Otmar Scherzer
Computational Science Center
Universität Wien
Nordbergstr. 15
A-1090 Wien

Numerik II

Vorlesungsskriptum WS 2012/2013

This lecture notes are based on the excellent course book of M. Hanke-Bourgeois [8].

Inhaltsverzeichnis

Kapitel 1

Splines

Historisch wurde zunächst die Polynominterpolation zur Approximation skalarer Funktionen verwendet. Diese Art der Interpolation führt aber zu starken Oszillationen. Die Alternative ist die gesuchte Funktion durch stückweise zusammengesetzte Funktionen zu approximieren, die sogenannten *Splines*.

1.1 Treppenfunktionen

Gegeben sei ein reelles Intervall $[a, b]$ und ein Gitter

$$\Delta = \{a = x_0 < x_1 < \ldots < x_l = b\} \ . \tag{1.1}$$

Wir definieren

$$h := \max_{i=1,\ldots,l} h_i \, , \quad h_i = x_i - x_{i-1} \ .$$

Unter einer *Treppenfunktion* versteht man eine von rechts stetige Funktion s mit der Eigenschaft

$$s(x) = s_i \, , \quad x_{i-1} \leq x < x_i \, , \quad i = 1, \ldots, l \ .$$

Diese Treppenfunktionen bilden einen linearen Raum $S_{0,\Delta}$ der dimension l. Als Basisfunktion verwendet man üblicherweise die characteristischen Funktionen χ_i der l Teilintervalle. Es gilt somit

$$s(x) = \sum_{i=1}^{l} s_i \chi_i(x) \ .$$

Wir fassen einige trivialle Eigenschaften dieser Treppenfunktionen zusammen:

Bemerkung 1.1. 1. $S_{0,\Delta}$ ist ein Teilraum von $L^2[a,b]$ der quadratisch integrierbaren Funktionen auf dem Intervall $[a,b]$). Alle Funktionen sind der Einfachheit halber reell. Bei komplexen Funktionen approximiert man Real- und Imaginärteil getrennt.

2. Die Basisfunktionen $\chi_i/\sqrt{h_i}$ bilden eine Orthonormalbasis von $L^2[a,b]$.

3. Die beste Approximation von einer Funktion in $L^2[a,b]$ kann somit mit den Vorkenntnissen der Numerik I bestimmt werden, und wird im folgenden Satz ausgeführt.

Satz 1.2. *Sei $f \in L^2[a,b]$. Dann ist $s = \sum_{i=1}^{l} s_i\chi_i$ mit*

$$s_i = \frac{1}{h_i}\int_{x_{i-1}}^{x_i} f(x)\,dx\,, \quad i = 1,\ldots,l \tag{1.2}$$

bezüglich der $L^2[a,b]$-Norm die beste Approximation in $S_{0,\Delta}$ von f.
Ist $f \in H^1[a,b]$, dann gilt

$$\|f - s\|_{L^2[a,b]} \le h\|f'\|_{L^2[a,b]}\,. \tag{1.3}$$

Beweis. Wie wir schon in Numerik 1 gesehen haben sind die Koeffizienten $\hat{s}_i$ der best approximierenden Treppenfunktion s bzgl der Basis $\frac{\chi_i}{\sqrt{h_i}}$, $i = 1,\ldots,l$ gegeben sind durch

$$\hat{s}_i = \left\langle \frac{\chi_i}{\sqrt{h_i}}, f \right\rangle_{L^2[a,b]} = \frac{1}{\sqrt{h_i}}\int_{x_{i-1}}^{x_i} f(x)\,dx\,,$$

woraus folgt

$$s = \sum_{i=1}^{l} \hat{s}_i\frac{\chi_i}{\sqrt{h_i}} = \sum_{i=1}^{l}\frac{1}{h_i}\left(\int_{x_{i-1}}^{x_i} f(x)\,dx\right)\chi_i =: \sum_{i=1}^{l} s_i\chi_i\,.$$

Nun widmen wir uns der Fehlerabschätzung (1.3). Wir verwenden dazu die Fehlerdarstellung

$$\|f - s\|^2_{L^2[a,b]} = \int_a^b (f(x) - s(x))^2\,dx \;=\; \sum_{i=1}^{l}\int_{x_{i-1}}^{x_i}(f(x) - s_i)^2\,dx\,.$$

It nun $f \in H^1[a,b]$ (so ist sie auch stetig) und es existiert in jedem Teilintervall eine Zwischenstelle $\xi_i \in (x_{i-1}, x_i)$ mit $f(\xi_i) = s_i$. Somit gilt

$$\begin{aligned}(f(x) - s_i)^2 &= (f(x) - f(\xi_i))^2 \\ &\leq \left(\int_x^{\xi_i} f'(t)\, dt \right)^2 \\ &\leq \left(\int_{x_{i-1}}^{x_i} |f'(t)|\, dt \right)^2 \\ &\underbrace{\leq}_{\text{CS}} \int_{x_{i-1}}^{x_i} f'(t)^2\, dt \int_{x_{i-1}}^{x_i} 1\, dt \\ &= h_i \int_{x_{i-1}}^{x_i} f'(t)^2\, dt\ .\end{aligned}$$

Somit folgt also durch erneute Integration über das Intervall $[a,b]$:

$$\begin{aligned}\|f - s\|^2_{L^2[a,b]} &\leq \sum_{i=1}^{l} \int_{x_{i-1}}^{x_i} h_i \int_{x_{i-1}}^{x_i} f'(t)^2\, dt\, dx \\ &= \sum_{i=1}^{l} h_i^2 \int_{x_{i-1}}^{x_i} f'(t)^2\, dt \\ &\leq \max_{i=1,\ldots,l} h_i^2 \int_a^b f'(t)^2\, dt \\ &= h^2 \|f'\|^2_{L^2[a,b]}\ .\end{aligned}$$

□

1.2 Lineare Splines

Die Ordnung h Abschätzung in (1.3) ist optimal. Bessere Abschätzungen kann man mit stückweise linearen Ansatzfunktionen, oder noch glatteren Funktionen bekommen.

Definition 1.3. Ein *Spline vom Grad n* ist eine Funktion $s \in C^{n-1}[a,b]$, die auf jedem Intervall $[x_{i-1}, x_i)$ ein Polynom vom Grad n ist. Für den Raum der Splines vom Grad n schreiben wir $S_{n,\Delta}$.

Neben den Treppenfunktionen sind die stückweise linearen Funktionen, die sogenannten *linearen Splines* ($n = 1$), und die *kubischen Splines* ($n = 3$) wichtig. Bevor wir uns im Detail mit diesen Funktionen beschäftigen studieren wir einige allgemeine Resultate:

Bemerkung 1.4.
1. Der Raum $S_{n,\Delta}$ ist ein linearer Vektorraum, der die Polynome vom Grad n enthält.
2. Ist $s \in S_{n,\Delta}$, dann ist für $0 \leq k < n$, $s^k \in S_{n-k,\Delta}$ (die k-te Ableitung).
3. Die n-te Ableitung ist zwischen den Knoten konstant.

Proposition 1.5. *$S_{n,\Delta}$ ist ein $(n+l)$-dimensionaler Unterraum von $L^2[a,b]$.*

Beweis. Sei $\hat{\chi}_i$ eine beliebige n-te Stammfunktion der charakteristischen Funktion χ_i. Aus der Definition ergibt sich, dass $\hat{\chi}_i \in S_{n,\Delta}$. Wie wir schon festgestellt haben, sind die Polynome vom Grad kleiner gleich n, und damit alle Monome x^j, $j = 0, 1, \ldots, n-1$, in $S_{n,\Delta}$ enthalten. Wir zeigen nun, dass

$$\{\hat{\chi}_i : i = 1, \ldots, l\} \cup \left\{x^j : j = 0, \ldots, n-1\right\}$$

eine Basis des $S_{n,\Delta}$ bildet.

Sei $s \in S_{n,\Delta}$ beliebig, so is $s^n \in S_{0,\Delta}$ und kann daher geschrieben werden als $s^n = \sum_{i=1}^{l} s_i \chi_i$. Folglich stimmen s und $\sum_{i=1}^{l} s_i \hat{\chi}_i$ bis auf ein Polynom p vom Grad $n-1$ überein, d.h.:

$$s = p + \sum_{i=1}^{l} s_i \hat{\chi}_i \text{ in } [a,b] \,. \tag{1.4}$$

Zum Nachweis der linearen Unabhängigkeit verwenden wir (1.4), und zeigen, dass für ein $s = 0$ auch $s_i = 0$ und $p = 0$ gilt.

Aus (1.4) folgt aus $s = 0$ somit auch

$$0 = s^n = \sum_{i=1}^{l} s_i \chi_i \text{ in } [a,b] \,,$$

woraus aus den Basiseigenschaften von χ_i auch folgt, dass $s_i = 0$ für alle $i = 1, \ldots, l$. Folglich folgt aus (1.4) und unserer Annahme $s = 0$, dass $s = p = 0$, und somit die Behauptung. □

Bemerkung 1.6.

- Für lineare Splines ergeben sich genau $(l-1)$ Freiheitsgrade.
- Anstatt (1.4) verwendet man in der Numerik die sogenannte *nodale Basis der Hutfunktionen* Λ_i, $i = 0, 1, \ldots, l-1$.
- Die Hutfunktionen erfüllen

$$\Lambda_i(x_j) = \delta_{ij}\,, \quad i, j = 0, \ldots, l\,. \tag{1.5}$$

- $S_{1,\Delta} \subseteq H^1[a,b]$.

(1.5) erlaubt eine einfache Bestimmung der Koeffizienten des interpolierenden linearen Splines:

Satz 1.7. *Seien $y_0, \ldots, y_l$ vorgegebene Werte. Dann ist $s = \sum_{i=0}^{l} y_i \Lambda_i \in S_{1,\Lambda}$ der eindeutig bestimmte lineare Spline mit*

$$s(x_i) = y_i\,, \quad i = 0, \ldots, l\,.$$

Beweis. Wegen (1.5) erfüllt $s = \sum_{i=0}^{l} y_i \Lambda_i$ die Interpolationseigenschaft an den Knoten. Die Eindeutigkeit folgt nun daraus, dass eine lineare Funktion auf $[x_{i-1}, x_i]$ durch die beiden Randwerte in diesem Intervall eindeutig bestimmt ist. □

Im folgenden beschäftigen wir uns mit Fehlerabschätzungen für lineare Splines. Dazu verwenden wir einige Hilfsmittel der Approximationstheorie:

Lemma 1.8. *Seien $f, \phi \in H^1[a,b]$, die auf dem Gitter Δ übereinstimmen. Dann gilt*

$$\|f - \phi\|_{L^2[a,b]} \leq \frac{h}{\sqrt{2}} \|f' - \phi'\|_{L^2[a,b]}\,. \tag{1.6}$$

Beweis. Sei $x \in [x_{i-1}, x_i)$ beliebig. Da f und ϕ am Gitter identisch sind gilt:

$$f(x) - \phi(x) = \int_{x_{i-1}}^{x} (f'(t) - \phi'(t))\, dt.$$

Damit folgt aus der Cauchy-Schwarz Ungleichung:

$$\begin{aligned}\int_{x_{i-1}}^{x_i} |f(x)-\phi(x)|^2\,dx &= \int_{x_{i-1}}^{x_i}\left(\int_{x_{i-1}}^{x}(f'(t)-\phi'(t))\,dt\right)^2 dx\\ &\le \int_{x_{i-1}}^{x_i}\left(\int_{x_{i-1}}^{x}(f'(t)-\phi'(t))^2\,dt\int_{x_{i-1}}^{x}dt\right)dx\\ &= \int_{x_{i-1}}^{x_i}(x-x_{i-1})\int_{x_{i-1}}^{x}(f'(t)-\phi'(t))^2\,dt\,dx\\ &\le \int_{x_{i-1}}^{x_i}(x-x_{i-1})dx\int_{x_{i-1}}^{x_i}(f'(t)-\phi'(t))^2\,dt\\ &= \frac{1}{2}h_i^2\int_{x_{i-1}}^{x_i}(f'(t)-\phi'(t))^2\,dt\,.\end{aligned}$$

Summiert man über $i = 1,\dots,l$, dann erhält man die gewünschte Ungleichung. □

Mit diesem Lemma können wir nun Fehlerabschätzungen für interpolierende lineare Splines beweisen:

Satz 1.9. *Sei $f \in H^1[a,b]$ und s der interpolierende Spline zu f auf Δ. Dann ist*

$$\|f-s\|_{L^2[a,b]} \le \frac{h}{\sqrt{2}}\|f'\|_{L^2[a,b]}\,.$$

Beweis. Wir wissen bereits, dass jeder lineare Spline in $H^1[a,b]$ liegt. Damit existiert insbesondere die schwache Ableitung, und folgende Identität ist wohldefiniert:

$$\begin{aligned}\|f'-s'\|^2_{L^2[a,b]} &= \|f'\|^2_{L^2[a,b]} - \langle 2f'-s', s'\rangle_{L^2[a,b]}\\ &= \|f'\|^2_{L^2[a,b]} - \int_a^b (2f'-s')(x)s'(x)\,dx\,.\end{aligned}$$

Wir zeigen nun, dass $\int_a^b (2f'-s')(x)s'(x)\,dx \ge 0$ ist, woraus aus der obigen Ungleichung folgt, dass

$$\|f'-s'\|_{L^2[a,b]} \le \|f'\|_{L^2[a,b]}\,.$$

Zusammen mit (1.6) folgt dann:

$$\|f-s\|_{L^2[a,b]} \le \frac{h}{\sqrt{2}}\|f'-s'\|_{L^2[a,b]} \le \frac{h}{\sqrt{2}}\|f'\|_{L^2[a,b]}\,,$$

und somit die Behauptung des Satzes.

Wir zeigen also nun $\int_a^b (2f' - s')(x)s'(x)\,dx \geq 0$: s ist eine stückweise lineare Funktion, weshalb gilt

$$s'(x) = \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}\,, \quad x \in (x_{i-1}, x_i)\,.$$

Damit folgt

$$\begin{aligned}
&\int_a^b (2f' - s')(x)s'(x)\,dx \\
&= \sum_{i=1}^{l} \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} \int_{x_{i-1}}^{x_i} (2f' - s')(x)\,dx \\
&= \text{(Integration)} \\
&\qquad \sum_{i=1}^{l} \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} (2f(x_i) - s(x_i) - 2f(x_{i-1}) + s(x_{i-1})) \\
&= \text{(Interpolatierende Spline Eigenschaft)} \\
&\qquad \sum_{i=1}^{l} \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} (f(x_i) - f(x_{i-1})) \\
&= \sum_{i=1}^{l} \frac{f(x_i) - f(x_{i-1})^2}{x_i - x_{i-1}} \\
&\geq 0\,.
\end{aligned}$$

□

Nun beweisen wir typische Abschätzung für Splines. Typische ist, dass die Abschätzungen von der Ordnung her umso besser sind, je glatter die zu interpoliernde Funktion ist:

Satz 1.10. *Sei $f \in H^2[a,b]$ und s der interpolierende Spline, dann gilt:*

$$\begin{aligned}
\|f - s\|_{L^2[a,b]} &\leq \frac{h^2}{2} \|f''\|_{L^2[a,b]}\,, \\
\|f' - s'\|_{L^2[a,b]} &\leq \frac{h}{\sqrt{2}} \|f''\|_{L^2[a,b]}\,.
\end{aligned} \tag{1.7}$$

Beweis. Wir zeigen zuerst die zweite Identität

$$
\begin{aligned}
&\|f' - s'\|^2_{L^2[a,b]} \\
&= \sum_{i=1}^{l} \int_{x_{i-1}}^{x_i} (f' - s')^2(x)\, dx \\
&= (\text{Partielle Integration auf } (x_{i-1}, x_i)) \\
&\quad \sum_{i=1}^{l} \left((f-s)(x)(f'-s')(x)|_{x_{i-1}}^{x_i} - \int_{x_{i-1}}^{x_i} (f-s)(x)(f''-s'')(x)\, dx \right) \\
&= (s'' = 0 \text{ in jedem Teilintervall und } s \text{ interpoliert an Randpunkten}) \\
&\quad - \sum_{i=1}^{l} \int_{x_{i-1}}^{x_i} (f-s)(x) f''(x)\, dx \\
&= - \int_a^b (f-s)(x) f''(x)\, dx \\
&\leq (\text{Cauchy Schwarz Ungleichung}) \\
&\quad \|f-s\|_{L^2[a,b]} \|f''\|_{L^2[a,b]} \\
&\underbrace{\leq}_{(1.6)} \frac{h}{\sqrt{2}} \|f' - s'\|_{L^2[a,b]} \|f''\|_{L^2[a,b]} \,.
\end{aligned}
$$

Durch Division mit $\|f' - s'\|_{L^2[a,b]}$ folgt somit der zweite Teil von (1.7).

Aus Lemma 1.8 und der zweiten Ungleichung folgt

$$\|f - s\|_{L^2[a,b]} \leq \frac{h}{\sqrt{2}} \|f' - s'\|^2_{L^2[a,b]} = \frac{h^2}{2} \|f''\|^2_{L^2[a,b]} \,.$$

□

Bemerkung 1.11.

Beachte, dass die Erhöhung des Splinegrades von $n = 0$ zu $n = 1$ zu einer entsprechenden Erhöhung der Ordnung um einen Faktor 1 geführt hat. Vergleiche (1.7) und (1.3).

Die beste Approximierende von f im L^2-Sinn erfüllt natürlich dann auch die erste Abschätzung von (1.7). Beachte, dass diese Funktion nicht in H^1 sein muss.

Im Abschluß zu den linearen Splines beschäftigen wir uns noch mit der Berechnung des approximierenden linearen Splines: Dazu verwenden wir die *Gramsche Matrix* der linearen Splines

$$G = \left[\langle \Lambda_i, \Lambda_j \rangle_{L^2[a,b]}\right]_{i,j=0,1,\ldots,l} . \tag{1.8}$$

Diese Matrix hat folgende Eigenschaften:

1. Die Dimension der Matrix ist $(l+1)\times(l+1)$ (also Knoten zum Quadrat).
2. Die Matrix ist regulär, da die Λ_i linear unabhängig sind.
3. Da für $|i-j| > 2$ $\langle \Lambda_i, \Lambda_j \rangle_{L^2[a,b]} = 0$ gilt, ist G eine symmetrische Tridiagonalmatrix.
4. Die Diagonal- und Nebendiagonaleinträge ergeben sich wie folgt:

   Diagonale: Für $i = j \in \{1, \ldots, l\}$ gilt:

   $$\int_{x_{i-1}}^{x_i} \Lambda_i^2(x)\, dx = \int_0^{h_i} \frac{t^2}{h_i^2}\, dt = \frac{h_i}{3} .$$

   Analog sieht man, dass für $i = j \in \{0, \ldots, l-1\}$

   $$\int_{x_i}^{x_{i+1}} \Lambda_i^2(x)\, dx = \frac{h_{i+1}}{3} .$$

   Zusammengefaßt gilt also für die Diagonale von G

   $$g_{i,i} = \begin{cases} \frac{h_1}{3} & i = 0\,, \\ \frac{h_i + h_{i+1}}{3} & i = 1, \ldots, l-1\,, \\ \frac{h_l}{3} & i = l\,. \end{cases}$$

   Nebendiagonalen: Wegen der Symmetrie reicht es, die untere Nebendiagonale zu bestimmen, also ist $j = i-1$, $i = 1, \ldots, l$:

   $$\begin{aligned} g_{i,i-1} &= \int_{x_{i-1}}^{x_i} \Lambda_i(x)\Lambda_{i-1}(x)\, dx \\ &= \int_0^{h_i} \frac{t}{h_i}\frac{h_i - t}{h_i}\, dt \\ &= \frac{1}{h_i^2}\left(\frac{h_i^3}{2} - \frac{h_i^3}{3}\right) \\ &= \frac{h_i}{6} . \end{aligned}$$

Zusammengefasst:

$$G := \frac{1}{6}\begin{bmatrix} 2h_1 & h_1 & & & 0 \\ h_1 & 2(h_1+h_2) & h_2 & & \\ & h_2 & \ddots & \ddots & \\ & & \ddots & 2(h_{l-1}+h_l) & h_l \\ 0 & & & h_l & 2h_l \end{bmatrix}$$

Das folgende allgemeine Resultat gibt Auskunft, wie die beste approximierende Näherung mit der Gramschen Matrix berechnet werden kann.

Satz 1.12. *Sei $f \in X$, wobai X ein Hilbertraum ist, und $\{\phi_i : i \in \mathbb{N}\}$ eine Basis von X_n. $\mathcal{G}$ bezeichne die zu dieser Basis gehörige Gramsche Matrix. Dann ist $f_n \sum_{i=1}^{n} x_i\phi_i \in X_n$ genau dann die beste Approximation an f aus X_n, wenn*

$$Gx = b \text{ für } x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \text{ und } \begin{pmatrix} \langle\phi_1, f\rangle \\ \vdots \\ \langle\phi_n, f\rangle \end{pmatrix}.$$

1.3 Kubische Splines

Kubische Splines, das sind die Elemente von $S_{3,\Delta}$, werden häufig in der Computergraphik verwendet.

Wir fassen einige allgemeine Resultate, die wir vorher kennengelernt haben zusammen:

1. Ein kubischer Spline ist zweimal stetig differenzierbar. Dies ist auch der Grund, dass diese Funktionen in der Computergraphik so beliebt sind.

2. Wir wissen aus Bemerkung 1.4, dass wir zur eindeutigen Bestimmung eines kubischen Splines $(l+3)$ Bedingungen brauchen.

3. Ist $s \in S_{3,\Delta}$, dann ist $s'' \in S_{1,\Delta}$. Damit kann man s'' mit Hilfe von Hutfunktionen darstellen:

$$s'' = \sum_{i=0}^{l} \gamma_i \Lambda_i \,, \tag{1.9}$$

wobei $\gamma_i = s''(x_i)\,, i = 0, \ldots, l\,$, gilt. Man nennt γ_i die *Momente* des kubischen Splines.

Im folgenden bestimmen wir die Bedingungen, die zu einer eindeutigen Berechnung des kubischen Splines führen.

