Mitschrift

Numerische Methoden der Elektrotechnik

TUM EI M.Sc. Wahlpflichtmodul

Markus Hofbauer Kevin Meyer Benedikt Schmidt

SS 2014

Dozent
Prof. Dr.-Ing. Ulf Schlichtmann

Liste der noch zu erledigenden Punkte

Inhaltsverzeichnis

1 Einführung

1.1 Zusammenfassung

1.1.1 Überblick

Definitionsversuch: Entwicklung und mathematisches Verständnis von numerischen Algorithmen, als von Rechenmethoden zur zahlenmäßigen Lösung mathematischer Probleme.

Zusammenspiel mit Informatik:

1.1.2 Vorgehen

1.1.3 Aspekte für die Lösung einer mathematischen Aufgabenstellung

- Kondition eines Problems (Empfindlichkeit für Störungen)
- Numerische Lösungsverfahren
- Stabilität des Lösungsverfahrens (Empfindlichkeit für Störungen)
- Effizienz des Lösungsverfahrens
- Genauigkeit der Lösung

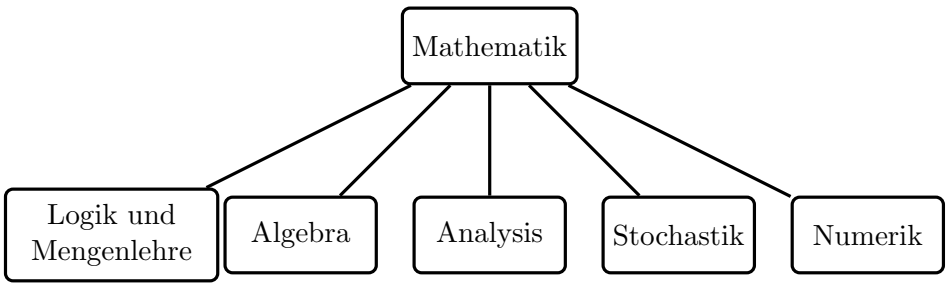


Abbildung 1.1: Teilgebiete der Mathematik

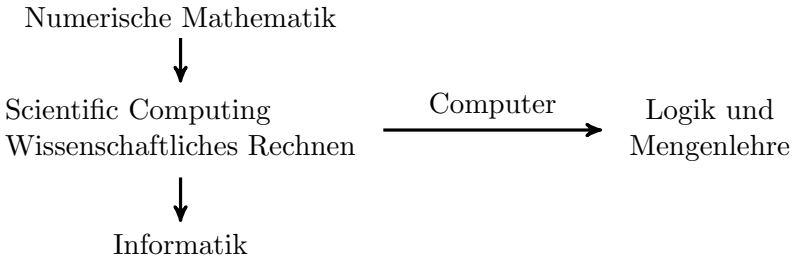


Abbildung 1.2: Zusammenspiel Mathematik und Information in der Numerik

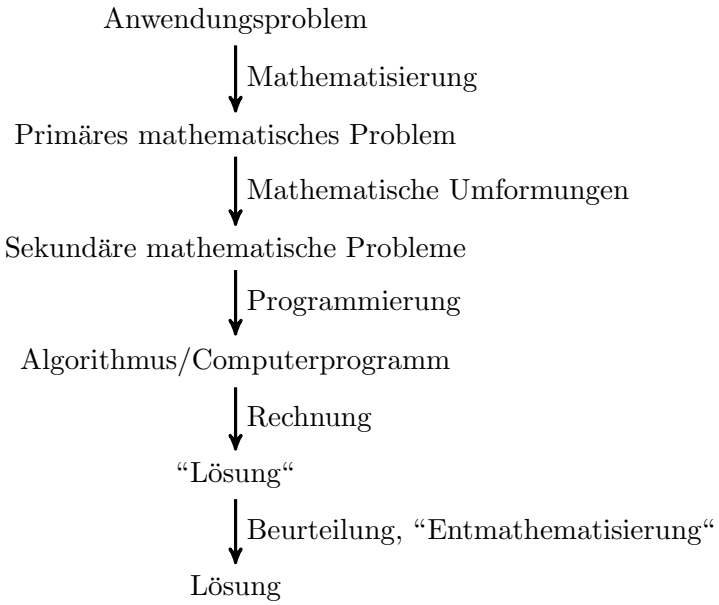


Abbildung 1.3: Vorgehen zur Lösung eines mathematischen Problems

1.2 Wiederholung

1.2.1 Lineares Gleichungssystem

$$a_{11}\,x_1 + \ldots + a_{1n}\,x_n = b_1 \tag{1.1}$$
$$a_{21}\,x_1 + \ldots + a_{2n}\,x_n = b_2 \tag{1.2}$$
$$\vdots \tag{1.3}$$
$$a_{m1}\,x_1 + \ldots + a_{mn}\,x_n = b_m \tag{1.4}$$
$$\boldsymbol{A} \cdot \vec{x} = \vec{b} \tag{1.5}$$
$$\boldsymbol{A} \in \mathbb{R}^{m\times n},\ \vec{x} \in \mathbb{R}^n,\ \vec{b} \in \mathbb{R}^m \tag{1.6}$$

1.2.2 Matrizen

Elementare Matrix-Operationen

Addition

$$\boldsymbol{A}_{<m\times n>} + \boldsymbol{B}_{<m\times n>} = \boldsymbol{C}_{<m\times n>} \tag{1.7}$$
$$c_{\mu\nu} = a_{\mu\nu} + b_{\mu\nu} \tag{1.8}$$

Multiplikation mit Skalar

$$k \cdot \boldsymbol{A} = \boldsymbol{A} \cdot k = B \tag{1.9}$$
$$b_{\mu\nu} = k \cdot a_{\mu\nu} \tag{1.10}$$

Matrix-Multiplikation

$$\boldsymbol{A}_{<m\times n>} \cdot \boldsymbol{B}_{<n\times l>} = \boldsymbol{C}_{<m\times l>} \tag{1.11}$$
$$c_{\mu\nu} = \sum_{\nu=1}^{n} a_{\mu\nu} \cdot b_{\mu\lambda};\quad \mu = 1\ldots m,\ \lambda = 1\ldots l \tag{1.12}$$

Multiplikation Matrix-Vektor

$$\text{Spezialfall der Matrix-Multiplikation: } < n\times 1 > \text{ bzw. } < 1\times n > \tag{1.13}$$

Auffassen als Linearkombination der Spalten der Matrix

$$\boldsymbol{A} \cdot \vec{x} = \vec{b} \tag{1.14}$$
$$\begin{bmatrix}\vec{a}_1 & \vec{a}_2 & \ldots & \vec{a}_n\end{bmatrix} \cdot \vec{x} = \vec{b} \tag{1.15}$$
$$\vec{a}_1\,\vec{x} + \ldots + \vec{a}_n\,\vec{x}_n = \vec{b} \tag{1.16}$$

Vektoren

$$\text{Spaltenvektor:} \quad \vec{a}_{<n\times 1>} : \vec{a} \tag{1.17}$$

$$\text{Zeilenvektor:} \quad \vec{b}^T_{<1\times m>} : \vec{b}^T \tag{1.18}$$

Vektormultiplikation

$$\vec{b}^T \cdot \vec{a} = c \quad \text{(Skalarprodukt } m = n!\text{)} \tag{1.19}$$

$$\vec{a} \cdot \vec{b}^T = \boldsymbol{C} \quad \text{(Vektorprodukt, dyadisches Produkt)} \tag{1.20}$$

Matrix als Kombination von Zeilen- und Spaltenvektoren

$$\boldsymbol{A}_{<m\times n>} = \begin{bmatrix} \vec{a}_1^T \\ \vec{a}_2^T \\ \vdots \\ \vec{a}_m^T \end{bmatrix} \tag{1.21}$$

$$\boldsymbol{B}_{<n\times l>} = \begin{bmatrix} \vec{b}_1 & \vec{b}_2 & \dots & \vec{b}_l \end{bmatrix} \tag{1.22}$$

$$\boldsymbol{A} \cdot \boldsymbol{B} = \boldsymbol{C} \tag{1.23}$$

$$\vec{c}_{\mu\lambda} = \vec{a}_\mu^T \cdot \vec{b}_\lambda \tag{1.24}$$

Rechenregeln

$$(\boldsymbol{A} \cdot \boldsymbol{B}) \cdot \boldsymbol{C} = \boldsymbol{A} \cdot (\boldsymbol{B} \cdot \boldsymbol{C}) \quad \text{Assoziativität} \tag{1.25}$$

$$\boldsymbol{A} \cdot (\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{A}\,\boldsymbol{B} + \boldsymbol{A}\,\boldsymbol{B} \quad \text{Distributivität} \tag{1.26}$$

$$\boldsymbol{A} \cdot \boldsymbol{B} \neq \boldsymbol{B} \cdot \boldsymbol{A} \quad \text{Kumutativität gilt im Allgemeinen nicht} \tag{1.27}$$

Diagonalmatrix

$$\boldsymbol{D} = \operatorname{diag}\begin{bmatrix} d_1 & d_2 & \dots & d_n \end{bmatrix} = \begin{bmatrix} d_1 & \dots & 0 \\ 0 & \dots & d_n \end{bmatrix} \tag{1.28}$$

$$\boldsymbol{D}_1 \cdot \boldsymbol{D}_2 = \boldsymbol{D}_2 \cdot \boldsymbol{D}_1 \tag{1.29}$$

Einheitsmatrix

$$\boldsymbol{I} = \boldsymbol{E} = \boldsymbol{1} = \operatorname{diag}\begin{bmatrix} 1 & \dots & 1 \end{bmatrix} \tag{1.30}$$

$$\boldsymbol{A} \cdot \boldsymbol{I} = \boldsymbol{I} \cdot \boldsymbol{A} \tag{1.31}$$

$$\boldsymbol{I}^n = \boldsymbol{I} \tag{1.32}$$

$$\boldsymbol{I}^{-1} = \boldsymbol{I} \tag{1.33}$$

Transponierte Matrix

$$\boldsymbol{A}^T_{<m\times n>} = \boldsymbol{B}_{<n\times m>},\ b_{\nu\mu} = a_{\mu\nu} \tag{1.34}$$

$$(\boldsymbol{A}\cdot\boldsymbol{B})^T = \boldsymbol{B}^T\cdot\boldsymbol{A}^T \tag{1.35}$$

$$\boldsymbol{A} = \boldsymbol{A}^T \quad \Rightarrow \text{ symmetrische Matrix} \tag{1.36}$$

Inverse Matrix

$$\boldsymbol{A} \in \mathbb{R}^{n\times n} \tag{1.37}$$

$$\boldsymbol{A}^{-1}\cdot\boldsymbol{A} = \boldsymbol{A}\cdot\boldsymbol{A}^{-1} = \boldsymbol{I} \tag{1.38}$$

$\boldsymbol{A}^{-1}$ existiert nur für nicht singuläre $\boldsymbol{A}$ und ist eindeutig. (1.39)

$$\boldsymbol{A} = \text{diag}\begin{bmatrix} d_1 & \dots & d_n \end{bmatrix} \Rightarrow \boldsymbol{A}^{-1} = \text{diag}\begin{bmatrix} \frac{1}{d_1} & \dots & \frac{1}{d_n} \end{bmatrix} \tag{1.40}$$

$$(\boldsymbol{A}\cdot\boldsymbol{B}) = \boldsymbol{B}^{-1}\cdot\boldsymbol{A}^{-1} \tag{1.41}$$

Matrix- und Vektornormen

$$||\cdot||:\ \mathbb{R}^n \to \mathbb{R};\ \mathbb{R}^{n\times n} \to \mathbb{R}$$

Bedingungen für **Vektornormen**:

- $$||\vec{x}|| \geq 0,\ ||\vec{x}|| = 0 \text{ nur für } \vec{x} = \vec{0}$$
- $$||c\cdot\vec{x}|| = |c|\cdot||\vec{x}||,\ c\in\mathbb{R}$$
- $$||\vec{x}+\vec{y}|| \leq ||\vec{x}|| + ||\vec{y}||$$

  Dreiecksungleichung

Bedingungen für **Matrixnormen**:

- $$||\boldsymbol{A}|| \geq 0,\ ||\boldsymbol{A}|| = 0 \text{ nur für } \boldsymbol{A} = \boldsymbol{0}$$
- $$||c\cdot\boldsymbol{A}|| = |c|\cdot||\boldsymbol{A}||,\ c\in\mathbb{R}$$
- $$||\boldsymbol{A}+\boldsymbol{B}|| \leq ||\boldsymbol{A}|| + ||\boldsymbol{B}||$$
- $$||\boldsymbol{A}\cdot\boldsymbol{B}|| \leq ||\boldsymbol{A}||\cdot||\boldsymbol{B}||$$

  Multiplikativitätsbedingung

- $$||\boldsymbol{A} \cdot \vec{x}|| \leq ||\boldsymbol{A}|| \cdot ||\vec{x}||$$

  Kompatibilitätsbedingung

Vektornormen

$\|\vec{x}\|_1 = \sum_{i=1}^{n} \|x_i\|$	Betragssummennorm, l_1-Norm
$\|\vec{x}\|_2 = \sqrt{\sum_{i=1}^{n} \|x_i\|^2}$	Euklidnorm, l_2-Norm, Vektorlänge
$\|\vec{x}\|_\infty = \max_i \|x_i\|$	Maximumsnorm, l_∞-Norm, Tschebychefnorm
$\|\vec{x}\|_p = \sqrt[p]{\sum_{i=1}^{n} \|x_i\|^p},\ p \geq 1$	Höldernormnorm, l_p-Norm

Matrixnormen ($A \in \mathbb{R}^{n \times n}$)

$\|\|\boldsymbol{A}\|\|_M = n \cdot \max_{i,j} \|a_{ij}\|$	Gesamtnorm, Matrixnorm ($\|\|\boldsymbol{I}\|\| = n$)
$(\|\|\boldsymbol{A}\|\|_\infty =) \|\|\boldsymbol{A}\|\|_Z = \max_i \sum_{j=1}^{n} \|a_{ij}\|$	Zeilennorm ($\|\|\boldsymbol{I}\|\|_Z = 1$)
$(\|\|\boldsymbol{A}\|\|_1 =) \|\|\boldsymbol{A}\|\|_S = \max_j \sum_{i=1}^{n} \|a_{ij}\|$	Spaltennorm ($\|\|\boldsymbol{I}\|\|_S = 1$)
$\|\|\boldsymbol{A}\|\|_E = \sqrt{\sum_{i=1} \sum_{j=1} \|a_{ij}\|^2}$	Euklidnorm, Schurnorm, Frobeniusnorm ($\|\|\boldsymbol{I}\|\|_E = \sqrt{n}$)
$(\boldsymbol{A}\|\|_\lambda =) \|\|\boldsymbol{A}\|\|_\lambda = \sqrt{\lambda_{\max}(\boldsymbol{A}^T \boldsymbol{A})}$	Spektralnorm, Hilbertnorm ($\|\|\boldsymbol{I}\|\|_\lambda = 1$)

Kompatibilität zwischen Vektor und Matrixnorm

l_1 :	$\|\|\boldsymbol{A}\vec{x}\|\|_1 \leq \|\|\boldsymbol{A}\|\|_M \cdot \|\|\vec{x}\|\|_1$
	$\|\|\boldsymbol{A}\vec{x}\|\|_1 \leq \|\|\boldsymbol{A}\|\|_S \cdot \|\|\vec{x}\|\|_1$
l_2 :	M, λ, E
l_∞ :	M, Z

1.3 Schaltungsanalyse

$\vec{i}_{<k>}$	Kantenstromvektor	$\vec{u}_{<k>}$	Kantenspannungsvektor
$\vec{i}_{0<k>}$	Kantenquellenstromvektor	$\vec{u}_{0<k>}$	Kantenquellenspannungsvektor
$\vec{i}_{n<k>}$	Knotenquellenstromvektor	$\vec{u}_{n<k>}$	Knotenquellenspannungsvektor

$\boldsymbol{A}_{<n \times k>}$	Kontenmatrix, Knoteninzidenzmatrix
$\boldsymbol{Y}_{<k \times k>}$	Kantenadmittanzmatrix
$\boldsymbol{Y}_{n<n \times n>}$	Kontenadmittanzmatrix

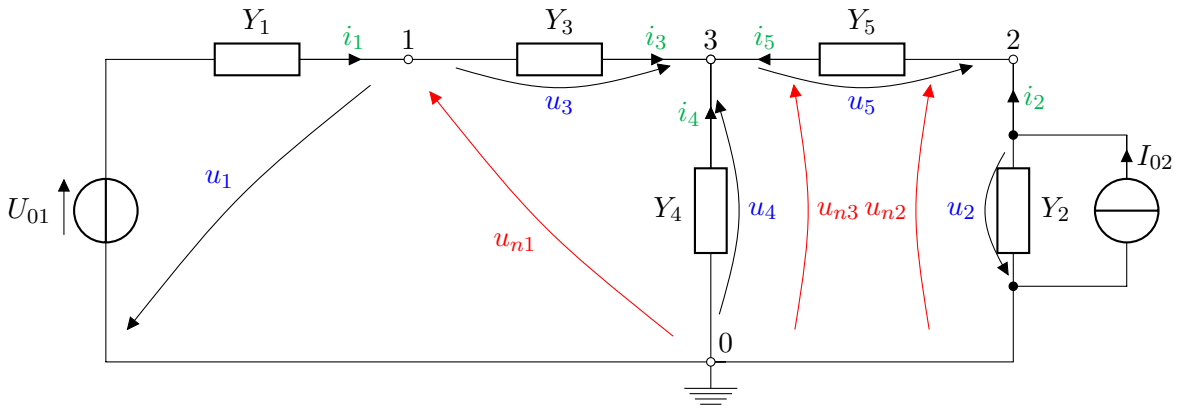


Abbildung 1.4: Beispielschaltung

1.3.1 Gerichteter Graph der Schaltung

1.3.2 Inzidenz Matrix

$$\boldsymbol{A}_{<m\times n>} = \begin{bmatrix} 1 & & & -1 \\ & 1 & & -1 \\ -1 & & 1 & \\ & & 1 & -1 \\ & -1 & 1 & \end{bmatrix} \tag{1.42}$$

Summen der Spaltenvektoren $= \vec{0}$.
$\Rightarrow \boldsymbol{A}$ hat linear abhängige Spalten.

