Freie Universität Berlin

FB Mathematik und Informatik

Numerische Mathematik/Scientific Computing

Notes on Numerical Methods for ODEs and Numerical Linear Algebra

(Numerics II)

Ralf Kornhuber

1. Edition: Wintersemester 07/08
– corrected version from October 2011 –

Compiled and typeset by Hanne Hardering

Contents

1 Stiff Differential Equations **1**
- 1.1 Stability of Solutions of ODEs . . . 1
- 1.2 Stability of Linear Recursions . . . 10
- 1.3 Preserving Stability: Linear Systems . . . 11
- 1.4 Collocation and Gauß Methods . . . 15
- 1.5 Dissipative Systems and A-stability of Gauß Methods . . . 20
- 1.6 Preserving Asymptotic Stability: Nonlinear Systems . . . 24
- 1.7 Algorithmic Aspects of Implicit RK's (Gauß Methods) . . . 28
  - 1.7.1 Fixed point iteration . . . 29
  - 1.7.2 Newton iteration and simplifications . . . 30
- 1.8 Linearly Implicit One-Step-Methods . . . 32
- 1.9 Extrapolation Methods . . . 35
- 1.10 Gradient Flows and Parabolic PDEs . . . 37

2 Differential Algebraic System **43**
- 2.1 Motivation . . . 43
- 2.2 Linear DAEs: Existence and Uniqueness . . . 45
- 2.3 Nonlinear Semi-explicit DAEs . . . 51

3 Hamiltonian Systems **54**
- 3.1 Energy and Symplecticity . . . 54
- 3.2 Symplectic Runge-Kutta-Methods . . . 58

4 Iterative Methods for Linear Systems **62**
- 4.1 Motivation (Why Iterative Solutions?) . . . 62
- 4.2 Linear Iterative Schemes . . . 66
- 4.3 Preconditioning and Linear Iterations . . . 68
- 4.4 Linear Descent Methods . . . 73
- 4.5 Nonlinear Descent Methods . . . 77
  - 4.5.1 Gradient Methods (Steepest Descent) . . . 77
  - 4.5.2 Conjugate Gradient Methods (CG Methods) . . . 80
  - 4.5.3 Generalized minimal residual method (GMRes) . . . 87

1 Stiff Differential Equations

We already know ordinary differential equations (ODEs) from Numerics I where we dealt with existence, uniqueness, and condition of solutions as well as the numerical treatment of ODEs by explicit and implicit Runge-Kutta methods. We saw that implicit methods are not always applicable, e.g. $f(x) = x^2$. We will now consider the following examples to motivate the use of implicit methods.

Example (Population of bacteria) The growth of a population of bacteria is described by the ODE [9]

$$x' = qx - kx^2, \qquad t > 0, \qquad x(0) = 1$$

The exact solution takes the form

$$x(t) = \frac{qe^{qt}}{q + (e^{qt} - 1)k} \to \frac{q}{k} \text{ for } t \to \infty$$

Using different matlab routines yields results of different quality

ode45: based on an explicit Runge-Kutta formula: very small timesteps $\to$ inefficient
ode23t: based on the implicit trapezoidal rule: adapted timesteps $\to$ efficient

Example (Chemical reaction system) We consider a chemical reaction system described by

$$\begin{array}{rcll} A & \overset{c_1=0.04}{\to} & B & \text{slow} \\ 2B & \overset{c_2=3\cdot 10^7}{\to} & B + C & \text{very fast} \\ B + C & \overset{c_3=10^4}{\to} & A + C & \text{fast} \end{array}$$

Thus, we get the ODE

$$\begin{array}{rrrrr} A: & x_1' = & -c_1 x_1 & & +c_3 x_2 x_3 \\ B: & x_2' = & c_1 x_1 & -c_2 x_2^2 & -c_3 x_2 x_3 \\ C: & x_3' = & & c_2 x_2^2 & \end{array}$$

A differential equation is called stiff if it is not efficiently solvable by explicit discretization methods. This is the case if the solution being sought is varying slowly but there are nearby solutions that vary rapidly, so the numerical method must take small steps to obtain satisfactory results.

1.1 Stability of Solutions of ODEs

We consider the initial value problem (IVP)

$$x' = f(x), \qquad 0 < t < \infty, \qquad x(0) = x_0, \qquad f : \mathbb{R}^d \to \mathbb{R}^d \text{ continuously differentiable} \tag{1.1}$$

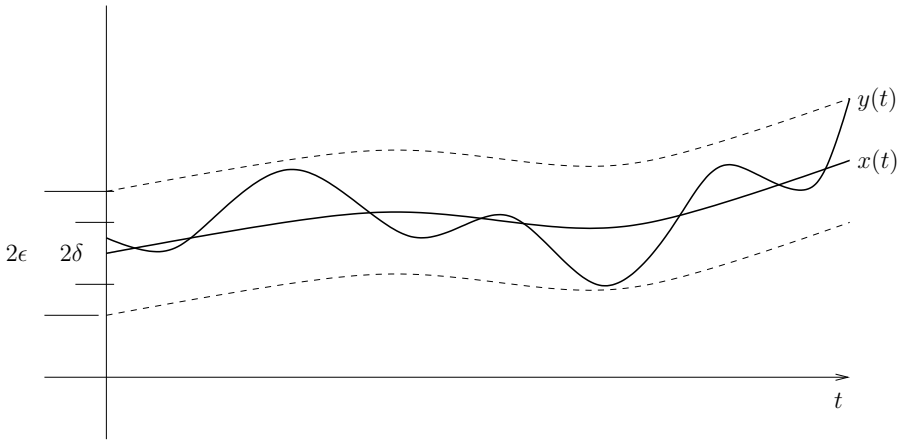
and the perturbed IVP:

$$y' = f(y), \qquad y(0) = x_0 + \delta x_0 \tag{1.2}$$

Definition 1.1.1 (stability of solutions) *Let $x(t)$ be the solution of (1.1).*

1. *$x(t)$ is called* stable *if*

$$\forall \epsilon > 0 \; \exists \delta > 0 : \; |\delta x_0| < \delta \;\Rightarrow\; \begin{array}{l}(i) \text{ solution } y(t) \text{ of (1.2) exists} \\ (ii) \;\; \sup_{t\in[0,\infty)} \|x(t) - y(t)\| < \epsilon\end{array}$$



2. *$x(t)$ is called* asymptotically stable *if*
   - *$x(t)$ is stable*
   - $\exists \delta_0 : |\delta x_0| < \delta_0 \;\Rightarrow \lim_{t\to\infty} \|x(t) - y(t)\| = 0$

Example The solution of the bacteria problem is asymptotically stable:

$$x(t) = \frac{x_0 q e^{qt}}{q + (e^{qt} - 1)k x_0} \to \begin{cases} 0 & q < 0 \\ \frac{q}{k} & q > 0 \end{cases}$$

Observation: The bacteria problem has the *fixed point* $x^* = \frac{q}{k}$ for $q > 0$ and $\|x(t) - x^*\| \to 0$.

Definition 1.1.2 *$x^* \in \mathbb{R}^d$ is called* fixed point *of (1.1) if $x(t) \equiv x^*$ solves (1.1) for $x_0 = x^*$. x^* is (asymptotically) stable if and only if $x(t) \equiv x^*$ is (asymptotically) stable.*

From now on we will only consider the stability of fixed points.
Without loss of generality we can assume that $x^* = 0$: For $x^* \neq 0$ we consider the ODE

$$y' = g(y), \quad y := x - x^*, \quad g(y) := f(y + x^*)$$

with fixed point $y^* = 0$.

Linear case

We start with the linear IVP

$$x' = Ax, \quad x(0) = x_0, \quad A \in \mathbb{R}^{d\times d}, \quad x_0 \in \mathbb{R}^d \tag{1.3}$$

with fixed point $x^* = 0$.

Theorem 1.1.3 *The flow operator Φ^t of (1.3) takes the form*

$$\Phi^t = e^{tA}$$

denoting

$$\exp(tA) = e^{tA} := \sum_{k=0}^{\infty} \frac{1}{k!}(tA)^k \tag{1.4}$$

This series converges uniformly on finite time intervals $[0, T]$.

Proof Let t be fixed. The sequence (s_n) defined by

$$s_n = \sum_{k=0}^{n} \frac{1}{k!}(tA)^k$$

is a Cauchy sequence in $\mathbb{R}^{d\times d}$. Since $\mathbb{R}^{d\times d}$ is complete, we get

$$s_n \to s = e^{tA}.$$

The majorant criterion yields uniform convergence for all $t \in [0, T]$.
It remains to show that $e^{tA}x_0$ is a solution of (1.3).
Termwise differentiation yields:

$$\begin{aligned}\frac{d}{dt}e^{tA}x_0 &= \sum_{k=0}^{\infty} \frac{1}{k!}\frac{d}{dt}(tA)^k x_0 \\ &= \sum_{k=1}^{\infty} \frac{1}{(k-1)!}A(tA)^{k-1}x_0 \\ &= Ae^{tA}x_0\end{aligned}$$

The initial value satisfies

$$e^{tA}x_0|_{t=0} = x_0 + (tA)|_{t=0} + \frac{1}{2}(tA)^2|_{t=0} + \ldots = x_0.$$

□

Lemma 1.1.4 *The matrix exponential defined in (1.4) has the properties*

(i) $e^{t(TAT^{-1})} = Te^{tA}T^{-1}, \quad \forall T \in \mathbb{R}^{d\times d}$ *regular*

(ii) $e^{t(A+B)} = e^{tA}e^{tB}, \ \forall B \in \mathbb{R}^{d\times d}$ *with* $AB = BA$

(iii) $A = \text{blockdiag}(A_1, \ldots, A_k) \;\Rightarrow\; e^{tA} = \text{blockdiag}(e^{tA_1}, \ldots, e^{tA_k})$

(iv) $e^{\alpha I} = e^{\alpha} I, \quad \alpha \in \mathbb{R}, \quad I = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} \in \mathbb{R}^{d \times d}$

Proof We show *(i)*:

$$(TAT^{-1})^k = TA^kT^{-1}$$

$$e^{t(TAT^{-1})} = \sum_{k=0}^{\infty} \frac{1}{k!}(t(TAT^{-1}))^k = T\left(\sum_{k=0}^{\infty} \frac{1}{k!}(tA)^k\right)T^{-1} = Te^{tA}T^{-1}$$

(ii)-(iv) Exercise. □

Let

$$p(\lambda) = \det(A - \lambda I)$$

denote the characteristic polynomial of A.
Then, by definition,

$$p(\lambda) = 0 \;\Leftrightarrow\; \lambda \text{ eigenvalue of } A$$

and

$$e \in \ker(A - \lambda I) \;\Leftrightarrow\; e \text{ eigenvector of } A \Leftrightarrow\; Ae = \lambda e$$

Let further $s(\lambda_0)$ denote the algebraic multiplicity of λ_0, i.e.,

$$p(\lambda) = (\lambda - \lambda_0)^{s(\lambda_0)} q(\lambda), \qquad q(\lambda_0) \neq 0$$

and

$$r(\lambda_0) = \dim \ker(A - \lambda_0 I) > 0$$

denote the geometric multiplicity. Recall $r(\lambda_0) \leq s(\lambda_0)$.

Lemma 1.1.5 (Jordan normal form) *Let $\sigma(A) = \{\lambda_1, \ldots, \lambda_m\}$ be the spectrum of A, λ_k pairwise different.*
Then there is a regular matrix $T \in \mathbb{C}^{d \times d}$:

$$TAT^{-1} = J = \text{blockdiag}(J_1, \ldots, J_{m^*}), \quad d \geq m^* \geq m.$$

The Jordan blocks J_i, $i = 1, \ldots, m^*$, take the form

$$J_i = J_i(\lambda_{k_i}) = \begin{pmatrix} \lambda_{k_i} & 1 & & 0 \\ & \lambda_{k_i} & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_{k_i} \end{pmatrix} \in \mathbb{R}^{n_i \times n_i}$$

with corresponding $\lambda_{k_i} \in \sigma(A)$. Conversely, for each $\lambda_k \in \sigma(A)$ there are $r(\lambda_k)$ Jordan blocks $J_{i_j}(\lambda_k) \in \mathbb{C}^{n_{i_j} \times n_{i_j}}$, $j = 1, \ldots, r(\lambda_k)$, with $\sum_{j=1}^{r(\lambda_k)} n_{i_j} = s(\lambda_k)$.

Remark

$$J_i = J_i(\lambda_{k_i}) = \lambda_{k_i} I + N$$

with

$$N = \begin{pmatrix} 0 & 1 & & 0 \\ & 0 & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{pmatrix} \in \mathbb{R}^{n_i \times n_i}$$

nilpotent, i.e., $N^{n_i - 1} \neq 0,\ N^{n_i} = 0$.
If $r(\lambda_k) = s(\lambda_k)$, then $n_{ij} = 1$, $j = 1, \ldots, r(\lambda_k)$. Hence, $r(\lambda_k) = s(\lambda_k) \quad \forall \lambda_k \in \sigma(A)$ implies that J is diagonal.
Notation:

$$\lambda = \Re\lambda + i \cdot \Im\lambda \in \mathbb{C}, \quad \Re\lambda, \Im\lambda \in \mathbb{R}, \quad |e^{\lambda}| = |e^{\Re\lambda}| \cdot |e^{i \cdot \Im\lambda}| = |e^{\Re\lambda}|$$

Proposition 1.1.6 *The stability of fixed points can be characterized as follows*

(i) $x^* = 0$ *is stable if*

- $\Re\lambda \leq 0$ *for all* $\lambda \in \sigma(A)$
- $\Re\lambda = 0 \ \Rightarrow \ s(\lambda) = r(\lambda)$

(ii) $x^* = 0$ *is asymptotically stable if* $\Re\lambda < 0$ *holds for all* $\lambda \in \sigma(A)$

(iii) If $\Re\lambda < \alpha \in \mathbb{R}$ *holds for all* $\lambda \in \sigma(A)$, *then*

$$\exists C > 0 : \ \|e^{tA}\| \leq C e^{t\alpha} \quad \forall t \geq 0$$

Proof We show (iii): Lemma 1.1.5 and Lemma 1.1.4 yield

$$\begin{aligned} T e^{tA} T^{-1} &= e^{tJ} \\ &= \text{blockdiag}(e^{tJ_1}, \ldots, e^{tJ_{m^*}}) \\ \Rightarrow \|e^{tA}\| &\leq \|T\| \cdot \|T^{-1}\| \max_{i=1,\ldots,m^*} \|e^{tJ_i}\| \end{aligned}$$

and for a Jordan block $J_i = \lambda_{k_i} I + N \in \mathbb{R}^{n_i \times n_i}$ holds

$$\begin{aligned} e^{tJ_i} &= e^{\lambda_{ki} t I} e^{tN} \\ &= e^{\lambda_{ki} t} \left(I + tN + \frac{1}{2}(tN)^2 + \ldots + \frac{1}{(n_i - 1)!}(tN)^{n_i - 1} \right). \end{aligned}$$

Let $\Re\lambda_{ki} < \alpha \in \mathbb{R}$

$$\begin{aligned} \|e^{tJ_i}\| &\leq |e^{\lambda_{ki} t}| \left(1 + t\|N\| + \ldots + \frac{1}{(n_i - 1)!} t^{n_i - 1} \|N\|^{n_i - 1} \right) \\ &= e^{\Re\lambda_{ki} t} \left(1 + t\|N\| + \ldots + \frac{1}{(n_i - 1)!} t^{n_i - 1} \|N\|^{n_i - 1} \right) \\ &\leq C e^{\alpha t} \end{aligned}$$

$\Rightarrow$ *(iii)*.

We show (ii): Let $\Re\lambda < -\alpha < 0$

$$\|e^{tA}x_0\| \leq Ce^{-\alpha t}\|x_0\| \to 0 \text{ for } t \to \infty.$$

We show (i), second case: Assume that $\Re\lambda = 0$. By assumption, we have $s(\lambda) = r(\lambda)$ and therefore $J_k(\lambda) = (\lambda) \in \mathbb{R}^{1\times 1}$.

$$\Rightarrow \|e^{tJ_k}\| \leq e^{t\Re\lambda} \cdot 1 = 1$$
$$\Rightarrow \|e^{tA}x_0\| \leq \|T\| \cdot \|T^{-1}\| \cdot 1 \cdot \|x_0\| \;\Rightarrow\; (i).$$

□

Example We consider the scalar case $\lambda \in \mathbb{C}$

$$x' = \lambda x, \quad x(0) = x_0, \quad x(t) = x_0 e^{\lambda t}$$

$\Re\lambda < 0 \Rightarrow\; |x(t)| \leq |x_0|e^{tRe\lambda} \to 0$ for $t \to \infty \;\Rightarrow\; x^* = 0$ asymptotically stable
$\Re\lambda = 0 \Rightarrow\; x^* = 0$ stable
$\Re\lambda > 0 \Rightarrow\; x^* = 0$ unstable

Example (Monomolecular reaction) We consider the monomolecular reaction

$$A \xrightarrow{c} B$$

which can be described by the ODE

$$\begin{aligned} A: &\quad x_1' &&= -cx_1 \\ B: &\quad x_2' &&= cx_1 \end{aligned}$$

Rewritten in matrix formulation the ODE takes the form

$$x' = \begin{pmatrix} -c & 0 \\ c & 0 \end{pmatrix} x$$

with the eigenvalues

$$p(\lambda) = \lambda(\lambda + c) \;\Rightarrow\; \lambda_1 = -c, \quad \lambda_2 = 0$$

Therefore, $x^* = 0$ is stable but not asymptotically stable:

$$x(0) = \delta_0 = \begin{pmatrix} 0 \\ \epsilon \end{pmatrix} \;\Rightarrow\; x_1 = 0, \quad x_2 = \epsilon$$

Remark Properties of A characterize the stability of all fixed points of $x' = Ax$. We say $x' = Ax$ is (asymptotically) stable if and only if $x^* = 0$ is (asymptotically) stable.

Notation: $\nu(A) = \max_{\lambda\in\sigma(A)} \Re\lambda$ is called *spectral abscissa*.

Example We consider the following initial value problem

$$x' = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} x, \quad x(0) = \begin{pmatrix} 0 \\ 1 \end{pmatrix} = x_0, \quad \Phi^t x_0 = \begin{pmatrix} \sin t \\ \cos t \end{pmatrix}$$

with the eigenvalues

$$p(\lambda) = \lambda^2 + 1 \;\Rightarrow\; \lambda_1 = i, \quad \lambda_2 = -i$$

By Proposition (1.1.6) the fixed point $x^* = 0$ is stable but not asymptotically stable: Let $\varepsilon > 0$ be arbitrary. Then

$$\Phi^t \begin{pmatrix} 0 \\ \varepsilon \end{pmatrix} = \varepsilon \begin{pmatrix} \sin t \\ \cos t \end{pmatrix} \not\to 0 \text{ for } t \to \infty$$

Example We consider $x' = Ax$ with

$$A = \begin{pmatrix} -2 & 1 & & 0 \\ 1 & -2 & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & -2 \end{pmatrix} \in \mathbb{R}^{d \times d}$$

A has the eigenvalues

$$\lambda_i = -4 \sin^2 \left(\frac{i}{2(d+1)} \pi \right), \quad i = 1, \ldots, d$$

Thus,

$$\nu(A) = -4 \sin^2 \left(\frac{1}{2(d+1)} \pi \right) < 0.$$

Therefore, $x' = Ax$ is asymptotically stable.

Attention Our results do not carry over to $x' = A(t)x$ (exercise).

Nonlinear case

We consider

$$x' = f(x)$$

Our (natural) hope is that (asymptotic) stability is inherited from the linearization of $f(x)$ at a fixed point x^*

$$x' = \underbrace{f(x^*)}_{=0} + Ax, \quad A = Df(x)|_{x=x^*}$$

Example In general, stability is not inherited from the linearization. Consider the ODE

$$\begin{aligned} x_1' &= x_1^3 - x_2 \\ x_2' &= x_1 \end{aligned}$$

$x^* = 0$ is a fixed point.

$$A = Df(x)|_{x=x^*} = \begin{pmatrix} 3x_1^2 & -1 \\ 1 & 0 \end{pmatrix}|_{x=0} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

$$\lambda_1 = -i, \quad \lambda_2 = i$$

Therefore, $x' = Ax$ is stable.
Let $x_0 = \delta \in \mathbb{R}^2$, $x_0 \neq 0$. Then (exercise)

$$V(\Phi^t x_0) \to \infty \text{ for } t \to t_+ < \infty, \quad V(x) = x_1^2 + x_2^2$$

Conclusion If x^* is a stable fixed point of the linearized system, this does not imply that x^* is a stable fixed point of the original nonlinear system. Terms of higher order might dominate.

Theorem 1.1.7 *Let $x^* \in \mathbb{R}^d$ be a fixed point of*

$$x' = f(x),$$

i.e., $f(x^*) = 0$, and f is continuously differentiable. If $\nu(Df(x^*)) < 0$, then x^* is asymptotically stable.

Proof We can assume $x^* = 0$ without loss of generality.
By definition of the derivative $A = Df(x^*)$, we get $\|f(x) - (\underbrace{f(x^*)}_{=0} + Ax)\| = o(\|x\|)$.

Hence,

$$f(x) = Ax + g(x), \quad g(x) = o(\|x\|).$$

1. Let $x_0 \in \mathbb{R}^d$ and let $\Phi^t x_0$ exist for $t \in [0, T)$, $T \in \mathbb{R} \cup \{\infty\}$.
   Then
   $$\Phi^t x_0 := e^{tA} x_0 + \int_0^t \exp((t-s)A) g(\Phi^s x_0) ds$$
   is a solution (variation of constants, see, e.g., CoMa II).

   $$\begin{aligned} \frac{d}{dt}\Phi^t x_0 &= Ae^{tA} x_0 + \exp((t-s)A) g(\Phi^s x_0)|_{s=t} + \int_0^t A \exp((t-s)A) g(\Phi^s x_0) ds \\ &= A\phi^t x_0 + g(\Phi^t x_0) \\ \Phi^t x_0|_{t=0} &= x_0 \end{aligned}$$

2. To show:
   $$\|\Phi^t x_0\| \leq Ce^{-\beta t}, \quad \beta > 0$$
   We will use the Gronwall's lemma:

   $$\begin{aligned} & \Psi(t) \leq a + b \int_0^t \Psi(s) ds, \quad 0 \leq t \leq t^* \\ \Rightarrow \; & \Psi(t) \leq ae^{bt}, \quad 0 \leq t \leq t^* \end{aligned}$$

Choose $\beta \in \mathbb{R}$ such that $\nu(A) < -\beta < 0$.
Theorem 1.1.6 (iii) implies

$$\exists C > 0 : \ \| \exp(tA)\| \leq Ce^{-t\beta}.$$

Hence,

$$\|\Phi^t x_0\| \leq Ce^{-t\beta}\|x_0\| + C\int_0^t e^{-(t-s)}\|g(\Phi^s x_0)\|ds.$$

Since $g(x) = o(\|x\|)$

$$\exists \delta_0 > 0 : \ \|g(x)\| \leq \frac{\beta}{2C}\|x\| \quad \forall \|x\| < \delta_0.$$

Choose $\|x_0\| < \delta_0$ and $t^* \leq T$ such that $\|\Phi^t x_0\| < \delta_0 \quad \forall t \in [0, t^*]$. Then

$$\|\Phi^t x_0\| \leq Ce^{-t\beta}\|x_0\| + C\frac{\beta}{2C}\int_0^t e^{-(t-s)}\|\Phi^s x_0\|ds.$$

Use Gronwall's lemma for $\Psi(t) := e^{\beta t}\|\Phi^t x_0\|$

$$\Psi(t) \leq C\|x_0\| + \frac{\beta}{2}\int_0^t \Psi(s)ds, \quad 0 \leq t \leq t^*$$

$$\Rightarrow \ \Psi(t) \leq C\|x_0\|e^{\frac{\beta}{2}t}$$

$$\Rightarrow \ \|\Phi^t x_0\| \leq C\|x_0\|e^{-\frac{\beta}{2}t}$$

Choose $\|x_0\| < \min\left\{\delta_0, \frac{\delta_0}{C}\right\}$. Then

$$\|\Phi^t x_0\| < \delta_0 e^{-\frac{\beta}{2}t} \leq \delta_0 \quad \forall t \in [0, T].$$

This means that there is no blow-up for $t \to T$. Hence, $T = \infty$ and $\lim_{t\to\infty} \Phi^t x_0 = 0$.

□

Add-on The Theorem of Grobman/Hartman [4, p. 115] yields an even stronger result: If $\sigma_0 = \{\lambda \in \sigma | \Re\lambda = 0\} = \emptyset$, then there exists a continuous invertible coordinate transformation h such that

$$h(\Phi^t x) = \exp(tA)h(x) \quad x \in U(x^*)$$

Example We consider the growth of bacteria again

$$f(x) = qx - kx^2$$

$q > 0$: $x^* = \frac{q}{k}$

$$f'(x) = q - 2kx$$

$$f'(x)|_{x=x^*} = q - 2k\frac{q}{k} = -q < 0$$

Therefore, x^* is an asymptotically stable fixed point.

$q = 0$: $x^* = 0$

$$f'(x)|_{x=0} = 0$$

Therefore, 1.1.7 is not applicable.
$x^* = 0$ is not stable (exercise)

Generalizations

- quasi-stationary states: "almost" fixed points (van der Pool, reaction)
- metastable states (conformations): "almost" fixed subsets

1.2 Stability of Linear Recursions

Motivation Let $x' = Ax$ be (asymptotically) stable.
The explicit Euler discretization takes the form

$$\begin{aligned} x_{k+1} &= x_k + \tau A x_k = \Psi^\tau x_k \\ &= (\Psi^\tau)^{k+1} x_0, \qquad \Psi^\tau = I + \tau A \in \mathbb{R}^{d\times d} \end{aligned}$$

Any (explicit) Runge-Kutta method takes the form

$$x_{k+1} = Bx_k, \qquad \Psi^\tau = B \in \mathbb{R}^{d\times d}$$

Question Is stability inherited from the continuous problem by discretization?

Definition 1.2.1 *Consider the linear recursion*

$$x_{k+1} = Bx_k = B^{k+1}x_0, \qquad k = 0, 1, \ldots \tag{1.5}$$

(1.5) is called stable if $\sup_{k\in\mathbb{N}} \|B^k\| < C$.
(1.5) is called asymptotically stable if $\lim_{k\to\infty} \|B^k\| = 0$.

Remark (Asymptotic) stability is invariant under similarity transformations.

Proof

$$\begin{aligned} (TBT^{-1})^k &= TB^kT^{-1} \\ \|(TBT^{-1})^k\| &= \|TB^kT^{-1}\| \le \|T\|\|T^{-1}\|\|B^k\| \end{aligned}$$

□

Theorem 1.2.2 *Let* $\rho(B) = \max_{\lambda\in\sigma(B)} |\lambda|$ *be the spectral radius of* B.