Zuerst verwenden wir, dass für eine C^2-Funktion ρ in einem Intervall $[x_{i-1}, x_i]$ gilt:

$$\begin{aligned}\rho(x) - \rho(x_i) &= \int_{x_i}^{x} \rho'(t) \cdot 1 \, dt \\ &= (\text{partielle Integration}) \\ &= \rho'(t)(t-x)|_{t=x_i}^{x} - \int_{x_i}^{x} \rho''(t)(t-x)\, dt \\ &= -\rho'(x_i)(x_i - x) - \int_{x_i}^{x} \rho''(t)(t-x)\, dt\ .\end{aligned} \tag{1.10}$$

Weiters stellen wir fest, dass für $t \in [x_{i-1}, x_i]$

$$\begin{aligned}s''(t) &= \gamma_{i-1}\Lambda_{i-1}(t) + \gamma_i \Lambda_i(t) \\ &= \gamma_{i-1}\frac{x_i - t}{x_i - x_{i-1}} + \gamma_i \frac{t - x_{i-1}}{x_i - x_{i-1}} \\ &= -\frac{\gamma_{i-1}}{h_i}(t - x_i) + \frac{\gamma_i}{h_i}(t - x_{i-1}) \\ &= \frac{\gamma_i - \gamma_{i-1}}{h_i}(t - x_i) + \gamma_i\end{aligned} \tag{1.11}$$

gilt, woraus sich für $x \in [x_{i-1}, x_i)$ ergibt

$$\begin{aligned}&s(x) - s(x_i) + s'(x_i)(x_i - x) \\ =& -\int_{x_i}^{x} s''(t)(t-x)\, dt \\ =& -\frac{\gamma_i - \gamma_{i-1}}{h_i}\int_{x_i}^{x}(t - x_i)(t-x)\, dt - \gamma_i \int_{x_i}^{x} t - x\, dt \\ =& \frac{\gamma_i - \gamma_{i-1}}{h_i}\frac{(x - x_i)^3}{6} + \gamma_i \frac{(x - x_i)^2}{2}\ .\end{aligned} \tag{1.12}$$

Mit der Abkürzung

$$s_i = s(x_i) \text{ und } s_i' = s'(x_i) \text{ für } i = 0, \ldots, l$$

erhalten wir aud (1.12) somit eine kompakte Identität für den kubischen Spline: Für $x \in [x_{i-1}, x_i]$ und $i = 1, \ldots, l$ gilt

$$s(x) = s_i + s_i'(x - x_i) + \gamma_i \frac{(x - x_i)^2}{2} + \frac{\gamma_i - \gamma_{i-1}}{h_i} \frac{(x - x_i)^3}{6} \,. \tag{1.13}$$

Insbesondere gilt somit für $i = 1, \ldots, l$

$$\begin{aligned} s_{i-1} &= s_i - s_i' h_i + \frac{\gamma_i h_i^2}{2} - \frac{(\gamma_i - \gamma_{i-1}) h_i^2}{6} \\ &= s_i - s_i' h_i + \frac{h_i^2}{6}(2\gamma_i + \gamma_{i-1}) \,, \\ s_{i-1}' &= s_i' - \frac{h_i}{2}(\gamma_{i-1} + \gamma_i) \,. \end{aligned} \tag{1.14}$$

Durch Kombination dieser Gleichungen erhalten wir für $i = 1, \ldots, l-1$

$$\begin{aligned} &\frac{s_{i+1} - s_i}{h_{i+1}} - \frac{s_i - s_{i-1}}{h_i} \\ =& s_{i+1}' - \gamma_i \frac{h_{i+1}}{6} - \gamma_{i+1} \frac{h_{i+1}}{3} - s_i' + \gamma_{i-1} \frac{h_i}{6} + \gamma_i \frac{h_i}{3} \\ =& (\gamma_i + \gamma_{i+1}) \frac{h_{i+1}}{2} - \gamma_i \frac{h_{i+1}}{6} - \gamma_{i+1} \frac{h_{i+1}}{3} + \gamma_{i-1} \frac{h_i}{6} + \gamma_i \frac{h_i}{3} \\ =& \frac{1}{6}(h_{i+1}\gamma_{i+1} + 2\gamma_i(h_i + h_{i+1}) + \gamma_{i-1} h_i) \,. \end{aligned}$$

In Matrixschreibweise lautet das System:

$$\begin{aligned} &\frac{1}{6} \underbrace{\begin{bmatrix} h_1 & 2(h_1 + h_2) & h_2 & & & 0 \\ & h_2 & 2(h_2 + h_3) & \ddots & & \\ & & \ddots & \ddots & h_{l-1} & \\ & & & h_{l-1} & 2(h_{l-1} + h_l) & h_l \end{bmatrix}}_{\in \mathbb{R}^{(l-1)\times(l+1)}} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_{l-1} \\ \gamma_l \end{bmatrix} \\ = - & \underbrace{\begin{bmatrix} -h_1^{-1} & h_1^{-1} + h_2^{-1} & -h_2^{-1} & & & 0 \\ & -h_2^{-1} & h_2^{-1} + h_3^{-1} & \ddots & & \\ & & \ddots & \ddots & -h_{l-1}^{-1} & \\ & & & -h_{l-1}^{-1} & h_{l-1}^{-1} + h_l^{-1} & -h_l^{-1} \end{bmatrix}}_{\in \mathbb{R}^{(l-1)\times(l+1)}} \begin{bmatrix} s_0 \\ s_1 \\ \vdots \\ s_{l-1} \\ s_l \end{bmatrix} . \end{aligned} \tag{1.15}$$

Erfüllen umgekehrt die Momente γ_i und die Funktionswerte $s_i = s(x_i)$ die Gleichung (1.15), dann definieren diese Werte den interpolierenden kubischen Spline. Von dem sind die Ableitungen wegen (1.13) gegeben durch

$$s'_i = s'(x_i) = \frac{s_i - s_{i-1}}{h_i} + \gamma_{i-1}\frac{h_i}{6} + \gamma_i\frac{h_i^2}{3} \ .$$

Die beiden Matrizen in (1.15) haben die Dimension $(l-1)\times(l+1)$, und daher sind die Gleichungssysteme unterbestimmt. Man muß also zusätzliche Bedingungen aufnehmen. Bei *natürlichen kubischen Splines* fordert man zusätzlich, daß

$$s''(a) = s''(b) = 0 \tag{1.16}$$

gilt. Dies bedeutet aber wegen (1.9) gerade, dass

$$\gamma_0 = s''(a) = 0 \text{ and } \gamma_l = s''(b) = 0 \ . \tag{1.17}$$

Damit vereinfacht sich das Gleichungessystem (1.15) in diesem Fall zu

$$\frac{1}{6}\underbrace{\begin{bmatrix} 2(h_1+h_2) & h_2 & & 0 \\ h_2 & 2(h_2+h_3) & \ddots & \\ & \ddots & \ddots & h_{l-1} \\ & & h_{l-1} & 2(h_{l-1}+h_l) \end{bmatrix}}_{:=\mathcal{G}\in\mathbb{R}^{(l-1)\times(l-1)}} \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_{l-1} \end{bmatrix} \begin{bmatrix} d_1 \\ \vdots \\ d_{l-1} \end{bmatrix} , \tag{1.18}$$

wobei

$$d_i = \frac{s_{i+1}-s_i}{h_{i+1}} - \frac{s_i - s_{i-1}}{h_i} = \frac{s_{i+1}}{h_{i+1}} - s_i\left(\frac{1}{h_i}+\frac{1}{h_{i+1}}\right) + \frac{s_{i-1}}{h_{i-1}} , \quad i = 1,\dots,l-1 \ . \tag{1.19}$$

Die Matrix $\mathcal{G}$ in (1.18) ist die Gramsche Matrix der Hutfunktionen $\{\Lambda_1,\dots,\Lambda_{l-1}\}$. Sie ist somit positiv definit und damit hat Gleichung (1.18) eine eindeutige Lösung. Wir fassen das Gesagte in einem Satz zusammen:

Satz 1.13. *Seien $y_0,\dots,y_l$ vorgegebene Daten, dann gibt es genau einen natürlichen Spline $s \in S_{3,\Delta}$ mit*

$$s(x_i) = y_i \, , \qquad i = 0,\dots,l \ .$$

Beispiel 1.14. Auf einem äquidistanten Gitter mit Gitterweite h. Wir bestimmen den natürlichen kubischen Spline s, der für ein festes $j = 0, 1, \ldots, l$ die Lagrange-Interpolationsaufgabe

$$s(x_i) = \delta_{ij}\,, \quad i = 0, \ldots, l\,, \tag{1.20}$$

löst. Das Gleichungssystem (1.18), (1.19) hat die Gestalt

$$\begin{aligned}
&\begin{bmatrix} 4 & 1 & & 0 \\ 1 & 4 & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & 4 \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_{l-1} \end{bmatrix} \\
&= s_{i+1} - 2s_i + s_{i-1} \\
&= \frac{6}{h^2}(\delta_{i+1,j} - 2\delta_{i,j} + \delta_{i-1,j}) \\
&= \frac{6}{h^2}(e_{j+1} - 2e_j + e_{j-1})\,,
\end{aligned} \tag{1.21}$$

wobei $e_j, j = 1, \ldots, l-1$ die kartesischen Basisvektoren in $\mathbb{R}^{l-1}$ sind. Formal setzt man $e_0 = e_l = 0$.

Kapitel 2

Gewöhnliche Differentialgleichungen

Wir werden im ersten Teil dieser Vorlesung Theorie und Numerik für Systeme von gewöhnlichen Differentialgleichungen der Form

$$y' = f(t, y)\ , t \in [0, T] \text{ mit der Anfangsbedingung } y(0) = y_0$$

behandeln. Dabei ist zu beachten, dass y eine vektorwertige Funktion sein kann. Wir sprechen in diesem Fall von einem *System erster Ordnung*.

Beispiel 2.1. (Exponentielles Wachstum) Eine Population einer Art lebe von einem unerschöpflichen Nahrungsvorrat und hat zur Zeit t die Größe $P(t)$. Nehmen wir nun an, dass sich die Population gleichmäßig verdoppelt. Eine gleichmäßige änderung der Bevölkerung, bedeutet, dass die Bevolkerungsänderungsrate in jedem Zeitpunkt konstant ist, also

$$\frac{P'(t)}{P(t)} = \alpha = \text{const.} \tag{2.1}$$

Die Konstante ergibt sich, wenn man die Verdoppelungen in einem Jahr erreichen will wie folgt: Da $(\log P)'(t) = \alpha$ gilt, folgt

$$\log P(t) = \alpha t + \beta\ .$$

Woraus, wenn die Zeiteinheit für t Jahre gewählt wird:

$$2 = \frac{P(t+1)}{P(t)} = e^{\alpha t} \text{ und } P(0) = e^{\beta}\ .$$

Mehr Beispiele können in dem Buch von von Heuser [9] nachgelesen werden.

Wir weisen nun auf einen Zusammenhang zwischen partiellen Differentialgleichungen, die Evolutionsprozesse beschreiben, und gewöhnlichen Differentialgleichungen hin. Mit Hilfe dieser einfachen Feststellung kann man Lösungsverfahren für gewöhnliche Differentialgleichungen zur Lösung von partiellen Differentialgleichungen heranziehen.

Beispiel 2.2. Sei $u(x,t)$, $-1 \leq x \leq 1$, die Temperaturverteilung zum Zeitpunkt t in einem Stab der Länge $l = 2$. Bei konstanter Wärmeleitfähigkeit σ im Stab, genügt u der *Wärmeleitungsgleichung*:

$$u_t = \sigma u_{xx}\,, \quad -1 < x < 1\,, 0 < t < T\,. \tag{2.2}$$

Dies ist eine partielle Differentialgleichung (weil sie von zwei Variablen, x, t abhängt). Durch Diskretisierung der Ortsvariablen x kann man diese partielle Differentialgleichung in ein System von gewöhnlichen Differentialgleichungen überführen. Um dies zu demonstrieren, setzen wir der Einfachheit $\sigma = 1$. Sei v eine stetige, stückweise differenzierbare Funktion mit $v(-1) = v(1) = 0$, dann folgt mit partieller Integration

$$\int_{-1}^{1} u_t(t,x)v(x)\,dx = \int_{-1}^{1} u_{xx}(t,x)v(x)\,dx = -\int_{-1}^{1} u_x(t,x)v_x(x)\,dx\,. \tag{2.3}$$

Nehmen wir nun an, dass am linken und rechten Ende des Stabes Randtemperaturen $u_0(t)$ und $u_1(t)$ gegeben sind, dann kann man $u(t,x)$ für jedes t etwa stückweise Geraden über dem Gitter $\Delta = \{-1 = x_0 < x_1 < ... < x_n = 1\}$ approximieren, d.h.

$$u(t,x) = \sum_{i=0}^{n} y_i(t)\Lambda_i(x)\,, \tag{2.4}$$

wobei Λ_i ein linearer Ansatz-Spline ist, der an x_i den Wert 1 und an seinen Nachbarpunkten den Wert 0 (sofern sie im Intervall $[-1,1]$ liegen; außerdem ist der Spline stetig. y_0 und y_n sind bekannt, das sind nämlich die Randtemperaturvorgaben $u_0(t)$ und $u_1(t)$. Alle anderen Funktionen y_i sind unbekannt.

Durch Einsetzen von (2.4) in (2.3) ergibt sich ein Differentialgleichungssytem für die Funktionen $y_1, ..., y_{n-1}$:

$$\sum_{i=0}^{n} y_i'(t) \int_{-1}^{1} \Lambda_i(x)v(x)\,dx = -\sum_{i=0}^{n} y_i(t) \int_{-1}^{1} \Lambda_i'(x)v_x(x)\,dx\,,$$

wobei $v(x) \in \{\Lambda_j(x) : j = 1, ..., n-1\}$ - also kann v irgendeine Hutfunktion sein, die homogene Randbedingungen erfüllt. Bezeichnet

$$G := [\langle \Lambda_i, \Lambda_j \rangle]_{1 \le i,j \le n-1} = \frac{h}{6} \begin{pmatrix} 4 & 1 & 0 & . & . & 0 \\ 1 & 4 & 1 & 0 & . & 0 \\ . & . & . & . & . & . \\ 0 & . & . & 1 & 4 & 1 \\ . & . & . & 0 & 1 & 4 \end{pmatrix}$$

und $A = [\langle \Lambda_i', \Lambda_j' \rangle]_{1 \le i,j \le n-1}$ so gilt

$$Gy'(t) + Ay(t) = b\,, \tag{2.5}$$

und b ein Vektor, der sich aus den Größen u_0 und u_1 ergibt.

Zur Lösung des Systems (2.5) braucht man noch Anfangswerte für die Funktionen $y_1, ..., y_{n-1}$, die sich etwa durch Interpolation der Anfangstemperatur $u(0, x)$ ergeben.

2.1 Lösungstheorie

Für eine umfassende Lösungstheorie von gewöhnlichen Differentialgleichungen verweisen wir auf Walter [11].

Die Grundlage vieler Existenz- und Eindeutigkeitsaussagen für die Lösung von gewöhnlichen Differentialgleichungen ist der Satz von Picard-Lindelöf.

In diesem Kapitel treffen wir die folgenden allgemeinen Annahmen:

- $y = y(t) \in \mathbb{R}^d$,
- die Funktion f sei in dem offenen Gebiet $\Omega = I \times J$ definiert, mit $(0, T) \subseteq I$,
- $J \subseteq \mathbb{R}^d$. J darf auch ein unbeschränktes Gebiet sein.

Satz 2.3. *(Picard-Lindelöf) Die Funktion f sei stetig auf Ω und auf allen kompakten Teilmengen $K \subseteq \Omega$ gelte*

$$\|f(t, y) - f(t, z)\|_2 \le L_K \|y - z\|_2 \quad \text{für alle } (t, y), (t, z) \in K\,. \tag{2.6}$$

Dann existiert zu jedem $y_0 \in J$ ein nicht leeres Teilintervall $I_0 \subseteq I$ mit 0 im Abschluss von I_0 und eine eindeutig bestimmte stetig differenzierbare Lösung $y : I_0 \to J$ des Anfangswertproblems

$$y' = f(t, y), t \in I_0 \text{ und } y(0) = y_0\,. \tag{2.7}$$

Der Beweis des Satzes von Picard-Lindelöf wird mit dem Banachschen Fixpunktsatz geführt, wobei geeignete Normen eingeführt werden müssen. Den Beweis diesen Satzes findet man etwa in Walter [11].

Die Voraussetzungen des Satzes von Picard-Lindelöf sind etwa erfüllt, falls die Funktion f auf Ω stetig differenzierbar ist. Die Folgerung der Existenz eines Teilintervalls I_0 ist klärungsbedürftig. Ist $I = (0, T)$ das maximale Interval, in dem f die Voraussetzungen des Satzes 2.3 erfüllt, dann folgt, dass entweder eine eindeutig bestimmte Lösung der Differentialgleichung im gesamten Intervall $[0, T]$ existiert, oder dass die Lösung im Inneren des Intervalls $[0, T]$ gegen den Rand der Menge J konvergiert.

Neben der Existenz und Eindeutigkeit der Lösung einer Problemstellung ist noch ein weitere wichtige Fragestellung, die nach der stetigen Abhängigkeit der Lösung von den Eingangsdaten. Der nächste Satz garantiert stetige Abhängigkeit der Lösung von den Anfangsdaten.

Satz 2.4. *f sei stetig und erfülle*

$$\langle f(t, y) - f(t, z), y - z\rangle_2 \leq l\|y - z\|_2^2 \text{ für alle } (t, y), (t, z) \in \Omega \,. \tag{2.8}$$

Sind $y, z : I \to J$ stetig differenzierbare Lösungen der Differentialgleichungen $y' = f(t, y)$ und $z' = f(t, z)$ zu verschiedenen Anfangswerten $y_0, z_0 \in J$. Dann gilt

$$\|y(t_0) - z(t_0)\|_2 \leq \exp(lt)\|y_0 - z_0\|_2 \text{ für alle } t_0 \in I \,.$$

Beweis. Sei $t_0 \in I$ beliebig. Wir nehmen ohne Beschränkung der Allgemeinheit an, dass $y(t_0) \neq z(t_0)$ ist. Da y und z als stetig differenzierbar vorausgesetzt wurden, ist $x(t) := \|y(t) - z(t)\|_2^2$ in einer Umgebung von t_0 strikt positiv und daher ist die Funktion $\log x(t)$ in dieser Umgebung wohldefiniert. Dort gilt also

$$\begin{aligned} x'(t) &= \frac{d}{dt}\|y(t) - z(t)\|_2^2 \\ &= 2\,\langle y'(t) - z'(t), y(t) - z(t)\rangle_2 \\ &= 2\,\langle f(t, y(t)) - f(t, z(t)), y(t) - z(t)\rangle_2 \\ &\leq 2l\|y(t) - z(t)\|_2^2 \\ &= 2lx(t) \,. \end{aligned}$$

Daher ist

$$\frac{d}{dt}\log x(t) = \frac{x'(t)}{x(t)} \leq 2l \,.$$

Sei $t < t_0$, dann folgt durch Aufintegrieren von t nach t_0, dass

$$\log(x(t_0)/x(t)) = \log x(t_0) - \log x(t) \leq 2l(t_0 - t) \ ,$$

bzw.

$$\begin{aligned} \|y(t_0) - z(t_0)\|_2^2 = x(t_0) &\leq x(t) \exp(2l(t_0 - t)) \\ &= \|y(t) - z(t)\|_2^2 \exp(2l(t_0 - t)) \ . \end{aligned} \tag{2.9}$$

Diese Ungleichung gilt für jedes beliebige t aus dem größtmöglichen Intervall um t_0, in dem $x(t)$ positiv bleibt. Da aber nach unserer Annahme $x(t_0)$ positiv bleibt, kann es kein $t \in [0, t_0)$ geben mit $x(t) = 0$. Daher kann in obiger Gleichung der Grenzübergang $t \to 0$ durchgeführt werden, woraus die Behauptung folgt. □

Aus Satz 2.4 folgt insbesondere, dass unter der Bedingung (2.8) Lösungen des Anfangswertproblems (2.7) (falls existent) eindeutig bestimmt sind.

Die Voraussetzungen (2.8) und (2.6) sind nicht direkt vergleichbar.

Gilt die Lipschitz-Bedingung (2.6) global, d.h. für alle $(t, y), (t, z) \in \Omega$, dann gilt (2.8) mit $l = L = L_K$. Deshalb bezeichnen wir (2.8) im folgenden als *schwache Lipschitz-Bedingung*. Die Bedingung (2.8) gleichmäßig in Ω gelten, während (2.6) nur lokal gelten muss.

(2.8) hat gegenüber (2.6) den Vorteil, dass eine negative Konstante l zulässig ist, während $L = L_K$ zwangsläufig immer positiv sein muss. Differentialgleichungen, die einer schwachen Lipschitz-Bedingung mit einem negativem l genügen, nennt man *strikt dissipativ*.

Beispiel 2.5. Die Differentialgleichung

$$y' = \lambda y \, , y(0) = y_0$$

hat die Lösung $y(t) = y_0 \exp(\lambda t)$. Wegen

$$\langle f(t, y) - f(t, z), y - z \rangle = \lambda \|y - z\|_2^2$$

ist die Voraussetzung von Satz 2.8 für $l = \lambda$ in ganz $\mathbb{R}^+ \times \mathbb{R}^d$ erfüllt. Für negative Werte von λ werden Fehler in den Startwerten mit einem Faktor $\exp(\lambda t)$ gedämpft. Darüberhinaus gehen alle Lösungen gegen Null. Eine Differentialgleichung mit dieser Eigenschaft nennt man *asymptotisch stabil*. Für $\lambda > 0$ werden Fehler in den Startwerten verstärkt. Die Lösungen sind instabil.

Satz 2.4 besagt, dass die Zuordnung $y_0 \to y(t)$ stetig ist, genauer Lipschitz-stetig mit Lipschitz-Konstante $\kappa = \exp(lt)$. Wir können daher in Ω die Größe κ als ein Maß für die lokale Fehlerverstärkung des absoluten Datenfehlers ansehen. κ hat die Rolle einer absoluten (lokalen bzgl. t) *Konditionszahl* der Abbildung

$$y_0 \to y(t)\,, \quad 0 \le t \le \Delta t\,.$$

2.2 Numerische Verfahren

Für weiterführende Literatur zu diesem Kapitel sei auf [3, 2] verwiesen.

2.2.1 Euler Verfahren

Als erstes Beispiel eines Verfahrens zur Lösung von gewöhnlichen Differentialgleichungen betrachten wir das klassische *Euler Verfahren.*

Auf einem vorgegebenen Gitter

$$\Delta = \{0 = t_0 < t_1 < t_2 < ... < t_n\} \subseteq I$$

ist das Euler-Verfahren wie folgt definiert:

$$y_{i+1} = y_i + (t_{i+1} - t_i) f(t_i, y_i)\,.$$

Es beruht auf der Beobachtung, dass die Funktion f in jedem Punkt $(t, y) \in \Omega$ die Steigung der Lösungskurve definiert. Darum kann das Euler-Verfahren auch dadurch erklärt werden kann, dass man die Zeitableitung durch einen *Vorwärtsdifferenzenquotienten* approximiert.

Beispiel 2.6. Wir studieren die einfache gewöhnliche Differentialgleichung

$$y' = y, \quad y(0) = 1\,.$$

Die exakte Lösung ist $y(t) = \exp(t)$. Mit dem Euler-Verfahren ergibt sich in einem äquidistanten Gitter $(t_i = ih)$ (man beachte $f(t, y) = y$)

$$\begin{aligned} y_0 &= 1\,, \\ y_1 &= y_0 + hf(h, y_0) = 1 + h\,, \\ y_2 &= y_1 + hf(h, y_1) = 1 + h + h(1 + h) = (1 + h)^2\,, \\ y_3 &= y_2 + hf(h, y_2) = (1 + h)^2 + h(1 + h)^2 = (1 + h)^3\,. \end{aligned}$$

Mit Induktion sieht man, dass $y_n = (1 + h)^n$. Für $t_n = T$ bedeutet dies

$$y_n = (1 + T/n)^n \to \exp(T) = y(T)\,, \quad n \to \infty\,.$$

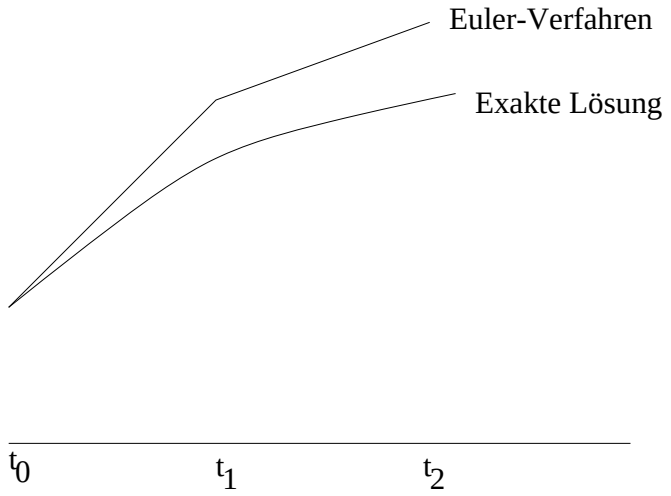

Abbildung 2.1: Schematische Darstellung des Euler-Verfahrens

Klarerweise löst das Euler-Verfahren die gewöhnliche Differentialgleichung *nicht* exakt. Wir untersuchen nun wie gut das Euler-Verfahren die Lösung der gewöhnlichen Differentialgleichung approximiert.

Der Eifachheit wegen beschränken wir uns auf ein äquidistantes Gitter Δ mit konstanten Schrittweite $h = t_{i+1} - t_i$, $i = 0, ..., n$.

Satz 2.7. *Sei $I = [0, T]$ und $f : I \times \mathbb{R}^d \to \mathbb{R}^d$ stetig differenzierbar und bezüglich y global Lipschitz-stetig,*

$$\|f(t, y) - f(t, z)\|_2 \leq L\|y - z\|_2 \text{ für alle } t \in I \text{ und } y, z \in \mathbb{R}^d \, .$$

Ist nun y die eindeutige Lösung des Anfangswertproblems (2.7) *und sind y_i, $i = 1, ..., n$, die berechneten Näherungen des Euler-Verfahren an den äquidistanten Gitterpunkten $t_i = ih \in I$, dann gilt*

$$\begin{aligned} \varepsilon_i &:= \|y(t_i) - y_i\|_2 \\ &\leq \frac{(1 + Lh)^i - 1}{2L} \|y''\|_{[0,T]} h \\ &\leq \frac{\exp(Lt_i) - 1}{2L} \|y''\|_{[0,T]} h \, , \quad i = 0, ..., n \, . \end{aligned} \tag{2.10}$$

Hierbei ist $\|y''\|_{[0,T]} = \max_{0 \leq t \leq T} \|y''(t)\|_2$ (man beachte $y'' \in \mathbb{R}^d$).

