Vertiefung Numerische Mathematik für den Masterstudiengang UTRM

Vorlesungsskriptum SS 2010

R. Verfürth

Fakultät für Mathematik, Ruhr-Universität Bochum

Inhaltsverzeichnis

Einleitung

Das vorliegende Skript baut auf demjenigen der Vorlesung „Numerische Mathematik für Maschinenbauer, Bauingenieure und Umwelttechniker“ [6] auf. Dort werden u.a. numerische Verfahren für

- lineare und nichtlineare Gleichungssysteme,
- die Integration,
- Anfangswertprobleme für gewöhnliche Differentialgleichungen

behandelt. Hier hingegen werden numerische Verfahren für

- Randwertprobleme für gewöhnliche Differentialgleichungen,
- elliptische, parabolische und hyperbolische partielle Differentialgleichungen

vorgestellt. Die meisten Verfahren aus [6] sind in dem Demonstrations-Applet `Numerics` implementiert, die adaptiven Finite-Element-Methoden für elliptische Differentialgleichungen dieses Skriptums in dem Applet `ALF`. Beide Applets stehen zusammen mit englisch-sprachigen Bedienungsanleitungen auf der Seite

`www.rub.de/num1/demoapplets.html`

zur Verfügung. Analytische Grundlagen zu den betrachteten Problemen findet man u.a. in [5].

KAPITEL I

Randwertprobleme für gewöhnliche Differentialgleichungen

I.1. Anfangswertprobleme

In diesem und dem nächsten Abschnitt erinnern an einige Eigenschaften von Anfangswertproblemen für gewöhnliche Differentialgleichungen und deren numerische Approximation. Für weitere Details verweisen wir auf [6, Kapitel VI].

Gegeben sind ein Intervall I, eine Teilmenge D des $\mathbb{R}^d$, eine Funktion $f(t,y) : I \times D \to \mathbb{R}^d$, eine *Anfangszeit* $t_0 \in I$ und ein *Anfangswert* $y_0 \in D$. Bei einem *Anfangswertproblem* ist eine differenzierbare Funktion $y : I \to D$ gesucht mit

$$\boxed{\begin{aligned} y'(t) &= f(t, y(t)) \quad \text{für alle } t \in I \quad \textit{(Differentialgleichung)} \\ y(t_0) &= y_0 \qquad\qquad\qquad\qquad\quad \textit{(Anfangsbedingung)} \end{aligned}}$$

Beispiel I.1.1. Die Funktion y beschreibe die Größe einer Population mit konstanter Sterbe- oder Wachstumsrate λ zur Zeit t. Dann erfüllt y das Anfangswertproblem

$$\begin{aligned} y'(t) &= \lambda y(t), \\ y(0) &= c. \end{aligned}$$

Hier ist

$$I = \mathbb{R},\ D = \mathbb{R},\ f(t,y) = \lambda y,\ t_0 = 0,\ y_0 = c.$$

Die exakte Lösung ist

$$y(t) = ce^{\lambda t}.$$

Beispiel I.1.2. Das Anfangswertproblem

$$\begin{aligned} y'(t) &= \begin{pmatrix} \lambda & -\omega \\ \omega & \lambda \end{pmatrix} y(t) \\ y(0) &= \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \end{aligned}$$

beschreibt eine *gedämpfte Schwingung*. Hier ist

$$I = \mathbb{R},\ D = \mathbb{R}^2,\ f(t,y) = \begin{pmatrix} \lambda & -\omega \\ \omega & \lambda \end{pmatrix} y,\ t_0 = 0,\ y_0 = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}.$$

Die exakte Lösung lautet

$$y(t) = e^{\lambda t} \begin{pmatrix} c_1 \cos(\omega t) - c_2 \sin(\omega t) \\ c_1 \sin(\omega t) + c_2 \cos(\omega t) \end{pmatrix}.$$

Beispiel I.1.3. Für das Anfangswertproblem

$$\begin{aligned} y'(t) &= y(t)^2 \\ y(0) &= 1 \end{aligned}$$

ist

$$I = \mathbb{R},\ D = \mathbb{R},\ f(t,y) = y^2,\ t_0 = 0,\ y_0 = 1.$$

Die exakte Lösung ist

$$y(t) = \frac{1}{1-t}.$$

Die Lösung *explodiert*, wenn sich die Zeit t von 0 kommend dem Wert 1 nähert.

Beispiel I.1.4. Für das Anfangswertproblem

$$\begin{aligned} y'(t) &= \sqrt{|y(t)|} \\ y(0) &= 0 \end{aligned}$$

ist

$$I = \mathbb{R},\ D = \mathbb{R},\ f(t,y) = \sqrt{|y|},\ t_0 = 0,\ y_0 = 0.$$

Es hat *unendlich viele Lösungen*; zwei davon sind gegeben durch

$$y(t) = 0 \quad \text{für alle } t,$$

$$y(t) = \begin{cases} 0 & \text{für } t < 0, \\ \frac{1}{4}t^2 & \text{für } t \geq 0. \end{cases}$$

Abbildung I.1.1 zeigt eine typische Lösung.

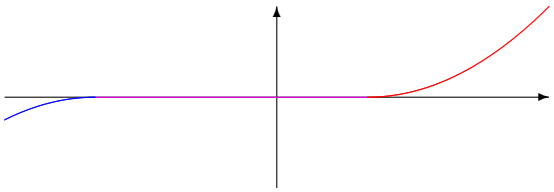


Abbildung I.1.1. Typische Lösung von Beispiel I.1.4

Beispiel I.1.5. Differentialgleichungen höherer Ordnung können durch Einführen neuer Unbekannter in Differentialgleichungen erster Ordnung transformiert werden. Für das *mechanische System*

$$\begin{aligned} Mx''(t) + Rx'(t) + Kx(t) &= F(t) \\ x(0) &= x_0 \\ x'(0) &= v_0 \end{aligned}$$

im $\mathbb{R}^d$ führt man z.B. die Funktion $v(t) = x'(t)$ als neue Unbekannte ein und erhält für diese die Gleichung

$$v'(t) = M^{-1}F(t) - M^{-1}Rv(t) - M^{-1}Kx(t).$$

Dadurch wird die ursprüngliche Differentialgleichung zweiter Ordnung im $\mathbb{R}^d$ in eine erster Ordnung im $\mathbb{R}^{2d}$ überführt, die den Daten

$$y(t) = \begin{pmatrix} x(t) \\ v(t) \end{pmatrix},$$
$$f(t,y) = \begin{pmatrix} 0 \\ M^{-1}F(t) \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ -M^{-1}K & -M^{-1}R \end{pmatrix} y$$

entspricht.

Wie die obigen Beispiele zeigen, ist nicht jedes Anfangswertproblem eindeutig lösbar und hat nicht jedes Anfangswertproblem eine Lösung, die auf dem ganzen Intervall I, auf dem die rechte Seite f definiert ist, existiert. Das folgende Ergebnis gibt ein einfach nachprüfbares Kriterium für die Existenz einer eindeutigen Lösung und deren Definitionsbereich. Es ist nicht das schärfst mögliche. Die Bedingung an f kann dahingehend abgeschwächt werden, dass es eine Konstante L geben muss, so dass für alle $t \in I$ und alle $y_1, y_2 \in D$ gilt

$$|f(t,y_1) - f(t,y_2)| \leq L|y_1 - y_2|. \quad \textit{(Lipschitz-Bedingung)}$$

Stetig differenzierbare Funktionen mit beschränkter Ableitung erfüllen die Lipschitz-Bedingung. Die Funktion $y \mapsto |y|$ ist ein Beispiel für eine nicht differenzierbare Funktion, die die Lipschitz-Bedingung erfüllt. Die Funktionen f aus den Beispielen I.1.3 und I.1.4 erfüllen die Lipschitz-Bedingung nicht.

