|\operatorname{grad}(w(x) - v(x))\|_{\ell_2}^2\ dx = m\,|w - v|_1^2 \geq \frac{m}{c_F^2 + 1}\,\|w - v\|_1^2.
\end{aligned}$$

□

The Lax-Milgram Theorem

Theorem 1.12 (Lax-Milgram). *Let V be a Hilbert space, $F \in V^*$ and $A : V \longrightarrow V^*$ a strongly monotone and Lipschitz continuous operator. Then there exists a unique solution $u \in V$ of the equation $A(u) = F$.*

Proof. Let $\mathcal{J} \colon V^* \longrightarrow V$ denote the Riesz isomorphism. With $\widetilde{A}(u) = \mathcal{J}A(u)$ and $\widetilde{f} = \mathcal{J}F$ one obtains the equivalent problem:

$$\widetilde{A}(u) = \widetilde{f},$$

which can also be written in fixed point form

$$u = u - \tau\,(\widetilde{A}(u) - \widetilde{f}) \equiv K_\tau(u) + g_\tau.$$

In the following it will be shown that K_τ is contractive for a proper choice of τ:

$$\begin{aligned}
\|K_\tau(w) - K_\tau(v)\|^2 &= (K_\tau(w) - K_\tau(v), K_\tau(w) - K_\tau(v)) \\
&= ((w-v) - \tau\,(\widetilde{A}(w) - \widetilde{A}(v)), (w-v) - \tau\,(\widetilde{A}(w) - \widetilde{A}(v))) \\
&= (w-v, w-v) - 2\tau\,(\widetilde{A}(w) - \widetilde{A}(v), w-v) \\
&\quad + \tau^2\,(\widetilde{A}(w) - \widetilde{A}(v), \widetilde{A}(w) - \widetilde{A}(v)) \\
&= \|w-v\|^2 - 2\tau\,\langle A(w) - A(v), w-v\rangle + \tau^2\,\|A(w) - A(v)\|^2 \\
&\leq (1 - 2\mu_1\tau + \mu_2^2\tau^2)\|w-v\|^2.
\end{aligned}$$

The rest follows like in the linear case. □

Discretization:

Same construction:

$$a(g_h + \sum_{j=1}^{N_h} u_j\varphi_j, \varphi_i) = \langle F, \varphi_i\rangle \quad \text{for all } i = 1, 2, \ldots, N_h.$$

Hence

$$\underline{K}_h(\underline{u}_h) = \underline{f}_h$$

with the nonlinear map $\underline{K}_h : \mathbb{R}^{N_h} \longrightarrow \mathbb{R}^{N_h}$, given by

$$\underline{K}_h(\underline{w}_h) = (\underline{K}_i(\underline{w}_h))_{i=1,2,\ldots,N_h} \quad \text{with } \underline{K}_i(\underline{w}_h) = a(g_h + \sum_{j=1}^{N_h} w_j\varphi_j, \varphi_i)$$

Remark: In the linear (homogenous) case we have:

$$\underline{K}_h(\underline{w}_h) = K_h\underline{w}_h.$$

Iterative methods:

From the constructive proof of the Lax-Milgram Theorem we obtain the following fixed point iteration:

$$\underline{u}_h^{(n+1)} = \underline{u}_h^{(n)} + \tau\,B_h^{-1}(\underline{f}_h - \underline{K}_h(\underline{u}_h^{(n)})),$$

which converges q-linear with convergence rate $q = \sqrt{1 - (\mu_1/\mu_2)^2}$.

More generally, we consider iterative methods of the form:

$$\underline{u}_h^{(n+1)} = \underline{u}_h^{(n)} + \tau\,C_h^{-1}(\underline{f}_h - \underline{K}_h(\underline{u}_h^{(n)})).$$

In the following, we use the simplified notation:

$$u^{(n+1)} = u^{(n)} + \tau\, C^{-1}(f - K(u^{(n)})).$$

The conditions

$$(K(w) - K(v), w - v)_{\ell_2} \geq \nu_1 (C[w - v], w - v)_{\ell_2}$$

and

$$(C^{-1}[K(w) - K(v)], K(w) - K(v))_{\ell_2}^{1/2} \leq \nu_2 \, (C[w - v], w - v)_{\ell_2}^{1/2}$$

guarantee the q-linear convergence with convergence rate $q = \sqrt{1 - (\nu_1/\nu_2)^2}$ by the Lax-Milgram Theorem.

In the neighborhood of the exact solution u we have:

$$K(w) - K(v) \approx K'(u)[(w - u) - (v - u)] = K'(u)[w - v].$$

This motivates the choice $C = K'(u)$, which guarantees that $\nu_1 \approx 1$ and $\nu_2 \approx 1$ and, consequently, $q \approx 0$ (super-linear convergence).

However, the exact solution u is not available.

Remedy: Preconditioner C or, more generally, a variable preconditioner C_n, which approximates $K'(u)$ well:

$$u^{(n+1)} = u^{(n)} + \tau\, C_n^{-1}(f - K(u^{(n)})).$$

Important examples:

$C = K'(u^{(0)})$, $\tau = 1$: simplified Newton method:

$$u^{(n+1)} = u^{(n)} + K'(u^{(0)})^{-1}(f - K(u^{(n)})).$$

$C_n = K'(u^{(n)})$, $\tau = 1$: Newton's method.

$$u^{(n+1)} = u^{(n)} + K'(u^{(n)})^{-1}(f - K(u^{(n)})).$$

With $F(u) = K(u) - f$ the equation reads $F(u) = 0$ and Newton's method has the form:

$$u^{(n+1)} = u^{(n)} - F'(u^{(n)})^{-1} F(u^{(n)}).$$

1.6.1 Newton's method

Algorithm:

1. Compute $r = f - K(u^{(n)})$ and the Jacobian matrix $C_n = K'(u^{(n)})$.
2. Solve $C_n w = r$.
3. Compute $u^{(n+1)} = u^{(n)} + w$.

In order to apply the Lax-Milgram Theorem (in the case of a fixed preconditioner), the preconditioner would have to be symmetric and positive definite, a condition, which is too restrictive in many cases. Convergence analysis can also be done without this restriction:

Theorem 1.13 (Local convergence of Newton's method). *Let $F : D \longrightarrow V$, $D \subset V$ open, be differentiable in D and let $u \in D$ be a zero of F with non-singular Jacobian $F'(u)$. Then we have:*

1. *If F' is continuous at u, then Newton's method*

$$u^{(n+1)} = u^{(n)} - F'(u^{(n)})^{-1} F(u^{(n)})$$

   converges locally and q-super-linearly.

2. *If there is a constant ω with*

$$\|F'(u)^{-1}\big(F'(v) - F'(u)\big)\| \leq \omega \|v - u\| \quad \text{for all } v \in D, \tag{1.12}$$

   then Newton's method converges q-quadratically.

Proof. For

$$G(v) = v - F'(v)^{-1} F(v)$$

we have

$$\begin{aligned} G(v) - u &= v - F'(v)^{-1}F(v) - u \\ &= F'(v)^{-1}\big[F(u) - F(v) - F'(v)(u - v)\big] \\ &= \int_0^1 F'(v)^{-1}\big[F'(v + t(u - v)) - F'(v)\big](u - v)\, dt. \end{aligned}$$

Since F' is continuous at the point u, the q-super-linear convergence follows.

From

$$F'(v)^{-1}F'(u) = \Big[I - \big(I - F'(u)^{-1}F'(v)\big)\Big]^{-1}$$

it follows

$$\|F'(v)^{-1}F'(u)\| \leq \frac{1}{1 - \omega \|v - u\|}.$$

From

$$\begin{aligned} &F'(u)^{-1}\big[F'(v + t(u - v)) - F'(v)\big] \\ &= F'(u)^{-1}\big[F'(v + t(u - v)) - F'(u)\big] - F'(u)^{-1}\big[F'(v) - F'(u)\big] \end{aligned}$$

it follows

$$\|F'(u)^{-1}\big[F'(v + t(u - v)) - F'(v)\big]\| \leq \omega(2 - t)\|v - u\|.$$

Therefore

$$\|G(v) - u\| \leq \frac{3\omega \|v - u\|^2}{2(1 - \omega \|v - u\|)},$$

which immediately implies the q-quadratic convergence. □

Computation of the Jacobian matrix

Starting point:

$$\underline{K}'(\underline{u})\underline{w} = \lim_{t\to 0}\frac{1}{t}\Big(\underline{K}(\underline{u}+t\underline{w}) - \underline{K}(\underline{u})\Big)$$

So

$$\begin{aligned}
[\underline{K}'(u)w]_i &= \lim_{t\to 0}\frac{1}{t}\Big([\underline{K}(\underline{u}+t\underline{w})]_i - [\underline{K}(\underline{u})]_i\Big)\\
&= \lim_{t\to 0}\frac{1}{t}\Big(a(u+tw,\varphi_i) - a(u,\varphi_i)\Big)\\
&= \lim_{t\to 0}\frac{1}{t}\langle A(u+tw) - A(u),\varphi_i\rangle = \langle A'(u)w,\varphi_i\rangle = a'(u)(w,\varphi_i)\\
&= \sum_j a'(u)(\varphi_j,\varphi_i)w_j
\end{aligned}$$

with $w(x) = \sum_i w_i\varphi_i(x)$.

Here, $A'(u)$ is the (Gâteaux-)derivative of the nonlinear operators $A : V \longrightarrow V^*$ at the point u: $A'(u) : V \longrightarrow V^*$ with corresponding bilinear form $a'(u)$.

So: The Jacobian matrix $\underline{K}'(u)$ of the nonlinear function $\underline{K} : \mathbb{R}^N \longrightarrow \mathbb{R}^N$ is obtained as stiffness matrix associated to that bilinear form $a'(u)$, which is obtained by linearizing a at the point u:

$$\underline{K}'(u)_{ij} = a'(u)(\varphi_j,\varphi_i).$$

Example: For the one-dimensional model problem we obtain

$$a(u,w) = \int_\Omega \alpha(|u'(x)|)u'(x)w'(x)\ dx.$$

Hence

$$\begin{aligned}
&\lim_{t\to 0}\frac{1}{t}\Big(a(u+tw,\varphi) - a(u,\varphi)\Big)\\
&= \lim_{t\to 0}\frac{1}{t}\int_\Omega \Big[\alpha(\pm u'(x)\pm tw'(x))\big(u(x)+tw(x)\big)' - \alpha(\pm u'(x))u'(x)\Big]\varphi'(x)\ dx\\
&= \lim_{t\to 0}\frac{1}{t}\int_\Omega \Big[\big(\alpha(\pm u'(x)) \pm t\alpha'(\pm u'(x))w'(x)\big)\big(u'(x)+tw'(x)\big) - \alpha(\pm u'(x))u'(x)\Big]\varphi'(x)\ dx\\
&= \int_\Omega \Big(\alpha(|u'(x)|) + \alpha'(|u'(x)|)|u'(x)|\Big)w'(x)\varphi'(x)\ dx = \langle A'(u)w,\varphi\rangle = a'(u)(w,\varphi).
\end{aligned}$$

If, for an arbitrary u, the $\widetilde{a}(u) : V\times V \longrightarrow \mathbb{R}$ is defined by

$$\widetilde{a}(u)(w,\varphi) = \int_\Omega \alpha(|u'(x)|)w'(x)\varphi'(x)\ dx,$$

then, of course,

$$a(u,\varphi) = \widetilde{a}(u)(u,\varphi).$$

Therefore

$$\underline{K}_h(\underline{u}_h) = \widetilde{K}_h(\underline{u}_h)\underline{u}_h,$$

where $\widetilde{K}_h(\underline{u}_h)$ denotes the stiffness matrix with respect tot the bilinear form $\widetilde{a}(u_h)$:

$$\widetilde{K}_{i,j} = \widetilde{a}(u_h)(\varphi_j, \varphi_i).$$

This type of representation is called a quasi-linear form.

Remark: In the quasi-linear case the following alternative to Newton's method is available:

$$C_n = \widetilde{K}_h(\underline{u}_h^{(n)}).$$

The computation of the next iterate reduces to solving the linear system

$$\widetilde{K}_h(\underline{u}_h^{(n)})\underline{u}^{(n+1)} = \underline{f}_h$$

by using $\underline{K}_h(\underline{u}_h^{(n)}) = \widetilde{K}_h(\underline{u}_h^{(n)})\underline{u}_h^{(n)}$. This technique is called linearization by freezing the coefficients. In comparison with Newton's method one omits that term, which contains the derivative α'.

Concluding remarks to FEM:

h-FEM, p-FEM, hp-FEM.

1.7 Finite Difference Methods

Starting point of the Finite Difference Method (FDM) is the classical formulation. The method will be explained for the example of a two-dimensional linear boundary value problem

$$\begin{aligned} -\operatorname{div}(A(x)\operatorname{grad} u(x)) + b(x)\cdot\operatorname{grad} u(x) + c(x)u(x) &= f(x) \quad x \in \Omega, \\ u(x) &= g_D(x) \quad x \in \Gamma_D, \\ A(x)\operatorname{grad} u(x)\cdot n(x) &= g_N(x) \quad x \in \Gamma_N. \end{aligned}$$

Firstly, we select two families of straight lines, which are parallel to the x-axis and the y-axis, respectively. By intersection we obtain grid points in the interior of Ω (Ω_h) and boundary grid points ($\Gamma_h = \Gamma_{Dh} \cup \Gamma_{Nh}$).

In each grid point $x \in \Omega_h$ the derivatives appearing in the differential equations are replaced by finite difference approximations. Analogously, in each boundary grid point $x \in \Gamma_h$ the derivatives involved in the Neumann boundary condition are discretized. We obtain a system of difference equations, which defines the approximation $u_h(x)$ of the exact solution in the grid points $x \in \Omega_h \cup \Gamma_h$. Hence the approximation u_h is considered as a function defined on $\Omega_h \cup \Gamma_h$ (grid function). If the differential equation and the Neumann

boundary conditions are discretized in some grid point x, only x and a few neighboring grid points appear in the difference equations. The set of these few involved grid points are collected in some set $S_h(x)$ (stencil). In the linear case the difference equations are of the following form:

$$\begin{aligned} \sum_{\xi \in S_h(x)} k_h(x,\xi) u_h(\xi) &= f_h(x) \quad x \in \Omega_h \cup \Gamma_{Nh}, \\ u_h(x) &= g_D(x) \quad x \in \Gamma_{Dh}. \end{aligned}$$

After elimination of the unknowns which are given by Dirichlet boundary conditions a system of linear equations results

$$K_h \underline{u}_h = \underline{f}_h$$

for the vector $\underline{u}_h$ of the approximations in the remaining grid points.

Advantages (compared to FEM):

- Simple generation of the grid (except for boundary grid points in the case of complex geometries)
- Simple assembling of and simple access to the discretization matrix
- Under appropriate conditions on the difference approximations (which are relatively easy to satisfy)) stability and convergence with respect to the supremum norm can be guaranteed.

Disadvantages (compared to FEM):

- Geometrically less flexible.
- Local refinements have global impact (additional grid points appear far away).
- Properties of the continuous problem (like e.g. symmetry or coerciveness) do not always carry over to the discretized problem.

The notion of monotonicity is of essential importance for FDM, if analyzing the methods in the supremum norm:

Definition 1.3. *A finite difference method is called monotone, if*

1. $k_h(x,x) > 0$ *for all* $x \in \Omega_h \cup \Gamma_{Nh}$ *and*
2. $k_h(x,\xi) < 0$ *for all* $x \in \Omega_h \cup \Gamma_{Nh}$ *and all* $\xi \in S_h(x)$ *with* $\xi \neq x$ *and*
3. $\sum_{\xi \in S_h(x)} k_h(x,\xi) \geq 0$ *for all* $x \in \Omega_h \cup \Gamma_{Nh}$.