Beweis. Die Voraussetzung an f garantiert, dass y zweimal stetig differenzierbar ist: Es gilt nämlich, dass

$$y'' = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial y} y' = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial y} f \,,$$

und alle terme auf der rechten Seite sind wegen Voraussetung stetig.

Der eigentliche Beweis dieses Satzes gliedert sich nun in drei Teile:

Lokaler Fehler: Nehmen wir zunächst an, dass für t_i das Euler-Verfahren auf dem Punkt $(t_i, y(t_i))$ der *exakten* Lösungskurve initiiert ist. Sei z_{i+1} die Approximation für $y(t_{i+1})$, die man mit dem Euler-Verfahren berechnet, dann gilt

$$\begin{aligned}
\|y(t_{i+1}) - z_{i+1}\|_2 &= \|y(t_{i+1}) - (y(t_i) + hf(t_i, y(t_i)))\|_2 \\
&\quad \text{(Def. des Eulerverfahrens)} \\
&= \|y(t_{i+1}) - y(t_i) + hy'(t_i)\|_2 \\
&\quad \text{(Def. der gewöhnlichen Differentialgleichung)} \\
&= \left\| \int_{t_i}^{t_{i+1}} y'(\tau)\, d\tau - y'(t_i)h \right\|_2 \\
&\quad \text{(Hauptsatz der Integralrechnung)} \\
&= \left\| \int_{t_i}^{t_{i+1}} y'(\tau) - y'(t_i)\, d\tau \right\|_2 \\
&\leq \|y''\|_{[0,T]} \int_{t_i}^{t_{i+1}} (\tau - t_i)\, d\tau \\
&\quad \text{(Mittelwertsatz der Integralrechnung)} \\
&\leq \frac{1}{2} \|y''\|_{[0,T]} h^2 \,.
\end{aligned}$$

Lokale Fehlerfortpflanzung: Tatsächlich ist das Euler-Verfahren nach i Schritten nicht auf der exakten Lösungskurve. Dieser Fehler lässt sich wie folgt darstellen:

$$y_{i+1} = y_i + hf(t_i, y_i) \text{ bzw. } z_{i+1} = y(t_i) + hf(t_i, y(t_i)) \,.$$

Damit folgt also

$$\begin{aligned}
&\|y_{i+1} - z_{i+1}\|_2 \\
&\leq \quad \|y_i - y(t_i)\|_2 + h\|f(t_i, y_i) - f(t_i, y(t_i))\|_2 \\
&\leq \quad (1 + hL)\|y_i - y(t_i)\|_2 \,.
\end{aligned}$$

Kummulierter Fehler: Jetzt leiten wir eine obere Schranke für die Norm des Gesamtfehlers ε_i nach i Zeitschritten her.

Wir führen den Beweis dieser Behauptung mit Induktion: Für $i = 0$ gilt per Definition des Euler-Verfahrens $y(t_0) = y_0$ und $\varepsilon_i = 0$; damit ist ((2.10)) in diesem Fall trivialerweise erfüllt. Aus den ersten beiden Beweisschritten ergibt sich nun

$$\begin{aligned}\varepsilon_{i+1} &\leq \|z_{i+1} - y_{i+1}\|_2 + \|y(t_{i+1}) - z_{i+1}\|_2 \\ &\leq (1 + hL)\varepsilon_i + \frac{1}{2}\|y''\|_{[0,T]}h^2 \\ &\leq \frac{1}{2L}\left((1 + hL)^{i+1} - 1 - hL + hL\right)\|y''\|_{[0,T]}h \\ &= \frac{(1 + hL)^{i+1} - 1}{2L}\|y''\|_{[0,T]}h \ ,\end{aligned} \tag{2.11}$$

was zu zeigen war. Wegen der elementaren Ungleichung $1 + hL \leq \exp(hL)$ folgt auch der Rest der Behauptung.

□

Aus Satz 2.7 folgt, dass der Fehler des Euler-Verfahrens *linear* in h gegen Null konvergiert, falls das Gitter sukzessive verfeinert wird. Die Auswirkungen des Datenfehlers wird aber in dem Moment kritisch, in dem der Rundungsfehler die Größenordnung des lokalen Fehlers erreicht. Die folgende heuristische Überlegung mag das belegen: Nehmen wir an, im $(i + 1)$-ten Schritt kommt zu den bereits untersuchten Fehlern (lokaler Fehler und fortgepflanzter Fehler) noch ein additiver Rundungsfehler der Größenordnung ε (Maschinengenauigkeit) hinzu. Dann erhalten wir anstelle von (2.11) die Ungleichung

$$\varepsilon_{i+1} \leq (1 + hL)\varepsilon_i + \frac{1}{2}\|y''\|_{[0,T]}h^2 + \varepsilon \ .$$

Induktiv ergibt sich entsprechend

$$\varepsilon_i \leq \frac{\exp(Lih) - 1}{2L}\left(\|y''\|_{[0,T]}h + 2\frac{\varepsilon}{h}\right) \quad i = 0, ..., n \ . \tag{2.12}$$

Das bedeutet: Der Gesamtfehler des Euler-Verfahrens setzt sich aus einem (für $h \to 0$ konvergenten) Verfahrensfehler und einem (für $h \to 0$ divergentem) fortgepflanztem Rundungsfehler zusammen. Man sieht leicht, dass die Schranke auf der rechten Seite von (2.12) für $h \sim \sqrt{\varepsilon}$ ihren minimalen Wert von der Größenordnung $\sqrt{\varepsilon}$ annimmt.

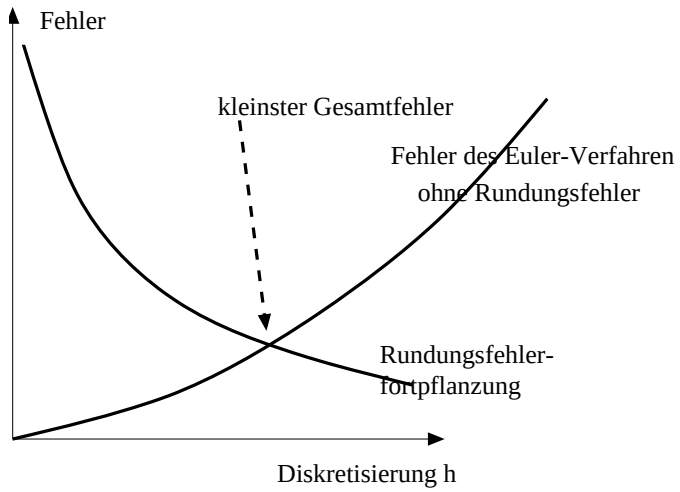

Abbildung 2.2: Gesamtfehler des Euler-Verfahrens

Bemerkung 2.8. Die vorangegangenen Überlegungen zeigen also, dass es keinen Sinn macht, Schrittweiten zu wählen, die kleiner als die Wurzel der Rechengenauigkeit sind.

Im weiteren gehen wir von der allgemeinen Voraussetzung aus, dass f die schwache Lipschitz-Bedingung (2.8) erfüllt.

2.2.2 Implizites Euler Verfahren

Wieder approximiert man die exakte Lösung durch einen linearen Spline, aber im Unterschied zum explizitem Euler-Verfahren, fordert man nun, dass die *linksseitige Ableitung* des Splines im Gitterknoten mit dem Wert von $f(t_i, y_i)$ übereinstimmt. Wie der Name des Verfahrens besagt, kann die Bestimmung des Splines nun nicht mehr explizit erfolgen. Statt dessen ergibt sich y_{i+1} als Lösung des folgenden (im allgemeinen nichtlinearen) Gleichungssystems

$$y_{i+1} = y_i + hf(t_{i+1}, y_{i+1}) \; . \tag{2.13}$$

Beispiel 2.9. Sei $\lambda < 0$. Wir betrachten wiederum die einfache Differentialgleichung

$$y' = \lambda y \, , \qquad y(0) = 1 \; .$$

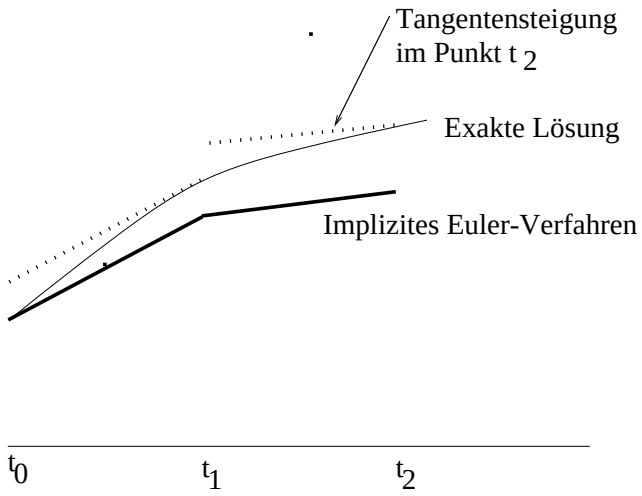

Abbildung 2.3: Schematische Darstellung des impliziten Euler-Verfahrens

Die exakte Lösung ist $y(t) = \exp(\lambda t)$.

Das implizite Euler-Verfahren mit fester Schrittweite hat in diesem einfachen Fall die Form

$$y_{i+1} = y_i + h\lambda y_{i+1} \tag{2.14}$$

oder äquivalent

$$y_{i+1} = \frac{1}{1 - h\lambda} y_i \, . \tag{2.15}$$

Also gilt

$$y_n = \frac{1}{(1 - h\lambda)^n} \, .$$

Mit $T = nh$ ergibt sich

$$y_n = \left(1 - \frac{T}{n}\lambda\right)^{-n} \to \exp(\lambda T) = y(T) \, , \quad n \to \infty \, .$$

Aus (2.15) erkennt man, dass die lokale Fehlerverstärkung in einem Zeitschritt durch (beachte $l = \lambda < 0$)

$$\kappa_{IE} = (1 - \lambda h)^{-1} \in (\exp(\lambda h), 1) = (\kappa, 1)$$

gegeben ist. Da diese Beziehung unabhängig von h gilt, ist das implizite Euler-Verfahren *immer gut konditioniert*. Im Gegensatz zum Euler-Verfahren, welches nur dann gut konditioniert ist, falls $|l|\, h = O(1)$ ist.

Dieses Beispiel zeigt das typische Konvergenzverhalten des impliziten Euler-Verfahren; es ist in vielen Problemen unabhängig von der Wahl von h konvergent. Die Nachteile des impliziten Euler-Verfahrens liegen auf der Hand, es ist die Lösbarkeit des nichtlinearen Gleichung in jedem Iterationsschritt.

Im nächsten Satz studieren wir die Lösbarkeit dieses Gleichungssystems.

Satz 2.10. *Sei $I = [0, T]$, und $f : I \times \mathbb{R}^d \to \mathbb{R}^d$ sei stetig differenzierbar und erfülle die schwache Lipschitz-Bedingung (2.8) für ein $l \in \mathbb{R}$. Dann existiert zu jedem $y \in \mathbb{R}^d$ und jedes $t \in (0, T]$ eine eindeutige Lösung Y der nichtlinearen Gleichung*

$$Y = y + hf(t, Y)\,, \tag{2.16}$$

vorausgesetzt, dass $hl < 1$ ist.

Beweis. Lösungen von (2.16) sind offensichtlich Nullstellen der Funktion

$$F(Y) := y + hf(t, Y) - Y\ .$$

Dabei erfüllt die Funktion F ebenfalls eine schwache Lipschitz-Bedingung:

$$\begin{aligned}\langle F(Y) - F(Z), Y - Z\rangle_2 &= h\,\langle f(t, Y) - f(t, Z), Y - Z\rangle_2 - \|Y - Z\|_2^2 \\ &\le -(1 - hl)\|Y - Z\|_2^2\ .\end{aligned}$$

Aus $1 - hl > 0$ folgt unmittelbar, dass F höchstens eine Nullstelle Y haben kann, also die Eindeutigkeit der Lösung von Y.

Nullstellen Y der Funktion F sind auch die stationären Lösungen $u(t)$ der Differentialgleichung

$$u'(\tau) = F(u(\tau))\ .$$

Beachte: τ ist eine künstlich eingeführte "Zeit".

Eine stationäre Lösung v ist definiert als

$$v = \lim_{t\to\infty} u(t)$$

falls $\lim_{t\to\infty} u'(t) = 0$.

Ist die Funktion f stetig differenzierbar, so ist auch F stetig differenzierbar und daher insbesondere lokal Lipschitz-stetig. Also existieren zu jedem $u_0, v_0 \in \mathbb{R}^d$ Lösungen u und v der Differentialgleichung

$$u' = F(u), u(0) = u_0 \quad v' = F(v), v(0) = v_0\ . \tag{2.17}$$

(vgl. Satz von Picard-Lindelöf 2.3). Ferner gilt nach Satz 2.4 für beliebiges t_0 die Ungleichung

$$\|u(t_0) - v(t_0)\|_2 \leq \exp(-(1 - hl)t_0)\|u_0 - v_0\|_2. \tag{2.18}$$

Da $q := \exp(-(1 - hl)t_0) < 1$ gilt, ist die Abbildung

$$u_0 \to u(t_0)$$

eine kontrahierende Selbstabbildung des $\mathbb{R}^d$, und nach dem Banachschen Fixpunktsatz existiert eine eindeutiger Fixpunkt dieser Abbildung, den wir mit Y bezeichnen. Man beachte, dass sich für jedes t_0 prinzipiell verschiedene Y ergeben.

Jede Lösung der Differentialgleichung (2.17) ist t_0 periodisch

$$v'(\tau) = u'(\tau + t_0) = F(u(\tau + t_0)) = F(v(\tau)) \; .$$

Also sind u und v beides Lösungen von (2.17) mit gleichem Anfangswert $v(0) = u(t_0) = Y = u_0$. Aus der bereits gezeigten Eindeutigkeit folgt nun, dass $u(\tau) = v(\tau) = u(\tau + t_0)$.

Ebenso sieht man, dass für festes $t_1 > 0$ die Funktion $v_1(t) = u(t + t_1)$ eine Lösung der Differentialgleichung (2.17) mit Anfangswert $v_1(0) = u(t_1)$ ist. Aus der t_0 Peridodizität folgt auch die t_0 Periodizität von v_1. Also folgt aus Satz 2.4:

$$\begin{aligned} \|u(t_1) - Y\|_2 &= \|v_1(0) - u(0)\|_2 \\ &= \|v_1(t_0) - u(t_0)\|_2 \\ &\leq \exp(-(1 - hl)t_0)\|v_1(0) - u(0)\|_2 \\ &= q\|u(t_1) - Y\|_2 \; . \end{aligned}$$

Da $q < 1$ ist, muss also $u(t_1) = Y$ sein. Da aber t_1 beliebig war, ergibt sich $u \equiv Y$ und damit ist Y die gesuchte Nullstelle von F. □

Die Bedingung $hl < 1$ ist insbesondere dann erfüllt, wenn l negativ ist. Damit gilt die Bedingung unabhängig von der Konstante L in einer *starken* Lipschitz-Bedingung an f. Im Fall einer positiven Konstante l ergeben sich Einschränkungen an die Schrittweite h.

Nun kommt der angekündigte Konvergenzsatz für das implizite Euler-Verfahren.

Satz 2.11. *Es gelten die Voraussetzungen des Satzes 2.10 an f. Darüberhinaus gelte für die Konstante l im Satz 2.10 hl ¡ 1. Dann gilt für das implizite Euler-Verfahren die Fehlerabschätzung*

$$\|y(t_i) - y_i\|_2 \leq \frac{1}{2l}\left(\left(\frac{1}{1-lh}\right)^i - 1\right)\|y''\|_{[0,T]}h\,, \quad t_i = ih \in I\,. \tag{2.19}$$

Beweis. Der Aufbau des Beweis dieses Satzes ist ähnlich wie der Beweis zu Satz 2.7.

1. **Lokaler Fehler:** Wir betrachten einen Schritt des impliziten Euler-Verfahrens, ausgehend von einem Punkt $(t_i, y(t_i))$ auf der exakten Lösungskurve. Durch Taylorentwicklung ergibt sich

   $$y(t_i) = y(t_{i+1}) - hy'(t_{i+1}) + r_i \text{ mit } \|r_i\|_2 \leq \frac{1}{2}\|y''\|_{[0,T]}h^2\,.$$

   Damit ergibt für die Approximation z_{i+1} aus dem Euler-Verfahren ein Fehler

   $$\begin{aligned} z_{i+1} - y(t_{i+1}) &= y(t_i) + hf(t_{i+1}, z_{i+1}) - y(t_{i+1}) \\ &= y(t_{i+1}) - hy'(t_{i+1}) + r_i + hf(t_{i+1}, z_{i+1}) - y(t_{i+1} \\ &= h(f(t_{i+1}, z_{i+1}) - f(t_{i+1}, y(t_{i+1}))) + r_i\,. \end{aligned}$$

   Damit gilt

   $$\|z_{i+1} - y(t_{i+1})\|^2 \leq h\,\langle f(t_{i+1}, z_{i+1}) - f(t_{i+1}, y(t_{i+1})), z_{i+1} - y(t_{i+1})\rangle_2 + \langle r_i, z_{i+1} - y(t_{i+1})\rangle_2$$

   Durch Verwendung der schwachen Lipschitz-Bedingung (2.8) ergibt sich dann die Ungleichung

   $$\|z_{i+1} - y(t_{i+1})\|_2^2 \leq lh\|z_{i+1} - y(t_{i+1})\|_2^2 + \|r_i\|_2\|z_{i+1} - y(t_{i+1})\|_2\,,$$

   woraus folgt

   $$\|z_{i+1} - y(t_{i+1})\|_2 \leq \frac{1}{2(1-lh)}\|y''\|_{[0,T]}h^2\,.$$

2. **Lokale Fehlerfortpflanzung:** Ausgehend von y_i bzw. $y(t_i)$ ergeben sich im i-ten Schritt zwei verschiedene Näherungen von $y(t_{i+1})$,

   $$y_{i+1} = y_i + hf(t_{i+1}, y_{i+1}) \text{ bzw. } z_{i+1} = y_i + hf(t_{i+1}, z_{i+1})\,.$$

Damit ist

$$y_{i+1} - z_{i+1} = h(f(t_{i+1}, y_{i+1}) - f(t_{i+1}, z_{i+1})) + (y_i - y(t_i)) \ .$$

Wie im ersten Beweisschritt ergibt sich durch Multiplikation mit $y_{i+1} - z_{i+1}$ schließlich

$$\|y_{i+1} - z_{i+1}\|_2 \leq \frac{1}{1 - lh} \|y_i - y(t_i)\|_2 \ .$$

3. **Kumulierter Fehler:** Für den Gesamtfehler ε_i nach i-Zeitschritten ergibt sich daher beim impliziten Euler-Verfahren die Rekursion

$$\varepsilon_{i+1} \leq \frac{1}{1 - lh} \varepsilon_i + \frac{1}{2(1 - lh)} \|y''\|_{[0,T]} h^2 \ .$$

Durch Induktion ergibt sich die Behauptung.

□

Aus diesem Satz folgt unmittelbar

Korollar 2.12. *Es gelten die Voraussetzungen von Satz 2.11 mit einem $l < 0$. Dann gilt für alle $t_i \in [0, T]$ die Abschätzung*

$$\|y(t_i) - y_i\|_2 \leq \frac{1}{2\,|l|} \|y''\|_{[0,T]} h \ .$$

Zur Implementierung des impliziten Euler-Verfahrens muss man das nichtlineare Gleichungssystem (2.16) effizient lösen. Typischerweise verwendet man ein Newton (-artiges) Verfahren

$$y_{i+1}^{(n+1)} = y_{i+1}^{(n)} - \left(I - h f_y(t_{i+1}, y_{i+1}^{(n)})\right)^{-1} \left(y_{i+1}^{(n)} - y_i - h f(t_{i+1}, y_{i+1}^{(n)})\right) \ ,$$
$$n = 0, 1, 2, ...$$

Die Berechnung der **Jacobi-Matrix** $J = f_y(t_{i+1}, y_{i+1}^{(n)})$ und der Inversen von $\mathbb{1} - hJ$ sind meist sehr aufwendig. Deshalb verwendet man anstelle des Newton-Verfahrens zumeist **Quasi-Newton Verfahren**, bei der die Jacobi-Matrix geeignet approximiert wird.

2.2.3 Runge-Kutta Verfahren

Der Nachteil der beiden Euler-Verfahren ist ihre langsame Konvergenz (in Abhängigkeit der Zeitdiskretisierung). Konvergenzverbesserungen kann man mit einem Ansatz der Form

$$y_{i+1} = y_i + h \sum_{j=1}^{s} b_j f(t_i + c_j h, n_j)\,, \quad \sum_{j=1}^{s} b_j = 1\,, \tag{2.20}$$

gewährleisten, wobei η_j Näherungen für $y(t_i + c_j h)$ sind. s nennt man dabei die *Stufenzahl* des Verfahrens.

Speziell gilt:

Explizites Euler-Verfahren: $s = 1$ und $c_1 = 0$, $\eta_1 = y_i$,

Implizites Euler-Verfahren: $s = 1$ und $c_1 = 1$, $\eta_1 = y_{i+1}$.

Da bei (2.20) jeweils ausgehend von $y_i \approx y(t_i)$ die nächste Näherung $y_{i+1} \approx y(t_{i+1})$ berechnet wird, spricht man bei Verfahren dieser Art von *Einschrittverfahren*. Im Gegensatz dazu verwenden *Mehrschrittverfahren* auch ältere Näherungen $y_{i-1}, ...,$ zur Berechnung von y_{i+1}.

Nehmen wir nun an, dass $y_i = y(t_i)$ auf der exakten Lösungskurve liegt und bestimmen davon ausgehend, wie im ersten Schritt des Beweises von Satz 2.7, den lokalen Fehler des Verfahrens (2.20). In diesem Fall ergibt sich mit dem Hauptsatz der Differentialrechnung:

$$\begin{aligned} y(t_{i+1}) - y_{i+1} &= y(t_{i+1}) - y(t_i) - h \sum_{j=1}^{s} b_j f(t_i + c_j h, n_j) \\ &\qquad \text{(Annahme, dass } y(t_i) = y_i\text{)} \\ &= \int_{t_i}^{t_{i+1}} y'(t)\, dt - h \sum_{j=1}^{s} b_j f(t_i + c_j h, n_j) \\ &= (\text{Dgl.}) \int_{t_i}^{t_{i+1}} f(t, y(t))\, dt \; - h \sum_{j=1}^{s} b_j f(t_i + c_j h, n_j)\,. \end{aligned}$$

Wir sehen daher, dass der lokale Fehler klein wird, falls die Summe $h \sum_{j=1}^{s} b_j f(t_i + c_j h, n_j)$ eine gute Approximation des entsprechenden Integrals $\int_{t_i}^{t_{i+1}} f(t, y(t))\, dt$ ist. Daher liegt es nahe, *Quadraturformeln* zur Wahl der Parameter $\{b_j\}$, $\{c_j\}$ und $\{\eta_j\}$ heranzuziehen.

Beispiel 2.13. Mit der Mittelpunktsformel ergibt sich der Ansatz

$$y_{i+1} = y_i + hf(t_i + h/2, \eta_1) \,, \tag{2.21}$$

wobei idealerweise $\eta_1 = y(t_i + h/2)$ sein sollte; allerdings ist dieser Wert nicht bekannt. Eine Näherung kann jedoch leicht gefunden werden durch

$$\eta_1 = y(t_i) + \frac{h}{2} y'(t_i) \approx y_i + \frac{h}{2} f(t_i, y_i) \,.$$

Das ist das *Verfahren von Runge* aus dem Jahre 1895.

Ein andere Alternative ist die Trapezregel. Sie ergibt sich aus dem Ansatz

$$y_{i+1} = y_i + \frac{h}{2} f(t_i, y_i) + \frac{h}{2} f(t_{i+1}, \tilde{\eta}_1) \,,$$

wobei nun $\tilde{\eta}_i = y(t_i + h)$ sein sollte. Geht man wie beim Verfahren vom Runge vor und ersetzt man

$$\tilde{\eta}_1 = y_i + hy'(t_i) \,,$$

dann ergibt sich das *Verfahren von Heun*.