Falls f stetig differenzierbar ist bzgl. der Variablen y, gibt es ein Intervall $J = (t_-, t_+)$ mit $t_0 \in J$ und eine eindeutige auf J stetig differenzierbare Funktion y, die das Anfangswertproblem

$$y'(t) = f(t, y(t)),$$
$$y(t_0) = y_0.$$

löst.
Es ist $J = I$ oder $y(t)$ strebt für $t \to t_\pm$ gegen den Rand von D.
Ist die Ableitung von f bzgl. der Variablen y auf $I \times D$ beschränkt, so ist $J = I$.

In Hinblick auf Randwertprobleme ist die Abhängigkeit der Lösung eines Anfangswertproblems vom Anfangswert zentral. Das folgende Ergebnis gibt ein einfaches Kriterium an, wann die Lösung eines Anfangswertproblems differenzierbar vom Anfangswert abhängt und wie

die Ableitung bestimmt werden kann. Letzteres wird für die praktische Lösung mit dem *Schießverfahren* in Abschnitt I.4 und der *Mehrzielmethode* in Abschnitt I.5 wesentlich sein.

Falls f zweimal stetig differenzierbar ist bzgl. der Variablen y, hängt die Lösung y des Anfangswertproblems

$$y'(t) = f(t, y(t))$$
$$y(t_0) = y_0$$

differenzierbar vom Anfangswert y_0 ab, d.h.

$$y(t) = y(t; y_0).$$

Die Ableitung $Z(t)$ der Funktion $y_0 \mapsto y(t; y_0)$ löst das Anfangswertproblem

$$Z'(t) = D_y f(t, y(t; y_0))Z(t),$$
$$Z(t_0) = I.$$

Dabei ist $D_y f(t, y)$ die *Jacobi-Matrix* von f bzgl. der Variablen y und I die *Einheitsmatrix*.

Beispiel I.1.6. Für die gedämpfte Schwingung aus Beispiel I.1.2 ist

$$f(t, y) = \begin{pmatrix} \lambda & -\omega \\ \omega & \lambda \end{pmatrix} y.$$

Die Jacobi-Matrix von f ist

$$D_y f(t, y) = \begin{pmatrix} \lambda & -\omega \\ \omega & \lambda \end{pmatrix}.$$

Die Funktion Z hat ihre Werte in der Menge der 2×2 Matrizen. Das Anfangswertproblem für Z lautet

$$Z'(t) = \begin{pmatrix} \lambda & -\omega \\ \omega & \lambda \end{pmatrix} Z(t),$$
$$Z(0) = I.$$

In Komponenten-Schreibweise hat dieses die Form

$$\begin{aligned}
z'_{1,1}(t) &= \lambda z_{1,1}(t) - \omega z_{2,1}(t), \quad z_{1,1}(0) = 1, \\
z'_{1,2}(t) &= \lambda z_{1,2}(t) - \omega z_{2,2}(t), \quad z_{1,2}(0) = 0, \\
z'_{2,1}(t) &= \omega z_{1,1}(t) + \lambda z_{2,1}(t), \quad z_{2,1}(0) = 0, \\
z'_{2,2}(t) &= \omega z_{1,2}(t) + \lambda z_{2,2}(t), \quad z_{2,2}(0) = 1.
\end{aligned}$$

I.2. Numerische Verfahren für Anfangswertprobleme

Die numerische Lösung von Anfangswertproblemen beruht auf folgender Grundidee:

Approximiere die Lösung y des Anfangswertproblems zu Zeiten $t_0 < t_1 < t_2 < \dots$ und bezeichne mit η_i die Approximation für $y(t_i)$. Berechne für $i = 0, 1, \dots$ sukzessive η_{i+1} aus f und η_i *(Einschrittverfahren)* oder aus f und $\eta_i, \dots, \eta_{i-m}$ *(Mehrschrittverfahren)*.

Viele Verfahren, insbesondere *Runge-Kutta-Verfahren*, erhält man durch Anwenden einer *Quadraturformel* auf das Integral in der Identität

$$\eta_{i+1} - \eta_i \approx y(t_{i+1}) - y(t_i) = \int_{t_i}^{t_{i+1}} f(s, y(s))ds.$$

Zur Abkürzung bezeichnet man mit $h_i = t_{i+1} - t_i$ die i-te Schrittweite. Im einfachsten Fall sind die Punkte $t_0, t_1, \dots$ *äquidistant*, d.h. $h_i = h$ und $t_i = t_0 + ih$ für alle i.

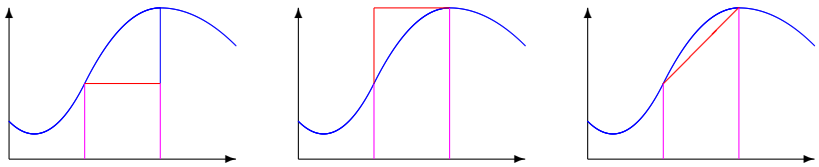


ABBILDUNG I.2.1. Quadraturformeln für das explizite Euler-Verfahren, das implizite Euler-Verfahren und die Trapezregel

Die einfachsten Verfahren sind das *explizite Euler-Verfahren*, das *implizite Euler-Verfahren* und die *Trapezregel*, die auch unter dem Namen *Verfahren von Crank-Nicolson* firmiert. Abbildung I.2.1 skizziert die zugehörigen Quadraturformeln.

Explizites Euler-Verfahren:

$$\begin{aligned} \eta_0 &= y_0, \\ \eta_{i+1} &= \eta_i + h_i f(t_i, \eta_i), \\ t_{i+1} &= t_i + h_i. \end{aligned}$$

Implizites Euler-Verfahren:

$$\begin{aligned} \eta_0 &= y_0, \\ \eta_{i+1} &= \eta_i + h_i f(t_{i+1}, \eta_{i+1}), \\ t_{i+1} &= t_i + h_i. \end{aligned}$$

Trapezregel, Verfahren von Crank-Nicolson:

$$\begin{aligned}\eta_0 &= y_0,\\ \eta_{i+1} &= \eta_i + \frac{h_i}{2}\Big[f(t_i,\eta_i) + f(t_{i+1},\eta_{i+1})\Big],\\ t_{i+1} &= t_i + h_i.\end{aligned}$$

Die drei genannten Verfahren sind spezielle Vertreter einer größeren Klasse von Verfahren, den *Runge-Kutta-Verfahren*. Diese haben folgende allgemeine Form:

Runge-Kutta-Verfahren:

$$\begin{aligned}\eta_0 &= y_0,\\ \eta_{i,j} &= \eta_i + h_i \sum_{k=1}^{r} a_{jk} f(t_i + c_k h, \eta_{i,k}) \quad \text{für } j = 1,\dots,r,\\ \eta_{i+1} &= \eta_i + h_i \sum_{k=1}^{r} b_k f(t_i + c_k h, \eta_{i,k}),\\ t_{i+1} &= t_i + h_i.\end{aligned}$$