Rang der Matrix $\boldsymbol{A}$: $r = \text{rang}\,(\boldsymbol{A}) = 3 = n - 1$.
Dimension des Nullraums: Zahl der Spalten $- \; r = 1$.
Vektor im Nullraum von $\boldsymbol{A}$:

$$\boldsymbol{A} \cdot \vec{u} = \vec{0} \Rightarrow \vec{u} \in; \quad \vec{u} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \tag{1.43}$$

1.3.3 Laplace Matrix

$$\boldsymbol{A}^T \cdot \boldsymbol{A} = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 1 & 1 \\ -1 & -1 & 0 & -1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & -1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 0 & -1 & -1 \\ 0 & 2 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & 3 & -1 \end{bmatrix} \tag{1.44}$$

- singulär

- $r = 3$
- symmetrisch

$$\boldsymbol{A}^T \cdot \boldsymbol{A} = \underbrace{\boldsymbol{D}}_{\text{Grad (degree)}} - \underbrace{\boldsymbol{W}}_{\text{Adjazenz}} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \tag{1.45}$$

1.3.4 Kirchhoffsches Stromgesetz (KCL)

$\boldsymbol{A}^T \cdot \vec{w} = \vec{0}$ (keine Stromquellen) $\vec{w}$: Kantenströme
$\boldsymbol{A}^T \cdot \vec{w} = \vec{f}$ (mit Stromquellen) $\vec{f}$: Stromquellen

1.3.5 Ohmsches Gesetz

$\vec{w} = \boldsymbol{C} \cdot \vec{e}$ $\boldsymbol{C}$: Diagonalmatrix der Kantenleitwerte
$\vec{e}$: Kantenspannungen

1.3.6 Kirchhoffsches Spannungsgesetz (KVL)

$\vec{e} = \vec{b} - \boldsymbol{A} \cdot \vec{u}$ $\vec{u}$: Knotenspannungen - GESUCHT
$\vec{b}$: Spannungsquellen

$$\boldsymbol{A}^T \cdot \vec{w} = \vec{f} \tag{1.46}$$

$$\boldsymbol{A}^T \cdot \boldsymbol{C} \cdot \vec{e} = \vec{f} \tag{1.47}$$

$$\boldsymbol{A}^T \cdot \boldsymbol{C} \left(\vec{b} - \boldsymbol{A} \cdot \vec{u} \right) = \vec{f} \tag{1.48}$$

$$\underbrace{\boldsymbol{A}^T \cdot \boldsymbol{C} \cdot \boldsymbol{A}}_{\text{Systemmatrix, gewichtete Laplace Matrix}} \vec{\sqcap} = \boldsymbol{A}^{\mathcal{T}} \cdot \boldsymbol{C} \cdot \vec{\lfloor} - \vec{\{} \tag{1.49}$$

$$\boldsymbol{Y}_{<n \times n>} \cdot \vec{u}_{<n>} = \vec{d}_{<n>} \tag{1.50}$$

singulär $\Rightarrow$ nicht invertierbar
$\Rightarrow$ keine Lösung für Gleichungssystem
Abhilfe: Festlegung eines Bezugspunktes, z.B. $u_0 = 0$
...

1.3.7 Alternative Darstellung

$$\boldsymbol{C}^{-1} \cdot \vec{w} + \boldsymbol{A} \cdot \vec{u} = \vec{b} \tag{1.51}$$

$$\boldsymbol{A}^T \cdot \vec{w} = \vec{f} \tag{1.52}$$

$$\begin{bmatrix} \boldsymbol{C}^{-1} & \boldsymbol{A} \\ \boldsymbol{A}^T & \boldsymbol{0} \end{bmatrix} \cdot \begin{bmatrix} \vec{w} \\ \vec{w} \end{bmatrix} = \begin{bmatrix} \vec{b} \\ \vec{f} \end{bmatrix} \tag{1.53}$$

1.4 Maschinendarstellung von Zahlen

(IEEE 754-2008-“Binary Floating Point Arithmetic Standard“)
Darstellung reeller Zahlen: 64 bit:

s	1 bit	Vorzeichen
c	11 bit	Exponent
f	52 bit	Mantisse

$$x = (-1)^s \cdot 2^{c-1023} \cdot (1 + f)$$

z.B.:

$$10 : \underbrace{0}_{s} \quad \underbrace{100000000010}_{c\text{: } 1026-1023=3} \quad \underbrace{0100\ldots010}_{f\text{: } 1{,}25} \tag{1.54}$$

$$-0{,}8 : \underbrace{1}_{s} \quad \underbrace{01111111110}_{c\text{: } 1022-1023=-1} \quad \underbrace{10011001\ldots10011010}_{f\text{: } 1{,}599\,99} \tag{1.55}$$

1.4.1 Gleitkomma-Arithmetik: Eigenschaften und Fehlerarten

x: beliebige reelle Zahl (mathematisch exakt)
z_1, z_2: Gleitkomma-Maschinenzahlen (endliche Stellenzahl)
Gleitkomma - Grund**op**erationen
$\mathrm{gl}(z_1 \,\mathrm{op}\, z_2)$, op := $+$, $-$, $\times$, $/$

a) $z_1 \,\mathrm{op}\, z_2 = x$; x nicht notwendigerweise Gleitkomma-Maschinenzahl

b) $+$, $-$: Exponentenangleich erforderlich - Mantissenstellen können verloren gehen.

c) Subtraktion nahezu gleichgroßer Zahlen:

- Ergebnis mit wesentlich kleinerer Mantisse
- Normalisierung (Linksverschiebung Mantisse, Exponentenangleich): Nachziehen nicht signifikanter Ziffern, Verlust signifikanter Ziffern

d) Unterlauf/Überlauf

...

2 Nichtlineare Gleichungen

Dies ist nötig beispielsweise in der Schaltungsanalyse inklusiver nichtlinearer Elemente (z.B. Diode, Transistor).

Gesucht sei hierbei das stationäre Verhalten bei konstanter Erregung (*DC-Arbeitspunkt*). Die Diode kann mithilfe von $i_D = I_S \cdot \left(e^{\frac{u_D}{U_T}} - 1\right)$ beschrieben werden. Daraus folgt dann

$$I_0 - \frac{u_D}{R} = I_S \cdot \left(e^{\frac{u_D}{U_T}} - 1\right)$$

Gesucht ist somit die Nullstelle von

$$F(u_D) = f(u_D) - g(u_D) = 0$$

Nullstellensuchen sind meist iterative Verfahren welche aus einem Startwert $\vec{x}$ eine Lösung $\vec{x}^\star$ mit $F(\vec{x}^\star) = 0$ ermitteln. Zuerst einmal wollen wir das Problem eindimensional betrachten:

$$F(x) = 0$$

Gegeben ist dabei ein $F(x)$, welches stetig auf $[a, b]$ ist mit $F(a) \cdot F(b) < 0$ und gesucht ist $x^\star$ mit $F(x^\star) = 0$, $a \leq x^\star \leq b$.

2.1 Intervallhalbierung

1. Startintervall $[a^{(0)}, b^{(0)}] = [a, b]$, $k = 0$
2. Intervallmitte $m = \frac{a^{(k)} + b^{(k)}}{2}$

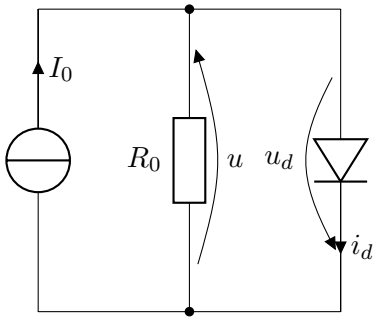


Abbildung 2.1: Beispiel einer Schaltung

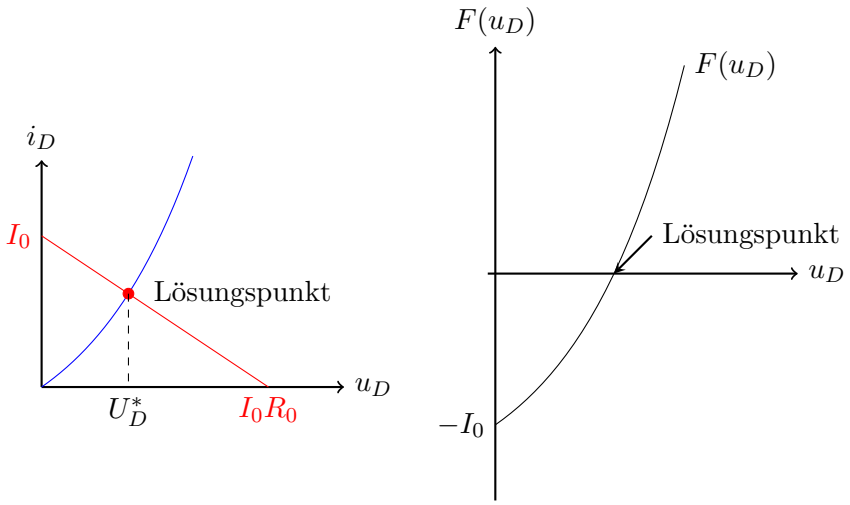


Abbildung 2.2: Arbeitspunktermittlung durch Nullstellensuche

3. $[a^{k+1)}, b^{(k+1)}] = \begin{cases} [m, b^{(k)}] & \text{fuer } F(m) \cdot F(a^{(k)}) > 0 \\ [a^{(k)}, m] & \text{fuer } F(m) \cdot F(a^{(k)}) < 0 \end{cases}$

4. Falls $\left|a^{(k+1)} - b^{k+1)}\right| > \epsilon$ dann gehe zu Schritt 2, $k = k + 1$

Die Konvergenz eines iterativen Verfahrens wird mithilfe der Parameter L, dem Konvergenzfaktor, und p, der Konvergenzordnung, beschrieben.

$$\Delta^{(k+1)} = \frac{1}{2}\Delta^{(k)} = (\frac{1}{2})^{k+1}\Delta^{(0)}$$

$$\epsilon = x - x^\star$$

$$\left|\epsilon^{(k+1)}\right| \leq L \left|\epsilon^{(k)}\right|^p$$

Die Intervallhalbierung liegt bei $p = 1$, also linearer Konvergenz, und $L = 1/2$. Die Anzahl der Iterationsschritte bis zu einem Restfehler Δ^R ist damit

$$\kappa = \left\lceil \mathrm{ld}\left(\frac{\Delta^{(0)}}{\Delta^R}\right)\right\rceil$$

2.2 Newton-Raphson-Verfahren

Die Idee hinter dem Newton-Raphson-Verfahren ist eine Taylorreihe erster Ordnung:

$$F(x) = F(x^{(k)}) + F'(x^{(k)}) \cdot (x - x^{(k)}) + F''(x^{(k)}) \cdot \frac{(x - x^{(k)})^2}{2} + \ldots$$

$$x^{(k+1)} = x^{(k)} - \frac{F(x^{(k)})}{F'(x^{(k)})}$$

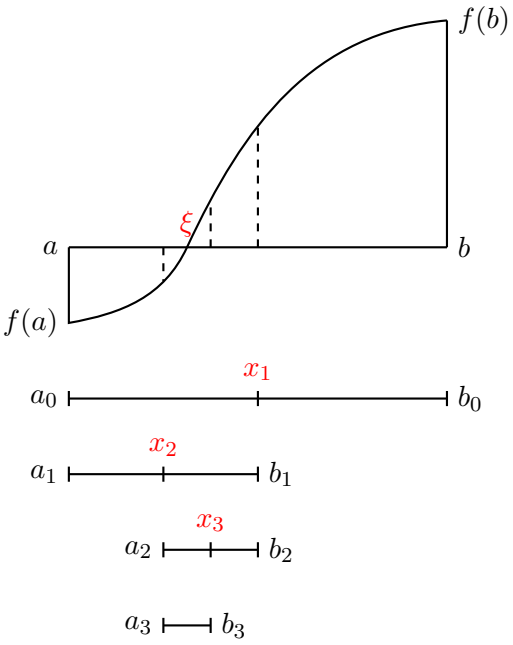


Abbildung 2.3: Intervallhalbierung

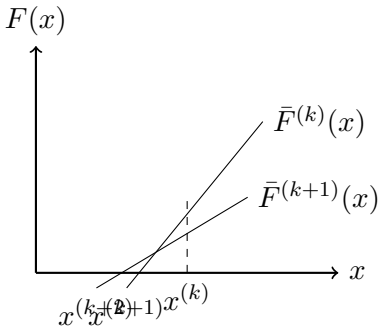


Abbildung 2.4: Newton-Iteration

Das Konvergenzverhalten des Newton-Raphson-Verfahrens erhält man aus einer Taylorreihe:

$$x^{(k+1)} = x^{(k)} - \frac{F(x^\star) + F'(x^\star)(x^{(k)} - x^\star) + \frac{1}{2}F''(x^\star) \cdot \epsilon^{k^2} + ... + \frac{1}{n!}F^{(n)}(x^\star) \cdot \epsilon^{k^n}}{F'(x^\star) + F''(x^\star) \cdot \epsilon^{(k)} + \frac{1}{(n-1)!} \cdot F^{(n)} {\epsilon^{(k)}}^{n-1}}$$

$$\epsilon^{(k+1)} = \epsilon^{(k)} \left(1 - \frac{F' + \frac{1}{2}F'' \cdot \epsilon^{(k)} + \dots}{F' + F'' \cdot \epsilon^{(k)} + \dots}\right) \approx \epsilon^{(k)} \left(\left(1 + \frac{1}{2}\epsilon^{(k)}\frac{F''}{F'}\right) \cdot \left(1 - \frac{1}{2}\epsilon^{(k)}\frac{F''}{F'}\right)\right)$$

$$\epsilon^{(k+1)} \approx \epsilon^{(k)} \left(1 - \left(1 + \frac{1}{2} \cdot \epsilon^{(k)}\frac{F''}{F'} - \epsilon^{(k)}\frac{F''}{F'}\right)\right) = \frac{1}{2} \cdot {\epsilon^{(k)}}^2 \frac{F''}{F'}$$

Probleme beim Newton-Raphson-Verfahren treten bei n-fachen Nullstellen ($F'(x^\star) = \dots = F^{(n-1)}(x^\star) = 0$), $F^{(n)}(x^\star) \neq 0$) auf.

$$\epsilon^{(k+1)} \approx \epsilon^{(k)} \cdot \left(1 - \frac{(n-1)!}{n!}\right) - \epsilon^{(k)} \cdot \left(1 - \frac{1}{n}\right) = \epsilon^{(k)}\frac{n-1}{n}$$

Somit erhält man nur mehr lineare Konvergenz. z.B.: $F(x) = e^x - 1 \Rightarrow x^{(k+1)} = x^{(k)} - \frac{e^{x^{(k)}}-1}{e^{x^{(k)}}}$

2.2.1 Variation: Sekanten-Methode

$$F'(x^{(k)}) = \lim_{x \to x^{(k)}} \frac{F(x) - F(x^{(k)})}{x - x^{(k)}}$$

$$F'(x^{(k)}) = \frac{F(x^{(k-1)}) - F(x^{(k)})}{x^{(k-1)} - x^{(k)}}$$

Eingesetzt in die Newton-Gleichung erhält man somit

$$x^{(k+1)} = x^{(k)} - \frac{F(x^{(k)}) \cdot (x^{(k)} - x^{(k-1)})}{F(x^{(k)}) - F(x^{(k-1)})}$$

Dadurch ist pro Schritt nur eine Funktionsauswertung nötig.

2.3 Fixpunktiteration

Fixpunkt x^* einer Funktion $g(x)$

$$x^* = g(x^*) \tag{2.1}$$

Nullstellenproblem und Fixpunktproblem können ineinander überführt werden., z.B.

$$g(x) = x - F(x)$$

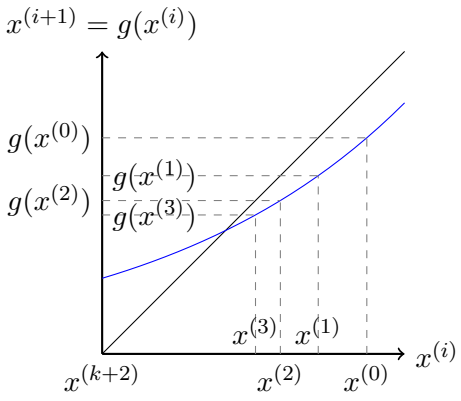


Abbildung 2.5: Fixpunktiteration

2.3.1 Fixpunkttheorem

1. Sei $g(x)$ stetig auf $[a, b]$, sowie $g(x) \in [a, b]$ für alle $x \in [a, b]$.
   Dann hat $g(x)$ mindestens einen Fixpunkt auf $[a, b]$
2. Existiere zusätzlich $g'(x)$ auf (a, b), sowie eine positive Konstante $k < 1$ mit $|g'(x)| \leq k$, für alle $x \in (a, b)$
   - Dann hat $g(x)$ exakt einen Fixpunkt auf $[a, b]$
   - Gilt $0 < k < 1$, dann konvergiert für jedes $x^{(0)} \in [a, b]$ die Folge

     $$x^{(n)} = g(x^{(n-1)}) \quad , n \geq 1$$

     zum eindeutigen Fixpunkt $x^* \in [0, b]$.