Wir studieren das Verfahren von Runge etwas genauer. Bei diesem Verfahren ergibt sich die Taylorentwicklung für hinreichend glattes f unter der Voraussetzung $y_i = y(t_i)$

$$\begin{aligned} y_{i+1} &= y_i + hf(t_i, y_i) + \frac{h^2}{2} f_t(t_i, y_i) + \frac{h^2}{2} f_y(t_i, y_i) f(t_i, y_i) + O(h^3) \,, \\ y(t_{i+1}) &= y(t_i) + hy'(t_i) + \frac{h^2}{2} y''(t_i) + O(h^3) \\ &= y_i + hf(t_i, y_i) + \frac{h^2}{2} \left(f_t(t_i, y_i) + f_y(t_i, y_i) f(t_i, y_i) \right) + O(h^3) \,. \end{aligned}$$

Damit gilt für den lokalen Fehler

$$\| y(t_{i+1}) - y_{i+1} \|_2 = O(h^3) \,.$$

Das Verfahren von Runge hat also einen kleineren lokalen Fehler als die beiden Euler-Verfahren.

Definition 2.14. Ein Einschrittverfahren hat die *Ordnung* q, falls für jede Differentialgleichung $y' = f(t, y)$ mit $f \in C^{q+1}(I \times J)$ und jedes $t_i \in I$ für den lokalen Fehler gilt:

$$y_i = y(t_i) \in J \to \| y_{i+1} - y(t_{i+1}) \|_2 = O(h^{q+1}) \,, \quad h \to 0 \,.$$

Beachte: Die Ordnung ist q (und nicht $q+1$), obwohl die entsprechende h-Potenz $q+1$ ist. Wie wir später sehen werden, ist die Konvergenzordnung an einem festen Punkt $t_0 \in (0,T]$ bei einem Verfahren der Ordnung q nämlich $O(h^q)$.

Beispiel 2.15. Die beiden Euler-Verfahren haben die Ordnung $q=1$ und das Verfahren von Runge hat die Ordnung $q=2$.

Die *Runge-Kutta Verfahren* beruhen auf folgender Wahl der Koeffizienten η_j:

$$\eta_j \approx y(t_i + c_j h) = y(t_i) + \int_{t_i}^{t_i + c_j h} y'(t)\,dt \; = y(t_i) + \int_{t_i}^{t_i + c_j h} f(t, y(t))\,dt \; . \tag{2.22}$$

Zur Auswertung des Integrals bieten sich Quadraturformeln an, wobei wir uns darauf einschränken $f(t,y)$ nur an den gleichen Knotenpunkten $f(t_i + c_j h, \eta_j)$, $j = 1, ..., s$, auszuwerten, wie sie zur Berechnung von y_{i+1} herangezogen werden. Das ergibt folgenden Ansatz:

$$\eta_j = y_i + h \sum_{k=1}^{s} a_{jk} f(t_i + c_k h, \eta_k)\,, \quad \sum_{k=1}^{s} a_{jk} = c_j \; . \tag{2.23}$$

Falls $a_{jk} = 0$ für $j \leq k$ ist die Rechenvorschrift *explizit*, das führt zu *expliziten Runge-Kutta* Verfahren. Ansonsten ergibt sich ein *implizites Runge-Kutta* Verfahren. Die Bedingung $\sum_{k=1}^{s} a_{jk} = c_j$ ist aus Quadraturformelmethoden motiviert, wird aber in der Literatur nicht einheitlich verwendet.

Üblicherweise werden die Koeffizienten $\{a_{jk}, b_j, c_j\}$ in einem quadratischen Tableau zusammengefasst (das so genannte **Runge-Kutta Abc**),

$$\begin{array}{c|c} \mathrm{c} & \mathrm{A} \\ \hline & \mathrm{b}^t \end{array} \quad = \quad \begin{array}{c|ccccc} c_1 & a_{1,1} & ... & ... & ... & a_{1,s} \\ c_2 & a_{2,1} & a_{2,2} & ... & ... & ... \\ c_3 & a_{3,1} & a_{3,2} & ... & ... & ... \\ ... & ... & ... & ... & ... & ... \\ c_s & a_{s,1} & ... & ... & a_{s,s-1} & a_{s,s} \\ \hline & b_1 & b_2 & ... & b_{s-1} & b_s \end{array}$$

wobei wir $A = [a_{j,k}] \in \mathbb{R}^{s \times s}$, $b = [b_1, ..., b_s]^t \in \mathbb{R}^s$ und $c = [c_1, ..., c_s]^t \mathbb{R}^s$ gesetzt haben. Wir sprechen ab nun von dem Runge-Kutta Verfahren (A, b, c).

Beispiel 2.16. Für das explizite und implizite Euler Verfahren ergeben sich folgende Tableaus:

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \qquad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

Um das Verfahren von Runge (hier in zwei Stufen aufgeschrieben)

$$\begin{aligned} \eta_1 &= y_i + \tfrac{h}{2} f(t_i, y_i) \\ y_{i+1} &= y_i + h f(t_i + h/2, \eta_1) \end{aligned}$$

in Runge-Kutta Form zu bringen setzen wir $c_0 = 0$, $c_1 = 1/2$, $y_i = \eta_0$, und wir erhalten die äquivalente Schreibweise:

$$\begin{aligned} \eta_0 &= y_i + h \textstyle\sum_{k=0}^{1} 0 \cdot f(t_i + c_k h, \eta_k) \\ \eta_1 &= y_i + \tfrac{h}{2} f(t_i + h/2, \eta_0) \\ y_{i+1} &= y_i + h f(t_i + h/2, \eta_1) \ . \end{aligned}$$

Diese drei Gleichungen werden mit folgendem Runge-Kutta Tableau beschrieben

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 0 & 1 \end{array}$$

Wir konstruieren nun ein Verfahren der Ordnung 3 und leiten uns dazu Bedingungen für die Parameter des Runge-Kutta Schemas her.

Satz 2.17. *Runge-Kutta Verfahren haben mindestens die Ordnung 1. Ein Runge-Kutta Verfahren* (2.20), (2.23) *ist von zweiter Ordnung, wenn*

$$\sum_{j=1}^{s} b_j c_j = \frac{1}{2} \tag{2.24}$$

Es ist von dritter Ordnung, wenn zusätzlich

$$\sum_{j=1}^{s} b_j c_j^2 = \frac{1}{3} \text{ und } \sum_{j=1}^{s} b_j \sum_{k=1}^{s} a_{jk} c_k = \frac{1}{6} \ . \tag{2.25}$$

Beweis. Falls die Funktion f in der Differentialgleichung hinreichend oft differenzierbar ist, dann gilt aufgrund einer Taylorentwicklung

$$\begin{aligned} y(t_i + h) &= y(t_i) + h y'(t_i) + \frac{1}{2} h^2 y''(t_i) + \frac{1}{6} h^3 y''(t_i) + O(h^4) \\ &= y(t_i) + h f(t_i, y(t_i)) + \\ &\quad \frac{1}{2} h^2 (f_t + f_y f)(t_i, y(t_i)) + \\ &\quad \frac{1}{6} h^3 (f_{tt} + 2 f_{ty} f + f_t f_y + f^t f_{yy} f + f_y^2 f)(t_i, y(t_i)) + \\ &\quad O(h^4) \ . \end{aligned} \tag{2.26}$$

Man beachte, dass f, f_t, f_{tt} Vektoren sind, f_{ty} eine Matrix ist und f_{yy} ein Tensor ist;

$$(f^t f_{yy} f)_i = f f_{iyy} f \,,$$

wobei f_{iyy} die Hessematrix der Komponente f_i ist.

Wir entwickeln nun die y_{i+1} in Abhängigkeit von h in eine Taylorreihe. Dazu verwenden wir, dass wegen (2.23) gilt

$$\begin{aligned}
\eta_j - y_i &= h \sum_{k=1}^{s} a_{jk} f(t_i + c_k h, \eta_k) \\
&= h \sum_{k=1}^{s} a_{jk} f(t_i, y_i) + O(h^2) \\
&= h c_j f(t_i, y_i) + O(h^2) \,.
\end{aligned}$$

Damit gilt also

$$\begin{aligned}
\eta_j &= y_i + h \sum_{k=1}^{s} a_{jk} f(t_i + c_k h, \eta_k) \\
&= y_i + h \sum_{k=1}^{s} a_{jk} \left(f(t_i, y_i) + f_t(t_i, y_i) c_k h + f_y(t_i, y_i)(\eta_k - y_i) + O(h^2) \right) \\
&= y_i + h f(t_i, y_i) \sum_{k=1}^{s} a_{jk} + h^2 \Bigg(f_t(t_i, y_i) \sum_{k=1}^{s} a_{jk} c_k + \ldots \\
&\quad f_y(t_i, y_i) f(t_i, y_i) \sum_{k=1}^{s} a_{jk} c_k \Bigg) + O(h^3) \,.
\end{aligned}$$

Damit gilt

$$\eta_j = y_i + h f(t_i, y_i) c_j + h^2 (f_t + f_y f)(t_i, y_i) \sum_{k=1}^{s} a_{jk} c_k + O(h^3) \,.$$

Damit ergibt sich schließlich aus (2.20), wobei wir jetzt zur Vereinfachung der Notation immer wenn wir von f oder einer Ableitung sprechen, die Auswer-

tung an der Stelle (t_i, y_i)) meinen:

$$\begin{aligned}
y_{i+1} &= y_i + h\sum_{j=1}^{s} b_j f(t_i + c_j h, \eta_j) \\
&= y_i + h\sum_{j=1}^{s} b_j \Big(f + f_t c_j h + f_y(\eta_j - y_i) + \frac{1}{2} f_{tt} c_j^2 h^2 \\
&\quad + c_j h f_{ty}(\eta_j - y_i) + \frac{1}{2}(\eta_j - y_i) f_{yy}(\eta_j - y_i) + ... \Big) \\
&= y_i + fh\sum_{j=1}^{s} b_j \\
&\quad + h^2 f_t \sum_{j=1}^{s} b_j c_j \\
&\quad + h^2 f_y f \sum_{j=1}^{s} b_j c_j \\
&\quad + h^3 f_y (f_t + f_y f) \sum_{j=1}^{s} b_j \sum_{k=1}^{s} a_{jk} c_k \\
&\quad + h^3 f_{tt} \frac{1}{2} \sum_{j=1}^{s} b_j c_j^2 \\
&\quad + h^3 f_{ty} f \sum_{j=1}^{s} b_j c_j^2 \\
&\quad + h^3 f^t f_{yy} f \frac{1}{2} \sum_{j=1}^{s} b_j c_j^2 \\
&\quad + O(h^4) \\
&= y_i + hf \\
&\quad + h^2 (f_t + f_y f) \sum_{j=1}^{s} b_j c_j \\
&\quad + h^3 (f_{tt} + 2 f_{ty} f + f^t f_{yy} f) \frac{1}{2} \sum_{j=1}^{s} b_j c_j^2 \\
&\quad + h^3 f_y (f_t + f_y f) \sum_{j=1}^{s} b_j \sum_{k=1}^{s} a_{jk} c_k \\
&\quad + O(h^4) \ ,
\end{aligned}$$

wobei in der letzten Identität verwendet haben, dass $\sum_{j=1}^{s} b_j = 1$ gilt.

Um die Konsistenzordnung des Runge-Kutta Verfahrens bestimmen zu können, vergleichen die Darstellung für y_{i+1} aus obiger Formel mit (2.26) unter der Voraussetzung $y_i = y(t_i)$. Demnach ist jedes Runge-Kutta Verfahren von erster Ordnung, es hat die Ordnung 2, falls (2.24) gilt, und es hat die Ordnung 3 falls zusätzlich (2.25) erfüllt ist. □

Man überprüft sofort, dass beim Verfahren von Runge (2.21) die Bedingung (2.24) erfüllt ist, nicht aber die beiden Bedingungen in (2.25).

Beispiel 2.18. Wann immer in der Literatur von dem Runge-Kutta Verfahren gesprochen wird, dann ist das folgende Verfahren von *Kutta* (1901) auf der Basis der Simpson-Formel gemeint. Die Koeffizienten dieses Verfahrens lauten:

$$c_1 = 0\,, \quad c_2 = 1/2\,, \quad c_3 = 1/2, \quad c_4 = 1$$

mit den Gewichten

$$b_1 = 1/6\,, \quad b_2 = 1/3\,, \quad b_3 = 1/3, \quad b_4 = 1/6\,.$$

Berücksichtigt man, dass das Verfahren explizit ist, so reduziert sich die verbleibende Ordnungsbedingung in (2.25) zu

$$\frac{1}{6}a_{32} + \frac{1}{12}a_{42} + \frac{1}{12}a_{43} = \frac{1}{6}\,,$$

sodass die Bestimmung der $\{a_{jk}\}$ in dieser Weise unterbestimmt ist. Erweitert man Satz 2.17, dann ergibt sich eine eindeutige Lösung aller Ordnungsbedingungen für ein explizites Verfahren vierter Ordnung, die im folgenden Tableau dargestellt ist:

0				
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	1/3	1/3	1/6

Ein weiteres häufig verwendetes explizites Runge-Kutta Verfahren läuft unter dem Namen *dopri5*. Es ist ein sechsstufiges Verfahren mit noch höherer

Konsistenzordnung (aber es ist auch sehr stabil). Das Runge-Kutta Tableau findet man z.B. in [6, 7])

0						
$\frac{1}{5}$	$\frac{1}{5}$					
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$				
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$			
$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$		
1	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$	
	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$

Im folgenden beweisen wir Konvergenzaussagen für allgemeine Runge-Kutta Verfahren:

Satz 2.19. *Sei $I = [0, T]$ und sei $f : I \times \mathbb{R}^d \to \mathbb{R}^d$ $q + 1$ mal stetig differenzierbar und gleichmäßig Lipschitz-stetig bzgl. der zweiten Komponente und das (explizite oder implizite) Runge-Kutta Verfahren (A, b, c) habe die Ordnung q. Dann existiert ein $h_0 > 0$, sodass für alle $h \in (0, h_0)$ alle Näherungen des Runge-Kutta Verfahrens für $t_i = ih \in I$ eindeutig definiert sind. Ferner gilt*

$$\|y(t_i) - y_i\|_2 \leq Ch^q\,, \quad h < h_0\,,$$

wobei die Konstante unabhängig von i und h ist. [1]

Beweis. Wir gehen analog zu den Beweisen für die Konvergenzordnung von Euler und impliziten Euler vor. Die Aussage über den lokalen Fehler ist natürlich nicht mehr zu beweisen, weil wir vorausgesetzt haben, dass die Ordnung $O(h^q)$ ist. Es gilt also, mit der gleichen Notation wie bei Euler-Verfahren, dass

$$\|y(t_{i+1}) - z_{i+1}\|_2 \leq C_1 h^{q+1}\,. \tag{2.27}$$

Im folgenden studieren wir die Existenz der Koeffizienten $\{\eta_k\}$ des Runge-Kutta Verfahren: Wir bezeichnen

$$\phi_j(u_j, \eta) = u_j + h \sum_{\nu=1}^{s} a_{j\nu} f(t_i + c_\nu h, \eta_\nu)\,, \quad j = 1, \ldots, s\,.$$

[1] Man beachte: In Definition 2.14 ist vorausgesetzt, dass aus $y(t_i) = y_i$ folgt, dass $y(t_{i+1}) - y_{i+1} = O(h^{q+1})$. Die Konsistenzordnung berücksichtigt also keine kummulierten Fehler. Diese Kummulation der Fehler bewirkt schließlich, dass der *Konvergenzfehler* von der Ordnung q ist.

Damit sieht man, dass das Runge-Kutta Verfahren wohldefiniert ist, wenn das Gleichungssystem

$$\eta = \Phi(y, \eta) = (\Phi_i(y_i, \eta))_{i=1,\dots,s}$$

lösbar ist. Dazu zeigen wir, dass $\Phi(y, \cdot)$ eine Kontraktion auf $\mathbb{R}^{sd}$ ist, indem wir zeigen, dass der Betrag von Φ_η (Ableitung von Φ bezüglich η) gleichmäßig durch 1 beschränkt ist. Da

$$\Phi_\eta(y, \eta) = h \begin{bmatrix} a_{11} f_y(t_i + c_1 h, \eta_1) & \cdots & a_{1s} f_y(t_i + c_s h, \eta_s) \\ \vdots & & \vdots \\ a_{s1} f_y(t_i + c_1 h, \eta_1) & \cdots & a_{ss} f_y(t_i + c_s h, \eta_s) \end{bmatrix}.$$

Ist L eine obere Schranke für $\|f_y\|_\infty$, so gilt somit

$$\|\Phi_\eta(y, \eta)\|_\infty \leq hL\|A\|_\infty \,.$$

Das zeigt, dass die Funktion $\Phi(y, \cdot)$ für $h < h_0 = 1/(2L\|A\|_\infty$ eine Kontraktion mit Kontraktionsfaktor $1/2$ ist. Somit existiert eine Lösung der Fixpunktgleichung.

Lokale Fehlerfortpflanzung: Bezeichnen wir wieder mit z_{i+1} und y_{i+1} die Näherungen des Runge-Kutta Verfahrens, wenn wir mit $y(t_i)$ und y_i starten. Dann gilt:

$$\begin{aligned} \|y_{i+1} - z_{i+1}\|_\infty &= \|y(t_i) - y_i + h \sum_{j=1}^{s} b_j \left(f(t_i + c_j h, \eta_j(y(t_i))) - f(t_i + c_j h, \eta_j(y_i))\right)\|_\infty \\ &\leq \|y(t_i) - y_i\|_\infty + h\|b\|_1 L \|\eta_j(y(t_i)) - \eta_j(y_i)\|_\infty \\ &\leq (1 + hC_2)\|y(t_i) - y_i\|_\infty \,, \end{aligned}$$

mit $C_2 = 2\|b\|_1 L$.

Kummulierter Fehler: Für den Gesamtfehler gilt somit mit $h = T/n$ für alle $i = 0, \dots, n$:

$$\begin{aligned} \|y_{i+1} - z_{i+1}\|_\infty &\leq (1 + hC_2)\|y(t_i) - y_i\|_\infty + C_1 h^{q+1} \\ &\underbrace{\leq}_{\text{Induktion}} \frac{C_1}{C_2}(1 + 2C_2 h)^i h^q \\ &\leq \frac{C_1}{C_2}(1 + 2C_2 T/n)^n h^q \\ &\leq \frac{C_1}{C_2} e^{2C_2 T} h^q \,. \end{aligned}$$

□

2.2.4 Stabilitätstheorie

Die Motivation zur Betrachtung der Runge-Kutta Verfahren war die langsame Konvergenzgeschwindigkeit der Euler-Verfahren zu verbessern. Nicht berücksichtigt wurde die Stabilität des Verfahrens, d.h., die Anfälligkeit der Rekursion (2.20) gegenüber Fehlern. Wir werden nun untersuchen, wie Approximationsfehler in y_i in den $(i+1)$-ten Punkt fortgepflanzt werden. Dazu verfolgen wir einem allgemeinen Prinzip bestehend aus drei Schritten.

Linearisierung: Wir interessieren uns für den Einfluss einer *kleinen* Störung u_i des exakten Wertes $y(t_i)$. Bezeichnen wir mit y_u die Lösung von

$$y_u' = f(t, y_u) \text{ mit } y_u(0) = y(0) + u_0 \,,$$

so gilt

$$u' := y_u' - y' = f(t, y_u) - f(t, y) \approx f_y(t, y)(y_u - y) = f_y(t, y)u'$$

mit $u(0) = u_0$. Dies ist eine *lineare* Differentialgleichung für u.

Einfrieren der Zeit: Wir untersuchen ein kurzes Zeitintervall $\Delta = h$ und vernachlässigen den Einfluss der Zeit in $f_y(t, y)$, gehen also davon aus, dass sich die Lösung der linearisierten Gleichung wie die Lösung einer stationären Differentialgleichung

$$u' = Au \text{ mit } A = f_y(t_i, y_i) \in \mathbb{R}^{d \times d} \,,$$

mit Anfangswerten $u(t_i) = u_i$ ist.

Diagonalisierung: Wir setzen voraus, dass die Matrix A diagonalisierbar ist: Es existiert also eine Basis $\{x_1, ..., x_d\} \subseteq \mathbb{R}^d$ mit $Ax_n = \lambda_n x_n \,, n = 1, ..., d$. Entwickelt man $u(t) = \sum_{n=1}^{d} \eta_n(t) x_n$, dann ergibt sich

$$\eta_n{}'(t) = \lambda_n \eta_n \,, \quad n = 1, ..., d \,.$$

Die Differentialgleichungen für η liefern eine Approximation der ursprünglichen Gleichung. Stabilitätsaussagen werden nun nur für die Funktionen η untersucht, und dann wird angenommen, dass die ursprüngliche Problemstellung die gleiche Stabilität aufweist.

Im Rest dieses Abschnitts betrachten wir daher nur noch die *eindimensionale* Testgleichung

$$y' = \lambda y\,, \quad y(0) = 1\,, \quad \lambda \in \mathbb{C}\,. \tag{2.28}$$

Wir gehen davon aus, dass das, was wir für diese einfache Gleichung beweisen, viel allgemeiner gilt. Das stimmt in vielen Fällen, aber natürlich nicht immer.

Die Lösung $y(t) = \exp(\lambda t)$ verhält sich dabei in Abhängigkeit von λ wie folgt:

$\Re(\lambda) > 0 : |y(t)| \to \infty$ für $t \to \infty$.

$\Re(\lambda) < 0 : |y(t)| \to 0$ für $t \to \infty$.

$\Re(\lambda) = 0 : |y(t)| = |y_0|$ für alle $t \in \mathbb{R}_0^+$.

Wir führen nun einige Stabilitätsbegriffe ein. Die Sinnhaftigkeit dieser Definition wird sich im nächsten Kapitel zeigen.

Definition 2.20. Seien $\{y_n\}$ die Näherungen eines numerischen Verfahrens zur Lösung der Testgleichung (2.28). Dann bezeichnet man das Verfahren als

- *A-stabil*, falls für jedes $\lambda \in \mathbb{C}$ mit $\Re(\lambda) \leq 0$ und $h > 0$ mit $\frac{1}{\lambda h} \notin \sigma(A)$ gilt:
$$|y_{n+1}| \leq |y_n| \text{ für alle } n \in \mathbb{N}_0\,.$$
- Das Verfahren heißt *Isometrie erhaltend* wenn für jedes $\lambda \in \mathbb{C}$ mit $\Re(\lambda) = 0$ und $h > 0$ mit $\frac{1}{\lambda h} \notin \sigma(A)$ gilt:
$$|y_{n+1}| = |y_n| = |y_0| \text{ für alle } n \in \mathbb{N}.$$

Mit

$$R : \mathbb{C} \backslash \left\{ \frac{1}{\sigma} : 0 \neq \sigma \in \sigma(A) \right\} \to \mathbb{C}\,,$$
$$\zeta \to 1 + \zeta b^* (\mathbb{1} - \zeta A)^{-1} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

bezeichnet man *Stabilitätsfunktion* des Runge-Kutta Verfahrens mit den Koeffizienten (A, b, c).

Bemerkung 2.21. Man beachte, dass die Stabilitätsfunktion nicht von der Wahl der Knoten $\{c_j\}$ abhängt. Das macht Sinn, da Stabilität nur für die autonome Differentialgleichung (2.28) untersucht wird. Weiters gilts

$$R(0) = 1\,,$$

und für $\zeta \in B_\rho(0)$, mit $\rho < 1/(\max|\sigma(A)|)$, ist R wohldefiniert.

Wenn das Runge-Kutta Verfahren explizit ist, dann ist $A \in \mathbb{R}^{s\times s}$ eine strikte untere Dreiecksmatrix und folglich gilt $A^s = 0$. Demnach ergibt sich

$$(\mathbb{1} - \zeta A)^{-1} = \mathbb{1} + \zeta A + ... + \zeta^{s-1}A^{s-1}\,, \tag{2.29}$$

sodass $R(\zeta) = 1 + \zeta b^*(\mathbb{1} - \zeta A)^{-1}\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$ ein Polynom vom Grad s ist. Damit ist $\sigma(A) = \{0\}$ uns die Funktion R ist auf ganz $\mathbb{C}$ definiert.

Also, ist das Runge-Kutta Verfahren explizit, dann ist R ein Polynom.

Wir verwenden im folgenden die Bezeichnung

$$\mathbb{C}^- = \{\lambda : \Re(\lambda) \le 0\}\ .$$

Satz 2.22. *Sie $h > 0$ und $\lambda \in \mathbb{C}$ mit $\frac{1}{h\lambda} \notin \sigma(A)$. Für die Testgleichung* (2.28) *kann das Runge-Kutta Verfahren geschrieben werden als*

$$y_n = (R(h\lambda))^n \quad n = 0, 1, ... \tag{2.30}$$

Darüberhinaus ist die Funktion R eine rationale Funktion mit Zähler- und Nennergrad höchstens s.