Dabei ist $0 \le c_1 \le \ldots \le c_r \le 1$. Die Zahl r heißt *Stufe* des Runge-Kutta-Verfahrens.
Das Verfahren heißt *explizit*, wenn $a_{jk} = 0$ ist für alle $k \ge j$; es heißt *implizit*, wenn $a_{j,k} \neq 0$ ist für mindestens ein $k \ge j$.
Bei einem expliziten Verfahren können $\eta_{i,1},\dots,\eta_{i,r}$ sukzessive berechnet werden. Bei einem impliziten Verfahren erfordert die Berechnung von $\eta_{i,1},\dots,\eta_{i,r}$ das Lösen eines (i.a. nichtlinearen) Gleichungssystems mit $r \cdot d$ Gleichungen und Unbekannten.
Für die beiden Euler-Verfahren ist $r = 1$ und $c_1 = 0$, $a_{11} = 0$, $b_1 = 1$ für das explizite Euler-Verfahren und $c_1 = 1$, $a_{11} = 1$, $b_1 = 1$ für das implizite Euler-Verfahren. Für die Trapezregel ist $r = 2$ und $c_1 = 0$, $c_2 = 1$, $a_{11} = a_{12} = 0$, $a_{21} = a_{22} = \frac{1}{2}$, $b_1 = b_2 = \frac{1}{2}$.
Für die Praxis besonders wichtig sind die *stark diagonal impliziten Runge-Kutta-Verfahren*, kurz *SDIRK-Verfahren* (vgl. [6, §VI.4]). Bei ihnen ist die Matrix $(a_{ij})_{1\le i,j\le r}$ eine untere Dreiecksmatrix mit identischen Diagonalelementen. Die Berechnung von $\eta_{i,1},\dots,\eta_{i,r}$ erfordert das Lösen von r (i.a. nichtlinearen) Gleichungssystemen mit jeweils d Gleichungen und Unbekannten. Diese Verfahren haben besonders gute Genauigkeits- und Stabilitätseigenschaften.

Die Qualität eines Einschrittverfahrens wird durch seine *Ordnung* gemessen. Sie ist ein Maß für den Fehler eines *einzelnen* Schrittes des Verfahrens und wie folgt definiert:

Ein Einschrittverfahren hat die Ordnung $p > 0$, wenn gilt

$$|y(t_1) - \eta_1| = O(h_1^{p+1}).$$

Für den Fehler nach einer *beliebigen* Zahl von Schritten eines Einschrittverfahrens gilt:

Hat das Einschrittverfahren die Ordnung p und ist f bzgl. der Variablen y stetig differenzierbar mit beschränkter Ableitung, gilt für alle i

$$|y(t_i) - \eta_i| = O\Big(\big(\max_{1\le j\le i} h_j\big)^p\Big)$$

Die beiden Euler-Verfahren haben die Ordnung 1. Das Verfahren von Crank-Nicolson hat die Ordnung 2. Es gibt Runge-Kutta-Verfahren beliebig hoher Ordnung.

Die Ordnung eines Einschrittverfahrens beschreibt sein *asymptotisches* Verhalten für immer kleiner werdende Schrittweiten. In der Praxis rechnet man natürlich mit einer endlichen Schrittweite. Das Einschrittverfahren sollte dann für einen möglichst großen Bereich von Schrittweiten eine Näherungslösung liefern, die das gleiche *qualitative* Verhalten hat wie die exakte Lösung des Anfangswertproblems. Dieses Phänomen wird durch die *Stabilität* beschrieben (vgl. [6, §VI.5]).

Beispiel I.2.1. Für das Anfangswertproblem

$$y'(t) = -100y(t)$$
$$y(0) = 1$$

mit exakter Lösung $y(t) = e^{-100t}$ gilt:

- Das explizite Euler-Verfahren liefert nur dann eine abklingende Lösung, wenn $h_i \le \frac{1}{50}$ ist für alle i.
- Das implizite Euler-Verfahren und das Verfahren von Crank-Nicolson liefern für jede Schrittweite eine abklingende Lösung.

Explizite Verfahren können nicht stabil sein. Aber es gibt stabile implizite Runge-Kutta-Verfahren beliebig hoher Ordnung.

Die folgenden beiden Beispiele illustrieren den Effekt guter Stabilitätseigenschaften.

Beispiel I.2.2. Betrachte die gedämpfte Schwingung

$$y'(t) = \begin{pmatrix} -0.9 & -6.3 \\ 6.3 & -0.9 \end{pmatrix} y(t)$$
$$y(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

mit der exakten Lösung

$$y(t) = e^{-0.9t} \begin{pmatrix} \cos(6.3t) \\ \sin(6.3t) \end{pmatrix}.$$

Die Funktion $t \mapsto y(t)$ beschreibt eine sich im Ursprung zusammenziehende Spirale. Der linke Teil von Abbildung I.2.2 zeigt das Ergebnis der beiden Euler-Verfahren (rot und blau), der Trapezregel (grün), des klassischen Runge-Kutta-Verfahrens (gelb) [6, S. 89] und zweier SDIRK-Verfahren der Ordnung 3 und 4 (türkis und orange) für jeweils 100 Schritte mit konstanter Schrittweite $h = 0.1$. Das explizite Euler-Verfahren (rot) explodiert innerhalb weniger Schritte und verlässt den Ausschnitt der Abbildung, weil die Schrittweite zu groß ist. Um mit diesem Verfahren eine qualitativ richtige Lösung zu erhalten, müsste die Schrittweite um etwa den Faktor 10 kleiner sein. Das implizite Euler-Verfahren (blau) dämpft die Lösung zu stark. Dieser Effekt mindert sich für kleiner werdende Schrittweite, bleibt aber im Prinzip bestehen. Die anderen Verfahren liefern qualitativ akzeptable Lösungen.

Beispiel I.2.3. Betrachte die ungedämpfte Schwingung

$$y'(t) = \begin{pmatrix} 0 & -6.3 \\ 6.3 & 0 \end{pmatrix} y(t)$$
$$y(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

mit exakter Lösung

$$y(t) = \begin{pmatrix} \cos(6.3t) \\ \sin(6.3t) \end{pmatrix}.$$

Die Funktion $t \mapsto y(t)$ beschreibt einen Kreis mit Radius 1 um den Ursprung. Der rechte Teil von Abbildung I.2.2 zeigt das Ergebnis der beiden Euler-Verfahren (rot und blau), der Trapezregel (grün), des klassischen Runge-Kutta-Verfahrens (gelb) [6, S. 89] und zweier SDIRK-Verfahren der Ordnung 3 und 4 (türkis und orange) für jeweils 100 Schritte mit konstanter Schrittweite $h = 0.1$. Das explizite Euler-Verfahren (rot) explodiert innerhalb weniger Schritte und verlässt den Ausschnitt der Abbildung, weil die Schrittweite zu groß ist. Dieser Effekt bleibt auch bei kleiner werdender Schrittweite bestehen, wenn auch zunehmend abgeschwächt. Das implizite Euler-Verfahren (blau) dämpft die Lösung zu stark. Dieser Effekt mindert sich für kleiner werdende Schrittweite, bleibt aber im Prinzip bestehen. Das SDIRK-Verfahren der Ordnung 3 (türkis) dämpft die Lösung ebenfalls. Die anderen Verfahren liefern qualitativ akzeptable Lösungen.