1. *If* $\rho(B) \le 1$ *and every eigenvalue* $\lambda \in \sigma(B)$ *with* $|\lambda| = 1$ *fulfills* $r(\lambda) = s(\lambda)$, *then* $x_{k+1} = Bx_k$ *is stable.*
2. *If* $\rho(B) < 1$, *then* $x_{k+1} = Bx_k$ *is asymptotically stable.*

Proof [4, Theorem 3.33]

Example The explicit Euler method

$$\begin{aligned} x_{k+1} &= x_k + \tau A x_k = (I + \tau A)x_k, \qquad k = 0, 1, \dots \\ \Rightarrow\ \Psi^\tau &= B = (I + \tau A) \\ \lambda \in \sigma(A) \Rightarrow\ 1 + \tau\lambda &\in \sigma(B) \end{aligned}$$

is asymptotically stable if

$$|1 + \tau\lambda|^2 = (1 + \tau\Re\lambda)^2 + (\tau\Im\lambda)^2 < 1.$$

$\Re\lambda < 0 \quad \forall\lambda \in \sigma(A) \ \Rightarrow\ x' = Ax$ asymptotically stable $\not\Rightarrow\ |1 + \tau\lambda|^2 < 1$.

Example

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

$x' = Ax$ is stable

$$\sigma(A) = \{i, -i\}$$

$$\rho(I + \tau A) = \{|1 + \tau i|, |1 - \tau i|\} = (1 + \tau^2)^{\frac{1}{2}} > 1$$

Therefore, $x_{k+1} = (I + \tau A)x_k$ is unstable for all $\tau > 0$. Thus, stability is not inherited by the explicit Euler method.

1.3 Preserving Stability: Linear Systems

We consider the linear system

$$x' = Ax, \qquad A \in \mathbb{R}^{d\times d} \tag{1.6}$$

and a Runge-Kutta method

$$x_{k+1} = \Psi^\tau x_k \tag{1.7}$$

We are interested in the question of inheritance of stability, i.e.,

$$\text{(1.6) (asymp.) stable } \Rightarrow \text{ (1.7) (asymp.) stable?}$$

Proposition 1.1.6 and Theorem 1.2.2 provide the result that stability is governed by the eigenvalues of A and Ψ^τ. This motivates to consider Dahlquist's test equation

$$x' = \lambda x \tag{1.8}$$

$$\begin{aligned} \text{(1.8) stable } &\Leftrightarrow\ \lambda \in \mathbb{C}_- = \{z \in \mathbb{C} | \Re z \le 0\} \\ \text{(1.8) asymp. stable } &\Leftrightarrow\ \lambda \in \mathring{\mathbb{C}}_- = \{z \in \mathbb{C} | \Re z < 0\} \end{aligned}$$

Application of a Runge-Kutta method Ψ^τ with uniform step size $\tau > 0$ to (1.8) leads to

$$x_{k+1} = R(\lambda, \tau)x_k, \qquad \Psi^\tau = R(\lambda, \tau) \in \mathbb{R}$$

Theorem 1.3.1 *We consider the Runge-Kutta method Ψ^τ of stage s given by the Butcher scheme* $\begin{array}{c|c} \mathbb{A} \\ \hline b^T \end{array}$, *i.e.,*

$$\Psi^\tau x = x + \tau \sum_{i=1}^{s} b_i k_i$$
$$k_i = f\left(x + \tau \sum_{j=1}^{s} a_{ij} k_j \right)$$

There exists $\tau^* > 0$ such that application of Ψ^τ to (1.8) yields

$$x_{k+1} = R(\lambda\tau) x_k \quad \forall \tau < \tau^*$$

where

$$R(z) = \frac{P(z)}{Q(z)}$$

with uniquely determined mutually prime polynomials P and Q with $\deg P, \deg Q \leq s$ normalized by $P(0) = Q(0) = 1$.

Proof Inserting $f(x) = \lambda x$ we get the linear system

$$\Psi^\tau x = x + \tau \sum_{i=1}^{s} b_i k_i$$
$$k_i = \lambda \left(x + \tau \sum_{j=1}^{s} a_{ij} k_j \right), \quad i = 1, \ldots, s$$

1. $\lambda = 0 \;\Rightarrow\; k_i = 0 \;\Rightarrow\; \Psi^\tau = 1$
2. $\lambda \neq 0$
   Let $y = (y_i)_{i=1}^s$ with $y_i := \frac{1}{\lambda} k_i$. Then

$$\Psi^\tau = 1 + \tau\lambda \sum_{i=1}^{s} b_i y_i = 1 + z \sum_{i=1}^{s} b_i y_i$$
$$y_i = 1 + \tau\lambda \sum_{j=1}^{s} a_{ij} y_j = 1 + z \sum_{j=1}^{s} a_{ij} y_j, \quad i = 1, \ldots, s$$

   with $z := \tau\lambda$.
   Consider the linear system in matrix form

$$My := (I - z\mathbb{A})y = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

$$\det(I - z\mathbb{A}) = \det(I - \tau\lambda\mathbb{A}) =: g(\tau)$$

Then $g \in C(\mathbb{R})$ and $g(0) = 1$ and therefore,

$$\exists \tau^* : \ \det(I - \tau\lambda\mathbb{A}) \neq 0 \quad \forall \tau < \tau^*$$

Cramer's rule yields

$$\begin{aligned} y_i \ &= \tfrac{\det M^{(i)}}{\det M}, \qquad \text{with} \quad M = (M_1, \ldots, M_s) = (I - z\mathbb{A}) \\ &\qquad\qquad\qquad\qquad M^{(i)} = (M_1, \ldots, M_{i-1}, \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, M_{i+1}, \ldots, M_s) \\ &=: \tfrac{P_i(z)}{\tilde{Q}(z)} \end{aligned}$$

where $\deg P_i \leq s - 1$, $\deg \tilde{Q} \leq s$. Hence,

$$\Psi^\tau = 1 + z \sum_{i=1}^{s} b_i y_i = \frac{\tilde{Q}(z) + z \sum_{i=1}^{s} b_i P_i(z)}{\tilde{Q}(z)} =: \frac{P(z)}{Q(z)}.$$

If Ψ^τ is explicit, then $\det(I - z\mathbb{A}) = 1$

□

Definition 1.3.2 *The rational function $R(z)$, $z \in \mathbb{C}$, associated with the Runge-Kutta method Ψ^τ according to Theorem 1.3.1 is called* stability function *of Ψ^τ.*

Example

1) explicit Euler: $\Psi^\tau x = x + \tau f(x)$
application to (1.8): $\Psi^\tau = R(\tau\lambda) = 1 + \tau\lambda$
$\Rightarrow \ R(z) = 1 + z$

2) implicit Euler: $\Psi^\tau x = x + \tau f(\Psi^\tau x)$
application to (1.8): $R(\tau\lambda) = 1 + \tau\lambda R(\tau\lambda)$
$\Rightarrow \ R(z) = \dfrac{1}{1 - z}$

3) trapezoidal rule: $\Psi^\tau x = x + \dfrac{1}{2}\tau\big(f(x) + f(\Psi^\tau x)\big)$
application to (1.8): $R(\tau\lambda) = 1 + \dfrac{1}{2}\tau\lambda(1 + R(\tau\lambda))$
$\Rightarrow \ R(z) = \dfrac{1 + \frac{z}{2}}{1 - \frac{z}{2}}$

4) Runge-Kutta-4:
$\Rightarrow \ R(z) = 1 + z + \dfrac{1}{2}z^2 + \dfrac{1}{3!}z^3 + \dfrac{1}{4!}z^4$

Proposition 1.3.3 *If Ψ^τ is consistent with order p, then*

$$R(z) = e^z + O(z^{p+1}) \text{ for } z \to 0$$

Proof Exercise

Theorem 1.3.4 *The condition*

$$\mathbb{C}_- \subset S = \{z \in \mathbb{C} || R(z)| \leq 1\} \tag{1.9}$$

implies

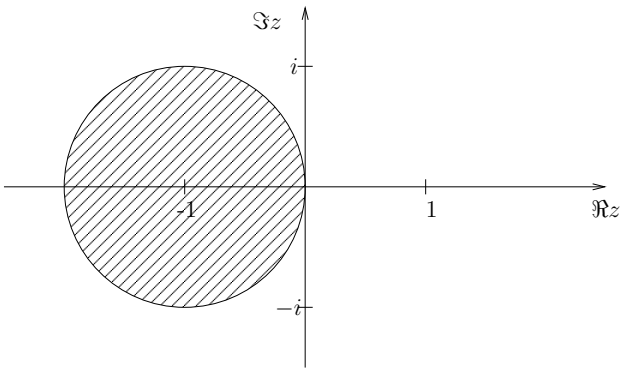
$$\textit{stability of } x' = \lambda x \;\Rightarrow\; \textit{stability of } x_{k+1} = \Psi^\tau x_k, \qquad \forall \lambda \in \mathbb{C}.$$

Proof Obvious

Definition 1.3.5 *S is called* stability domain *of Ψ^τ.*
If the stability domain satisfies (1.9), then Ψ^τ is called A-stable.

Example 1. The stability function of the explicit Euler method is $R(z) = 1 + z$. Thus, the stability domain is given by

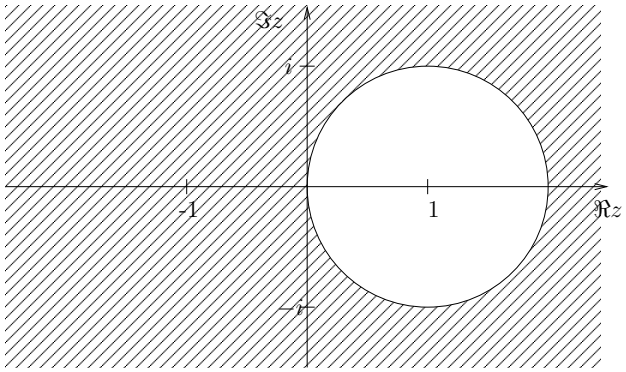
$$S = \{z \in |1 + z| \leq 1\}.$$



Hence, the explicit Euler method is not A-stable.

2. The stability function of the implicit Euler method is $R(z) = \frac{1}{1-z}$. Thus, the stability domain is given by

$$S = \left\{ z \in \left| \frac{1}{1-z} \right| \leq 1 \right\}.$$



Hence, the implicit Euler method is A-stable.

Remark 1. (1.9) even implies

$$\text{(asymp.) stability of } x' = \lambda x \;\Rightarrow\; \text{(asymp.) stability of } x_{k+1} = \Psi^\tau x_k$$

2. This assertion directly extends to linear systems. [4, Theorem 6.13]

Proposition 1.3.6 *If Ψ^τ is explicit, then it is not A-stable.*

Proof Let $\mathbb{C}_- \subset S$. Then S is unbounded. Let $(z_k) \subset S$ with $|z_k| \to \infty$. Because Ψ^τ is explicit, $R(z) = P(z)$ is a polynomial. Therefore, $|R(z)| \leq 1$ cannot be true contradicting $(z_k) \subset S$. □

Definition 1.3.7 *Ψ^τ is called* L-stable *if Ψ^τ is A-stable and*

$$R(\infty) := \lim_{z\to\infty} R(z) = 0$$

Example 1. For the implicit Euler method holds

$$R(z) = \frac{1}{1-z} \to 0 \text{ for } z \to \infty$$

Thus the implicit Euler method is L-stable.

2. For the implicit trapezoidal rule holds

$$R(z) = \frac{1+\frac{z}{2}}{1-\frac{z}{2}}$$

Thus the implicit trapezoidal rule is A-stable but not L-stable.

Remark For a general criterion for L-stability of RK-methods see [4, Lemma 6.32].

1.4 Collocation and Gauß Methods

We consider

$$x' = f(t,x), \qquad f : \mathbb{R}\times\mathbb{R}^d \to \mathbb{R}^d$$

Basic idea of collocation Construct a vector-valued function $u : \mathbb{R} \to \mathbb{R}^d$ of d polynomials of degree s such that

$$u(t) = x$$

and the collocation conditions

$$u'(t+c_i\tau) = f(t+c_i\tau, u(t+c_i\tau)), \qquad i = 1,\dots,s, \qquad 0 \leq c_1 \leq \dots \leq c_s \leq 1$$

hold. Set

$$\Psi^\tau x := u(t+\tau)$$

Remark u fulfills the ODE exactly in the collocation points $t + c_i\tau$, $i = 1,\dots,s$.

Construction of u Consider the Lagrange basis

$$L_i(\theta) = \prod_{\substack{j=1 \\ j\neq i}}^{s} \frac{\theta - c_j}{c_i - c_j}.$$

Let

$$k_i = u'(t + c_i\tau), \quad i = 1, \ldots, s.$$

Lagrange interpolation formula [9]:

$$u'(t + \theta\tau) = \sum_{j=1}^{s} k_j L_j(\theta)$$

Set

$$\begin{aligned} g(\theta) &:= u(t + \theta\tau) \\ g'(\theta) &= \tau u'(t + \theta\tau) \end{aligned}$$

The fundamental theorem yields

$$\begin{aligned} g(c_i) &= g(0) + \int_0^{c_i} g'(\theta)d\theta \\ u(t + c_i\tau) &= u(t) + \tau \int_0^{c_i} u'(t + \theta\tau)d\theta \\ &= x + \tau \sum_{j=1}^{s} \int_0^{c_i} L_j(\theta)d\theta \; k_j \\ &= x + \tau \sum_{j=1}^{s} a_{ij}k_j \end{aligned}$$

where $a_{ij} = \int_0^{c_i} L_j(\theta)d\theta$, $\mathbb{A} = (a_{ij})_{i,j=1}^{s}$.
Substitution into the collocation condition provides

$$k_i = f(t + c_i\tau, x + \tau \sum_{j=1}^{s} a_{ij}k_j), \quad i = 1, \ldots, s. \tag{1.10}$$

Substitution into the ansatz yields

$$\begin{aligned} \Psi^\tau x &= u(t + \tau) \\ &= x + \tau \int_0^1 u'(t + \theta\tau)d\theta \\ &= x + \tau \sum_{j=1}^{s} \int_0^1 L_j(\theta)d\theta \; k_j \\ &= x + \tau \sum_{j=1}^{s} b_j k_j \end{aligned}$$

where $b_j = \int_0^1 L_j(\theta)d\theta$, $b = (b_j)_{j=1}^s$.

Thus, we obtain a Runge-Kutta method $\begin{array}{c|c} & \mathbb{A} \\ \hline & b^T \end{array}$ of stage s.

Remark The method is characterized by $c_1, \ldots, c_s$.
We did not check whether Ψ^τ is feasible, i.e., whether $k_i = k_i(\tau)$ fulfilling (1.10) exist or not.

Proposition 1.4.1 *Let $f : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ be sufficiently smooth. Then for all $x \in \mathbb{R}^d$ and $t \in \mathbb{R}$ exists $\tau^* > 0$ such that the nonlinear system (1.10) is uniquely solvable in a neighborhood of $k_i(0) = f(t, x)$.*

Proof For convenience we assume $d = 1$. The system (1.10) can be written as

$$F(k, \tau) = 0, \qquad F = \begin{pmatrix} F_1 \\ \vdots \\ F_s \end{pmatrix}, \qquad k = \begin{pmatrix} k_1 \\ \vdots \\ k_s \end{pmatrix}$$

$$F_i = k_i - f(t + c_i\tau, x + \tau \sum_{j=1}^s a_{ij}k_j).$$

Resolution with respect to k:

$$F(k(\tau), \tau) = 0, \quad \tau < \tau^*$$

$$F(k(0), 0) = 0 \Rightarrow \; k(0) = \begin{pmatrix} f(t, x) \\ \vdots \\ f(t, x) \end{pmatrix}$$

Differentiation with respect to k yields

$$D_k F = I - \tau \left(f'(t + c_i\tau, x + \tau \sum_{j=1}^s a_{ij}k_j)a_{ij} \right)_{ij}$$

$$D_k F(k(0), 0) = I, \quad \det I > 0.$$

The implicit function theorem provides the assertion:
There exist $\tau^* > 0$ and $\delta > 0$ such that there exists a unique $k = k(\tau)$ fulfilling $F(k(\tau), \tau) = 0$ for $0 < \tau < \tau^*$ and $\|k(\tau) - k(0)\| < \delta$. □

Remark Proposition 1.4.1 can be applied to arbitrary implicit Runge-Kutta methods.

Example Consider the ODE

$$x' = x^2.$$

The implicit Euler method takes the form

$$x_{k+1} = x_k + \tau x_{k+1}^2.$$

For simplicity we assume $x_k = 0$.

$$x_{k+1} - \tau x_{k+1}^2 = 0$$

$$\Rightarrow x_{k+1,1} = 0, \quad x_{k+1,2} = \frac{1}{\tau}$$

The correct local value is $x_{k+1} = 0$.

Special case $x' = f(t)$

$$k_i = f(x + c_i\tau), \qquad a_{ij} \text{ does not enter!}$$
$$\Psi^\tau x = x + \tau \sum_{j=1}^{s} b_j k_j, \qquad b_j = \int_0^1 L(\theta) d\theta$$

Equidistant nodes $c_i = \frac{i-1}{s},\ i = 1, \ldots, s$ lead to well-known Newton-Côtes formulas. [9]

Example $s = 1$: $c_1 = 0$

$$L_1(\theta) \equiv 1$$

$$a_{11} = \int_0^{c_1} 1 \; d\theta = 0$$
$$b_1 = \int_0^1 1 \; d\theta = 1$$

$$\begin{array}{c|c} 0 \\ \hline 1 \end{array}$$ explicit Euler

$s = 2$: $c_1 = 0, \quad c_2 = 1$

$$L_1(\theta) = \frac{\theta - c_2}{c_1 - c_2} = 1 - \theta, \quad L_2(\theta) = \frac{\theta - c_1}{c_2 - c_1} = \theta$$

$$a_{11} = \int_0^{c_1} 1 - \theta d\theta = 0 \quad a_{12} = \textstyle\int_0^{c_1} \theta d\theta = 0$$
$$a_{21} = \int_0^{c_2} 1 - \theta d\theta = \frac{1}{2} \quad a_{22} = \textstyle\int_0^{c_2} \theta d\theta = \frac{1}{2}$$

$$b_1 = b_2 = \frac{1}{2}$$

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$ implicit trapezoidal rule

Collocation methods have s free parameters c_i instead of $s + s^2$ free parameters b_i and a_{ij} in general.
Hope: some nice conditions on b, $\mathbb{A}$ are fulfilled automatically.

Proposition 1.4.2 *Collocation methods are consistent.*

Proof A criterion for consistency is [4, Theorem 4.18]

$$\sum_{i=1}^{s} b_i = 1.$$

It holds

$$\begin{aligned}\sum_{i=1}^{s} b_i &= \sum_{i=1}^{s} \int_0^1 L_i(\theta)d\theta \\ &= \int_0^1 \sum_{i=1}^{s} L_i(\theta)d\theta \\ &= \int_0^1 1 d\theta \\ &= 1.\end{aligned}$$

□

Now we investigate the order of consistency of collocation methods.

Theorem 1.4.3 *A collocation method is consistent with order p if and only if the corresponding quadrature rule*

$$\int_t^{t+\tau} \phi(\theta)d\theta = \tau \sum_{j=1}^{s} b_j \phi(t + c_j\tau)$$

is consistent with order p.

Proof We only sketch a proof from [4, Theorem 6.40].

$$\begin{aligned}\text{exact ODE: } & x'(t) = f(x) \\ \text{perturbed ODE: } & u'(t) = f(u) + \underbrace{u' - f(u)}_{\delta f(u)}\end{aligned}$$

effect of perturbation: $\exists M : \mathbb{R}^d \to \mathbb{R}^d$

$$\Phi^t x - \Psi^\tau x = x(t+\tau) - u(t+\tau) = \int_t^{t+\tau} M(\theta)\delta f(\theta)d\theta$$

For the quadrature rule holds

$$\int_t^{t+\tau} M(\theta)\delta f(\theta)d\theta - \tau \sum_{j=1}^{s} b_j M(t + c_j\tau)\delta f(t + c_j\tau) = O(\tau^{p+1})$$

exploiting

$$\delta f(t + c_j\tau) = u'(t + c_j\tau) - f(u(t + c_j\tau)) = 0.$$

□

Consequence To obtain maximal order, select $c_1, \ldots, c_s$ such that the corresponding quadrature rule has optimal order: Select Gauß points [9].

Reminder *Gauß points* are the zero points of the polynomial $p \in P_s$ with

$$\int_0^1 p(x)q(x)dx = 0 \quad \forall q \in P_{s-1}.$$

The corresponding quadrature rule has order $p = 2s$.

Remark We know how to construct a Runge-Kutta method with optimal order $p = 2s$.

Example $s = 1$: $c_1 = \frac{1}{2}$

$$L_1(\theta) \equiv 1$$

$$a_{11} = \int_0^{\frac{1}{2}} 1 d\theta = \frac{1}{2}$$
$$b_1 = \int_0^1 1 d\theta = 1$$

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

$$\Psi^\tau x = x + \tau f\left(\frac{1}{2}(x + \Psi^\tau x)\right)$$

midpoint rule; order $p = 2$

$s = 2$: $c_1 = \frac{1}{2} - \frac{\sqrt{3}}{6}$, $\quad c_2 = \frac{1}{2} + \frac{\sqrt{3}}{6}$

$$\begin{array}{c|cc} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{4} + \frac{\sqrt{3}}{6} & & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Remark Gauß methods are implicit.

Proof The order of explicit Runge-Kutta methods of stage s is bounded by s [9]. □

1.5 Dissipative Systems and A-stability of Gauß Methods

We consider

$$x' = f(x).$$

Definition 1.5.1 *A mapping $f : \mathbb{R}^d \to \mathbb{R}^d$ is called* dissipative *with respect to the scalar product $\langle \cdot, \cdot \rangle$ if it satisfies*

$$\langle f(x) - f(y), x - y \rangle \leq 0 \quad \forall x, y, \in \mathbb{R}^d.$$

Example 1. Let $d = 1$ then f is dissipative if and only if f is monotonically decreasing. Hence, dissipativity is a generalization of "monotonically decreasing" to vector fields.

2. Let $f(x) = Ax, \quad A \in \mathbb{R}^{d \times d}$. Then f is dissipative if and only if A is negative semi-definite, i.e.,

$$\langle Ax, x \rangle \leq 0 \quad \forall x \in \mathbb{R}^d.$$

As an example consider

$$A = \begin{pmatrix} -2 & 1 & & 0 \\ 1 & -2 & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & -2 \end{pmatrix}.$$

A is diagonalizable, $\sigma(A) \subset \mathbb{R}, \quad \lambda < 0 \quad \forall \lambda \in \sigma(A)$.

Lemma 1.5.2 *The phase flow Φ^t of $x' = f(x)$ is nonexpansive in the sense that*

$$\exists t^* > 0 : |\Phi^t x - \Phi^t y| \leq |x - y| \quad \forall t \leq t^* \quad \forall x, y \in \mathbb{R}^d$$

if and only if f is dissipative.

Proof Let

$$g(t) := |\Phi^t x - \Phi^t y|^2 = \langle \Phi^t x - \Phi^t y, \Phi^t x - \Phi^t y \rangle.$$

Then

$$\begin{aligned} g'(t) &= \langle \frac{d}{dt}(\Phi^t x - \Phi^t y), \Phi^t x - \Phi^t y \rangle + \langle \Phi^t x - \Phi^t y, \frac{d}{dt}(\Phi^t x - \Phi^t y) \rangle \\ &= 2\langle f(\Phi^t x) - f(\Phi^t y), \Phi^t x - \Phi^t y \rangle. \end{aligned}$$

1. Let f be dissipative. Then

$$\begin{aligned} |\Phi^t x - \Phi^t y|^2 = g(t) &= g(0) + \int_0^t g'(s) ds \\ &= g(0) + \int_0^t 2 \underbrace{\langle f(\Phi^s x) - f(\Phi^s y), \Phi^s x - \Phi^s y \rangle}_{\leq 0} ds \\ &\leq g(0) = |x - y|^2. \end{aligned}$$

2. Let Φ^t be nonexpansive. Then we have for sufficiently small τ^*

$$\begin{aligned} & g(t) \leq g(0) \quad \forall t \leq t^* \\ \Rightarrow \quad & 0 \geq g'(0) = 2\langle f(x) - f(y), x - y \rangle. \end{aligned}$$

□

The concept of inheritance of nonexpansivity leads us to the following definition.

Definition 1.5.3 *A Runge-Kutta method is called* B-stable *if*

$$|\Psi^\tau x - \Psi^\tau y| \leq |x - y| \quad \forall x, y \in \mathbb{R}^d, \ \tau > 0$$

holds for dissipative f.

Proposition 1.5.4 *B-stable Runge-Kutta methods are A-stable.*

Proof Consider

$$x' = \lambda x, \qquad x(0) = 1, \qquad \Re\lambda \leq 0$$

Reformulation in real functions yields

$$\begin{aligned} x &= u + iv, \qquad \lambda = \alpha + i\beta, \qquad \alpha \leq 0 \\ x' &= u' + iv' = \lambda x = (\alpha + i\beta)(u + iv) = \alpha u - \beta v + i(\beta u + \alpha v) \end{aligned}$$

which can be rewritten in matrix form

$$\begin{pmatrix} u \\ v \end{pmatrix}' = \underbrace{\begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}}_{A} \begin{pmatrix} u \\ v \end{pmatrix}. \tag{1.11}$$

Then

$$\begin{aligned} \langle A\begin{pmatrix} u \\ v \end{pmatrix}, \begin{pmatrix} u \\ v \end{pmatrix} \rangle &= (\alpha u - \beta v, \beta u + \alpha v) \begin{pmatrix} u \\ v \end{pmatrix} \\ &= \alpha u^2 - \beta uv + \beta uv + \alpha v^2 \\ &= \alpha(u^2 + v^2) \leq 0. \end{aligned}$$

Therefore, (1.11) is dissipative.
Exploiting B-stability we get

$$|R(\tau\lambda)||x - y| = |\Psi^\tau x - \Psi^\tau y| \leq |x - y|$$

and thus $|R(\tau\lambda)| \leq 1$.
Insert $\tau = 1$:

$$|R(\lambda)| \leq 1 \;\Leftrightarrow\; \lambda \in S$$

Hence, $\lambda \in \mathbb{C}_-$ arbitrary implies $\mathbb{C}_- \subset S$. □

It remains to answer the question which methods inherit nonexpansivity.