2.3.2 Konvergenzverhalten iterativer Verfahren

$$x^{(k+1)} = g(x^{(k)}) \qquad \text{Iterationsvorschrift} \tag{2.2}$$

$$x^* = g(x^*) \qquad \text{Fixpunkt} \tag{2.3}$$

Taylorreihe von $g(x)$ um x^*

$$\left[g(x^{(k)})\right] x^{(k+1)} = \underbrace{g(x^*)}_{x^*} + g'(x^*) \cdot (x^{(k)} - x^*) + \frac{1}{2} g''(x^*) \cdot (x^{(k)} - x^*)^2 + \dots \tag{2.4}$$

$$\varepsilon^{(k+1)} = g'(x^*) \cdot \varepsilon^{(k)} + \frac{1}{2} g''(x^*) \cdot {\varepsilon^{(k)}}^2 + \dots \tag{2.5}$$

- $g'(x^*) \neq 0$, $|g'(x^*)| < 1$ $\quad \rightarrow$ lineare Konvergenz
- $g'(x^*) = 0$, $g''(x^*) \neq 0$ $\quad \Rightarrow$ quadratische Konvergenz (”Konstruktionshinweis”)

2.4 Mehrdimensionales Newton-Raphson-Verfahren

$$\vec{F}(\vec{x}) = \vec{0} \quad , \vec{F} = (F_1, \ldots, F_r)^T \tag{2.6}$$

$$\vec{F}(\vec{x}^{(k+1)}) = \vec{F}(\vec{x}^{(k)}) + \underbrace{\frac{\partial \vec{F}(\vec{x}^{(k)})}{\partial \vec{x}^T}} \cdot \left(\vec{x}^{(k+1)} - \vec{x}^{(k)}\right) \tag{2.7}$$

Jacobi-Matrix
Fundamentalmatrix

$$\boldsymbol{J}(\vec{x}) = \frac{\partial \vec{F}(\vec{x})}{\partial \vec{x}^T} = \left(\frac{\partial \vec{F}}{\partial \vec{x}_1}, \ldots, \frac{\partial \vec{F}}{\partial \vec{x}_r}\right) = \begin{pmatrix} \frac{\partial \vec{F}_1}{\partial \vec{x}_1} & \cdots & \frac{\partial \vec{F}_1}{\partial \vec{x}_r} \\ \vdots & & \vdots \\ \frac{\partial \vec{F}_r}{\partial \vec{x}_1} & \cdots & \frac{\partial \vec{F}_r}{\partial \vec{x}_r} \end{pmatrix} \tag{2.8}$$

$$\vec{F}(\vec{x}^{(k+1)}) = \vec{0} \Rightarrow \boldsymbol{J}(\vec{x}^{(k)}) \cdot \left(\vec{x}^{(k+1)} - \vec{x}^{(k)}\right) = -\vec{F}(\vec{x}^{(k)}) \tag{2.9}$$

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} - \boldsymbol{J}^{-1}(\vec{x}^{(k)}) \cdot \vec{F}(\vec{x}^{(k)})$$

Praktisch:

$$\boldsymbol{J}(\vec{x}^{(k)}) \cdot \vec{x}^{(k+1)} = \boldsymbol{J}(\vec{x}^{(k)}) \cdot \vec{x}^{(k)} - \vec{F}(\vec{x}^{(k)})$$

2.4.1 Konvergenz

$$\vec{x}^{(k+1)} = \vec{\Phi}(\vec{x}^{(k)}) = \vec{\Phi}(\vec{x}^{\star}) + \left.\frac{\partial \vec{\Phi}}{\partial \vec{x}^T}\right|_{\vec{x}=\vec{x}^{\star}} \cdot \left(\vec{x}^{(k)} - \vec{x}^{\star}\right) + \mathcal{O}\left(\left(\vec{x}^{(k)} - \vec{x}^{\star}\right) \cdot \left(\vec{x}^{(k)} - \vec{x}^{\star}\right)^T\right)$$

$$\vec{\epsilon}^{(k+1)} = \left.\frac{\partial \vec{\Phi}}{\partial \vec{x}^T}\right|_{\vec{x}=\vec{x}^{\star}} \vec{\epsilon}^{(k)} + \mathcal{O}\left(\vec{\epsilon}^{(k)} \cdot \vec{\epsilon}^{(k)^T}\right)$$

$$\vec{\Phi}(\vec{x}) = \vec{x} - \boldsymbol{J}^{-1} \cdot \vec{F}$$

$$\frac{\partial \vec{\Phi}}{\partial \vec{x}^T} = \underbrace{\boldsymbol{E} - \underbrace{\boldsymbol{J}^{-1} \underbrace{\frac{\partial \vec{F}}{\partial \vec{x}^T}}_{\boldsymbol{J}^{-1}}}_{\boldsymbol{E}}}_{\mathbf{0}} - \underbrace{\frac{\partial \boldsymbol{J}^{-1}}{\partial \vec{x}^T} \cdot \vec{F}}_{\mathbf{0}}$$

2.4.2 Abbruchkriterien

Gegeben sei eine Nullstellen-Suche: $\vec{F}(\vec{x}) = \vec{0} \Rightarrow \vec{x}^{\star}$

Als Abbruchkriterium wäre schön

$$||\vec{x}^{(k)} - \vec{x}^{\star}|| = ||\vec{\epsilon}^{(k)}|| \leq ||\vec{\epsilon}^{(A)}||$$

oder

$$\frac{||\vec{\epsilon}^{(k)}||}{||\vec{x}^{\star}||} \leq ||\vec{\epsilon}^{(R)}||$$

Problematisch dabei ist, dass $\vec{x}^{\star}$ unbekannt ist. Daher muss man in der Anwendung ein anderes Kriterium wählen:

- $||\vec{x}^{(k+1)} - \vec{x}^{(k)}|| \leq ||\vec{\epsilon}^{(A)}||$ bzw. $\frac{||\vec{x}^{(k+1)} - \vec{x}^{(k)}||}{||\vec{x}^{(k+1)}||} \leq ||\vec{\epsilon}^{(R)}||$
- $||\vec{F}(\vec{x}^{(k+1)}) - \vec{F}(\vec{x}^{(k)})|| \leq \Delta_F^{(A)}$
- $||\vec{F}(\vec{x}^{(k)})|| \leq F_{min}$
- $k > k_{max}$

3 Lineare Gleichungssysteme

$$\boldsymbol{A} \cdot \vec{x} = \vec{b}, \qquad \boldsymbol{A} \in \mathbb{R}^{n \times n}; \quad \vec{b}, \vec{x} \in \mathbb{R}^n, \quad \boldsymbol{A} \neq 0 \tag{3.1}$$

$$\text{Theoretische: } \vec{x} = \boldsymbol{A}^{-1} \cdot \vec{b} \tag{3.2}$$

$$\text{Alternativ: Lösung durch „Division“-Eliminierung} \tag{3.3}$$

3.1 Methode zur Bestimmung von A^{-1}

$$\boldsymbol{A} \cdot \boldsymbol{A}^{-1} = \boldsymbol{I} \tag{3.4}$$

$$\text{Sei } \boldsymbol{X} = \boldsymbol{A}^{-1} \tag{3.5}$$

$$\boldsymbol{A} \cdot \boldsymbol{X} = \boldsymbol{I} \tag{3.6}$$

$$\boldsymbol{A} \cdot \begin{bmatrix} \vec{x}_1 & \vec{x}_2 & \dots & \vec{x}_n \end{bmatrix} = \begin{bmatrix} \vec{l}_1 & \vec{l}_2 & \dots & \vec{l}_n \end{bmatrix}; \quad \vec{l}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \dots, \vec{l}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \tag{3.7}$$

$\boldsymbol{A} \cdot \vec{x}_1 = \vec{l}_1; \quad \boldsymbol{A} \cdot \vec{x}_2 = \vec{l}_2; \quad \dots; \boldsymbol{A} \cdot \vec{x}_n = \vec{l}_n;$
Nur rechte Seite ändert sich; $\boldsymbol{A} = \boldsymbol{L} \cdot \boldsymbol{U}$ nur einmal erforderlich.
Aufwand $n \cdot 2 \cdot \frac{1}{2}(n^2 + n) = n^3 + n^2 \to O(n^3)$ für Substitution.

Probleme:

- Pivot $p = 0$ ↯
- Pivot p “sehr klein“

Abhilfe: “Pivotisierung“- Zeilen-/Spaltentausch

- “partial pivoting“: Wähle Zeile mit betragsmäßig größtem Element in Pivotspalte
- “complete pivoting“: Zeilen- und Spaltentausch
- Pivotisierung auch zur Aufwandreduktion (Nullelemente erhalten)

Beispiel $p = 0$

$$\begin{pmatrix} 10 & -7 & 0 \\ -3 & 2{,}1 & 6 \\ 5 & -1 & 5 \end{pmatrix} \cdot \vec{x} = \begin{pmatrix} 7 \\ 3{,}9 \\ 6 \end{pmatrix} \qquad p_{21} = \frac{5}{10} = \frac{1}{2} \quad p_{31} = \frac{-3}{10} \tag{3.8}$$

$$\begin{pmatrix} 10 & -7 & 0 \\ 0 & 0 & 6 \\ 0 & 2{,}5 & 5 \end{pmatrix} \cdot \vec{x} = \begin{pmatrix} 7 \\ 6 \\ 2{,}5 \end{pmatrix} \qquad p_{32} = \frac{0}{2{,}5} \quad \lightning \tag{3.9}$$

$$\Longrightarrow \text{ Zeilentausch} \tag{3.10}$$

$$\begin{pmatrix} 10 & -7 & 0 \\ 0 & 2{,}5 & 5 \\ 0 & 0 & 6 \end{pmatrix} \cdot \vec{x} = \begin{pmatrix} 7 \\ 2{,}5 \\ 6 \end{pmatrix} \tag{3.11}$$

Zeilentausch durch Permutationsmatrix $\boldsymbol{P}$.
$\boldsymbol{PA}\vec{x} = \boldsymbol{P}\vec{b} \Rightarrow$ dann $\boldsymbol{PA} \rightarrow \boldsymbol{LU}$
hier: $\boldsymbol{P} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$

3.2 Kondition eines Gleichungssystems

$$\begin{pmatrix} 1 & 1 \\ 1 & 1{,}0001 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \qquad \Rightarrow \vec{x} = \begin{pmatrix} 2 \\ 0 \end{pmatrix} \tag{3.12}$$

$$\begin{pmatrix} 1 & 1 \\ 1 & 1{,}0001 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 2{,}0001 \end{pmatrix} \qquad \Rightarrow \vec{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{3.13}$$

Änderung in der 5. Stelle von $\vec{b}$ wird zu einer Änderung in der ersten Stelle der Lösung $\vec{x}$ "verstärkt". System reagiert sehr sensitiv auf kleine Änderungen der Ausgangsdaten. Kein Lösungsalgorithmus kann etwas daran ändern.

$$\operatorname{cond}(\boldsymbol{A}) = 4{,}0002 \cdot 10^4 \tag{3.14}$$

3.3 Einfluss der Pivotisierung auf die Ergebnisgenauigkeit

$$\boldsymbol{B} \rightarrow \begin{pmatrix} 0{,}0001 & 1 \\ 1 & 1 \end{pmatrix} \cdot \vec{x} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \tag{3.15}$$

$$\begin{pmatrix} 0{,}0001 & 1 \\ 0 & -9999 \end{pmatrix} \cdot \vec{x} = \begin{pmatrix} 1 \\ -9998 \end{pmatrix} \quad \Rightarrow \begin{matrix} x_2 = 0{,}999\,899\,98 \\ x_1 = 1{,}000\,100\,01 \end{matrix} \tag{3.16}$$

Annahme: 3 Stellen Genauigkeit

$$\begin{pmatrix} 0{,}0001 & 1 \\ 0 & -10\,000 \end{pmatrix} \cdot \vec{x} = \begin{pmatrix} 1 \\ -10\,000 \end{pmatrix} \quad \Rightarrow \begin{matrix} x_2 = 1 \\ x_1 = 0 \ \lightning \end{matrix} \tag{3.17}$$

Zerlegung $\boldsymbol{B} = \boldsymbol{LDU}$ (ohne Genauigkeitsbeschränkung)

$$\begin{pmatrix}\text{“out of}\\ \text{scale}\\ \text{with } \boldsymbol{B}\text{ “)}\end{pmatrix} = \begin{pmatrix}1 & 0\\ 10\,000 & 1\end{pmatrix} \cdot \begin{pmatrix}0{,}0001 & 0\\ 0 & -9999\end{pmatrix} \cdot \begin{pmatrix}1 & 10\,000\\ 0 & 1\end{pmatrix} \tag{3.18}$$

Änderung der Pivotisierungsreihenfolge

$$\begin{pmatrix}1 & 1\\ 0{,}0001 & 1\end{pmatrix} \cdot \vec{x} = \begin{pmatrix}2\\ 1\end{pmatrix} \tag{3.19}$$

$$\begin{pmatrix}1 & 1\\ 0 & 0{,}9999\end{pmatrix} \cdot \vec{x} = \begin{pmatrix}2\\ 0{,}9998\end{pmatrix} \quad \Rightarrow \begin{matrix}x_2 = 0{,}999\,899\,98\\ x_1 = 1{,}000\,100\,01\end{matrix} \tag{3.20}$$

Annahme: 3 Stellen Genauigkeit

$$\left.\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix} \cdot \vec{x} = \begin{pmatrix}2\\ 1\end{pmatrix} \quad \Rightarrow \begin{matrix}x_2 = 1\\ x_1 = 0\end{matrix}\right\} \quad \begin{matrix}\text{nur sehr geringer}\\ \text{Genauigkeitsverlust}\end{matrix} \tag{3.21}$$

NB: $\text{cond}\,(\boldsymbol{B}) = 2{,}618\,385\,27$

Uns interessieren also zwei Themen:

- **Kondition** der Gleichungssystems
- **Stabilität** des Lösungsverfahrens

3.4 Kondition einer Matrix

Ausgehend von einem linearen Gleichungssystem

$$\boldsymbol{A} \cdot \vec{x} = \vec{b}$$

kann man auf der rechten Seite einen Fehler $\Delta\vec{b}$ hinzufügen.

$$\boldsymbol{A}\,(\vec{x} + \Delta\vec{x}) = \vec{b} + \Delta\vec{b}$$

Dieser resultiert dann in einem Fehler $\Delta\vec{x}$ der Lösung in $\vec{x}$. Der Zusammenhang zwischen diesen Größen lautet

$$\boldsymbol{A}\Delta\vec{x} = \Delta\vec{b}$$

und kann umgeformt werden zu

$$\Delta\vec{x} = \boldsymbol{A}^{-1}\Delta\vec{b}$$

$$||\Delta\vec{x}|| \leq ||\boldsymbol{A}^{-1}|| \cdot ||\Delta\vec{b}||$$

Gleichzeitig kann

$$\vec{b} = \boldsymbol{A}\vec{x}$$

umgeformt werden zu

$$||\vec{b}|| \leq ||\boldsymbol{A}|| \cdot ||\vec{x}||$$

$$\frac{1}{||\vec{x}||} \leq ||\boldsymbol{A}|| \cdot \frac{1}{||\vec{b}||}$$

Diese Gleichungen zusammengefasst ergeben dann

$$\frac{||\Delta\vec{x}||}{||\vec{x}||} \leq ||\boldsymbol{A}^{-1}|| \cdot ||\boldsymbol{A}|| \cdot \frac{||\Delta\vec{b}||}{||\vec{b}||}$$

Darin stellen dann $\frac{||\Delta\vec{x}||}{||\vec{x}||}$ den relativen Fehler des Ergebnisses, $||\boldsymbol{A}^{-1}|| \cdot ||\boldsymbol{A}||$ den *Verstärkungsfaktor* für Fehler in $\vec{b}$ und $\frac{||\Delta\vec{b}||}{||\vec{b}||}$ den relativen Fehler in $\vec{b}$ dar. Wir definieren die Kondition von A über

$$\text{cond}\,(\boldsymbol{A}) = ||\boldsymbol{A}^{-1}|| \cdot ||\boldsymbol{A}||$$

welche offensichtlich abhängig von der gewählten Matrixnorm ist. Eine $\text{cond}\,(\boldsymbol{A}) \to \infty$ bedeutet dabei eine schlechte, $\text{cond}\,(\boldsymbol{A}) \to 0$ eine gute *Kondition*.