I.3. Randwertprobleme

Gegeben sind ein Intervall I in $\mathbb{R}$, zwei verschiedene Punkte a und b in I, eine Teilmenge D des $\mathbb{R}^d$, eine Funktion $f(t, y) : I \times D \to \mathbb{R}^d$

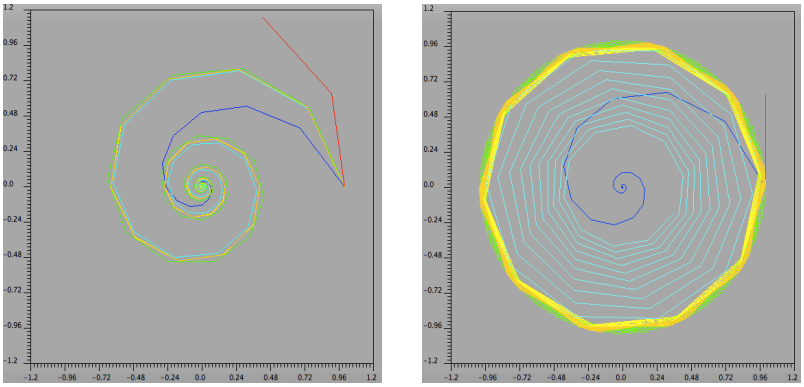


ABBILDUNG I.2.2. Jeweils 100 Schritte mit Schrittweite $h = 0.1$ des expliziten Euler-Verfahrens (rot), impliziten Euler-Verfahrens (blau), der Trapezregel (grün), des klassischen Runke-Kutta-Verfahrens (gelb), eines SDIRK-Verfahrens der Ordnung 3 (türkis) und eines SDIRK-Verfahrens der Ordnung 4 (orange) angewandt auf das Anfangswertproblem einer gedämpften Schwingung aus Beispiel I.2.2 (links) und einer ungedämpften Schwingung aus Beispiel I.2.3 (rechts). Das explizite Euler-Verfahren verlässt den Ausschnitt der Abbildung nach wenigen Schritten.

und eine Funktion $r(u, v) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$. Bei einem *Randwertproblem* ist eine differenzierbare Funktion $y : I \to D$ gesucht mit

$$\begin{aligned} y'(t) &= f(t, y(t)) \quad \text{für alle } t \in I \text{ } \textit{(Differentialgleichung)} \\ r(y(a), y(b)) &= 0 \qquad\qquad\qquad\qquad\qquad \textit{(Randbedingung)} \end{aligned}$$

Beispiel I.3.1. Das Randwertproblem

$$\begin{aligned} y'(t) &= \begin{pmatrix} \lambda & -\omega \\ \omega & \lambda \end{pmatrix} y(t) \\ y_1(0) &= 1 \\ y_1(\frac{\pi}{2\omega}) &= 0 \end{aligned}$$

für eine *gedämpfte Schwingung* entspricht den Daten

$$\begin{aligned} I = \mathbb{R}, \quad a = 0, \quad b = \frac{\pi}{2\omega}, \quad D = \mathbb{R}^2, \\ f(t, y) = \begin{pmatrix} \lambda & -\omega \\ \omega & \lambda \end{pmatrix} y, \\ r(u, v) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} u + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} v - \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \end{aligned}$$

Die exakte Lösung ist

$$y(t) = e^{\lambda t} \begin{pmatrix} \cos(\omega t) \\ \sin(\omega t) \end{pmatrix}.$$

Beispiel I.3.2. Randwertprobleme für Differentialgleichungen höherer Ordnung können analog zu Anfangswertproblemen durch Einführen neuer Unbekannter in Randwertprobleme für Differentialgleichungen erster Ordnung überführt werden. Für das *mechanische System*

$$\begin{aligned} Mx''(t) + Rx'(t) + Kx(t) &= F(t) \\ x(0) &= x_0 \\ x(L) &= x_L \end{aligned}$$

im $\mathbb{R}^d$ führt man z.B. die Funktion $v(t) = x'(t)$ als neue Unbekannte ein und erhält so ein Randwertproblem, das den Daten

$$\begin{aligned} I = \mathbb{R}, \quad a = 0, \quad b = L, \quad D &= \mathbb{R}^{2d}, \\ y(t) &= \begin{pmatrix} x(t) \\ v(t) \end{pmatrix}, \\ f(t,y) &= \begin{pmatrix} 0 \\ M^{-1}F(t) \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ -M^{-1}K & -M^{-1}R \end{pmatrix} y, \\ r(u,v) &= \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} u + \begin{pmatrix} 0 & 0 \\ I & 0 \end{pmatrix} v - \begin{pmatrix} x_0 \\ x_L \end{pmatrix} \end{aligned}$$

entspricht.

Manche Probleme, die auf den ersten Blick keine Randwertprobleme sind, können in ebensolche überführt werden. Hierzu zwei Beispiele:

Beispiel I.3.3. Gesucht seien eine Funktion $u : [a,b] \to \mathbb{R}$ und eine Zahl $\lambda \in \mathbb{R}$ mit

$$\begin{aligned} u'(t) &= g(t, u(t)), \\ \rho(u(a), u(b), \lambda) &= 0, \end{aligned}$$

wobei g und ρ gegebene Funktionen sind. Wenn man die Zahl λ als konstante Funktion interpretiert, entspricht dieses *Eigenwertproblem* einem Randwertproblem mit den Daten

$$\begin{aligned} I = \mathbb{R}, \quad D &= \mathbb{R}^2, \\ y(t) &= \begin{pmatrix} u(t) \\ \lambda \end{pmatrix}, \\ f(t,y) &= \begin{pmatrix} g(t, y_1) \\ 0 \end{pmatrix}, \\ r(u,v) &= \rho(u_1, v_1, v_2). \end{aligned}$$

Beispiel I.3.4. Gesucht seien eine Zahl $\beta > 0$ und eine Funktion $u : [0, \beta] \to \mathbb{R}$ mit

$$\begin{aligned} u'(s) &= g(s, u(s)), \\ \rho(u(0), u(\beta)) &= 0, \end{aligned}$$

wobei g und ρ wieder gegebene Funktionen sind. Dies ist ein *freies Randwertproblem*, da der rechte Randpunkt β Teil der Lösung ist. Wenn man die Zahl β als konstante Funktion interpretiert und in geschickter Weise eine neue Variable t einführt, entspricht dieses Problem einem Randwertproblem mit den Daten

$$\begin{aligned} I = \mathbb{R}, \quad a = 0, \quad b = 1, \quad D = \mathbb{R}^2, \\ y(t) &= \begin{pmatrix} u(t\beta) \\ \beta \end{pmatrix}, \\ t &= \frac{s}{y_2}, \\ f(t, y) &= \begin{pmatrix} y_2 g(t y_2, y_1) \\ 0 \end{pmatrix}, \\ r(u, v) &= \rho(u_1, v_1). \end{aligned}$$

Für Randwertprobleme gibt es keine allgemeine Existenz- und Eindeutigkeitsaussage wie für Anfangswertprobleme. Die Lösbarkeit und die Zahl allfälliger Lösungen hängt von dem konkreten Beispiel und dem Zusammenspiel von Differentialgleichung und Randbedingung ab. Dies illustriert das folgende Beispiel.