Example: Consider the differential equation

$$-u''(x) + b(x)u'(x) = f(x).$$

On an equidistant grid central difference approximations can be used:

$$\frac{1}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) + b(x_i)\frac{1}{2h}(u_{i+1} - u_{i-1}) = f(x_i).$$

so

$$-\left[\frac{1}{h^2} + b(x_i)\frac{1}{2h}\right] u_{i-1} + \left[\frac{1}{h^2}\right] u_i - \left[\frac{1}{h^2} - b(x_i)\frac{1}{2h}\right] u_{i+1} = f(x_i)$$

This FDM is monotone only if h is small enough:

$$|b(x_i)|h < 2$$

However, if upwind differencing is used for the first order term, i.e.:

$$\begin{aligned} b(x_i)u'(x_i) &\approx \begin{cases} b(x_i)\dfrac{1}{h}(u_i - u_{i-1}) & \text{for } b(x_i) > 0 \\ b(x_i)\dfrac{1}{h}(u_{i+1} - u_i) & \text{for } b(x_i) \le 0 \end{cases} \\ &= \frac{1}{h}\left[-b(x_i)^+ u_{i-1} + |b(x_i)|u_i + b(x_i)^- u_{i+1}\right] \end{aligned}$$

the resulting FDM is always monotone:

$$-\left[\frac{1}{h^2} + b(x_i)^+\frac{1}{h}\right] u_{i-1} + \left[\frac{1}{h^2} + |b(x_i)|\frac{1}{h}\right] u_i - \left[\frac{1}{h^2} - b(x_i)^-\frac{1}{h}\right] u_{i+1} = f(x_i).$$

(Notation: $b^+ = \max(b, 0)$, $b^- = \min(b, 0)$, observe $b = b^+ + b^-$ and $|b| = b^+ - b^+$.)

A comment on the convergence analysis:

A difference equation for the error $e_h(x) = u_h(x) - u(x)$ can easily be derived:

$$\begin{aligned} \sum_{\xi \in S_h(x)} k_h(x, \xi)e_h(\xi) &= f_h(x) - \sum_{\xi \in S_h(x)} k_h(x, \xi)u(\xi) \quad x \in \Omega_h \cup \Gamma_{Nh} \\ e_h(x) &= 0 \quad x \in \Gamma_{Dh}. \end{aligned}$$

In short in matrix-vector notation:

$$K_h \underline{e}_h = \underline{\psi}_h.$$

Hence

$$\|\underline{e}_h\|_\infty \le C_S \|\underline{\psi}_h\|_\infty \quad \text{with} \quad \|K_h^{-1}\|_\infty \le C_S.$$

So, on the one hand the so-called consistency error

$$\psi_h(x) = f_h(x) - \sum_{\xi \in S_h(x)} k_h(x,\xi) u(\xi)$$

has to be analyzed (e.g., by Taylor expansion for sufficiently smooth solutions).

On the other hand, it has to be shown that

$$\max_{x \in \Omega_h \cup \Gamma_{Nh}} |e_h(x)| \leq C_S \max_{x \in \Omega_h \cup \Gamma_{Nh}} |\psi_h(x)|.$$

(C_h–C_h stability). For monotone FDMs a stability constant C_S can be determined relatively easily.

Consistency and stability then imply convergence.

Remark: Background: discrete maximum principle, monotone FDMs imply the following properties of the discretization matrices $K_h = K$ under weak conditions:

1. $K_{ii} > 0$.
2. $K_{ij} \leq 0$ for $i \neq j$.
3. $K^{-1} \geq 0$ (element-wise).

(M-matrices).

1.8 Finite Volume Methods

Starting point for a finite volume method (FVM) is a balance law derived from the differential equation. For a differential equation

$$-\operatorname{div}\Big(q(x, u(x), \operatorname{grad} u(x))\Big) = f(x) \quad x \in \Omega$$

the associated balance law is obtained by first integrating over some domain $\omega \subset \Omega$ and then using Gauss' Theorem:

$$-\int_{\partial\omega} q(x, u(x), \operatorname{grad} u(x)) \cdot n(x)\ ds = \int_{\omega} f(x)\ dx \quad \text{for all } \omega \subset \Omega.$$

The vector-valued function q is called the flux and right hand side f is called the source term.

Let $\mathcal{O}_h = \{\Omega_k : k = 1, 2, \ldots\}$ be a subdivision of the domain Ω in polygonal sub-domains (cells, control volumes, finite volumes). We have

$$-\int_{\partial\Omega_k} q(x, u(x), \operatorname{grad} u(x)) \cdot n_k(x)\ ds = \int_{\Omega_k} f(x)\ dx \quad \text{for all } k = 1, 2, \ldots$$

or

$$-\sum_{j\in N(k)}\int_{\Gamma_{kj}} q(x,u(x),\operatorname{grad} u(x))\cdot n_{kj}(x)\ ds = \int_{\Omega_k} f(x)\ dx \quad \text{for all } k=1,2,\ldots$$

where Γ_{kj} denotes the common edge, shared by Ω_k and a neighboring sub-domain Ω_j, $j \in N(k)$.

For a finite volume method the integrals on the left hand side, which can be interpreted as the total flux through an edge of the control volume, are approximated by a so-called numerical flux (in dependence of the chosen degrees of freedom):

$$g_{kj} \approx -\frac{1}{|\Gamma_{kj}|}\int_{\Gamma_{kj}} q(x,u(x),\operatorname{grad} u(x))\cdot n_{kj}\ ds.$$

It is reasonable to assume that

$$g_{kj} = -g_{jk}.$$

For edges on the boundary of the domain approximations of the fluxes follow from the boundary conditions. The integral over the source term is also approximated (e.g. by some quadrature rule):

$$f_k \approx \frac{1}{|\Omega_k|}\int_{\Omega_k} f(x)\ dx.$$

Then the following typical form of a discretized balance law is obtained:

$$\sum_{j\in N(k)} g_{kj}\,|\Gamma_{kj}| = f_k\,|\Omega_k| \quad \text{for } k=1,2,\ldots,$$

construction of the subdivision $\mathcal{O}_h$:

Starting from a (primary) grid $\mathcal{T}_h = \{T_1, T_2, \ldots\}$ in triangles two different methods of construction the (secondary) grid $\mathcal{O}_h = \{\Omega_1, \Omega_2, \ldots\}$ will be discussed:

- Voronoi diagrams (for the case of non-obtuse triangles):

  In each node x_i a polygonal control volume Ω_i is constructed by perpendicular bisection of each edge containing the node x_i.

- Donald diagrams:

  In each node x_i a polygonal control volume Ω_i is constructed by connecting the midpoints of all edges containing the node x_i and the center of gravity of all triangles T_k containing the node x_i.

FVMs are usually constructed either in the spirit of FDMs or in the spirits of FEMs.

An example of a FVM: approximation by quadrature rule and finite differencing

Special case $q = \operatorname{grad} u$:

Secondary grid: Voronoi diagram

Degrees of freedom in the nodes of the triangles of the primary grid (cell-centered).

$$-\frac{1}{|\Gamma_{kj}|}\int_{\Gamma_{kj}} \frac{\partial u(x)}{\partial n_{kj}}\, ds \approx -\frac{\partial u(x)}{\partial n_{kj}}\Big|_{m_{kj}} \approx \frac{1}{\|x_k - x_j\|}(u_k - u_j) = g_{kj}$$

Sign condition of an M-matrix is fulfilled.

An example of a FVM: the box method

Let $\mathcal{T}_h$ be a subdivision of the domain Ω in triangles, the construction of the subdivision in control volumes are done either by Voronoi diagrams or Donald diagrams. Let Ω_i by the control volume associated with the node x_i. For the approximation u_h piecewise linear and continuous functions are used (Courant element):

$$u_h \in V_h = \operatorname{span}(\varphi_i : i = 1, 2, \ldots, N_h)$$

Requirement

$$-\int_{\partial\Omega_i} q(x, u_h(x), \operatorname{grad} u_h(x)) \cdot n_i(x)\, ds = \int_{\Omega_i} f(x)\, dx \quad \text{for } i = 1, 2, \ldots, N_h.$$

Let $\chi_i = \chi_{\Omega_i}$ be the characteristic function of the set Ω_i for $i = 1, 2, \ldots, N_h$. For

$$v_h \in W_h = \{\sum_{i=1}^{N_h} v_i \chi_i : v_i \in \mathbb{R}\}$$

it follows:

$$-\sum_{i=1}^{N_h} \int_{\partial\Omega_i} \Big(q(x, u_h(x), \operatorname{grad} u_h(x)) \cdot n_i(x)\Big) v_h(x)\, ds = \int_{\Omega} f(x) v_h(x)\, dx.$$

Hence: Find $u_h \in V_h$ such that

$$a_h(u_h, v_h) = \langle F, v_h \rangle \quad \text{for all } v_h \in W_h$$

with

$$a_h(w, v) = -\sum_{i=1}^{N_h} \int_{\partial\Omega_i} \Big(q(x, w(x), \operatorname{grad} w(x)) \cdot n_i(x)\Big) v(x)\, ds.$$

Remark: Let v be a test function. Then the following variational problem can be derived:

$$\begin{aligned}
&-\int_\Omega \operatorname{div}\Big(q(x,u(x),\operatorname{grad}u(x))\Big)v(x)\ dx \\
&= -\sum_{i=1}^{N_h}\int_{\Omega_i} \operatorname{div}\Big(q(x,u(x),\operatorname{grad}u(x))\Big)v(x)\ dx \\
&= -\sum_{i=1}^{N_h}\int_{\partial\Omega_i} q(x,u(x),\operatorname{grad}u(x))\cdot n_i(x)v(x)\ ds \\
&\quad +\sum_{i=1}^{N_h}\int_{\Omega_i} q(x,u(x),\operatorname{grad}u(x))\cdot \operatorname{grad}v(x)\ dx = \int_\Omega f(x)v(x)\ dx,
\end{aligned}$$

i.e.:

$$a_h(u,v) = \langle F, v\rangle$$

with

$$\begin{aligned}
a_h(w,u) = &= -\sum_{i=1}^{N_h}\int_{\partial\Omega_i} q(x,u(x),\operatorname{grad}u(x))\cdot n_i(x)v(x)\ ds \\
&\quad +\sum_{i=1}^{N_h}\int_{\Omega_i} q(x,u(x),\operatorname{grad}u(x))\cdot \operatorname{grad}v(x)\ dx, \\
\langle F, v\rangle &= \int_\Omega f(x)v(x)\ dx
\end{aligned}$$

The box method discussed above is obtained from this variational setting by choosing piecewise linear and continuous trial functions u_h and piecewise constant (and, therefore, discontinuous test functions v_h. This is an example of a so-called Petrov-Galerkin method, for which the trial space V_h and the test space W_h are different.

Another approach based on this variational setting is to use also discontinuous trial functions. This leads to the so-called discontinuous Galerkin methods.

Chapter 2

Parabolic Differential Equations

2.1 Initial-Boundary Value Problems for Parabolic Differential Equations

Classical formulation:

Let $Q_T = \Omega \times (0,T)$ (space-time cylinder). We consider the following problem:
Find $u : \overline{Q}_T \longrightarrow \mathbb{R}$ such that the differential equation

$$\frac{\partial u}{\partial t}(x,t) + Lu(x,t) \;=\; f(x,t) \quad (x,t) \in \; Q_T$$

where

$$Lv(x) = -\operatorname{div}(A(x)\operatorname{grad} v(x)) + b(x)\cdot \operatorname{grad} v(x) + c(x)v(x)$$

and the boundary conditions

$$\begin{aligned} u(x,t) &= g_D(x,t) \quad (x,t) \in \Gamma_D \times (0,T), \\ A(x)\operatorname{grad} u(x,t)\cdot n(x) &= g_N(x,t) \quad (x,t) \in \Gamma_N \times (0,T) \end{aligned}$$

and the initial condition

$$u(x,0) = u_0(x) \quad x \in \overline{\Omega}$$

are satisfied.

Special case heat equation:

$$\frac{\partial u}{\partial t}(x,t) - \Delta u(x,t) = f(x,t).$$

Model problem:

$$\begin{aligned} u_t(x,t) - u_{xx}(x,t) &= f(x,t) \quad x \in (0,1), \\ u(0,t) &= 0, \\ u(1,t) &= 0, \\ u(x,0) &= u_0(x) \quad x \in [0,1] \end{aligned}$$

Variational formulation:

For the model problem one obtains analogously to th elliptic case:
Find $u : [0,1] \longrightarrow H_0^1(0,1)$ such that

$$\int_0^1 u_t(x,t)v(x)\ dx + \int_0^1 u_x(x,t)v_x(x)\ dx = \int_0^1 f(x,t)v(x)\ dx$$

for all $v \in V = H_0^1(0,1)$.

Generally: Find $u : [0,T] \longrightarrow V$ such that

$$(u'(t), v)_H + a(u(t), v) = \langle F(t), v\rangle \quad \text{for all } v \in V$$

and the initial condition

$$u(0) = u_0.$$

Notation: $u(t)(x) = u(x,t)$ and $u'(t)(x) = u_t(x,t)$, ordinary differential equation in Banach space. In addition to the Hilbert space $V = H_0^1(\Omega)$ also the Hilbert space $H = L^2(\Omega)$ (scalar product $(.,.)_H = (.,.)_0$) is involved.

Function spaces:

$X = L^2((0,T),V),\ X^* = (L^2((0,T),V))^* = L^2((0,T),V^*)$.

$$\|v\|_X = \left(\int_0^T \|v(t)\|_V^2\ dt\right)^{1/2}, \qquad \|w\|_{X^*} = \left(\int_0^T \|w(t)\|_{V^*}^2\ dt\right)^{1/2}.$$

Generalized derivative u': Motivation: For $\varphi \in C_0^\infty(0,T)$ it follows for classical derivatives:

$$\begin{aligned} \int_0^T \varphi(t) \int_\Omega u_t(x,t)v(x)\ dx\ dt &= \int_0^T \varphi(t)\frac{d}{dt}\left[\int_\Omega u(x,t)v(x)\ dx\right]\ dt \\ &= -\int_0^T \varphi'(t)\int_\Omega u(x,t)v(x)\ dx\ dt. \end{aligned}$$

For the function $u'(t) : v \mapsto \int_\Omega u_t(x,t)v(x)\ dx$, hence $u'(t) \in V^*$, we have:

$$\int_0^T \varphi(t)\langle u'(t), v\rangle\ dt = -\int_0^T \varphi'(t)(u(t),v)_H\ dt \quad \text{for all } v \in V,\ \varphi \in C_0^\infty(0,T).$$

Definition 2.1. *Let $u \in L^2((0,T),V)$. A function $w \in L^2((0,T),V^*)$ is called generalized derivative if and only if*

$$\int_0^T \varphi(t)w(t)\ dt = -\int_0^T \varphi'(t)u(t)\ dt \quad \textit{for all } \varphi \in C_0^\infty(0,T).$$

Here, $u(t)$ is to be interpreted as the functional $v \mapsto (u(t),v)_H$. Herewith $H = L^2(\Omega)$ is identified with H^*. Integral: Bochner integral: generalization of the Lebesgue integral.

Notation $w = u'$ (uniqueness).

Assumptions on V and H:

$$V \subset H, \quad V \text{ dense in } H, \quad \|v\|_H \le c\,\|v\|_V.$$

Then we have:

$$V \subset H \equiv H^* \subset V^*.$$

Definition 2.2. $H^1((0,T),V;H) = \{v \in L^2((0,T),V) : v' \in L^2((0,T),V^*)\}$. $\|v\|_1^2 = \|v\|_X^2 + \|v'\|_{X^*}^2$.

Theorem 2.1. $H^1((0,T),V;H) \subset C([0,T],H)$. *There is a constant c with*

$$\max_{t\in[0,T]} \|v(t)\|_H \le c\,\|v\|_1.$$

This justifies the notation $u(t)$, in particular $u(0)$.