Theorem 1.5.5 *Gauß methods are B-stable and therefore A-stable.*

Proof Let f be dissipative and sufficiently smooth, $x, y \in \mathbb{R}^d$. Let $\tau > 0$ be small enough such that the collocation polynomials

$$\begin{aligned} u(0) &= x \qquad u(\tau) = \Psi^\tau x \\ v(0) &= y \qquad v(\tau) = \Psi^\tau y \end{aligned}$$

exist. Set

$$g(\theta) := |u(\theta\tau) - v(\theta\tau)|^2.$$

g is a polynomial of degree at most $2s$. Then

$$g'(\theta) = 2\tau\langle u'(\theta\tau) - v'(\theta\tau), u(\theta\tau) - v(\theta\tau)\rangle. \tag{1.12}$$

The fundamental theorem of calculus yields

$$\begin{aligned}|\Psi^\tau x - \Psi^\tau y|^2 &= g(1)\\ &= g(0) + \int_0^1 g'(\theta)d\theta\\ &= |x-y|^2 + \int_0^1 g'(\theta)d\theta.\end{aligned}$$

It is sufficient to show

$$\int_0^1 g'(\theta)d\theta \le 0.$$

g' is a polynomial of degree $2s-1$. Therefore, Gauß quadrature is exact.

$$\int_0^1 g'(\theta)d\theta = \sum_{j=1}^{s} b_j g'(c_j) \tag{1.13}$$

Now we use the collocation conditions

$$\begin{aligned}u'(c_j\tau) &= f(u(c_j\tau))\\ v'(c_j\tau) &= f(v(c_j\tau)).\end{aligned}$$

Insert into (1.13) using (1.12)

$$\begin{aligned}g'(c_j) &= 2\tau\langle u'(c_j\tau) - v'(c_j\tau), u(c_j\tau) - v(c_j\tau)\rangle\\ &= 2\tau\langle f(u(c_j\tau)) - f(v(c_j\tau)), u(c_j\tau) - v(c_j\tau)\rangle\\ &\le 0 \text{ because } f \text{ is dissipative.}\end{aligned}$$

As $b_j \ge 0$ (stability of Gauß quadrature [9]), this concludes the proof. □

Example 1. The midpoint rule $\Psi^\tau x = x + \tau f\left(\frac{1}{2}(x+\Psi^\tau x)\right)$ is a 1-stage Gauß method and therefore B-stable. Its stability function is

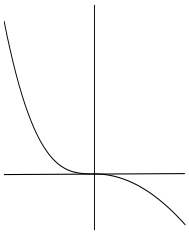
$$R(z) = \frac{1+\frac{z}{2}}{1-\frac{z}{2}}.$$

2. The trapezoidal rule $\Psi^\tau x = x + \frac{\tau}{2}\left(f(x) + f(\Psi^\tau x)\right)$ has the same stability function

$$R(z) = \frac{1+\frac{z}{2}}{1-\frac{z}{2}}.$$

and is A-stable but it is not B-stable:

Consider $$f(x) = \begin{cases} |x|^3 & x \le 0\\ -x^2 & x \ge 0\end{cases}$$



f is monotonically decreasing and therefore dissipative.
$x' = f(x)$ has the fixed point $x^* = 0$.
If the trapezoidal rule were B-stable, then

$$|\Psi^\tau x| = |\Psi^\tau x - \Psi^\tau x^*| \leq |x - x^*| = |x|.$$

But $x = -2$ and $\tau = \frac{36}{7}$ gives

$$\Psi^\tau x = 2.5 > 2 = |x|.$$

Theorem 1.5.6 *Let f be dissipative and sufficiently smooth. Then the nonlinear system*

$$k_i = f\left(x + \tau \sum_{j=1}^{s} a_{ij} k_j\right), \qquad i = 1, \ldots, s$$

associated with a Gauß method has a unique solution for any $\tau \geq 0$ and all $x \in \mathbb{R}^d$.

Proof [4, Theorem 6.54]
Illustration (very special case): $d = 1$, implicit Euler

$$\underbrace{k_1 \quad \underbrace{-f(x + \tau k_1)}_{\text{mon. increasing}}}_{\text{strictly mon. increasing}} \qquad = 0$$

Therefore uniquely solvable. □

1.6 Preserving Asymptotic Stability: Nonlinear Systems

We consider the nonlinear autonomous system

$$x' = f(x), \qquad f : \mathbb{R}^d \to \mathbb{R}^d$$

For (technical) simplicity we assume $d = 1$.
Our aim is to find criteria for Ψ^τ which guarantee asymptotic stability of fixed points of the nonlinear recursion

$$x_{k+1} = \Psi^\tau x_k.$$

Theorem 1.6.1 *Let $\Psi : \mathbb{R} \to \mathbb{R}$ be continuously differentiable with fixed point x^*, i.e., $\Psi(x^*) = x^*$, and $|\Psi'(x^*)| < 1$. Consider*

$$x_{k+1} = \Psi(x_k), \qquad x_0 \in \mathbb{R} \text{ given.}$$

Then x^* is asymptotically stable in the sense that

$$\exists \delta > 0 : \lim_{k \to \infty} x_k = x^* \text{ for } |x_0 - x^*| < \delta.$$

Proof Consider the Taylor expansion

$$\Psi(x) = \Psi(x^*) + \Psi'(x^*)(x - x^*) + g(x - x^*) \tag{1.14}$$

with

$$\lim_{x \to 0} \frac{g(x)}{x} = 0.$$

Inserting $\Psi(x^*) = x^*$ and (1.14) into the recursion yields

$$x_{k+1} - x^* = \Psi'(x^*)(x_k - x^*) + g(x_k - x^*).$$

Choose $\beta > 0$ such that

$$|\Psi'(x^*)| + \beta < 1$$

and $\delta > 0$ such that

$$|g(x - x^*)| \leq \beta |x - x^*| \quad \forall x : \ |x - x^*| < \delta.$$

Assuming $|x_k - x^*| < \delta$ we get

$$\begin{aligned} |x_{k+1} - x^*| &\leq |\Psi'(x^*)||x_k - x^*| + |g(x_k - x^*)| \\ &\leq (|\Psi'(x^*)| + \beta)|x_k - x^*|. \end{aligned}$$

Hence $|x_0 - x^*| < \delta$ inductively leads to

$$|x_{k+1} - x^*| \leq (|\Psi'(x^*)| + \beta)^{k+1} |x_0 - x^*| \to 0 \text{ for } k \to \infty.$$

□

Remark The result and the proof directly extend to $d > 1$ if $|\Psi'(x^*)| < 1$ is replaced by $\rho(D\Psi(x^*)) < 1$. The main ingredient for this is that $\rho(A) < 1$ implies that there is a vector norm $\|\cdot\|$ and an associated matrix norm $\|\cdot\|$ such that $\|A\| < 1$ [13].

We will now consider the inheritance of asymptotic stability of fixed points.
Let $x' = f(x)$ be a scalar equation with $f : \mathbb{R} \to \mathbb{R}$ sufficiently smooth (continuosly differentiable). Let further $x^* \in \mathbb{R}$ be a fixed point of f.
We know that $f'(x^*) < 0$ implies that x^* is asymptotically stable (see Theorem 1.1.7).
Application of a Runge-Kutta method yields

$$x_{k+1} = \Psi^\tau x_k. \tag{1.15}$$

Question Is x^* an asymptotically stable fixed point of (1.15)?
In the linear case $f(x) = \lambda x$ the answer to this question is yes if Ψ^τ is A-stable.

Remark Let R be the stability function of Ψ^τ.

$$\begin{aligned} \Psi^\tau \text{ A-stable } &\Leftrightarrow \ \mathbb{C}_- \subset S = \{z \in \mathbb{C} || R(z)| \leq 1\} \\ &\Rightarrow \ |R(z)| < 1 \quad \forall z : \Re z < 0 \end{aligned}$$

see [4, Theorem 6.13].

Question Does this result extend to the nonlinear case?

Lemma 1.6.2 (Invariance under linearization) *Let x^* be a fixed point of $x' = f(x)$, i.e., $f(x^*) = 0$, and let Ψ^τ be a Runge-Kutta method. Then*

(i) $\Psi^\tau(x^*) = x^*$

(ii) $\Psi^\tau(x) = x^* + R(\tau f'(x^*))(x - x^*) + g(x - x^*)$ *with* $g(x) = o(|x|)$.

Proof 1. Consider the Runge-Kutta method

$$\begin{aligned}\Psi^\tau(x) &= x + \tau \sum_{i=1}^{s} b_i k_i \\ k_i &= f(x + \tau \sum_{j=1}^{s} a_{ij} k_j), \quad i = 1, \ldots, s.\end{aligned}$$

Insert $k_i = 0$ for $i = 1, \ldots, s$ to see that $k_i(x^*)$ is a solution for $x = x^*$.

2. Consider the Taylor expansion

$$\begin{aligned} f(x) &= f(x^*) + f'(x^*)(x - x^*) + u(x, x - x^*) \\ &= f'(x^*)(x - x^*) + u(x, x - x^*) \end{aligned} \tag{1.16}$$

with $|u(x, x - x^*)| = \frac{1}{2} f''(\rho(x))|x - x^*|^2 \leq c|x - x^*|^2$. Insert (1.16) into the Runge-Kutta method to obtain

$$\begin{aligned}\Psi^\tau(x) &= x + \tau \sum_{i=1}^{s} b_i k_i \\ k_i &= f(x + \tau \sum_{j=1}^{s} a_{ij} k_j) \\ &= f'(x^*)(x - x^* + \tau \sum_{j=1}^{s} a_{ij} k_j) + u(x + \tau \sum_{j=1}^{s} a_{ij} k_j, \underbrace{x - x^* + \tau \sum_{j=1}^{s} a_{ij} k_j}_{=: y_i(x,k)}).\end{aligned} \tag{1.17}$$

Let

$$\begin{aligned}\tilde{\Psi}^\tau(x) &= x + \tau \sum_{i=1}^{s} b_i \tilde{k}_i \\ \tilde{k}_i &= f'(x^*)(x - x^* + \tau \sum_{j=1}^{s} a_{ij} \tilde{k}_j).\end{aligned} \tag{1.18}$$

Theorem 1.3.1 implies

$$\tilde{\Psi}^\tau(x) = x^* + R(\tau f'(x^*))(x - x^*), \quad \tau \leq \tau^*$$

where $R(z) = \frac{P(z)}{Q(z)}$ is the stability function of Ψ^τ.
It is still to be shown that

$$\Psi^\tau x - \tilde{\Psi}^\tau x = o(|x - x^*|).$$

Matrix formulation of (1.17) and (1.18), respectively, with

$$k = \begin{pmatrix} k_1 \\ \vdots \\ k_s \end{pmatrix}, \quad \tilde{k} = \begin{pmatrix} \tilde{k}_1 \\ \vdots \\ \tilde{k}_s \end{pmatrix}, \quad e = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

yields

$$(I - \tau f'(x^*)\mathbb{A})k - f'(x^*)(x - x^*)e = U(x,k) \tag{1.19}$$

$$(I - \tau f'(x^*)\mathbb{A})\tilde{k} - f'(x^*)(x - x^*)e = 0 \tag{1.20}$$

with

$$U(x,k) = \begin{pmatrix} u(y_1(x,k) + x^*, y_1(x,k)) \\ \vdots \\ u(y_s(x,k) + x^*, y_s(x,k)) \end{pmatrix}.$$

(1.19)-(1.20) and multiplication with $(I - \tau f'(x^*)\mathbb{A})^{-1}$ $(\tau < \tau^*!)$ yields

$$k - \tilde{k} = (I - \tau f'(x^*)\mathbb{A})^{-1} U(x,k).$$

It is sufficient to show $\|U(x,k)\| = o(|x - x^*|)$

$$|u(y_i(x,k) + x^*, y_i(x,k))| \le c|x - x^* + \tau \sum_{j=1}^{s} a_{ij} k_j|^2. \tag{1.21}$$

Now $k_i(x^*) = 0$ and the implicit function theorem provides

$$\exists \delta > 0 : \ |k_j(x)| \le c|x - x^*|, \quad j = 1, \ldots, s \text{ if } |x - x^*| < \delta.$$

Insert into (1.21) to conclude the proof.

□

Theorem 1.6.3 *Let $x^* \in \mathbb{R}$ be a fixed point of $x' = f(x)$ with $f \in C^2(\mathbb{R})$ and $f'(x^*) < 0$, i.e., x^* is asymptotically stable. Let Ψ^τ be an A-stable Runge-Kutta method. Then x^* is an asymptotically stable fixed point of Ψ^τ.*

Proof x^* is a fixed point of Ψ^τ by Lemma 1.6.2 (i). To use Theorem 1.6.1 we have to show

$$\left| \frac{d}{dx} \Psi^\tau x \right|_{x=x^*} < 1.$$

Lemma 1.6.2 (ii) yields

$$\begin{aligned} \frac{\Psi^\tau x - \Psi^\tau x^*}{x - x^*} &= \frac{x^* + R(\tau f'(x^*))(x - x^*) + g(x - x^*) - x^*}{x - x^*} \\ &= R(\tau f'(x^*)) + \frac{g(x - x^*)}{x - x^*} \overset{x \to x^*}{\to} R(\tau f'(x^*)). \end{aligned}$$

Hence,

$$\frac{d}{dx}\Psi^\tau x|_{x=x^*} = R(\tau f'(x^*)).$$

Since Ψ^τ is A-stable,

$$\tau f'(x^*) < 0 \;\Rightarrow\; |R(\tau f'(x^*))| < 1.$$

□

Remark Theorem 1.6.3 can be directly extended to systems (see [4, Theorem 6.23]).

A Roadmap of Notions

Stiffness: (asymptotic) stability of fixed points (continuous problem)

- sufficient criteria for (asymptotic) stability
  linear case (eigenvalues) → nonlinear case by linearization (only asymp. stability)

(asymptotic) stability of recursions (discrete problem)

- sufficient criteria for (asymptotic) stability
  linear case → nonlinear case by linearization

inheritance of (asymptotic) stability

- linear case: $x' = \lambda x$ ($x' = Ax$)
  - stability function, stability domain
    criterion for inheritance: A-stability
  - construction of A-stable methods: collocation methods, Gauß methods
- nonlinear case: $x' = f(x)$, only asymptotic stability
  criterion for inheritance: A-stability

Question Could we get something like A-stability "cheaper"?

1.7 Algorithmic Aspects of Implicit RK's (Gauß Methods)

We consider

$$x' = f(x), \quad f : \mathbb{R}^d \to \mathbb{R}^d$$

and the implicit Runge-Kutta method

$$\Psi^\tau x = x + \tau \sum_{i=1}^{s} b_i k_i$$

$$k_i = f(x + \tau \sum_{j=1}^{s} a_{ij} k_j), \quad i = 1, \ldots, s$$

with $d \cdot s$ unknowns and $d \cdot s$ equations.
Reformulation in *symmetric form*:

$$g_i := x + \tau \sum_{j=1}^{s} a_{ij} k_j \Rightarrow k_j = f(g_j)$$

$$\Psi^\tau x = x + \tau \sum_{i=1}^{s} b_i f(g_i)$$

$$g_i = x + \tau \sum_{j=1}^{s} a_{ij} f(g_j), \quad i = 1, \ldots, s$$

The advantage of this form is that differentiating g_i does not produce inner derivatives. The disadvantage is that there are s additional f-evaluations.
To avoid cancellations, we solve for corrections $z_i = g_i - x, \quad i = 1, \ldots, s$

$$\Psi^\tau x = x + \tau \sum_{i=1}^{s} b_i f(x + z_i) \tag{1.22}$$

$$z_i = \tau \sum_{j=1}^{s} a_{ij} f(x + z_j), \quad i = 1, \ldots, s. \tag{1.23}$$

To save f-evaluations in (1.22), we can write

$$z = \tau \mathbb{A} \begin{pmatrix} f(x + z_1) \\ \vdots \\ f(x + z_s) \end{pmatrix}, \quad z = \begin{pmatrix} z_1 \\ \vdots \\ z_s \end{pmatrix}.$$

Assume that $\mathbb{A} = (a_{ij})_{i,j=1}^s$ is invertible and compute $\mathbb{A}^{-1}$ (s is moderate!). Then (1.22) can be rewritten as

$$\Psi^\tau x = x + \sum_{i=1}^{s} b_i (\mathbb{A}^{-1} z)_i.$$

Vector form of (1.23):

$$z = \tau F(z), \quad F = (F_i)_{i=1}^s, \quad F_i(z) = \sum_{j=1}^{s} a_{ij} f(x + z_j)$$

1.7.1 Fixed point iteration

$$z^{\nu+1} = \tau F(z^\nu), \quad z^0 = 0 \tag{1.24}$$

Why is $z^0 = 0$ a good initial iterate?

Proposition 1.7.1 *Let f be continuously differentiable. Then there is a $\tau^* > 0$ such that (1.24) converges to the solution z for $\tau \leq \tau^*$.*

Proof 1. Find $K \subset \mathbb{R}^{sd}$ closed and bounded such that $\tau F(K) \subset K$:
Choose arbitrary $c > 0$. Then

$$\exists c_1 > 0 : \ \|F(z)\| \leq c_1 \quad \forall z : \|z\| < c.$$

Let $\tau_1^* = \frac{c}{c_1}$. Then

$$z \in K := \{z \in \mathbb{R}^{sd} | \|z\| \leq c\} \ \Rightarrow \ \tau F(z) \in K \quad \forall \tau \leq \tau_1^*.$$

Hence, $\tau F(K) \subset K$.

2. Contractivity: $\|\tau F(x) - \tau F(y)\| \leq q\|x - y\|$ with $q < 1 \quad \forall x, y \in K$
Jacobian:

$$DF(z) = (B_{ij})_{i,j=1}^{s} \in \mathbb{R}^{sd \times sd}, \quad B_{ij} = a_{ij} Df(x + z_j) \in \mathbb{R}^{d \times d}$$

Mean value theorem:

$$\|\tau F(x) - \tau F(y)\| \leq \tau \|DF(\xi)\| \|x - y\| \leq q\|x - y\|$$

with $q = \tau \max_{z \in K} \|DF(z)\| < 1$ for $\tau < \tau^* < \min\{\tau_1^*, \frac{1}{\max_{z \in K} \|DF(z)\|}\}$

3. $x^0 = 0 \in K$

The assertion follows from Banach's fixed point theorem. □

Example The bacteria equation

$$x' = f(x) = \alpha x - \beta x^2$$

is stiff for $\alpha, \beta >> 1$ and has the fixed point $x^* = \frac{\alpha}{\beta}$. The function F corresponding to the implicit Euler scheme satisfies

$$\begin{aligned} F(z) &= f(x + z) = \alpha(x + z) - \beta(x + z)^2 \\ F'(z) &= \alpha - 2\beta(x + z) \\ \|F'(z)\| &\approx \alpha >> 1 \text{ for } x \approx x^*, \ z \approx 0 \\ \Rightarrow \ \tau^* &\approx \frac{1}{\alpha} << 1 \end{aligned}$$

Stiffness causes stepsize reduction!

Consequence Do not use simple fixed point iteration for implicit Runge-Kutta methods. Leninger/Willoughby: Timestep restriction of this form arise for any iteration involving only f-evaluations.

1.7.2 Newton iteration and simplifications

We consider the ordinary Newton method for

$$z - \tau F(z) = 0$$

$$
\begin{aligned}
&z^0 = 0\\
&(I - \tau DF(z^\nu))\Delta z^\nu = -(z^\nu - \tau F(z^\nu))\\
&z^{\nu+1} = z^\nu + \Delta z^\nu
\end{aligned}
$$

where the Jacobian $I - \tau DF(z)$ is given by

$$DF(z) = (a_{ij} Df(x + z_j))_{i,j=1}^s.$$

Then the computational effort for each step is

- $(s \cdot d)^2$ scalar function evaluations to obtain $DF(z)^\nu$
- solution of a linear system

Convergence properties local quadratic convergence, globalization by damping

Simplified Newton

Basic idea Trade local quadratic convergence in for reduced computational effort. Replace $Df(x + z_j^\nu)$ by $J = Df(x)$ to obtain

$$
\begin{aligned}
B &= I - \tau (a_{ij} Df(x))_{i,j=1}^s\\
&= I - \begin{pmatrix} \tau a_{11} J & \tau a_{12} J & \dots & \tau a_{1s} J \\ \tau a_{21} J & & & \vdots \\ \vdots & & & \tau a_{s-1s} J \\ \tau a_{s1} J & \dots & \tau a_{ss-1} J & \tau a_{ss} J \end{pmatrix}\\
&= I - \tau \mathbb{A} \otimes J
\end{aligned}
$$

with the tensorproduct $A \otimes B \in \mathbb{R}^{nk \times ml}$ of $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{k \times l}$ defined by

$$A \otimes B = \begin{pmatrix} a_{11} B & \dots & a_{1m} B \\ \vdots & & \vdots \\ a_{n1} B & \dots & a_{nm} B \end{pmatrix}.$$

Consequence Only *one* LU-decomposition is sufficient.

Remark If $\mathbb{A}$ is invertible, then the tensorproduct structure can be used to reduce the computational effort for the LU-decomposition [7, IV.8].

Convergence properties Local linear convergence for sufficiently small τ [12, Section 12.6] (see also [3]).

Heuristic stopping criterion for contractions Stop the iteration as soon as

$$\|z - z^\nu\| = O(\tau^p) = c\tau^p.$$

Very rough but simple estimate:
Under the assumption $\|z - z^{\nu+1}\| \leq q\|z - z^\nu\|$ with $q < 1$ we get the a posteriori error estimate

$$\|z - z^{\nu+1}\| \leq \frac{q}{1-q}\|z^{\nu+1} - z^\nu\|.$$

Proof

$$\|z - z^{\nu+1}\| \leq q\|z - z^\nu\| \leq q(\|z - z^{\nu+1}\| + \|z^{\nu+1} - z^\nu\|)$$

□

Approximation of unknown convergence rate q

$$q \approx \theta = \frac{\|z^{\nu+1} - z^\nu\|}{\|z^\nu - z^{\nu-1}\|}$$

motivated by

$$\|z^{\nu+1} - z^\nu\| \leq q\|z^\nu - z^{\nu-1}\|.$$

Unfortunately, $\theta \leq q$ is equivalent to $\frac{\theta}{1-\theta} \leq \frac{q}{1-q}$ rather than $\frac{q}{1-q} \leq \frac{\theta}{1-\theta}$. Therefore, the theoretical justification of the upper bound is lost.

Bold question What happens if only *one* Newton step is performed in each time step?

1.8 Linearly Implicit One-Step-Methods

Sections 1.1 and 1.6 suggest that stability can be obtained by linearization.

Basic idea

- rewrite $x' = f(x)$ as
$$x' = Jx + (f(x) - Jx), \quad J = Df(x)$$
- use an implicit discretization only for the leading linear term

General form of a s-stage linearly implicit Runge-Kutta method (sometimes called Rosenbrock-Wanner method):

$$\Psi^\tau x = x + \tau \sum_{i=1}^{s} b_i k_i \tag{1.25}$$

$$k_i = J(x + \tau \sum_{j=1}^{i} \beta_{ij} k_j) + (f(x + \tau \sum_{j=1}^{i-1} \alpha_{ij} k_j) - J(x + \tau \sum_{j=1}^{i-1} \alpha_{ij} k_j))$$

with coefficients $\mathbb{B} = (\beta_{ij})_{i,j=1}^{s}, \quad \mathbb{A} = (\alpha_{ij})_{i,j=1}^{s}, \quad b = (b_i)_{i=1}^{s}$.
The computational effort comprises the solution of s linear systems of the form

$$(I - \tau\beta_{ii}J)k_i = \tau \sum_{j=1}^{i-1}(\beta_{ij} - \alpha_{ij})Jk_j + f(x + \tau \sum_{j=1}^{i-1} \alpha_{ij}k_j).$$

Therefore, it contains s LU-decompositions of $(I - \tau\beta_{ii}J)$ in $\mathbb{R}^{d\times d}$ (not in $\mathbb{R}^{sd\times sd}$ like the LU-composition of $(I - \tau\mathbb{A} \otimes Df)$). If we additionally assume $\beta_{11} = \beta_{22} = \ldots = \beta$, then only *one* LU-decomposition of $I - \tau\beta J$ is needed.

Example We consider the linearly implicit Euler method: $b_1 = 1, \quad \beta_{11} = 1, \alpha_{11} = 0$

$$\begin{aligned} \Psi^\tau x &= x + \tau k_1 \\ k_1 &= J(x + \tau k_1) + f(x) - J(x) \\ \Leftrightarrow (I - \tau J)k_1 &= f(x) \end{aligned}$$

and the implicit Euler with a single Newton step starting from $z^0 = 0$. The implicit Euler scheme can be written as

$$\begin{aligned} \Psi^\tau x &= x + z_1 \\ z_1 &= \tau f(x + z_1). \end{aligned}$$

A Newton step on $z - \tau f(x + z)$ yields

$$\begin{aligned} \Psi^\tau x &= x + z^1 \\ (I - \tau J)(z^1 - z^0) &= -(z^0 - \tau f(x + z^0)) \\ \iff (I - \tau J)z^1 &= \tau f(x) \quad \text{with } z^0 = 0 \end{aligned}$$

The two methods are identical.

Proposition 1.8.1 *Assume that* $\max_{\lambda\in\sigma(J)} \Re\lambda = \nu(J) \leq 0$ *and* $\beta \geq 0$. *Then* $(I - \tau\beta J)$ *is invertible for all* $\tau \geq 0$.

Proof Let $\lambda \in \sigma(J)$. We have to show

$$1 - \tau\beta\lambda \neq 0$$

but obviously

$$\Re(1 - \tau\beta\lambda) \geq 1.$$

□

Consequence For stiff systems no timestep restriction is required for solvability.

Proposition 1.8.2 *A Rosenbrock-Wanner method* $\begin{array}{c|c} \mathbb{B} & \mathbb{A} \\ \hline & b^T \end{array}$ *is A-stable if and only if the Runge-Kutta method* $\begin{array}{c|c} & \mathbb{B} \\ \hline & b^T \end{array}$ *is A-stable.*

Proof Application of the Rosenbrock-Wanner method to $x' = \lambda x$ with $J = \lambda$ leads to

$$\Psi^\tau x = x + \tau \sum_{i=1}^{s} b_i k_i$$
$$k_i = \lambda(x + \tau \sum_{j=1}^{i} \beta_{ij} k_j)$$

and thus $\Psi^\tau = R(\tau\lambda)$ with the stability function R of $\begin{array}{c|c} \mathbb{B}. \\ \hline b^T \end{array}$. □

Construction of higher order methods

Proposition 1.8.3 *The method (1.25) is consistent with order $p = 1$ if*

$$\sum_{j=1}^{s} b_j = 1.$$

It is consistent with order $p = 2$ if additionally

$$\sum_{j,k=1}^{s} b_j(\alpha_{jk} + \beta_{jk}) = \frac{1}{2}.$$

It is consistent with order $p = 3$ if additionally

$$\sum_{j,k,l=1}^{s} b_j \beta_{jk} \beta_{jl} = \frac{1}{3}$$
$$\sum_{j,k,l=1}^{s} b_j(\alpha_{jk} + \beta_{jk})(\alpha_{kl} + \beta_{kl}) = \frac{1}{6}.$$

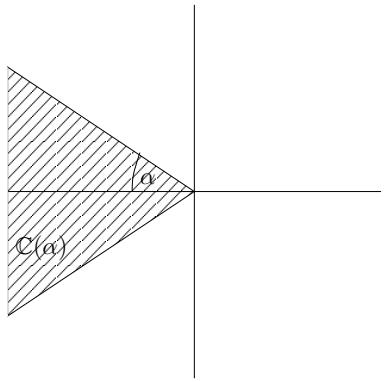
Proof [7, IV.7]

Stability

It is difficult to have high order, $\beta_{ii} = \beta > 0$ and A-stability. The following definition provides a compromise.

Definition 1.8.4 *A Runge-Kutta method is called* A(α)-stable *if*

$$\mathbb{C}(\alpha) := \{z = re^{i\phi} | r \geq 0, |\pi - \phi| \leq \alpha\} \subset S.$$



The following basic schemes are A(α)-stable:

GRK4	$\alpha = 90$
GRK4T	$\alpha = 89,3$

see [7, IV].