Beispiel I.3.5. Betrachte das Randwertproblem

$$\begin{aligned} y'(t) &= \begin{pmatrix} 0 & -\omega \\ \omega & 0 \end{pmatrix} y(t) \\ \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} y(0) + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} y(L) &= \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \end{aligned}$$

für eine *ungedämpfte Schwingung*. Die allgemeine Lösung der Differentialgleichung ist

$$y(t) = \begin{pmatrix} c_1 \cos(\omega t) - c_2 \sin(\omega t) \\ c_1 \sin(\omega t) + c_2 \cos(\omega t) \end{pmatrix}.$$

Die Daten $L = \frac{2\pi}{\omega}$, $\alpha = 0$, $\beta = 1$ führen auf die widersprüchlichen Bedingungen $c_1 = 0$ und $c_1 = 1$, so dass das entsprechende Randwertproblem keine Lösung hat. Die Daten $L = \frac{2\pi}{\omega}$, $\alpha = 0$, $\beta = 0$ hingegen führen auf die einzige Bedingung $c_1 = 0$, so dass c_2 beliebig ist und das Randwertproblem unendlich viele Lösungen hat.

I.4. Schießverfahren

Das *Schießverfahren* ist vielleicht das einfachste Verfahren zur Lösung von Randwertproblemen. Die zugrunde liegende Idee lässt sich wie folgt beschreiben:

- Bezeichne mit $y(t;s)$ die Lösung des *Anfangswertproblems*
$$\begin{aligned} y'(t) &= f(t, y(t)), \\ y(a;s) &= s. \end{aligned}$$
- Dann löst $y(t;s)$ das *Randwertproblem*
$$\begin{aligned} y'(t) &= f(t, y(t)), \\ r(y(a), y(b)) &= 0 \end{aligned}$$
genau dann, wenn gilt
$$r(s, y(b;s)) = 0.$$
- Bestimme mit dem *Newton-Verfahren* [6, §II.5] eine Nullstelle der Funktion
$$F(s) = r(s, y(b;s)).$$
- Die Ableitung $DF(s)$ von F an der Stelle s ist gegeben durch
$$DF(s) = D_u r(s, y(b;s)) + D_v r(s, y(b;s)) Z(b;s),$$
wobei Z das *Anfangswertproblem*
$$\begin{aligned} Z'(t;s) &= D_y f(t, y(t;s)) Z(t;s) \\ Z(a;s) &= I \end{aligned}$$
löst und I die Einheitsmatrix bezeichnet.
- Löse die beiden Anfangswertprobleme für $y(t;s)$ und $Z(t;s)$ näherungsweise mit einem numerischen Verfahren für Anfangswertprobleme, wobei beide Male die gleichen Gitterpunkte t_i benutzt werden.

Diese Idee führt auf den folgenden Algorithmus:

Algorithmus I.4.1. (Schießverfahren)

(0) *Gegeben sei ein Startwert $s^{(0)} \in \mathbb{R}^d$. Setze $i = 0$.*

(1) *Berechne mit einem numerischen Verfahren für Anfangswertprobleme eine Näherung $\eta^{(i)}(t)$ für die Lösung $y^{(i)}$ des Anfangswertproblems*
$$\begin{aligned} y^{(i)'}(t) &= f(t, y^{(i)}(t)), \\ y^{(i)}(a) &= s^{(i)}. \end{aligned}$$
Setze
$$F^{(i)} = r(s^{(i)}, \eta^{(i)}(b)).$$

(2) *Berechne mit dem gleichen Verfahren wie in Schritt* (1) *und den gleichen Gitterpunkten eine Näherung* $\zeta^{(i)}(t)$ *für die Lösung* $Z^{(i)}$ *des Anfangswertproblems*

$$Z^{(i)'}(t) = D_y f(t, \eta^{(i)}(t)) Z^{(i)}(t),$$
$$Z^{(i)}(a) = I.$$

Setze

$$D^{(i)} = D_u r(s^{(i)}, \eta^{(i)}(b)) + D_v r(s^{(i)}, \eta^{(i)}(b)) \zeta^{(i)}(b).$$

(3) *Löse das lineare Gleichungssystem*

$$D^{(i)} \Delta s^{(i)} = -F^{(i)}.$$

Setze

$$s^{(i+1)} = s^{(i)} + \Delta s^{(i)},$$

erhöhe i *um* 1 *und gehe zu Schritt* (1) *zurück.*

Das Schießverfahren hat folgende Eigenschaften:

Die Anfangswertprobleme in Schritt (1) haben d Unbekannte.
Die Anfangswertprobleme in Schritt (2) haben d^2 Unbekannte.
Die Anfangswertprobleme in Schritt (2) sind linear.
Die linearen Gleichungssysteme in Schritt (3) haben d Gleichungen und Unbekannte.
Das Newton-Verfahren sollte wie in [6, §II.6] beschrieben gedämpft werden.
Falls das Newton-Verfahren konvergiert, ist die Konvergenz quadratisch.

Selbst wenn das Randwertproblem eine eindeutige Lösung hat, kann das Schießverfahren völlig versagen. Dies illustriert das folgende Beispiel. Das katastrophale Verhalten liegt daran, dass Lösungen zu verschiedenen Anfangswerten ein und derselben Differentialgleichung exponentiell schnell auseinander laufen können. In diesem Sinne können Randwertprobleme schlecht konditioniert sein.

Beispiel I.4.2. Betrachte das Randwertproblem

$$y'(t) = \begin{pmatrix} 0 & 1 \\ 110 & 1 \end{pmatrix} y(t),$$
$$y_1(0) = 1,$$
$$y_1(10) = 1.$$

Die exakte Lösung ist

$$y(t) = c_1 e^{-10t} \begin{pmatrix} 1 \\ -10 \end{pmatrix} + c_2 e^{11t} \begin{pmatrix} 1 \\ 11 \end{pmatrix}$$

mit

$$c_1 = \frac{e^{110} - 1}{e^{110} - e^{-100}}, \quad c_2 = \frac{1 - e^{-100}}{e^{110} - e^{-100}}.$$

Die Lösung des Anfangswertproblems zum Anfangswert s ist

$$y(t; s) = \frac{11s_1 - s_2}{21} e^{-10t} \begin{pmatrix} 1 \\ -10 \end{pmatrix} + \frac{10s_1 + s_2}{21} e^{11t} \begin{pmatrix} 1 \\ 11 \end{pmatrix}.$$

Der korrekte Anfangswert für die Lösung des Randwertproblemes ist

$$s^* = \begin{pmatrix} 1 \\ -10 + 21 \cdot \frac{1-e^{-100}}{e^{110}-e^{-100}} \end{pmatrix}.$$

Der falsche Anfangswert

$$\widetilde{s} = \begin{pmatrix} 1 \\ -10 + 10^{-9} \end{pmatrix}$$

mit einem realtiven Fehler von 10^{-10} liefert den falschen Randwert

$$y_1(10; \widetilde{s}) \approx 10^{37}.$$

Man verliert also 47 Dezimalstellen!