Final formulation:

Find $u \in X = L^2((0,T),V)$ with

$$u' \in X^* = L^2((0,T),V^*),$$

such that

$$\begin{aligned} \langle u'(t),v\rangle + a(u(t),v) &= \langle F(t),v\rangle \quad \text{for all } v \in V, \\ u(0) &= u_0. \end{aligned}$$

So:

$$\begin{aligned} u'(t) + Au(t) &= F(t), \\ u(0) &= u_0. \end{aligned}$$

We have:

$$\begin{aligned} \int_0^T \varphi(t)\langle u'(t),v\rangle\ dt &= -\int_0^T \varphi'(t)(u(t),v)_H\ dt \\ &= \int_0^T \varphi(t)\frac{d}{dt}(u(t),v)_H\ dt \quad \text{for all } v \in V,\ \varphi \in C_0^\infty(0,T) \end{aligned}$$

and, therefore:

$$\langle u'(t), v\rangle = \frac{d}{dt}(u(t), v)_H.$$

Hence:

$$\begin{aligned} \frac{d}{dt}(u(t), v)_H + a(u(t), v) &= \langle F(t), v\rangle \quad \text{for all } v \in V, \\ u(0) &= u_0. \end{aligned}$$

It is easy to show that the solution is unique:

Lemma 2.1. *Assume that there is a constant $\mu_1 \geq 0$ with*

$$a(v, v) \geq \mu_1 \|v\|_V^2 \quad \textit{for all } v \in V.$$

Then there exists at most one solution of the initial value problem

$$\begin{aligned} \langle u'(t), v\rangle + a(u(t), v) &= \langle F(t), v\rangle \quad \textit{for all } v \in V, \\ u(0) &= u_0. \end{aligned}$$

Proof. Assume that $u_1(t)$ and $u_2(t)$ are solutions of the initial value problem. Then $u(t) = u_2(t) - u_1(t)$ is a solution of the initial value problem

$$\begin{aligned} \langle u'(t), v\rangle + a(u(t), v) &= 0 \quad \text{for all } v \in V, \\ u(0) &= 0. \end{aligned}$$

Now we have

$$\begin{aligned} \frac{d}{dt}\left[\frac{1}{2}\|u(t)\|_H^2\right] &= \langle u'(t), u(t)\rangle = -a(u(t), u(t)) \\ &\leq -\mu_1 \|u(t)\|_V^2 \leq -\mu_1 c^{-2} \|u(t)\|_H^2. \end{aligned}$$

Hence

$$\frac{d}{dt}\|u(t)\|_H + \mu_1 c^{-2} \|u(t)\|_H \leq 0.$$

Therefore,

$$\frac{d}{dt}\left(e^{\mu_1 t/c^2}\|u(t)\|_H\right) \leq 0.$$

It follows by integrating:

$$\|u(t)\|_H \leq e^{-\mu_1 t/c^2}\|u(0)\|_H = 0.$$

So: $u(t) = 0$. □

Remark: Observe that uniqueness follows also for $\mu_1 = 0$.

The following so-called a-priori estimates are an important part of the existence proof:

Lemma 2.2. *Assume that there exist constants $\mu_2 \geq \mu_1 > 0$ with*

$$a(v,v) \geq \mu_1 \|v\|_V^2 \quad \textit{for all } v \in V$$

and

$$|a(w,v)| \leq \mu_2 \|w\|_V \|v\|_V \quad \textit{for all } v, w \in V.$$

Let $u \in X$ be a solution of the initial value problem

$$\begin{aligned} \langle u'(t), v\rangle + a(u(t), v) &= \langle F(t), v\rangle \quad \textit{for all } v \in V, \\ u(0) &= u_0. \end{aligned}$$

Then there are constants C_1, C_2 and C_3, only depending on μ_1, μ_2, $\|u_0\|_H$ and $\|F\|_{X^*}$, such that:

$$\|u\|_X \leq C_1, \qquad \|Au\|_{X^*} \leq C_2, \qquad \|u(t)\|_H \leq C_3.$$

Proof. We have

$$\begin{aligned} \frac{1}{2}\|u(t)\|_H^2 &= \frac{1}{2}\|u_0\|_H^2 + \int_0^t \left[\langle F(s), u(s)\rangle - a(u(s), u(s))\right]\, ds \\ &\leq \frac{1}{2}\|u_0\|_H^2 + \int_0^t \left[\|F(s)\|_{V^*}\|u(s)\|_V - \mu_1 \|u(s)\|_V^2\right]\, ds. \end{aligned}$$

For $t = T$ it follows:

$$\begin{aligned} 0 &\leq \frac{1}{2}\|u_0\|_H^2 + \int_0^T \left[\|F(s)\|_{V^*}\|u(s)\|_V - \mu_1 \|u(s)\|_V^2\right]\, ds \\ &\leq \frac{1}{2}\|u_0\|_H^2 + \|F\|_{X^*}\|u\|_X - \mu_1 \|u\|_X^2. \end{aligned}$$

This implies

$$\|u\|_X \leq \frac{1}{2\mu_1}(\|F\|_{X^*} + \sqrt{\|F\|_{X^*}^2 + 2\mu_1 \|u_0\|_H}) \equiv C_1.$$

From the boundedness of A it immediately follows that

$$\|Au\|_{X^*} \leq \mu_2 \|u\|_X \leq \mu_2 C_1 \equiv C_2.$$

From the first equation it follows:

$$\|u(t)\|_H^2 \leq \|u_0\|_H^2 + 2\|F\|_{X^*}\|u\|_X \leq \|u_0\|_H^2 + 2\mu_2 C_1 \|F\|_{X^*} \equiv C_3^2.$$

□

Next: Semi-discretization, then existence theory:

2.2 Semi-discretization: the vertical method of lines

V is replaced by a finite-dimensional subspace V_h:

Find: $u_h : [0,T] \longrightarrow V_h$ such that

$$\begin{aligned} \frac{d}{dt}(u_h(t), v_h)_H + a(u_h(t), v_h) &= \langle F(t), v_h \rangle \quad \text{for all } v_h \in V_h, \\ (u_h(0), v_h)_H &= (u_0, v_h)_H \quad \text{for all } v_h \in V_h. \end{aligned}$$

Let $\{\varphi_i : i = 1, 2, \ldots, N_h\}$ be a basis of V_h. With

$$u_h(t)(x) = \sum_{j=1}^{N_h} u_j(t)\varphi_j(x)$$

one obtains

$$\begin{aligned} \sum_{j=1}^{N_h} (\varphi_j, \varphi_i)_H \, u_j'(t) + \sum_{j=1}^{N_h} a(\varphi_j, \varphi_i) \, u_j(t) &= \langle F(t), \varphi_i \rangle \quad \text{for all } i = 1, 2, \ldots, N_h, \\ \sum_{j=1}^{N_h} (\varphi_j, \varphi_i)_H \, u_h(0) &= (u_0, \varphi_i)_H \quad \text{for all } i = 1, 2, \ldots, N_h. \end{aligned}$$

So

$$\begin{aligned} M_h \underline{u}_h'(t) + K_h \underline{u}_h(t) &= \underline{f}_h(t), \\ M_h \underline{u}_h(0) &= \underline{g}_h \end{aligned}$$

with the mass matrix

$$M_h = (M_{ij}), \quad M_{ij} = (\varphi_j, \varphi_i)_H,$$

the stiffness matrix

$$K_h = (K_{ij}), \quad K_{ij} = a(\varphi_j, \varphi_i)$$

and the vectors

$$\underline{u}_h(t) = (u_i(t)), \qquad \underline{f}_h(t) = (f_i), \quad f_i = \langle F(t), \varphi_i \rangle, \qquad \underline{g}_h = (g_i), \quad g_i = (u_0, \varphi_i)_H.$$

So we obtain an initial value problem. Standard form:

$$\begin{aligned} u'(t) &= f(t, u(t)), \\ u(0) &= u_0. \end{aligned}$$

here with $u(t) = \underline{u}_h(t)$, $f(t, u(t)) = M_h^{-1}(\underline{f}_h(t) - K_h \underline{u}_h(t))$ and $u_0 = M_h^{-1} g_h$.

The existence and uniqueness of a solution easily follows by the Picard-Lindelöf Theorem.

For the initial value we have:

$$\|u_h(0)\|_H^2 = (u_h(0), u_h(0))_H = (u_0, u_h(0))_H \leq \|u_0\|_H \|u_h(0)\|_H,$$

so

$$\|u_h(0)\|_H \leq \|u_0\|_H.$$

Therefore, the a-priori estimates are also valid for the approximations with the same bounds:

$$\|u_h\|_X \leq C_1, \qquad \|Au_h\|_{X^*} \leq C_2, \qquad \|u_h(t)\|_H \leq C_3.$$

The existence of approximate solutions and their a-priori bounds guarantee the existence of a solution of the continuous problem:

Theorem 2.2. *Let V and H be separable Hilbert spaces with $V \subset H$, V dense in H and $\|v\| \leq c\,\|v\|_V$. Assume that there are constants $\mu_2 \geq \mu_1 > 0$ with*

$$a(v,v) \geq \mu_1 \|v\|_V^2 \quad \textit{for all } v \in V$$

and

$$|a(w,v)| \leq \mu_2 \|w\|_V \|v\|_V \quad \textit{for all } v, w \in V.$$

Then there exists a unique solution $u \in X$ of the initial value problem

$$\begin{aligned} \langle u'(t), v\rangle + a(u(t), v) &= \langle F(t), v\rangle \quad \textit{for all } v \in V, \\ u(0) &= u_0. \end{aligned}$$

Proof. (sketch) There is a sequence $(\varphi_i)_{i\in\mathbb{N}}$ in V such that $\bigcup_{n\in\mathbb{N}} V_n$ is dense in V, where $V_n = \operatorname{span}(\varphi_1, \varphi_2, \ldots, \varphi_n)$. The corresponding approximate solutions are denoted by $u_n : [0,T] \longrightarrow V_n$:

$$\begin{aligned} \langle u_n'(t), v_n\rangle + a(u_n(t), v_n) &= \langle F(t), v_n\rangle \quad \text{for all } v_n \in V_n, \\ (u_n(0), v_n)_H &= (u_0, v_n)_H \quad \text{for all } v_n \in V_n. \end{aligned}$$

Because of the a-priori bounds

$$\|u_n\|_X \leq C_1, \qquad \|Au_n\|_{X^*} \leq C_2, \qquad \|u_n(t)\|_H \leq C_3$$

it follows by a compactness argument (for a sub-sequence):

$$\begin{aligned} \langle f, u_n\rangle &\longrightarrow \langle f, u\rangle \quad \text{for all } f \in X^*, \\ \langle Au_n, v\rangle &\longrightarrow \langle w, v\rangle \quad \text{for all } v \in X, \\ (u_n(0), v)_H &\longrightarrow (u_0, v)_H \quad \text{for all } v \in H, \\ (u_n(T), v)_H &\longrightarrow (z, v)_H \quad \text{for all } v \in H. \end{aligned}$$

It can be shown that

$$u' = F - w, \quad u(0) = u_0, \quad u(T) = z$$

and that:

$$Au = w.$$

□

Remark: The spaces $V \subset H \equiv H^* \subset V^*$ with

1. V a separable and reflexive Banach space
2. H a separable Hilbert space
3. V is dense in H with $\|v\|_H \leq c\,\|v\|_H$ for all $v \in V$

are called an evolution triple (Gelfand triple).

2.2.1 The Discretization Error

Definition 2.3. *Let a be a bounded and coercive bilinear form on V. The Ritz projection $R_h : V \longrightarrow V_h$ is given by*

$$a(R_h w, v_h) = a(w, v_h) \quad \textit{for all } v_h \in V_h,\ w \in V.$$

The Ritz projection is a linear and continuous operator describing the approximate solution u_h of a variational problem

$$a(u, v) = \langle F, v \rangle \quad \text{for all } v \in V$$

by the Galerkin method in the form

$$u_h = R_h u.$$

Then Cea's lemma reads:

$$\|w - R_h w\|_V \leq \frac{\mu_2}{\mu_1} \inf_{w_h \in V_h} \|w - w_h\|_V.$$

Definition 2.4. *The projection $P_h : H \longrightarrow V_h$ is given by*

$$(P_h w, v_h)_H = (w, v_h)_H \quad \textit{for all } v_h \in V_h,\ w \in H.$$

For $H = L^2(\Omega)$ the operator P_h is called the L^2-projection on V_h.
For the initial condition we obtain:

$$(u_h(0), v_h)_H = (u_0, v_h)_H = (P_h u_0, v_h)_H,$$

hence

$$u_h(0) = P_h u_0 \equiv u_{0h}.$$

The discretization error is divided into two parts:

$$u_h(t) - u(t) = u_h(t) - R_h u(t) + R_h u(t) - u(t) = \theta_h(t) + \rho_h(t).$$

The second part $\rho_h(t)$ can be estimated by the approximation error using Cea's lemma.

For the first part $\theta_h(t)$ it follows from

$$\langle u'(t), v_h\rangle + a(u(t), v_h) = \langle u'(t), v_h\rangle + a(R_h u(t), v_h) = \langle F(t), v_h\rangle$$

and

$$\langle u_h'(t), v_h\rangle + a(u_h(t), v_h) = \langle F(t), v_h\rangle$$

by substraction

$$\langle u_h'(t) - u'(t), v_h\rangle + a(\theta_h(t), v_h) = 0.$$

Hence

$$\langle \theta_h'(t), v_h\rangle + a(\theta_h(t), v_h) = -\langle \rho_h'(t), v_h\rangle.$$

It follows

Theorem 2.3. *Assume that a is bounded and there is a constant $\mu_1 > 0$ with*

$$a(v, v) \geq \mu_1 \|v\|_V^2 \quad \textit{for all } v \in V.$$

Then we have:

$$\begin{aligned} \|u_h(t) - u(t)\|_H &\leq \|u_{0h} - R_h u_0\|_H \, e^{-\mu_1 t/c^2} + \|(I - R_h)u(t)\|_H \\ &\quad + \int_0^t \|\big[(I - R_h)u(s)\big]'\|_H \; e^{-\mu_1 (t-s)/c^2} \; ds. \end{aligned}$$

Proof.

$$\begin{aligned} \|\theta_h(t)\|_H \frac{d}{dt}\|\theta_h(t)\|_H &= \frac{1}{2}\frac{d}{dt}\|\theta_h(t)\|_H^2 = \langle \theta_h'(t), \theta_h(t)\rangle \\ &= -a(\theta_h(t), \theta_h(t)) - \langle \rho_h'(t), \theta_h(t)\rangle \\ &\leq -\mu_1 \|\theta_h(t)\|_V^2 + \|\rho_h'(t)\|_H \, \|\theta_h(t)\|_H \\ &\leq -\mu_1 c^{-2} \|\theta_h(t)\|_H^2 + \|\rho_h'(t)\|_H \, \|\theta_h(t)\|_H. \end{aligned}$$

So

$$\frac{d}{dt}\|\theta_h(t)\|_H + \mu_1 c^{-2} \|\theta_h(t)\|_H \leq \|\rho_h'(t)\|_H.$$

By multiplication with $e^{\mu_1 t/c^2}$ it follows:

$$\frac{d}{dt}\Big(\|\theta_h(t)\|_H e^{\mu_1 t/c^2}\Big) \leq \|\rho_h'(t)\|_H \; e^{\mu_1 t/c^2}.$$

By integration one obtains:

$$\begin{aligned} \|\theta_h(t)\|_H e^{\mu_1 t/c^2} &\leq \|\theta_h(0)\|_H + \int_0^t \|\rho_h'(s)\|_H \, e^{\mu_1 s/c^2} \; ds \\ &= \|u_{0h} - R_h u_0\|_H + \int_0^t \|\rho_h'(s)\|_H \, e^{\mu_1 s/c^2} \; ds. \end{aligned}$$

□

For the initial error we have:

$$\|u_{0h} - R_h u_0\|_H \leq \|u_{0h} - u_0\|_H + \|(I - R_h)u_0\|_H. = \|(I - P_h)u_0\|_H + \|(I - R_h)u_0\|_H.$$

For sufficiently smooth functions it follows: $(R_h u(t))' = R_h u'(t)$. Hence:

$$\begin{aligned}\|u_h(t) - u(t)\|_H &\leq [\|(I - P_h)u_0\|_H + \|(I - R_h)u_0\|_H]\ e^{-\mu_1\, t/c^2} + \|(I - R_h)u(t)\|_H \\ &\quad + \int_0^t \|(I - R_h)u'(s)\|_H\ e^{-\mu_1\,(t-s)/c^2}\ ds\end{aligned}$$

This shows that the analysis of the discretization error for the parabolic problem can be reduced to the error analysis of the corresponding elliptic problem.

Example: For the Courant element we have (under appropriate assumptions):

$$\|(I - R_h)v\|_0 \leq c_0\, h^2\, \|v\|_2 \quad \text{and} \quad \|(I - P_h)v\|_0 \leq c_0\, h^2\, \|v\|_2.$$

Therefore, (under appropriate conditions):

$$\|u_h(t) - u(t)\|_0 \leq C\, h^2 \left[\|u_0\|_2\, e^{-\mu_1\, t/c^2} + \|u(t)\|_2 + \int_0^t \|u'(s)\|_2\ e^{-\mu_1\,(t-s)/c^2}\ ds\right]$$

2.3 Runge-Kutta Methods for Initial Value Problems for Ordinary Differential Equations

In this chapter initial value problems for ordinary differential equations are discussed. In particular, the consequences for semi-discretized parabolic initial-boundary value problems will be studied.