3.5 Allgemeines Iterationsverfahren für lineare Gleichungssysteme

$$\boldsymbol{A} \cdot \vec{x} = \vec{b} \tag{3.22}$$

$$\vec{x}^{(k+1)} = \vec{\Phi}\left(\vec{x}^{(k)}\right) \tag{3.23}$$

$$\boldsymbol{B} \text{ beliebig} \tag{3.24}$$

$$\boldsymbol{B}\vec{x} + (\boldsymbol{A} - \boldsymbol{B})\,\vec{x} = \vec{b} \tag{3.25}$$

$$\Rightarrow \text{ Iterationsvorschrift:} \tag{3.26}$$

$$\boldsymbol{B} \cdot \vec{x}^{(k+1)} + (\boldsymbol{A} - \boldsymbol{B}) \cdot \vec{x}^{(k)} = \vec{b} \tag{3.27}$$

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} - \boldsymbol{B}^{-1} \cdot \left(\boldsymbol{A} \cdot \vec{x}^{(k)} - \vec{b}\right) \tag{3.28}$$

$$= \underbrace{\left(\boldsymbol{I} - \boldsymbol{B}^{-1} \cdot \boldsymbol{A}\right)}_{\text{Konvergenz, wenn } |\lambda|<1} \cdot \vec{x}^{(k)} + \boldsymbol{B}^{-1} \cdot \vec{b} \tag{3.29}$$

3.5.1 Iterative Verfahren zum Lösen linearer Gleichungssysteme

$$\boldsymbol{A}\vec{x} = \vec{b},\ \boldsymbol{A} \in \mathrm{R}^{n\times n},\ \vec{x}, \vec{b} \in \mathrm{R}^{n},\ \boldsymbol{A} \text{ nicht singulär}$$

$$\textbf{Splitting: } \boldsymbol{A} = \left[a_{ij}\right] = \boldsymbol{D} + \boldsymbol{L} + \boldsymbol{U} \tag{3.30}$$

$$= \begin{bmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{nn} \end{bmatrix} + \begin{bmatrix} 0 & & & \\ a_{21} & 0 & & \\ \vdots & & \ddots & \\ a_{n1} & \dots & a_{n\,n-1} & 0 \end{bmatrix} + \begin{bmatrix} 0 & a_{12} & \dots & a_{1n} \\ & 0 & & \vdots \\ & & \ddots & a_{n-1\,n} \\ & & & 0 \end{bmatrix} \tag{3.31}$$

$$\boldsymbol{A} = \boldsymbol{S} - \boldsymbol{T} \tag{3.32}$$
$$\boldsymbol{A} \cdot \vec{x} = (\boldsymbol{S} - \boldsymbol{T}) \cdot \vec{x} = \vec{b} \tag{3.33}$$
$$\Rightarrow \boldsymbol{S} \cdot \vec{x} = \boldsymbol{T} \cdot \vec{x} + \vec{b} \tag{3.34}$$
$$= (\boldsymbol{S} - \boldsymbol{A}) \cdot \vec{x} + \vec{b} \tag{3.35}$$
$$\Rightarrow \vec{x} = \boldsymbol{S}^{-1} \cdot (\boldsymbol{S} - \boldsymbol{A}) \cdot \vec{x} + \boldsymbol{S}^{-1} \cdot \vec{b} \tag{3.36}$$

...

3.5.2 Verschiedene Iterationsverfahren

1. $\boldsymbol{S} = \boldsymbol{D}$: Jacobi (Gesamtschrittverfahren)
2. $\boldsymbol{S} = \boldsymbol{D} + \boldsymbol{L}$: Gauß-Seidl (Einzelschrittverfahren)
3. $\boldsymbol{S} = \frac{1}{w}(\boldsymbol{D} + w\boldsymbol{L})$: sukzessive Überrelaxation (SOR, successive Over-Relaxation)

1. Jacobi: $\boldsymbol{S} = \boldsymbol{D}; \boldsymbol{D}$ nicht singulär (d.h. $a_{ii} \neq 0$)

$$A\vec{x} = \vec{b} \Rightarrow \boldsymbol{D} \cdot \vec{x}^{(k+1)} = (\boldsymbol{D} - \boldsymbol{A}) \cdot \vec{x}^{(k)} + \vec{b} = (-\boldsymbol{L} - \boldsymbol{U}) \cdot \vec{x}^{(k)} + \vec{b} \tag{3.37}$$
$$\vec{x}^{(k+1)} = \underbrace{\boldsymbol{D}^{-1} \cdot (-\boldsymbol{L} - \boldsymbol{U})}_{\boldsymbol{K}_j} \cdot \vec{x}^{(k)} + \underbrace{\boldsymbol{D}^{-1}\vec{b}}_{\vec{c}_j} \tag{3.38}$$
$$= \boldsymbol{K}_j \cdot \vec{x}^{(k)} + \vec{c}_j \tag{3.39}$$

2. Gauss-Seidel: $\boldsymbol{S} = \boldsymbol{D} + \boldsymbol{L}$ (nicht singulär)

$$\boldsymbol{A}\vec{x} + \vec{b} = \vec{0} \Rightarrow (\boldsymbol{D} + \boldsymbol{L})\vec{x}^{(k+1)} = (\boldsymbol{D} + \boldsymbol{L} - \boldsymbol{A})\vec{x}^{(k)} + \vec{b} = -\boldsymbol{U}\vec{x}^{(k)} + \vec{b}$$
$$\vec{x}^{(k+1)} = \underbrace{-(\boldsymbol{D} + \boldsymbol{L})^{-1} \cdot \boldsymbol{U}}_{\boldsymbol{K}_g} \vec{x}^{(k)} + \underbrace{(\boldsymbol{D} + \boldsymbol{L})^{-1}\vec{b}}_{\vec{c}_g}$$
$$\vec{x}^{(k+1)} = \boldsymbol{K}_g\vec{x}^{(k)} + \vec{c}_g$$

3. Successive Over-Relaxation: $\boldsymbol{S} = \frac{1}{\omega}(\boldsymbol{D} + \omega\boldsymbol{L})$

$$\boldsymbol{A}\vec{x} = \vec{b}$$
$$\frac{1}{\omega}(\boldsymbol{D} + \omega\boldsymbol{L})\vec{x}^{(k+1)} = \frac{1}{\omega}(\boldsymbol{D} + \omega\boldsymbol{L} - \omega\boldsymbol{A})\vec{x}^{(k)} + \vec{b}$$
$$(\boldsymbol{D} + \omega\boldsymbol{L})\,\vec{x}^{(k+1)} = (\boldsymbol{D} + \omega\boldsymbol{L} - \omega\boldsymbol{A})\vec{x}^{(k)} + \omega\vec{b}$$
$$(\boldsymbol{D} + \omega\boldsymbol{L})\,\vec{x}^{(k+1)} = (\boldsymbol{D}(1-\omega) - \omega\boldsymbol{U})\vec{x}^{(k)} + \omega\vec{b}$$
$$\vec{x}^{(k+1)} = \underbrace{(\boldsymbol{D} + \omega\boldsymbol{L})^{-1}(\boldsymbol{D}(1-\omega) - \omega\boldsymbol{U})}_{\boldsymbol{K}_\omega}\vec{x}^{(k)} + \underbrace{(\boldsymbol{D} + \omega\boldsymbol{L})^{-1}\,\omega\vec{b}}_{\vec{c}_\omega}$$
$$\vec{x}^{(k+1)} = \boldsymbol{K}_\omega\vec{x}^{(k)} + \vec{c}_\omega$$

Die jeweilige Wahl von ω bei SOR beeinflusst dann das Konvergenzverhalten. Für Spezialfälle können Empfehlungen gegeben werden:

a) Falls $a_{ii} \neq 0$, $i = 1, \ldots, n \Rightarrow \varrho(\boldsymbol{K}_\omega) \geq |1 - \omega| \Rightarrow$ Konvergenz für $0 < \omega < 2$

b) Falls $\boldsymbol{A}$ positiv definit und $0 < \omega < 2$, dann konvergiert SOR für jede Initiallösung $\vec{x}^{(0)}$

c) Falls $\boldsymbol{A}$ positiv definit und tridiagonal ist, dann $\varrho(\boldsymbol{K}_g) = \varrho(\boldsymbol{K}_j)^2 < 1$ und die optimale Wahl für ω ist:

$$\omega = \frac{2}{1 + \sqrt{1 - \varrho(\boldsymbol{K}_j)^2}}$$

Ein Beispiel zum Vergleich zwischen Jacobi und Gauß-Seidel

$$10x_1 - x_2 + 2x_3 = 6 \tag{3.40}$$
$$-x_1 + 11x_2 - x_3 + 3x_4 = 25 \tag{3.41}$$
$$2x_1 - x_2 + 10x_3 - x_4 = -11 \tag{3.42}$$
$$3x_2 - x_3 + 8x_4 = 15 \tag{3.43}$$

Jacobi:

$$x_1^{(k+1)} = \frac{1}{10}x_2^{(k)} - \frac{1}{5}x_3^{(k)} + \frac{3}{5} \tag{3.44}$$
$$x_2^{(k+1)} = \frac{1}{11}x_1^{(k)} + \frac{1}{11}x_3^{(k)} - \frac{3}{11}x_4^{(k)} + \frac{25}{11} \tag{3.45}$$
$$x_3^{(k+1)} = -\frac{1}{5}x_1^{(k)} + \frac{1}{10}x_2^{(k)} + \frac{1}{10}x_4^{(k)} - \frac{11}{10} \tag{3.46}$$
$$x_4^{(k+1)} = -\frac{3}{8}x_2^{(k)} + \frac{1}{8}x_3^{(k)} + \frac{15}{8} \tag{3.47}$$

Gauß-Seidel:

$$x_1^{(k+1)} = \frac{1}{10}x_2^{(k)} - \frac{1}{5}x_3^{(k)} + \frac{3}{5} \tag{3.48}$$
$$x_2^{(k+1)} = \frac{1}{11}x_1^{(k+1)} + \frac{1}{11}x_3^{(k)} - \frac{3}{11}x_4^{(k)} + \frac{25}{11} \tag{3.49}$$
$$x_3^{(k+1)} = -\frac{1}{5}x_1^{(k+1)} + \frac{1}{10}x_2^{(k+1)} + \frac{1}{10}x_4^{(k)} - \frac{11}{10} \tag{3.50}$$
$$x_4^{(k+1)} = -\frac{3}{8}x_2^{(k+1)} + \frac{1}{8}x_3^{(k+1)} + \frac{15}{8} \tag{3.51}$$

Man erkennt bei Gauß-Seidel die Abhängigkeit von Werten, welche erst im aktuellen Iterationsschritt berechnet werden. In der Praxis bietet Gauß-Seidel dadurch in der Regel eine schneller Konvergenz. Vorteil von Jacobi ist jedoch die bessere Parallelisierbarkeit eines Iterationsschrittes.

Konvergenzbetrachtung

$$\boldsymbol{S}\vec{x} = \boldsymbol{T}\vec{x} + \vec{b} \tag{3.52}$$
$$\boldsymbol{S}\vec{x}^{(x+1)} = \boldsymbol{T}\vec{x}^{(k)} + \vec{b} \tag{3.53}$$

3.52 - 3.53:

$$\begin{aligned}\boldsymbol{S}(\vec{x}-\vec{x}^{(k+1)}) &= \vec{T}(\vec{x}-\vec{x}^{(k)})\\ \boldsymbol{S}\vec{e}^{(k+1)} &= \boldsymbol{T}\vec{e}^{(k)}\\ \vec{e}^{(k+1)} &= \boldsymbol{S}^{-1}\boldsymbol{T}\vec{e}^{(k)}\\ \vec{e}^{(k+1)} &= \boldsymbol{G}\vec{e}^{(k)} = \boldsymbol{G}^k\vec{e}^{(0)}\end{aligned}$$

3.5.3 Gradientenverfahren

Sei $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ mit

- $\boldsymbol{A}^T = \boldsymbol{A}$
- $\vec{x}^T \cdot \boldsymbol{A} \cdot \vec{x} > 0$ für $\vec{x} \neq 0$ (positiv definit)
- $\vec{b} \in \mathbb{R}$

Dann existiert genau ein Minimum für die quadratische Form:

$$\Phi(\vec{x}) = 0{,}5 \cdot \vec{x}^T \cdot \boldsymbol{A} \cdot \vec{x} - \vec{x}^T \cdot \vec{b} \tag{3.54}$$

$$\nabla\Phi(\vec{x}) = 0{,}5\left(\boldsymbol{A}^T + \boldsymbol{A}\right) \cdot \vec{x} - \vec{b} = \boldsymbol{A} \cdot \vec{x} - \vec{b} \tag{3.55}$$

$$\text{Minimum: } \nabla\Phi(\vec{x}) = \vec{0} \Rightarrow \boldsymbol{A} \cdot \vec{x} = \vec{b} \tag{3.56}$$

$\Rightarrow$ Minimierung der quadratischen Form entspricht Lösung des linearen Gleichungssystem.

3.6 Iterativer Ansatz für Minimierungsproblem

Allgemein: $\vec{x}^{(k+1)} = \vec{x}^{(k)} + \alpha_K \cdot \vec{d}^{(k)}$
Bestimme: $\alpha_k,\ \vec{d}$ =?
Idee: Wähle Richtung des steilsten Abstiegs ("steepest descent", siehe Abbildung 3.1)

$$\vec{d}^{(k)} = -\nabla\Phi(\vec{x}^{(k)}) = \vec{b} - \boldsymbol{A}\vec{x}^{(k)} =: \vec{r}^{(k)} \text{ Residium} \tag{3.57}$$

$$\tag{3.58}$$

α_k =? $\Rightarrow$ Minimiere $\vec{\Phi}$

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} + \alpha_k\vec{r}^{(k)} \tag{3.59}$$

$$\vec{\Phi}(x^{(k+1)}) = \frac{1}{2}\left(\vec{x}^{(k)} + \alpha_k\vec{r}^{(k)}\right)^T \cdot \boldsymbol{A} \cdot \left(\vec{x}^{(k)} + \alpha_k\vec{r}^{(k)}\right) - \left(\vec{x}^{(k)} + \alpha_k\vec{r}^{(k)}\right)^T \cdot \vec{b} \tag{3.60}$$

$$\frac{\partial\vec{\Phi}(x^{(k+1)})}{\partial\alpha_k} = \frac{1}{2}\vec{r}^{(k)^T} \cdot \boldsymbol{A} \cdot \vec{x}^{(k)} + \frac{1}{2}\vec{x}^{(k)^T} \cdot \boldsymbol{A} \cdot \vec{r}^{(k)} + \alpha_k \cdot \vec{r}^{(k)^T} \cdot \boldsymbol{A} \cdot \vec{r}^{(k)} - \vec{r}^{(k)^T} \cdot \vec{b} \tag{3.61}$$

$$= \vec{r}^{(k)^T} \cdot \boldsymbol{A}\vec{x}^{(k)} - \vec{r}^{(k)^T} \cdot \vec{b} + \alpha_k \cdot \vec{r}^{(k)^T} \cdot \boldsymbol{A} \cdot \vec{r}^{(k)} \neq 0 \tag{3.62}$$

$$\dots \tag{3.63}$$

$$\Rightarrow \alpha_k = \frac{\vec{r}^{(k)^T} \cdot \vec{r}^{(k)}}{\vec{r}^{(k)^T} \cdot \boldsymbol{A} \cdot \vec{r}^{(k)}} \tag{3.64}$$

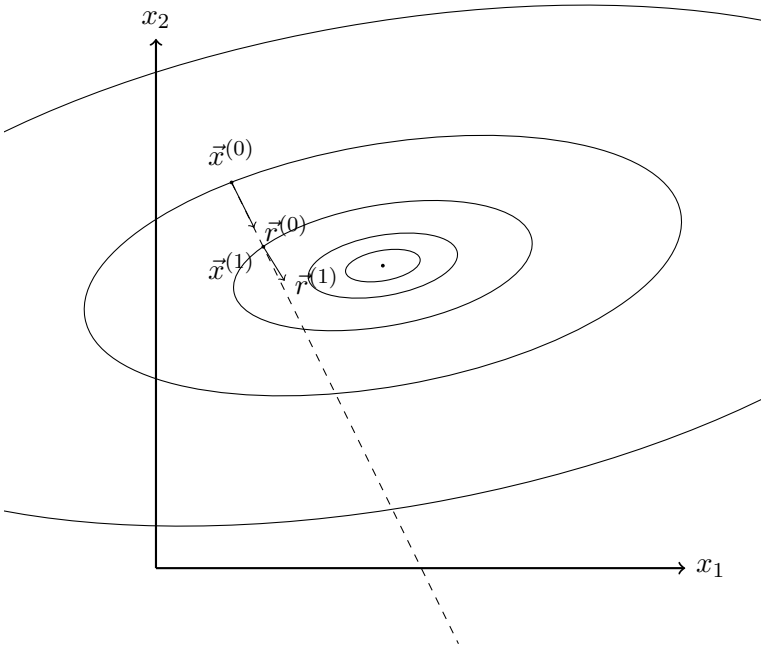


Abbildung 3.1: Gradientenverfahren - steepest descent

Verfahren:
Für $k = 0, 1, 2, \ldots$

$$\vec{r}^{(k)} = \vec{b} - \boldsymbol{A}\vec{x}^{(k)} \tag{3.65}$$

$$\alpha_k = \frac{\vec{r}^{(k)^T} \cdot \vec{r}^{(k)}}{\vec{r}^{(k)^T} \cdot \boldsymbol{A} \cdot \vec{r}^{(k)}} \tag{3.66}$$

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} + \alpha_k \cdot \vec{r}^{(k)} \tag{3.67}$$

2 Matrix-Vektormultiplikationen (dominant)
2 Vektor-Skalarprodukte
Alternative:

$$-\boldsymbol{A} \cdot \vec{x}^{(k+1)} + \vec{b} = -\boldsymbol{A} \cdot \vec{x}^{(k)} - \alpha_k \cdot \boldsymbol{A} \cdot \vec{r}^{(k)} + \vec{b} \tag{3.68}$$

$$\vec{r}^{(k+1)} = \vec{r}^{(k)} - \alpha_k \cdot \boldsymbol{A}\vec{r}^{(k)} \tag{3.69}$$