I.5. Mehrzielmethode

Beispiel I.4.2 zeigt, dass das Schießverfahren versagen kann, weil Lösungen zu verschiedenen Anfangswerten exponentiell auseinander laufen können. Es zeigt aber auch, dass dieser Effekt vermieden werden kann, indem man Anfangswertprobleme nur auf kleinen Intervallen löst. Dies führt auf folgende Idee vom Typ „Teile und Herrsche“:

- Unterteile das Intervall $[a, b]$ durch Zwischenpunkte $a = \tau_1 < \tau_2 < \ldots < \tau_m = b$.
- Für $s_1, \ldots, s_m \in \mathbb{R}^d$ bezeichne mit $y(t; \tau_k, s_k)$ die Lösung des Anfangswertproblems
$$y'(t) = f(t, y(t)),$$
$$y(\tau_k; s_k) = s_k.$$
- Definiere die stückweise Funktion $\widetilde{y}$ durch
$$\widetilde{y}(t) = y(t; \tau_k, s_k) \text{ für } \tau_k \le t < \tau_{k+1},\ 1 \le k \le m-1,$$
$$\widetilde{y}(\tau_m) = s_m.$$
- Dann löst $\widetilde{y}$ das Randwertproblem
$$y'(t) = f(t, y(t))$$
$$r(y(a), y(b)) = 0$$
genau dann, wenn gilt
$$y(\tau_{k+1}; \tau_k, s_k) = s_{k+1} \quad \text{für } 1 \le k \le m-1,$$
$$r(s_1, s_m) = 0.$$

- Dies definiert ein Gleichungssystem

$$F(s_1, \ldots, s_m) = 0,$$

  das mit dem Newton-Verfahren gelöst werden kann.
- Die Berechnung der Ableitung von F erfordert das Lösen von Anfangswertproblemen auf den Intervallen $[\tau_k, \tau_{k+1}]$.

In jedem Newton-Schritt ist ein lineares Gleichungssystem der Form

$$DF\Delta s = -F$$

mit $m \cdot d$ Gleichungen und Unbekannten zu lösen. Da DF die Form

$$DF = \begin{pmatrix} G_1 & -I & & & \\ & G_2 & -I & & 0 \\ 0 & & \ddots & \ddots & \\ & & & G_{m-1} & -I \\ A & 0 & & 0 & B \end{pmatrix}$$

hat mit geeigneten $d \times d$ Matrizen G_1, ..., G_{m-1}, A und B, hat dieses Gleichungssystem die Form

$$\begin{aligned} G_1\Delta s_1 - \Delta s_2 &= -F_1 \\ G_2\Delta s_2 - \Delta s_3 &= -F_2 \\ \vdots &= \vdots \\ G_{m-1}\Delta s_{m-1} - \Delta s_m &= -F_{m-1} \\ A\Delta s_1 + B\Delta s_m &= -F_m. \end{aligned}$$

Bei diesem Gleichungssystem können die Unbekannten Δs_2, ..., Δs_m sukzessive eliminiert werden und man erhält das Gleichungssystem

$$(A + BG_{m-1} \ldots G_1)\Delta s_1 = -F_m - B\sum_{j=1}^{m-1}(\prod_{i=j+1}^{m-1} G_i)F_j$$

mit d Gleichungen für die d Komponenten von Δs_1.

Diese Ideen und Beobachtungen führen auf folgenden Algorithmus:

Algorithmus I.5.1. (Mehrzielmethode)

(0) *Gegeben seien m Punkte $a = \tau_1 < \ldots < \tau_m = b$ und m Vektoren $s_1^{(0)}, \ldots, s_m^{(0)} \in \mathbb{R}^d$. Setze $i = 0$.*

(1) *Bestimme mit einem numerischen Verfahren für Anfangswertprobleme Näherungen $\eta^{(i,j)}(t)$, $1 \le j \le m-1$, für die Lösungen $y^{(i,j)}$ der Anfangswertprobleme*

$$\begin{aligned} y^{(i,j)'}(t) &= f(t, y^{(i,j)}(t)), \\ y^{(i,j)}(\tau_j) &= s_j^{(i)} \end{aligned}$$

für $1 \le j \le m-1$. *Setze*

$$F_j^{(i)} = \eta^{(i,j)}(\tau_{j+1}) - s_{j+1}^{(i)} \quad \text{für } 1 \le j \le m-1,$$
$$F_m^{(i)} = r(s_1^{(i)}, s_m^{(i)}).$$

(2) *Bestimme mit dem gleichen Verfahren wie in Schritt* (1) *und den gleichen Gitterpunkten Näherungen* $\zeta^{(i,j)}(t)$ *für die Lösungen* $Z^{(i,j)}$ *der Anfangswertprobleme*

$$Z^{(i,j)'}(t) = D_y f(t, \eta^{(i,j)}(t)) Z^{(i,j)}(t)$$
$$Z^{(i,j)}(\tau_j) = I$$

für $1 \le j \le m-1$. *Setze*

$$G_j^{(i)} = \zeta^{(i,j)}(\tau_{j+1})$$

für $1 \le j \le m-1$ *und*

$$A^{(i)} = D_u r(s_1^{(i)}, s_m^{(i)}),$$
$$B^{(i)} = D_v r(s_1^{(i)}, s_m^{(i)}).$$

(3) *Berechne die Matrix*

$$H^{(i)} = A^{(i)} + B^{(i)} G_{m-1}^{(i)} \cdot \ldots \cdot G_1^{(i)}$$

und den Vektor

$$\varphi^{(i)} = -F_m^{(i)} - B^{(i)} \sum_{j=1}^{m-1} \Big(\prod_{l=j+1}^{m-1} G_l^{(i)} \Big) F_j^{(i)}.$$

Löse das lineare Gleichungssystem

$$H^{(i)} \Delta s_1^{(i)} = \varphi^{(i)}$$

und berechne rekursiv die Vektoren

$$\Delta s_{k+1}^{(i)} = G_k^{(i)} \Delta s_k^{(i)} + F_k^{(i)}$$

für $1 \le k \le m-1$. *Setze*

$$s_k^{(i+1)} = s_k^{(i)} + \Delta s_k^{(i)}$$

für $1 \le k \le m$, *erhöhe* i *um* 1 *und gehe zu Schritt* (1) *zurück.*

Die Mehrzielmethode hat folgende Eigenschaften:

Bei gleicher Zahl von Gitterpunkten auf dem gesamten Intervall $[a, b]$ erfordert die Lösung der Anfangswertprobleme beim Schießverfahren und bei der Mehrzielmethode den gleichen Aufwand.
Die Anfangswertprobleme auf den Teilintervallen können parallel gelöst werden.
Wenn keine zusätzlichen Informationen bekannt sind, können die Punkte $\tau_1, \ldots, \tau_m$ äquidistant gewählt werden.

I.6. Differenzenverfahren

In diesem und dem nächsten Abschnitt betrachten wir spezielle Randwertprobleme, sog. *Sturm-Liouville-Probleme*. Bei ihnen sind eine stetig differenzierbare Funktion $p : [0,1] \to \mathbb{R}$ mit

$$\underline{p} = \min_{0 \le x \le 1} p(x) > 0$$

und eine stetige Funktion $q : [0,1] \to \mathbb{R}$ mit

$$\underline{q} = \min_{0 \le x \le 1} q(x) > 0$$

gegeben. Gesucht ist eine zweimal stetig differenzierbare Funktion $u : [0,1] \to \mathbb{R}$ mit

$$\boxed{\begin{aligned} -(pu')' + qu &= f \quad \text{in } (0,1) \qquad \textit{(Differentialgleichung)} \\ u(0) = 0,\ u(1) &= 0 \qquad\qquad\qquad\qquad\quad \textit{(Randbedingung)} \end{aligned}}$$

Beispiel I.6.1. Häufig liegen Strum-Liouville-Probleme in der allgemeineren Form

$$\begin{aligned} -(pu')' + qu &= f \quad \text{in } (a,b) \\ u(a) = \alpha,\ u(b) &= \beta \end{aligned}$$

vor. Diese allgemeine Form kann wie folgt in die obige spezielle Form mit $a = 0$, $b = 1$, $\alpha = 0$, $\beta = 0$ transformiert werden: Suche u in der Form

$$u(x) = \alpha + \frac{\beta - \alpha}{b - a}(x - a) + v(x)$$

mit

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

und führe eine neue Variable durch

$$t = \frac{x - a}{b - a}$$

ein.