We have the following general form of an initial value problem:

Find a function $u : [0, T] \longrightarrow \mathbb{R}^N$ such that:

$$\begin{aligned}u'(t) &= f(t, u(t)) \quad t \in (0, T), \\ u(0) &= u_0.\end{aligned} \tag{2.1}$$

In the following the symbols $\|.\|$ and $(.,.)$ are used to denote a norm and a scalar product on $\mathbb{R}^N$, respectively.

Special case: Assume that the right hand side of the differential equation does not explicitly depend on u:

$$\begin{aligned}u'(t) &= f(t) \quad t \in [0, T], \\ u(0) &= u_0.\end{aligned}$$

Then the solution can be represented in the following way:

$$u(t) = u_0 + \int_0^t f(s)\ ds.$$

So, the computation of the solution of the initial value problem reduces to the integration of the right hand side $f(t)$ in this case.

2.3.1 Euler's method

Other names: Euler polygon method, forward Euler method, explicit Euler method.

Assume that the interval $[0, T]$ is discretized by a sequence of grid points

$$0 = t_0 < t_1 < \ldots < t_m = T$$

e.g.: $t_j = j \cdot \tau$ for $j = 0, 1, \ldots, m$ and a given step size $\tau = T/m$.

Motivation by Taylor expansion:

By Taylor expansion at the point $t_0 = 0$ one obtains

$$u(t) \approx u(t_0) + u'(t_1)(t - t_0) = u_0 + f(t_0, u_0)(t - t_0) \equiv u_\tau(t) \quad \text{for } t \in [t_0, t_1].$$

Therefore, we obtain the following approximation at the point t_1:

$$u_1 = u_0 + \tau\, f(t_0, u_0).$$

By Taylor expansion at the point $t_0 = 0$ one obtains

$$\begin{aligned} u(t) &\approx u(t_1) + u'(t_1)(t - t_1) \\ &= u(t_1) + f(t_1, u(t_1))(t - t_1) \approx u_1 + f(t_1, u_1)(t - t_1) \equiv u_\tau(t) \quad \text{for } t \in [t_1, t_2]. \end{aligned}$$

Therefore, we obtain the following approximation at the point t_2:

$$u_2 = u_1 + \tau\, f(t_1, u_1).$$

If this process is continued analogously, one obtains a polygonal approximation $u_\tau(t)$ for the exact solution by connecting the approximations, given by

$$u_{j+1} = u_j + \tau\, f(t_j, u_j) \quad j = 0, 1, \ldots, m - 1$$

linearly.

Motivation as FDM:

If the derivative at the point t_j is replaced by a forward difference quotient

$$u'(t_j) \approx \frac{1}{\tau}(u(t_{j+1}) - u(t_j)),$$

one obtains Euler's method as a finite difference method:

$$\frac{1}{\tau}(u_{j+1} - u_j) = f(t_j, u_j) \quad j = 0, 1, \ldots, m - 1.$$

Motivation by quadrature rule:

By integrating the differential equation

$$u'(t) = f(t, u(t))$$

over the interval $[t, t+\tau]$ one obtains the relation

$$u(t+\tau) = u(t) + \int_t^{t+\tau} f(s, u(s))\, ds.$$

If the integral is approximated by the left rectangular rule, i.e.:

$$\int_t^{t+\tau} f(s, u(s))\, ds \approx \tau\, f(t, u(t)),$$

one obtains

$$u(t+\tau) \approx u(t) + \tau\, f(t, u(t)). \tag{2.2}$$

This motivates the formula

$$u_{j+1} = u_j + \tau\, f(t_j, u_j), \quad j = 0, 1, \ldots, m-1 \tag{2.3}$$

for the successive computation of approximations $u_j = u_\tau(t_j)$ of the exact values $u(t_j)$.

2.3.2 The classical convergence analysis

The hope is, of course, to obtain only small deviations from the exact solution by choosing sufficiently small step sizes. That is, we expect the following property:

$$e_\tau(t) \to 0 \quad \text{für } \tau \to 0 \tag{2.4}$$

for all $t \in [0, T]$, where $e_\tau(t)$ is called the **global (discretization) error**:

$$e_\tau(t) = u(t) - u_\tau(t).$$

Usually the global error is considered only at the grid points. The method is called **convergent**, if condition (2.4) is satisfied in some suitable norm.

The global error consists of contributions which can be interpreted as propagations of local errors $d_\tau(t_j)$, $j = 0, 1, \ldots$, by the method.

The local error at the point t_{j+1} for Euler's method is given by

$$d_\tau(t_{j+1}) = u(t_{j+1}) - u_\tau(t_{j+1}) = u(t_{j+1}) - \Big(u(t_j) + \tau\, f(t_j, u(t_j))\Big),$$

that is the difference between the exact solution of the differential equation and the approximate solution after one step of the method starting from the initial values $(t_j, u(t_j))$.

A possible error for the initial value defines the value of $d_\tau(t_0)$:

$$d_\tau(t_0) = u(0) - u_\tau(0).$$

The so-called consistency error of Euler's method interpreted as FDM is given by

$$\psi_\tau(t_{j+1}) = \frac{u(t_{j+1}) - u(t_j)}{\tau} - f(t_j, u(t_j))$$

which leads to the following simple relation with the local error:

$$\psi_\tau(t_{j+1}) = \frac{1}{\tau} d_\tau(t_{j+1}).$$

The investigation of the magnitude of the local errors is called consistency analysis. A method is called **consistent**, if the consistency error vanishes for $\tau \to 0$, i.e.: if

$$d_\tau = o(\tau).$$

in some suitable norm.

In order to estimate the propagation of local errors, one has to study the behavior of the difference $w_j - v_j$ in dependence of the initial difference $w_{j_0} - v_{j_0}$ for $j \geq j_0$, where the sequences v_j and w_j are generated by the numerical method, here Euler's method:

$$\begin{aligned} v_{j+1} &= v_j + \tau\, f(t_j, v_j), \\ w_{j+1} &= w_j + \tau\, f(t_j, w_j), \end{aligned}$$

starting from the initial values v_{j_0} and w_{j_0} at the point t_{j_0}. This investigation is called stability analysis.

If an estimation of the form

$$\|w_j - v_j\| \leq C \|w_{j_0} - v_{j_0}\|$$

can be shown in some suitable norm with a constant C independent of the step size τ, then the method is called **stable** and C is called the stability constant of the method.

So, in order to analyze the convergence of a method, its consistency and its stability have to be analyzed.

For Euler's method the consistency analysis and the stability analysis are quite simple:

Consistency analysis:

By Taylor series expansion one obtains

$$\begin{aligned} d_\tau(t+\tau) &= u(t+\tau) - u(t) - \tau\, f(t, u(t)) \\ &= u(t) + u'(t)\tau + R_2(t,\tau) - u(t) - \tau\, f(t, u(t)) \\ &= \Big[u'(t) - f(t, u(t))\Big]\, \tau + R_2(t,\tau) = R_2(t,\tau) = O(\tau^2) \end{aligned}$$

Or, more precisely for $u'' \in L^\infty(0,T)$:

$$\|d_\tau(t+\tau)\| \;=\; \|R_2(t,\tau)\| = \left\| \int_t^{t+\tau} (t+\tau-s) u''(s)\, ds \right\| \le \frac{\tau^2}{2} \sup_{s\in[t,t+\tau]} \|u''(s)\|$$

Therefore, there is a constant $K = \|u''\|_{L^\infty(0,T)}/2$ such that

$$\|d_\tau\|_{L^\infty(0,T)} \le K\,\tau^2$$

or, in short:

$$d_\tau = O(\tau^2).$$

So, the local error converges to 0, as τ^2 converges to 0, if τ converges to 0.

If the local error of a numerical method converges to 0 like $K\tau^{p+1}$ with some $p > 0$ for $\tau \to 0$ approaching 0, or, in short

$$d_\tau = O(\tau^{p+1}),$$

then the numerical method is called **consistent** with **consistency order** p.

With these notations and considerations Euler's method is consistent with consistency order 1.

Stability analysis:

Assume that the following Lipschitz condition for the right hand side of the differential equation: There is a constant $L \ge 0$ such that

$$\|f(t,w) - f(t,v)\| \le L\,\|w - v\| \quad \text{for all } t, v, w. \tag{2.5}$$

Then:

$$\begin{aligned} \|w_{j+1} - v_{j+1}\| &= \|\,[w_j + \tau\, f(t_j, w_j)] - [v_j + \tau\, f(t_j, v_j)]\,\| \\ &\le \|w_j - v_j\| + \tau\,\|f(t_j,w_j) - f(t_j,v_j)\| \le (1+\tau\,L)\,\|w_j - v_j\| \\ &\le e^{\tau\,L}\,\|w_j - v_j\|. \end{aligned}$$

By applying this estimate repeatedly one obtains the estimation:

$$\|w_j - v_j\| \le \underbrace{e^{\tau\,L}\cdot e^{\tau\,L}\cdots e^{\tau\,L}}_{(j-j_0)\text{ times}} \|w_{j_0} - v_{j_0}\| = e^{(j-j_0)\tau\,L}\,\|w_{j_0} - v_{j_0}\| \le e^{(t_j - t_{j_0})L}\,\|w_{j_0} - v_{j_0}\|.$$

Therefore, there is a constant $C = e^{(t_j - t_{j_0})L} \le e^{TL}$ independent of τ such that

$$\|w_j - v_j\| \le C\,\|w_{j_0} - v_{j_0}\|.$$

So the condition (2.5) implies the stability of Euler's method with a stability constant of the form $C = e^{tL}$.

Based on these considerations the convergence of the method can easily be analyzed:

Convergence analysis

The global error consists of the propagated local errors. For a consistent and stable method of consistency order $p > 0$ we have:

$$\begin{aligned} \|e_\tau(t_j)\| &= \|u(t_j) - u_\tau(t_j)\| \\ &\leq \sum_{k=1}^{j} C\|d_\tau(t_k)\| \leq \sum_{k=1}^{j} C\,K\,\tau^{p+1} = C\,K\,\tau^p \sum_{k=1}^{j} \tau \leq C\,K\,t_j\,\tau^p = C'\,\tau^p \end{aligned}$$

with $C' = C\,K\,t_j \leq C\,K\,T$, hence:

$$e_\tau = O(\tau^p).$$

So the method is convergent. More precisely, the method is called convergent with **convergence order** p.

Remark: In short:

$$\text{consistency} \;+\; \text{stability} \;=\; \text{convergence}$$

In particular, it follows that Euler's method is convergent with convergence order 1:

Theorem 2.4. *If $u'' \in L^\infty(0,T)$ and if*

$$\|f(t,w) - f(t,v)\| \leq L\|w - v\| \quad \textit{for all } t, v, w,$$

then Euler's method converges and we have

$$\|u_j - u(t_j)\| \leq e^{t_j L} \left[\|u_0 - u(t_0)\| + \frac{\tau}{2}\, t_j\, \|u''\|_{L^\infty((0,T),H)}\right].$$

Remark: Under the weaker condition $u'' \in L^1(0,T)$ one obtains

$$\|d_\tau(t+\tau)\| \leq \tau \int_t^{t+\tau} \|u''(s)\|\, ds$$

and, therefore,

$$\|u_j - u(t_j)\| \leq e^{t_j L} \left[\|u_0 - u(t_0)\| + \tau\, \|u''\|_{L^1((0,T),H)}\right].$$

2.3.3 Explicit Runge-Kutta Methods

If the integral in

$$u(t+\tau) = u(t) + \int_t^{t+\tau} f(s, u(s))\, ds \tag{2.6}$$

is approximated by more accurate quadrature rules, more accurate methods are obtained.

For example, one could use the midpoint rule:

$$\int_t^{t+\tau} f(s, u(s))\, ds \approx \tau\, f(t + \frac{\tau}{2}, u(t + \frac{\tau}{2})).$$

However, the value $u(t+\tau/2)$ is not available. But it can be approximated from the relation

$$u(t+\frac{\tau}{2}) = u(t) + \int_t^{t+\tau/2} f(s,u(s))\; ds$$

be using another quadrature rule: the left rectangular rule:

$$u(t+\frac{\tau}{2}) \approx u(t) + \frac{\tau}{2} f(t,u(t)).$$

In summary, one obtains the following method (Runge's second order method):

$$u_{j+1} = u_j + \tau\, f(t_j + \frac{\tau}{2}, g_2)$$

with

$$\begin{aligned} g_1 &= u_j, \\ g_2 &= u_j + \frac{\tau}{2} f(t_j, g_1). \end{aligned}$$

Consistency analysis:

$$\begin{aligned} d_\tau(t+\tau) &= u(t+\tau) - u_\tau(t+\tau) \\ &= u(t+\tau) - u(t) - \tau\, f(t+\frac{\tau}{2}, u(t) + \frac{\tau}{2} f(t,u(t))) \\ &= u(t) + u'(t)\tau + \frac{1}{2}u''(t)\tau^2 + O(\tau^3) - u(t) \\ &\quad - \tau\left[f(t,u(t)) + f_t(t,u(t))\frac{\tau}{2} + f_u(t,u(t))\frac{\tau}{2} f(t,u(t)) + O(\tau^2)\right] \\ &= [u'(t) - f(t,u(t))]\tau + [u''(t) - f_t(t,u(t)) - f_u(t,u(t))f(t,u(t))]\frac{\tau^2}{2} + O(\tau^3). \end{aligned}$$

From

$$u'(t) = f(t,u(t))$$

we obtain by differentiation

$$u''(t) = f_t(t,u(t)) + f_u(t,u(t))u'(t) = f_t(t,u(t)) + f_u(t,u(t))f(t,u(t)).$$

Hence, it follows for the local error:

$$d_\tau(t+\tau) = O(\tau^3),$$

i.e.: the consistency order of this method is 2.

The construction of numerical methods can be easily generalized. Starting point is a quadrature rule of the form:

$$\begin{aligned} &\int_t^{t+\tau} f(s,u(s))\; ds \\ &\quad\approx\; \tau\left[b_1\, f(t,u(t)) + b_2\, f(t+c_2\,\tau, u(t+c_2\,\tau)) + \cdots + b_s\, f(t+c_s\,\tau, u(t+c_s\,\tau))\right]. \end{aligned}$$

The values b_i, $i = 1, 2, \ldots, s$ are called the weights, the values $t + c_i\,\tau$, $i = 1, 2, \ldots, s$ with $c_1 = 0$ are called the nodes of the quadrature rule.

Instead of the unknown quantities $u(t + c_i\,\tau)$ approximations

$$g_i \approx u(t + c_i\,\tau)$$

are recursively computed by using quadrature rules applied to

$$u(t + c_i\,\tau) = u(t) + \int_t^{t+c_i\,\tau} f(s, u(s))\; ds.$$

Then one obtains:

$$\begin{aligned}
g_1 &= u_j, \\
g_2 &= u_j + \tau\, a_{21}\, f(t_j, g_1), \\
g_3 &= u_j + \tau \Big[a_{31}\, f(t_j, g_1) + a_{32}\, f(t_j + c_2\,\tau, g_2) \Big], \\
&\vdots \\
g_s &= u_j + \tau \Big[a_{s1}\, f(t_j, g_1) + a_{s2}\, f(t_j + c_2\,\tau, g_2) + \cdots + a_{s,s-1}\, f(t_j + c_{s-1}\,\tau, g_{s-1}) \Big].
\end{aligned} \tag{2.7}$$

Then the next approximations has the form:

$$u_{j+1} = u_j + \tau \Big[b_1\, f(t_j, g_1) + b_2\, f(t_j + c_2\,\tau, g_2) + \cdots + b_s\, f(t_j + c_s\,\tau, g_s) \Big]. \tag{2.8}$$

The method (2.7), (2.8) is called an s-stage explicit Runge-Kutta method. For describing the method it is sufficient to specify the coefficients a_{ij}, b_j and c_i in form of the following (Butcher) tableau:

$$\begin{array}{c|ccccc}
0 & & & & & \\
c_2 & a_{21} & & & & \\
c_3 & a_{31} & a_{32} & & & \\
\vdots & \vdots & \vdots & \ddots & & \\
c_s & a_{s1} & a_{s2} & \ldots & a_{s,s-1} & \\
\hline
 & b_1 & b_2 & \ldots & b_{s-1} & b_s
\end{array}$$

or in compact dorm:

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array}.$$

Examples: Euler's method is a 1-stage Runge-Kutta method with tableau

$$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array}$$

and consistency order 1.