Modifications

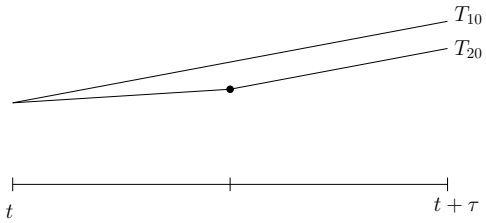
- inexact Jacobian J (W-methods)
- sporadically recomputed J

1.9 Extrapolation Methods

Basic idea Consider a low order scheme Ψ_*^τ and a partition

$$t < t + \tau_j < t + 2\tau_j < \ldots < t + (n_j - 1)\tau_j < t + \tau, \quad \tau_j = \frac{\tau}{n_j}$$

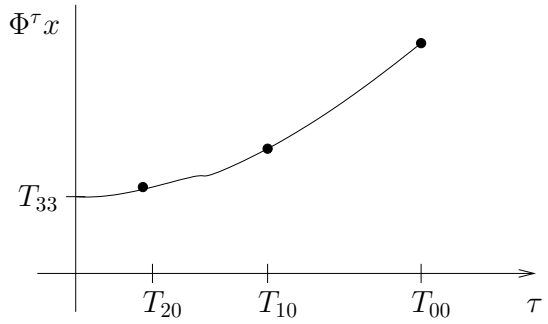
with a step number sequence $n_0 < n_1 < \ldots$.



Compute

$$T_{j0} = \left(\Psi_*^{\tau_j}\right)^{n_j} x.$$

Extrapolate to $\tau = 0$.



Use method of Aitken-Neville (Numerics I) for the computation

$$\begin{array}{ccccccc}
T_{00} & & & & & & \\
 & \searrow & & & & & \\
T_{10} & \to & T_{11} & & & & \\
 & \searrow & & \searrow & & & \\
T_{20} & \to & T_{21} & \to & T_{22} & & \\
\vdots & & \vdots & & \vdots & \ddots & \\
T_{m0} & \to & T_{m1} & \to & T_{m2} & \dots & T_{mm}
\end{array}$$

with

$$T_{j,k} = T_{j,k-1} + \frac{T_{j,k-1} - T_{j-1,k-1}}{\left(\frac{n_{j-k}}{n_k}\right)^2 - 1}.$$

Define

$$\Psi^\tau x := T_{mm}.$$

Proposition 1.9.1 *Assume there is an asymptotic expansion*

$$\Psi_*^\tau x = \Phi^\tau x + \sum_{k=1}^{n+1} c_k \tau^k + r_{n+2}(\tau)\tau^{n+2}, \quad \tau \le \tau^*$$

with $\|r_{n+2}\|_\infty \le C$. *Then*

$$|\Phi^\tau x - \Psi^\tau x| \le \frac{c_{n+1}}{n_0 \dots n_m}\tau^{m+1} + O(\tau^{m+2})$$

i.e., Ψ^τ *is consistent with order* m.

Proof See [9] or any other textbook.

Example Consider the linearly implicit Euler method

$$\begin{aligned}
\Psi_*^\tau x &= x + \tau k_1 \\
(I - \tau J)k_1 &= f(x), \quad J = Df(x).
\end{aligned}$$

An asymptotic expansion exists for sufficiently smooth f [5].
The extrapolation of the linearly implicit Euler method is not A-stable, but A(α)-stable

$$\begin{aligned} T_{11} &: \text{A-stable} \\ T_{22} &: \text{A(89,85)-stable} \\ &\vdots \\ T_{77} &: \text{A(89,81)-stable} \end{aligned}$$

It is tempting to use the trapezoidal rule

$$\Psi_*^\tau x = x + \frac{\tau}{2}(f(x) + f(\Psi_*^\tau x))$$

because it has an asymptotic expansion in τ^2.

$$\hookrightarrow \; T_{mm} = \Phi^\tau x + O(\tau^{2(m+1)})$$

Unfortunately, extrapolation *destroys* stability properties.
Remedy: Linearly implicit midpoint rule

$$(I - \tau J)x_{k+1} - (I + \tau J)x_{k-1} = 2\tau(f(x_k) - Jx_k)$$

See [1], [4, 6.4.2].

1.10 Gradient Flows and Parabolic PDEs

We first consider discrete gradient flows as an example for a class of stiff ODEs.
Continuous gradient flows are leading to parabolic PDEs. By discretization in space (method of lines) we will discover a large class of arbitrary large and arbitrary stiff systems of ODEs.

Definition 1.10.1 *A functional $E : \mathbb{R}^d \to \mathbb{R}$ is called* convex *if*

$$E(\omega x + (1-\omega)y) \le \omega E(x) + (1-\omega)E(y) \quad \forall \omega \in [0,1]. \tag{1.26}$$

E is called strictly convex *if the equality in (1.26) holds only at $\omega = 0$ and $\omega = 1$.*

Example 1. E defined by

$$E(x) = \sum_{i=1}^{d} x_i b_i \text{ for } x \in \mathbb{R}^d, \quad b \in \mathbb{R}^d \text{ fixed}$$

is convex but not strictly convex.

2. E defined by

$$E(x) = \frac{1}{2}\sum_{i=1}^{d}(x_i^2 - b_i x_i) \text{ for } x \in \mathbb{R}^d, \quad b \in \mathbb{R}^d \text{ fixed}$$

is strictly convex.

Example We consider the following four examples of scalar, convex functions:

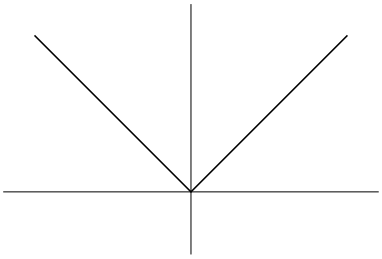


Figure 1.1: convex

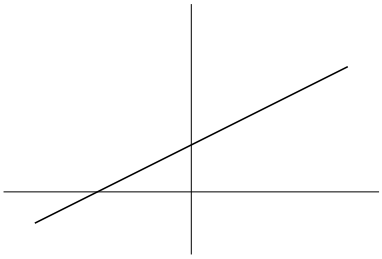


Figure 1.3: convex

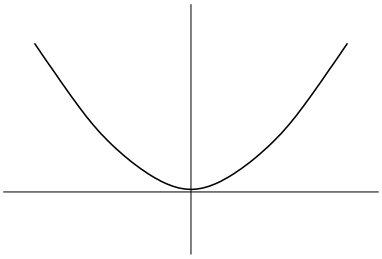


Figure 1.2: strictly convex

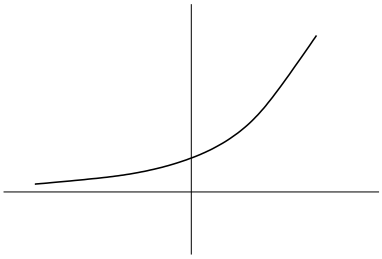


Figure 1.4: strictly convex

Definition 1.10.2 *A functional $E : \mathbb{R}^d \to \mathbb{R}$ is called* coercive *if*

$$\lim_{\|x\|\to\infty} E(x) = \infty.$$

The functions in Figure 1.1 and Figure 1.2 are coercive. The functions in Figure 1.3 and Figure 1.4 are not coercive.

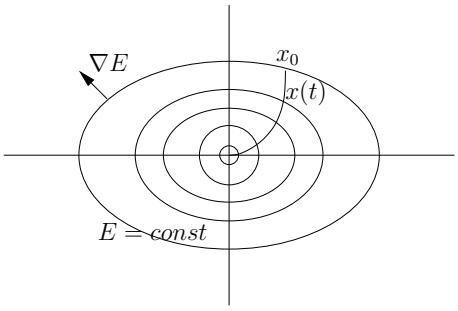
Lemma 1.10.3 *A coercive, continuous, convex functional $E : \mathbb{R}^d \to \mathbb{R}$ has at least one minimum. If E is strictly convex, then the minimum is unique.*

Proof Exercise

Definition 1.10.4 *Let $E : \mathbb{R}^d \to \mathbb{R}$ be convex and continuously differentiable. The initial value problem*

$$x'(t) = -\nabla E(x), \quad x(0) = x_0, \quad t > 0$$

is called the gradient flow *associated with E.*



Lemma 1.10.5 *A gradient flow has the following properties*

1. *For any $t > 0$, $E(x(t)) \leq E(x_0)$ holds.*
2. *$f(x) = -\nabla E(x)$ is dissipative, i.e.,*
$$\langle -\nabla E(x) + \nabla E(y), x - y \rangle \leq 0$$
3. *Every fixed point of the gradient flow is a minimum of E and vice versa.*
4. *Let E be strictly convex. Then all fixed points of the gradient flow are asymptotically stable.*
5. *Let E be strictly convex and coercive. Then $x^* = \lim_{t\to\infty} x(t)$ is a fixed point of the gradient flow for all x_0.*

Proof Exercise

Example Let $f : \mathbb{R}^d \to \mathbb{R}^d$ such that there exists a strictly convex, coercive, and continuously differentiable $E : \mathbb{R}^d \to \mathbb{R}$ with $f = -\nabla E$. We are looking for a zero point of f:
Integrate the gradient flow of E using a "good" initial value until you reach a fixed point. Then Lemma (1.10.5) states that x^* is a minimum of E which must be a zero of f.
This procedure might be very slow.

Continuous gradient flows: The heat equation

Let Ω be a compact subset of $\mathbb{R}^2$ and $C^1(\Omega)$ the set of all continuously differentiable functionals $v : \Omega \to \mathbb{R}$. Then the "energy" is given by

$$E(u) = \int_\Omega \left(\frac{1}{2} |\nabla u|^2 - bu \right) dx, \quad u \in C^1(\Omega),\ b \in C(\Omega)$$

where u denotes the temperature distribution and b the density of the heat source.
Let $\langle \cdot, \cdot \rangle$ denote the L_2 scalar product $\langle v, w \rangle = \int_\Omega v \cdot w \, dx$. The derivative of E at $u \in C^1(\Omega)$ is a 1-form $\nabla E(u)(\cdot) : C^1(\Omega) \to \mathbb{R}$

$$\nabla E(u)(v) = \langle \nabla u, \nabla v \rangle - \langle b, v \rangle.$$

Lemma 1.10.6 *Let $u, \bar{u} \in C^1(\Omega)$ such that*

$$\langle u, v \rangle = \langle \bar{u}, v \rangle \quad \forall v \in C^1(\Omega).$$

Then $u = \bar{u}$.

Proof It is sufficient to show

$$\langle u, v \rangle = 0 \quad \forall v \ \Rightarrow \ u = 0.$$

Assume

$$\exists x \in \Omega : u(x) > 0$$

then there is a neighborhood V of x with

$$u(\bar{x}) > 0 \quad \forall \bar{x} \in V.$$

Then there is a $v_0 \in C^1(\Omega)$ with $v_0 \neq 0$, $v_0 \geq 0$ and $v_0(\bar{x}) \quad \forall \bar{x} \notin V$. Inserting v_0 we obtain

$$\langle u, v_0 \rangle = \int_V u(x) v_0(x)\ dx > 0.$$

□

Gradient flow of E:

$$\begin{aligned} \frac{d}{dt}\langle u, v\rangle = \langle u_t, v\rangle &= -\nabla E(u)(v) \\ &= -\langle \nabla u, \nabla v\rangle + \langle b, v\rangle \quad \forall v \in C^1(\Omega) \end{aligned}$$

This is the weak form of the heat equation. Green's identity and Lemma 1.10.6 provide

$$\langle u_t, v\rangle = \langle \Delta u, v\rangle + \langle b, v\rangle \ \Rightarrow\ u_t = \Delta u + b \ \wedge\ \frac{\partial}{\partial n} v = 0 \text{ on } \partial\Omega$$

which is the strong form of the heat equation.
If $b \equiv 0$, then the heat energy $\int u\ dx$ is preserved:

$$\langle u_t, v\rangle = -\langle \nabla u, \nabla v\rangle \quad \forall v \in C^1(\Omega)$$

Insertion of $v = 1$ yields

$$0 = \langle u_t, 1\rangle = \int_\Omega \frac{\partial}{\partial t} u\ dx = \frac{\partial}{\partial t} \int_\Omega u\ dx.$$

An initial-boundary-value problem for the heat equation

We consider the heat equation

$$u_t = u_{xx} + b$$

with initial conditions

$$u(x, 0) = u_0(x) \quad \forall x \in \Omega = [a, b]$$

and boundary conditions

$$u(a, t) = u(b, t) = 0 \quad \forall t \in [0, T].$$

For the discretization of the problem we consider the following two options.

Method of lines (first space, then time) The basic idea is to discretize in space in order to obtain an ODE.
We choose an equidistant mesh

$$x_i = a + ih, \quad i = 0, \ldots, n, \quad h = \frac{b-a}{n}.$$

Then, assuming $u(\cdot,t) \in C^4(a,b) \quad \forall t \in (0,T]$, we have

$$u_{xx}(x_i) = \frac{1}{h^2}(u(x_{i-1}) - 2u(x_i) + u(x_{i+1})) + O(h^2), \quad i = 1,\dots,n-1.$$

Compute approximations $U_i(t)$ of $u(x_i,t)$, $t \in [0,T]$ to obtain the ODE:

$$\begin{aligned} U_i'(t) &= \frac{1}{h^2}(U_{i-1}(t) - 2U_i(t) + U_{i+1}(t)) + B_i \\ U_0(t) &= U_n(t) = 0 \end{aligned}$$

with $B_i(t) = b(x_i,t)$. This can be rewritten in matrix form

$$\begin{aligned} U' &= AU + B \\ U &= (U_i)_{i=1}^{n-1}, \quad B = (B_i)_{i=1}^{n-1} \\ A &= \frac{1}{h^2}\begin{pmatrix} -2 & 1 & & 0 \\ 1 & -2 & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & -2 \end{pmatrix} \in \mathbb{R}^{n-1\times n-1}. \end{aligned}$$

A has the following eigenvalues:

$$\lambda_i = -\frac{4n^2}{(b-a)^2}\sin^2\left(\frac{i}{2n}\pi\right).$$

Thus,

$$-4h^{-2} = -\frac{4n^2}{(b-a)^2} \approx \lambda_{n-1} < \lambda_{n-2} < \ldots < \lambda_1 \approx -\frac{4n^2}{(b-a)^2}\frac{\pi^2}{4n^2} = -\left(\frac{\pi}{(b-a)}\right)^2 < 0.$$

The ODE is arbitrarily stiff because

$$\frac{|\lambda_{n-1}|}{|\lambda_1|} \to \infty \text{ for } n \to \infty.$$

Hence, severe timestep restrictions would arise for explicit schemes.
For example, the timestep restriction for discretization by the explicit Euler method is

$$\tau < \frac{2}{|\lambda_{n-1}|} \approx \frac{1}{2}h^2.$$

Consequence Use implicit schemes.
advantage: reuse of ODE software
disadvantage: fixed spatial mesh

Rothe's method (first time, then space) The basic idea is the following

1. Consider

$$u' = Lu$$
$$Lu = u'' + b$$

   as an ODE in functionspace, i.e., $u(t) \in C^2(\Omega)$.

2. Apply existing ODE theory (discretization in time, stepsize control,...).

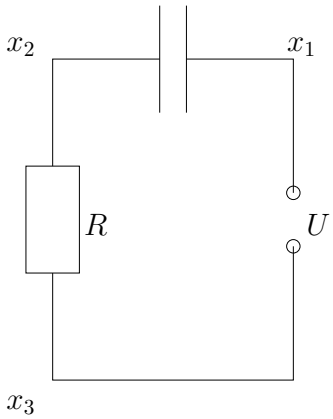
3. Approximate arising boundary value problems in each timestep.

advantage: adaptive spatial mesh
disadvantage: theoretically demanding (see [2])

2 Differential Algebraic System

2.1 Motivation

Example (Charging a capacitor (electric circuits)) We consider an electric circuit of the following form



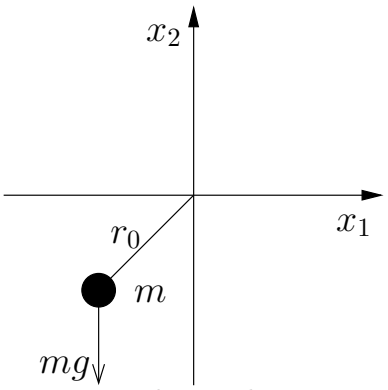
The problem can be described by the equations

applied voltage U:	$x_1 - x_3 - U = 0$
Kirchhoff's law:	$c(x_1' - x_2') + \frac{x_3 - x_2}{R} = 0$
reference value:	$x_3 = 0$

Rewriting this in matrix form yields

$$\begin{pmatrix} 0 & 0 & 0 \\ c & -c & 0 \\ 0 & 0 & 0 \end{pmatrix} x' = \begin{pmatrix} 1 & 0 & -1 \\ 0 & \frac{1}{R} & -\frac{1}{R} \\ 0 & 0 & 1 \end{pmatrix} x + \begin{pmatrix} -U \\ 0 \\ 0 \end{pmatrix}. \tag{2.1}$$

Example (Pendulum (multibody dynamics)) We consider the mathematical pendulum as depicted in the following draft



The energy of the system is composed of

kinetic energy:	$T = \frac{1}{2} m \left((x_1')^2 + (x_2')^2 \right)$
potential energy:	$U = mgx_2, \quad g$ gravity

We denote the Lagrangian by

$$L = T - U - \frac{1}{2}\lambda(x_1^2 + x_2^2 - r_0^2)$$

where λ is a Lagrange multiplier (virtual force enforcing $r_0^2 = x_1^2 + x_2^2$).

The corresponding Euler-Lagrange equations

$$\frac{d}{dt}\left(\frac{\partial L}{\partial q'}\right) - \frac{\partial L}{\partial q} = 0 \qquad q = x_1, x_2, \lambda$$

take the form

$$\begin{aligned} mx_1'' + x_1\lambda &= 0 \\ mx_2'' + x_2\lambda + mg &= 0 \\ x_1^2 + x_2^2 - r_0^2 &= 0. \end{aligned}$$

Rewriting this in vector form yields

$$\begin{pmatrix} m & 0 & 0 \\ 0 & m & 0 \\ 0 & 0 & 0 \end{pmatrix} x'' = \begin{pmatrix} -x_1\lambda \\ -x_2\lambda - mg \\ x_1^2 + x_2^2 - r_0^2 \end{pmatrix}. \tag{2.2}$$

Example (Chemical engineering) For another example see [4].

Remark (2.1) and (2.2) are mixtures of differential and algebraic equations. Therefore, they are called *differential-algebraic equations* (DAEs).

How to treat DAEs

First option: Eliminate one or more unknowns by the algebraic equation

We consider the first example and insert $x_3 = 0$, $x_1 = U$ to obtain

$$(2.1) \Leftrightarrow -cx_2' - \frac{x_2}{R} = 0.$$

Second option: Reformulate the problem in suitable unknown functions

We consider the second example and introduce polar coordinates

$$\begin{aligned} x_1 &= r\cos\phi \\ x_2 &= r\sin\phi. \end{aligned}$$

The derivatives are

$$\begin{aligned} x_1' &= -r\phi'\sin\phi \\ x_2' &= r\phi'\cos\phi \\ x_1'' &= -r\phi''\sin\phi - r(\phi')^2\cos\phi \\ x_2'' &= r\phi''\cos\phi - r(\phi')^2\sin\phi. \end{aligned}$$

Inserting into (2.2) we get

$$-mr\phi''\sin\phi - mr(\phi')^2\cos\phi + \lambda r\cos\phi = 0 \tag{2.3}$$

$$mr\phi''\cos\phi - mr(\phi')^2\sin\phi + \lambda r\sin\phi + mg = 0 \tag{2.4}$$

$$r^2(\cos^2\phi + \sin^2\phi) - r_0^2 = 0 \qquad \Rightarrow r(t) \equiv r_0. \tag{2.5}$$

Computing $((2.3)\cdot\sin\phi - (2.4)\cdot\cos\phi)$ we get

$$mr\phi'' + mg\cos\phi = 0$$

which is an ODE again.

Polar coordinates are minimal coordinates for the pendulum.

Unfortunately in most practical applications

- the algebraic equation part cannot be converted in closed form.
- minimal coordinates are not available.

2.2 Linear DAEs: Existence and Uniqueness

We consider the implicit system

$$Ex' = Ax - b, \quad E, A \in \mathbb{R}^{d\times d},\ b \in \mathbb{R}^d. \tag{2.6}$$

First we will take a look at the two **extreme cases:**

1. If E is regular, then multiplication by E^{-1} yields

   $$x' = E^{-1}Ax - E^{-1}b.$$

   This is a standard ODE system.

2. If $E = 0$, then (2.6) takes the form

   $$Ax = b.$$

   This is a standard linear system.

Hence, existence results for DAEs generalize existence results for ODEs and linear algebraic systems.
The basic idea for the analysis of (2.6) is the decoupling by suitable transformation of E and A.

Definition 2.2.1 *The pairs* (E_1, A_1) *and* (E_2, A_2) *are* equivalent*, i.e.,*

$$(E_1, A_1) \sim (E_2, A_2)$$

if there exist $P, Q \in \mathbb{R}^{d\times d}$ *regular such that*

$$E_2 = PE_1Q, \quad A_2 = PA_1Q. \tag{2.7}$$

(2.7) is called equivalence transformation*.*

Remark The relation $\sim$ is an equivalence relation (reflexive, symmetric, transitive).

Proof Exercise

Example Let $A \in \mathbb{R}^{d\times d}$, $I = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} \in \mathbb{R}^{d\times d}$. Then

$$(I, A) \sim (I, J)$$

with Jordan normal form $J = \operatorname{diag}(J_1, \ldots, J_m)$.

Proposition 2.2.2 *Let* $(E_1, A_1) \sim (E_2, A_2)$*, i.e.,*

$$E_2 = PE_1Q$$
$$A_2 = PA_1Q.$$

Then x_1 *solves* $E_1x_1' = A_1x_1$ *if and only if* $x_2 = Q^{-1}x_1$ *solves* $E_2x_2' = A_2x_2$*.*

Proof

$$\begin{aligned} E_1x_1' = A_1x_1 &\Leftrightarrow PE_1QQ^{-1}x_1' = PA_1QQ^{-1}x_1 \\ &\Leftrightarrow E_2(Q^{-1}x_1)' = A_2Q^{-1}x_1 \end{aligned}$$

□

We will need some facts from linear algebra for later use.

Definition 2.2.3 *Let* $E,\ A \in \mathbb{R}^{d\times d}$*. The polynomial*

$$p(\lambda) = \det(\lambda E - A)$$

is called characteristic polynomial *of* (E, A)*.*
The pair (E, A) *is called* singular *if* $p(\lambda) \equiv 0$ *and* regular *otherwise.*

Remark Regularity/singularity is invariant under equivalence transformation.

Examples

1. $(0, A)$ is regular $\Leftrightarrow$ A is regular.
2. $\left(\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right)$ is singular.

Proposition 2.2.4 *Let* (E, A) *be singular. Then the homogeneous initial value problem*

$$Ex'(t) = Ax(t), \quad t > 0, \qquad x(0) = 0,$$

has at least two solutions.

Proof Obviously, $x \equiv 0$ solves the homogeneous initial value problem. We now show that there is also another solution $x \not\equiv 0$.

Let $\lambda_1, \ldots, \lambda_{d+1} \in \mathbb{R}$, $\lambda_i \neq \lambda_j$ for $i \neq j$.
By assumption $\lambda_iE - A$ is singular for $i = 1, \ldots, d+1$. Hence,

$$\exists v_i \in \mathbb{R}^d, v_i \neq 0 : (\lambda_iE - A)v_i = 0.$$

$v_1, \ldots, v_{d+1}$ cannot be linearly independent. Hence,

$$\exists \alpha_i \in \mathbb{R} : \sum_{i=1}^{d+1} \alpha_i v_i = 0.$$

Define

$$x(t) := \sum_{i=1}^{d+1} \alpha_i v_i e^{\lambda_i t} \not\equiv 0.$$

Then $x(0) = 0$ and

$$Ex' = \sum_{i=1}^{d+1} \alpha_i e^{\lambda_i t} \lambda_i E v_i = \sum_{i=1}^{d+1} \alpha_i e^{\lambda_i t} A v_i = Ax.$$

□

Does regularity of (E, A) imply existence of solutions of (2.6)?

Proposition 2.2.5 *Let $\mathcal{L} = \{M = (m_{ij}) \in \mathbb{R}^{d\times d} | m_{ij} = 0,\ j > i\}$ (lower triangular matrices) and $\mathcal{L}_0 = \{M \in \mathcal{L} | m_{ii} = 0\}$ (strictly lower triangular matrices).*

1. *$L \in \mathcal{L}$ regular $\Rightarrow L^{-1} \in \mathcal{L}$*
2. *$L_0 \in \mathcal{L}_0 \Rightarrow L_0^d = 0$ (nilpotent)*
3. *$L \in \mathcal{L},\ L_0 \in \mathcal{L}_0 \Rightarrow L \cdot L_0 \in \mathcal{L}_0$*

Proof Exercise

Proposition 2.2.6 *Let (E, A) be regular. Then*

$$(E, A) \sim \left(\begin{pmatrix} I & 0 \\ 0 & N \end{pmatrix}, \begin{pmatrix} J & 0 \\ 0 & I \end{pmatrix} \right)$$

where $J \in \mathbb{R}^{n\times n}$, $N \in \mathbb{R}^{m\times m}$ with $d = n + m$, both J and N have Jordan normal form and $N^\nu = 0$, $\nu \le m$.

Proof (E, A) regular $\Rightarrow \exists \lambda_0 \in \mathbb{R} : \det(\lambda_0 E - A) \neq 0$,
i.e., $(A - \lambda_0 E)^{-1} \in \mathbb{R}^{d\times d}$ exists.
Hence,

$$\begin{aligned} (E, A) &\sim (E, A - \lambda_0 E + \lambda_0 E) \\ &\sim \left((A - \lambda_0 E)^{-1} E, I + \lambda_0 (A - \lambda_0 E)^{-1} E\right). \quad \Big| \begin{matrix} P = (A - \lambda_0 E)^{-1} \\ Q = I \end{matrix} \end{aligned}$$

Let $\operatorname{diag}(\bar{J}, \bar{N}) = T\left((A - \lambda_0 E)^{-1} E\right) T^{-1}$ be the Jordan normal form of $(A - \lambda_0 E)^{-1} E$ where $\bar{J}$ is regular (non-zero eigenvalues) and $\bar{N} \in \mathcal{L}_0$ (eigenvalue zero). Then

$$\begin{aligned} (E, A) &\sim \left(\begin{pmatrix} \bar{J} & 0 \\ 0 & \bar{N} \end{pmatrix}, \begin{pmatrix} I + \lambda_0 \bar{J} & 0 \\ 0 & I + \lambda_0 \bar{N} \end{pmatrix} \right) \quad \Big| \begin{matrix} P = T(A - \lambda_0 E)^{-1} \\ Q = T^{-1} \end{matrix} \\ &\bar{N} \in \mathcal{L}_0 \Rightarrow I + \lambda_0 \bar{N} \in \mathcal{L} \text{ regular} \\ &\sim \left(\begin{pmatrix} I & 0 \\ 0 & (I + \lambda_0 \bar{N})^{-1} \bar{N} \end{pmatrix}, \begin{pmatrix} \bar{J}^{-1} + \lambda_0 I & 0 \\ 0 & I \end{pmatrix} \right) \quad \Big| \begin{matrix} P = \begin{pmatrix} \bar{J}^{-1} & 0 \\ 0 & (I + \lambda_0 \bar{N})^{-1} \end{pmatrix} \\ Q = I \end{matrix} \end{aligned}$$

$(I + \lambda_0 \bar{N})^{-1} \in \mathcal{L}, \bar{N} \in \mathcal{L}_0 \Rightarrow (I + \lambda_0 \bar{N})^{-1} \bar{N} \in \mathcal{L}_0 \Rightarrow \left((I + \lambda_0 \bar{N})^{-1} \bar{N}\right)^\nu = 0, \qquad \nu \le m.$

Transformation of $\bar{J}^{-1} + \lambda_0 I$ and $(I + \lambda_0 \bar{N})^{-1} \bar{N}$ to Jordan normal form concludes the proof. □

Using Proposition 2.2.6 provides the decoupling

$$\begin{aligned} y' &= Jy + f, \qquad y(0) = y_0 \quad (2.8)\\ Nz' &= z + g, \qquad z(0) = z_0 \quad (2.9) \end{aligned}$$

with suitable f, g.
Existence and uniqueness for (2.8) is clear. We consider (2.9).