Die Differenzenverfahren dieses Abschnittes beruhen auf dem *symmetrischen Differenzenquotienten*:

$$\boxed{\partial_h \varphi(x) = \frac{1}{h}\big[\varphi(x + \frac{h}{2}) - \varphi(x - \frac{h}{2})\big].}$$

Taylor-Entwicklung liefert

$$\boxed{\partial_h \varphi(x) = \varphi'(x) + \frac{h^2}{24}\varphi'''(x + \theta h)}$$

mit einem geeigneten $\theta \in (-\frac{1}{2}, \frac{1}{2})$.

Die Idee der Differenzendiskretisierung lässt sich wie folgt beschreiben:

- Ersetze Ableitungen durch Differenzenquotienten ∂_h

$$\begin{aligned}
&\big(-(pu')'\big)(x)\\
&\approx \big(-\partial_h(pu')\big)(x)\\
&= \frac{1}{h}\big[p(x-\frac{h}{2})u'(x-\frac{h}{2}) - p(x+\frac{h}{2})u'(x+\frac{h}{2})\big]\\
&\approx \frac{1}{h}\big[p(x-\frac{h}{2})\partial_h u(x-\frac{h}{2}) - p(x+\frac{h}{2})\partial_h u(x+\frac{h}{2})\big]\\
&= \frac{1}{h^2}\big[p(x-\frac{h}{2})(u(x)-u(x-h))\\
&\qquad\qquad - p(x+\frac{h}{2})(u(x+h)-u(x))\big]
\end{aligned}$$

- Fordere die resultierenden Gleichungen nur in *Gitterpunkten* ih mit $h = \frac{1}{n+1}$ und $1 \le i \le n$.

Algorithmus I.6.2. (Differenzendiskretisierung)

(0) *Wähle eine Gitterweite* $h = \frac{1}{n+1}$.

(1) *Für* $1 \le i \le n$ *setze*

$$f_i = f(ih),\ q_i = q(ih),\ p_{i\pm\frac{1}{2}} = p(ih \pm \frac{h}{2}).$$

(2) *Bestimme* $u_0, \ldots, u_{n+1}$, *so dass*

$$u_0 = 0,\ u_{n+1} = 0$$

und

$$\begin{aligned}
f_i = &-\frac{1}{h^2}p_{i-\frac{1}{2}}u_{i-1} + \Big(\frac{1}{h^2}\big[p_{i-\frac{1}{2}} + p_{i+\frac{1}{2}}\big] + q_i\Big)u_i\\
&- \frac{1}{h^2}p_{i+\frac{1}{2}}u_{i+1}
\end{aligned}$$

ist für $1 \le i \le n$.

(3) *Bezeichne mit* u_h *die stetige, stückweise lineare Funktion, die an den Stellen* ih *mit* u_i *übereinstimmt (vgl. Abbildung I.6.1).*

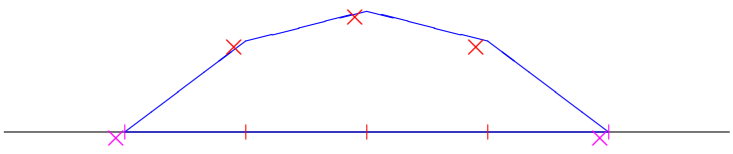


ABBILDUNG I.6.1. Stetige stückweise lineare Interpolation

Die Differenzendiskretisierung hat folgende Eigenschaften:

Die Differenzendiskretisierung führt auf ein lineares Gleichungssystem mit n Gleichungen für die n Unbekannten u_1, …, u_n.
Die Matrix des Gleichungssystems ist symmetrisch, positiv definit und tridiagonal mit positiven Diagonalelementen und negativen Nebendiagonalelementen.
Das Gleichungssystem hat eine eindeutige Lösung.
Die Lösung des Gleichungssystems mit dem Gaußschen Eliminationsverfahren [6, §I.2] oder der Cholesky-Zerlegung [6, §I.6] erfordert $O(n)$ Operationen.

Für die Differenzendiskretisierung gilt die folgende *a priori Fehlerabschätzung*:

Ist $\underline{q} > 0$, $\underline{p} > 0$, die Funktion p dreimal stetig differenzierbar und die Lösung u des Sturm-Liouville-Problems viermal stetig differenzierbar, gilt die *a priori Fehlerabschätzung*

$$\max_{0\le x\le 1} |u(x) - u_h(x)| \le ch^2.$$

Die Konstante c hängt von der unteren Schranke $\underline{q}$ für q, den Ableitungen bis zur Ordnung 3 von p und den Ableitungen bis zur Ordnung 4 von u ab.

I.7. Variationsmethoden

Die Voraussetzungen

- $\underline{q} > 0$,
- p dreimal stetig differenzierbar,
- u viermal stetig differenzierbar

des vorigen Abschnitte sind für viele praktische Probleme zu restriktiv. Diese Einschränkung werden bei den Variationsmethoden dieses Abschnittes umgangen. Außerdem erlauben sie eine a posteriori Fehlerkontrolle und eine optimale adaptive Anpassung der Diskretisierung. Schließlich bereiten die Techniken dieses Abschnittes die Finite-Element-Methoden für partielle Differentialgleichungen aus Kapitel IV vor.

Die Variationsmethoden beruhen auf einer geeigneten Variationsformulierung des Sturm-Liouville-Problems. Deren Grundidee lässt sich wie folgt beschreiben:

- Multipliziere die Differentialgleichung mit einer stetig differenzierbaren Funktion v mit $v(0) = 0$ und $v(1) = 0$:

  $$-(pu')'(x)v(x) + q(x)u(x)v(x) = f(x)v(x)$$

  für $0 \le x \le 1$.

- Integriere das Ergebnis von 0 bis 1:

$$\int_0^1 \big[-(pu')'(x)v(x) + q(x)u(x)v(x)\big]dx = \int_0^1 f(x)v(x)dx.$$

- Integriere den Ableitungsterm partiell:

$$\begin{aligned}
&-\int_0^1 (pu')'(x)v(x)dx \\
&= p(0)u'(0)\underbrace{v(0)}_{=0} - p(1)u'(1)\underbrace{v(1)}_{=0} + \int_0^1 p(x)u'(x)v'(x)dx \\
&= \int_0^1 p(x)u'(x)v'(x)dx.
\end{aligned}$$

Damit diese Idee auf einem verlässlichen mathematischen Fundament steht, müssen die Eigenschaften der Funktionen u und v präziser gefasst werden. Klassische Eigenschaften wie stetige Differenzierbarkeit sind hierfür zu restriktiv; der Begriff der Ableitung muss geeignet verallgemeinert werden. In Hinblick auf die Diskretisierung sollten insbesondere stückweise differenzierbare Funktionen im erweiterten Sinn differenzierbar sein.