$\Rightarrow$ nur eine Matrix-Vektormultiplikation je Iteration

3.6.1 Konvergenz

$$\vec{x}^{(k)} = \vec{x}^* + \vec{e}^{(k)}, \text{ Annahme: } \vec{e} \text{ is EV von } \boldsymbol{A}: \tag{3.70}$$
$$\boldsymbol{A} \cdot \vec{e} = \lambda_e \cdot \vec{e} \tag{3.71}$$
$$\vec{r}^{(k)} = \vec{b} - \boldsymbol{A} \cdot \vec{x}^{(k)} \tag{3.72}$$
$$= \vec{b} - \boldsymbol{A} \cdot \left(\vec{x}^* + \vec{e}^{(k)}\right) \tag{3.73}$$
$$= \underbrace{\vec{b} - \boldsymbol{A} \cdot \vec{x}^*}_{\vec{0}} - \boldsymbol{A} \cdot \vec{e}^{(k)} \tag{3.74}$$
$$\vec{r}^{(k)} = -\boldsymbol{A} \cdot \vec{e}^{(k)} = -\lambda_e \cdot \vec{e}^{(k)} \Rightarrow \vec{r}^{(k)} \text{ ist ebenfalls EV} \tag{3.75}$$

$$\text{Aus Gleichung 3.67: } \vec{x}^{(k+1)} = \vec{x}^{(k)} + \alpha_k \cdot \vec{r}^{(k)} \tag{3.76}$$
$$\vec{e}^{(k+1)} = \vec{e}^{(k)} + \alpha_k \cdot \vec{r}^{(k)} \tag{3.77}$$
$$= \vec{e}^{(k)} + \frac{{\vec{r}^{(k)}}^T \cdot \vec{r}^{(k)}}{{\vec{r}^{(k)}}^T \cdot \boldsymbol{A} \cdot \vec{r}^{(k)}} \cdot \left(-\lambda_e \cdot \vec{e}^{(k)}\right) \tag{3.78}$$
$$= \vec{e}^{(k)} + \frac{{\vec{r}^{(k)}}^T \cdot \vec{r}^{(k)}}{{\vec{r}^{(k)}}^T \cdot \vec{r}^{(k)} \cdot \lambda_e} \cdot \left(-\lambda_e \cdot \vec{e}^{(k)}\right) \tag{3.79}$$
$$= \vec{e}^{(k)} - \vec{e}^{(k)} = \vec{0} \Rightarrow \text{Exakte Lösung in einem Schritt} \tag{3.80}$$

Allgemein

$$||\vec{x}||_A = \sqrt{\vec{x}^T \cdot \boldsymbol{A} \cdot \vec{x}} \text{ Energienorm} \tag{3.81}$$
$$\text{Es gilt: } ||\vec{e}^{(k+1)}||_A = \frac{\kappa_2(\boldsymbol{A}) - 1}{\kappa_2(\boldsymbol{A}) + 1} \cdot ||\vec{e}^{(k)}||_A \tag{3.82}$$

mit κ_2 als die Kondition bezüglich der Spektralnorm

3.7 CG - Konjugierte Gradientenmethode

$\vec{p}^{(k)}$ ist eine Menge mit $\underbrace{\vec{p}^{(i)^T}}_{\text{paarweise A-orthogonal bzw. A-konjugiert}} \cdot \boldsymbol{A} \cdot \vec{p}^{(j)} = 0, \forall i, j = 1, \ldots, n \quad i \neq j$

$\Rightarrow \vec{p}^{(k)}$ bilden eine Basis des $\mathbb{R}^n$

$$\text{Daher: } \vec{x}^* = \sum_{i=1}^{n} \alpha_i \cdot \vec{p}^{(i)} \tag{3.83}$$
$$\Rightarrow \vec{b} = \boldsymbol{A} \cdot \vec{x}^* = \sum_{i=1}^{n} \alpha_i \cdot \boldsymbol{A} \cdot \vec{p}^{(i)} \tag{3.84}$$

3.7.1 Schrittweite

$$0 = \frac{\partial \vec{\Phi}}{\partial \alpha_k} \left(\vec{x}^{(k)} + \alpha_k \vec{p}^{(k)} \right) \tag{3.85}$$

$$\Rightarrow \alpha_k = \frac{\vec{p}^{(k)^T} \cdot \vec{r}^{(k)}}{\vec{p}^{(k)^T} \cdot \boldsymbol{A} \cdot \vec{p}^{(k)}} \tag{3.86}$$

Wie Gradientenverfahren, aber Suchrichtung $\vec{p}$

3.7.2 Suchrichtung

Definition: Lösung $\vec{x}^{(k)}$ heißt optimal bzgl. einer Richtung $\vec{p} \neq 0$, wenn $\vec{\Phi}(\vec{x}^{(k)}) \leq \vec{\Phi}(\vec{x}^{(k)} + \lambda\vec{p})$, $\forall \lambda \in \mathbb{R}$

$\Leftrightarrow$ $\vec{\Phi}$ besitzt lokales Minimum entlang $\vec{p}$ für $\lambda = 0$

$\Leftrightarrow$ $\frac{\partial \vec{\Phi}}{\partial \lambda} \left(\vec{x}^{(k)} + \lambda \vec{p}^{(k)} \right) |_{\lambda=0} = \vec{p}^T \cdot \boldsymbol{A} \cdot \vec{x}^{(k)} - \vec{p}\vec{b}^T + \lambda \vec{p}^T \cdot \boldsymbol{A} \cdot \vec{p}|_{\lambda=0}$

$\Leftrightarrow$ $\vec{p}\vec{I}\vec{r}^{(k)}$

$$\text{Sei } \vec{x}^{(k+1)} = \vec{x}^{(k)} + \vec{q} \Rightarrow \vec{r}^{(k+1)} = \vec{b} \cdot \boldsymbol{A} \cdot \vec{x}^{(k+1)} \tag{3.87}$$

$$= \vec{b} \cdot \boldsymbol{A} \cdot \vec{x}^{(k)} - \boldsymbol{A}\vec{q} \tag{3.88}$$

$$= \vec{r}^{(k)} - \boldsymbol{A}\vec{q} \tag{3.89}$$

Forderung $\vec{x}^{(k+1)}$ ebenfalls optimal bzgl. $\vec{p}$, also

$$0 = \vec{p}^T \cdot \vec{r}^{(k+1)} = \vec{p}^T \cdot (\vec{r}^{(k)} - \boldsymbol{A}\vec{q}) \tag{3.90}$$

$$= \underbrace{\vec{p}^T \cdot \vec{r}^{(k)}}_{0} - \vec{p}^T \cdot \boldsymbol{A} \cdot \vec{q} = \underbrace{-\vec{p}^T \cdot \boldsymbol{A} \cdot \vec{q}}_{\vec{p},\vec{q},\boldsymbol{A}\text{orthogonal}} = 0 \tag{3.91}$$

Konstruktion der Suchrichtung

- Start $\vec{p}^{(0)} = \vec{r}^{(0)}$ (Ausgehend von Initiallösung $\vec{x}^{(0)}$, Gradient an $\vec{x}^{(0)}$)
- Suche Richtungen $\vec{p}^{(k+1)} = \vec{r}^{(k+1)} - \beta_k \cdot \vec{p}^{(k)}$, $k = 0, 1, 2, \ldots$ mit $\beta_k \in \mathbb{R}$, so dass gilt:

$$\vec{p}^{(j)^T} \cdot \boldsymbol{A} \cdot \vec{p}^{(k+1)} = 0 = \left(\boldsymbol{A} \cdot \vec{p}^{(j)} \right)^T \cdot \vec{p}^{(k+1)} = 0 \text{ , } j = 0, 1, \ldots, k \tag{3.92}$$

$$\Rightarrow \left(\boldsymbol{A} \cdot \vec{p}^{(j)} \right)^T \cdot \left(\vec{r}^{(k+1)} - \beta_k \cdot \vec{p}^{(k)} \right) = 0 \underset{(j=k)}{\Rightarrow} \beta_k = \frac{\left(\boldsymbol{A} \cdot \vec{p}^{(k)} \right)^T \cdot \vec{r}^{(k+1)}}{\left(\boldsymbol{A} \cdot \vec{p}^{(k)} \right)^T \cdot \vec{p}^{(k)}} \tag{3.93}$$

Zu zeigen: $\vec{p}^{(k+1)}$ ist $\boldsymbol{A}$ orthogonal zu $\vec{p}^{(j)}$ $\forall j < k$
Beweisidee: vollständige Induktion

3.7.3 Zusammenfassung des Verfahrens

$$\vec{r}^{(0)} = \vec{b} - \boldsymbol{A}\vec{x}^{(0)} \tag{3.94}$$

$$\vec{p}^{(0)} = \vec{r}^{(0)} \tag{3.95}$$

Für $k = 0, 1, \ldots$

$$\alpha_k = \frac{\vec{p}^{(k)^T} \cdot \vec{r}^{(k)}}{\vec{p}^{(k+1)^T} \boldsymbol{A}\vec{p}^{(k)}} \tag{3.96}$$

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} + \alpha_k \vec{p}^{(k)} \tag{3.97}$$

$$\vec{b} - \boldsymbol{A}\vec{x}^{(k+1)} = \vec{r}^{(k+1)} = \vec{r}^{(k)} - \alpha_k \boldsymbol{A}\vec{p}^{(k)} \tag{3.98}$$

$$\beta_k = \frac{(\boldsymbol{A}\vec{p}^{(k)})^T \cdot \vec{r}^{(k+1)}}{(\boldsymbol{A}\vec{p}^{(k)})^T \cdot \vec{p}^{(k)}} \tag{3.99}$$

$$\vec{p}^{(k+1)} = \vec{r}^{(k+1)} - \beta_k \vec{p}^{(k)} \tag{3.100}$$

Der Aufwand je Iteration besteht maßgeblich aus

- 1 Matrix-Vektorprodukt
- 2 Vektor-Skalarprodukte (nach einer Umstellung der obigen Formeln)

3.7.4 Konvergenz

Sei $\boldsymbol{A}$ symmetrisch und positiv definit, dann gilt:

- Das CG-Verfahren bricht nach höchstens n Schritten mit der exakten Lösung ab
- $||\vec{e}^{(k)}||_A \leq \frac{2c^k}{1+c^{2k}} ||\vec{e}^{(0)}||_A$ mit $c = \frac{\sqrt{K_2(\boldsymbol{A})}-1}{\sqrt{K_2(\boldsymbol{A})}+1}$

In der Praxis mit endlicher Rechengenauigkeit erhält man allerdings nicht die exakte Lösung nach n Schritten, anstelle dessen bricht man nach einem gewissen Kriterium (zum Beispiel Betrag des Residiums) ab.

Verbessert werden kann das Konvergenzverhalten durch eine Vorkonditionierung.

3.7.5 Vorkonditionierung

Für eine Linksvorkonditionierung wird $\boldsymbol{A}\vec{x} = \vec{b}$ transformiert in

$$\boldsymbol{P}^{-1}\boldsymbol{A}\vec{x} = \boldsymbol{P}^{-1}\vec{b} \tag{3.101}$$

Um eine Rechtsvorkonditionierung vorzunehmen (welche seltener benutzt wird) wird nach

$$(\boldsymbol{A}\boldsymbol{P}^{-1})\boldsymbol{P}\vec{x} = \vec{b} \tag{3.102}$$

transformiert und in zwei Schritten gelöst, zuerst

$$\boldsymbol{A}\boldsymbol{P}^{-1}\vec{y} = \vec{b} \tag{3.103}$$

und danach

$$\boldsymbol{P}\vec{x} = \vec{y} \tag{3.104}$$

Gilt $K(\boldsymbol{P}^{-1}\boldsymbol{A}) < K(\boldsymbol{A})$, dann erhält man ein besseres Konvergenzverhalten. Ideal wäre $\boldsymbol{P}^{-1} = \boldsymbol{A}^{-1}$, allerdings ist die Berechnung der Inversen aufwändiger als die Lösung des ursprünglichen Problems (die Lösung des Gleichungssystems). Möglichst einfach wäre der Einsatz von $\boldsymbol{P} = \boldsymbol{I}$, allerdings erhält man damit keinen Gewinnung durch die Vorkonditionierung. Praktischerweise wählt man eine möglichst einfach ermittelbare Matrix P welche zwischen diesen beiden Extrema liegt.

Eine mögliche Lösung für dieses Problem ist der Einsatz von *Preconditioned Conjugate Gradient (PCG)*: Sei $\boldsymbol{P}$ symmetrisch und positiv definit $\Rightarrow \exists$ Zerlegung

$$\boldsymbol{P} = \boldsymbol{P}^{\frac{1}{2}}\boldsymbol{P}^{\frac{1}{2}^T} \tag{3.105}$$

CG wird dann angewandt auf

$$\tilde{\boldsymbol{A}}\tilde{\vec{x}} = \tilde{\vec{b}} \tag{3.106}$$

$$\tilde{\boldsymbol{A}} = \boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{P}^{-\frac{1}{2}^T} \tag{3.107}$$

$$\tilde{\vec{x}} = \boldsymbol{P}^{\frac{1}{2}^T}\vec{x} \tag{3.108}$$

$$\tilde{\vec{b}} = \boldsymbol{P}^{-\frac{1}{2}}\vec{b} \tag{3.109}$$

Das Verfahren sieht dann wie folgt aus

$$\vec{r}^{(0)} = \vec{b} - \boldsymbol{A}\vec{x}^{(0)} \tag{3.110}$$

$$\boldsymbol{P}\vec{z}^{(0)} = \vec{r}^{(0)} \Rightarrow \vec{z}^{(0)} \tag{3.111}$$

$$\vec{p}^{(0)} = \vec{z}^{(0)} \tag{3.112}$$

Für $k = 0, 1, 2, \ldots$

$$\alpha_k = \frac{\vec{p}^{(k)^T} \cdot \vec{r}^{(k)}}{\vec{p}^{(k)^T} \cdot \boldsymbol{A}\vec{p}^{(k)}} \tag{3.113}$$

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} + \alpha_k\vec{p}^{(k)} \tag{3.114}$$

$$\vec{r}^{(k+1)} = \vec{r}^{(k)} - \alpha_k\boldsymbol{A}\vec{p}^{(k)} \tag{3.115}$$

$$\boldsymbol{P}\vec{z}^{(k+1)} = \vec{r}^{(k+1)} \Rightarrow \vec{z}^{(k+1)} \tag{3.116}$$

$$\beta_k = \frac{(\boldsymbol{A}\vec{p}^{(k)})^T \cdot \vec{z}^{(k+1)}}{(\boldsymbol{A}\vec{p}^{(k)})^T \cdot \vec{p}^{(k)}} \tag{3.117}$$

$$\vec{p}^{(k+1)} = \vec{z}^{(k+1)} - \beta_k\vec{p}^{(k)} \tag{3.118}$$

4 Interpolation

Aufgabe: Eine (stetige) Funktion zu bestimmen, welche gegebene Datenpunkte *bestmöglichst* abbildet.

Eine möglicher Herangehensweise an diese Problem ist eine Entwicklung in eine Taylorreihe, welche allerdings gewisse Nachteile aufweist:

- konzentriert sich auf einen Punkt
- keine Aussage über Genauigkeit an anderen Punkten
- Approximation verbessert sich nicht mit höheren Graden des Approximationspolynoms: z.B.: e^x, $\frac{1}{x}$

Eine Interpolation hingegen zielt auf eine exakte Abbildung an den gegebenen Stützstellen eine Näherung dazwischen ab.

Allgemein formuliert lautet das Interpolationsproblem: Gegeben $n+1$ Paare von reellen oder komplexen Zahlen $((x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_n, f(x_n)))$. Mit Stützstellen bezeichnet man x_i, mit Stützwert $f(x_i)$ und mit Stützpunkt $(x_i, f(x_i))$.