Beweis. Wir führen die folgenden s-dimensionalen Vektoren ein:

$$\eta = \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_s \end{bmatrix}, \quad \tau = \begin{bmatrix} t_i + c_1 h \\ \vdots \\ t_i + c_s h \end{bmatrix}, \quad f(\tau, \eta) = \begin{bmatrix} f(\tau_1, \eta_1) \\ \vdots \\ f(\tau_s, \eta_s) \end{bmatrix} = \lambda\eta\ .$$

Aus der Definition des Runge-Kutta Verfahren (2.20), (2.23) ergibt sich y_{n+1} aus y_n durch Lösen des Gleichungssystems

$$\begin{aligned} \eta &= y_n \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + hAf(\tau, \eta) = y_n \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + h\lambda A\eta\,, \\ y_{n+1} &= y_n + hb^* f(\tau, \eta) = y_n + h\lambda b^*\eta\ . \end{aligned} \tag{2.31}$$

Oder anders angeschrieben $(\mathbb{1} - h\lambda A)\eta = y_n \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$ und somit

$$y_{n+1} = y_n + h\lambda b^*(\mathbb{1} - h\lambda A)^{-1} \left(y_n \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right) = R(h\lambda) y_n \, .$$

Mit Induktion folgt die Darstellung (2.30).

Um zu sehen, dass, im allgemeinen, R eine rationale Funktion ist, verwenden wir die Cramersche Regel, wonach sich die Komponenten von

$$\eta = (\mathbb{1} - h\lambda A)^{-1} \left(y_n \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right)$$

in folgender Form schreiben lassen

$$\eta_k = \frac{p_k(h\lambda)}{det(\mathbb{1} - h\lambda A)} \, ,$$

wobei p_k, $k = 1, ..., s-1$ jeweils ein Polynom vom Grad maximal $s-1$ ist. Damit folgt die Behauptung aus (2.31). □

Beispiel 2.23.

- Zum impliziten Euler Verfahren gehört die Stabilitätsfunktion $R(\zeta) = \frac{1}{1-\zeta}$.
- Zum expliziten Euler Verfahren gehört die Stabilitätsfunktion $R(\zeta) = 1 + \zeta$.
- Für *das* Runge-Kutta Verfahren ergibt sich mit Hilfe von (2.29) die Stabilitätsfunktion

$$\begin{aligned} R_4(\zeta) &= 1 + \zeta[1/6, 1/3, 1/3, 1/6] \begin{bmatrix} 1 & 0 & 0 & 0 \\ \zeta/2 & 1 & 0 & 0 \\ \zeta^2/4 & \zeta/2 & 1 & 0 \\ \zeta^3/4 & \zeta^2/2 & \zeta & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \\ &= 1 + \zeta[1/6, 1/3, 1/3, 1/6] \begin{bmatrix} 1 \\ 1 + \zeta/2 \\ 1 + \zeta/2 + \zeta^2/4 \\ 1 + \zeta/2 + \zeta^2/2 + \zeta^3/4 \end{bmatrix} \\ &= 1 + \zeta + \zeta^2/2 + \zeta^3/6 + \zeta^4/24 \, . \end{aligned}$$

Man erkennt, dass $R_4(\zeta)$ gerade aus den ersten fünf Summanden der Taylorreihe $\exp(\zeta)$ besteht. Das hat einen tieferen Hintergrund, wie wir aus folgendem Resultat sehen.

Satz 2.24. *Ist R die Stabilitätsfunktion eines Einschrittverfahrens der Ordnung q, dann gilt*

$$R(\zeta) = \exp(\zeta) + O(|\zeta|^{q+1}), \quad \zeta \to 0 .$$

Beweis. $R(\zeta)$ und $\exp(\zeta)$ können beide in eine lokal konvergente Reihe um $\zeta = 0$ entwickelt werden. Anwendung des Runge-Kutta Verfahrens auf die Testgleichung (2.28) mit $\lambda = 1$ (beachte, dass die Lösung dieser Testgleichung $\exp(\zeta)$ ist) und Schrittweite $h > 0$ liefert, unter Berücksichtigung von Satz 2.22

$$y_1 - \underbrace{\exp(h)}_{=y(h)} \underbrace{=}_{\text{Ordnung } q} = O(h^{q+1}) \text{ for } h \to 0 .$$

Damit müssen die ersten q Terme der beiden Taylorentwicklung übereinstimmen, woraus die Behauptung folgt. □

Im folgenden charakterisieren wir die Stabilitätseigenschaften des Runge-Kutta Verfahrens über die Stabilitätsfunktion:

Satz 2.25. *Gegeben sei ein Runge-Kutta Verfahren mit Stabilitätsfunktion R, die auf $\mathbb{C}^-$ definiert ist. Dann gilt:*

1. *Das Verfahren ist genau dann A-stabil, wenn $|R(\zeta)| \leq 1$ für alle $\zeta \in \mathbb{C}^-$ gilt.*

2. *Das Verfahren ist genau dann Isometrie erhaltend, wenn $|R(\zeta)| = 1$ für alle ζ mit $\Re(\zeta) = 0$ gilt.*

Beweis. Der Beweis folgt unmittelbar aus (2.30). □

Aus diesem Satz lässt sich sofort schließen, dass alle expliziten Runge-Kutta Verfahren nicht A-stabil sein können, da für Polynome

$$\lim_{a \to -\infty} |R(a + \mathrm{i}b)| = \infty$$

gilt.

Das implizite Euler-Verfahren ist hingegen A-stabil. In diesem Fall ist $R(\zeta) = (1-\zeta)^{-1}$ und

$$|1-\zeta|^2 = (1-\Re(\zeta))^2 + (\Im(\zeta))^2 = 1 - 2\Re(\zeta) + |\zeta|^2 \geq 1 \text{ für } \Re(\zeta) \leq 0\,.$$

Für weitere Untersuchungen der Stabilität von Runge-Kutta Verfahren führen wir den Begriff des Stabilitätsgebietes eines Verfahrens ein.

Definition 2.26. Mit $\mathcal{S} := \{\zeta \in \mathbb{C} : |R(\zeta)| \leq 1\}$ wird das Stabilitätsgebiet eines Runge-Kutta Verfahrens bezeichnet.

Mit unserer Notationsübereinkunft schließen wir die Spetralwerte von A aus dem Stabilitätsbereich aus.

Für A-stabile Runge-Kutta Verfahren ist die abgeschlossene linke Halbebene $\mathbb{C}^-$ in $\mathcal{S}$ enthalten.

Satz 2.27. *Für jedes Runge-Kutta Verfahren ist* $0 \in \mathcal{S}$ *.*

Beweis. Da jedes Runge-Kutta Verfahren mindestens die Ordnung 1 hat, gilt nach Satz 2.24,

$$R(\zeta) = 1 + \zeta + O(|\zeta|^2)\,, \quad \zeta \to 0\,.$$

Demnach ist $R(0) = 1$, also $0 \in \mathcal{S}$. □

Für jedes Runge-Kutta Verfahren gibt es aber auch ein offenes Teilintervall $(0, \varepsilon) \notin \mathcal{S}$, denn

$$|R(\zeta)| > 1 + \frac{\zeta}{2} > 1 \text{ für } \zeta \in (0, \varepsilon)\,,$$

für ein hinreichend kleines $\varepsilon > 0$. Das bedeutet, dass 0 am Rand des Stabilitätsbereiches liegt!

Satz 2.28. *Das Stabilitätsgebiet eines expliziten Runge-Kutta Verfahrens ist immer beschränkt.*

Beweis. Die Stabilitätsfunktion eines expliziten Runge-Kutta Verfahrens ist ein Polynom. Also existiert eine Kugel, die die Menge $\{\zeta : |R(\zeta)| \leq 1\}$ enthält. □

Satz 2.29. *Das Stabilitätsgebiet $\mathcal{S}$ des Runge-Kutta Verfahrens enthalte den Halbkreis*

$$\mathcal{B}_\tau^- := \left\{\zeta \in \mathbb{C}^- : |\zeta| \leq \tau\right\} .$$

Dann gilt für die Näherungen y_n des Runge-Kutta Verfahrens, angewendet auf die Testgleichung (2.28) *mit $\lambda \in \mathbb{C}^-$:*

$$|y_{n+1}| \leq |y_n| , \text{ sobald } h \leq \tau/|\lambda| .$$

Beweis. Für $\Re(\lambda) \leq 0$ und $h \leq \frac{\tau}{|\lambda|}$ liegt $\zeta := h\lambda$ in $\mathcal{B}_\tau^-$. Aus der Rekursion $y_{n+1} = R(h\lambda)y_n$ und da $h\lambda$ im Stabilitätsbereich liegt, folgt aus Satz 2.25

$$|y_{n+1}| = |R(h\lambda)|\,|y_n| \leq 1\,|y_n|$$

und somit die Aussage des Satzes. □

Das Stabilitätsgebiet des expliziten Euler Verfahrens ist

$$S = \{\zeta : |1 + \zeta| \leq 1\} . \tag{2.32}$$

Das ist ein Kreis mit Mittelpunkt -1 und Radius 1.

2.2.5 Steife Differentialgleichungen

Wir betrachten die Wärmeleitungsgleichung

$$u_t(t,x) = u_{xx}(t,x) , \quad u(0,x) = g(x) , \tag{2.33}$$

mit Randbedingungen $u(t,-1) = u(t,1) = 0$. Beachte, dass (2.33) keine gewöhnliche Differentialgleichung ist, sondern eine partielle.

Zur analytischen Lösung der Differentialgleichung verwenden wir *Separation der Variablen*, dazu machen wir den Ansatz

$$u(x,t) = \sum_{n\in\mathbb{N}} \eta_n(t)v_n(x) .$$

Erfüllt diese Funktion (2.33), dann muß

$$\sum_{n\in\mathbb{N}} \eta_n'(t)v_n(x) = \sum_{n\in\mathbb{N}} \eta_n(t)v_n''(x)$$

gelten, was, wiederum, gilt, wenn

$$\eta_n'(t)v_n(x) = \eta_n(t)v_n''(x) \text{ für alle } t > 0, x \in (-1,1)\,.$$

Da η_n eine Funktion in t ist und v_n in x folgt somit:

$$\frac{\eta_n'(t)}{\eta_n} = C = \frac{v_n''(x)}{v_n(x)} \text{ für alle } t > 0, x \in (-1,1)\,.$$

Also müssen wir die beiden Eigenwertprobleme $v_n'' = Cv_n(x)$ und $\eta' = C\eta$ parallel lösen. Die Eigenwerte des Operators

$$Av = v_{xx} = \lambda v\,, \quad v(-1) = v(1) = 0\,, \tag{2.34}$$

sind $\lambda_n = -\left(\frac{\pi}{2}n\right)^2, n \in \mathbb{N}$ mit den zugehörigen Eigenfunktionen

$$v_n(x) = \begin{cases} \cos\left(\frac{\pi}{2}nx\right)\,, & n \text{ ungerade} \\ \sin\left(\frac{\pi}{2}nx\right)\,, & n \text{ gerade}\,. \end{cases}$$

Für ungerade n erfüllen die cos-Funktionen und für gerade n erfüllen die sin-Funktionen die Randbedingungen von (2.34). Also ist die Lösung von (2.34) wie folgt darstellbar

$$u(t,x) = \sum_{n=1}^{\infty} \eta_{2n-1}(t)\cos\left(\frac{\pi}{2}(2n-1)x\right) + \sum_{n=1}^{\infty} \eta_{2n}(t)\sin\left(\pi n x\right)\,,$$

wobei die Koeffizientenfunktionen η_n Lösungen der gewöhnlichen Differentialgleichung

$$\eta_n' = -\left(\frac{\pi}{2}n\right)^2 \eta_n\,, \quad n \in \mathbb{N}\,, \tag{2.35}$$

sind. Die Lösung der partiellen Differentialgleichung wird somit auf die Lösung eines System von gewöhnlichen Differentialgleichungen zurückgeführt.

Zur Lösung dieser gewöhnlichen Differentialgleichungen benötigten wir noch Anfangswerte. Wir nehmen konkret an, dass

$$u(0,x) = g(x) = 1 - x^2\,, -1 \le x \le 1\,,$$

und entwickeln g in eine Fourierreihe

$$g(x) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{32}{\pi^3(2n-1)^3}\cos\left(\frac{\pi}{2}(2n-1)x\right)\,.$$

Demnach ergeben sich für die Koeffizientenfunktionen η_n die Anfangswerte

$$\eta_{2n}(0) = 0\,, \quad \eta_{2n-1}(0) = (-1)^{n+1}\frac{32}{\pi^3(2n-1)^3}\,, \quad n > 0\,,$$

und durch Lösen der skalaren Differentialgleichung (2.35) erhält man eine analytische Darstellung der Lösung von (2.33)

$$u(t,x) = \sum_{n=1}^{\infty} \frac{32(-1)^{n+1}}{\pi^3(2n-1)^3} \exp\left(-\pi^2\frac{(2n-1)^2}{4}t\right)\cos\left(\frac{\pi}{2}(2n-1)x\right)\,. \quad (2.36)$$

Nehmen wir an, wir haben in der n-ten Komponente einen Fehler ε_n. Nach i weiteren Zeitschritten mit einem Runge-Kutta Verfahren hat sich der Fehler in der n-ten Komponente zu

$$R(h\lambda_n)^i\varepsilon_n = R(-\pi^2 h(n/2)^2)^i\varepsilon_n$$

verändert.

Für explizite Verfahren liegt für hinreichend großes n $-\pi^2 h(n-1/2)^2$ nicht im Stabilitätsbereich $\mathcal{S}$ und der fortgepflanzte Fehler explodiert. Hingegen spielen bei der exakten Lösung (2.36) die hochfrequenten Anteile praktisch keine Rolle, da sie mit die Vorfaktoren $\exp(-\pi^2(n-1/2)^2 t)$ gedämpft werden, oder sogar Null sind.

Solche Aufgabenstellungen können nur mit A-stabilen Verfahren gelöst werden. Eine Differentialgleichung mit (stark) negativen Eigenwerten in der Stabilitätsanalyse nennt man *steife* Differentialgleichung.

2.2.6 Implizite Runge-Kutta Verfahren

Das Ziel ist die Konstruktion von impliziten Runge-Kutta Verfahren mit großem Stabilitätsbereich und hoher Ordnung. Für solch ein Verfahren verwenden wir die s-Gauß-Legendre Quadraturformel,

$$Q[g] = \sum_{j=1}^{s} b_j g(c_j)\,, \quad (2.37)$$

welche den maximal möglichen Exaktheitsgrad $2s-1$ für das Integral $\int_0^1 g(t)\,dt$ hat.

Für eine implizites Runge-Kutta Verfahren wählt man nun für $\{b_j\}$ die zugehörigen Gewichte der s-stufigen Gauß-Quadraturformel [2] und für $\{c_j\}$ die Nullstellen des s-ten Legendre-Polynoms (skaliert auf das Intervall $[0,1]$). Die verbleibenden Koeffizienten $\{a_{jk}\}$ werden wie in (2.23) gewählt, dass die Approximation

$$h\sum_{k=1}^{s} a_{jk} f(t_i+c_k h, \eta_k) \approx \int_{t_i}^{t_i+c_j h} f(t, y(t))\, dt = h\int_0^{c_j} f(t_i+\tau h, y(t_i+\tau h))\, d\tau\ .$$

Polynome mit möglichst hohem Grad exakt integriert, also

$$\sum_{k=1}^{s} a_{jk} p(c_k) \approx \int_0^{c_j} p(t)\, dt\ . \tag{2.38}$$

Da die Koeffizienten $\{c_k\}$ bereits vorgegeben sind, ist die optimale Wahl der Gewichte $\{a_{jk}\}$ durch Lagrange-Grundpolynome gegeben, also

$$a_{jk} = \int_0^{c_j} l_k(t)\, dt\ , \quad l_k(t) = \prod_{i=1, i\neq k}^{s} \frac{t-c_i}{c_k-c_i}\,, \tag{2.39}$$

und die resultierende Quadraturformel hat Exaktheitsgrad $s-1$.

Das aus (2.37) und (2.39) resultierende implizierte Runge-Kutta Verfahren heißt auch s-stufiges **(Runge-Kutta-)Gauß-Verfahren**.

Beispiel 2.30. Das einfachste Gauß-Verfahren ($s=1$) führt auf die so genannte **implizite Mittelpunktsregel**. Das erste Legrende Polynom hat eine Nullstelle genau in der Mitte des Intervalls $c_1 = 1/2$ und das zugehörige Gewicht ist gerade $b_1 = 1$. a_{11} ergibt sich aus (2.39) zu $a_{11} = \int_0^{1/2} 1dt\ = 1/2$. Das zugehörige Tableau ist

1/2	1/2
	1

und das Einschrittverfahren hat die Form

$$y_{i+1} = y_i + hf(t_i + h/2, \eta_1)\,, \quad \eta_1 = y_i + \frac{h}{2} f(t_i + h/2, \eta_1)\ . \tag{2.40}$$

[2] Das Resultat über die Gauß-Quadraturformel über Numerische Mathematik, wie etwa in Hämmerlin-Hofmann [5]

Durch Kombination der beiden Gleichungen erhält man

$$y_{i+1} = y_i + hf(t_i + h/2, (y_i + y_{i+1})/2) \ .$$

Die Stabilitätsfunktion der impliziten Mittelpunktsregel ist leicht ausgerechnet: Nach Definition 2.20 gilt:

$$R(\zeta) = 1 + \zeta \left(1 - \frac{\zeta}{2}\right)^{-1} = \frac{1 + \zeta/2}{1 - \zeta/2} = 1 + \zeta + \zeta^2/2 + \zeta^3/4 + ...$$

Die Stabilitätsfunktion R ist eine Möbiusfunktion.

- Es gilt für $\zeta = a + \mathrm{i}b \in \mathbb{C}^-$ (also $a < 0$)

$$|R(\zeta)|^2 = \frac{(1 + a/2)^2 + b^2}{(1 - a/2)^2 + b^2} \leq 1 \text{ da } a < 0 \ .$$

- Ist $\lambda = \mathrm{i}b$ mit $b \in \mathbb{R}$, so gilt:

$$|R(\zeta)|^2 = \frac{1 + b^2}{1 + b^2} = 1 \ .$$

Also ist die implizite Mittelpunktsregel nach Satz 2.25 A-stabil und Isometrie erhaltend.

Man kann die implizite Mittelpunktsregel auch als eine Kombination des impliziten und des expliziten Euler-Verfahrens interpretieren. Demnach wird in (2.40) zunächst in einem impliziten Zeitschritt eine erste Näherung η_1 für $y(t_i + h/2)$ berechnet; diese Näherung dient dann einem zweiten Halbschritt zur Berechnung von y_{i+1} mit dem expliziten Euler-Verfahren als Ausgangspunkt.

Die Stabilität der impliziten Gauß-Runge-Kutta Verfahren ist eine allgemeine Eigenschaft:

Satz 2.31. *Alle impliziten Gauß-Runge-Kutta Verfahren sind A-stabil und Isometrie erhaltend.*

Einen Beweis findet man z.B. in [5].

Wir fassen nun einige Gauß-Runge-Kutta Tableaus zusammen:

1.

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

2.

$$\begin{array}{c|ccc} \frac{1}{2}-\frac{\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9}-\frac{\sqrt{15}}{15} & \frac{5}{36}-\frac{\sqrt{15}}{30} \\ \frac{1}{2} & \frac{5}{36}+\frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36}-\frac{\sqrt{15}}{24} \\ \frac{1}{2}+\frac{\sqrt{15}}{10} & \frac{5}{36}+\frac{\sqrt{15}}{30} & \frac{2}{9}+\frac{\sqrt{15}}{15} & \frac{5}{36} \\ \hline & \frac{5}{18} & \frac{4}{9} & \frac{5}{18} \end{array}$$

Wir skizzieren nun die effiziente Implementierung von impliziten Gauß-Runge-Kutta Verfahren. Der Hauptaufwand steckt dabei in der Lösung des nichtlinearen Gleichungssystems

$$\eta_j = y_i + h \sum_{\nu=1}^{s} a_{j\nu} f(t_i + c_\nu h, \eta_\nu) \,, \qquad j = 1, ..., s \,.$$

Hat man Näherungen η_ν berechnet, dann kann man den eigentlichen Iterationsschritt durchführen:

$$y_{i+1} = y_i + h \sum_{\nu=1}^{s} b_\nu f(t_i + c_\nu h, \eta_\nu) \,. \tag{2.41}$$

In der Praxis führt man Hilfsvariablen ein:

$$k_j = f(t_i + c_j h, \eta_j) \,,$$

und erhält auf diese Weise das äquivalente Gleichungssystem

$$k_j = f\left(t_i + c_j h, y_i + h \sum_{\nu=1}^{s} a_{j\nu} k_\nu \right) \,, \quad j = 1, ..., s \,, \tag{2.42}$$

mit denen man dann den eigentlichen Iterationsschritt wie folgt implementiert:

$$y_{i+1} = y_i + h \sum_{\nu=1}^{s} b_\nu k_\nu \,.$$

Die Lösung des $sd \times sd$-dimensionalen Gleichungssystems (2.42) für die k_j kann mit einem Newton-artigen Verfahren erfolgen. Man verwendet häufig Quasi-Newton-Verfahren, bei denen anstelle der exakten Ableitungen

$$f_y\left(t_i + c_j h, y_i + h\sum_{\nu=1}^{s} a_{j\nu} k_\nu\right)$$

von f die Näherung $J = f_y(t_i, y_i)$ verwendet wird.

2.3 Randwertprobleme

bei gewöhnlichen Differentialgleichungen Wir betrachten zuerst das einfache Modellproblem:

$$\begin{aligned} L[u] = -u'' + bu' + cu = f \text{ in } (0,1)\,, \\ u(0) = u(1) = 0\,. \end{aligned} \tag{2.43}$$

Es lässt sich zeigen, dass diese Differentialgleichung eine eindeutige Lösung besitzt, falls

$$c(x) \geq 0 \text{ für alle } x \in (0,1)\,,$$

was wir im folgenden immer voraussetzen.

Zur numerischen Lösung verwendet man sehr häufig *Differenzenverfahren*, die wir im folgenden untersuchen. Alternativ werden wir später *finite Element Methoden* kennenlernen. Wir schränken uns auf ein äquidistantes Gitter ein

$$\Delta_h = \{x_i = ih : i = 1, \ldots, n-1, h = 1/n\} \subseteq (0,1)\,. \tag{2.44}$$

Mit

$$\vec{u} = (u(x_1), \ldots, u(x_n))^t \in \mathbb{R}^{n-1} \tag{2.45}$$

bezeichnen wir den Vektor der exakten Lösung u von (2.43) auf dem Gitter Δ_h (2.44). Wir denken uns zusätzlich

$$0 = u(x_0) = u(x_n) = 0\,,$$

was die Randbedingungen wiedergibt. Den erweiterten Vektor werden wir aber nie verwenden. Gesucht ist nun eine Näherungsvektor

$$\vec{u}_h = (u_1, \ldots, u_n)^t \in \mathbb{R}^{n-1} \tag{2.46}$$

für $\vec{u}$. Zu diesem Zweck wird L aus (2.43) diskretisiert, indem wir die Ableitungen von u an den Stellen $x = x_i$ durch Differenzenquotienten ersetzen. Wir haben folgende Alternativen:

- **Einseitiger Vorwärtsdifferenzenquotient:**

$$D_h^+[u](x) = \frac{u(x+h) - u(x)}{h} \sim u'(x) \; .$$

- **Einseitiger Rückwärtsdifferenzenquotient:**

$$D_h^-[u](x) = \frac{u(x) - u(x-h)}{h} \sim u'(x) \; .$$

- **Zentraler Differenzenquotient:**

$$D_h[u](x) = \frac{u(x+h) - u(x-h)}{2h} \sim u'(x) \; . \tag{2.47}$$

Die zweite Ableitung kann durch den zentralen Differenzenquotienten

$$D_h^2[u](x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} \sim u''(x) \tag{2.48}$$

approximiert werden.