Proposition 2.2.7 *Assume that $g \in C^{\nu}([0,T],\mathbb{R}^m)$ and N from Proposition 2.2.6. Then, without any initial conditions, the differential equation*

$$Nz' = z + g$$

has the unique solution

$$z = -\sum_{i=0}^{\nu-1} N^i g^{(i)}$$

with ν denoting the size of the largest Jordan block of N.

Proof We denote

$$N = \operatorname{diag}(N_i); \qquad N_i = \begin{pmatrix} 0 & & & 0 \\ 1 & 0 & & \\ & \ddots & \ddots & \\ 0 & & 1 & 0 \end{pmatrix}.$$

The different blocks N_i decouple. Hence, it is sufficient to consider $N = \begin{pmatrix} 0 & & & 0 \\ 1 & 0 & & \\ & \ddots & \ddots & \\ 0 & & 1 & 0 \end{pmatrix}$.

Componentwise reformulation yields

$$\begin{aligned} 0' &= z_1 + g_1 \quad &\Rightarrow \quad z_1 &= -g_1\\ z_1' &= z_2 + g_2 \quad &\Rightarrow \quad z_2 &= -g_1' - g_2\\ z_2' &= z_3 + g_3 \quad &\Rightarrow \quad z_3 &= -g_1'' - g_2' - g_3\\ &\vdots \\ z_{m-1}' &= z_m + g_m \quad &\Rightarrow \quad z_m &= -\sum_{i=1}^{m} g_i^{(m-i)} \end{aligned}$$

with unique solutions z_i. After taking a closer look at the powers N^i, $i = 1,\dots,m-1$, of N, elementary calculations yield

$$z = -\big(Ig + Ng' + N^2 g'' + \cdots + N^{m-1} g^{(m-1)}\big) = -\sum_{i=0}^{m-1} N^1 g(i).$$

□

Remark Existence requires consistent initial data

$$z_0 = -\sum_{i=0}^{\nu-1} N^i g^{(i)}(0).$$

Remark For $\nu > 1$, existence requires smoothness of the right hand side

$$g \in C^\nu\left([0,T], \mathbb{R}^m\right).$$

Definition 2.2.8 *The index ν occurring in Proposition 2.2.7 is called* (differentiation) index *of (2.6). We set $\nu = 0$ if $m = 0$, i.e., if E is regular.*

Remark The index is invariant under equivalence transformations (exercise).

Examples

$$\begin{aligned} Ex' &= Ax - b && E \text{ regular} && \Rightarrow \nu = 0 \\ 0 &= Ax - b && && \Rightarrow \nu = 1 \\ \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} x' &= Ax - b && && \Rightarrow \nu = 2 \end{aligned}$$

Definition 2.2.9 *Let*

$$PEQ = \begin{pmatrix} I & 0 \\ 0 & N \end{pmatrix}, \quad PAQ = \begin{pmatrix} J & 0 \\ 0 & I \end{pmatrix}, \quad Pb = \begin{pmatrix} f \\ g \end{pmatrix}$$

with N, J as in Proposition 2.2.6. The initial condition $x_0 \in \mathbb{R}^d$ is called consistent *with (2.6) if*

$$Q^{-1}x_0 = \begin{pmatrix} y_0 \\ z(0) \end{pmatrix}$$

with arbitrary $y_0 \in \mathbb{R}^n$ and $z = -\sum_{i=0}^{\nu-1} N^i g^{(i)}$.

Theorem 2.2.10 *Let (E, A) be regular with index $\nu \leq d$, $b \in C^\nu\left([0,T], \mathbb{R}^d\right)$, $x_0 \in \mathbb{R}^d$ consistent with (2.6). Then the initial value problem*

$$Ex'(t) = Ax(t) - b, \quad t > 0, \qquad x(0) = x_0$$

for (2.6) has a unique solution.

Proof Proposition 2.2.6, 2.2.7

Index-1-problems

As an important special case we consider the semi-explicit DAE

$$\begin{aligned} y' &= Ay + Bz + f && A \in \mathbb{R}^{n\times n}, \quad B \in \mathbb{R}^{n\times m}, \quad f : [0,1] \to \mathbb{R}^n \\ 0 &= Cy + Dz + g && C \in \mathbb{R}^{m\times n}, \quad D \in \mathbb{R}^{m\times m}, \quad g : [0,1] \to \mathbb{R}^m. \end{aligned} \tag{2.10}$$

Proposition 2.2.11 *Let $D \in \mathbb{R}^{m\times m}$ be regular. Then (2.10) has index $\nu = 1$.*

Proof

$$\begin{aligned}
\left(\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} A & B \\ C & D \end{pmatrix}\right) &\sim \left(\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} A & B \\ D^{-1}C & I \end{pmatrix}\right) && \mid P = \begin{pmatrix} I & 0 \\ 0 & D^{-1} \end{pmatrix} \\
&&& \quad Q = I \\
&\sim \left(\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} A - BD^{-1}C & 0 \\ D^{-1}C & I \end{pmatrix}\right) && \mid P = \begin{pmatrix} I & -B \\ 0 & I \end{pmatrix} \\
&&& \quad Q = I \\
&\sim \left(\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} A - BD^{-1}C & 0 \\ 0 & I \end{pmatrix}\right) && \mid P = I \\
&&& \quad Q = \begin{pmatrix} I & 0 \\ -D^{-1}C & I \end{pmatrix} \\
&\sim \left(\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} J & 0 \\ 0 & I \end{pmatrix}\right) && \mid P = \begin{pmatrix} T^{-1} & 0 \\ 0 & I \end{pmatrix} \\
&&& \quad Q = \begin{pmatrix} T & 0 \\ 0 & I \end{pmatrix}
\end{aligned}$$

where $J = T^{-1}(A - BD^{-1}C)T$ is a Jordan decomposition. □

Remark The elimination of z is possible:

$$z = -D^{-1}(Cy + g).$$

Thus, we obtain the ODE

$$y' = (A - BD^{-1}C)y - BD^{-1}g + f$$

(state space form, "model reduction").

Remark Consistent initial data $z_0 \in \mathbb{R}^m$ can be computed from

$$-Dz_0 = Cy_0 + g(0)$$

with given $y_0 \in \mathbb{R}^n$.

Index reduction

Differentiation leads to

$$\begin{aligned}
y' &= Ay + Bz + f \\
-Cy' - Dz' &= g'
\end{aligned}$$

which can be rewritten as the ODE

$$\begin{pmatrix} y' \\ z' \end{pmatrix} = \begin{pmatrix} I & 0 \\ -C & -D \end{pmatrix}^{-1} \begin{pmatrix} A & B \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} + \begin{pmatrix} f \\ g' \end{pmatrix}. \tag{2.11}$$

Each solution of the DAE (2.10) solves (2.11) provided that $g \in C^1[0,T]$. For the converse we refer to Lemma 2.3.3 later.

Remark There is a special perturbation theory (perturbation index) for DAEs (see [4, 3.1.3], [11, 3.4]).

2.3 Nonlinear Semi-explicit DAEs

We consider the initial value problem for the semi-explicit DAE

$$\begin{array}{ll} y' = f(y,z) & y(0) = y_0 \\ 0 = g(y,z) & z(0) = z_0 \end{array} \tag{2.12}$$

for $t \in (0,T]$.
Let z_0 be consistent with the DAE in the sense that $0 = g(y_0, z_0)$.

Theorem 2.3.1 (local existence) *Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ and $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be continuously differentiable with*

$$g_z(y,z) \text{ invertible } \forall y \in \mathbb{R}^n, z \in \mathbb{R}^m. \tag{2.13}$$

Then (2.12) has a unique solution for all $y_0 \in \mathbb{R}^n$, consistent data $z_0 \in \mathbb{R}^m$ and sufficiently small T.

Proof Consistency implies $g(y_0, z_0) = 0$. By the implicit function theorem, there are $U \subset \mathbb{R}^n$ open with $y_0 \in U$, $V \subset \mathbb{R}^m$ open with $z_0 \in V$ and $G : U \to V$ differentiable such that

$$g(y,z) = 0 \quad (y,z) \in U \times V \Leftrightarrow z = G(y) \quad y \in U.$$

Inserting yields the state space form (model reduction)

$$y' = f\left(y, G(y)\right) \quad y \in U. \tag{2.14}$$

$F(y) := f\left(y, G(y)\right)$ differentiable implies that F is locally Lipschitz. Therefore, there exists $T > 0$ such that (2.14) has a unique solution. □

Remark DAEs (2.12) with property (2.13) are called index-1. See [11, Chapter 4] for a more detailed discussion.

Remark Computation of consistent initial data z_0 requires the solution of the nonlinear system

$$z_0 \in \mathbb{R}^m : \quad g(y_0, z_0) = 0.$$

The state space form (model reduction)

Basic idea:

- Eliminate $z = G(y) \Leftrightarrow g(y,z) = 0$.
- Apply an arbitrary discretization for the resulting state space form

$$y' = f\left(y, G(y)\right) = F(y).$$

State space RK-method in symmetric form

$$\Psi^\tau y = y + \tau \sum_{i=1}^{s} b_i F(Y_i)$$
$$Y_i = y + \tau \sum_{j=1}^{s} a_{ij} F(Y_j)$$

Introduction of $Z_i = G(Y_i) \Leftrightarrow 0 = g(Y_i, Z_i)$

$$\Psi^\tau y = y + \tau \sum_{i=1}^{s} b_i f(Y_i, Z_i)$$
$$Y_i = y + \tau \sum_{j=1}^{s} a_{ij} f(Y_j, Z_j)$$
$$0 = g(Y_i, Z_i) \qquad i = 1, \ldots, s$$

Proposition 2.3.2 *Assume that $g_z(y,z)$ is invertible for all $y \in \mathbb{R}^n$, $z \in \mathbb{R}^m$. Then the state space RK-method converges with order p of Ψ^τ.*

Proof Direct application of existing convergence theory

Advantages and drawbacks

- $\oplus$ reuse of existing discretizations and theory
- $\ominus$ explicit resolution of $g(y,z) = 0$ might cause very small time steps (implicit function theorem)
- $\ominus$ each time step requires the solution of at least s nonlinear systems with m unknowns

The index reduction method

We differentiate the algebraic equation

$$0 = g_y(y,z)y' + g_z(y,z)z'.$$

Rewriting this in explicit matrix form, assuming $g_z(y,z) \neq 0$, we obtain

$$\begin{pmatrix} y \\ z \end{pmatrix}' = \begin{pmatrix} I & 0 \\ g_y(y,z) & g_z(y,z) \end{pmatrix}^{-1} \begin{pmatrix} f(y,z) \\ 0 \end{pmatrix} \quad \begin{matrix} y(0) = y_0 \\ z(0) = z_0. \end{matrix} \tag{2.15}$$

Lemma 2.3.3 *For sufficiently smooth g and consistent intial data z_0 the initial value problems (2.12) and (2.15) are equivalent.*

Proof (2.12) $\Rightarrow$ (2.15): differentiation
(2.15) $\Rightarrow$ (2.12): $g(y,z) = 0$ follows from

$$\begin{aligned} g(y,z) &= g(y_0,z_0) + \int_0^t (g(y,z))'\, ds \\ &= 0 + \int_0^t \left(g_y y' + g_z z'\right) ds = 0. \end{aligned}$$

□

Basic idea:

- Reformulate (2.12) as ODE (2.15) with consistent initial data.
- Apply suitable discretization to (2.15).

Advantages and drawbacks

⊕ simple reuse of existing discretizations

⊖ differentiation necessary (problem if g is not available in closed form)

⊖ structure of the problem is destroyed

Alternative: Linearly implicit extrapolation methods [4, 6.4.2]

3 Hamiltonian Systems

3.1 Energy and Symplecticity

In physics, classical mechanics are described by differential equations. The concept of energy is of fundamental importance in this domain and mostly considered in the context of closed Hamiltonian systems. The conservation of mechanical energy is a principle which states that under certain conditions, the total mechanical energy of a system is constant. This leads us to the concept of symplecticity.
As the topic of this section originated in this field we introduce this chapter by stating a fundamental law of classical mechanics.

Example (Newton's second law: conservation of momentum) Newton's second law states that the rate of change of momentum of a body is proportional to the resultant force acting on the body and is in the same direction. In symbolic notation this can be written as

$$Mx'' = F(x)$$
$$x : [0, T] \to \mathbb{R}^d$$

where
x denotes the location of the center of gravity of one or more bodies in space,
x' denotes the velocity,
x'' denotes the acceleration,
$M \in \mathbb{R}^{d\times d}$ denotes the symmetric, positive definite mass matrix,
Mx'' denotes the momentum,
$F : \mathbb{R}^d \to \mathbb{R}^d$ denotes the force field.

Thus, Newton's second law is described by a differential equation typically fulfilling certain characteristics.

Definition 3.1.1 *F is called* conservative *or* potential force *if there is a potential $U : \mathbb{R}^d \to \mathbb{R}$ of F, i.e., $F = -\nabla U$.*

Remark If the Jacobian F' is symmetric, then F is conservative.
F conservative:

$$Mx'' = -\nabla U(x) \tag{3.1}$$

We will now introduce some examples taken from the natural sciences.

Example (Pendulum) A classical example is the mathematical pendulum which is described by the equation

$$m\phi'' = -m\frac{g}{r_0}\cos\phi = -U'(\phi)$$
$$U(\phi) = m\frac{g}{r_0}\sin\phi$$

where ϕ denotes the angle, g the gravity, m the mass and r_0 the radius.
For this problem there is an analytic solution available (using elliptic integrals).

Example (Kepler problem) The Kepler problem is a two-body problem arising in celestial mechanics. There are two bodies (planets) which motion is affected by the attraction between them. Let x_1 and x_2 denote their state in space, m_1 and m_2 their masses, $r_{12} = \|x_1 - x_2\|$ the distance between them, and $U = -g\frac{m_1 m_2}{r_{12}}$ the gravitational potential. Newton's law yields

$$\begin{aligned} m_1 x_1'' &= -\frac{\partial}{\partial x_1} U \\ m_2 x_2'' &= -\frac{\partial}{\partial x_2} U \\ Mx'' &= -\nabla U \\ M &= \begin{pmatrix} m_1 I & 0 \\ 0 & m_2 I \end{pmatrix} (\in \mathbb{R}^6)\,. \end{aligned}$$

Example (Classical molecular dynamics) Another example is the interaction of a vast number of N atoms which spatial coordinates are denoted by $X_i \in \mathbb{R}^3$, $i = 1, ..., N$ and the distances in between by $r_{ij} = \|x_i - x_j\|$. The potential of the force field consists of

$$U = U_B + U_A + U_T + U_Q + U_{VdW}$$

where

$$U_B = \sum_{\substack{i,j=1 \\ i>j}}^{N} \frac{1}{2} b_{ij} \left(r_{ij} - r_{ij}^*\right)^2 : \text{ bond deformation}$$

(deviation from a reference state x_i^* with $r_{ij}^* = \left\|x_i^* - x_j^*\right\|$ (analogue of springs))

U_T : torsion potential

U_A : angle deformation

U_Q : Coulomb potential

U_{VdW} : van-der-Waals interaction (quantum effects).

For the evaluation of $-\nabla U$ fast multipole methods can be used (Greengard/Rokhlin).

We consider the conservative system

$$Mx'' = -\nabla U(x). \tag{3.2}$$

The *energy* of a state x at time t is

$$E(x(t)) = \underbrace{\frac{1}{2}\langle x'(t), Mx'(t)\rangle}_{\text{kinetic}} + U(x(t)).$$

The energy of the pendulum is for example

$$E(\phi) = \frac{1}{2} m(\phi')^2 + m\frac{g}{r_0}\sin(\phi).$$

Proposition 3.1.2 *The energy $E(x(t))$ of (3.2) is conserved throughout the evolution.*

Proof It holds

$$\begin{aligned}\frac{d}{dt}E(x(t)) &= \langle x'(t), Mx''(t)\rangle + \langle \nabla U(x(t)), x'(t)\rangle \\ &= \langle x'(t), Mx''(t) + \nabla U(x(t))\rangle \\ &= \langle x'(t), 0\rangle \\ &= 0.\end{aligned}$$

□

Hamiltonian systems

Let $p = Mx'$; $q = x$; $p, q \in \mathbb{R}^d$. Then the *Hamiltonian* is defined by

$$H(p,q) = \frac{1}{2}\langle M^{-1}p, p\rangle + U(q) = \frac{1}{2}\langle x', Mx'\rangle + U(x) = E(x).$$

Hamiltonian system:

$$\begin{aligned} q' &= \quad H_p = \tfrac{\partial}{\partial p}H(p,q) = M^{-1}p \\ p' &= -H_q = -\tfrac{\partial}{\partial q}H(p,q) = -\nabla U(q) \end{aligned} \tag{3.3}$$

We assume from now on that (3.3) has a unique solution for all initial values $x_0 = \binom{p}{q} \in \mathbb{R}^{2d}$ such that $\Phi^t : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ exists.
Example (pendulum):

$$\begin{aligned} q' &= m^{-1}p \\ p' &= -m\tfrac{g}{r_0}\cos q \end{aligned} \quad \Leftrightarrow \quad mq'' = -m\frac{g}{r_0}\cos q$$

Remark (3.3) can be rewritten as

$$y' = -J\nabla H(y)$$

where

$$y = \begin{pmatrix} p \\ q \end{pmatrix}, \quad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}, \quad \nabla H(y) = \begin{pmatrix} H_p \\ H_q \end{pmatrix}.$$

Notice that for J holds

$$J^{-1} = -J = J^T.$$

Reminder (condition number of initial value problems) Let the flow $\Phi^t y_0$ be the solution of (3.3) with initial condition $y_0 = \binom{p_0}{q_0}$.

1. For fixed $t > 0$, the pointwise condition number $\kappa(t)$ is the smallest number with the property

$$\|\Phi^t(y_0 + \Delta y_0) - \Phi^t y_0\| \le \kappa(t)\|\Delta y_0\| + o(\|\Delta y_0\|) \quad \Delta y_0 \to 0.$$

2. The derivative of $\Phi^t y_0$ w.r.t. y_0 takes the form $\Psi(t) := D_y \Phi^t y|_{y=y_0}$ (Wronski matrix) and
$$\kappa(t) \leq \|\Psi(t)\|.$$

3. Ψ solves the equation
$$\Psi' = W(\Phi^t y_0)\Psi$$
with
$$W(\Phi^t y_0) = -J\nabla^2 H(y) = -J \begin{pmatrix} H_{pp} & H_{pq} \\ H_{qp} & H_{qq} \end{pmatrix}.$$

Theorem 3.1.3 (Poincaré 1899) *Let $H(p,q)$ be twice continuously differentiable. Then $\Psi(t)^T J \Psi(t) = J$ holds for each $t > 0$ where Ψ is defined.*

Proof We have
$$\begin{aligned} \frac{d}{dt}\left(\Psi^T J \Psi\right) &= \Psi'^T J \Psi + \Psi^T J \Psi' \\ &= (W\Psi)^T J \Psi + \Psi^T J W \Psi \\ &= -(J\nabla^2 H \Psi)^T J \Psi + \Psi^T J(-J)\nabla^2 H \Psi \\ &= -\Psi^T (\nabla^2 H)^T \underbrace{J^T J}_{I} \Psi + \Psi^T \underbrace{J(-J)}_{I} \nabla^2 H \Psi \\ &= -\Psi^T \nabla^2 H \Psi + \Psi^T \nabla^2 H \Psi \\ &= 0. \end{aligned}$$

Finally
$$\Psi(0) = \frac{d}{dy}\Phi^0 y|_{y=y_0} = I$$
provides
$$\Psi(0)^T J \Psi(0) = J.$$

□

Definition 3.1.4 *A transformation $\Phi : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ satisfying $(D\Phi)^T J (D\Phi) = J$ is called* symplectic.

Remark The flow Φ^t of a Hamiltonian system is symplectic by Theorem 3.1.3. Conversely if $y' = f(y)$ is symplectic then it is locally Hamiltonian, i.e.,
$$\forall y_0 \in \mathbb{R}^{2d}\ \exists H :\ f(y) = -\nabla H(y)$$
in a neighborhood of y_0. [7, VI.2.6]

Proposition 3.1.5 *Symplectic transformations are area-preserving.*

Proof We take $d = 1$ for simplicity and let

$$\Phi\begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} f(p,q) \\ g(p,q) \end{pmatrix} \qquad \text{with the Jacobian } D\Phi = \begin{pmatrix} f_p & f_q \\ g_p & g_q \end{pmatrix}.$$

From symplecticity

$$(D\Phi)^T J (D\Phi) = \begin{pmatrix} 0 & f_p g_q - f_q g_p \\ -(f_p g_q - f_q g_p) & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

we can conclude

$$\det D\Phi = f_p g_q - f_q g_p = 1.$$

For each measurable $\Omega \subset \mathbb{R}^2$, we have

$$|\Omega| = \int_\Omega |\det D\Phi| \, ds = \int_{\Phi^{-1}\Omega} ds = \left|\Phi^{-1}\Omega\right|.$$

□

Example (The pendulum and V.I. Arnold's cats) [7, VI.2. Fig 2.2]

3.2 Symplectic Runge-Kutta-Methods

We consider the Hamiltonian system

$$\begin{aligned} y' &= -J\nabla H(y) \\ H(p,q) &= \frac{1}{2}\langle p, Mp\rangle + U(q), \quad y = \begin{pmatrix} p \\ q \end{pmatrix}. \end{aligned} \tag{3.4}$$

As an example to illustrate the behavior of different numerical methods we consider the pendulum

$$q' = p; \quad p' = -\tfrac{g}{r_0}\cos q; \quad q(0) = \pi; \quad p(0) = 0$$

and apply the explicit Euler method, the implicit Euler method and the midpoint rule.

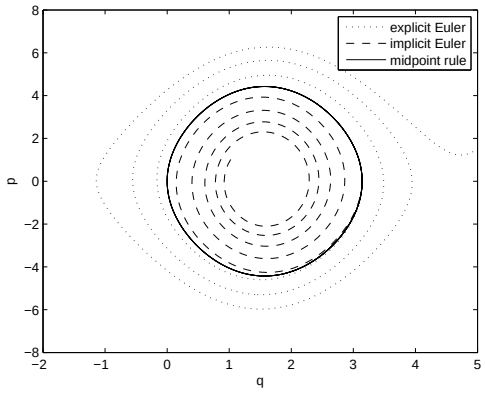


Figure 3.1: Solutions of the pendulum

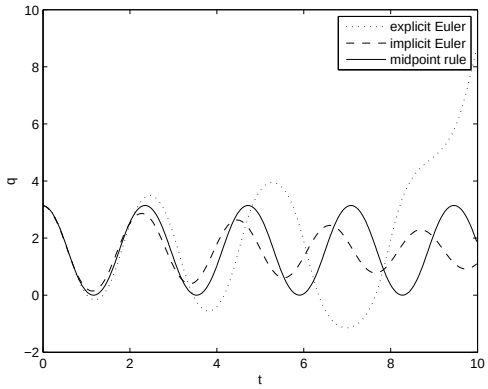


Figure 3.2: Solutions being subject to time

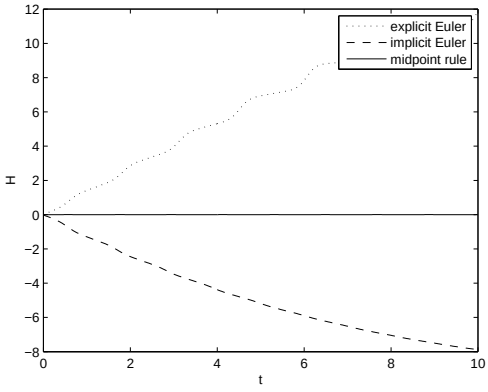


Figure 3.3: Hamiltonians

We observe that the explicit Euler seems to generate energy while the implicit Euler seems to consume energy and the midpoint rule seems to preserve energy.

Definition 3.2.1 *A Runge-Kutta-method is called* symplectic *if its discrete flow* Ψ^τ *is sym-*

plectic.

Which RK-methods are symplectic?
Consider

$$x' = f(x) \tag{3.5}$$

and let

$$G(x) = x^T Ax + b^T x + c$$

be a quadratic first integral of (3.5), i.e.,

$$G(x) = G(\Phi^t x), \quad t > 0.$$

Lemma 3.2.2 *G is a first integral if and only if $\nabla G(x) \cdot f(x) = 0 \quad \forall x \in \mathbb{R}^d$.*

Proof Exercise

Theorem 3.2.3 *The discrete flow Ψ^τ of every Gauß method preserves quadratic first integrals, i.e.,*

$$G(x) = G(\Psi^\tau x) \quad \forall x \in \mathbb{R}^d.$$

Proof Gauß methods are RK-methods, say of stage s.
Let u be the collocation polynomial, $\deg u = s$, with

$$u(0) = x, \quad u(\tau) = \Psi^\tau x, \quad u'(c_j\tau) = f(u(c_j\tau)), \quad j = 1, \dots, s$$

where $c_j\tau$, $j = 1, \dots, s$ denote the Gauß points. Then it holds

$$\int_0^1 p(\theta)d\theta = \sum_{j=1}^s b_j p(c_j), \quad \deg p \leq 2s - 1.$$

Denote $q(\theta) = G(u(\tau\theta))$, which is a polynomial of degree $\leq 2s$. Then

$$\begin{aligned} G(\Psi^\tau x) &= G(u(\tau)) \\ &= q(1) \\ &= q(0) + \int_0^1 q'(\theta)d\theta \\ &= q(0) + \sum_{j=1}^s b_j q'(c_j). \end{aligned}$$

Use chain rule, collocation conditions and Lemma 3.2.2 to obtain

$$\begin{aligned} q'(c_j) &= \tau \nabla G(u(\tau c_j))u'(\tau c_j) \\ &= \tau \nabla G(u(\tau c_j))f(u(\tau c_j)) \\ &= \tau \cdot 0 \\ &= 0. \end{aligned}$$

Hence,

$$G(\Psi^\tau x) = q(0) = G(u(0)) = G(x).$$

□

Corollary 3.2.4 *Every Gauß method is symplectic.*

Proof Consider the augmented Hamiltonian system

$$\begin{array}{lll} y' & = -J\nabla H(y) & y(0) = y_0 \\ \Psi' & = W(y)\Psi & \Psi(0) = I. \end{array} \tag{3.6}$$

By Theorem 3.1.3

$$G(y, \Psi) = (\Psi)^T J \Psi = J$$

is a quadratic first integral of (3.6). Hence, G is preserved by any Gauß method. □

Of course we approximate y without approximating Ψ in practice.