Ziel ist die Bestimmung von $\Phi(x, a_0, \ldots, a_n), a_0, a_1, \ldots, a_n$, so dass $\forall_{i=0}^{n} f(x_i) = \Phi(x_i)$

4.1 Lagrange Interpolationspolynom

$$L_{n,k}(x) = \frac{(x - x_0) \cdot (x - x_1) \cdot \ldots \cdot (x - x_{k-1}) \cdot (x - x_{k+1}) \cdot \ldots \cdot (x - x_n)}{(x_k - x_0) \cdot (x_k - x_1) \cdot \ldots \cdot (x_k - x_{k-1}) \cdot (x_k - x_{k+1}) \cdot \ldots \cdot (x_k - x_n)} \tag{4.1}$$

$$L_k(x) = L_{n,k} = \prod_{i=0, i \neq k}^{n} \frac{(x - x_i)}{x_k - x_i)} \tag{4.2}$$

$L_{n,k}(x_k) = 1$, $L_{n,k}(x_i) = 0$ für $x_i \neq x_k$

$$P(x) = \sum_{k=0}^{n} f(x_k) \cdot L_{n,k}(x) \tag{4.3}$$

Theorem: Seien $n+1$ paarweise unterschiedliche Stützstellen x_i sowie die zugehörigen Stützwerte $f(x_i)$ gegeben. Dann existiert ein eindeutiges Interpolationspolynom $P(x)$ vom Grad höchstens n mit $P(x_i) = f(x_i)$

Fehlerabschätzung: Gegeben seien $n+1$ paarweise unterschiedliche Stützstellen $x_i \in [a,b]$ sowie eine Funktion f die auf diesem Intervall $[a,b]$ $(n+1)$-mal stetig differenzierbar ist. Dann existiert für $x \in [a,b]$ ein $\xi(x) \in (a,b)$ mit der Eigenschaft

$$f(x) = P(x) + \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \prod_{i=0}^{n} (x - x_i) \tag{4.4}$$

4.2 Dividierte Differenzen

Sei $P(x)$ das Lagrange-Polynom n-ter Ordnung. Wähle

$$P_n(x) = a_0 + a_1(x-x_0) + a_2(x-x_0)(x-x_1) + \ldots + a_n(x-x_0)(x-x_1)\cdot\ldots\cdot(x-x_{n-1}) \tag{4.5}$$

Setzt man nun kontiniuerlich die Stützstellen x_i in $P_n(x)$ erhält man

$$P_n(x_0) = a_0 = f(x_0) \tag{4.6}$$

und somit den Wert für a_0. Aus

$$P_n(x_1) = f(x_0) + a_1(x_1 - x_0) = f(x_1) \tag{4.7}$$

folgt dann die Bestimmung von a_1

$$a_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0} \tag{4.8}$$

Dividierte Differenzen-Notation:

- 0. dividierte Differenz

$$f[x_i] = f(x_i) \tag{4.9}$$

- 1. dividierte Differenz

$$f[x_i, x_{i+1}] = \frac{f[x_{i+1}] - f[x_i]}{x_{i+1} - x_i} = \frac{f(x_{i+1} - f(x_i)}{x_{i+1} - x_i} \tag{4.10}$$

- k. dividierte Differenz

$$f[x_i, x_{i+1}, \ldots, x_{i+k}] = \frac{f[x_{i+1}, x_{i+2}, \ldots x_{i+k}] - f[x_i, x_{i+1}, x_{i+2}, \ldots x_{i+k-1}]}{x_{i+k} - x_i} \tag{4.11}$$

Mithilfe dieser Notation erhält man $P_n(x)$ als sogenannte Newtons dividierte Differenz

$$P_n(x) = f[x_0] + \sum_{k=1}^{n} f[x_0, x_1, \ldots, x_k] \cdot (x - x_0) \cdot (x - x_{k-1}) \tag{4.12}$$

Durch eine tabellarische Darstellung erhält man eine systematische Vorgehensweise zur Berechnung der dividierten Differenzen.

x	$f(x)$	1. div. Diff.	2. div. Diff.
x_0	$f[x_0]$	$f[x_0,x_1]=\frac{f[x_1]-f[x_0]}{x_1-x_0}$	$f[x_0,x_1,x_2]=\frac{f[x_1,x_2]-f[x_1,x_0]}{x_2-x_0}$
x_1	$f[x_1]$	$f[x_1,x_2]=\frac{f[x_2]-f[x_1]}{x_2-x_1}$	$\vdots$
x_2	$f[x_2]$	$f[x_2,x_3]=\frac{f[x_3]-f[x_2]}{x_3-x_2}$	
$\vdots$	$\vdots$	$\vdots$	

Bei äquidistanten Stützstellen bietet sich eine vereinfachte Notation an

$$h = x_{i+1} - x_i \tag{4.13}$$

$$x = x_0 + sh \tag{4.14}$$

$$x - x_i = (s - i) \cdot h \tag{4.15}$$

Damit ist dann

$$P_n(x) = P_n(x_0+s\cdot h) = s\cdot h\cdot f[x_0,x_1]+s\cdot(s-1)\cdot h^2 f[x_0,x_1,x_2]+\ldots+(s-n+1)h^n f[x_0,x_1,\ldots,x_n] \tag{4.16}$$

Mithilfe von Binomialkoeffizienten

$$\binom{s}{k} = \frac{s(s-1)\cdot\ldots\cdot(s-k+1)}{k!} \tag{4.17}$$

$$P_n(x) = f[x_0] + \sum_{k=1}^{n} \binom{s}{k} k! h^k f[x_0,x_1,\ldots,x_k] \tag{4.18}$$

Δ-Notation: $\Delta f(x_i) = f(x_{i+1}) - f(x_i)$

$$f[x_0,x_1] = \frac{f(x_1)-f(x_0)}{x_1-x_0} = \frac{\Delta f(x_i)}{h} \tag{4.19}$$

$$f[x_0,x_1,x_2] = \frac{1}{2h}\frac{\Delta f(x_1)-\Delta f(x_0)}{h} = \frac{1}{2}\Delta^2 f(x_0) \tag{4.20}$$

$$f[x_0,x_1,\ldots,x_k] = \frac{1}{k!h!}\Delta^k f(x_0) \tag{4.21}$$

Damit erhält man dann die Newton'sche Vorwärtsdifferenz

$$P_n(x) = f(x_0) + \sum_{k=1}^{n} \binom{s}{k} \Delta^k f(x_0) \tag{4.22}$$

4.3 Spline - Interpolation

Grundidee: Interpolation durch stückweise stetige Polynome.
Einfachste Variante: stückweise lineare Interpolation (siehe Abbildung 4.1)
Nachteil: keine Differenzierbarkeit an Endpunkten der einzelnen linearen Polynome gewährleistet.

Alternative: stückweise quadratische Polynome

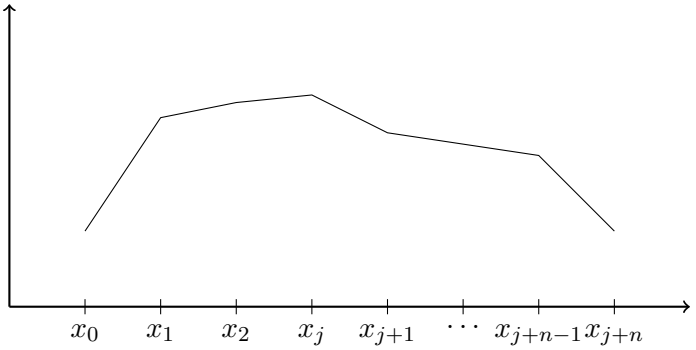


Abbildung 4.1: Beispielhafte Darstellung einer stückweise linearen Interpolation

- je 3 Koeffizienten
  $\Rightarrow$ nur 2 für Bestimmung der Endpunkte erforderlich
  $\Rightarrow$ 3. Koeffizient kann stetige Differenzierbarkeit auf $[x_0, x_n]$ sicherstellen

- ABER: Anforderung an Ableitung in $[x_0, x_n]$ können nicht berücksichtigt werden.

Weitere Alternative: stückweise kubische Polynome ("cubic splines")

- je 4 Koeffizienten
  $\Rightarrow$ stetige Differenzierbarkeit auf $[x_0, x_n]$
  $\Rightarrow$ stetige 2. Ableitung

Definition: Gegeben sei eine auf $[a, b]$ definierte Funktion f sowie Stützstellen $a = x_0 < x_1 < \ldots < x_n = b$. Ein kubischer Spline-Interpolant S erfüllt folgende Bedingungen:

1. $S(x)$ ist ein kubisches Polynom, auf $[x_j, x_{j+1}]$ als $S_j(x)$ bezeichnet, für $j = 0, 1, \ldots, n-1$
2. $S_j(x_j) = f(x_j)$, $S_j(x_{j+1}) = f(x_{j+1})$, für $j = 0, 1, \ldots, n-1$
3. $S_j(x_{j+1}) = S_{j+1}(x_{j+1})$, für $j = 0, 1, \ldots, n-2$
4. $S'_{j+1}(x_{j+1}) = S'_j(x_{j+1})$, für $j = 0, 1, \ldots, n-2$
5. $S''_{j+1}(x_{j+1}) = S''_j(x_{j+1})$, für $j = 0, 1, \ldots, n-2$
6. Eine der folgenden Randbedingungen ist erfüllt
   - $S''(x_0) = S''(x_n) = 0$ (freier bzw. natürlicher Rand)
   - $S'(x_0) = f'(x_0)$, $S'(x_n) = f'(x_n)$ (eingespannter Rand)

$$S_j(x) = a_j + b_j\,(x - x_j) + c_j\,(x - x_j)^2 + d_j\,(x - x_j)^3\ ,\ j = 0, 1, \ldots, n-1 \tag{4.23}$$

$$S_j(x_j) = a_j = f(x_j)\ ,\ j = 0, 1, \ldots, n-1 \tag{4.24}$$

$$a_{j+1} = S_{j+1}(x_{j+1}) = S_j(x_{j+1}) = a_j + b_j\,(x_{j+1} - x_j) + c_j\,(x_{j+1} - x_j)^2 + d_j\,(x_{j+1} - x_j)^3\ ,\ j = 0, 1, \ldots, n-1 \tag{4.25}$$

$$\text{mit } h_j = x_{j+1} - x_j:\ a_{j+1} = a_j + b_j \cdot h_j + c_j \cdot h_j^2 + d_j \cdot d_j \cdot h_j^3\ ,\ j = 0, 1, \ldots, n-1 \tag{4.26}$$

$$a_n := f(x_n) \text{ (Hilfsdefinition)} \tag{4.27}$$

$$b_n := S'(x_n) \tag{4.28}$$

$$S'_j(x) = b_j + 2\,c_j \cdot (x - x_j) + 3\,d_j \cdot (x - x_j)^2 \ , \ j = 0, 1, \ldots, n-1 \tag{4.29}$$

$$S'_j(x_j) = b_j \tag{4.30}$$

$$b_{j+1} = b_j + 2\,c_j \cdot h_j + 3\,d_j \cdot h_j^2 \tag{4.31}$$

$$c := \frac{S''(x_n)}{2} \tag{4.32}$$

$$S''_j(x) = 2\,c_j + 6\,d_j \cdot (x - x_j) \tag{4.33}$$

$$S''_j(x) = 2\,c_j \tag{4.34}$$

$$S''_{j+1}(x_{j+1} = 2\,c_{j+1}) \tag{4.35}$$

$$S''_{j+1}(x_{j+1} = S''_j(x_{j+1}) \tag{4.36}$$

$$c_{j+1} = c_j + 3\,d_j \cdot h_j \tag{4.37}$$

$$d_j = \frac{1}{3\,h_j} \cdot (c_{j+1} - c_j) \ , \ j = 0, 1, \ldots, n-1 \tag{4.38}$$

Gleichung 4.26 in Gleichung 4.31:

$$a_{j+1} = a_j + b_j \cdot h_j + \frac{h_j^2}{3} \cdot (2\,c_j + c_{j+1}) \tag{4.39}$$

$$b_{j+1} = b_j + h_j \cdot (c_j + c_{j+1}) \tag{4.40}$$

$$b_j = b_{j-1} + h_j \cdot (c_{j+1} + c_j) \tag{4.41}$$

Aus Gleichung 4.39:

$$b_j = \frac{1}{h_j} \cdot (a_{j+1} - a_j) - \frac{h_j}{3} \cdot (2\,c_j + c_{j+1}) \tag{4.42}$$

$$b_{j-1} = \frac{1}{h_{j-1}} \cdot (a_j - a_{j-1}) - \frac{h_{j-1}}{3} \cdot (2\,c_{j-1} + c_j) \tag{4.43}$$

In Gleichung 4.42:

$$h_{j-1}\,c_{j-1} + 2 \cdot (h_{j-1} + h_j) \cdot c_j + h_j\,c_{j+1} = \frac{3}{h_j} \cdot (a_{j+1} - a_j) - \frac{3}{h_{j-1}} \cdot (a_j - a_{j-1}) \ , \ j = 0, 1, \ldots, n-1 \tag{4.44}$$

Natürliche Splines: $S''(x_0) = S''(x_n) = 0$

$$\Rightarrow c_n = \frac{S'(x_n)}{2} = 0 \tag{4.45}$$

$$0 = S''(x_0) = 2\,c_0 + 6\,d_0 \cdot (x_0 - x_0) \Rightarrow c_0 = 0 \tag{4.46}$$

Mit Gleichung 4.44:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \dots & & \\ h_0 & 2\cdot(h_0+h_1) & h_1 & 0 & & & \\ 0 & h_1 & 2\cdot(h_1+h_2) & h_2 & 0 & & \\ 0 & \dots & & & & & \\ 0 & \dots & & & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ \\ c_n \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{3}{h_j}\cdot(a_2-a_1) - \frac{3}{h_0}\cdot(a_1-a_0) \\ \vdots \\ \vdots \\ 0 \end{bmatrix} \tag{4.47}$$

Matrix streng diagonaldominant (Diagonalelemente betragsmäßig größer als Summe aller anderen Elemente einer Zeile)
$\Rightarrow$ eindeutige Lösung

5 Numerische Infinitesimalrechnung

5.1 Numerische Differentiation

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} \tag{5.1}$$

Taylor-Polynom:

$$f(x_0 + h) = f(x_0) + h \cdot f'(x_0) + \frac{h^2}{2} \cdot f''(x_0) + \frac{h^3}{3} \cdot f^{(3)}(x_0) + \dots$$

Abbruch nach linearem Term:

$$f(x_0 + h) = f(x_0) + h \cdot f'(x_0) + R(h^2)$$

5.1.1 Vorwärtsdifferenz

$$f'(x_0) \approx \frac{f(x_0 + h) - f(x_0)}{h} + R(h) \tag{5.2}$$

5.1.2 Rückwärtsdifferenz

$$f'(x_0) \approx \frac{f(x_0) - f(x_0 - h)}{h} \tag{5.3}$$

5.1.3 Zentrierte Differenz

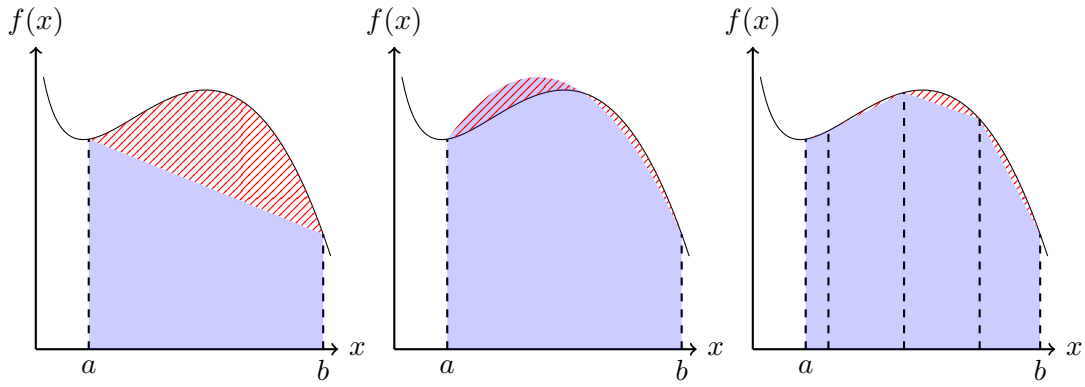
$$f'(x_0) \approx \frac{f(x_0 + h) - f(x_0 - h)}{2h}$$

Fehleranalyse

$$\begin{aligned} f(x_0 + h) &= \underbrace{\tilde{f}(x_0 - h)}_{\text{gerundeter Wert}} + e_{-1} \\ f(x_0 - h) &= \tilde{f}(x_0 + h) + e_{+1} \end{aligned} \tag{5.4}$$

$$f'(x_0) = \frac{\tilde{f}(x_0 + h) - \tilde{f}(x_0 - h)}{2h} + \frac{e_{+1} - e_{-1}}{2h} - \frac{f^{(3)}(\xi)}{6} h^2 \tag{5.5}$$

Annahmen: $|e_{-1}|, |e_{+1}| \leq \varepsilon$ und $\left|f^{(3)}(\xi)\right| \leq M$
Damit $\left| f'(x_0) - \frac{\tilde{f}(x_0+h) - \tilde{f}(x_0-h)}{2h} \right| \leq \frac{\varepsilon}{h} + \frac{h^2 M}{6}$
$\Rightarrow h_{\text{opt}} = \sqrt[3]{\frac{3\xi}{M}}$



(a) Polynom 1. Ordnung (b) Polynom 2. Ordnung (c) mehrere Polynome

Abbildung 5.1: Unterschiedliche Methoden eine Funktion bei einer numerischen Integration anzunähern

5.2 Numerische Integration

Alternativer Begriff: „Numerische Quadratur“ (bestimmte Integrale)

$$J = \int_a^b f(x)\,\mathrm{d}x \cong \int_a^b P_h(x)\,\mathrm{d}x \tag{5.6}$$
$$P_n(x) = a_0 + a_1 x + \ldots + a_n x^n$$

5.2.1 Trapezregel

Lineares Lagrange Polynom

$$P_1(x) = \frac{x - x_1}{x_0 - x_1} \cdot f(x_0) + \frac{x - x_0}{x_1 - x_0} \cdot f(x_1)$$
$$\int_a^b = \int_{x_0}^{x_1} P_1(x)\,\mathrm{d}x + \underbrace{\frac{1}{2}\int_{x_0}^{x_1} f''(\xi(x)) \cdot (x - x_0) \cdot (x - x_1)\,\mathrm{d}x}_{\text{Fehlerterm } \epsilon}$$

Mittelwert Integralrechnung

$$\epsilon = \frac{1}{2} f''(\xi) \int_{x_0}^{x_1} (x - x_0) \cdot (x - x_1)\,\mathrm{d}x = \frac{1}{2} f''(\xi) \left[\frac{x^3}{3} - \frac{(x_1 + x_0)}{2} x^2 + x_0 x_1 x\right]_{x_0}^{x_1} = \frac{-h^3}{12} f''(\xi) \tag{5.7}$$

mit $h = b - a = x_1 - x_0$

Also:

$$\int_a^b f(x)\,\mathrm{d}x = \left[\frac{(x-x_1)^2}{2(x_0-x_1)}\cdot f(x_0) + \frac{(x-x_0)^2}{2(x_1-x_0)}\cdot f(x_1)\right]_{x_0}^{x_1} - \frac{h^3}{12}f''(\xi) = \frac{h}{2}\left[f(x_0)+f(x_1)\right] - \frac{h^3}{12}f''(\xi) \tag{5.8}$$

Exakt, sofern $f(x)$ höchstens linear ist

5.2.2 Simpson Regel

Quadratisches Lagrange Polynom

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x &= \int_{x_0}^{x_2} \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}f(x_1) + \frac{(x-x_0)(x-x_1)}{(x_2-x_1)(x_2-x_1)}f(x_2)\,\mathrm{d}x \\ &\quad + \int_{x_0}^{x_2} \frac{(x-x_0)(x-x_1)(x-x_2)}{6}f^{(3)}(\xi(x))\,\mathrm{d}x \\ &= \frac{h}{3}\left[f(x_0)+4f(x_1)+f(x_2)\right] - \frac{h^5}{90}f^{(4)}(\xi)\end{aligned} \tag{5.9}$$

mit $h = x_1 - x_0 = x_2 - x_1$

Simpson $\frac{3}{8}$ Regel

$$\int_a^b f(x)\,\mathrm{d}x = \frac{3h}{8}\left[f(x_0)+3f(x_1)+3f(x_2)+f(x_3)\right] - \frac{h^5}{6480}f^{(4)}(\xi) \tag{5.10}$$

- benötigt $3m$ Segmente bzw. $3m+1$ Stützpunkte

- Grad der Genauigkeit/„Degree of precision“
  Größte ganze Zahl n, so dass eine Integrationsformel für $x^k,\ k = 0, 1, \ldots, n$ exakte Resultate liefert.