Runge's second order method is a 2-stage Runge-Kutta method with tableau

$$\begin{array}{c|cc} 0 & & \\ 1/2 & 1/2 & \\ \hline & 0 & 1 \end{array}$$

and consistency order 2.

By an appropriate choice of the coefficients one obtains corresponding high consistency orders. Let $p(s)$ denote the maximally attainable consistency order of an s-stage Runge-Kutta method. Then

s	1	2	3	4	5	6	7	8	9	$s \geq 10$
$p(s)$	1	2	3	4	4	5	6	6	7	$\leq s-3$

where the symbols | denote the so-called Butcher barriers.

The best known representative of a 4-stage Runge-Kutta method of order 4 is the "classical" Runge-Kutta method, given by the tableau

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}$$

Under the condition (2.5) stability cam be shown for the whole class of explicit Runge-Kutta methods.

2.3.4 Stiff Differential Equations and A-Stability

For a semi-discretized parabolic initial-boundary value problem one obtains the following Lipschitz condition for the right hand side with respect to the norm $\|.\| = \|.\|_{M_h}$:

$$\|f(t,w) - f(t,v)\| = \|M_h^{-1} K_h(\underline{w}_h - \underline{v}_h)\|_{M_h} \leq L \, \|\underline{w}_h - \underline{v}_h\|_{M_h}$$

with the sharp bound

$$L = \|M_h^{-1} K_h\|_{M_h}.$$

It is easy to see that

$$\|M_h^{-1} K_h\|_{M_h} \geq \lambda_{\max}(M_h^{-1} K_h).$$

In the symmetric case $K_h^T = K_h$ we have even equality:

$$\|M_h^{-1} K_h\|_{M_h} = \lambda_{\max}(M_h^{-1} K_h).$$

Example: For the one-dimensional model problem one obtains:

$$\frac{3}{\max_k h_k^2} \leq \lambda_{\max}(M_h^{-1} K_h) \leq \frac{12}{\min_k h_k^2}. \tag{2.9}$$

Proof. We have

$$\begin{aligned}
(K_h \underline{v}_h, \underline{v}_h)_{\ell_2} &= \sum_{k=1,N_h} \left(K_h^{(k)} \begin{pmatrix} v_{k-1} \\ v_k \end{pmatrix}, \begin{pmatrix} v_{k-1} \\ v_k \end{pmatrix} \right)_{\ell_2} = \sum_{k=1,N_h} \frac{1}{h_k} \left(\hat{K} \begin{pmatrix} v_{k-1} \\ v_k \end{pmatrix}, \begin{pmatrix} v_{k-1} \\ v_k \end{pmatrix} \right)_{\ell_2} \\
&\leq 12 \sum_{k=1,N_h} \frac{1}{h_k} \left(\hat{M} \begin{pmatrix} v_{k-1} \\ v_k \end{pmatrix}, \begin{pmatrix} v_{k-1} \\ v_k \end{pmatrix} \right)_{\ell_2} \\
&\leq \frac{12}{\min_k h_k^2} \sum_{k=1,N_h} h_k \left(\hat{M} \begin{pmatrix} v_{k-1} \\ v_k \end{pmatrix}, \begin{pmatrix} v_{k-1} \\ v_k \end{pmatrix} \right)_{\ell_2} \\
&= \frac{12}{\min_k h_k^2} \sum_{k=1,N_h} \left(M_h^{(k)} \begin{pmatrix} v_{k-1} \\ v_k \end{pmatrix}, \begin{pmatrix} v_{k-1} \\ v_k \end{pmatrix} \right)_{\ell_2} = \frac{12}{\min_k h_k^2} (M_h \underline{v}_h, \underline{v}_h)_{\ell_2}
\end{aligned}$$

and

$$(K_h e_i, e_i)_{\ell_2} = \frac{1}{h_i} + \frac{1}{h_{i+1}} \qquad (M_h e_i, e_i)_{\ell_2} = \frac{1}{3}(h_i + h_{i+1}).$$

Hence

$$\frac{3}{\max_k h_k^2} \leq \frac{3}{h_i h_{i+1}} = \frac{(K_h e_i, e_i)_{\ell_2}}{(M_h e_i, e_i)_{\ell_2}} \leq \lambda_{\max}(M_h^{-1} K_h) = \sup_{\underline{v}_h \neq 0} \frac{(K_h \underline{v}_h, \underline{v}_h)_{\ell_2}}{(M_h \underline{v}_h, \underline{v}_h)_{\ell_2}} \leq \frac{12}{\min_k h_k^2}$$

□

This implies, e.g., for an equidistant subdivision

$$L = O\left(\frac{1}{h^2}\right) \gg 1.$$

Therefore, a stability bound of the form $C = e^{tL}$ is completely useless.

Stability estimates for initial value problems

Consider the initial value problem

$$\begin{aligned}
u'(t) &= f(t, u(t)) \quad t \in (0, T), \\
u(0) &= u_0.
\end{aligned}$$

Under the Lipschitz condition

$$\|f(t, w) - f(t, v)\| \leq L \|w - v\|$$

it can be shown that

$$\|w(t) - v(t)\| \leq e^{(t-t_0)\,L} \|w(t_0) - v(t_0)\|$$

for two solutions $w(t)$ and $v(t)$ of the differential equation and $0 \leq t_0 \leq t \leq T$. Compare with the analogous stability estimate of Euler's method under the same Lipschitz condition on f.

Now assume that a so-called one-sided Lipschitz condition is available:

$$(f(t,w) - f(t,v), w - v) \leq L \|w - v\|^2$$

Then, for two solutions $w(t)$ and $v(t)$ of the differential equation, we have

$$\begin{aligned}
\|w(t) - v(t)\| \frac{d}{dt}\|w(t) - v(t)\| &= \frac{d}{dt}\frac{1}{2}\|w(t) - v(t)\|^2 \\
&= \frac{d}{dt}\frac{1}{2}(w(t) - v(t), w(t) - v(t)) \\
&= (w'(t) - v'(t), w(t) - v(t)) \\
&= (f(t, w(t)) - f(t, v(t)), w(t) - v(t)) \\
&\leq L \|w(t) - v(t)\|^2
\end{aligned}$$

Hence

$$\frac{d}{dt}\|w(t) - v(t)\| \leq L \|w(t) - v(t)\|,$$

which easily implies

$$\|w(t) - v(t)\| \leq e^{(t-t_0)\,L} \|w(t_0) - v(t_0)\|.$$

Of course, the Lipschitz condition implies the one-sided Lipschitz condition with the same Lipschitz constant L. However, In some interesting cases the Lipschitz constant in the one-sided condition is considerably smaller than in the original Lipschitz condition.

Of particular interest is the case of a vanishing one-sided Lipschitz constant:

$$(f(t,w) - f(t,v), w - v) \leq 0$$

Differential equations with this property are (sometimes) called dissipative. Then, of course, we have

$$\|w(t) - v(t)\| \leq \|w(t_0) - v(t_0)\|.$$

Example: The discussed semi-discretized parabolic initial-boundary value problem leads to a right hand side of the following form:

$$f(t,u) = M_h^{-1}\left[\underline{f}_h(t) - K_h \underline{u}_h\right]$$

with $u = \underline{u}_h$. Hence:

$$\begin{aligned}
(f(t, \underline{w}_h - f(t, \underline{w}_h), \underline{w}_h - \underline{v}_h)_{M_h} &= -\left(M_h M_h^{-1} K_h \left[\underline{w}_h - \underline{v}_h\right], \underline{w}_h - \underline{v}_h\right)_{\ell_2} \\
&= -\left(K_h \left[\underline{w}_h - \underline{v}_h\right], \underline{w}_h - \underline{v}_h\right)_{\ell_2} \leq 0
\end{aligned}$$

So, while the Lipschitz constant $\|M_h^{-1} K_h\|_{M_h}$ is large, the one-sided Lipschitz constant is 0, i.e.: the system is dissipative.

Stability estimates for Runge-Kutta methods

The stability estimates for dissipative systems suggest that an numerical method for computing an approximate solution should satisfy a similar stability property:

$$\|w_{j+1} - v_{j+1}\| \leq \|w_j - v_j\| \quad \text{for all } j = 0, 1, \ldots, m-1.$$

Methods with this property are called contractive.

For contractive methods we have the stability constant $C = 1$ and the convergence statements follow accordingly.

We start the analysis with a very simple class of differential equations: consider scalar linear differential equations of the form

$$u'(t) = \lambda\, u(t). \tag{2.10}$$

with $\lambda \in \mathbb{C}$.

Remark: Such simple model problems result from more general linear systems of differential equations of the form

$$u'(t) = Ju(t),$$

where J is a constant matrix, if the initial value is an eigenvector v of J with eigenvalue $\lambda \in \mathbb{C}$.

For a Runge-Kutta method, applied to (2.10), one obtains:

$$\begin{aligned} g &= u_j e + \tau\lambda A g, \\ u_{j+1} &= u_j + \tau\lambda b^T g \end{aligned}$$

with $g = (g_1, g_2, \ldots, g_s)^T$ and $e = (1, 1, \ldots, 1)^T$, hence

$$u_{j+1} = R(\tau\lambda)\, u_j$$

with

$$R(z) = 1 + z b^T (I - zA)^{-1} e.$$

$R(z)$ is called the stability function of the Runge-Kutta method.

Examples: Euler's method:

$$R(z) = 1 + z$$

Runge's second order method:

$$R(z) = 1 + z + \frac{1}{2} z^2.$$

Classical fourth order Runge-Kutta method:

$$R(z) = 1 + z + \frac{1}{2} z^2 + \frac{1}{6} z^3 + \frac{1}{24} z^4.$$

Remark: For the exact solution of the differential equation (2.10) one obtains

$$u(t) = u_0 e^{\lambda t}$$

and, therefore:

$$u(t_{j+1}) = e^{\lambda \tau} u(t_j).$$

The accuracy of a Runge-Kutta method corresponds to the accuracy of the approximation of the exponential function e^z by the stability function $R(z)$ in a neighborhood of $z = 0$.

The differential equation (2.10) is dissipative if and only if

$$\operatorname{Re} \lambda \leq 0.$$

The corresponding property of contractivity for the Runge-Kutta method leads to the condition

$$|R(\tau\lambda)| \leq 1.$$

With the help of the stability function the so-called stability domain of a Runge-Kutta method is defined by:

$$S = \{z \in \mathbb{C} : |R(z)| \leq 1\}.$$

With this notation the above-mentioned condition on the step size τ (contractivity) reads:

$$\tau\lambda \in S.$$

Example: The one-dimensional model problem of a semi-discretized parabolic initial-boundary-value problem leads to a linear system of the form

$$\underline{u}_h'(t) = J\underline{u}_h(t) + M_h^{-1}\underline{f}_h(t) \quad \text{with} \quad J = -M_h^{-1}K_h.$$

One immediately sees that, in the symmetric case $K_h^T = K_h$, the matrix J has only real and negative eigenvalues. For Euler's method the stability domain is the interior and the boundary of a circle with radius 1 and center -1:

$$S = \{z \in \mathbb{C} : |z - (-1)| \leq 1\}.$$

The step size must satisfy the condition

$$\tau \leq \frac{2}{\lambda_{\max}(M_h^{-1}K_h)}.$$

This condition is satisfied if

$$\tau \leq \frac{h^2}{6},$$

see (2.9). This is a strong restriction on the time step size τ, in particular if the spatial step size h is small.

This initial-value problem is an example of a so-called stiff differential equation: The existence of eigenvalues with a large negative real part is the cause, that Euler's method produces reasonable approximations only for very small step sizes τ.

It is desirable that, for all values of λ which lead to stable (bounded) solutions, the numerical method also produces stable (bounded) approximations. This leads to the notion of A-stability:

Definition 2.5. *A Runge-Kutta method is called A-stable if and only if*

$$|R(z)| \leq 1 \quad \textit{for all } \lambda \in \mathbb{C} \textit{ with } \operatorname{Re}\lambda \leq 0$$

or, in short:

$$\mathbb{C}^- \subset S$$

with $\mathbb{C}^- = \{z \in \mathbb{C} : \operatorname{Re}\lambda \leq 0\}$.

The stability function of an explicit s-stage Runge-Kutta method is a polynomial of degree $\leq s$, therefore, the method can not be A-stable.

2.3.5 Implicit Runge-Kutta methods

Euler's method (the explicit Euler method) was obtained by using the left rectangular rule in (2.6). By using the right rectangular rule instead

$$\int_t^{t+\tau} f(s, u(s))\, ds \approx \tau\, f(t+\tau, u(t+\tau))$$

the so-called implicit Euler method (backward Euler method) is obtained:

$$u_{j+1} = u_j + \tau\, f(t_j + \tau, u_{j+1}).$$

In order to actually calculate u_{j+1} this equation (in general a nonlinear system of equations) must be solved.

If one uses the midpoint rule

$$\int_t^{t+\tau} f(s, u(s))\, ds \approx \tau\, f(t+\frac{\tau}{2}, u(t+\frac{\tau}{2}))$$

for improving the accuracy and if the needed solution $u(t+\tau/2)$ is replaced by an approximation g_1 obtained by the implicit Euler method, one obtains the so-called implicit midpoint rule:

$$\begin{aligned} g_1 &= u_j + \frac{\tau}{2} f(t_j + \frac{\tau}{2}, g_1), \\ u_{j+1} &= u_j + \tau\, f(t_j + \frac{\tau}{2}, g_1). \end{aligned}$$

These two methods are examples of the so-called implicit Runge-Kutta methods.

The general form of an s-stage Runge-Kutta method can by written in the following way:

$$\begin{aligned} g_1 &= u_j + \tau\,[a_{11}\,f(t+c_1\,\tau, g_1) + a_{12}\,f(t+c_2\,\tau, g_2) + \cdots + a_{1s}\,f(t+c_s\,\tau, g_s)]\,, \\ g_2 &= u_j + \tau\,[a_{21}\,f(t+c_1\,\tau, g_1) + a_{22}\,f(t+c_2\,\tau, g_2) + \cdots + a_{2s}\,f(t+c_s\,\tau, g_s)]\,, \\ &\vdots \\ g_s &= u_j + \tau\,[a_{s1}\,f(t+c_1\,\tau, g_1) + a_{s2}\,f(t+c_2\,\tau, g_2) + \cdots + a_{ss}\,f(t+c_s\,\tau, g_s)] \end{aligned} \tag{2.11}$$

and

$$u_{j+1} = u_j + \tau\,[b_1\,f(t+c_1\,\tau, g_1) + b_2\,f(t+c_2\,\tau, g_2) + \cdots + b_s\,f(t+c_s\,\tau, g_s)] \tag{2.12}$$

The method is uniquely determined by the following tableau of coefficients:

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \dots & a_{1,s-1} & a_{1s} \\ c_2 & a_{21} & a_{22} & \dots & a_{2,s-1} & a_{2s} \\ \vdots & & & & & \\ c_s & a_{s1} & a_{s2} & \dots & a_{s,s-1} & a_{ss} \\ \hline & b_1 & b_2 & \dots & b_{s-1} & b_s \end{array}$$

or in compact form:

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array}. \tag{2.13}$$

Definition 2.6. *A Runge-Kutta method given by the tableau (2.13) is called*

1. *explicit, if A is a strictly lower triangular matrix,*
2. *implicit, if it is not explicit.*

This definition of an explicit method is slightly more general as before. So far is was always assumed that $c_1 = 0$ for an explicit method.