Remark The midpoint rule is a Gauß method and therefore symplectic.

There are other symplectic RK-methods.

Proposition 3.2.5 *The so-called symplectic Euler method*

$$y_{n+1} = \Psi^\tau y_n = \begin{pmatrix} p_{n+1} \\ q_{n+1} \end{pmatrix} = y_n + \tau \begin{pmatrix} -H_q(p_{n+1}, q_n) \\ H_p(p_{n+1}, q_n) \end{pmatrix}$$

is symplectic.

Proof Exercise

Proposition 3.2.6 *The trapezoidal rule* $\Psi^\tau x = x + \frac{\tau}{2}(f(x) + f(\Psi^\tau x))$ *is not symplectic.*

Proof Exercise

Do symplectic integrators preserve H?
No, but almost!

Theorem 3.2.7 *Assume that* H *is analytic,* Ψ *is symplectic and of order* p *and*

$$\exists K \subset \mathbb{R}^{2d}, K \, compact: \ \Phi^t y_0 \in K \ \forall t \geq 0$$

Then there is a $\tau_0 > 0$ *such that*

$$H(\Psi^{n\tau} y_0) = H(y_0) + O(\tau^p)$$

for exponentially long time intervals $n\tau \leq e^{\frac{\tau_0}{2\tau}}$.

Proof [7, IX.8.1]

Challenges

1. multibody system: equality constraints, inequality constraints
2. classical molecular dynamics: preserving adiabatic invariants over exponentially long times

4 Iterative Methods for Linear Systems

We consider the linear system

$$Ax = b, \qquad A \in \mathbb{R}^{n\times n} \text{ regular}, \ b \in \mathbb{R}^n$$

To solve this system we already know the following direct methods [9]:

- Gaussian elimination (LR decomposition)
- QR decomposition

  $$A = QR, \ Ax = b \ \Rightarrow \ Qy = b, \ Rx = y$$
  $$\kappa(A) = \kappa(R), \ \kappa(Q) = 1$$

  with $\kappa(A)$ the condition number of A.

4.1 Motivation (Why Iterative Solutions?)

Numerical example:

$$A_n \in \mathbb{R}^{n\times n}, \qquad (A_n)_{ij} = \frac{1}{i+j-1} \text{ Hilbertmatrix}$$

$$\kappa(A_n) \to \infty \text{ as } n \to \infty \text{ (very fast)}$$

Observation: meaningless result of QR-algorithm for $\kappa(A) \gg 1$

Simplified stability anlysis of direct QR solution : Due to round-off errors, we only compute approximations $\tilde{y}$ and $\tilde{x}$ of $y = Q^T b$ and $x = R^{-1}y$. We assume that

- $\dfrac{\|y - \tilde{y}\|}{\|y\|} \leq eps$
- $\tilde{x} = R^{-1}\tilde{y}$

with eps denoting the machine accuracy. This means that $\tilde{y}$ is obtained just by rounding y and all other round-off errors are ignored. As $\kappa(R) = \kappa(A)$ this leads to (see [10])

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \kappa(R)\frac{\|y - \tilde{y}\|}{\|y\|} + o(eps) = \kappa(A) \cdot eps + o(eps).$$

Hence, the stability of the QR algorithm is $\sigma_{\text{direct}} := \kappa(A)$. This is in agreement with our numerical experiments.

Iterative solution We consider an iteration function $G : \mathbb{R}^n \to \mathbb{R}^n$ and assume that

- $\|x - G(y)\| \le \rho\|x - y\| \qquad \forall y \in \mathbb{R}^n$

holds with the exact solution x of $Ax = b$ and the convergence rate $\rho < 1$. Then, for any initial iterate x^0, the sequence x^k, $k = 0, 1, \ldots$, produced by the iteration $x^{k+1} = G(x^k)$, $k = 0, 1, \ldots$, satisfies

$$\|x - x^k\| \le \rho^k \|x - x^0\|, \qquad k = 1, 2, \ldots.$$

In particular, the sequence (x^k) converges to x.

Simplified stability anlysis of iterative solution We assume that

- the convergence rate ρ is independent of $\kappa(A)$.

Due to round-off errors, we only compute a perturbed evaluation of the iteration function $\tilde{G}$. We assume

- $\dfrac{\|G(y) - \tilde{G}(y)\|}{\|G(y)\|} \le eps.$

This means that $\tilde{G}(y)$ is obtained just by rounding $G(y)$ and all other round-off errors are ignored. We investigate the perturbed sequence

$$\tilde{x}^{k+1} = \tilde{G}(\tilde{x}^k), \quad k = 0, 1, \ldots.$$

Note that (x^k) is bounded, because it is a convergent sequence. We assume that G is continuous and $(\tilde{x}^k)$ is bounded. As a consequence, $(G(\tilde{x}^k))$ is also bounded so that there is a constant $C > 0$ such that

$$\|G(\tilde{x}^k)\| \le C\|x\|.$$

Denoting $\epsilon^k = G(\tilde{x}^k) - \tilde{G}(\tilde{x}^k)$ we then get by induction

$$\begin{aligned}
\|x - \tilde{x}^k\| &\le \|x - G(\tilde{x}^{k-1})\| + \|\epsilon^k\| \\
&\le \rho\|x - \tilde{x}^{k-1}\| + \|\epsilon^k\| \\
&\le \rho^2\|x - \tilde{x}^{k-2}\| + \rho\|\epsilon^{k-1}\| + \|\epsilon^k\| \\
&\le \rho^k\|x - x^0\| + \sum_{i=0}^{k-1} \rho^i \|\epsilon^{k-i}\| \\
&\le \rho^k\|x - x^0\| + (1-\rho)^{-1} \max_{i=0,\ldots,k-1} \|\epsilon^{k-i}\| \\
&\le \rho^k\|x - x^0\| + (1-\rho)^{-1} \max_{i=0,\ldots,k-1} \|G(\tilde{x}^{k-i})\| eps \\
&\le \rho^k\|x - x^0\| + (1-\rho)^{-1} C\|x\| eps
\end{aligned}$$

Hence

$$\frac{\|x - \tilde{x}^k\|}{\|x\|} \le \rho^k \frac{\|x - x^0\|}{\|x\|} + (1-\rho)^{-1} C eps,$$

indicating that $\sigma_{\text{iterative}} \ll \kappa(A) = \sigma_{\text{direct}}$, if $\kappa(A)$ is large enough.

Upshot Iterative schemes can have better stability properties than direct methods.

Complexity

The computational effort (complexity) of QR factorization is $O(n^3)$.
We assume that

- the convergence rate ρ is independent of n,
- the evaluation of the iteration function G has complexity $O(n^2)$.

Note that a matrix-vector multiplication has complexity $O(n^2)$. To calculate the computational effort to achieve $\|x-x^k\| \leq \text{tol}$ where tol denotes the prescribed tolerance we compute k_0 such that $\|x - x^{k_0}\| \leq \rho^{k_0}\|x - x^0\| \leq \text{tol}$:

$$k_0 \geq \frac{\log \frac{\text{tol}}{\|x-x^0\|}}{\log \rho}.$$

The complexity of the evaluation of x^{k_0} is bounded by

$$\frac{\log \frac{\text{tol}}{\|x-x^0\|}}{\log \rho} n^2 = O(n^2) \ll O(n^3)$$

for large n.

Upshot For large systems, iterative schemes can be more efficient than direct methods.

Conclusion Iterative solvers can be beneficial for *large, ill-conditioned systems*. They are particularly attractive, if only approximate solutions up to a prescribed tolerance are of interest anyway. This is the case for large, ill-conditioned linear systems that typically arise from the discretization of partial differential equations.
The complexity of the evaluation of the iteration function G together with the robustness of the convergence rate with respect to the condition number $\kappa(A)$ and the size n of the coefficient matrix A is crucial for the quality of an iterative scheme.

Modell problem (M)

As a model problem we consider the heat equation in d space dimensions ($d = 1, 2, 3$). Find $u(\cdot,\cdot) : \Omega \times [0,T] \to \mathbb{R}$ such that

$$u_t = \sum_{i=1}^{d} u_{x_i x_i} + f \qquad t > 0,\ x \in \Omega$$

where

computational domain:	$\Omega = (0,1)^d$	
boundary conditions:	$u(x,t) = 0$	$x \in \partial\Omega,\ t > 0$
initial conditions:	$u(x,0) = u_0(x)$	$x \in \Omega.$

Let $f : \Omega \to \mathbb{R}$ be independent of t. Then a stationary distribution of temperature u is obtained if $u_t = 0$, i.e.,

$$\begin{aligned} -\Delta u := -\sum_{i=1}^{d} u_{x_i x_i} &= f \qquad x \in \Omega \\ u(x,0) &= 0 \qquad x \in \partial\Omega. \end{aligned}$$

This is an elliptic boundary value problem.
Discretization ($d = 2$):
Choose a mesh size $h := \frac{1}{n+1}$ with $n \in \mathbb{N}$ fixed and the associated grid

$$\Omega_h := \left\{(x_i, y_j) \in \bar{\Omega} \mid x_i = ih;\ y_j = jh;\ 0 \le i, j \le n+1\right\}$$

A finite difference approximation of Δu can be obtained by setting

$$\begin{aligned} u_{xx}(x_i, y_j) &\approx \frac{1}{h^2}\left(u(x_{i-1}, y_j) - 2u(x_i, y_j) + u(x_{i+1}, y_j)\right) \\ \Delta u(x_i, y_j) &\approx \Delta_h u(x_i, y_j) \\ &= \frac{1}{h^2}\left(u(x_{i-1}, y_j) + u(x_{i+1}, y_j) - 4u(x_i, y_j) + u(x_i, y_{j-1}) + u(x_i, y_{j+1})\right) \end{aligned}$$

and consequently a finite difference approximation of the PDE: $U_{ij} \approx u(x_i, y_j)$

$$-\Delta_h U_{ij} = f_{ij} := f(x_i, y_j)$$

$$U_{0j} = U_{n+1j} = U_{i0} = U_{in+1} = 0$$

This can be written in matrix form $Ax = b$ by line-wise ordering:

$$U = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix} \quad U_i = \begin{pmatrix} U_{i1} \\ \vdots \\ U_{in} \end{pmatrix} \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \quad b_i = h^2 \begin{pmatrix} f_{i1} \\ \vdots \\ f_{in} \end{pmatrix}$$

$$A = \frac{1}{h^2}\begin{pmatrix} A_n & I_n & & 0 \\ I_n & A_n & \ddots & \\ & \ddots & \ddots & I_n \\ 0 & & I_n & A_n \end{pmatrix}$$

$$A_n = \begin{pmatrix} 4 & -1 & & 0 \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 4 \end{pmatrix} \quad I_n = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$$

The eigenvalues of A are

$$\lambda_{ij} = \frac{1}{h^2} 4\left(\sin^2\left(i\frac{\pi}{2}h\right) + \sin^2\left(j\frac{\pi}{2}h\right)\right) \qquad i, j = 1, \ldots, n.$$

The eigenvectors of A are

$$(e_{ij})_{lk} = \sin(i\pi lh) \cdot \sin(j\pi kh) \qquad i,j,k,l = 1,\dots,n.$$

The condition number $\kappa(A)$ is

$$\begin{aligned}\kappa(A) &= \|A\|_2 \|A^{-1}\|_2 \\ &= \lambda_{\max}(A) \cdot (\lambda_{\min}(A))^{-1} \\ &= \frac{2\sin^2\left(\frac{\pi}{2} \cdot \frac{n}{n+1}\right)}{2\sin^2\left(\frac{\pi}{2} \cdot \frac{1}{n+1}\right)} \\ &\approx \frac{1}{\left(\frac{\pi}{2}\right)^2 h^2} \\ &= \left(\frac{2(n+1)}{\pi}\right)^2 \overset{n\to\infty}{\longrightarrow} \infty.\end{aligned}$$

Remark A has a tridiagonal block structure.
A is *sparse*, i.e., the numbers of non-zero coefficients in each row is bounded independently of n.

Remark For a sparse matrix A the complexity of computing Ax is $O(n)$. Hence, we try to achieve linear complexity for each iteration step.
Direct solvers also try to reduce the complexity by exploiting sparsity (direct sparse solvers).

4.2 Linear Iterative Schemes

We consider

$$Ax = b, \qquad A \in \mathbb{R}^{n\times n} \text{ regular, sparse, } b \in \mathbb{R}^n. \tag{4.1}$$

Basic idea Replace (4.1) by a sequence of "simpler" systems

$$Bw = r, \qquad B \in \mathbb{R}^{n\times n} \text{ regular, } r \in \mathbb{R}^n \tag{4.2}$$

which can be solved with complexity $O(n)$. Examples are sparse diagonal or triangular matrices.
This idea can be written in fixed point formulation

$$Bx = Bx + b - Ax, \qquad B \in \mathbb{R}^{n\times n} \text{ arbitrary.}$$

Fixed point iteration:

$$\begin{aligned}Bx^{k+1} &= Bx^k + \underbrace{b - Ax^k}_{\text{residual of } x^k}, \qquad B \in \mathbb{R}^{n\times n} \text{ regular} \\ x^{k+1} &= (I - B^{-1}A)x^k + B^{-1}b \\ &= Gx^k + B^{-1}b.\end{aligned} \tag{4.3}$$

B is called the *preconditioner* and $G = I - B^{-1}A$ the *iteration matrix*.

Remark

$$B = A \;\Rightarrow\; x^1 = x \quad \forall x^0 \in \mathbb{R}^n$$

However, $B^{-1} = A^{-1}$ in (4.3) is usually difficult to compute.

Theorem 4.2.1 *Assume that there are consistent norms $\|\cdot\|$ of $\mathbb{R}^n$ and $\mathbb{R}^{n\times n}$ such that*

$$\|I - B^{-1}A\| = \rho < 1.$$

Then the fixed point iteration

$$x^{k+1} = (I - B^{-1}A)x^k + B^{-1}b$$

converges globally, i.e., for each initial value $x^0 \in \mathbb{R}^n$

$$\|x - x^{k+1}\| \le \rho\|x - x^k\|$$

with convergence rate ρ.

Proof Apply Banach's fixed point theorem to

$$\begin{aligned} T &: \mathbb{R}^n \to \mathbb{R}^n \\ Tx &= (I - B^{-1}A)x + B^{-1}b. \end{aligned}$$

We have to show contractivity:

$$\|Tx - Ty\| = \|(I - B^{-1}A)(x - y)\| \le \rho\|x - y\|$$

Hence, T has a fixed point. □

Example Decompose $A = L + D + R$ where L is lower triangular, D is diagonal, and R is upper triangular. By different choices of B we obtain different classical iterative methods.

1. $B = I$: Richardson iteration
2. $B = D$: Jacobi iteration
3. $B = D + L$: Gauß-Seidel iteration

Proposition 4.2.2 (strong row criterion) *Assume that A is strongly diagonal dominant, i.e.,*

$$\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| < |a_{ii}| \qquad \forall i = 1, \ldots, n.$$

Then the Jacobi iteration is globally convergent.

Proof Exercise

Proposition 4.2.3 *For the model problem (M) the convergence rates ρ_J and ρ_{GS} of Jacobi and Gauß-Seidel iteration, respectively, satisfy*

$$\rho_J^2 = \rho_{GS} < 1.$$

Proof We will show later (Theorem 4.4.4) that $\rho_{GS} < 1$. The remainder follows from [13, Corollary 8.3.16] (weak row criterion,...). □

4.3 Preconditioning and Linear Iterations

We concentrate on linear systems

$$Ax = b$$

with $A \in \mathbb{R}^{n\times n}$ symmetric and positive definite (s.p.d.), $b \in \mathbb{R}^n$ and $\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \gg 1$.

Definition 4.3.1 *A matrix $B \in \mathbb{R}^{n\times n}$ with the properties*

- *B s.p.d.*
- *$Bw = r$ "easily" solvable (with complexity $O(n)$)*
- *$\kappa(B^{-1}A) \ll \kappa(A)$*

is called preconditioner. *B is called* (quasi)-optimal *if $\kappa(B^{-1}A)$ is independent of n.*

Remark $\kappa(B^{-1}A) = \kappa(AB^{-1})$ because $B(B^{-1}A)B^{-1} = AB^{-1}$ is an equivalence transformation of $B^{-1}A$.

Lemma 4.3.2 *Let $C \in \mathbb{R}^{n\times n}$ be s.p.d. and let $\langle\cdot,\cdot\rangle$ denote the Euclidean scalar product in $\mathbb{R}^n$.*
Then

$$\langle x, y\rangle_C := \langle Cx, y\rangle \qquad x, y \in \mathbb{R}^n$$

is a scalar product on $\mathbb{R}^n$.

Lemma 4.3.3 *Let $A \in \mathbb{R}^{n\times n}$ be symmetric.*

1. *There is an orthonormal basis of eigenvectors e_i of A, i.e.,*

$$Ae_i = \lambda_i e_i, \qquad \langle e_i, e_j\rangle = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}.$$

2. *Let A be s.p.d. Then there is an s.p.d. matrix $A^{\frac{1}{2}} \in \mathbb{R}^{n\times n}$ satisfying*

$$A^{\frac{1}{2}}A^{\frac{1}{2}} = A.$$

Proof We only show 2:
Let $T := (e_1, e_2, \ldots, e_n)$ columnwise. Then $T^T = T^{-1}$. Let D be diagonal with

$$D = T^{-1}AT, \qquad D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

Since A is positive definite, $\lambda_i > 0$ for $i = 1, \ldots, n$. Let

$$D^{\frac{1}{2}} := \operatorname{diag}(\lambda^{\frac{1}{2}}, \ldots, \lambda^{\frac{1}{2}})$$

and define

$$A^{\frac{1}{2}} := TD^{\frac{1}{2}}T^{-1}.$$

Then $A^{\frac{1}{2}}$ is symmetric because

$$\begin{aligned}\langle A^{\frac{1}{2}}x, y\rangle &= \langle TD^{\frac{1}{2}}T^{-1}x, y\rangle \\ &= \langle D^{\frac{1}{2}}T^{-1}x, T^{-1}y\rangle \\ &= \langle T^{-1}x, D^{\frac{1}{2}}T^{-1}y\rangle \\ &= \langle x, TD^{\frac{1}{2}}T^{-1}y\rangle \\ &= \langle x, A^{\frac{1}{2}}y\rangle.\end{aligned}$$

$A^{\frac{1}{2}}$ is positive definite because

$$\langle A^{\frac{1}{2}}x, x\rangle = \langle D^{\frac{1}{2}}T^{-1}x, T^{-1}x\rangle = \langle D^{\frac{1}{2}}y, y\rangle \geq 0.$$

$A^{\frac{1}{2}}A^{\frac{1}{2}} = A$ because

$$A^{\frac{1}{2}}A^{\frac{1}{2}} = TD^{\frac{1}{2}}T^{-1}TD^{\frac{1}{2}}T^{-1} = TDT^{-1} = A.$$

□

Lemma 4.3.4 *Let $C \in \mathbb{R}^{n\times n}$ be s.p.d. and let $A \in \mathbb{R}^{n\times n}$ be symmetric with respect to $\langle\cdot,\cdot\rangle_C$, i.e.,*

$$\langle Ax, y\rangle_C = \langle x, Ay\rangle_C \qquad \forall x, y \in \mathbb{R}^n.$$

Then

$$\lambda_{\min}(A) = \min_{\substack{x\in\mathbb{R}^n \\ x\neq 0}} \frac{\langle Ax, x\rangle_C}{\langle x, x\rangle_C} \leq \max_{\substack{x\in\mathbb{R}^n \\ x\neq 0}} \frac{\langle Ax, x\rangle_C}{\langle x, x\rangle_C} = \lambda_{\max}(A).$$

The term

$$\frac{\langle Ax, x\rangle_C}{\langle x, x\rangle_C}$$

is called Rayleigh quotient *of A with respect to $\langle\cdot,\cdot\rangle_C$.*

Proof 1. Let $C = I$. Let e_i denote the orthonormal eigenvectors and λ_i the eigenvalues of A.
For $x \in \mathbb{R}^n$ arbitrary,

$$x = \sum_{i=1}^{n} x_i e_i, \qquad x_i = \langle x, e_i\rangle$$

holds. Hence, if $x \neq 0$,

$$\frac{\langle Ax, x\rangle}{\langle x, x\rangle} = \frac{\sum_{i,j=1}^{n} \lambda_i x_i x_j \langle e_i, e_j\rangle}{\sum_{i,j=1}^{n} x_i x_j \langle e_i, e_j\rangle} = \frac{\sum_{i=1}^{n} \lambda_i x_i^2}{\sum_{i=1}^{n} x_i^2}.$$

This leads to

$$\lambda_{\min}(A) \leq \frac{\langle Ax, x\rangle}{\langle x, x\rangle} \leq \lambda_{\max}(A).$$

Insert $x = e_{\min}$ to obtain

$$\frac{\langle Ae_{\min}, e_{\min}\rangle}{\langle e_{\min}, e_{\min}\rangle} = \lambda_{\min}(A)$$

and $x = e_{\max}$ to conclude the proof.

2. Let $C \in \mathbb{R}^{n\times n}$ be s.p.d. and $C^{\frac{1}{2}} \in \mathbb{R}^{n\times n}$ s.p.d. with $C^{\frac{1}{2}}C^{\frac{1}{2}} = C$ according to Lemma 4.3.3. Then

$$\begin{aligned}\langle C^{\frac{1}{2}}AC^{-\frac{1}{2}}x, y\rangle &= \langle C^{-\frac{1}{2}}CAC^{-\frac{1}{2}}x, y\rangle \\ &= \langle CAC^{-\frac{1}{2}}x, C^{-\frac{1}{2}}y\rangle \\ &= \langle AC^{-\frac{1}{2}}x, C^{-\frac{1}{2}}y\rangle_C \\ &= \langle C^{-\frac{1}{2}}x, AC^{-\frac{1}{2}}y\rangle_C \\ &= \langle C^{\frac{1}{2}}x, AC^{-\frac{1}{2}}y\rangle \\ &= \langle x, C^{\frac{1}{2}}AC^{-\frac{1}{2}}y\rangle.\end{aligned}$$

This means $C^{\frac{1}{2}}AC^{-\frac{1}{2}}$ is symmetric.
Since $C^{\frac{1}{2}}AC^{-\frac{1}{2}}$ is an equivalence transformation of A, we get

$$\lambda(A) = \lambda(C^{\frac{1}{2}}AC^{-\frac{1}{2}}).$$

Furthermore,

$$\frac{\langle Ax, x\rangle_C}{\langle x, x\rangle_C} = \frac{\langle C^{\frac{1}{2}}AC^{-\frac{1}{2}}C^{\frac{1}{2}}x, C^{\frac{1}{2}}x\rangle}{\langle C^{\frac{1}{2}}x, C^{\frac{1}{2}}x\rangle} = \frac{\langle C^{\frac{1}{2}}AC^{-\frac{1}{2}}y, y\rangle}{\langle y, y\rangle}.$$

Hence, 1. provides

$$\begin{aligned}\lambda_{\min}(A) &= \lambda_{\min}(C^{\frac{1}{2}}AC^{-\frac{1}{2}}) \\ &= \min_{\substack{y\in\mathbb{R}^n \\ y\neq 0}} \frac{\langle C^{\frac{1}{2}}AC^{-\frac{1}{2}}y, y\rangle}{\langle y, y\rangle} \\ &= \min_{\substack{x\in\mathbb{R}^n \\ x\neq 0}} \frac{\langle Ax, x\rangle_C}{\langle x, x\rangle_C}.\end{aligned}$$

The analogue for $\lambda_{\max}(A)$ concludes the proof.

□

Corollary 4.3.5 *Let $B \in \mathbb{R}^{n\times n}$ be a preconditioner satisfying*

$$\mu_0\langle Bx, x\rangle \leq \langle Ax, x\rangle \leq \mu_1\langle Bx, x\rangle \qquad \forall x \in \mathbb{R}^n$$

for some $0 \leq \mu_0, \mu_1 \in \mathbb{R}$. Then $\kappa(B^{-1}A) \leq \frac{\mu_1}{\mu_0}$.

Proof Exercise

Remark The linear iteration

$$Bx^{k+1} = Bx^k + b - Ax^k$$

is equivalent to the Richardson iteration applied to the preconditioned system

$$B^{-1}Ax = B^{-1}b.$$

Does a good linear iteration provide a good preconditioner?

Proposition 4.3.6 *Let $B \in \mathbb{R}^{n\times n}$ be s.p.d.. We assume that the iteration matrix*

$$G = I - B^{-1}A$$

of the associated linear iteration satisfies

$$\|G\|_2 = \|I - B^{-1}A\|_2 = \max|\lambda(G)| = \rho < 1.$$

Then

$$\kappa(B^{-1}A) \leq \frac{1+\rho}{1-\rho}.$$

Proof $G = I - B^{-1}A$ is symmetric with respect to $\langle\cdot,\cdot\rangle_B$. Lemma 4.3.4 provides

$$\frac{|\langle Gx, x\rangle_B|}{\langle x, x\rangle_B} \leq \max|\lambda(G)| = \rho, \qquad \forall x \in \mathbb{R}^n.$$

As

$$\begin{aligned}\frac{|\langle Gx, x\rangle_B|}{\langle x, x\rangle_B} &= \frac{\left|\langle x, x\rangle_B - \langle B^{-1}Ax, x\rangle_B\right|}{\langle x, x\rangle_B} \\ &= \left|1 - \frac{\langle Ax, x\rangle}{\langle Bx, x\rangle}\right|,\end{aligned}$$

this is equivalent to

$$-\rho \leq 1 - \frac{\langle Ax, x\rangle}{\langle Bx, x\rangle} \leq \rho.$$

Rearranging terms we get

$$(1-\rho)\langle Bx, x\rangle \leq \langle Ax, x\rangle \leq (1+\rho)\langle Bx, x\rangle.$$

The assertion follows from Corollary 4.3.5. □

Does a good preconditioner provide a good linear iteration?