Beispiel 2.32. Wir betrachten den einfachsten Fall von (2.43) mit $b, c \equiv 0$, also $-u'' = f$ mit Dirichlet Randbedingungen. Wir approximieren u'' durch $D_h^2[u]$ an den Knotenpunkten Δ_h. Unter Berücksichtigung von $u(x_0) = u(x_n) = 0$ ergibt sich somit.

$$\underbrace{\begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{n-1}) \end{bmatrix}}_{=:\vec{f}} = -\begin{bmatrix} u''(x_1) \\ u''(x_2) \\ \vdots \\ u''(x_{n-1}) \end{bmatrix} \sim \underbrace{h^{-2}\begin{bmatrix} 2 & -1 & & 0 \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix}}_{=:L_h} \underbrace{\begin{bmatrix} u(x_1) \\ u(x_2) \\ \vdots \\ u(x_{n-1}) \end{bmatrix}}_{\vec{u}} \; .$$

Da $\vec{u}$ durch $\vec{u}_h$ approximiert werden soll, liegt es nahe folgendes Gleichungssytem zur Bestimmung von $\vec{u}_h$ zu verwenden:

$$L_h \vec{u}_h = \vec{f} \; . \tag{2.49}$$

Die Eigenwerte von L_h sind $4h^{-2}\sin^2(kh\pi/2)4$, $k = 1, \ldots, n-1$. Die Funktion $\mathrm{sinc}(x) := \frac{\sin(x)}{x}$ ist im Intervall $[0, \pi/2]$ monoton fallend, weshalb gilt

$$\mathrm{sinc}(x) \geq \mathrm{sinc}\left(\frac{\pi}{2}\right) = \frac{2}{\pi} \text{ für } x \in [0, \pi/2] \, ,$$

woraus sich ergibt:

$$\|L_h^{-1}\|_2 = \frac{1}{\lambda_{\min}(L_h)} = \max_{1\le k\le n-1} \frac{h^2}{4\sin^2(kh\pi/2)} \le \frac{1}{4}\,.$$

Folglich gilt

$$\begin{aligned} \|\vec{u}-\vec{u}_h\|_2 &= \|L_h^{-1}(L_h\vec{u}-\vec{f})\|_2 \\ &\le \|L_h^{-1}\|_2 \|L_h\vec{u}-\vec{f}\|_2 \\ &\le \frac{1}{4}\|L_h\vec{u}-\vec{f}\|_2\,. \end{aligned} \tag{2.50}$$

Der letzte Fehlerterm bezeichnet die *Konsistenz* des Operators L_h. Gilt eine Abschätzung dieser Form, so wird *Stabilität* durch *Konsistenz* impliziert.

Wir beschäftigen uns nun mit Fehlerabschätzungen für die Differenzenquotienten:

Lemma 2.33. *Sei $u \in C^2[0,1]$ und $x \in [h, 1-h]$. Dann gelten für die einseitigen Differenzenquotienten die Abschätzungen*

$$\left|D_h^{\pm}[u](x) - u'(x)\right| \le \frac{1}{2}\|u''\|_\infty h\,.$$

Für den zentralen Differenzenquotienten gilt für $u \in C^3[0,1]$ sogar:

$$|D_h[u](x) - u'(x)| \le \frac{1}{6}\|u'''\|_\infty h^2\,.$$

Für D_h^2 gilt: Sei $u \in C^4[0,1]$ und $x \in [h, 1-h]$, dann ist:

$$\left|D_h^2[u](x) - u''(x)\right| \le \frac{1}{12}\|u''''\|_\infty h^2\,. \tag{2.51}$$

Beweis. Wir beweisen die Behauptung für den zentralen Differenzenquotienten. Für $u \in C^3[0,1]$ folgt aus Taylorentwicklung um $x \in (0,1)$:

$$\begin{aligned} u(x+h) &= u(x) + hu'(x) + \frac{1}{2}h^2u''(x) + \frac{1}{6}h^3u'''(\zeta_+)\,, \\ u(x-h) &= u(x) - hu'(x) + \frac{1}{2}h^2u''(x) - \frac{1}{6}h^3u'''(\zeta_-)\,, \end{aligned}$$

für zwei $\zeta_\pm$ mit der Eigenschaft $x-h < \zeta_- < x < \zeta_+ < x+h$. Daraus folgt

$$u(x+h) - u(x-h) = 2hu'(x) + \frac{1}{6}h^3(u'''(\zeta_+) + u'''(\zeta_-))\,,$$

und somit

$$\left|\frac{u(x+h)-u(x-h)}{2h}-u'(x)\right| \leq \frac{1}{6}h^2 \sup\left\{|u'''(\zeta)| : \zeta \in [0,1]\right\} ,$$

und somit die Behauptung. □

Beispiel 2.34. Angewendet auf Beispiel 2.32 ergibt sich, falls die Lösung der Differentialgleichung 4× stetig differenzierbar ist:

$$\|\underbrace{L_h}_{=D_h^2}\vec{u}-\vec{f}\|_\infty \leq \frac{1}{12}\|u''''\|_\infty h^2 = \frac{1}{12}\|f''\|_\infty h^2 .$$

Wir diskretisieren nun den allgemeinen Differentialoperator L definiert in (2.43). Dazu ersten wir die zweite Ableitung durch $D_h^2[u]$ und die erste Ableitung durch einen der drei Differenzenquotienten $D_h^+[u]$, $D_h^-[u]$, oder $D_h[u]$. Bei Verwendung der verschiedenen Differenzenquotienten ergeben sich unterschiedliche Diagonalmatrizen

$$L_h = h^{-2}\begin{bmatrix} d_1 & s_1 & & 0 \\ r_2 & d_2 & \ddots & \\ & \ddots & \ddots & s_{n-2} \\ 0 & & r_{n-1} & d_{n-1} \end{bmatrix} \in \mathbb{R}^{(n-1)\times(n-1)} , \tag{2.52}$$

wobei im Fall

- D_h^+:

$$\begin{aligned} d_i &= 2 - hb(x_i) + h^2 c(x_i) , \\ r_i &= -1 , \\ s_i &= -1 + hb(x_i) , \end{aligned} \tag{2.53}$$

- D_h^-:

$$\begin{aligned} d_i &= 2 + hb(x_i) + h^2 c(x_i) , \\ r_i &= -1 - hb(x_i) , \\ s_i &= -1 , \end{aligned} \tag{2.54}$$

- D_h:

$$\begin{aligned} d_i &= 2 + h^2 c(x_i) , \\ r_i &= -1 - hb(x_i)/2 , \\ s_i &= -1 + hb(x_i)/2 . \end{aligned} \tag{2.55}$$

Die Näherungslösung ergibt sich aus der Lösung des Gleichungssystems (2.49).

Definition 2.35. Ein Differenzenverfahren hat die *Konsistenzordnung* q bezüglich der Maximumnorm, wenn

$$\|L_h\vec{u} - \vec{f}\|_\infty \leq Ch^q \,.$$

Die Definition ist nur insofern präzise, da diese Abschätzung nur unter Glattheitsvoraussetzungen an u, und die Koeffizienten der Differentialgleichung b, c, f gelten kann. Diese müssten als extra Voraussetzungen spezifiziert werden.

Satz 2.36. *Die Lösung des Randwertproblems* (2.43) *sei* $4\times$ *stetig differenzierbar (das gilt z.B. unter den Voraussetzungen, dass* b, c, f $2\times$ *stetig differenzierbar sind). Dann hat das Differenzenverfahren* (2.49) *die folgenden Konsistenzordnungen* q*:*

- $q = 2$*, falls* D_h *(zentraler Differenzenquotient) zur Approximation für* u' *verwendet wird, bzw.*
- $q = 1$*, falls* $D_h^\pm$ *(Vorwärts- bzw. Rückwärts Differenzenquotient) zur Approximation für* u' *verwendet wird.*

Beweis. Wir verwenden die Zerlegung

$$L_h = \underbrace{h^{-2}\begin{bmatrix} 2 & -1 & & 0 \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix}}_{=:A_h} + B_h + \underbrace{\begin{bmatrix} c(x_1) & 0 & & 0 \\ 0 & c(x_2) & \ddots & \\ & \ddots & \ddots & 0 \\ 0 & & 0 & c(x_{n-1}) \end{bmatrix}}_{=:C_h} .$$

Nur die Matrix B_h hängt von den verwendeten Differenzenverfahren ab.

Aus Lemma 2.33 und (2.51) folgt, dass im Fall der Verwendung von zentralen Differenzenquotienten,

$$\begin{aligned} |(A_h\vec{u})_i - (-u'')| &\leq \frac{1}{12}\|u''''\|_\infty h^2 \,, \\ |(B_h\vec{u})_i - b(x_i)u'(x_i)| &\leq \frac{1}{6}\|b\|_\infty\|u'''\|_\infty h^2 \,, \\ |(C_h\vec{u})_i - c(x_i)u(x_i)| &= 0 \,, \end{aligned}$$

gilt. Für einseitige Differenzenquotienten gilt dementsprechend

$$|(B_h\vec{u})_i - b(x_i)u'(x_i)| \leq \frac{1}{2}\|b\|_\infty\|u''\|_\infty h \ .$$

Damit gilt bei der Verwendung der zentralen Differenzenquotienten:

$$\begin{aligned}\|L_h\vec{u} - \vec{f}\|_\infty &= \|A_h\vec{u} + B_h\vec{u} + C_h\vec{u} - \vec{f}\|_\infty \\ &= \max_{1\leq i\leq n-1} |(A_h\vec{u} + B_h\vec{u} + C_h\vec{u})_i - (-u''(x_i) + b(x_i)u'(x_i) + c(x_i)u(x_i)| \\ &\leq \left(\frac{1}{12}\|u''''\|_\infty + \frac{1}{6}\|b\|_\infty\|u'''\|_\infty\right) h^2 \,,\end{aligned} \tag{2.56}$$

beziehungsweise bei der Verwendung eines einseitigen Differenzenquotienten

$$\|L_h\vec{u} - \vec{f}\|_\infty \leq \left(\frac{1}{12}\|u''''\|_\infty h + \frac{1}{2}\|b\|_\infty\|u''\|_\infty\right) h \ .$$

□

2.4 Stabilitätsabschätzungen

Wir leiten Abschätzungen für $\|L_h^{-1}\|_\infty$ her, wobei L_h eine der Matrizen aus (2.52) sind. Mit Satz 2.36 folgt somit, Folglich gilt, wenn $\|L_h^{-1}\|_\infty < C$ (für hinreichend kleine h) ist,

$$\begin{aligned}\|\vec{u} - \vec{u}_h\|_\infty &= \|L_h^{-1}(L_h\vec{u} - \vec{f})\|_\infty \\ &\leq \|L_h^{-1}\|_\infty\|L_h\vec{u} - \vec{f}\|_\infty \\ &\leq C\|L_h\vec{u} - \vec{f}\|_\infty \\ &\leq C\left(\frac{1}{12}\|u''''\|_\infty + \frac{1}{6}\,|b(x_i)|\,\|u'''\|_\infty\right) h^2\end{aligned} \tag{2.57}$$

bei Verwendung von zentralen Differenzenverfahren. Bei Verwendung von einseitigen Differenzenverfahren bekommt man eine Ordnung h Abschätzung.

Dieses Beispiel legt folgende allgemeine Definition nahe:

Definition 2.37. Ein Differenzenverfahren für das Randwertproblems (2.43) heisst *stabil*, wenn Konstanten $C, h_0^* > 0$ existieren, sodass für alle $0 < h < h_0^*$ die Matrix L_h invertierbar und $\|L_h^{-1}\|_\infty \leq C$ gilt.

Zum Nachweis einer stabilen Diskretisierung verwendet man häufig M-Matrizen, die wie folgt definiert sind:

Definition 2.38. Sei $A = [a_{ij}] \in \mathbb{R}^{n\times n}$ mit $a_{ij} \leq 0$ für $i \neq j$, und $A^{-1} \geq 0$, dann nennt man A eine M-Matrix.

Bei M-Matrizen gilt die Monotonieeigenschaft (die auch den Namen M-Matrix impliziert)

$$x \leq y \text{ impliziert } A^{-1}x \leq A^{-1}y \,. \tag{2.58}$$

Die Verifikation der M-Matrix Eigenschaft ist schwierig, ausser wenn sie spezielle Struktur hat, etwa tridiagonale Form.

Satz 2.39. *Jede irreduzibel diagonaldominate Tridiagonalmatrix $T = [t_{ij}]$ mit positiven Diagonalelementen und negativen Nebendiagonalelementen ist eine M-Matrix.*

Beweis. Irreduzible Tridiagonalmatrix bedeutet, dass $t_{i,i-1} \neq 0$ und $t_{i-1,i} \neq 0$ gilt. Diagonaldominant bedeutet, dass

$$|a_{ii}| \geq \sum_{j\neq i} |a_{ij}|$$

gilt. Wir wissen aus der Grundnumerik, dass solche Matrizen invertierbar sind. Es bleibt zu zeigen, dass T^{-1} nichtnegativ ist. Dazu schreiben wir

$$T = D - N \,,$$

wobei D eine Diagonalmatrix und N eine Nebendiagonalmatrix ist. Wegen der Voraussetzungen an T gilt, $D \geq 0$ und $N \geq 0$.

Für beliebiges $\varepsilon > 0$ ist nun auch $T + \varepsilon I$ strikt diagonaldominant und es gilt

$$T + \varepsilon I = (D + \varepsilon I)(I - R) \text{ mit } R = (D + \varepsilon I)^{-1}N \,.$$

Da $D \geq 0$ ist auch $D + \varepsilon I \geq 0$, und damit auch $(D + \varepsilon I)^{-1} \geq 0$. Konsequenterweise ist dann auch $R \geq 0$. Zudem gilt wegen der Diagonaldominanz von T auch (die echte Ungleichung kommt durch die Störung mit εI)

$$\|R\|_\infty = \|(D + \varepsilon I)^{-1}N\|_\infty < 1 \,. \tag{2.59}$$

Damit konvergiert die Reihe

$$\begin{aligned}(T+\varepsilon I)^{-1} &= (I-R)^{-1}(D+\varepsilon I)^{-1} \\ &= \left(\sum_{k=0}^{\infty} R^k\right)(D+\varepsilon I)^{-1} \\ &= \sum_{k=0}^{\infty} R^k (D+\varepsilon I)^{-1} \\ &\geq 0\ .\end{aligned}$$

Durch Grenzübergang $\varepsilon \to 0$ erhält man somit auch $(T+\varepsilon I)^{-1} \to T^{-1}$, die somit auch nichtnegativ ist. □

Korollar 2.40. *Sei*

$$0 < h < h_0 = \frac{2}{\|b\|_\infty}\ . \tag{2.60}$$

Darüberhinaus setzen wir voraus, dass die Funktion c, die bei der Bestimmung von L_h aus (2.52) *verwendet wird, nichtnegativ ist.*

Dann gilt bei Verwendung des zentralen Differenzenquotienten $D_h[u](x_i)$ (2.47) *für $u'(x_i)$, $i = 1, 2, \ldots, n-1$, dass die Koeffizientenmatrix L_h aus* (2.52) *ein M-Matrix ist.*

Beweis. Aus (2.60) folgt, dass $|hb(x_i)/2| < 1$ ist. Damit sind die Nebendiagonaleinträge, r_i, s_i von L_h aus (2.52) negativ. Weiters gilt $|r_i| + |s_i| \leq 2 \leq |d_i|$; Für $i = 2, 3, \ldots, n-2$ gilt die erste Ungleichung sogar als Gleichheit. Somit ist die Matrix L_h diagonaldominant, und alle Eigenschaften betreffend der Vorzeichen der koeffizienten r_i, s_i und d_i zusammen, zeigen, dass die Matrix L_h eine M-Matrix ist. □

Satz 2.41. *Die Randwertaufgabe 2.43 habe ein Lösung in $C^4[0,1]$. Dann genügt die Lösung $\vec{u}_h$ von* (2.49) *(mit zentralem Differenzenquotienten) der Fehlerabschätzung*

$$\|\vec{u} - \vec{u}_h\|_\infty = O(h^2)\ .$$

Beweis. Wegen Korollar 2.40 ist für hinreichend kleine h L_h eine M-Matrix und damit insbesondere invertierbar. Also gilt für hinreichend kleine h (vgl.

(2.57)
$$\begin{aligned}\|\vec{u}-\vec{u}_h\|_\infty &= \|L_h^{-1}(L_h\vec{u}-L_h\vec{u}_h)\|_\infty \\ &= \|L_h^{-1}(L_h\vec{u}-\vec{f})\|_\infty \\ &\le \|L_h^{-1}\|_\infty\|L_h\vec{u}-\vec{f}\|_\infty \\ &=: C_1\|L_h\vec{u}-\vec{f}\|_\infty\,,\end{aligned}$$
wobei die Konstante C_1 unabhängig von h. Aus (2.56) folgt somit
$$\|\vec{u}-\vec{u}_h\|_\infty = O(h^2)\,.$$

□

2.5 Singulär gestörte Probleme

Die Schranke $h_0 = \frac{2}{\|b\|_\infty}$ aus Korollar 2.40 zeigt, dass das Differenzenverfahren für große Schrittweiten instabil wird.

Wir studieren für kleine ε die Lösung der Randwertaufgabe
$$u'' - \frac{u'}{\varepsilon} = -\frac{1}{\varepsilon} \text{ in } (0,1)\,, \quad u(0)=u(1)=0\,.$$
Die analytisch Lösung ist gegeben durch
$$u_\varepsilon(x) = x - v_\varepsilon(x) \text{ mit } v_\varepsilon(x) = \frac{e^{(x-1)/\varepsilon}-e^{-1/\varepsilon}}{1-\varepsilon^{-1/\varepsilon}}\,.$$
Die Funktion ist außerhalb des Grenzschichtintervalls $(1-\gamma\varepsilon,1)$ ungefähr die Funktion $u(x)=x$, und fällt in der Grenzschicht steil auf 0 ab. Die ganze Problematik zeigt sich, wenn man $\varepsilon \to 0+$ betrachtet, was dazu führt, dass die Differentialgleichung formal zu einer Differentialgleichung 1. Ordnung
$$u' = 1 \text{ in } (0,1)\,, \quad u(0)=u(1)=0\,.$$
Diese Gleichung ist aber durch die beiden Randbedingungen überbestimmt.

Das Differenzenverfahren mit zentralen Differenzenquotienten ergibt das lineare Gleichungssystem
$$L_h\vec{u}_h = \vec{f}\,, \tag{2.61}$$
wobei h^2L_h/ε eine tridiagonale Töplitzmatrix mit den Einträgen
$$d_i = 2\,, \quad r_i = -1-h/(2\varepsilon)\,, \quad s_i = -1+h/(2\varepsilon)\,,$$

ist.

Wendet man Korollar 2.40 an, so erkennt man, dass L_h nur für $h < 2\varepsilon$ eine M-Matrix ist, und nur dann Stabilität mit unseren Mitteln gezeigt werden kann. Bei der Verwendung einseitigen Differenzenquotienten verhält es sich etwas anders: Die Matrix L_h aus (2.61) kann zu einer M-matrix gemacht werden, wenn abhängig von $b(x_i) < 0$ bzw. $b(x_i) > 0$, D_h^+ (2.53) bzw. D_h^- (2.54) gewählt wird. Dies nennt man das *Upwind-Schema* zur Lösung der Differentialgleichung (2.43) und führt aud die Matrixgleichung 2.49 mit den Matrixeinträgen für L_h aus (2.52):

$$\begin{aligned} d_i &= 2 + h\,|b(x_i)| + h^2 c(x_i)\,, \\ r_i &= -1 - hb^+(x_i)\,, \quad b^+(x) = \max\{b(x), 0\}\,, \\ s_i &= -1 + hb^-(x_i)\,, \quad b^-(x) = \min\{b(x), 0\}\,. \end{aligned} \tag{2.62}$$

2.6 Schießverfahren

Wir studieren das nichtlineare Randwertproblem

$$u'' = f(x, u, u') \text{ in } (0,1)\,, \qquad u(0) = u(1) = 0\,. \tag{2.63}$$

Nehmen wir an, dass die Ableitung $\alpha = u'(0)$ am linken Rand bekannt wäre. Dann löst u auch das Anfangswertproblem

$$v'' = f(x, v, v') \text{ in } (0,1)\,, \qquad v(0) = 0 \text{ und } v'(0) = \alpha\,. \tag{2.64}$$

Beim Schießverfahren wählt man ein α, löst (2.64) und vergleicht die Lösung v_α an der Stelle 1 mit $u(1)$. Aus der Differenz bestimmt man sich ein optimiertes α.

Um das zu systematisieren formulieren wir die Abbildung

$$\begin{aligned} F : D(F) &\to \mathbb{R}\,, \\ \alpha &\to v_\alpha(1) \end{aligned} \tag{2.65}$$

wobei $D(F)$ der Bereich der α ist, wo (2.64) eine Lsöung in $[0,1]$ besitzt. Wir suchen nun eine Nullstelle von F und kann mit Newton- oder Sekantenverfahren gelöst werden. Dazu brauchen wir die Ableitung von F bzgl. α.

Proposition 2.42. *Sei f zweimal stetig differenzierbar bzgl. der Variablen u und u'. Dann ist F differenzierbar mit Ableitung $F'(\alpha) = w_\alpha(1)$, wobei w_α die Lösung von*

$$\begin{aligned} w'' &= f_u(x, v_\alpha, v'_\alpha)w + f_{u'}(x, v_\alpha, v'_\alpha)w' \,, \\ w(0) = 0 \,, &\quad w'(0) = 1 \,, \end{aligned} \tag{2.66}$$

ist

Beweis. Wir definieren

$$w_\beta(x) = \frac{v_\beta(x) - v_\alpha(x)}{\beta - \alpha} \,,$$

wenn v_α, v_β Lösungen von (2.64) mit $v'_\alpha(0) = \alpha$ und $v'_\beta(0) = \beta$ sind. Damit gilt insbesondere

$$\frac{F(\beta) - F(\alpha)}{\beta - \alpha} = \frac{v_\beta(1) - v_\alpha(1)}{\beta - \alpha} = w_\beta(1) \,,$$

und somit

$$F'(\alpha) = \lim_{\beta \to \alpha} w_\beta(1) \,, \tag{2.67}$$

falls der Grenzwert exisitiert.

Wir fassen diverse Eigenschaften von w_β zusammen.

1. $w_\beta(0) = 0$ and $w'_\beta(0) = 1$.
2. Sei $x \in (0, 1)$ beliebig, dann gilt

$$\begin{aligned} (\beta - \alpha) w''_\beta(x) &= v''_\beta(x) - v''_\alpha(x) \\ &= f(x, v_\beta, v'_\beta) - f(x, v_\alpha, v'_\alpha) \\ &=: \phi(1) - \phi(0) \,, \end{aligned} \tag{2.68}$$

wobei

$$\phi(\zeta) = f(x, v_\alpha + \zeta(v_\beta - v_\alpha), v'_\alpha + \zeta(v'_\beta - v'_\alpha)) \,.$$

Kürzt man $v = v_\alpha + \zeta(v_\beta - v_\alpha)$ ab, so ergibt sich

$$\begin{aligned} \phi'(\zeta) &= f_u(x, v, v')(v_\beta - v_\alpha) + f_{u'}(x, v, v')(v'_\beta - v'_\alpha) \\ &= (\beta - \alpha) \left(f_u(x, v, v') w_\beta + f_{u'}(x, v, v') w'_\beta \right) \,, \end{aligned}$$

Eingesetzt in (2.68) erhalten wir:

$$\begin{aligned} w''_\beta(x) &= \frac{\phi(1)-\phi(0)}{\beta-\alpha} \\ &= \frac{1}{\beta-\alpha}\int_0^1 \phi'(\zeta)\,d\zeta \\ &= \int_0^1 \left(f_u(x,v,v')w_\beta + f_{u'}(x,v,v')w'_\beta\right) d\zeta \\ &= \left(\int_0^1 f_u(x,v,v')d\zeta\right) w_\beta + \left(\int_0^1 f_{u'}(x,v,v')d\zeta\right) w'_\beta\,. \end{aligned} \tag{2.69}$$

Man kann zeigen, dass $v = v(\zeta)$ und $v' = v'(\zeta)$ gleichmäßig gegen v_α und v'_α konvergieren. Genauso zeigt man dann, dass w_β gleichmäßig gegen eine Funktion w konvergiert, die dann die Grenzgleichung von (2.69) erfüllt, was genau (2.66) ist. □

Schießverfahren haben Stabilitätsdefizite, die mit *Mehrzielmethoden* verbessert werden können. dazu gibt man sich ein Gitter

$$\Delta = \{0 = x_0 < x_1 < \cdots < x_n = 1\}\ .$$

Wir nehmen nun an, dass folgende Werte von u, der Lösung von (2.63), bekannt sind:

$$\begin{aligned} &\eta_0 = u(0) = 0\,, \quad \alpha_0 = u'(x_0)\,, \\ &\eta_i = u(x_i)\,, \quad \alpha_i = u'(x_i)\,, \qquad i = 0,1,\ldots,n-1\,. \end{aligned}$$

Dann kann die Lösung u aus den Lösungen

$$v_i : [x_{i-1}, x_i] \to \mathbb{R}$$

der Anfangswertprobleme

$$v''_i = f(x, v_i, v'_i)\,, \quad v_i(x_{i-1}) = \eta_{i-1}\,, \quad v'_i(x_{i-1}) = \alpha_{i-1}\,, \quad i = 1,\ldots,n\,. \tag{2.70}$$

Bei Mehrzielmethoden hat man nun also ein $(2n-1)$-dimensionales Gleichungssystem zu lösen:

$$F(z) = \begin{bmatrix} v_1(x_1) - \eta_1 \\ v_1'(x_1) - \alpha_1 \\ \vdots \\ v_{n-1}(x_{n-1}) - \eta_{n-1} \\ v_{n-1}'(x_{n-1}) - \alpha_{n-1} \\ v_n(1) \end{bmatrix} \text{ und } z = \begin{bmatrix} \alpha_0 \\ \eta_1 \\ \alpha_1 \\ \vdots \\ \eta_{n-1} \\ \alpha_{n-1} \end{bmatrix}.$$

Dieses Gleichunsgsystem kann wieder mit Newton Verfahren gelöst werden.