Examples:

1. The implicit Euler method is a 1-stage implicit Runge-Kutta method with tableau:

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}.$$

2. The implicit midpoint rule is a 1-stage implicit Runge-Kutta method with tableau:

$$\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array}.$$

3. Another possible quadrature rule for (2.6) is given by

$$\int_t^{t+\tau} f(s,u(s))\,ds \approx \tau\left[(1-\theta)\,f(t,u(t)) + \theta\,f(t+\tau,u(t+\tau))\right].$$

This leads to the method

$$u_{j+1} = u_j + \tau\left[(1-\theta)\,f(t_j,u_j) + \theta\,f(t_j+\tau,u_{j+1})\right],$$

so

$$\begin{aligned} g_1 &= u_j, \\ g_2 &= u_j + \tau\left[(1-\theta)\,f(t_j,g_1) + \theta\,f(t_j+\tau,g_2)\right], \\ u_{j+1} &= u_j + \tau\left[(1-\theta)\,f(t_j,g_1) + \theta\,f(t_j+\tau,g_2)\right]. \end{aligned}$$

This method is called the θ-method and it is (in general) a 2-stage Runge-Kutta method with tableau

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1-\theta & \theta \\ \hline & 1-\theta & \theta \end{array}.$$

This class of methods contains important special cases: the explicit Euler method ($\theta = 0$), the implicit Euler method ($\theta = 1$) and the implicit trapezoidal rule ($\theta = 1/2$).

Implementation of implicit Runge-Kutta Methods:

In order to calculate the next approximation u_{j+1} by an implicit Runge-Kutta method from (2.12), the values for $g_1, g_2, \ldots, g_s$ must be determined firstly by approximately solving the (in general non-linear) system of equations (2.11).

The equations are given in fixed point form. Therefore, one option would be to use simple fixed point iteration. The convergence of such an iterative method can be shown for initial values $g_i = u$ and sufficiently small step sizes τ. Better initial values can be calculated by using an appropriate explicit Runge-Kutta method.

For stiff differential equations a simplified Newton method is better suited for calculating $g_1, g_2, \ldots, g_s$ from (2.11). It is usually sufficient to evaluate the Jacobian matrix only once at the initial value $g_i = u$.

Example: The θ-method for the model problem of a semi-discretized parabolic differential equation

$$\begin{aligned} \underline{u}_h'(t) &= M_h^{-1}[\underline{f}_h(t) - K_h\underline{u}_h(t)], \\ \underline{u}_h(0) &= \underline{u}_{0,h} \end{aligned}$$

reads

$$\underline{u}_h^{j+1} = \underline{u}_h^j + \tau\left\{(1-\theta)\,M_h^{-1}\left[\underline{f}_h(t_j) - K_h\underline{u}_h^j\right] + \theta\,M_h^{-1}\left[\underline{f}_h(t_{j+1}) - K_h\underline{u}_h^{j+1}\right]\right\}.$$

The approximation $\underline{u}_h^{j+1}$ is obtained by solving the linear system of equations

$$[M_h + \tau\,\theta\,K_h]\,\underline{u}_h^{j+1} = [M_h - \tau(1-\theta)\,K_h]\,\underline{u}_h^{j} + \tau\,\left[(1-\theta)\,\underline{f}_h(t_j) + \theta\,\underline{f}_h(t_{j+1})\right].$$

In the case $\theta = 1/2$ (implicit trapezoidal rule) the method is also called Crank-Nicolson method.

Remark: For $\theta = 0$ one obtains the explicit Euler method. Nevertheless, it requires the solution of a linear system of equations. One can avoid this by replacing the mass matrix M_h by a diagonal matrix $\bar{M}_h$ (mass lumping). Example for the Courant-element:

$$\bar{M}_h = (\bar{M}_{ij}), \quad \bar{M}_{ij} = (\varphi_i, \varphi_j)_h,$$

where instead of the L^2-scalar product the following approximation based on the trapezoidal rule is used:

$$(v,w)_{L^2(\Omega)} = \sum_k \int_{T_k} v(x)w(x)\,dx \approx \sum_k h_k \left[\frac{1}{2}v(x_{k-1})w(x_{k-1}) + \frac{1}{2}v(x_k)w(x_k)\right] \equiv (v,w)_h.$$

Consistency order of implicit Runge-Kutta methods

As in the explicit case the global error, the local error and the consistency error can be introduced.

The consistency order can be determined by Taylor series expansions as in the explicit case.

Examples:

1. The θ-method has, in general, consistency order 1. For $\theta = 1/2$ one obtains consistency order 2.

2. For the 1-stage implicit midpoint rule one obtains consistency order 2.

These simple examples already show that a higher consistency order can be reached with an s-stage implicit Runge-Kutta method than with an s-stage explicit method.

It can be shown that the maximally attainable consistency order for an s-stage Runge-Kutta method is $p = 2s$. These methods with maximal consistency order are called s-stage Gauss methods. They are based on the Gaussian quadrature rule which are of maximal accuracy. The implicit midpoint rule the 1-stage Gauss method.

Stability of implicit Runge-Kutta methods

For analyzing the A-stability the stability function is needed:

Examples:

1. For the implicit Euler method one obtains

$$R(z) = \frac{1}{1-z}.$$

The stability domain is the exterior and the boundary of the circle with radius 1 and center 1:

$$S = \{z \in \mathbb{C} : |z-1| \geq 1\}.$$

Hence the method is A-stable.

2. For the θ-method one obtains

$$R(z) = \frac{1+(1-\theta)z}{1-\theta z}.$$

Hence the method is A-stable for $\theta \geq 1/2$.

So far, stability was studied only for the very simple scalar model problem

$$u'(t) = \lambda u(t).$$

Now we discuss the extension to linear systems of the form

$$u'(t) = Ju(t) + f(t)$$

for constant real matrices J. We can restrict the analysis to the case $f(t) \equiv 0$, i.e.:

$$u'(t) = Ju(t), \tag{2.14}$$

because, the concepts of a dissipative system and a contractive method do not depend on $f(t)$.

A Runge-Kutta method with stability function $R(z)$ applied to the system (2.14) can be written in the form

$$u_{j+1} = R(\tau J)u_j. \tag{2.15}$$

Then, of course, the system (2.14) is dissipative, if and only if

$$(Jv, v) \leq 0 \quad \text{for all}\, v \in \mathbb{R}^N,$$

and the Runge-Kutta method is contractive if and only if

$$\|w_{j+1} - v_{j+1}\| = \|R(\tau J)(w_j - v_j)\| \leq \|(w_j - v_j)\|,$$

i.e., if and only if

$$\|R(\tau J)\| \leq 1.$$

At first, we consider the case that J is a normal matrix, i.e.:

$$J^*J = JJ^*.$$

(J^* denotes the adjoint matrix of J with respect to the scalar product $(.,.)$, so, e.g., $J^* = J^T$ if the Euclidean scalar product $(.,.) = (.,.)_{\ell_2}$ is used.)

For normal matrices J the system (2.14) is dissipative if and only if

$$\operatorname{Re}\lambda \leq 0 \quad \text{for all}\ \lambda \in \sigma(J).$$

Proof. The scalar product $(.,.)$ on $\mathbb{R}^n$ can be extended to a scalar product on $\mathbb{C}^n$, denoted again by $(.,.)$:

$$(z, z') = (x, x') + (y, y') + i\left[(x', y) - (x, y')\right].$$

For normal matrices it follows:

$$J = UDU^* \quad \text{with } D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n),$$

where U is a unitary matrix ($U^*U = UU^* = I$). Or, in other words, there exists an orthonormal basis of eigenvectors. The system is dissipative if and only if

$$\operatorname{Re}(Jv, v) = (J \operatorname{Re} v, \operatorname{Re} v) + (J \operatorname{Im} v, \operatorname{Im} v) \leq 0 \quad \text{for all } v \in \mathbb{C}^n.$$

Now, for $v \in \mathbb{C}^n$, we have

$$\begin{aligned} \operatorname{Re}(Jv, v) &= \operatorname{Re}(UDU^*v, v) = \operatorname{Re}(DU^*v, U^*v) = \operatorname{Re}(Dw, w) \\ &= ((\operatorname{Re} D) \operatorname{Re} w, \operatorname{Re} w) + ((\operatorname{Re} D) \operatorname{Im} w, \operatorname{Im} w) \\ &= \sum_i \operatorname{Re} \lambda_i |w_i|^2 \end{aligned}$$

with $w = U^*v$. This immediately implies that the system is dissipative if and only in

$$\sum_i \operatorname{Re} \lambda_i |w_i|^2 \leq 0 \quad \text{for all } w \in \mathbb{C}^n.$$

i.e.:

$$\operatorname{Re} \lambda_i \leq 0 \quad \text{for all } i.$$

□

For normal matrices J we have

$$\|R(\tau J)\| = \max_{\lambda \in \sigma(J)} |R(\tau\lambda)|.$$

Proof. First we have by definition

$$\|R(\tau J\| = \sup_{0 \neq v \in \mathbb{R}^n} \frac{\|R(\tau J)v\|}{\|v\|}.$$

Then, with $v = v_1 + iv_2$ it follows

$$\sup_{0 \neq v \in \mathbb{C}^n} \frac{\|R(\tau J)v\|^2}{\|v\|^2} = \sup_{0 \neq v_1, v_2 \in \mathbb{R}^n} \frac{\|R(\tau J)v_1\|^2 + \|R(\tau J)v_2\|^2}{\|v_1\|^2 + \|v_2\|^2} = \|R(\tau J\|.$$

Let u_i denote the i-th column of U. Then a vector v can be written in the form

$$v = \sum_i \alpha_i u_i \quad \text{and} \quad \|v\|^2 = (v, v) = \sum_i |\alpha_i|^2.$$

Analogously, it follows that

$$R(\tau J)v = \sum_i \alpha_i \, R(\tau\lambda_i)u_i \quad \text{and} \quad \|R(\tau J)v\|^2 = \sum_i |\alpha_i|^2 \, |R(\tau\lambda_i)|^2.$$

Therefore,

$$\|R(\tau J)\|^2 = \sup_{0\neq v} \frac{\|R(\tau J)v\|^2}{\|v\|^2} = \sup_{0\neq\alpha} \frac{\sum_i |\alpha_i|^2 \, |R(\tau\lambda_i)|^2}{\sum_i |\alpha_i|^2} = \max_i |R(\tau\lambda_i)|^2.$$

□

Therefore, a Runge-Kutta method applied to (2.14) for a normal matrix J is contractive if and only if

$$|R(\tau\lambda)| \leq 1 \quad \text{for all } \lambda \in \sigma(J). \tag{2.16}$$

In general, this condition leads to restrictions for the step size τ.

Example: For the one-dimensional semi-discretized parabolic model problem in the symmetric case $K_h^T = K_T$ the corresponding matrix $J = -M_h^{-1}K_h$ is symmetric and, therefore normal in the scalar product $(.,.)_{M_h}$ and the eigenvalues $\lambda(J)$ of J are of the form

$$\lambda(J) = -\lambda(M_h^{-1}K_h) \leq 0.$$

Therefore, the system is dissipative and the explicit Euler method is contractive if

$$\tau \leq \frac{2}{\lambda_{\max}(M_h^{-1}K_h)}.$$

From (2.9) we know that $\lambda_{\max}(M_h^{-1}K_h) = O(1/h^2)$. Consequently,

$$\tau = O(h^2).$$

More precisely, it follows from (2.9) that it is sufficient to satisfy

$$\tau \leq \frac{h^2}{6}.$$

However, for an A-stable Runge-Kutta method applied to a dissipative linear system with a normal coefficient matrix, the condition (2.16) is always satisfied. Therefore, in this case the method is contractive.

Example: The implicit Euler method for the one-dimensional semi-discretized parabolic model problem in the symmetric case $K_h^T = K_T$ is contractive for all step sizes τ.

Surprisingly, the last statement that A-stable Runge-Kutta methods applied to dissipative linear systems with constant coefficient matrix are contractive for all step sizes τ is also correct for non-normal matrices J. This follows from the following theorem:

Theorem 2.5. *Let $(.,.)$ be a scalar product in $\mathbb{C}^n$, $A \in \mathbb{C}^{n\times n}$ and $R(z)$ a rational function, which is bounded on $\mathbb{C}^-$. Assume that*

$$\mathrm{Re}(Av, v) \leq 0 \quad \textit{for all } v \in \mathbb{C}^n.$$

Then:

$$\|R(A)\| \leq \sup_{z\in\mathbb{C}^-} |R(z)|.$$

Proof. Let A^* be the adjoint matrix of A

$$(A^*u, v) = (u, Av).$$

For

$$A(\omega) = \frac{\omega}{2}(A + A^*) + \frac{1}{2}(A - A^*)$$

we have:

$$\begin{aligned}
(A(\omega)v, v) &= \frac{\omega}{2}\left[(Av, v) + (A^*v, v)\right] + \frac{1}{2}\left[(Av, v) - (A^*v, v)\right] \\
&= \frac{\omega}{2}\left[(Av, v) + \overline{(Av, v)}\right] + \frac{1}{2}\left[(Av, v) - \overline{(Av, v)}\right] \\
&= \omega\ \mathrm{Re}(Av, v) + i\ \mathrm{Im}(Av, v) \\
&= \mathrm{Re}\,\omega\ \mathrm{Re}(Av, v) + i\ \left(\mathrm{Im}\,\omega\ \mathrm{Re}(Av, v) + \mathrm{Im}(Av, v)\right).
\end{aligned}$$

Hence

$$\mathrm{Re}(A(\omega)v, v) \leq 0 \quad \text{for all } v \in \mathbb{C}^n,$$

if $\mathrm{Re}\,\omega \geq 0$. Therefore, the rational function

$$\varphi(\omega) = (R(A(\omega))u, v)$$

has no poles in $\mathbb{C}^+ = \{z \in \mathbb{C} : \mathrm{Re}\,\lambda \geq 0\}$ for fixed vectors u and v . Then the maximum principle implies:

$$(R(A)u, v) = \varphi(1) \quad \leq \quad \sup_{y\in\mathbb{R}} |\varphi(iy)| \leq \sup_{y\in\mathbb{R}} \|R(A(iy))\|\|u\|\|v\|.$$

For the matrix $A(iy)$ we have:

$$A(iy) = i\,\frac{y}{2}(A + A^*) + \frac{1}{2}(A - A^*),$$

hence

$$A(iy)^* = -i\,\frac{y}{2}(A^* + A) + \frac{1}{2}(A^* - A) = -A(iy).$$

Therefore, the matrix $A(iy)$ is anti-symmetric and the eigenvalues are purely imaginary. So the matrix is normal and:

$$\|R(A(iy))\| \quad = \quad \sup_{z\in\sigma(A(iy))} |R(z)| \leq \sup_{z\in\mathbb{C}^-} |R(z)|.$$

In summary, we obtain:

$$(R(A)u, v) \leq \sup_{z \in \mathbb{C}^-} |R(z)| \, \|u\| \|v\|,$$

which immediately implies the statement

$$\|R(A)\| \leq \sup_{z \in \mathbb{C}^-} |R(z)|.$$

□

Example: For $J = -M_h^{-1} K_h$ we have

$$\begin{aligned} \operatorname{Re}\left(-M_h^{-1} K_h(\underline{v}_h + i\underline{w}_h), \underline{v}_h + i\underline{w}_h\right)_{M_h} &= -\operatorname{Re}(K_h(\underline{v}_h + i\underline{w}_h), \underline{v}_h + i\underline{w}_h)_{\ell_2} \\ &= -(K_h \underline{v}_h, \underline{v}_h)_{\ell_2} - (K_h \underline{w}_h, \underline{w}_h)_{\ell_2} \leq 0. \end{aligned}$$

Consequently an A-stable Runge-Kutta method for the system

$$M_h \underline{u}_h'(t) = \underline{f}_h(t) - K_h \underline{u}_h(t)$$

is contractive, if M_h is symmetric and positive definite and if

$$(K_h \underline{v}_h, \underline{v}_h)_{\ell_2} \geq 0 \quad \text{for all } \underline{v}_h \in \mathbb{R}^n.$$

For more general (in particular non-linear) systems the concept of A-stability is not sufficient.

Definition 2.7. *A Runge-Kutta method is called B-stable, if for all initial-value problems with*

$$(f(t, w) - f(t, v), w - v) \leq 0$$

the approximations satisfy:

$$\|w_{j+1} - v_{j+1}\| \leq \|w_j - v_j\|$$

for all step sizes $\tau > 0$.

Obviously, an B-stable method applied to a dissipative system is contractive for all step sizes $\tau > 0$.

Remark: A good implementation of a Runge-Kutta method requires an efficient step size control. In principle, one tries to choose the step size in each step such that the new local error does not exceed a prescribed tolerance.