Proposition 4.3.7 *Let $B \in \mathbb{R}^{n\times n}$ be a preconditioner for A. Then the* damped *linear iteration*

$$Bx^{k+1} = Bx^k + \omega(b - Ax^k)$$

converges for each damping parameter $\omega \in \left(0, \frac{2}{\lambda_{\max}(B^{-1}A)}\right)$.
The optimal damping parameter is

$$\omega_{opt} = 2\left(\lambda_{\min}(B^{-1}A) + \lambda_{\max}(B^{-1}A)\right)^{-1}$$

with the associated convergence rate

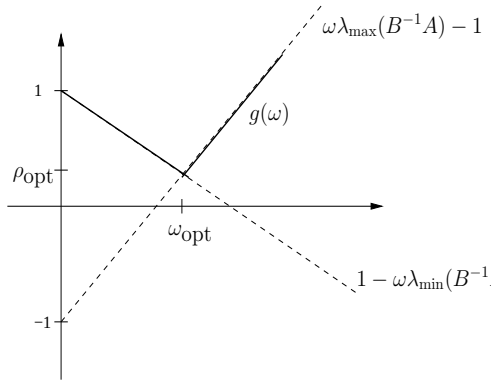
$$\rho_{opt} = \frac{\kappa(B^{-1}A) - 1}{\kappa(B^{-1}A) + 1}.$$

Proof By Lemma 4.3.4 we get

$$\begin{aligned} g(\omega) &:= \max\left|\lambda(I-\omega B^{-1}A)\right| \\ &= \max_{x\neq 0}\frac{\left|\langle x,x\rangle_B - \omega\langle B^{-1}Ax,x\rangle_B\right|}{\langle x,x\rangle_B} \\ &= \max_{x\neq 0}\left|1-\omega\frac{\langle Ax,x\rangle}{\langle Bx,x\rangle}\right| \\ &= \max\left\{1-\omega\min_{x\neq 0}\frac{\langle Ax,x\rangle}{\langle Bx,x\rangle}, \omega\max_{x\neq 0}\frac{\langle Ax,x\rangle}{\langle Bx,x\rangle}-1\right\}. \end{aligned}$$

Arguing by Lemma 4.3.4 we get

$$\begin{aligned} 0<\lambda_{\min}(B^{-1}A) &= \min_{x\neq 0}\frac{\langle B^{-1}Ax,x\rangle_B}{\langle x,x\rangle_B} = \min_{x\neq 0}\frac{\langle Ax,x\rangle}{\langle Bx,x\rangle} \\ 0<\lambda_{\max}(B^{-1}A) &= \max_{x\neq 0}\frac{\langle B^{-1}Ax,x\rangle_B}{\langle x,x\rangle_B} = \max_{x\neq 0}\frac{\langle Ax,x\rangle}{\langle Bx,x\rangle} \end{aligned}$$



$$g(\omega)<1 \Leftrightarrow \omega>0 \wedge \omega<\frac{2}{\lambda_{\max}(B^{-1}A)}$$

ω_{opt}:

$$\omega_{\text{opt}}\lambda_{\max}-1 = 1-\omega_{\text{opt}}\lambda_{\min} \Rightarrow \omega_{\text{opt}} = \frac{2}{\lambda_{\min}+\lambda_{\max}}$$

ρ_{opt}: evaluate $g(\omega_{\text{opt}})$

□

Example $A = L + D + R$

1. no preconditioner: $B = I$
2. Jacobi preconditioner: $B = D$
3. symmetric Gauß-Seidel preconditioner:
   1 step Gauß-Seidel $G_{GS} = I-(D+L)^{-1}(D+L+R) = -(D+L)^{-1}R$
   1 step Gauß-Seidel in reversed order $G_{GS}^{-} = I-(D+R)^{-1}(D+L+R) = -(D+R)^{-1}L$
   Hence,

$$\begin{aligned} G_{\text{sym}} &= G_{GS}^{-}G_{GS} \\ &= (D+R)^{-1}L(D+L)^{-1}R \\ &= (D+R)^{-1}(D+L-D)(D+L)^{-1}R \\ &= (D+R)^{-1}R-(D+R)^{-1}D(D+L)^{-1}R \\ &= (D+R)^{-1}R-(D+R)^{-1}D(D+L)^{-1}(A-(D+L)) \\ &= (D+R)^{-1}R-(D+R)^{-1}D(D+L)^{-1}A+(D+R)^{-1}D \\ &= I-(D+R)^{-1}D(D+L)^{-1}A \end{aligned}$$

$$\Rightarrow B_{\text{sym}} = (D+L)D^{-1}(D+R).$$

4. Modell Problem (M): $B = \text{blockdiag}(A_n, \ldots, A_n)$

4.4 Linear Descent Methods

We consider

$$Ax = b, \qquad A \in \mathbb{R}^{n\times n} \text{ s.p.d., } b \in \mathbb{R}^n \tag{4.4}$$

Lemma 4.4.1 *Problem (4.4) is equivalent to the minimization problem*

$$x \in \mathbb{R}^n : \ J(x) \le J(y) \quad \forall y \in \mathbb{R}^n,$$

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

Proof Gradient: $J'(v)(w) = \langle Av, w\rangle - \langle b, w\rangle$

$$\begin{aligned} J'(v) = 0 &\Leftrightarrow \langle Av - b, w\rangle = 0 \ \forall w \in \mathbb{R}^n \\ &\Leftrightarrow Av - b = 0 \end{aligned}$$

Hence,

$$Ax = b \ \Leftrightarrow \ x \text{ is local extremum of } J.$$

Hessian matrix: $J''(v) = A$

$$A \text{ positive definite } \Leftrightarrow \text{ local minimum in } x.$$

□

Basic idea of descent methods for linear systems:
Approximate x by minimization of J in the direction of certain descent directions e_i, $i = 1, \ldots, n$.

For given descent directions $e_i \neq 0$, $i = 1, \ldots, n$, and given iterate x^k we consider the following two algorithms:

Algorithm 1 Parallel directional correction (PDC)

for $i = 1, \ldots, n$ **do**
solve

$$\alpha_i \in \mathbb{R} : J(x^k + \alpha_i e_i) \le J(x^k + \alpha e_i) \ \forall \alpha \in \mathbb{R}$$

end for
new iterate: $x^{k+1} = x^k + \sum_{i=1}^n \alpha_i e_i$

Remark 1. SDC implies $J(v^k) \le J(v^{k-1}) \quad \forall k$

2. PDC does not imply the above but it is easier to parallelize.

Algorithm 2 Successive directional correction (SDC)

initialize $w_0 := x^k$
for $i = 1, \dots, n$ **do**
solve

$$\alpha_i \in \mathbb{R} : J(w_{i-1} + \alpha_i e_i) \leq J(w_{i-1} + \alpha e_i) \ \forall \alpha \in \mathbb{R}$$

update: $w_i := w_{i-1} + \alpha_i e_i$
end for
new iterate: $x^{k+1} = w_n$

3. The solution of the local 1D-minimization problem

$$\alpha_0 \in \mathbb{R} : \ J(w + \alpha_0 e) \leq J(w + \alpha e) \quad \forall \alpha \in \mathbb{R}$$

is available in closed form as $\alpha_0 = \frac{\langle b - Aw, e \rangle}{\langle Ae, e \rangle}$.

Proof of 3:

$$\begin{aligned}
g(\alpha) &:= J(w + \alpha e) \\
&= \frac{1}{2}\langle A(w + \alpha e), w + \alpha e \rangle - \langle b, w + \alpha e \rangle \\
&= \frac{1}{2}\alpha^2 \langle Ae, e \rangle + \frac{1}{2}\alpha \langle Aw, e \rangle + \frac{1}{2}\alpha \langle Ae, w \rangle + \frac{1}{2}\langle Aw, w \rangle - \langle b, w \rangle - \alpha \langle b, e \rangle \\
&= \frac{1}{2}\alpha^2 \langle Ae, e \rangle + \alpha \langle Aw, e \rangle - \alpha \langle b, e \rangle + \langle \frac{1}{2} Aw - b, w \rangle \\
g'(\alpha) &= \alpha \langle Ae, e \rangle + \langle Aw, e \rangle - \langle b, e \rangle \\
g'(\alpha_0) = 0 \ &\Leftrightarrow \ \alpha_0 = \frac{\langle b - Aw, e \rangle}{\langle Ae, e \rangle} \\
g''(\alpha) &= \langle Ae, e \rangle > 0
\end{aligned}$$

□

Proposition 4.4.2 *Assume that the e_i are linearly independent. Then the PDC and SDC methods are linear iterations.*

Proof We only consider PDC methods. Note that

$$\langle b - Aw, e_i \rangle = e_i^T (b - Aw).$$

Hence,

$$\begin{aligned}
x^{k+1} &= x^k + \sum_{i=1}^{n} \alpha_i e_i \\
&= x^k + \sum_{i=1}^{n} \frac{1}{\langle Ae_i, e_i \rangle} (e_i e_i^T)(b - Aw) \\
&= x^k + C(b - Aw)
\end{aligned}$$

with $C = \sum_{i=1}^{n} \frac{1}{\langle Ae_i, e_i\rangle}(e_i e_i^T)$. We show that C is invertible.
Assume

$$0 = Cv = \sum_{i=1}^{n} \frac{1}{\langle Ae_i, e_i\rangle} e_i (e_i^T v) = \sum_{i=1}^{n} \frac{e_i^T v}{\langle Ae_i, e_i\rangle} e_i$$

$$\begin{aligned} e_i \text{ lin. indep.} &\Rightarrow e_i^T v = 0 \quad \forall i = 1, \ldots, n \\ &\Rightarrow \langle v, w\rangle = 0 \quad \forall w \in \mathbb{R}^n \\ &\Rightarrow v = 0 \end{aligned}$$

Hence, PDC is a linear iteration with $B = C^{-1}$. □

Proposition 4.4.3 *Let e_i denote the Cartesian unit vectors. Then the resulting PDC and SDC methods are equivalent to the Jacobi method and the Gauß-Seidel method, respectively.*

Proof Exercise

Theorem 4.4.4 *The Gauß-Seidel iteration converges globally for all s.p.d. matrices A.*

Proof Let $x^0 \in \mathbb{R}^n$ and let $x^{k+1} = Mx^k$ describe one Gauß-Seidel step.
Recall

$$J(x^k) \leq J(x^0)$$

for all $k \geq 0$.

1. To show:

$$\exists C > 0 : \ \|x^k\| \leq C \ \forall k \in \mathbb{N}$$

$$\begin{aligned} J(x^0) &\geq J(x^k) \\ &= \frac{1}{2}\langle Ax^k, x^k\rangle - \langle b, x^k\rangle \\ &\geq c\|x^k\|_2^2 - \|b\|\|x^k\| \\ &=: g(\|x^k\|) \end{aligned}$$

   g is a parabola and $g(\|x^k\|)$ is bounded by $J(x^0)$. Hence, the arguments $\|x^k\|$ must also be bounded.

2. There exists an $x^* \in \mathbb{R}^n$, and a subsequence $\left(x^{k_j}\right)_{j\in\mathbb{N}}$: $x^{k_j} \to x^*$ for $j \to \infty$ (Heine-Borel).

3. $M : \mathbb{R}^n \to \mathbb{R}^n$ is continuous. This follows directly from Proposition 4.4.2.

4. Next we show that $J(x^*) = J(Mx^*)$. For this note that

$$J(x^{k_{j+1}}) \leq J(x^{k_j+1}) = J(Mx^{k_j}) \leq J(x^{k_j}).$$

   Let $j \to \infty$ and use continuity of J and M to obtain

$$J(x^*) \leq J(Mx^*) \leq J(x^*).$$

5. To show: $Ax = b \Leftrightarrow x^* = x$

$$J(x^*) = J(Mx^*) \Rightarrow 0 = \alpha_i = \frac{\langle b - Ax^*, e_i\rangle}{\langle Ae_i, e_i\rangle} \qquad i = 1, \ldots, n$$
$$\Rightarrow b - Ax^* = 0$$

6. To show: $x^k \to x$
   As x is the unique solution each convergent subsequence must converge to x.

□

Remark The Gauß-Seidel method is linearly convergent [6].

How can "better" directions e_i be constructed? A hint comes from the following result.

Proposition 4.4.5 *Assume that the search directions e_i are A-orthogonal, i.e.,*

$$\langle Ae_i, e_j\rangle = 0, \qquad i \neq j.$$

Then the corresponding PDC and SDC methods provide the exact solution in one step.

Proof (only for PDC method)
Let $x^0 \in \mathbb{R}^n$,

$$x^0 = \sum_{j=1}^{n} x_j^0 e_j,$$

and

$$x = \sum_{j=1}^{n} x_j e_j$$

be the exact solution. Then

$$\begin{aligned}\alpha_i &= \frac{\langle b - Ax^0, e_i\rangle}{\langle Ae_i, e_i\rangle} \\ &= \frac{\langle A(x - x^0), e_i\rangle}{\langle Ae_i, e_i\rangle} \\ &= \sum_{j=1}^{n} (x_j - x_j^0) \frac{\langle Ae_j, e_i\rangle}{\langle Ae_i, e_i\rangle} \\ &= x_i - x_i^0.\end{aligned}$$

Hence,

$$x^1 = x^0 + \sum_{i=1}^{n} \alpha_i e_i = x^0 + \sum_{i=1}^{n} (x_i - x_i^0) e_i = \sum_{i=1}^{n} x_i e_i = x.$$

□

Remark The eigenvectors e_i^* of A are A-orthogonal

$$\langle Ae_i^*, e_j^*\rangle = \lambda_i \langle e_i^*, e_j^*\rangle = 0 \qquad i \neq j.$$

This leads us to the basic idea of multigrid methods:
Select search directions e_i such that $e_i \approx e_i^*$.
How? See Numerics III.

4.5 Nonlinear Descent Methods

We consider

$$Ax^* = b, \qquad A \in \mathbb{R}^{n\times n} \text{ s.p.d., } b \in \mathbb{R}^n.$$

Then

$$J(x^*) \le J(x) \quad \forall x \in \mathbb{R}^n$$

Remark $J(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$ is called energy.
$\langle x, y\rangle_A = \langle Ax, y\rangle$ is the energy scalar product,
$\|x\|_A = \sqrt{\langle x, x\rangle_A}$ is the energy norm.

Basic idea of *nonlinear descent methods*:
Select a new descent direction in each iteration step.

4.5.1 Gradient Methods (Steepest Descent)

The direction of steepest descent of J at x is

$$-\nabla J(x) = b - Ax \qquad \text{(residual of } x\text{)}.$$

The directional derivative of J at x in the direction n, $\|n\| = 1$, is

$$\begin{aligned}\frac{\partial}{\partial n}J(x) &= \langle \nabla J(x), n\rangle \\ &\ge -\|\nabla J(x)\| \cdot \|n\| \qquad \text{(Cauchy-Schwarz inequality)} \\ &= \langle \nabla J(x), -\frac{\nabla J(x)}{\|\nabla J(x)\|}\rangle\end{aligned}$$

Algorithm 3 Gradient method (analytic formulation)

1: given iterate $x^k \in \mathbb{R}^n$, residual $r_k = b - Ax^k$
2: optimal reduction of energy in steepest descent direction

$$\alpha_k \in \mathbb{R}: \; J(x^k + \alpha_k r_k) \le J(x^k + \alpha r_k) \qquad \forall \alpha \in \mathbb{R}$$

3: new iterate $x^{k+1} = x^k + \alpha_k r_k$

Remark

$$\alpha_k = \frac{\langle r_k, r_k\rangle}{\langle Ar_k, r_k\rangle}$$

Proof

$$\begin{aligned}g(\alpha) &= J(x^k + \alpha r_k) \\ g'(\alpha) &= \langle \nabla J(x^k + \alpha r_k), r_k\rangle = \langle A(x^k + \alpha r_k) - b, r_k\rangle = \langle -r_k + \alpha A r_k, r_k\rangle = 0 \\ &\Leftrightarrow \; \alpha = \frac{\langle r_k, r_k\rangle}{\langle Ar_k, r_k\rangle}\end{aligned}$$

□

Remark Reduction of the energy is equivalent to reduction of the error, i.e.,

$$\alpha_k \in \mathbb{R}: \ \|x^* - x^{k+1}\|_A = \|x^* - (x^k + \alpha_k r_k)\|_A \leq \|x^* - (x^k + \alpha r_k)\|_A \quad \forall \alpha \in \mathbb{R}$$

Proof

$$\begin{aligned}\|x^* - x\|_A^2 &= \langle x^* - x, x^* - x\rangle_A \\ &= \langle x, x\rangle_A - 2\langle x^*, x\rangle_A + \langle x^*, x^*\rangle_A \\ &= \langle Ax, x\rangle - 2\langle b, x\rangle + \langle b, x^*\rangle \\ &= 2J(x) + \langle b, x^*\rangle\end{aligned}$$

Hence,

$$\begin{aligned}\|x^* - x^{k+1}\|_A^2 &= 2J(x^{k+1}) + \langle b, x^*\rangle \\ &\leq 2J(x^k + \alpha r_k) + \langle b, x^*\rangle \\ &= \|x^* - (x^k + \alpha r_k)\|_A \quad \forall \alpha \in \mathbb{R}.\end{aligned}$$

□

Algorithm 4 Gradient method (algorithmic formulation)

1: given iterate $x^k \in \mathbb{R}^n$, residual $r_k = b - Ax^k$
2: optimal reduction factor:

$$\alpha_k = \frac{\langle r_k, r_k\rangle}{\langle Ar_k, r_k\rangle}$$

3: new iterate $x^{k+1} = x^k + \alpha_k r_k$
4: stopping criterion
if $\|x^* - x^{k+1}\|_A \leq$ tol **then**
stop
else
$k := k + 1$ and continue with step 1
end if

Remark The computational effort consists of two matrix-vector multiplications and two scalar products.

Proposition 4.5.1 (Convergence rates) *Let A be s.p.d. with eigenvalues $0 < \lambda_1 \leq \ldots \leq \lambda_n$. Let $x^0 \in \mathbb{R}^n$ and let $x^1, \ldots, x^k$ be computed by the gradient method. Then*

$$\|x^* - x^{k+1}\|_A \leq \left(1 - \kappa(A)^{-2}\right)^{\frac{1}{2}} \|x^* - x^k\|_A,$$

where $\kappa(A) = \frac{\lambda_n}{\lambda_1}$ denotes the condition number of A.

Proof 1. Diagonalization of A:

$$A = QDQ^T, \quad Q^{-1} = Q^T, \quad D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \quad A^{\frac{1}{2}} = QD^{\frac{1}{2}}Q^T$$

2. Minimization property of the Rayleigh quotient: $\lambda_1 \leq \frac{\langle Ax,x\rangle_B}{\langle x,x\rangle_B} \leq \lambda_n$

$$\begin{aligned}
\langle r_k, x^* - x^k\rangle_A &= \langle Ar_k, x^* - x^k\rangle \\
&= \langle r_k, A(x^* - x^k)\rangle \\
&= \langle b - Ax^k, A(x^* - x^k)\rangle \\
&= \langle A(x^* - x^k), A(x^* - x^k)\rangle \\
&= \langle AA^{\frac{1}{2}}(x^* - x^k), A^{\frac{1}{2}}(x^* - x^k)\rangle \\
&\geq \lambda_1 \langle A^{\frac{1}{2}}(x^* - x^k), A^{\frac{1}{2}}(x^* - x^k)\rangle \\
&= \lambda_1 \|x^* - x^k\|_A^2 \\
\|r_k\|_A^2 &= \langle Ar_k, r_k\rangle \\
&= \langle AA(x^* - x^k), A(x^* - x^k)\rangle \\
&= \langle A^2 A^{\frac{1}{2}}(x^* - x^k), A^{\frac{1}{2}}(x^* - x^k)\rangle \\
&\leq \lambda_n^2 \|x^* - x^k\|_A^2
\end{aligned}$$

3. Upper bound for the convergence rate: Set $\alpha := \frac{\lambda_1}{\lambda_n^2}$

$$\begin{aligned}
\|x^* - (x^k + \alpha_k r_k)\|_A^2 &\leq \|x^* - (x^k + \alpha r_k)\|_A^2 \\
&= \|x^* - x^k\|_A^2 - 2\alpha\langle r_k, x^* - x^k\rangle + \alpha^2 \|r_k\|_A^2 \\
&\leq \|x^* - x^k\|_A^2 - 2\alpha\lambda_1 \|x^* - x^k\|_A^2 + \alpha^2 \lambda_n^2 \|x^* - x^k\|_A^2 \\
&= \left(1 - 2\frac{\lambda_1^2}{\lambda_n^2} + \frac{\lambda_1^2}{\lambda_n^2}\right) \|x^* - x^k\|_A^2 \\
&= \left(1 - \kappa(A)^{-2}\right) \|x^* - x^k\|_A^2
\end{aligned}$$

□

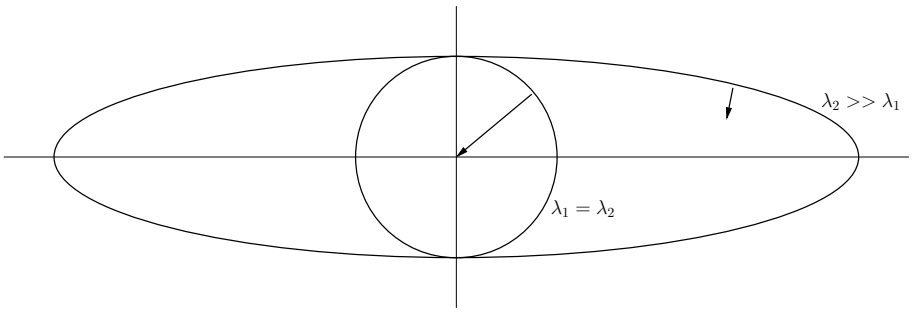
Remark

$$\left(1 - \kappa(A)^{-2}\right)^{\frac{1}{2}} \geq \frac{\kappa(A) - 1}{\kappa(A) + 1}$$

Example

$$n = 2, \qquad A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \qquad b = 0$$

We consider level sets of $J(x) = \frac{1}{2}(\lambda_1 x_1^2 + \lambda_2 x_2^2)$.



slow convergence for $\lambda_1 \ll \lambda_2$, i.e., $\kappa(A) \gg 1$

4.5.2 Conjugate Gradient Methods (CG Methods)

Lemma 4.5.2 *Let* $x^{k+1} = x^k + \alpha_k r_k$, $r_k = b - Ax^k$, $\alpha_k \in \mathbb{R}$. *Then*

$$x^{k+1} \in x^0 + V_k, \qquad V_k = \operatorname{span}\{r_0, Ar_0, \ldots, A^k r_0\}$$

V_k *is called a* Krylov space.

Proof Note that $V_1 \subset V_2 \subset \ldots \subset V_k$, $\quad AV_{k-1} \subset V_k$.

Induction over k:

1. $k = 0$:
$$x^1 = x^0 + \alpha_0 r_0 \in x^0 + V_0, \qquad V_0 = \operatorname{span}\{r_0\}$$

2. Let $k > 0$ and $x^k \in x^0 + v_{k-1}$, $v_{k-1} \in V_{k-1} = \operatorname{span}\{r_0, \ldots, A^{k-1} r_0\}$. Then

$$\begin{aligned} x^{k+1} &= x^k + \alpha_k r_k \\ &= x^k + \alpha_k (b - Ax^k) \\ &= x^0 + v_{k-1} + \alpha_k (b - A(x^0 + v_{k-1})) \\ &= x^0 + \underbrace{v_{k-1}}_{\in V_{k-1}} + \underbrace{\alpha_k r_0}_{\in V_0} - \underbrace{\alpha_k A v_{k-1}}_{\in V_k} \\ &= x^0 + v_k, \qquad v_k \in V_k. \end{aligned}$$

□

Algorithm 5 Conjugate gradient iteration (analytic version)

1: given iterate $x^k \in \mathbb{R}^n$
2: compute $x^{k+1} \in x^0 + V_k$ such that

$$\|x^* - x^{k+1}\|_A \leq \|x^* - x\|_A \qquad \forall x \in x^0 + V_k$$

How do we compute x^{k+1} cheaply?

Reminder (Best approximation) [9, Sec.2] Let H be a Hilbert space, $V_k \subset H$ finite dimensional.
Minimization: For given $f \in H$ find $v_k \in V_k$ such that

$$\|f - v_k\| \leq \|f - v\| \qquad \forall v \in V_k$$

Equivalent variational formulation:

$$\langle v_k, v \rangle = \langle f, v \rangle \qquad \forall v \in V_k$$

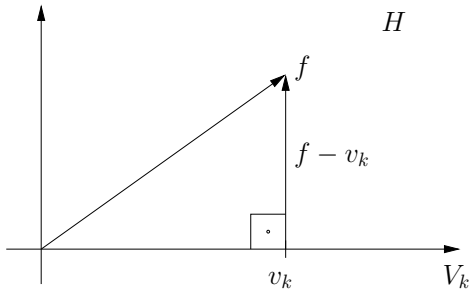


Figure 4.1: Galerkin orthogonality

Orthogonal basis $e_0, \ldots, e_{k-1} \in V_k$, i.e.,

$$v_k = \sum_{i=0}^{k-1} \frac{\langle f, e_i \rangle}{\|e_i\|^2} e_i.$$

Let $w_0, \ldots, w_{k-1}$ be a basis of V_k. We get an orthogonal basis $e_0, \ldots, e_{k-1}$ by Gram-Schmidt orthogonalization:

$$\begin{aligned} e_0 &= w_0 \\ e_{j+1} &= w_{j+1} - \sum_{i=0}^{j} \beta_i e_i, \qquad \beta_i = \frac{\langle w_{j+1}, e_i \rangle}{\langle e_i, e_i \rangle} \end{aligned}$$

Application to CG Let $H = \mathbb{R}^n$ with the energy scalar product $\langle x, x \rangle_A = \langle Ax, x \rangle$, let $V_k = \operatorname{span}\{r_0, \ldots, A^k r_0\}$ denote the kth Krylov space, and let $f := x^* - x^0$. Then for each $k > 0$, the minimization takes the form

$$v_k \in V_k : \|x^* - (x^0 + v_k)\|_A \leq \|x^* - (x^0 + v)\|_A \quad \forall v \in V_k$$

or, equivalently,

$$v_k \in V_k : \langle v_k, v \rangle_A = \langle x^* - x^0, v \rangle_A = \langle r_0, v \rangle \qquad \forall v \in V_k. \tag{4.5}$$

Let $e_0, \ldots, e_k$ be an orthogonal basis of $(\mathbb{R}^{k+1}, \langle \cdot, \cdot \rangle_A)$. Then

$$v_k = \sum_{i=0}^{k} \frac{\langle x^* - x^0, e_i \rangle_A}{\langle e_i, e_i \rangle_A} e_i = \sum_{i=0}^{k} \frac{\langle r_0, e_i \rangle}{\langle Ae_i, e_i \rangle} e_i$$

How do we get e_i?
Let us first exploit the Galerkin orthogonality (4.5).

Lemma 4.5.3 *For given $x^0 \in \mathbb{R}^n$ let $x^1, \ldots, x^{k+1}$ be computed by the CG-method, and let $r_{k+1} \neq 0$.*
Then

(i) $\langle r_{k+1}, v \rangle = 0 \quad \forall v \in V_k$

(ii) $V_k = \text{span}\{r_0, \ldots, r_k\}$

Proof (i) Rewriting (4.5) we get

$$\begin{aligned} 0 &= \langle x^* - (x^0 + v_k), v\rangle_A \\ &= \langle x^* - x^{k+1}, v\rangle_A \\ &= \langle b - Ax^{k+1}, v\rangle \\ &= \langle r_{k+1}, v\rangle \quad \forall v \in V_k \end{aligned}$$

(ii) $r_0, \ldots, r_k$ are linearly independent as a consequence of *(i)* ($\langle r_i, r_j\rangle = 0 \ \forall i > j$) and $r_{k+1} \neq 0 \ \Rightarrow \ r_i \neq 0 \ i \leq k+1$, because $V_i \subset V_k$.
Hence, it is sufficient to show

$$r_i \in V_k, \qquad i = 0, \ldots, k.$$

Let

$$x^i = x^0 + v_{i-1}, \qquad v_{i-1} \in V_{i-1}, \qquad i = 0, \ldots, k.$$

Then

$$\begin{aligned} r_i &= b - Ax^i \\ &= b - A(x^0 + v_{i-1}) \\ &= r_0 - Av_{i-1} \in V_i \subset V_k. \end{aligned}$$

□

As a consequence we can inductively construct an A-orthogonal basis of V_{k+1} by orthogonalizing $r_0, \ldots, r_{k+1}$.