Kapitel 3

Partielle Differentialgleichungen

Numerische Lösungsverfahren für partielle Differentialgleichungen richten sich nach dem Typ der Differentialgleichungen. Man unterscheidet zwischen *elliptischen*, *parabolischen*, und *hyperbolischen* Differentialgleichungen.

Im ersten Teil dieses Kapitels beschäftigen wir uns mit *finite Element Methoden* (FEM) zur Lösung von elliptischen Differentialgleichungen. Wir beschränken uns auf Raumdimension zwei. In höheren Dimensionen gehen viele Argumente analog, aber die Notation wird schwieriger.

Wir verwenden folgende Notation

$$x = (\zeta, \eta)^T \, , x_1 \cdot x_2 = \zeta_1\zeta_2 + \eta_1\eta_2 \text{ und } |x|^2 = \zeta^2 + \eta^2 \, .$$

Ableitungen bezeichnen wir mit

$$\nabla u = (u_\zeta, u_\eta)^T.$$

Das Gebiet Ω, auf dem wir die Differentialgleichung lösen, hat *stückweise linearen Rand* Γ ist *beschränkt* und *zusammenhängend*.

3.1 Finite Element Methoden

zur Lösung von elliptischen Differentialgleichungen in $\Omega \subseteq \mathbb{R}^2$.

Grundlage der *finite Element Methoden* (FEM) sind schwache Lösungen. Dazu braucht man die Theorie der Sobolevräume, die wir kurz zusammenfassen:

Definition 3.1. $H^1(\Omega)$ ist der Hilbertraum der quadratisch integrierbaren Funktionen mit quadratisch integrierbarer Ableitung und dem inneren Produkt

$$\langle u, v\rangle_{H^1(\Omega)} = \int_\Omega \nabla u \cdot \nabla v \, dx + \int_\Omega uv \, dx \; .$$

Neben der zugeörigen Norm $\|\cdot\|_{H^1(\Omega)}$ definieren wir auch die Halbnorm

$$|u|_{H^1(\Omega)} = \int_\Omega |\nabla u|^2 \; dx \; .$$

Funktionen $u \in H^1(\Omega)$ müssen in einzelnen Punkten in Ω nicht stetig sein. Trotzdem kann man sinnvolle Randdaten definieren:

Satz 3.2 ([1])**.** *Jede Funktion $u \in H^1(\Omega)$ besitzt eine eindeutig bestimmte Spur $u|_\Gamma \in L^2(\Gamma)$. Die Zuordnung $u \to u|_\Gamma$ ist eine stetige Abbildung mit folgender Eigenschaft:*

$$\| u|_\Gamma\|^2_{L^2(\Gamma)} \le c_\Omega \|u\|_{L^2(\Omega)} \|u\|_{H^1(\Omega)} \; .$$

Dieser Satz impliziert, dass

$$H^1_0(\Omega) = \left\{ u \in H^1(\Omega) : u|_\Gamma = 0 \right\}$$

ein abgeschlossener linearer Unterraum von $H^1(\Omega)$ ist. Für Funktionen in $H^1_0(\Omega)$ gilt die Poincare-Friedrich Ungleichung:

$$\gamma_\Omega \|u\|_{H^1(\Omega)} \le |u|_{H^1(\Omega)} \text{ für alle } u \in H^1_0(\Omega) \; . \tag{3.1}$$

Nach diesen Vorbereitungen betrachten wir nun die elliptische Differentialgleichung:

$$L[u] := -\nabla \cdot (\sigma \nabla u) + cu = f \text{ in } \Omega \tag{3.2}$$

mit Dirichlet Randdaten

$$u = 0 \text{ auf } \Gamma \; . \tag{3.3}$$

Wir diskutieren hier nicht die genauen Glattheitsanforderungen an c, σ, und f. Die essenziellen Voraussetzungen für uns sind, dass

$$0 < \sigma_0 \le \sigma(x) \le \sigma_\infty \text{ und } 0 \le c(x) \le c_\infty$$

gilt, was bewirkt, dass die Differentialgleichung elliptisch ist. Unter einer *klassischen Lösung* versteht man, wenn die Lösung in $C^2(\overline{\Omega})$ liegt.

Die mathematische Grundlage für schwache Lösungen ist partielle Integration, die wir unter der Annahme, dass die Lösung u von (3.2) und v glatt sind, wie folgt verwenden:

$$\int_\Omega fv\,dx \underbrace{=}_{(3.2)} -\int_\Omega \nabla\cdot(\sigma\nabla u)v\,dx + \int_\Omega cuv\,dx$$
$$= \int_\Omega \sigma\nabla u\nabla v\,dx + \int_\Omega cuv\,dx - \int_\Gamma v\sigma\frac{\partial u}{\partial n}\,ds\ .$$

Definition 3.3. Ein schwache Lösung des *homogenen Dirichletproblems*, der Gleichung (3.2) und (3.3), ist eine Lösungen der Gleichung

$$\int_\Omega fv\,dx = \int_\Omega \sigma\nabla u\nabla v\,dx + \int_\Omega cuv\,dx \text{ für alle } v\in H_0^1(\Omega)\ . \tag{3.4}$$

Bemerkung 3.4. Die schwache Lösung ist eindeutig, und liegt in $H_0^1(\Omega)$. Die Randbedingung ist in dem Sinn erfüllt, dass die Spur der Lösung auf Γ Null ist.

Das *inhomogene Dirichlet Problem* besteht in der Lösung von (3.2) mit Randbedingungen

$$u = g \text{ auf } \Gamma\ . \tag{3.5}$$

Wir erweitern die Funktion g von Γ auf Ω (was unter sehr allgemeinen Voraussetzungen geht – Inverser Spursatz). Mit dieser Erweiterung u_0 reduzieren wir das Problem (3.2), (3.5) auf ein homogenes Dirichlet Problem. Nämlich löst $w = u - u_0$ das homogene Dirichlet Problem

$$L[w] = f - L[u_0]\,,\qquad w|_\Gamma = 0\ .$$

Das *Dirichlet Problem* hat eine eindeutige Lösung:

Satz 3.5. *Es seinen σ, c und f beschränkte Funktionen mit*

$$0\le c(x)\le c_\infty \text{ und } 0<\sigma_0\le\sigma(x)\le c_\infty\ .$$

Dann hat das Dirichelt Problem für jedes $f\in L^2(\Omega)$ und $g\in L^2(\Gamma)$ eine eindeutige schwache Lösung $u\in H^1(\Omega)$.

Das *Neumann Problem* besteht in der Lösung von (3.2) mit Randbedingungen

$$\sigma\frac{\partial u}{\partial n} = g \text{ auf } \Gamma\ . \tag{3.6}$$

Die zugehörigen schwache Form leitet sich wieder durch partielle Integration ab

$$\begin{aligned}\int_\Omega f v\,dx &\underbrace{=}_{(3.2)} \int_\Omega \nabla\cdot(\sigma\nabla u)v\,dx + \int_\Omega cuv\,dx \\ &= \int_\Omega \sigma\nabla u\nabla v\,dx + \int_\Omega cuv\,dx - \int_\Gamma v\sigma\frac{\partial u}{\partial n}\,ds \\ &\underbrace{=}_{(3.6)} \int_\Omega \sigma\nabla u\nabla v\,dx + \int_\Omega cuv\,dx - \int_\Gamma vg\,ds\;.\end{aligned}$$

Existenz und Eindeutigkeit sind durch folgendes Resultat gegeben:

Satz 3.6. *Es seinen σ, c und f beschränkte Funktionen mit*

$$0 < c_0 \le c(x) \le c_\infty \text{ und } 0 < \sigma_0 \le \sigma(x) \le c_\infty\;.$$

Dann hat das Neumann Problem für jedes $f \in L^2(\Omega)$ und $g \in L^2(\Gamma)$ eine eindeutige schwache Lösung $u \in H^1(\Omega)$.

Für $c = 0$ existieren schwache Lösungen, falls zusätzlich gilt, dass

$$\int_\Omega f\,dx = -\int_\Gamma g\,ds\;. \tag{3.7}$$

In diesem Fall unterscheiden sich alle Lösungen um eine Konstante. Üblicherweise wird die ausgezeichnet mit $\int_\Omega u\,dx = 0$ berechnet, und als die *Lösung des Neumannproblems bezeichnet.*

Wir beschäftigen uns mit einer allgemeinen Strategie zur Lösung von elliptischen Differentialgleichungen. Dazu führen wir folgende Terminologie ein:

$$\begin{aligned}a(u,v) &:= \int_\Omega \sigma\nabla u\cdot\nabla v\,dx + \int_\Omega cuv\,dx\,, \\ l(v) &:= \int_\Omega fv\,dx\;.\end{aligned} \tag{3.8}$$

a ist eine Bilinearform (linear in beiden Komponenten) auf $V = H^1(\Omega)$, l ist ein linearer Operator auf V.

Definition 3.7. Sei V ein reeller Vektorraum mit Norm $\|\cdot\|_V$ und W ein abgeschlossener Unterraum von V. Eine Bilinearform $a : V \times V \to \mathbb{R}$ heißt

- symmetrisch, falls $a(u,v) = a(v,u)$ für alle $u, v \in V$ gilt,

- stetig, falls eine Zahl $a_\infty \in \mathbb{R}_+$ existiert, sodass

$$|a(u,v)| \leq a_\infty \|u\|_V \|v\|_V \text{ für alle } u, v \in V\,,$$

- W-elliptisch, falls eine Konstante $\alpha_0 > 0$ existiert mit

$$a(w,w) \geq a_0 \|w\|_V^2 \text{ für alle } w \in W\,.$$

Proposition 3.8. *Es seinen σ, c und f beschränkte Funktionen mit*

$$0 \leq c(x) \leq c_\infty \text{ und } 0 < \sigma_0 \leq \sigma(x) \leq \sigma_\infty$$

Dann ist a aus (3.8) symmetrisch und stetig auf $H_0^1(\Omega)$, mit

$$a_\infty = \max\{\sigma_\infty, c_\infty\}$$

und $H_0^1(\Omega)$-elliptisch mit

$$a_0 = \gamma_\Omega^2 \sigma_0$$

(γ_0 ist die Poincare-Friedrich Konstante (3.1)). Außerdem gilt $a(v,v) \geq 0$ für alle $v \in H^1(\Omega)$.

Beweis. • Für alle $u, v \in H_0^1(\Omega)$ gilt:

$$\begin{aligned}
&|a(u,v)| \\
&\underbrace{\leq}_{\Delta-\text{Ineq.}} \left|\int_\Omega \sigma \nabla u \nabla v \, dx\right| + \left|\int_\Omega cuv \, dx\right| \\
&\leq a_\infty \left(\int_\Omega |\nabla u \cdot \nabla v| \, dx + \int_\Omega |uv| \, dx\right) \\
&\underbrace{\leq}_{\text{Cauchy-Schwarz for functions}} a_\infty \left(\|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} + \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)}\right) \\
&\underbrace{\leq}_{\text{Cauchy-Schwarz for numbers}} a_\infty \left(\sqrt{\|u\|_{L^2(\Omega)}^2 + \|\nabla u\|_{L^2(\Omega)}^2} \cdot \sqrt{\|v\|_{L^2(\Omega)}^2 + \|\nabla v\|_{L^2(\Omega)}^2}\right)\,.
\end{aligned}$$

- Für $v \in H^1(\Omega)$ gilt:

$$\begin{aligned}
|a(v,v)| &= \int_\Omega \sigma |\nabla v|^2 \, dx + \int_\Omega cv^2 \, dx \\
&\geq \sigma_0 \int_\Omega \sigma |\nabla v|^2 \, dx \\
&\geq 0\,.
\end{aligned} \tag{3.9}$$

- From (3.9) it follows for $v \in H_0^1(\Omega)$.

$$|a(v,v)| \geq \sigma_0 \, |v|^2_{H_0^1(\Omega)} \underbrace{\geq}_{(3.1)} \sigma_0 \gamma_\Omega^2 \|v\|^2_{H^1(\Omega)} \, .$$

□

Damit läßt sich die schwache Formulierung der Differentialgleichung in kompakter Form schreiben:

$$a(u,v) = l(v) \text{ für alle } v \in H_0^1(\Omega) \, . \tag{3.10}$$

Um eine Näherungslösung zu bestimmen verwendet man die Galerkin-Methode: Hierzu wählt man einen (endlichdimensionalen) Teilraum $V_h \subseteq H_0^1(\Omega)$ und bestimmen $u_h \in V_h$ welches die Gleichung

$$a(u_h, v_h) = l(v_h) \text{ für alle } v_h \in H_0^1(\Omega) \, . \tag{3.11}$$

Ist nun $\{\phi_1, \ldots, \phi_n\}$ eine Basis von V_h, dann führt der Lösungsansatz

$$u_h = \sum_{i=1}^{n} u_i \phi_i$$

auf das lineare Gleichungssystem

$$\begin{aligned} A\vec{u}_h = b \text{ mit } \vec{u}_h = (u_1, \ldots, u_n)^T, \\ A = [a(\phi_i, \phi_j)]_{ij} \in \mathbb{R}^{n \times n} \, , \vec{b} = [l(\phi_j)]_j \in \mathbb{R}^n \, . \end{aligned} \tag{3.12}$$

Die Matrix A wird *Steifigkeitsmatrix* genannt.

Beispiel 3.9. Wir betrachten das homogene Dirichlet-Problem

$$-u'' = f \text{ in } (0,1) \text{ mit } u(0) = u(1) = 0 \, .$$

Wir wählen für V_h den Raum der linearen Splines über dem äquidistanten Gitter

$$\Delta_h = \{x_i = ih : 0 \leq i \leq n, h = 1/n\}$$

mit homogenen Randdaten. Der Raum der linear Splines besteht aus den Linearkombinationen der Hutfunktionen Λ_i, $i = 1, \ldots, n-1$, die am Knotenpunkt i/n,den Wert 1 haben und an allen Nachbarknoten verschwinden. Damit ergibt sich die Steifigkeitsmatrix

$$a(\Lambda_i, \Lambda_j) = \int_0^1 \Lambda_i'(x) \Lambda_j'(x) \, dx = \begin{cases} 2/h & i = j \\ -1/h & |i-j| = 1 \\ 0 & \text{sonst} \end{cases}$$

Somit lautet das Gleichungssystem ausgeschrieben in diesem Fall

$$\frac{1}{h}\begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \vdots & \\ & \vdots & \vdots & -1 \\ & & -1 & 2 \end{bmatrix}\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-1} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n-1} \end{pmatrix},$$

wobei

$$b_i = \int_0^1 f(x)\Lambda_i(x)\,dx\,, \qquad i = 1, \ldots, n-1\,.$$

Satz 3.10. *Es seinen σ, c und f beschränkte Funktionen mit $c(x) \geq c_0 > 0$ und $\sigma(x) \geq \sigma_0 > 0$. Dann ist die Stefigkeitsmatrix A symmetrisch und positiv definit.*

Beweis. $A = [a_{ij}]$ ist offensichtlich symmetrisch, da

$$a_{ij} = a(\phi_i, \phi_j) = a(\phi_j, \phi_i) = a_{ji}$$

gilt. Ferner gilt für einen beliebigen Vektor $\vec{v} = (v_1, \ldots, v_n)^T \in \mathbb{R}^n$ und $v = \sum_{i=1}^n v_i\phi_i \in V_h$

$$\vec{v}^T A\vec{v} = \sum_{i,j=1}^n v_i v_j a(\phi_i, \phi_j) = a\left(\sum_{i=1}^n v_i\phi_i, \sum_{j=1}^n v_j\phi_j\right) = a(v,v) \underbrace{\geq}_{3.8} 0\,. \quad (3.13)$$

Somit ist A positiv semidefinit.

a ist auch $H_0^1(\Omega)$ elliptisch und $V_h \subseteq H_0^1(\Omega)$, also ist $\vec{v}^T A\vec{v} = 0$ wegen (3.13) genau dann Null wenn die zugehörige Funktion $v \equiv 0$, also nur wenn $\vec{v} = 0$ gilt. □

Unter den obigen Voraussetzungen hat das Gleichungssystem

$$A\vec{u}_h = \vec{b}$$

für das Dirichletproblem eine eindeutige Lösung $\vec{u}_h$.

Satz 3.11 (Ceàs Lemma)**.** *Es seinen σ, c und f beschränkte Funktionen mit $c_\infty \geq c(x) \geq 0$ und $\sigma_\infty \geq \sigma(x) \geq \sigma_0 > 0$. Sei $f \in L^2(\Omega)$. Bezeichne u die schwache Lösung des Dirichlet Problems* (3.2), (3.3)*, und u_h die Galerkin Approximation. Dann gilt*

$$\|u - u_h\|_{H^1(\Omega)} \leq \frac{\max\{\sigma_\infty, c_\infty\}}{\gamma_\Omega^2 \sigma_0} \inf_{v_h \in V_h} \|u - v\|_{H^1(\Omega)}\,,$$

wobei γ_Ω die Konstante aus Poincare-Friedrich Ungleichung (3.1) *ist.*

Beweis. Sei u_h die Lösung von (3.11) beziehungsweise $u \in H_0^1(\Omega)$ die Lösung von (3.10), dann gilt für alle $v \in V_h$:

$$\begin{aligned} a(u-u_h, u-u_h) &= a(u-u_h, u-v) + a(u, v-u_h) - a(u_h, v-u_h) \\ &= a(u-u_h, u-v) + l(v-u_h) - l(v-u_h) \\ &= a(u-u_h, u-v)\ . \end{aligned}$$

Aus der Stetigkeit (mit Stetigkeitskonstante $a_\infty = \max\{\sigma_\infty, c_\infty\}$ und der H_0^1-Elliptizität (mit Elliptizitätskonstante $a_0 = \gamma_\Omega^2 \sigma_0$) von a folgt dann für jedes $v \in V_h$:

$$\begin{aligned} a_0 \|u-u_h\|_{H^1(\Omega)}^2 &\underbrace{\leq}_{\text{Pr. 3.8}} a(u-u_h, u-u_h) \\ &= a(u-u_h, u-v) \\ &\underbrace{\leq}_{\text{Pr. 3.8}} a_\infty \|u-u_h\|_{H^1(\Omega)} \|u-v\|_{H^1(\Omega)}\ . \end{aligned}$$

Mit folgt somit die Behauptung. □

Das Lemma von Ceà zeigt, dass der Fehler der Galerkin-Approximation in V_h höchstens um einen Faktor schlechter ist als die beste Approximation in V_h der Originallösung u.

Bemerkung 3.12. Bei dem inhomogenen Dirichletproblem definiert man sich zuerst eine Funktion $\rho : \Omega \to \mathbb{R}$, die die inhomogenen Randdaten fortsetzt. Dann such man eine Löung $u^\dagger$ von

$$a(\tilde{u} + \rho, v) = l(v) \text{ for all } v \in H_0^1(\Omega)\,, \tag{3.14}$$

Die Lösung des inhomogenen Problems ist $u^\dagger + \rho$.

Bei dem Neumannproblem ist die rechte Seite

$$l(v) = \int_\Omega f v\, dx \ + \ \int_\Gamma g v\, ds$$

und folgendes Gleichungssystem muß gelöst werden:

$$a(u, v) = l(v) \text{ für alle } v \in H^1(\Omega)\ .$$

Zur Realisierung der Galerkin-Methode brauchen wir geeignete *Ansatzräume* $V_h \subseteq H^1(\Omega)$. Üblicherweise verwendet man bei finite Element Methoden *Triangulierungen* um das Gebiet Ω zu zerteilen.

Definition 3.13. Ein Mengensystem von offenen Dreiecken $\Gamma = \{T_1, \ldots, T_m\}$ heißt eine *reguläre Triangulierung* von Ω, falls

1. $T_i \cap T_j = \emptyset$ für alle $i \neq j$,
2. $\bigcup_{i=1}^m \overline{T_i} = \overline{\Omega}$,
3. Für $i \neq j$ gilt enweder,
   (a) $\overline{T_i} \cap \overline{T_j} = \emptyset$,
   (b) $\overline{T_i} \cap \overline{T_j}$ ist eine gemeinsame Ecke von T_i oder T_j, oder
   (c) eine gemeinsame Kante.

Die Ecken der Dreiecke nennt man *Knoten*.

Auf der Triangulierung führt man dann lineare Ansatzfunktionen ein (analog zu linearen Splines).

Satz 3.14. *Sei Γ eine reguläre Triangulierung des polygonalen Gebietes Ω mit den Knoten x_i, $i = 1, \ldots, n$. Dann existieren stetige Funktionen $\Lambda_i : \Omega \to \mathbb{R}$, $i = 1, \ldots, n$, mit den Eigeschaften:*

1. $\Lambda_i(x_j) = \delta_{ij}, \quad i, j = 1, \ldots, n,$
2. $\Lambda_i(x) = \beta_{ik} + \alpha_{ik} \cdot x$ *für* $x \in T_k$ *mit* $\alpha_{ik} \in \mathbb{R}^2$, $\beta_{ik} \in \mathbb{R}$.

Die lineare Hülle $V^\Gamma = span\{\Lambda_1, \ldots, \Lambda_n\}$ besteht aus allen stückweise linearen Funktionen bzgl. der Triangulierung Γ.

Der Gradient einen Elements von V^Γ ist stückweise konstant, und es gilt $V^\Gamma \subseteq H^1(\Omega)$.

Definition 3.15. Das Tupel (Γ, V^Γ) heißt *finite Elemente*.

Das Analogon der Lagrange Interpolation für finite Elemente lautet:

Satz 3.16. *Sei Γ eine reguläre Triangulierung von $\Omega \subseteq \mathbb{R}^2$ mit Knoten $\{x_i : i = 1, \ldots, n\}$. Gegeben seien Werte $\{y_i : i = 1, \ldots, n\}$. Dann gilt $\psi = \sum_{i=1}^n y_i \Lambda_i \in V^\Gamma$ und*

$$\psi(x_i) = y_i, \quad i = 1, \ldots, n.$$

Insbesondere bedeutet unsere Annahme, dass das Gebiet stückweise linearen Rand hat. Im folgenden Beweisen wir ein typisches Resultat für finite Element Methoden:

Satz 3.17. *Sei Γ eine reguläre Triangulierung von $\Omega \subseteq \mathbb{R}^2$ mit Knoten $\{x_i : i = 1, \ldots, n\}$. Wir bezeichnen mit*

- *h die maximale Kantenlänge, und mit*
- *α_0 den minimalen Innenwinkel der Dreiecke T_i, $i = 1, \ldots, n$.*

Dann gilt für alle $f \in H^2(\Omega)$ und die zugehörigen stückweise lineare Interpolanten $\psi \in V^\Gamma$:

$$\|f - \psi\|_{L^2(\Omega)} \leq \sqrt{\frac{3}{8}} h^2 \|f\|_{H^2(\Omega)} \tag{3.15}$$

und

$$|f - \psi|_{H^1(\Omega)} \leq \frac{3}{\sqrt{8}\sin^2(\alpha_0)} h \|f\|_{H^2(\Omega)} \,. \tag{3.16}$$

Beweis. Da $C^2(\overline{\Omega})$ dicht in $H^2(\Omega)$ liegt, reicht es aus (3.15) und (3.16) nur für $f \in C^2(\overline{\Omega})$ zu zeigen.