If the derived estimates for the discretization error is applied to the fully discretized parabolic problem

$$\begin{aligned} \langle u'(t), v\rangle + a(u(t), v) &= \langle F(t), v\rangle \quad \text{for all } v \in V, \\ u(0) &= u_0 \end{aligned}$$

by the θ-method under the assumptions of contractivity and coerciveness of a, then one obtains the following estimate for the discretization error:

$$\|u_h^j - u(t_j)\|_H \le \|u_h^j - u_h(t_j)\|_H + \|u_h(t_j) - u(t_j)\|_H$$

with

$$\begin{aligned} \|u_h(t_j) - u(t_j)\|_H &\le \|u_{0h} - R_h u_0\|_H \, e^{-\mu_1 t_j/c} + \|(I - R_h)u(t_j)\|_H \\ &\quad + \int_0^{t_j} \|[(I - R_h)u(s)]'\|_H \, e^{-\mu_1 (t_j - s)/c} \, ds \\ \|u_h^j - u_h(t_j)\|_H &\le \|u_h^0 - u_{0h}\|_H + \tau \int_0^{t_j} \|u_h''(s)\|_H \, ds = \tau \int_0^{t_j} \|u_h''(s)\|_H \, ds. \end{aligned}$$

If one wants to estimate the discretization error only in dependence of the exact solution $u(t)$ of the continuous problem, then a similar strategy can be used as for the semi-discretized problem:

The error is split into to parts:

$$u_h^j - u(t_j) = \left[u_h^j - R_h u(t_j)\right] + \left[R_h u(t_j) - u(t_j)\right] = \theta_h^j + \rho_h^j,$$

which can be estimated separately. Then one obtains the following result:

$$\begin{aligned} \|u_h^j - u(t_j)\|_H &\le \|u_{0h} - R_h u_0\|_H + \|(I - R_h)u(t_j)\|_H \\ &\quad + \int_0^{t_j} \|\left[(I - R_h)u(s)\right]'\|_H \, ds + \tau \int_0^{t_j} \|u''(s)\|_H \, ds. \end{aligned}$$

Hence, e.g., for the θ-method with $\theta \ge 1/2$ and the Courant element:

$$\|u_h^j - u(t_j)\|_H = \begin{cases} O(\tau + h^2) & \text{for } \theta \in (1/2, 1], \\ O(\tau^2 + h^2) & \text{for } \theta = 1/2. \end{cases}$$

For $\theta < 1/2$ the estimates are valid only under the strong restriction on the step size τ, e.g.:

$$\tau \le \frac{h^2}{6(1 - 2\theta)}$$

for the symmetric one-dimensional model problem.

Chapter 3

Hyperbolic Differential Equations

3.1 Initial-Boundary Value Problems for Hyperbolic Differential Equations

Classical Formulation:

Let $Q_T = \Omega \times (0,T)$. Find $u : \overline{Q}_T \longrightarrow \mathbb{R}$ such that the differential equation

$$\frac{\partial^2 u}{\partial t^2}(x,t) + Lu(x,t) \;=\; f(x,t) \quad (x,t) \in \; Q_T$$

with

$$Lv(x) = -\operatorname{div}(A(x)\operatorname{grad} v(x)) + c(x)v(x)$$

and the boundary conditions

$$\begin{aligned} u(x,t) &= g_D(x,t) \quad (x,t) \in \Gamma_D \times (0,T), \\ A(x)\operatorname{grad} u(x,t) \cdot n(x) &= g_N(x,t) \quad (x,t) \in \Gamma_N \times (0,T) \end{aligned}$$

and the initial conditions

$$\begin{aligned} u(x,0) &= u_0(x) \quad x \in \overline{\Omega}, \\ \frac{\partial u}{\partial t}(x,0) &= v_0(x) \quad x \in \overline{\Omega} \end{aligned}$$

are satisfied.

Special case wave equation:

$$\frac{\partial^2 u}{\partial t^2}(x,t) - \Delta u(x,t) = f(x,t).$$

Model problem:

$$\begin{aligned} u_{tt}(x,t) - u_{xx}(x,t) &= f(x,t) \quad (x,t) \in (0,1) \times (0,T), \\ u(0,t) &= 0 \quad t \in (0,T), \\ u(1,t) &= 0 \quad t \in (0,T), \\ u(x,0) &= u_0(x) \quad x \in [0,1], \\ u_t(x,0) &= v_0(x) \quad x \in [0,1]. \end{aligned}$$

Variational Formulation:

Find $u \in L^2((0,T),V) = X$ with

$$u' \in L^2((0,T),H), \quad u'' \in L^2((0,T),V^*),$$

such that

$$\begin{aligned} \langle u''(t), v\rangle + a(u(t),v) &= \langle F(t), v\rangle \quad \text{for all } v \in V, \\ u(0) &= u_0, \\ u'(0) &= v_0. \end{aligned}$$

Hence:

$$\begin{aligned} u''(t) + Au(t) &= F(t), \\ u(0) &= u_0, \\ u'(0) &= v_0. \end{aligned}$$

Or in the following form:

$$\begin{aligned} \frac{d^2}{dt^2}(u(t),v)_H + a(u(t),v) &= \langle F(t), v\rangle \quad \text{for all } v \in V, \\ u(0) &= u_0, \\ u'(0) &= v_0. \end{aligned}$$

The derivatives are to be understood as generalized derivatives:

$$\begin{aligned} \int_0^T \varphi(t)u'(t)\,dt &= -\int_0^T \varphi'(t)u(t)\,dt \quad \text{for all } \varphi \in C_0^\infty(0,T). \\ \int_0^T \varphi(t)u''(t)\,dt &= \int_0^T \varphi''(t)u(t)\,dt \quad \text{for all } \varphi \in C_0^\infty(0,T). \end{aligned}$$

For studying the existence and uniqueness of the solution of an initial-boundary value problem for a hyperbolic differential equation the problem is transformed to a first order system by introducing

$$v(t) = u'(t)$$

With

$$u(t) = u_0 + \int_0^t v(s)\ ds \equiv (Sv)(t).$$

one obtains

$$\begin{aligned} v'(t) + (\mathcal{A}v)(t) &= F(t), \\ v(0) &= v_0. \end{aligned}$$

where

$$\mathcal{A} : X \longrightarrow X^*, \quad (\mathcal{A}v)(t) \equiv A(Sv)(t).$$

It is easy to see that $S : X \longrightarrow X$ is Lipschitz continuous:

$$\begin{aligned} \|Sw - Sv\|_X^2 &= \int_0^T \|(Sw - Sv)(t)\|_V^2\ dt = \int_0^T \left\| \int_0^t [w(s) - v(s)]\ ds \right\|_V^2\ dt \\ &\leq \int_0^T \left[\int_0^t \|w(s) - v(s)\|_V\ ds \right]^2\ dt \leq \int_0^T t \int_0^t \|w(s) - v(s)\|_V^2\ ds\ dt \\ &\leq \int_0^T t \int_0^T \|w(s) - v(s)\|_V^2\ ds\ dt \\ &= \frac{1}{2} T^2 \|w - v\|_X^2 \end{aligned}$$

If A is symmetric, bounded and coercive in V, then:

$$\begin{aligned} |\langle \mathcal{A}v - \mathcal{A}w, u \rangle| &= \left| \int_0^T \langle A(Sv - Sw)(t), u(t) \rangle\ dt \right| = \left| \int_0^T a((Sv - Sw)(t), u(t))\ dt \right| \\ &\leq \mu_2 \int_0^T \|(Sv - Sw)(t)\|_V \|u(t)\|_V\ dt \leq \mu_2\, \|Sv - Sw\|_X\, \|u\|_X \\ &\leq \frac{\mu_2 T}{\sqrt{2}}\, \|w - v\|_X \|u\|_X \end{aligned}$$

and

$$\begin{aligned}
\langle \mathcal{A}v - \mathcal{A}w, v - w\rangle &= \int_0^T \langle A(Sv - Sw)(t), v(t) - w(t)\rangle \, dt \\
&= \int_0^T a((Sv - Sw)(t), v(t) - w(t)) \, dt \\
&= \int_0^T a((Sv - Sw)(t), (Sv - Sw)'(t)) \, dt \\
&= \int_0^T \frac{d}{dt}\frac{1}{2} a((Sv - Sw)(t), (Sv - Sw)(t)) \, dt \\
&= \frac{1}{2}\, a((Sv - Sw)(T), (Sv - Sw)(T)) \geq 0.
\end{aligned}$$

That means that the operator $\mathcal{A} : X \longrightarrow X^*$ is monotone and Lipschitz continuous.

By the simple transformation

$$v_\lambda(t) = e^{-\lambda t} v(t)$$

with $\lambda > 0$ one obtains the system

$$\begin{aligned}
v_\lambda(t)' + (\mathcal{A}_\lambda v_\lambda)(t) &= F_\lambda(t) \\
v_\lambda(0) &= v_0
\end{aligned}$$

with

$$F_\lambda(t) = e^{-\lambda t} F(t) \quad \text{and} \quad (\mathcal{A}_\lambda v_\lambda)(t) = e^{-\lambda t}(\mathcal{A}v)(t) + \lambda\, v_\lambda(t).$$

It can easily be shown that $\mathcal{A}_\lambda : X \longrightarrow X^*$ is even strongly monotone and Lipschitz continuous. The special structure of $\mathcal{A}$ (Volterra operator: the value of $(\mathcal{A}v)(t)$ depends only on values $v(s)$, $s \in [0, t]$) and these properties are essential for an existence theory, which is analogous to the parabolic case: a-priori estimates, semi-discretization, compactness argument and limit process.

Semi-Discretization:

V is replaced by an finite-dimensional subspace V_h:

Find $u_h : [0, T] \longrightarrow V_h$ such that

$$\begin{aligned}
\frac{d^2}{dt^2}(u_h(t), v_h)_H + a(u_h(t), v_h) &= \langle F(t), v_h\rangle \quad \text{for all } v_h \in V_h, \\
(u_h(0), v_h)_H &= (u_0, v_h)_H \quad \text{for all } v_h \in V_h, \\
\frac{d}{dt}(u_h(0), v_h)_H &= (v_0, v_h)_H \quad \text{for all } v_h \in V_h.
\end{aligned}$$

If a basis $\{\varphi_i : i = 1, 2, \ldots, N_h\}$ of V_h is introduced, one obtains

$$\begin{aligned}
M_h \underline{u}_h''(t) + K_h \underline{u}_h(t) &= \underline{f}_h(t), \\
M_h u_h(0) &= \underline{g}_h, \\
M_h u_h'(0) &= \underline{h}_h
\end{aligned}$$

with the mass matrix

$$M_h = (M_{ij}), \quad M_{ij} = (\varphi_j, \varphi_i)_H,$$

the stiffness matrix

$$K_h = (K_{ij}), \quad K_{ij} = a(\varphi_j, \varphi_i)$$

and the vectors

$$\underline{u}_h(t) = (u_i(t)), \qquad \underline{f}_h(t) = (f_i), \quad f_i = \langle F(t), \varphi_i \rangle,$$

and

$$\underline{g}_h = (g_i), \quad g_i = (u_0, \varphi_i)_H \qquad \underline{h}_h = (h_i), \quad h_i = (v_0, \varphi_i)_H.$$

So one obtains an initial value problem. Standard form:

$$\begin{aligned} \underline{u}''(t) &= f(t, u(t)), \\ u(0) &= u_0, \\ u'(0) &= v_0. \end{aligned}$$

here with $u(t) = \underline{u}_h(t)$, $f(t, u(t)) = M_h^{-1}(\underline{f}_h(t) - K_h \underline{u}_h(t))$, $u_0 = M_h^{-1} g_h$ and $u_0 = M_h^{-1} g_h$.

3.2 Runge-Kutta Methods for Initial Value Problems of Second-Order Ordinary Differential Equations

In this section initial value problems for second-order ordinary differential equations are discussed. Typical applications are semi-discretized hyperbolic initial-boundary value problems.

The problem has the following general form:

Find a function $u(t)$ such that

$$\begin{aligned} u''(t) &= f(t, u(t)) \quad t \in [0, T], \qquad (3.1) \\ u(0) &= u_0, \\ u'(0) &= v_0. \end{aligned} \qquad (3.2)$$

By introducing $v(t) = u'(t)$ we obtain an equivalent first-order system:

$$\begin{aligned} \begin{pmatrix} u(t) \\ v(t) \end{pmatrix}' &= \begin{pmatrix} v(t) \\ f(t, u(t)) \end{pmatrix}, \\ \begin{pmatrix} u(0) \\ v(0) \end{pmatrix} &= \begin{pmatrix} u_0 \\ v_0 \end{pmatrix}. \end{aligned}$$

If the same Runge-Kutta method is applied to both first-order differential equations, then one obtains:

$$\begin{aligned} g_i &= u_j + \tau \sum_{k=1}^{s} a_{ik} h_k, \\ h_i &= v_j + \tau \sum_{k=1}^{s} a_{ik} f(t_j + c_k\tau, g_k), \\ u_{j+1} &= u_j + \tau \sum_{i=1}^{s} b_i h_i, \\ v_{j+1} &= v_j + \tau \sum_{i=1}^{s} b_i f(t_j + c_i\tau, g_i). \end{aligned}$$

Eliminating h_i results in:

$$\begin{aligned} g_i &= u_j + \tau \sum_{k=1}^{s} a_{ik}[v_j + \tau \sum_{l=1}^{s} a_{kl} f(t_j + c_l\tau, g_l)], \\ u_{j+1} &= u_j + \tau \sum_{i=1}^{s} b_i[v_j + \tau \sum_{k=1}^{s} a_{ik} f(t_j + c_k\tau, g_k)], \\ v_{j+1} &= v_j + \tau \sum_{i=1}^{s} b_i f(t_j + c_i\tau, g_i). \end{aligned}$$

Hence

$$\begin{aligned} g_i &= u_j + \tau\, c_i\, v_j + \tau^2 \sum_{k=1}^{s} \bar{a}_{ik} f(t_j + c_k\tau, g_k), \\ u_{j+1} &= u_j + \tau\, v_j + \tau^2 \sum_{i=1}^{s} \bar{b}_i f(t_j + c_i\tau, g_i), \\ v_{j+1} &= v_j + \tau \sum_{i=1}^{s} b_i f(t_j + c_i\tau, g_i) \end{aligned}$$

with

$$\bar{a}_{ik} = \sum_{l=1}^{s} a_{il} a_{lk}, \qquad \bar{b}_i = \sum_{k=1}^{s} b_k a_{ki}.$$

Example: For the θ-method one obtains:

$$\begin{aligned} u_{j+1} &= u_j + \tau\, v_j + \tau^2\theta\,[(1-\theta) f(t_j, u_j) + \theta f(t_{j+1}, u_{j+1})], \\ v_{j+1} &= v_j + \tau\,[(1-\theta) f(t_j, u_j) + \theta f(t_{j+1}, u_{j+1})]. \end{aligned}$$

The auxiliary variable v_k can be easily eliminated and one obtains:

$$u_{j+2} - 2u_{j+1} + u_j = \tau^2\,[\theta^2 f(t_{j+2}, u_{j+2}) + 2\theta(1-\theta) f(t_{j+1}, u_{j+1}) + (1-\theta)^2 f(t_j, u_j)].$$

Stability

We discuss the linear case (without loss of generality we can assume that $f(t) \equiv 0$):

$$Mu''(t) + Ku(t) = 0$$

under the assumptions $K^T = K > 0$ and $M^T = M > 0$. The equivalent first-order system reads:

$$\begin{pmatrix} u(t) \\ v(t) \end{pmatrix}' = \begin{pmatrix} 0 & I \\ -M^{-1}K & 0 \end{pmatrix} \begin{pmatrix} u(t) \\ v(t) \end{pmatrix}.$$

The matrix

$$J = \begin{pmatrix} 0 & I \\ -M^{-1}K & 0 \end{pmatrix}$$

is anti-symmetric with respect to the scalar product

$$\left(\begin{pmatrix} u_1 \\ v_1 \end{pmatrix}, \begin{pmatrix} u_2 \\ v_2 \end{pmatrix} \right)_{K,M} = (Ku_1, u_2)_{\ell_2} + (Mv_1, v_2)_{\ell_2}.$$

Therefore, J is a normal matrix, the eigenvalues are purely imaginary. Consequently, the system is dissipative.