Composite Simpson $\frac{1}{3}$

Teile $[a, b]$ in n Teilintervalle auf. Wende die Simpson Regel auf jedes Paar nacheinanderfolgender Teilintervalle an.
$h = \frac{b-a}{n}$; $x_j = a + j \cdot h$, $j = 0, 1, \ldots, n$

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x &= \sum_{j=1}^{\frac{n}{2}} \int_{x_{2j-2}}^{x_{2j}} f(x)\,\mathrm{d}x \\ &= \sum_{j=1}^{\frac{n}{2}} \frac{h}{3}\left[f(x_{2j-2}) + 4f(x_{2j-1}) + f(x_{2j})\right] \frac{h^5}{90} f^{(4)}(\xi_j) \\ &= \frac{h}{3}\left[f(x_0) + 2\sum_{j=1}^{\frac{n}{2}-1} f(x_{2j}) + 4\sum_{j=1}^{\frac{n}{2}} f(x_{2j-1} + f(x_n)\right] - \frac{h^5}{90}\sum_{j=1}^{\frac{n}{2}} f^{(4)}(\xi_0)\end{aligned} \tag{5.11}$$

mit $x_{2j-2} < \xi_j y x_{2j}, f$ auf $[a, b]$ 4-mal stetig differenzierbar.

Fehler

$E(f) = \frac{-h^5}{90} \sum_{j=1}^{\frac{n}{2}} f^{(4)}(\xi_j)$, mit $x_{2j-2} < \xi_j y x_{2j}, j = 1, 2, \ldots, \frac{n}{2}$

Extremwertsatz

: $f^{(4)}$ nimmt Max und Min auf $[a, b]$ an, also

$$\begin{aligned}\min_{x\in[a,b]} f^{(4)}(x) \le f^{(4)}(\xi_j) &\le \max_{x\in[a,b]} f^{(4)}(x) \\ \frac{n}{2}\min_{x\in[a,b]} f^{(4)}(x) \le \sum_{j=1}^{\frac{n}{2}} f^{(4)}(\xi_j) &\le \frac{n}{2}\max_{x\in[a,b]} f^{(4)}(x) \\ \min_{x\in[a,b]} f^{(4)}(x) \le \frac{n}{2}\sum_{j=1}^{\frac{n}{2}} f^{(4)}(\xi_j) &\le \max_{x\in[a,b]} f^{(4)}(x)\end{aligned} \tag{5.12}$$

Zwischenwertsatz

Es existiert ein $\mu \in (a, b)$, so dass

$$f^{(4)}(\mu) = \frac{2}{n}\sum_{j=1}^{\frac{n}{2}} f^{(4)}(\xi_j) \tag{5.13}$$

Damit

$$E(f) = \frac{-h^5}{90}\sum_{j=1}^{\frac{n}{2}} f^{(4)}(\xi_j) = \frac{-h^5}{180} n \cdot f^{(4)}(\mu) \overset{h=\frac{b-a}{n}}{=} \frac{-(b-a)}{180} h^4 \cdot f^{(4)}(\mu) \tag{5.14}$$

Also, mit $\mu \in (a, b)$:

$$\int_a^b f(x)\,\mathrm{d}x = \frac{h}{3}\left[f(a) + 2\sum_{j=1}^{n/2-1} f(x_{2j} + 4\sum_{j=1}^{n/2} f(x_{2j-1}) + f(b)\right] - \frac{b-a}{180}h^4 f^{(4)}(\mu) \quad (5.15)$$

Wie man sich nun vielleicht vorstellen kann gibt es unendlich viele mögliche numerische Integrationsverfahren. Das Composite Simpson-Verfahren ist allerdings ein häufig eingesetztes allgemeines Verfahren.

5.2.3 Genauigkeit

Zur Bestimmung der Genauigkeit gibt es zwei Ansätze:

1. Gegeben sei ein Verfahren, eine Funktion f, ein Interval $[a, b]$ und eine Stützstellenanzahl n und gesucht sei der Fehler
2. Gegeben sei eine Funktion f, ein Interval $[a, b]$ und ein gewünschter Fehler und gesucht sei ein Verfahren und eine Stützstellenanzahl n

Der Rundungsfehler wird definiert über

$$f(x_i) = \underbrace{\tilde{f}(x_i)}_{\text{gerundeter Wert}} + \underbrace{e_i}_{\text{Rundungsfehler}} \quad (5.16)$$

Damit erhält man dann einen akkumulierten Fehler $e(h)$ bei Composite-Simpson von

$$e(h) = \left|\frac{h}{3}\left[e_0 + 2\sum_{j=1}^{n/2-1} e_{2j} + 4\sum_{j=1}^{n/2} e_{2j-1} + e_n\right]\right| \leq \frac{h}{3}\left[|e_0| + 2\sum_{j=1}^{n/2-1} |e_{2j}| + 4\sum_{j=1}^{n/2} |e_{2j-1}| + |e_n|\right] \quad (5.17)$$

Seine alle Rundungsfehler $|e_i| \leq \varepsilon$, dann gilt

$$e(h) \leq \frac{h}{3}\left[\varepsilon + 2(\frac{n}{2} - 1)\varepsilon + 4\frac{n}{2}\varepsilon + \varepsilon\right] = nh\varepsilon = (b-a)\varepsilon \neq f(h, n) \quad (5.18)$$

Erstaunlich ist somit, dass der Rundungsfehler nicht wächst, wenn der zu integrierende Bereich in mehr Teilintervalle unterteilt wird.

6 Numerische Lösung von Differentialgleichungen

Bsp: Einschwinganalyse (TRansiente Analyse)

6.1 Analytische Lösung mittels Laplace-Transformation

$$
\begin{aligned}
C \cdot \dot{y}(t) + \frac{1}{R} \cdot y(t) &= i_0(t) \\
\dot{y}(t) + \underbrace{\frac{1}{RC}}_{-p_\infty} \cdot y(t) &= \frac{1}{RC} R \cdot i_0(t) \\
\dot{y}(t) - p_\infty y(t) &= -p_\infty R \cdot i_0(t) \\
\circ\!\!-\!\!\overset{\mathcal{L}}{\bullet} \text{ Laplace} & \\
p \cdot Y(p) - \underbrace{y(+0)}_{=0} - p_\infty Y(p) &= -p_\infty R \frac{1}{p} \cdot I_0 \\
Y(p) &= \frac{-p_\infty R \cdot I_0}{(p - p_\infty)} = \frac{A}{p} + \frac{B}{p - p_\infty}
\end{aligned}
\tag{6.1}
$$

$$
A = \lim_{p \to 0} p \cdot Y(p) = R \cdot I_0 \tag{6.2}
$$

$$
B = \lim_{p \to p_\infty} (p - p_\infty) \cdot Y(p) = -R \cdot I_0 \tag{6.3}
$$

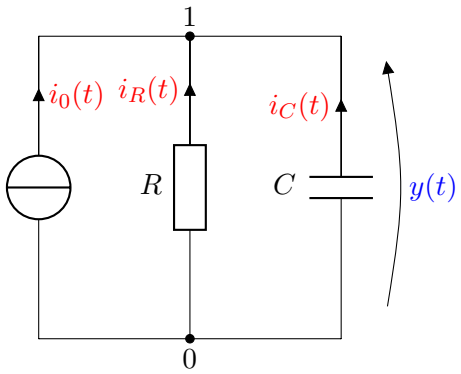


Abbildung 6.1: Beispielschaltung Transiente Analyse

$$y(t) = A + Be^{p_\infty t} = R \cdot I_0 \cdot (1 - e^{p_\infty t}) \quad \overset{\mathcal{L}}{\circ\!\!-\!\!\bullet} \tag{6.4}$$

6.2 Numerische Lösung mittels explizite Euler-Methode (linear Z-Transformation)

Diskretisierung der Zeit

$$y(\nu) \hat{=} y(\nu \cdot \Delta t) = y(t) \tag{6.5}$$

Diskretisierung der DGL

$$\dot{y}(\nu) \approx p_\infty \cdot y(\nu) - p_\infty - p_\infty R \cdot i_0(\nu) \tag{6.6}$$

Differenzengleichung (explizite Euler-Methode)

$$y(\nu + 1) \approx y(\nu) + \Delta t \cdot \dot{y}(\nu) \tag{6.7}$$

$$y(\nu + 1) \approx y(\nu) + \Delta t \cdot [p_\infty y(\nu) - p_\infty R \cdots i_0(\nu)] \quad , \nu = 0, 1, 2, \ldots \tag{6.8}$$

$$\hat{y}(\nu + 1) = (1 + p_\infty \Delta t) \cdot y(\nu) - \Delta t p_\infty R i_0(\nu) \tag{6.9}$$

sukzessive Berechnung ausgehend von $y(0),\ \nu = 1, 2, 3, \ldots$

Bei linearen Schaltungen (linearen Differenzengleichungen) ist geschlossene numerische Lösung möglich durch $\overset{z}{\circ\!\!-\!\!\bullet}$ von Gleichung 6.9

$$z \cdot \dot{Y}(z) - \underbrace{z \cdot y(0)}_{0} - (1 - p_\infty \Delta t) \cdot Y(z) = -p_\infty R \Delta t \frac{z}{z - 1} I_0 \tag{6.10}$$

$$Y(z) = \frac{-p_\infty \Delta t R}{z-(1+p_\infty \Delta t)} \frac{zI_0}{z-1} = \frac{zRI_0}{z-1} \frac{zRI_0}{z-(1+p_\infty \Delta t)} \tag{6.11}$$

$\circ\!\!-\!\!\bullet^{\,z}$

$$\hat{y}(\nu) = [1-(1+p_\infty \Delta t)^\nu] \cdot RI_0 \quad , \nu = 0,1,2,\ldots$$

Zahlenbeispiel: $I_0 R = 1;\ p_\infty = -1,\ t = 1$
exakte Lösung: $y(1) = 1 - e^{-1} = 0{,}632\,121$
numerische Lösung: $y(\nu) = 1 - (1 - \Delta t^\nu \quad , \nu \cdot \Delta t = 1$

Δt	ν	$\dot{y}(1)$	$\varepsilon^{(A)} = \lvert\hat{y}(1) - y(1)\rvert$
0,1	10	0,651 322	0,019 201
0,05	20	0,641 514	0,009 393
0,025	40	0,636 768	0,004 647

Fehler $\varepsilon^{(A)} \sim \Delta t$

Stabilität:

Für $|1 + p_\infty \cdot \Delta t| = |1 - \Delta t| < 1 \quad \Rightarrow \lim_{x\to\infty} \hat{y}(\nu) = 1$

$0 < \Delta t < 2$ Algorithmus stabil

Für $|1 - \Delta t| > 1 \quad \Rightarrow \lim_{v\to\infty} \hat{y}(\nu) \to \infty$

$\Delta t > 2$ Algorithmus instabil

6.3 Einfache numerische Integrationsverfahren

$$y(0) = 0, \quad y(t) = \int_0^t \dot{y}(\tau)\,\mathrm{d}\tau$$

$$y(t+\Delta t) = \int_0^t \dot{y}(\tau)\,\mathrm{d}\tau + \int_t^{t+\Delta t} \dot{y}(\tau)\,\mathrm{d}\tau = y(t) + \int_t^{t+\Delta t} \dot{y}(\tau)\,\mathrm{d}\tau \tag{6.12}$$

$$y(\nu+1) \approx y(\nu) + \tilde{\Delta} y = y(\nu) + \frac{\Delta t \cdot \tilde{\Delta} y}{\Delta t}$$

Expliziter Euler: $y(\nu+1) \approx y(\nu) + \Delta t \cdot \dot{y}(\nu)$ (FE, Forward Euler)

Impliziter Euler: $y(\nu+1) \approx y(\nu) + \Delta t \cdot \dot{y}(\nu+1)$ (BE, Backward Euler)

Trapez-Methode: $y(\nu+1) \approx y(\nu) + \frac{\Delta t}{2} \cdot (\dot{y}(\nu) + \dot{y}(\nu+1)$ (TR, Trapezvidal)

6.4 Eigenschaften numerischer Integrationsverfahren

- explizit/implizit: wird Wert aus aktuellem Zeitschritt $(\nu+1)$ verwendet?
- Schrittzahl: Anzahl der verwendeten vergangenen Zeitschritte
- Genauigkeit: $\varepsilon^{(A)}(\nu \cdot \Delta t) = \hat{x}(\nu) - x(\nu) \approx (\Delta t)^k$
  (für $|p_\infty \Delta t| << 1$) Fehler der Ordnung k
- Stabilität:
  - für $\mathrm{Re}\{p_\infty\} < 0$
  - für $\Delta t \to \infty$ (“asymptotische Stabilität“)

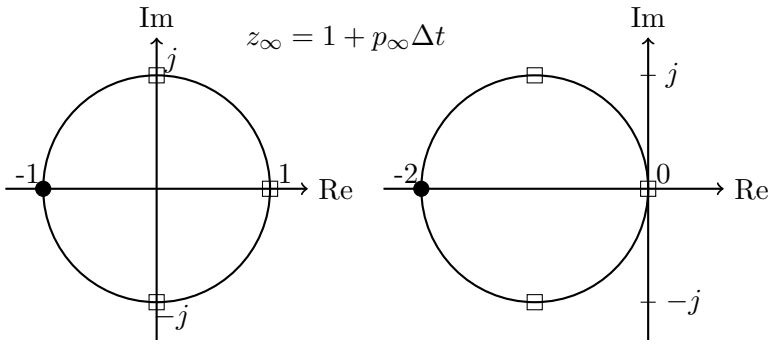


Abbildung 6.2: Stabilitätsbereich für den expliziten Euler

Test-DGL:	$\dot{x}(t) = p_\infty \cdot x(t)$	p_∞: Eigenwerte des Systems
	exakte Lösung:	$x(t) = x(0) \cdot e^{p_\infty t}$
		$x(\nu) = x(0) \cdot e^{p_\infty \cdot \nu \cdot \Delta t},\ x(0) \neq 0$
	numerische Lösung:	$\hat{x}(\nu) = \sum\{\hat{f}(\nu)\}$
	Fehler:	$\varepsilon^{(A)}(\Delta t) = \hat{x}(\nu) - x(\nu) = \hat{x}(\nu) - x(0) \cdot e^{p_\infty \nu \Delta t}$

6.5 Expliziter Euler Eigenschaften

Differenzengleichung: $\hat{x}(\nu + 1) = \hat{x}(\nu) + \Delta t \cdot \hat{f}(\nu),\ f(t) = \dot{x}(t)$
Einschritt-Verfahren

Test-DGL:	$f(t) = \dot{x}(t) = p_\infty \cdot x(t)$
	$x(0) \neq 0$
Differenzengleichung:	$\dot{x}(\nu + 1) = \hat{x}(\nu) + \Delta t p_\infty \cdot \hat{x}(\nu) = (1 + p_\infty \cdot \Delta t) \cdot \dot{x}(\nu)$
numerische Lösung:	$\hat{x}(\nu) = (1 + p_\infty \Delta t) \cdot x(\nu)$
Genauigkeit:	$\varepsilon^{(A)} = \hat{x}(\nu) - x(\nu) = \Delta t \cdot \frac{\overbrace{\nu \Delta t}^{=t}}{2} \cdot p_\infty^2 \cdot x(0) \cdot e^{p_\infty \overbrace{\nu \Delta t}^{=t}}$
Stabilität:	$\lim_{t \to \infty} x(t) \to 0$ für $\mathrm{Re}\{p_\infty\} < 0$
	$\lim_{\nu \to \infty} \hat{x}(\nu) = \lim_{\nu \to \infty} (1 + p_\infty \Delta t)^\nu x(0) \to 0$ falls $\lvert 1 + p_\infty \Delta t \rvert < 1 \rvert$
Ordnung	$k = 1$
Asymptotische Stabilität ($\Delta t \to \infty$):	$\hat{x}(\nu + 1) \approx p_\infty \Delta t \hat{x}(\nu) \Rightarrow$ instabiles Verhalten!

Somit ist die Wahl von Δt aus Stabilitätsgründen eingeschränkt. Bei zu großem Δt liegt das Produkt $p_\infty \Delta t$ nicht mehr im Stabilitätsbereicht. Praktisch beim expliziten Euler ist hingegen, dass es keine Stabilität in der rechten Halbebene gibt, es werden also instabile System als instabil detektiert.