- Wir zeigen zuerst die Ungleichung (3.15) eingeschränkt auf ein Dreieck $T \in \Gamma$ (anstelle von Ω). Die Ecken von T bezeichnen wir mit x_1, x_2 und x_3. Sei $x \in T$ dann folgt aus der Taylor-Entwicklung

$$f(x_k) = f(x) + \nabla f(x) \cdot d_k + \int_0^1 d_k^t \nabla^2 f(x + t d_k) d_k (1 - t)\, dt \,, \tag{3.17}$$

wobei

$$d_k := x_k - x$$

bezeichnet. Somit gilt für den lineare Interpolanten in diesem Dreieck:

$$\begin{aligned} \Psi(x) &= \sum_{k=1}^{3} f(x_k) \Lambda_k(x) \\ &= f(x) \sum_{k=1}^{3} \Lambda_k(x) + \nabla f(x) \cdot \sum_{k=1}^{3} d_k \Lambda_k(x) \\ &\quad + \sum_{k=1}^{3} \left(\int_0^1 d_k^t \nabla^2 f(x + t d_k) d_k (1 - t)\, dt \right) \Lambda_k(x) \,. \end{aligned} \tag{3.18}$$

Da

$$\sum_{k=1}^{3} \Lambda_k(x) = 1 \text{ und } \sum_{k=1}^{3} x_k \Lambda_k(x) = x \,, \tag{3.19}$$

und somit:

$$\sum_{k=1}^{3} d_k \Lambda_k(x) = \sum_{k=1}^{3} x_k \Lambda_k(x) - \sum_{k=1}^{3} x \Lambda_k(x) = x - x = 0 \,. \tag{3.20}$$

Insbesondere gilt auch

$$\sum_{k=1}^{3} \nabla \Lambda_k(x) = \vec{0} \text{ and } \sum_{k=1}^{3} x_k \nabla \Lambda_k(x) = \vec{1} \,. \tag{3.21}$$

Darüberhinaus gilt

$$\sum_{k=1}^{3} \Lambda_k^2(x) \leq \sum_{k=1}^{3} \Lambda_k(x) = 1 \,. \tag{3.22}$$

Damit ergibt sich aus (3.18):

$$\Psi(x) - f(x) = \sum_{k=1}^{3} \left(\int_0^1 d_k^t \nabla^2 f(x + t d_k) d_k (1-t)\, dt \right) \Lambda_k(x) \,,$$

Aus der Cauchy-Schwarz Ungleichung folgt somit

$$\begin{aligned} &(\Psi(x) - f(x))^2 \\ \leq &\sum_{k=1}^{3} \left(\int_0^1 d_k^t \nabla^2 f(x + t d_k) d_k (1-t)\, dt \right)^2 \underbrace{\sum_{k=1}^{3} \Lambda_k^2(x)}_{(3.22)\leq 1} \,. \end{aligned}$$

Integration über das Dreeick T liefert

$$\begin{aligned} &\int_T (\Psi(x) - f(x))^2 \, dx \\ \leq &\sum_{k=1}^{3} \underbrace{\int_T \left(\int_0^1 d_k^t \nabla^2 f(x + t d_k) d_k (1-t)\, dt \right)^2 dx}_{:= I_k} \,. \end{aligned} \tag{3.23}$$

- Als nächstes studieren wir das Integral I_k für ein festes $k = 1, 2, 3$. Wir definieren uns nun eine Kreissektor S_k mit Zentrum x_k und Öffnungswinkel γ_k und Radius h_T = maximale Seitenlänge des Dreiecks. In S_k führen wir Polarkoordinaten ein

$$x = x_k + r x_\theta \text{ mit } x_\theta = \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix} , \theta_0 \leq \theta \leq \theta_0 + \gamma_k \,, \quad 0 \leq r \leq h_T \,.$$

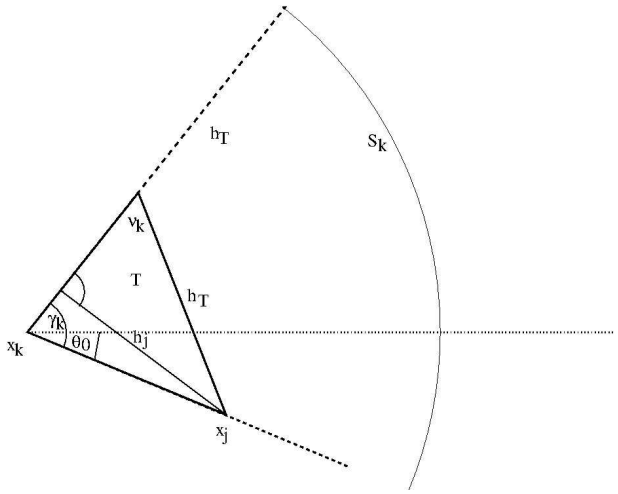

Abbildung 3.1: Wichtigste Information zum Beweis

Siehe Figure 3.1. Somit gilt

$$d_k = x_k - x = -rx_\theta \text{ und } x + td_k = x_k + (1-t)rx_\theta \ .$$

Damit und mit der Bezeichnung

$$\overline{\nabla^2 f}(x) := \begin{cases} \nabla^2 f(x) & \text{für } x \in T \\ 0 & \text{für } x \in \mathbb{R}^2 \backslash T \end{cases} \ ,$$

erhalten wir:

$$I_k = \int_{\theta_0}^{\theta_0+\gamma_k} \int_0^{h_T} \left(\int_0^1 r^2 x_\theta^t \overline{\nabla^2 f}(x_k + (1-t)rx_\theta) x_\theta (1-t)\, dt \right)^2 r\, dr d\theta \ .$$

Mit der Substitution $s = s(t) = (1-t)r$ und der Cauchy-Schwarz Ungleichung folgt dann:

$$\begin{aligned} I_k &= \int_{\theta_0}^{\theta_0+\gamma_k} \int_0^{h_T} \left(\int_0^r x_\theta^t \overline{\nabla^2 f}(x_k + sx_\theta) x_\theta \sqrt{s}\sqrt{s}\, ds \right)^2 r\, dr d\theta \\ &\leq \int_{\theta_0}^{\theta_0+\gamma_k} \int_0^{h_T} \left(\int_0^r \left| x_\theta^t \overline{\nabla^2 f}(x_k + sx_\theta) x_\theta \right|^2 s\, ds \right) \left(\int_0^r s\, ds \right) r\, dr d\theta \\ &= \frac{1}{2} \int_{\theta_0}^{\theta_0+\gamma_k} \int_0^{h_T} \left(\int_0^r \left| x_\theta^t \overline{\nabla^2 f}(x_k + sx_\theta) x_\theta \right|^2 s\, ds \right) r^3\, dr d\theta \end{aligned}$$

Nun verwenden wir, dass die Frobenius-Norm mit der Euklid-Norm verträglich ist, also, dass

$$\begin{aligned}\left|x_\theta^t \overline{\nabla^2 f}(x_k + sx_\theta)x_\theta\right|^2 &\underbrace{\leq}_{\text{C-S}} \underbrace{|x_\theta|^2}_{=1} \left|\overline{\nabla^2 f}(x_k + sx_\theta)x_\theta\right|^2 \\ &\leq \|\overline{\nabla^2 f}(x_k + sx_\theta)\|_F^2 \,.\end{aligned}$$

Aus den letzten beiden Ungleichungen folgt somit:

$$\begin{aligned} I_k &\leq \frac{1}{2}\int_{\theta_0}^{\theta_0+\gamma_k} \left(\int_0^{h_T} \|\overline{\nabla^2 f}(x_k + sx_\theta)\|_F^2 s\, ds\right)\left(\int_0^{h_T} r^3\, dr\right) d\theta \\ &= \frac{1}{8}h_T^4 \int_{\theta_0}^{\theta_0+\gamma_k}\int_0^{h_T} \|\overline{\nabla^2 f}(x_k + sx_\theta)\|_F^2 s\, ds d\theta \\ &\leq \frac{1}{8}h_T^4 \|f\|_{H^2(T)}^2 \\ &\leq \frac{1}{8}h^4 \|f\|_{H^2(T)}^2 \,. \end{aligned} \tag{3.24}$$

Aus (3.23) und (3.24) folgt somit

$$\begin{aligned}\|f-\psi\|_{L^2(\Omega)}^2 &= \sum_{T\in\Gamma} \|f-\psi\|_{L^2(T)}^2 \\ &\leq \sum_{T\in\Gamma}\sum_{k=1}^{3} I_k(T) \\ &\leq \frac{3}{8}h^4 \sum_{T\in\Gamma} \|f\|_{H^2(T)}^2 \\ &= \frac{3}{8}h^4 \|f\|_{H^2(\Omega)}^2 \,,\end{aligned}$$

und somit der erste Teil der Behauptung, (3.15)

- Die Hutfunktionen Λ_j, $j = 1, 2, 3$ verschwinden auf allen Knoten außer x_j. Der Gradient der Hutfunktion ist konstant auf T. Bezeichnen wir mit $c = |\nabla\Lambda_j|$, so gilt, weil die Hutfunktion auch linear auf der Normallinie Linie h_j ist (siehe Figure 3.1) und dort die Steigung c hat:

$$|\nabla\Lambda_j| = \frac{1}{h_j} \,.$$

Somit gilt

$$h_j = \cos(90 - \gamma_k)\,|x_k - x_j| = \sin(\gamma_k)\,|x_k - x_j| \ . \tag{3.25}$$

Da wir vorausgesetzt haben, dass alle Winkel der Triangulation innere Winkel sind alle Winkel zwischen 0° und 180° und somit sind die Sinuse der Winkel positiv. Aus dem Sinussatz ergibt sich

$$|x_k - x_j| = h_T \underbrace{\frac{\sin(\nu_k)}{\sin(\gamma_k)}}_{|\cdot|\leq 1} \geq h_T \sin(\nu_k) \geq h_T \sin(\alpha_0)\,,$$

und daraus ergibt sich somit:

$$\begin{aligned} |\nabla \Lambda_j| &= \frac{1}{h_j} \\ &= \frac{1}{|x_k - x_j| \sin\gamma_k} \\ &\leq \frac{1}{h_T \sin(\alpha_0)\sin(\gamma_k)} \\ &\leq \frac{1}{h_T \sin^2(\alpha_0)}\,, \qquad j = 1,2,3\,. \end{aligned} \tag{3.26}$$

- Wir verwenden nun folgende Hilfsidentität: Sei $z = (z_1, z_2)^t \in \mathbb{R}^2$ beliebig, dann gilt

$$\begin{aligned} &\sum_{k=1}^{3} (z \cdot x_k) \nabla \Lambda_k(x) \\ =& z_1 \nabla \left(\sum_{k=1}^{3} x_{k,1} \Lambda_k(x) \right) + z_2 \nabla \left(\sum_{k=1}^{3} x_{k,2} \Lambda_k(x) \right) \\ \underbrace{=}_{(3.19)}& z_1 \nabla x_1 + z_2 \nabla x_2 \\ =& z\,. \end{aligned} \tag{3.27}$$

Somit folgt aus (3.17),

$$\begin{aligned}
&\nabla\Psi(x)\\
=&\sum_{k=1}^{3} f(x_k)\nabla\Lambda_k(x)\\
=&f(x)\underbrace{\sum_{k=1}^{3}\nabla\Lambda_k(x)}_{=\vec{0}(3.21)}+\underbrace{\sum_{k=1}^{3}(\nabla f(x)\cdot x_k)\nabla\Lambda_k(x)}_{=\nabla f(x)(3.27)}-(\nabla f(x)\cdot x)\underbrace{\sum_{k=1}^{3}\nabla\Lambda_k(x)}_{=\vec{0}(3.21)}+\\
&\sum_{k=1}^{3}\left(\int_0^1 d_k^t\nabla^2 f(x+td_k)d_k(1-t)\,dt\right)\nabla\Lambda_k(x)\\
=&\nabla f(x)+\sum_{k=1}^{3}\left(\int_0^1 d_k^t\nabla^2 f(x+td_k)d_k(1-t)\,dt\right)\nabla\Lambda_k(x)\,.
\end{aligned}$$

Mit der Cauchy-Schwarz Ungleichung und (3.26) folgt somit:

$$\begin{aligned}
&|\nabla(\Psi-f)(x)|^2\\
\leq&\sum_{k=1}^{3}\left(\int_0^1 d_k^t\nabla^2 f(x+td_k)d_k(1-t)\,dt\right)^2\sum_{k=1}^{3}|\nabla\Lambda_k(x)|^2\\
\underbrace{\leq}_{(3.26)}&\frac{3}{\sin^4(\alpha_0)}\frac{1}{h_T^2}\sum_{k=1}^{3}\left(\int_0^1 d_k^t\nabla^2 f(x+td_k)d_k(1-t)\,dt\right)^2\,.
\end{aligned}$$

Integration über T liefert dann:

$$\begin{aligned}
|\Psi-f|^2_{H^1(T)}&=\int_T|\nabla(\Psi-f)(x)|^2\\
&\leq\frac{3}{\sin^4(\alpha_0)}\frac{1}{h_T^2}\sum_{k=1}^{3}\underbrace{I_k}_{=I_k(T)}\\
&\underbrace{\leq}_{(3.24)}\frac{9}{8\sin^4(\alpha_0)}h_T^2\|f\|^2_{H^2(T)}\,.
\end{aligned}$$

Durch Summmation über $T\in\Gamma$ erhält man somit die Behauptung.

□

3.2 Fehlerschranken für Finite-Element Methoden

Finite Elemente Methoden lösen partielle Differentialgleichungen mit Ansatzfunktionen (klassisch Hutfunktionen) auf einer regulären Triangulierung. Es gibt auch Methoden höherer Ordnung, auf die wir aber in der Vorlesung nicht eingehen.

Wir betrachten das elliptische Dirichletproblem:

$$-\nabla \cdot (\sigma \nabla u) + cu = f \text{ in } \Omega\,, \qquad u = 0 \text{ on } \Gamma\,, \tag{3.28}$$

mit $0 < \sigma_0 \leq \sigma(x) \leq \sigma_\infty$ und $0 \leq c(x) \leq c_\infty$. Sei u die Lösung von (3.28) und u_h die Galerkin-Approximation auf dem Raum V_h der linearen Hutfunktionen mit 0 Dirichlet-Randbedingungen.

Wir setzen voraus, dass Γ eine reguläre Triangulierung mit maximaler Kantenlänge h und minimalen Innenwinkel $\alpha_0 > 0$ ist.

Satz 3.18. *Sei $u \in H^2(\Omega) \cap H_0^1(\Omega)$ ein Lösung von (3.28). Dann existiert eine Konstante $c_1 > 0$ (die nicht von f abhängt), sodass:*

$$\|u - u_h\|_{H^1(\Omega)} \leq c_1 h \|u\|_{H^2(\Omega)}\,.$$

Beweis. Aus Ceàs Lemma, Theorem 3.11, folgt für die linear finite Element Funktion $\psi \in V_h$, die u an den Knoten von Γ interpoliert:

$$\begin{aligned}\|u - u_h\|_{H^1(\Omega)} &\leq \frac{\max\{\sigma_\infty, c_\infty\}}{\gamma_\Omega^2 \sigma_0} \inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)} \\ &\leq \frac{\max\{\sigma_\infty, c_\infty\}}{\gamma_\Omega^2 \sigma_0} \|u - \psi\|_{H^1(\Omega)}\,,\end{aligned}$$

Aus Satz 3.17 folgt, dass eine Konstante κ_1 existiert, die von α_0 abhängt, sodass

$$\|u - \psi\|_{H^1(\Omega)} \leq \kappa_1 h \|u\|_{H^2(\Omega)}\,,$$

woraus nun die Behauptung folgt. □

Dieser Satz benötigt Glattheit der Lösung ($u \in H^2(\Omega)$). Solche Lösungstheorie findet man in klassischen Bücher über partielle Differentialgleichungen, wie etwa [4]. Ein solches Resultat besagt: Ist Ω ein konvexes, polygonales Gebiet und ist $\sigma \in C^1(\overline{\Omega})$, so gibt es eine Konstante c_2 gibt, die nur von Ω abhängt, sodass

$$\|u\|_{H^2(\Omega)} \leq c_2 \|f\|_{L^2(\Omega)}\,. \tag{3.29}$$

Satz 3.19. *Sei Ω ein konvexes, polygonales Gebiet und $\sigma \in C^1(\overline{\Omega})$. Dann gilt*

$$\|u - u_h\|_{L^2(\Omega)} \leq c_0 h^2 \|u\|_{H^2(\Omega)} \text{ wobei } c_0 = \max\{\sigma_\infty, c_\infty\}\, c_1^2 c_2\,,$$

wobei c_1, c_2 die Konstanten aus Theorem 3.18 und (3.29) *sind.*

Beweis. Sei w eine Lösung von

$$-\nabla \cdot (\sigma \nabla w) + cw = u - u_h \text{ in } \Omega\,, \qquad w = 0 \text{ on } \Gamma\,. \tag{3.30}$$

Wegen (3.29) gilt $w \in H^2(\Omega) \cap H_0^1(\Omega)$. Außerdem löst sie das Variationsproblem:

$$a(w,v) := \sigma \int_\Omega \nabla w \cdot \nabla v + cwv\, dx \;=\; \int_\Omega (u - u_h) v\, dx \text{ für alle } v \in H_0^1(\Omega)\,. \tag{3.31}$$

Mit dieser Bezeichung gilt auch

$$a(u,v) = a(u_h, v) = \int_\Omega fv\, dx \text{ und } v \in V_h\,,$$

und die Galerkin-Approximation $w_h \in V_h$ von (3.30) erfüllt die Gleichung

$$a(w_h, v) = \int_\Omega (u - u_h) v\, dx \text{ für alle } v \in V_h\,. \tag{3.32}$$

Damit gilt insbesondere

$$0 = a(u - u_h, w_h) = a(w_h, u - u_h)\,. \tag{3.33}$$

Sei $a_\infty = \max\{\sigma_\infty, c_\infty\}$ wie in Proposition 3.8, dann gilt:

$$\begin{aligned}
\|u - u_h\|_{L^2(\Omega)}^2 &= \int_\Omega (u - u_h)^2\, dx \\
&\underbrace{=}_{(3.31)} a(w, u - u_h) \\
&\underbrace{=}_{(3.33)} a(w - w_h, u - u_h) \\
&\leq a_\infty \|w - w_h\|_{H^1(\Omega)} \|u - u_h\|_{H^1(\Omega)} \\
&\underbrace{\leq}_{\text{Satz 3.18}} a_\infty c_1 h \|w\|_{H^2(\Omega)} c_1 h \|u\|_{H^2(\Omega)} \\
&\underbrace{\leq}_{(3.29)} a_\infty c_1^2 c_2 h^2 \|u - u_h\|_{L^2(\Omega)} \|u\|_{H^2(\Omega)}\,,
\end{aligned}$$

und durch Durchdividieren die Behauptung. □

3.3 Steifigkeitsmatrix

Die finite Element Methode reduziert sich auf das Lösen des Gleichungssystems

$$A\vec{u}_h = b\,, \tag{3.34}$$

mit Steifigkeitsmatrix A, wobei

$$a_{ij} = a(\phi_i, \phi_j) = \int_\Omega \sigma \nabla\phi_i \cdot \nabla\phi_j + c\phi_i\phi_j \, dx\ . \tag{3.35}$$

Die Matrix ist dünn besetzt.

Für jedes Dreieck $T_k \in \Gamma$ ist

$$S_k = \left[\int_{T_k} \sigma \nabla\phi_i \cdot \nabla\phi_j + c\phi_i\phi_j \, dx\right]_{ij} \in \mathbb{R}^{n\times n} \tag{3.36}$$

eine Matrix, die aus allen Teilintegralen über dieses eine Dreieck besteht. Diese Matrizen werden Elementsteifigkeitsmatrizen genannt. Diese Matrizen können leichter analytisch berechnet werden als die Matrix A. Wegen

$$\begin{aligned} a(\phi_i, \phi_j) &= \int_\Omega \sigma \nabla\phi_i \cdot \nabla\phi_j + c\phi_i\phi_j \, dx \\ &= \sum_{k=1}^{m} \int_{T_k} \sigma \nabla\phi_i \cdot \nabla\phi_j + c\phi_i\phi_j \, dx \end{aligned}$$

folgt daraus

$$A = \sum_{k=1}^{m} S_k\ . \tag{3.37}$$

Zur Berechnung der Elementsteifigkeitsmatrizen wird mit der Transformation

$$\Phi(s,t) = x_1 + s(x_2 - x_1) + t(x_3 - x_1)\,, \tag{3.38}$$

das *Referenzdreieck*

$$D = \left\{z = (s,t)^t : s > 0, t > 0, s + t < 1\right\} \tag{3.39}$$

auf ein Dreieck $T \in \Gamma$ mit den Ecken $x_i = (\zeta_i, \eta_i)^t$, $i = 1, 2, 3$ abgebildet. Damit gilt

$$\Phi'(s,t) = [x_2 - x_1 \quad x_3 - x_1] = \begin{bmatrix} \zeta_2 - \zeta_1 & \zeta_3 - \zeta_1 \\ \eta_2 - \eta_1 & \eta_3 - \eta_1 \end{bmatrix}\ .$$

Die beiden Vektoren sind in einem nicht degenerierenden Dreieck linear unabhängig, woraus folgt, dass

$$d = \det\Phi' = (\zeta_2 - \zeta_1)(\eta_3 - \eta_1) - (\zeta_3 - \zeta_1)(\eta_2 - \eta_1) \neq 0 \; .$$

Somit gilt

$$\Phi'^{-1} = \frac{1}{d} \begin{bmatrix} \eta_3 - \eta_1 & \zeta_1 - \zeta_3 \\ \eta_1 - \eta_2 & \zeta_2 - \zeta_1 \end{bmatrix} \; .$$

Beispiel 3.20. Wir berechnen die Elementsteifigkeitsmatrix $S = [s_{ij}]$ für den Fall des Laplace-Operators $L[u] - \Delta u$ und ein Dreieck $T \in \Gamma$ mit dessen Ecken x_1, x_2 und x_3. Wir bezeichen mit Λ_i, $i = 1, 2, 3$ die Hutfunktionen, die an den Knoten x_i den Wert 1 haben, und sonst Null sind. Dann ergibt sich:

$$\begin{aligned} s_{ij} &= \int_\Omega \nabla_x \Lambda_i(x) \cdot \nabla_x \Lambda_j(x) \, dx \\ &= \int_D \nabla_x \Lambda_i(\Phi(z)) \cdot \nabla_x \Lambda_j(\Phi(z)) \, |\det\Phi'| \; dz \\ &= |d| \int_D (\Phi'^{-t} \nabla_z \Lambda_i(\Phi(z))) \cdot (\Phi'^{-t} \nabla_z \Lambda_j(\Phi(z)) \, |\det\Phi'|) \, dz \; . \end{aligned}$$

Die Funktion $\Lambda_i(\Phi(\cdot))$ ist wieder eine Hutfunktion über D mit Wert 1 an der Ecke z_i, und Null an den anderen Ecken. Daher ist

$$G := \begin{bmatrix} \nabla_z(\Lambda_1(\Phi(\cdot))^t \\ \nabla_z(\Lambda_2(\Phi(\cdot))^t \\ \nabla_z(\Lambda_3(\Phi(\cdot))^t \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \; .$$

Die Integranden von s_{ij} sind also konstant und der Flächeninhalt von D ist 0.5 folgt damit für $i, j = 1, 2, 3$:

$$\begin{aligned} s_{ij} &= \frac{|d|}{2} G \Phi'^{-1} \Phi'^{-t} G^t \\ &= \frac{1}{2\,|d|} \begin{bmatrix} \eta_2 - \eta_3 & \zeta_3 - \zeta_2 \\ \eta_3 - \eta_1 & \zeta_1 - \zeta_3 \\ \eta_1 - \eta_2 & \zeta_2 - \zeta_1 \end{bmatrix} \begin{bmatrix} \eta_2 - \eta_3 & \eta_3 - \eta_1 & \eta_1 - \eta_2 \\ \zeta_3 - \zeta_2 & \zeta_1 - \zeta_3 & \zeta_2 - \zeta_1 \end{bmatrix} \end{aligned}$$