We have for the stability function $R(z)$ of a Runge-Kutta method:

$$\|R(\tau\, J)\|_{K,M} = \max_{\mu \in \sigma(J)} |R(\tau\mu)|.$$

The eigenvalues of J:

$$\begin{pmatrix} 0 & I \\ -M^{-1}K & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \mu \begin{pmatrix} u \\ v \end{pmatrix}$$

Hence

$$Ku = -\mu^2\, Mu$$

If $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_N$ denote the eigenvalues of $M^{-1}K$, then one obtains for the eigenvalues of J:

$$\pm i\sqrt{\lambda_1},\ \pm i\sqrt{\lambda_2},\ \ldots\ \pm i\sqrt{\lambda_N}.$$

Since the system is dissipative, A-stable methods are contractive.

Example: The θ-method is A-stable (and, therefore, contractive) for $\theta \geq 1/2$. However, contrary to the parabolic case, the θ-method is never contractive for $\theta < 1/2$, since the stability domain contains, except for 0, no further points of the imaginary axis.

3.3 Partitioned Runge-Kutta Methods

If two not necessarily equal Runge-Kutta methods are applied to the two differential equations, one obtains a so-called partitioned Runge-Kutta method:

$$\begin{aligned} g_i &= u_j + \tau \sum_{k=1}^{s} a_{ik} h_k, \\ h_i &= v_j + \tau \sum_{k=1}^{s} a'_{ik} f(t_j + c_k\tau, g_k), \\ u_{j+1} &= u_j + \tau \sum_{i=1}^{s} b_i h_i, \\ v_{j+1} &= v_j + \tau \sum_{i=1}^{s} b'_i f(t_j + c_i\tau, g_i). \end{aligned}$$

By eliminating h_i one obtains:

$$\begin{aligned} g_i &= u_j + \tau \sum_{k=1}^{s} a_{ik}[v_j + \tau \sum_{l=1}^{s} a'_{kl} f(t_j + c_l\tau, g_l)], \\ u_{j+1} &= u_j + \tau \sum_{i=1} b_i[v_j + \tau \sum_{k=1}^{s} a'_{ik} f(t_j + c_k\tau, g_k)], \\ v_{j+1} &= v_j + \tau \sum_{i=1}^{s} b'_i f(t_j + c_i\tau, g_i). \end{aligned}$$

So

$$\begin{aligned} g_i &= u_j + \tau\, c_i\, v_j + \tau^2 \sum_{k=1}^{s} \bar{a}_{ik} f(t_j + c_k\tau, g_k), \\ u_{j+1} &= u_j + \tau\, v_j + \tau^2 \sum_{i=1}^{s} \bar{b}_i f(t_j + c_i\tau, g_i), \\ v_{j+1} &= v_j + \tau \sum_{i=1}^{s} b'_i f(t_j + c_i\tau, g_i) \end{aligned}$$

with

$$\bar{a}_{ik} = \sum_{l=1}^{s} a_{il} a'_{lk}, \qquad \bar{b}_i = \sum_{k=1}^{s} b_k a'_{ki}. \tag{3.3}$$

Example: For the θ method with parameter θ for the first equation and with parameter $1-\theta$ for the second equation one obtains:

$$\begin{aligned} u_{j+1} &= u_j + \tau\, v_j + \tau^2 \theta\, [\theta f(t_j, u_j) + (1-\theta) f(t_{j+1}, u_{j+1})], \\ v_{j+1} &= v_j + \tau\, [\theta f(t_j, u_j) + (1-\theta) f(t_{j+1}, u_{j+1})]. \end{aligned}$$

The auxiliary variables v_k can easily be eliminated:

$$u_{j+2} - 2u_{j+1} + u_j = \tau^2 \left[\theta(1-\theta)\, f(t_{j+2}, u_{j+2}) + 2\theta(1-\theta)\, f(t_{j+1}, u_{j+1}) + \theta(1-\theta)\, f(t_j, u_j)\right]$$

or with $\sigma = \theta(1-\theta)$:

$$u_{j+2} - 2u_{j+1} + u_j = \tau^2 \left[\sigma\, f(t_{j+2}, u_{j+2}) + (1-2\sigma)\, f(t_{j+1}, u_{j+1}) + \sigma\, f(t_j, u_j)\right].$$

Observe that, for $\theta = 0$ and $\theta = 1$ (i.e. for $\sigma = 0$) the total method is explicit, although it is the combination of an explicit and an implicit Euler method.

Stability analysis

Assume that the differential equation is of the form

$$Mu''(t) + Ku(t) = 0$$

with $M^T = M > 0$ and $K^T = K > 0$. Only the case $\theta = 0$ is considered. The equivalent first-order system reads

$$\begin{pmatrix} u(t) \\ v(t) \end{pmatrix}' = \begin{pmatrix} 0 & I \\ -M^{-1}K & 0 \end{pmatrix} \begin{pmatrix} u(t) \\ v(t) \end{pmatrix}.$$

The partitioned method then becomes the one-step method

$$\begin{aligned} u_{j+1} &= u_j + \tau\, v_j, \\ v_{j+1} &= v_j - \tau\, M^{-1}K u_{j+1} = -\tau\, M^{-1}K u_j + (I - \tau^2\, M^{-1}K) v_j, \end{aligned}$$

or after eliminating v_k, the two-step method ($\sigma = 0$):

$$u_{j+2} - 2u_{j+1} + u_j + \tau^2 M^{-1}K u_{j+1} = 0.$$

So

$$\begin{pmatrix} u_{j+1} \\ v_{j+1} \end{pmatrix} = \begin{pmatrix} I & \tau\, I \\ -\tau\, M^{-1}K & I - \tau^2\, M^{-1}K \end{pmatrix} \begin{pmatrix} u_j \\ v_j \end{pmatrix}. \tag{3.4}$$

The stability analysis can be reduced to scalar problems by transforming to the basis of eigenvectors of $M^{-1}K$: Assume that the matrix $M^{-1}K$ has the eigenvalues ν_i^2 with corresponding eigenvectors e_i, $i = 1, \ldots, N$ with $(Me_i, e_j)_{\ell_2} = \delta_{ij}$. From the ansatz

$$u_k = \sum_i \alpha_i^k\, e_i \quad \text{and} \quad v_k = \sum_i \beta_i^k\, e_i$$

one obtains

$$(Ku_k, u_k)_{\ell_2} = \sum_i (\nu_i \alpha_i^k)^2 \quad \text{and} \quad (Mv_k, v_k)_{\ell_2} = \sum_i (\beta_i^k)^2$$

and, therefore,

$$\left\| \begin{pmatrix} u_k \\ v_k \end{pmatrix} \right\|_{K,M}^2 = (Ku_k, u_k)_{\ell_2} + (Mv_k, v_k)_{\ell_2} = \sum_i \left\| \begin{pmatrix} \nu_i \alpha_i^k \\ \beta_i^k \end{pmatrix} \right\|_{\ell_2}^2 .$$

From (3.4) the following conditions are obtained

$$\begin{pmatrix} \nu_i \alpha_i^{j+1} \\ \beta_i^{j+1} \end{pmatrix} = \begin{pmatrix} 1 & \tau\nu_i \\ -\tau\nu_i & 1-(\tau\nu_i)^2 \end{pmatrix} \begin{pmatrix} \nu_i \alpha_i^j \\ \beta_i^j \end{pmatrix} = G_i \begin{pmatrix} \nu_i \alpha_i^j \\ \beta_i^j \end{pmatrix}.$$

Assume that there are symmetric and positive definite 2×2 matrices H_i with

$$\|G_i\|_{H_i} \leq 1 \quad \text{for all } i = 1, \ldots, N. \tag{3.5}$$

Then it follows:

$$\sum_i \left\| \begin{pmatrix} \nu_i \alpha_i^{j+1} \\ \beta_i^{j+1} \end{pmatrix} \right\|_{H_i}^2 = \sum_i \left\| G_i \begin{pmatrix} \nu_i \alpha_i^j \\ \beta_i^j \end{pmatrix} \right\|_{H_i}^2 \leq \sum_i \left\| \begin{pmatrix} \nu_i \alpha_i^j \\ \beta_i^j \end{pmatrix} \right\|_{H_i}^2 .$$

This shows that the method is contractive with respect to the norm

$$\left\| \begin{pmatrix} u_k \\ v_k \end{pmatrix} \right\|_*^2 = \sum_i \left\| \begin{pmatrix} \nu_i \alpha_i^k \\ \beta_i^k \end{pmatrix} \right\|_{H_i}^2 .$$

It remains to discuss the condition (3.5).

A necessary condition for (3.5) is that all eigenvalues of the matrices G_i

$$1 - \frac{\tau^2 \nu_i^2}{2} \pm \sqrt{\left(1 - \frac{\tau^2 \nu_i^2}{2}\right)^2 - 1}$$

are less or equal to 1. This is satisfied if and only if

$$\tau\nu_i \leq 2 \quad \text{for all } i = 1, \ldots, N.$$

Hence

$$\tau \leq \frac{2}{\sqrt{\lambda_{\max}(M^{-1}K)}}.$$

Example: Application to the semi-discretized one-dimensional hyperbolic model problem:

$$\lambda_{\max}(M_h^{-1} K_h) \leq \frac{12}{h^2}.$$

So, for

$$\tau \leq \frac{1}{\sqrt{3}}\, h$$

the eigenvalues of G_i are less than or equal to 1. This condition on the step size τ is less restrictive than in the parabolic case!

The following theorem describes necessary and sufficient conditions for (3.5):

Theorem 3.1 (Kreiss' Matrix-Theorem). *Let $\mathcal{F}$ be a family of d-by-d matrices. Then the following conditions are equivalent:*

(A) There is a constant C_A with

$$\|A^n\| \le C_A \quad \text{for all } n \in \mathbb{N},\ A \in \mathcal{F}.$$

(R) There is a constant C_R with

$$\|(z\,I - A)^{-1}\| \le \frac{C_R}{|z| - 1} \quad \text{for all } z \in \mathbb{C} \text{ with } |z| > 1,\ A \in \mathcal{F}.$$

(S) There are constants C_S and $C_B \ge 0$ and, for each matrix $A \in \mathcal{F}$ there is a non-singular matrix S with

(a) $\max(\|S\|, \|S^{-1}\|) \le C_S$,

(b) $B = SAS^{-1}$ is an upper triangular matrix and

$$|B_{ij}| \le C_B \min(1 - |\kappa_i|, 1 - |\kappa_j|) \tag{3.6}$$

for all $i \ne j$, where κ_j denote the diagonal elements of B, which are the eigenvalues of B and of A.

(H) There is a constant C_H and, for each matrix $A \in \mathcal{F}$ there is a Hermitian and positive definite matrix H which

$$C_H^{-1}\,\|v\| \le \|v\|_H \le C_H\,\|v\| \quad \text{for all } v \in \mathbb{C}^d$$

and

$$\|A\|_H \le 1.$$

If this theorem is applied to the family $\mathcal{F} = \{G_i : i = 1, \ldots, N\}$ of 2-by2 matrices, then it easily follows: If

$$\tau\nu_i \le 2 - \varepsilon \quad \text{for all } i = 1, \ldots, N$$

for an arbitrary but fixed number $\varepsilon > 0$, then the method is contractive in the corresponding norm.

Proof. The eigenvalues of G_i are given by

$$\lambda_\pm(c) = 1 - \frac{c^2}{2} \pm ic\sqrt{1 - \frac{c^2}{4}}$$

with $c = \tau\nu_i$. The corresponding eigenvectors define the transformation matrix

$$X = \begin{pmatrix} 1 & 1 \\ \dfrac{\lambda_+(c)-1}{c} & \dfrac{\lambda_-(c)-1}{c} \end{pmatrix}$$

with inverse

$$X^{-1} = \frac{c}{\lambda_+(c)-\lambda_-(c)} \begin{pmatrix} \dfrac{\lambda_-(c)-1}{c} & -1 \\ -\dfrac{\lambda_+(c)-1}{c} & 1 \end{pmatrix}$$

From

$$|\lambda_\pm(c)| = 1, \quad \left|\frac{\lambda_\pm(c)-1}{c}\right| = 1 \quad \text{and} \quad \left|\frac{c}{\lambda_+(c)-\lambda_-(c)}\right| = \frac{1}{\sqrt{4-c^2}} \le \frac{1}{4\varepsilon-\varepsilon^2}$$

it follows that

$$\|X\|_\infty = 2 \quad \text{and} \quad \|X^{-1}\|_\infty \le \frac{2}{4\varepsilon-\varepsilon^2}$$

and, therefore,

$$\|G_i^n\|_\infty = \|X\,\mathrm{diag}(\lambda_+(x)^n, \lambda_-(x)^n)X^{-1}\|_\infty \le \frac{4}{4\varepsilon-\varepsilon^2} = C_A.$$

□

The norm introduced above is equivalent to the original norm. Hence the convergence also holds in the original norm.

Remark: A further generalization leads to the larger class of Runge-Kutta-Nyström methods: For these methods the relation (3.3) is ignored. Then the tableau of coefficients has the form:

c_1	$\bar a_{11}$	$\bar a_{12}$	$\dots$	$\bar a_{1,s-1}$	$\bar a_{1s}$
c_2	$\bar a_{21}$	$\bar a_{22}$	$\dots$	$\bar a_{2,s-1}$	$\bar a_{2s}$
$\vdots$					
c_s	$\bar a_{s1}$	$\bar a_{s2}$	$\dots$	$\bar a_{s,s-1}$	$\bar a_{ss}$
	b_1	b_2	$\dots$	b_{s-1}	b_s
	$\bar b_1$	$\bar b_2$	$\dots$	$\bar b_{s-1}$	$\bar b_s$

or, in compact form:

c	$\bar A$
	b^T
	$\bar b^T$

The special case

0	0	0
1	$\bar b_1$	$\bar b_2$
	b_1	b_2
	b_1	b_2

leads to the method:

$$\begin{aligned} u_{j+1} &= u_j + \tau\, v_j + \tau^2\, [\bar{b}_1 f(t_j, u_j) + \bar{b}_2 f(t_{j+1}, u_{j+1})], \\ v_{j+1} &= v_j + \tau\, [b_1 f(t_j, u_j) + b_2 f(t_{j+1}, u_{j+1})]. \end{aligned}$$

The consistency order is 1, if

$$b_1 + b_2 = 1.$$

The consistency order is 2, if

$$b_1 = b_2 = \frac{1}{2} \qquad \text{and} \qquad \bar{b}_1 + \bar{b}_2 = \frac{1}{2}.$$

Consistency order 2 for the first equation and consistency order 1 for the second equation is obtained, if

$$b_1 + b_2 = 1 \qquad \text{and} \qquad \bar{b}_1 + \bar{b}_2 = \frac{1}{2}.$$

This leads to the so-called Newmark method:

$$\begin{aligned} u_{j+1} &= u_j + \tau\, v_j + \frac{\tau^2}{2}\, [(1 - 2\beta) f(t_j, u_j) + 2\beta f(t_{j+1}, u_{j+1})], \\ v_{j+1} &= v_j + \tau\, [(1 - \gamma) f(t_j, u_j) + \gamma f(t_{j+1}, u_{j+1})]. \end{aligned}$$

Bibliography

[1] Dietrich Braess. *Finite Elemente. Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie.* Berlin: Springer-Verlag, 2003.

[2] Herbert Gajewski, Konrad Gröger, and Klaus Zacharias. *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen.* Berlin: Akademie-Verlag, 1974.

[3] Christian Großmann and Hans-Görg Roos. *Numerik partieller Differentialgleichungen.* Stuttgart: B. G. Teubner, 1994.

[4] Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner. *Solving ordinary differential equations. I: Nonstiff problems.* Berlin: Springer-Verlag, 1993.

[5] Ernst Hairer and Gerhard Wanner. *Solving ordinary differential equations. II: Stiff and differential-algebraic problems.* Berlin: Springer-Verlag, 1996.

[6] Michael Jung and Ulrich Langer. *Methode der finiten Elemente für Ingenieure. Eine Einführung in die numerischen Grundlagen und Computersimulation.* . Stuttgart: Teubner, 2001.

[7] P. Knabner and L. Angermann. *Numerik partieller Differentialgleichungen. Eine anwendungsorientierte Einführung.* Berlin: Springer-Verlag, 2000.

[8] Eberhard Zeidler. *Nonlinear functional analysis and its applications. II/A: Linear monotone operators.* New York: Springer-Verlag, 1990.

[9] Eberhard Zeidler. *Nonlinear functional analysis and its applications. II/B: Nonlinear monotone operators.* New York: Springer-Verlag, 1990.