Lemma 4.5.4 *Let* $e_0 := r_0$ *and inductively*

$$e_{k+1} := r_{k+1} - \beta_k e_k, \qquad \beta_k = \frac{\langle Ar_{k+1}, e_k\rangle}{\langle Ae_k, e_k\rangle}, \qquad k \geq 0$$

Then $e_0, \ldots, e_{k+1}$ *is an A-orthogonal basis of* V_{k+1}.

Proof Let $e_0, \ldots, e_k$ be an A-orthogonal basis of V_k. Let $\tilde{e}_{k+1}$ be computed by Gram-Schmidt orthogonalization, i.e.,

$$\tilde{e}_{k+1} = r_{k+1} - \sum_{i=0}^{k} \beta_i e_i, \qquad \beta_i = \frac{\langle Ar_{k+1}, e_i\rangle}{\langle Ae_i, e_i\rangle}.$$

For $i = 0, \ldots, k-1$,

$$\beta_i = \langle Ar_{k+1}, e_i\rangle = \langle r_{k+1}, \underbrace{Ae_i}_{\in V_k}\rangle = 0.$$

Hence,

$$\tilde{e}_{k+1} = r_{k+1} - \beta_k e_k = e_{k+1}.$$

□

Algorithm 6 CG method

initialize

$$r_0 = b - Ax^0 \quad e_0 = r_0, \quad k = 0$$

if $r_k \neq 0$ **then**

compute

$$x^{k+1} = x^k + \alpha_k e_k, \qquad \alpha_k = \frac{\langle r_0, e_k \rangle}{\langle Ae_k, e_k \rangle}$$

$$r_{k+1} = b - Ax^{k+1}$$

$$e_{k+1} = r_{k+1} - \beta_k e_k, \qquad \beta_k = \frac{\langle Ar_{k+1}, e_k \rangle}{\langle Ae_k, e_k \rangle}$$

end if

Theorem 4.5.5 *For given $x^0 \in \mathbb{R}^n$ the iterates x^k of the CG-method can be computed inductively as given by algorithm 6.*

Proof By Lemma 4.5.4 $e_0, \ldots, e_k$ is an A-orthogonal basis of V_k. Hence, the solution of (4.5) is

$$v_{k-1} = \sum_{i=0}^{k-1} \alpha_i e_i$$

and

$$v_k = v_{k-1} + \alpha_k e_k.$$

Hence,

$$x^{k+1} = x^0 + v_k = x^0 + v_{k-1} + \alpha_k e_k = x^k + \alpha_k e_k.$$

The construction of e_{k+1} follows from Lemma 4.5.4. □

Remark One iteration step of algorithm 6 requires

- 3 multiplications with A,
- 3 scalar products,
- storage of r_0.

Lemma 4.5.6

$$\alpha_k = \frac{\langle r_k, r_k \rangle}{\langle Ae_k, e_k \rangle}$$

$$r_{k+1} = r_k - \alpha_k Ae_k$$

$$\beta_k = -\frac{\langle r_{k+1}, r_{k+1} \rangle}{\langle r_k, r_k \rangle}$$

Proof [6]

Remark If the algorithm is reformulated using Lemma 4.5.6, one iteration step requires

- 1 multiplication with A,
- 3 scalar products.

Remark It is guaranteed that $x^n = x^*$. However we hope for $\|x^* - x^k\| \leq \text{tol}$ for $k \ll n$.

Theorem 4.5.7 *The CG method satisfies the error estimate*

$$\|x^* - x^k\|_A \leq 2 \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right)^k \|x^* - x^0\|_A.$$

Proof [6]

Remark By Theorem 4.5.7 the average convergence rate is

$$\rho(A) = \frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \ll \frac{\kappa(A) - 1}{\kappa(A) + 1},$$

i.e., the CG method is much faster than the gradient method. However, we still have

$$\rho(A) \to 1 \text{ for } \kappa(A) \to \infty.$$

The remedy is

1. Find a good preconditioner B.
2. Apply CG to the preconditioned system

$$A_B x_B = b, \qquad A_B = AB^{-1}, \qquad x_B = Bx.$$

Remark A_B is symmetric with respect to $\langle \cdot, \cdot \rangle_{B^{-1}}$.

Replace

$$\begin{aligned} A &\to A_B = AB^{-1} \\ \langle \cdot, \cdot \rangle &\to \langle \cdot, \cdot \rangle_{B^{-1}} \\ x^k &\to x_B^k = Bx^k \end{aligned}$$

in the CG algorithm to obtain the preconditioned CG method (PCG).

Remark In addition to the matrix-vector operations needed by the CG method, each step of the PCG method requires requires 2 evaluations of B^{-1}.

Corollary 4.5.8 *The PCG method satisfies the error estimate*

$$\|x^* - x^k\|_A \leq 2 \left(\frac{\sqrt{\kappa(A_B)} - 1}{\sqrt{\kappa(A_B)} + 1} \right)^k \|x^* - x^0\|_A.$$

Algorithm 7 Preconditioned CG method (PCG)

1: initialization for given $x^0 \in \mathbb{R}^n$:

$$r_0 = b - AB^{-1}x^0 \qquad e_0 = r_0$$

2: minimization on V_k

$$\alpha_k = \frac{\langle B^{-1}r_k, r_k\rangle}{\langle AB^{-1}e_k, B^{-1}r_k\rangle}, \qquad x^{k+1} = x^k + \alpha_k B^{-1}e_k$$

3: stopping criterion: **if** $\|x^* - x^{k+1}\|_A \leq$ tol **then** stop
4: orthogonalization

$$\begin{aligned} r_{k+1} &= r_k - \alpha_k AB^{-1}e_k \\ \beta_k &= -\frac{\langle B^{-1}r_{k+1}, r_{k+1}\rangle}{\langle B^{-1}r_k, r_k\rangle} \\ B^{-1}e_{k+1} &= B^{-1}r_{k+1} - \beta_k B^{-1}e_k \end{aligned}$$

5: goto 2

Proof

$$\begin{aligned} \|x_B^* - x_B^k\|^2_{B^{-1}AB^{-1}} &= \langle B^{-1}AB^{-1}(Bx^* - Bx^k), (Bx^* - Bx^k)\rangle \\ &= \langle A(x^* - x^k), x^* - x^k\rangle \\ &= \|x^* - x^k\|_A^2 \end{aligned}$$

The rest follows from the theorem. □

Remark Each linear iteration can be used as a preconditioner.
The construction of preconditioners is a very active field. Optimal preconditioners are based on structural properties of the underlying partial differential equation (inheritance principle). This is analysis rather than linear algebra.

A posteriori estimate for the iterative error $\|x^* - x^k\|_A$

Lemma 4.5.9 *Let A, B be s.p.d. and $\mu_0, \mu_1 \in \mathbb{R}$. The following estimates are equivalent*

(i) $\mu_0\langle Ax, x\rangle \leq \langle AB^{-1}Ax, x\rangle \leq \mu_1\langle Ax, x\rangle$

(ii) $\mu_0\langle Bx, x\rangle \leq \langle Ax, x\rangle \leq \mu_1\langle Bx, x\rangle$.

Each of the estimates implies

$$\kappa(AB^{-1}) = \kappa(B^{-1}A) \leq \frac{\mu_1}{\mu_0}. \tag{4.6}$$

Proof 1. (i)⇒(4.6):

$$\langle B^{-1}Ax, y\rangle_A = \langle AB^{-1}Ax, y\rangle = \langle Ax, B^{-1}Ay\rangle$$

$B^{-1}A$ is symmetric with respect to $\langle\cdot,\cdot\rangle_A$.
Rayleigh quotient:

$$\lambda_{\min}(B^{-1}A) = \min_{x\neq 0}\frac{\langle B^{-1}Ax, x\rangle_A}{\langle x, x\rangle_A} = \min_{x\neq 0}\frac{\langle AB^{-1}Ax, x\rangle}{\langle Ax, x\rangle} \geq \mu_0$$
$$\lambda_{\max}(B^{-1}A) = \max_{x\neq 0}\frac{\langle B^{-1}Ax, x\rangle_A}{\langle x, x\rangle_A} = \max_{x\neq 0}\frac{\langle AB^{-1}Ax, x\rangle}{\langle Ax, x\rangle} \leq \mu_1$$

Hence,

$$\kappa(B^{-1}A) = \frac{\lambda_{\min}(B^{-1}A)}{\lambda_{\max}(B^{-1}A)} \leq \frac{\mu_1}{\mu_0}$$
$$\kappa(B^{-1}A) = \kappa(B^{-1}(AB^{-1})B) = \kappa(AB^{-1}).$$

2. (i)⇒(ii):

$$(i) \Rightarrow (\lambda_{\min}(B^{-1}A))^{-1} \leq \mu_0^{-1}$$
$$(\lambda_{\min}(B^{-1}A))^{-1} = \lambda_{\max}((B^{-1}A)^{-1}) = \lambda_{\max}(A^{-1}B)$$

$A^{-1}B$ symmetric with respect to $\langle\cdot,\cdot\rangle_A$. Hence,

$$\mu_0^{-1} \geq \lambda_{\max}(A^{-1}B) = \max_{x\neq 0}\frac{\langle A^{-1}Bx, x\rangle_A}{\langle x, x\rangle_A} = \max_{x\neq 0}\frac{\langle Bx, x\rangle}{\langle Ax, x\rangle}$$
$$\Rightarrow \langle Ax, x\rangle \geq \mu_0\langle Bx, x\rangle$$

analogously

$$\mu_1^{-1} \leq (\lambda_{\max}(B^{-1}A))^{-1} = \lambda_{\min}((B^{-1}A)^{-1}) = \lambda_{\min}(A^{-1}B) = \min_{x\neq 0}\frac{\langle Bx, x\rangle}{\langle Ax, x\rangle}$$
$$\Rightarrow \langle Ax, x\rangle \leq \mu_1\langle Bx, x\rangle$$

3. (ii)⇒(i): $B^{-1}A$ is symmetric with respect to $\langle\cdot,\cdot\rangle_B$

$$\mu_1 \geq \max_{x\neq 0}\frac{\langle Ax, x\rangle}{\langle Bx, x\rangle} = \max_{x\neq 0}\frac{\langle B^{-1}Ax, x\rangle_B}{\langle x, x\rangle_B} = \lambda_{\max}(B^{-1}A)$$
$$= \max_{x\neq 0}\frac{\langle B^{-1}Ax, x\rangle_A}{\langle x, x\rangle_A} = \max_{x\neq 0}\frac{\langle AB^{-1}Ax, x\rangle}{\langle Ax, x\rangle}$$
$$\Rightarrow \langle AB^{-1}Ax, x\rangle \leq \mu_1\langle Ax, x\rangle$$

analogously

$$\mu_0\langle Ax, x\rangle \leq \langle AB^{-1}Ax, x\rangle$$

□

Theorem 4.5.10 *Assume that the preconditioner satisfies*

$$\mu_0\langle Ax, x\rangle \leq \langle AB^{-1}Ax, x\rangle \leq \mu_1\langle Ax, x\rangle \qquad \forall x \in \mathbb{R}^n,$$

and let

$$Bd = r_k.$$

Then

$$\mu_1^{-1}\|d\|_B^2 \leq \|x^* - x^k\|_A^2 \leq \mu_0^{-1}\|d\|_B^2.$$

Proof

$$\|d\|_B^2 = \langle Bd, d\rangle = \langle r_k, B^{-1}r_k\rangle = \langle B^{-1}r_k, r_k\rangle = \langle AB^{-1}A(x^* - x^k), x^* - x^k\rangle$$

Then Lemma 4.5.9 implies for $x := x^* - x^k$

$$\mu_0\langle Ax, x\rangle \leq \|d\|_B^2 \leq \mu_1\langle Ax, x\rangle.$$

□

Remark $d = B^{-1}r_k$ is computed in step 2 of PCG. This means we get an error estimate for free.
Only good preconditioners give good error estimates.

4.5.3 Generalized minimal residual method (GMRes)

Let $A \in \mathbb{R}^{n\times n}$ be regular but not s.p.d.. Then the solution of $Ax^* = b$ is not equivalent to the minimization of $J(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$ as we can observe in the following example.

Example

$$A = \begin{pmatrix} 1 & 0 \\ -10 & 1 \end{pmatrix},\ b = 0 \ \Rightarrow\ x^* = 0,\ J(x^*) = 0$$

but

$$J(x) = \frac{1}{2}(x_1^2 + x_2(-10x_1 + x_2))$$
$$J\left(\begin{pmatrix}1\\1\end{pmatrix}\right) = \frac{1}{2}(2 - 10) = -4 < J(x^*)$$

One way out of this problem is by replacing J with another functional:
Let $B \in \mathbb{R}^{n\times n}$ be s.p.d. Then the solution of $Ax^* = b$ minimizes

$$\|b - Ax\|_B^2 \quad \forall x \in \mathbb{R}^n.$$

This leads us to the following generalization of CG:

Example Let A be s.p.d.. Choose $B = A^{-1}$. Then Algorithm 8 is the CG method.

Algorithm 8 Minimal residual method

choose $x^0 \in \mathbb{R}^n$, $r_0 = b - Ax^0$
for $k = 0, 1, \ldots$ **do**
  Krylov space: $V_k = \text{span}\{r_0, \ldots, A^k r_0\}$
  solve

$$x^{k+1} \in x^0 + V_k : \ \|b - Ax^{k+1}\|_B^2 \leq \|b - Ax\|_B^2 \ \forall x \in x^0 + V_k$$

end for

Proof

$$\|b - Ax\|_{A^{-1}}^2 = \langle A^{-1}A(x^* - x), A(x^* - x)\rangle = \|x^* - x\|_A^2$$

□

Definition 4.5.11 *Algorithm 8 with $B = I$ is called* generalized minimal residual method (GMRes).

Lemma 4.5.12 *If $V_k = V_{k+1}$, then $x^{k+1} = x^*$.*

Proof From $AV_k \subset V_{k+1} = V_k$ follows that $V_k \ni x \mapsto Ax \in V_k$ is a linear injective mapping. Since V_k is finite dimensional the mapping is also surjective. Therefore,

$$\begin{aligned} &\exists w^* \in V_k : Aw^* = r_0 = b - Ax^0 \\ &\quad \Rightarrow 0 \leq \|b - Ax^{k+1}\| \leq \|b - A(x^0 + w^*)\| = \|b - A(x^0) - b + A(x^0)\| = 0 \\ &\quad \Leftrightarrow x^{k+1} = x^*. \end{aligned}$$

□

Corollary 4.5.13 *GMRes terminates after $k_0 \leq n$ iteration steps.*

How can we compute x^{k+1} cheaply?
The minimization

$$x^{k+1} \in x^0 + V_k : \ \|b - Ax^{k+1}\| \leq \|b - Ax\| \quad \forall x \in x^0 + V_k \tag{4.7}$$

is equivalent to the minimization

$$v^k \in V_k : \ \|r_0 - Av^k\| \leq \|r_0 - Av\| \quad \forall v \in V_k, \tag{4.8}$$

and $x^{k+1} = x^0 + v^k$. This is a least square problem. Therefore, we obtain the following solution [9]: Let $A_k = A|_{V_k}$ and $Q_k R_k$ be a decomposition of A_k with

$$A_k = Q_k R_k, \qquad Q_k^{-1} = Q_k^T, \qquad R_k \text{ upper triangular}, \qquad v^k = R_k^{-1} Q_k^T r_0.$$

Our task is to compute A_k, Q_k, and R_k as cheaply as possible.

1. step Inductive construction of an orthonormal basis $e_0, \ldots, e_k$, i.e., $\langle e_i, e_j \rangle = \delta_{ij}$ by Gram-Schmidt orthogonalization (which in this case is often called Arnoldi method):

$$\tilde{e}_{j+1} = Ae_j - \sum_{i=0}^{j} \underbrace{\langle Ae_j, e_i \rangle}_{:=h_{ij}} e_i, \qquad j = 0, \ldots, k$$

$$e_{j+1} = \frac{\tilde{e}_{j+1}}{\|\tilde{e}_{j+1}\|}$$

provides

$$H_k = (h_{ij}) \in \mathbb{R}^{k_0 \times k_0}, \qquad k_0 = k + 1.$$

Observation:

$$H_k = \begin{pmatrix} * & * & \ldots & * \\ * & * & & * \\ & \ddots & \ddots & \vdots \\ 0 & & * & * \end{pmatrix}, \qquad h_{k+1,k} = \|\tilde{e}_{k+1}\|$$

is a Hessenberg-matrix.

2. step Reformulation of (4.8) as

$$y \in \mathbb{R}^{k_0} : \ \|\beta e_0 - \bar{H}_k y\| \leq \|\beta e_0 - \bar{H}_k \tilde{y}\| \quad \forall \tilde{y} \in \mathbb{R}^{k_0}, \tag{4.9}$$

where

$$\|b - Ax\| = \|\beta e_0 - \bar{H}_k y\| \text{ for } x = x_0 + \sum_{i=0}^{k} y_i e_i,$$

$$\beta = \|r_0\|, \quad \bar{H}_k = \begin{pmatrix} H_k \\ 0 \ldots 0 \ h_{k+1,1} \end{pmatrix} = \begin{pmatrix} * & * & \ldots & * \\ \boxed{*} & * & \ldots & * \\ & \boxed{*} & & \vdots \\ & & \ddots & * \\ 0 & & & \boxed{*} \end{pmatrix} \in \mathbb{R}^{(k_0+1) \times k_0}.$$

3. step QR-decomposition of $\bar{H}_k$

Elimination of $k + 1$ sub-"diagonal" elements by Givens rotation [9]:

$$G_{k_0+1,k_0} \cdots G_{32} G_{21} \bar{H}_k = \begin{pmatrix} * & * & \ldots & * \\ & * & \ldots & * \\ & & \ddots & \vdots \\ 0 & & & * \\ 0 & 0 & \ldots & 0 \end{pmatrix} =: \bar{R}_k = \begin{pmatrix} R_k \\ 0 \end{pmatrix}$$

Remark: Each step is cheaper than the preceding one.

$$\bar{H}_{k+1} = \left(\begin{array}{ccc|c} & \bar{H}_k & & * \\ & & & \vdots \\ & & & * \\ \hline 0 & \ldots & 0 & \boxed{*} \end{array}\right) \in \mathbb{R}^{(k_0+2) \times (k_0+1)}$$

- application of $G_{k_0+1,k_0} \cdots G_{32} G_{21}$ to last column
- elimination of $\boxed{*}$

4. step Solution of (4.9)

$$\|b - Ax\| = \|\beta e_0 - \bar{H}_k y\| = \|\bar{z}_k - \bar{R}_k y\|$$

$$\begin{aligned}
\bar{z}_k &= G_{k_0+1,k_0} \cdots G_{21}(\beta e_1) \\
&= G_{k_0+1,k_0} \begin{pmatrix} \bar{z}_k \\ 0 \end{pmatrix} = \begin{pmatrix} z_k \\ \zeta_k \end{pmatrix} \min_{y \in \mathbb{R}^{k_0}} \|\bar{z}_k - \bar{R}_k y\|^2 \qquad = \min_{y_k \in \mathbb{R}^{k_0}} \|z_k - R_k y\|^2 + |\zeta_k|^2 \\
&= |\zeta_k|^2 \text{ for } y_k = R_k^{-1} z_k
\end{aligned}$$

5. step

$$x^{k+1} = x^0 + \sum_{i=0}^{k} y_{k,i} e_i$$

only necessary if $\|b - Ax^{k+1}\| = |\zeta_k| \leq \text{tol}$

An algorithmic version of GMRes can be found on the web (Wikipedia) or in literature, for example [8, p. 232].
GMRes requires the storage of $e_0, \ldots, e_k$. Since this might be to much, the restart of GMRes after m steps is motivated.

Algorithm 9 GMRes(m)

```
initial iterate x^0
for k = 0, 1, ... do
  y^0 = x^k
  for i = 0, ..., m − 1 do
    compute y^{i+1} from y^i by GMRes
    if ‖x* − y^{i+1}‖ ≤ tol then
      stop
    end if
  end for
  x^{k+1} = y^m
end for
```

Remark In general, GMRes(m) does not terminate after a finite number of steps. Hence, we have to investigate the convergence of GMRes(m).

Lemma 4.5.14 *Let A be positive definite, i.e.,*

$$\langle Ax, x \rangle > 0 \quad \forall x \in \mathbb{R}^n, \quad x \neq 0.$$

Let $x^k,\ k \geq 1$, be computed by GMRes from $x^0 \neq x^*$. Then

$$\|r_k\| \leq \left(1 - \frac{\mu^2}{\sigma^2}\right)^{\frac{1}{2}} \|r_0\|,$$

where

$$r_k = b - Ax^k, \quad r_0 = b - Ax^0$$

and

$$\mu = \lambda_{\min}\left(\frac{1}{2}(A + A^T)\right) > 0, \quad \sigma = \|A\|_2.$$

Proof We have

$$x^0 + \alpha r_0 \in x^0 + V_{k-1} \quad \forall \alpha \in \mathbb{R}.$$

Set $\alpha = \frac{\langle r_0, Ar_0\rangle}{\|Ar_0\|^2}$. Then

$$\begin{aligned}
\|r^k\|^2 &= \min_{x \in x^0 + V_{k-1}} \|b - Ax\|^2 \\
&\leq \|b - A(x^0 + \alpha r_0)\|^2 \\
&= \|r_0 - \alpha A r_0\|^2 \\
&= \|r_0\|^2 - 2\alpha\langle r_0, Ar_0\rangle + \alpha^2\|Ar_0\|^2 \\
&= \|r_0\|^2 - \frac{\langle r_0, Ar_0\rangle^2}{\|Ar_0\|^2} \\
\langle r_0, Ar_0\rangle &= \frac{1}{2}\left(\langle A^T r_0, r_0\rangle + \langle r_0, Ar_0\rangle\right) \\
&= \langle r_0, \frac{1}{2}(A + A^T) r_0\rangle \\
&= \frac{\langle r_0, \frac{1}{2}(A + A^T) r_0\rangle}{\langle r_0, r_0\rangle}\langle r_0, r_0\rangle \\
&\geq \lambda_{\min}\left(\frac{1}{2}(A + A^T)\right)\|r_0\|^2 \\
&= \mu\|r_0\|^2 \geq 0 \\
\|Ar_0\| &\leq \|A\|\|r_0\| \\
&= \sigma\|r_0\| \\
\|r_k\|^2 &\leq \|r_0\|^2 - \frac{\langle r_0, Ar_0\rangle^2}{\|Ar_0\|^2} \\
&\leq \|r_0\|^2 - \frac{\mu^2\|r_0\|^4}{\sigma^2\|r_0\|^2} \\
&= \left(1 - \frac{\mu^2}{\sigma^2}\right)\|r_0\|^2.
\end{aligned}$$

□

Remark $A_S := \frac{1}{2}(A + A^T)$ is symmetric for all $A \in \mathbb{R}^{n\times n}$. A_S is called the *symmetric part* of A and $A_A := \frac{1}{2}(A - A^T)$ is the *antisymmetric part*.

Theorem 4.5.15 (Elman 1992) *Assume that* $\frac{1}{2}(A + A^T)$ *is s.p.d. Then GMRes(m) converges for any* $m \geq 1$.

Proof Lemma 4.5.14 implies

$$\|r_m\| \leq \rho \|r_0\| \qquad \rho = \left(1 - \frac{\mu^2}{\sigma^2}\right)^{\frac{1}{2}} < 1$$

By induction we get

$$\|r_{jm}\| \leq \rho^j \|r_0\| \to 0 \text{ for } j \to \infty.$$

Therefore,

$$\|x^* - x^k\| = \|A^{-1} r_k\| \leq \|A^{-1}\| \|r_k\| \to 0.$$

Hence,

$$\|x^* - x^k\| \leq \text{tol if } \|r_k\| \leq \frac{1}{\|A^{-1}\|} \text{tol}.$$

□

Remark If A is s.p.d., then $\mu = \lambda_{\min}(A)$ and $\sigma = \lambda_{\max}(A)$. Hence,

$$\rho = (1 - \kappa(A)^{-2})^{\frac{1}{2}}.$$

Thus, the convergence deteriorates with $\kappa(A) \to \infty$ (similar to steepest descent).
Way out: preconditioning: replace $Ax^* = b$ by

$$\begin{array}{rll} (AB^{-1})Bx^* & = b & \text{left} \\ (AB^{-1})x^* & = B^{-1}b & \text{right} \end{array}$$

where $B \approx A$ in the sense that

$$\kappa(B^{-1}A) = \kappa(AB^{-1}) \ll \kappa(A)$$

and the solution of $By = w$ can be obtained by one matrix-vector-multiplication.

Bibliography

[1] G. Bader and P. Deuflhard. *A Semi-Implicit Mid-Point Rule for Stiff Systems of Ordinary Differential Equations.* Numer. Math. 41, pp. 373-398, 1983.

[2] F. Bornemann. *An Adaptive Multilevel Approach for Parabolic Equations in Two Space Dimensions.* Dissertation, Freie Universität Berlin; published as TR 91-07, Konrad-Zuse-Zentrum Berlin, 1991.

[3] P. Deuflhard. *Newton Methods for Nonlinear Problems.* Springer, 2004.

[4] P. Deuflhard and F. Bornemann. *Scientific Computing with Ordinary Differential Equations.* Springer, 2002.

[5] P. Deuflhard, E. Hairer, and J. Zugck. *One-Step and Extrapolation Methods for Differential-Algebraic Systems.* Numer. Math. 51, pp. 501-516, 1987.

[6] P. Deuflhard and A. Hohmann. *Numerische Mathematik I - Eine algorithmisch orientierte Einführung.* Walter de Gruyter, 2002.

[7] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations.* Springer, 1991.

[8] C. Kanzow. *Numerik linearer Gleichungssysteme. Direkte und iterative Verfahren.* Springer, 2004.

[9] R. Kornhuber and C. Schütte. *Einführung in die Numerische Mathematik - Lecture notes Numerik I.* 2001.

[10] R. Kornhuber and C. Schütte. *Mit Zahlen rechnen - Lecture notes CoMa I.* 2005.

[11] P. Kunkel and V. Mehrmann. *Differential-Algebraic Equations: Analysis and Numerical Solution.* European Mathematical Society Publishing House, 2006.

[12] J. Ortega and W. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables.* Academic Press, 1972.

[13] J. Stoer and R. Bulirsch. *Numerische Mathematik – eine Einführung, Band 1. & 2.* Springer, 2005.