Im folgenden noch einmal die Eigenschaften der Verfahren kurz zusammengefasst:

	explizit/implizit	Schritte	Ordnung	stabil für $\mathrm{Re}\{p_\infty\} < 0$	asymptotisch stabil
expliziter Euler	explizit	1	1	nein	nein
impliziter Euler	implizit	1	1	ja	ja
Trapez	implizit	1	2	ja	nein
Gear	implizit	2	2	ja	ja

Zudem haben Einschrittverfahren Vorteile beim Starten, mit Mehrschrittverfahren kann nicht direkt gestartet werden.

6.6 Taylor-Verfahren höherer Ordnung

Die eher im technischen Bereich gängige Schreibweise

$$y(\nu+1) = y(\nu) + \Delta t T^{(n)}(\nu, y(\nu)) \tag{6.13}$$

ist äquivalent zur in der Mathematik üblichen Variante

$$y_{i+1} = y_i + h T^{(n)}(t_i, y_i) \tag{6.14}$$

mit

$$T^{(n)}(t_i, y_i) = f(t_i, y_i) + \frac{h}{2} f'(t_i, y_i) + \cdots + \frac{h^{n-1}}{n!} f^{(n-1)}(t_i, y_i) \tag{6.15}$$

Dadurch erhält man einer bessere Genauigkeit, genauer gesagt: $\epsilon O(h^n)$. Nachteilig kann sich allerdings auswirken, dass höhere Ableitungen der Funktion benötigt werden.

6.7 Prädiktor-Korrektor-Ansätze

Die Grundidee dahinter ist ein explizites Verfahren zur Vorhersage des Wertes y_{i+1} zu nutzen, und diesen Werte nachträglich mithilfe eines weiteren Verfahrens zu verbessern. Dieser Ansatz kann auch iterativ verwendet werden.

Ein Beispiel für so einen Prädiktor-Korrektor-Ansatz ist das Verfahren von Heun, welches auf einem expliziten Euler basiert. Der explizite Euler liefert

$$y^0_{i+1} = y_i + h \cdot f(t_i, y_i) \tag{6.16}$$

und daraus berechnet Heun ein verbessertes

$$y^1_{i+1} = f(t_{i+1}, y^0_{i+1}). \tag{6.17}$$

Der iterativer Ansatz dazu wäre dann

$$y^0_{i+1} = y^m_i + h \cdot f(t_i, y_i) \tag{6.18}$$

$$y^j_{i+1} = y^m_i + h \cdot \frac{f(t, y^m_i) + f\left(t_{i+1}, y^{j-1}_{i+1}\right)}{2} \tag{6.19}$$

6.8 Adams-Bashforth/Adams-Moulton

Die Grundidee hinter diesen Verfahren beruht darauf f durch ein Interpolationspolynom zu ersetzen.

Ausgangspunkt ist die Differentialgleichung

$$y' = f(t, y) \tag{6.20}$$

wobei $a \leq t \leq b$ und $y(a) = \alpha$.

Adams-Bashforth und Adams-Moulton sind m-schrittige Mehrschrittverfahren, in denen y_{i+1} ersetzt wird durch

$$y_{i+1} = a_{m-1}y_i + a_{m-2}y_{i-1} + \cdots + a_0 y_{i+1-m} + h\left[b_m f(t_{i+1}, y_{i+1}) + b_{m-1}f(t_i, y_i) + \cdots + b_0 f(t_{i+1-m}, y_{i+1-m})\right] \tag{6.21}$$

mit $i = m-1, m, \ldots N-1$, $h = \frac{b-a}{N}$. Falls $b_n = 0$ spricht man von einem expliziten Verfahren, ansonsten von einem impliziten.

Herleitung

$$y(t_{i+1}) = y_{i+1} \tag{6.22}$$

$$y_{i-1} - y_i = \int_{t_i}^{t_{i+1}} y'(t)\,\mathrm{d}t = \int_{t_i}^{t_{i+1}} f(t_i, y(t))\,\mathrm{d}t \tag{6.23}$$

$$y_{i+1} = y_i + \int_{t_i}^{t_{i+1}} f(t_i, y(t))\,\mathrm{d}t \tag{6.24}$$

Ersetze $f(t_i, y(t))$ durch Interpolationspolynom $P(t)$

$$y_{i+1} \approx y_i + \int_{t_i}^{t_{i+1}} P(t)\,\mathrm{d}t \tag{6.25}$$

Für das Interpolationspolynom bieten sich an

- Taylor
- Lagrange
- Newton'sche Rückwärtsdifferenzen

6.8.1 Adams-Bashforth (explizit)

Die explizite Einschrittvariante ist äquivalent zum expliziten Euler:

$$y_{i+1} = y_i + hf(t_i, y_i) \tag{6.26}$$

Die Variante mit zwei Schritten lautet

$$y_{i+1} = y_i + \frac{h}{2}\left[3f(t_i, y_i) - f(t_{i-1}, y_{i-1}\right] \tag{6.27}$$

und hat eine quadratisches Fehlerverhalten.

6.8.2 Adams-Moulton (implizit)

Die Einschrittvariante hiervon ist für $m = 0$ der implizite Euler, für $m = 1$ das Trapezverfahren. Wenn zwei Schritten verwendet werden sieht das Verfahren wie folgt aus

$$y_{i+1} = y_i + \frac{h}{12}\left[5f(t_{i+1}, y_{i+1} + 8f(t_i, y_i) - f(t_{i-1}, y_{i-1})\right] \tag{6.28}$$

und besitzt ein kubisches Fehlerverhalten.

7 Ausgleichsrechnung

Auch bekannt unter der Bezeichnung Approximation oder Least Square **Ziele der Approximation:** („curve fitting“)

- Bestgeeignete Funktion eines bestimmten Typs, um gegebenen Daten anzunähern
- Für gegebene Funktionen „bestmögliche“, einfachere Funktionen zu finden.

„Bestmögliche“ Funktion zur Annäherung gegebener Stützpunkte

- Gegeben: Wertepaare $(x_i, y_i),\ i = 1, \ldots, m$
- Annahme: Gerade beschriebt Wertepaare $y = f(x) = ax + b$

Ansätze:

- $E = \sum_{i=1}^{m} l_i = \sum_{i=1}^{m} [y_i - (ax_i + b)]$
- $E = \sum_{i=1}^{m} |l_i| = \sum_{i=1}^{m} |[y_i - (ax_i + b)]|$
- Minmax: $\min E_\infty = \min \max_{i=1,\ldots,m} \{|y_i - (ax_i + b)|\}$

7.1 Linear Least Squares

Minimiere: $E_2(a, b) = \sum_{i=1}^{m} (y_i - (ax_i + b))^2$
Bestimmung von a, b: $\frac{\partial E_2}{\partial a} = 0,\ \frac{\partial E_2}{\partial b} = 0$
$\Rightarrow$ Normalengleichung:

- $a = ...$
- $b = ...$

7.2 Polynominale Least Squares

Approximation durch $P_n(x) = a_n x^n + \ldots + a_1 x + a_0$

$$E_2(a_0, a_1, \ldots, a_n)) = \sum_{i=1}^{m} (y_i - P(x_i)^2) \tag{7.1}$$

Minimierung $\frac{\partial E_2}{\partial a_i} = 0 \Rightarrow n + 1$ Normalengleichungen in den $n + 1$ Unbekannten $a_0, \ldots, a_n$

7.3 Linear Least Squares - Sichtweise Lineare Algebra

$$\boldsymbol{A}_{<m\times n}\cdot\vec{x}_n=\vec{b}_n\quad,m>n \tag{7.2}$$

Beispiel:

- $m=2,n=1\quad\begin{bmatrix}a_1\\a_2\end{bmatrix}\cdot x=\begin{bmatrix}b_1\\b_2\end{bmatrix}$
- $m=3,n=2\quad\begin{bmatrix}a_{11}&a_{12}\\a_{21}&a_{22}\\a_{31}&a_{32}\end{bmatrix}\cdot\begin{bmatrix}x_1\\x_2\end{bmatrix}=\begin{bmatrix}b_1\\b_2\\b_3\end{bmatrix}$

Fehler je Zeile: $e_i=b_i-\vec{a}_i\cdot\vec{x}$
Vektor der Fehler: $\vec{e}=\vec{b}-\boldsymbol{A}\cdot\vec{x}$
Summe der Fehlerquadrate: $E_2=\|e\|^2=\left\|\vec{b}-\boldsymbol{A}\vec{x}\right\|^2$
Gesucht $\hat{\vec{x}}$, so dass E_2 minimiert wird.

$$\boldsymbol{A}\hat{\vec{x}}=\vec{p}\quad;\quad\vec{e}=\vec{b}-\vec{p} \tag{7.3}$$

$\vec{p}$ ist die Projektion von $\vec{b}$ in den Spaltenraum von col($\boldsymbol{A}$)
$\vec{e}$ steht senkrecht auf den Spaltenraum col($\boldsymbol{A}$)

$$\begin{aligned}E(x)&=\left\|\vec{b}-\boldsymbol{A}\vec{x}\right\|^2=\left\|\boldsymbol{A}\vec{x}-\vec{b}\right\|\\&=\vec{x}^T\boldsymbol{A}^T\boldsymbol{A}\vec{x}-(\boldsymbol{A}\vec{x})^T\vec{b}-\vec{b}^T\boldsymbol{A}\vec{x}-\vec{b}^T\vec{b}\\&=\vec{x}^T\underbrace{\boldsymbol{A}^T\boldsymbol{A}}_{\boldsymbol{K}}\vec{x}-2\vec{x}^T\underbrace{\boldsymbol{A}^T\vec{b}}_{\vec{f}}-\vec{b}^T\vec{b}\\&=\vec{x}^T\boldsymbol{K}\vec{x}-2\vec{x}^T\vec{f}+\vec{b}^T\vec{b}\quad\text{zu minimieren}\\&=(\vec{x}-\boldsymbol{K}^{-1}\vec{f})^T\boldsymbol{K}(\vec{x}-\boldsymbol{K}^{-1}\vec{f})-\vec{f}^T\boldsymbol{K}^{-1}\vec{f}+\vec{b}^T\vec{b}\end{aligned} \tag{7.4}$$

Ausdruck kann minimal 0 werden für $\hat{\vec{x}}=\boldsymbol{K}^{-1}\vec{f}\Rightarrow$ dort also Minimum vom $E(\vec{x})$

$$\begin{aligned}E_{\min}=E(\hat{\vec{x}})&=(\vec{b}-\boldsymbol{A}\hat{\vec{x}})^T\cdot(\vec{b}-\boldsymbol{A}\hat{\vec{x}})\\&=\vec{b}^T\vec{b}-\vec{f}^T\boldsymbol{K}^{-1}\vec{f}\\&=\vec{b}^T\vec{b}-\vec{b}^T\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\vec{b}\end{aligned} \tag{7.5}$$

$\vec{e}=\vec{b}-\boldsymbol{A}\vec{x}$ steht senkrecht auf col($\boldsymbol{A}$) $\Rightarrow\boldsymbol{A}^T\vec{e}=\vec{0}$

$$\Rightarrow\boldsymbol{A}^T(\vec{b}-\boldsymbol{A}\hat{\vec{x}})=0\Rightarrow\boldsymbol{A}^T\boldsymbol{A}\hat{\vec{x}}=\boldsymbol{A}^T\vec{b} \tag{7.6}$$

$\underbrace{\boldsymbol{A}^T\boldsymbol{A}\hat{\vec{x}}=\boldsymbol{A}^T\vec{b}}_{\text{quadratisch, symmetrisch}}$ LGS $\Rightarrow$ Bestimmung von $\hat{\vec{x}}$

Lösungsverfahren:

- Gauß LU: OK, aber cond $\left(\boldsymbol{A}^T\boldsymbol{A}\right)=\text{cond}\left(\boldsymbol{A}\right)^2\Rightarrow$ Problematisch, wenn $\boldsymbol{A}$ schlecht konditioniert
- Orthogonalzerlegung

7.4 QR-Zerlegung

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R} \text{ mit } \begin{cases} \boldsymbol{A} & \in \mathbb{R}^{n\times n} \\ \boldsymbol{Q} & \in \mathbb{R}^{n\times n} \text{ orthogonale Matrix}(\boldsymbol{Q}^T = \boldsymbol{Q}^{-1}) \\ \boldsymbol{R} & \in \mathbb{R}^{n\times n} \text{ rechte obere Dreiecksmatrix} \end{cases}$$

Damit:

$$\begin{aligned} \boldsymbol{A}\vec{x} &= \vec{b} \\ \boldsymbol{A}^T\boldsymbol{A}\vec{x} &= \boldsymbol{A}^T\vec{b} \\ (\boldsymbol{Q}\boldsymbol{R})^T\boldsymbol{Q}\boldsymbol{R}\vec{x} &= (\boldsymbol{Q}\boldsymbol{R})^T\vec{b} \\ \boldsymbol{R}^T\underbrace{\boldsymbol{Q}^T\boldsymbol{Q}}_{\boldsymbol{I}}\boldsymbol{R}\vec{x} &= \boldsymbol{R}^T\boldsymbol{Q}^T\vec{b} \\ \boldsymbol{R}^T\boldsymbol{R}\vec{x} &= \boldsymbol{R}^T\boldsymbol{Q}^T\vec{b} \quad |\cdot \boldsymbol{R}^{-1^{-1}} \\ \boldsymbol{R}\vec{x} &= \boldsymbol{Q}^T\vec{b} \Rightarrow \text{simple Rücksubstitution} \end{aligned} \tag{7.7}$$

OK, wie bestimme ich nun $\boldsymbol{Q}, \boldsymbol{R}$?

3 wesentliche Ansätze:

- Gram-Schmidt
- Hausholder-Transformation
- Givens-Rotation

7.4.1 Givens-Rotation

auch bekannt unter der Bezeichung Jacobi-Rotation
Eine Rotation ist gegeben durch

$$\boldsymbol{G}(l,k,\Theta) = \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & & & & & & 0 \\ \vdots & & c & & & -s & & \vdots \\ \vdots & & & & & & & \vdots \\ \vdots & & -s & & & c & & \vdots \\ \vdots & & & & & & 1 & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & 0 & 1 \end{bmatrix} \tag{7.8}$$

wobei $c = \cos\Theta$, $s = \sin\Theta$. Komponentenweise $\boldsymbol{G}(l,k,\Theta) = (g_{i,j}(l,k,\Theta))$ dargestellt ergibt das

$$g_{i,j} = \begin{cases} \cos\Theta & \text{für } i = l, j = l \vee i = k, j = k \\ \sin\Theta & \text{für } i = l, j = k \\ -\sin\Theta & \text{für } i = k, j = l \\ 1 & \text{für } i = j \text{ außer } i, j = l, k \\ 0 & \text{sonst} \end{cases} \tag{7.9}$$

Damit bedeutet $\boldsymbol{G} \cdot \vec{x}$ bzw. $\boldsymbol{G}^T \cdot \vec{x}$ eine Drehung des Vektors $\vec{x}$ um $\pm\Theta$ in der (l,k)-Ebene. Die Hauptanwendung hierfür ist ein iteratives Vorgehen um Nulleinträge in Matrizen oder Vektoren zu erreichen. Desweiteren ist $\boldsymbol{G}\boldsymbol{G}^T = \boldsymbol{I}$ und somit die Rotationsmatrix $\boldsymbol{G}$ orthogonal.

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} c & -s \\ s & c \end{pmatrix} = \begin{pmatrix} c^2+s^2 & 0 \\ 0 & c^2+s^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{7.10}$$

Damit gilt

$$\boldsymbol{G}_r \boldsymbol{G}_{r-1} \cdot \ldots \cdot \boldsymbol{G}_2 \boldsymbol{G}_1 \boldsymbol{A} = \boldsymbol{R} \tag{7.11}$$

$$\boldsymbol{A} = \boldsymbol{G}^T \boldsymbol{R} = \boldsymbol{Q}^T \boldsymbol{R} \tag{7.12}$$

Die Frage ist nun, wie Θ gewählt werden muss. Hierfür genügt die Betrachtung einer zweidimensionalen Struktur

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} x_l \\ x_k \end{pmatrix} = \begin{pmatrix} r \\ 0 \end{pmatrix} \tag{7.13}$$

Daraus kann gefolgert werden, dass

$$s = c\frac{x_k}{x_l} \tag{7.14}$$

und aus

$$c^2 + s^2 = 1 \tag{7.15}$$

$$c = \sqrt{1-s^2} = \sqrt{\frac{x_l^2 - c^2 x_k^2}{x_l^2}} \tag{7.16}$$

$$c^2 x_l^2 = x_l^2 - c^2 - x_k^2 \tag{7.17}$$

$$c = \pm\frac{x_l}{\sqrt{x_l^2 + x_k^2}} \tag{7.18}$$

Das ganze an dem Beispiel

$$\boldsymbol{A} = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 3 \end{pmatrix} \tag{7.19}$$

führt über

$$\boldsymbol{G} = \begin{pmatrix} c & s & 0 \\ -s & c & 1 \\ 0 & 0 & 1 \end{pmatrix} \tag{7.20}$$

zu

$$\boldsymbol{G}\boldsymbol{A} = \begin{pmatrix} \sqrt{10} & 6/\sqrt{10} & 1/\sqrt{10} \\ 0 & 3/\sqrt{10} & 3/\sqrt{10} \\ 0 & 1 & 3 \end{pmatrix} \tag{7.21}$$

Abbildungsverzeichnis