Dies leistet der Begriff der *schwachen Ableitung*. Er ist durch folgende Beobachtung motiviert: Partielle Integration liefert für stetig differenzierbare Funktionen u und v mit $v(0) = 0$ und $v(1) = 0$

$$\begin{aligned}
\int_0^1 u'(x)v(x)dx &= u(1)\underbrace{v(1)}_{=0} - u(0)\underbrace{v(0)}_{=0} - \int_0^1 u(x)v'(x)dx \\
&= -\int_0^1 u(x)v'(x)dx.
\end{aligned}$$

Die Funktion u heißt *schwach differenzierbar* mit *schwacher Ableitung* w, wenn für jede stetig differenzierbare Funktion v mit $v(0) = 0$ und $v(1) = 0$ gilt

$$\int_0^1 w(x)v(x)dx = -\int_0^1 u(x)v'(x)dx.$$

Beispiel I.7.1. Jede stetig differenzierbare Funktion ist schwach differenzierbar und die schwache Ableitung stimmt mit der klassischen Ableitung überein.
Jede stetige, stückweise stetig differenzierbare Funktion ist schwach differenzierbar und die schwache Ableitung stimmt mit der stückweisen klassischen Ableitung überein.

Die Funktion $u(x) = 1 - |2x - 1|$ ist schwach differenzierbar mit schwacher Ableitung

$$w(x) = \begin{cases} 2 & \text{für } 0 < x < \frac{1}{2} \\ -2 & \text{für } \frac{1}{2} < x < 1 \end{cases}$$

(vgl. Abbildung I.7.1). Man beachte, dass der Wert $w(\frac{1}{2})$ beliebig ist.

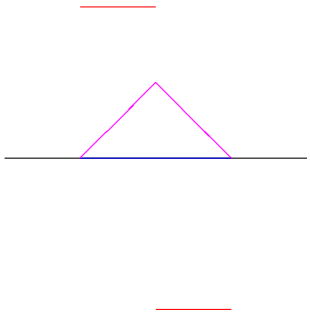


ABBILDUNG I.7.1. Funktion $u(x) = 1 - |2x - 1|$ (magenta) mit schwacher Ableitung (rot)

Die Variationsmethoden dieses Abschnittes, genauso wie die Finite-Element-Methoden aus Kapitel IV basieren ganz wesentlich auf den Sobolev-Räumen.

Die L^2-*Norm* ist definiert durch

$$\|v\| = \left\{ \int_0^1 |v(x)|^2 dx \right\}^{\frac{1}{2}}.$$

$L^2(0,1)$ ist der *Lebesgue-Raum* aller Funktionen v mit endlicher L^2-Norm $\|v\|$.
$H^1(0,1)$ ist der *Sobolev-Raum* aller Funktionen v in $L^2(0,1)$, deren schwache Ableitung existiert und ebenfalls in $L^2(0,1)$ ist.
$H_0^1(0,1)$ ist der *Sobolev-Raum* aller Funktionen v in $H^1(0,1)$ mit $v(0) = 0$ und $v(1) = 0$.

Beispiel I.7.2. Jede beschränkte Funktion ist in $L^2(0,1)$.
Die Funktion $v(x) = \frac{1}{\sqrt{x}}$ ist nicht in $L^2(0,1)$, da das Integral von $\frac{1}{x} = v(x)^2$ nicht endlich ist.
Jede stetig differenzierbare Funktion ist in $H^1(0,1)$.
Jede stetige, stückweise stetig differenzierbare Funktion ist in $H^1(0,1)$.
Die Funktion $v(x) = 1 - |2x - 1|$ ist in $H_0^1(0,1)$ (vgl. Abbildung I.7.1).
Die Funktion $v(x) = 2\sqrt{x}$ ist nicht in $H^1(0,1)$, da das Integral von $\frac{1}{x} = \big(v'(x)\big)^2$ nicht endlich ist.
Im Gegensatz zu mehrdimensionalen Problemen sind Funktionen in $H^1(0,1)$ *immer stetig.*

Mit diesen Begriffen können wir jetzt das zum Sturm-Liouville-Problem gehörende *Variationsproblem* mathematisch sauber formulieren:

Finde $u \in H_0^1(0,1)$ so, dass für alle $v \in H_0^1(0,1)$ gilt

$$\int_0^1 \big[p(x)u'(x)v'(x) + q(x)u(x)v(x)\big]dx = \int_0^1 f(x)v(x)dx.$$

Es hat folgende Eigenschaften:

Das Variationsproblem hat eine eindeutige Lösung.
Die Lösung des Variationsproblems ist das eindeutige *Minimum* in $H_0^1(0,1)$ der *Energiefunktion*

$$\frac{1}{2}\int_0^1 \big[p(x)u'(x)^2 + q(x)u(x)^2\big]dx - \int_0^1 f(x)u(x)dx.$$

Für die Diskretisierung des Variationsproblems benötigen wir die Finite-Element-Räume. Dazu bezeichne $\mathcal{T}$ eine beliebige *Unterteilung* des Intervalls $(0,1)$ in nicht überlappende Teilintervalle und $k \geq 1$ einen beliebigen Polynomgrad.

$S^{k,0}(\mathcal{T})$ ist der *Finite-Element-Raum* aller stetigen Funktionen, die stückweise auf den Intervallen von $\mathcal{T}$ Polynome vom Grad höchstens k sind.
$S_0^{k,0}(\mathcal{T})$ ist der *Finite-Element-Raum* aller Funktionen v in $S^{k,0}(\mathcal{T})$ mit $v(0) = 0$ und $v(1) = 0$.

Mit diesen Bezeichnungen wird das Variationsproblem durch das folgende *Finite-Element-Problem* diskretisiert:

Finde eine *Ansatzfunktion* $u_\mathcal{T} \in S_0^{k,0}(\mathcal{T})$ so, dass für alle *Testfunktionen* $v_\mathcal{T} \in S_0^{k,0}(\mathcal{T})$ gilt

$$\int_0^1 \big[p(x)u_\mathcal{T}'(x)v_\mathcal{T}'(x) + q(x)u_\mathcal{T}(x)v_\mathcal{T}(x)\big]dx = \int_0^1 f(x)v_\mathcal{T}(x)dx.$$

Das Finite-Element-Problem hat folgende Eigenschaften:

Das Finite-Element-Problem hat eine eindeutige Lösung.
Die Lösung des Finite-Element-Problems ist das eindeutige *Minimum* in $S_0^{k,0}(\mathcal{T})$ der *Energiefunktion*.

Nach Wahl einer Basis für $S_0^{k,0}(\mathcal{T})$ führt das Finite-Element-Problem auf ein lineares Gleichungssystem mit $k \cdot \sharp\mathcal{T} - 1$ Unbekannten und einer symmetrischen, positiv definiten, tridiagonalen Matrix, der sog. *Steifigkeitsmatrix*.
Meistens werden die Integrale durch Quadraturformeln [6, Kapitel IV] angenähert.
Meistens wird k gleich 1 (*lineare Elemente*) oder 2 (*quadratische Elemente*) gewählt.
In der Regel wird eine *nodale Basis* für $S_0^{k,0}(\mathcal{T})$ benutzt (vgl. Abbildung I.7.2).

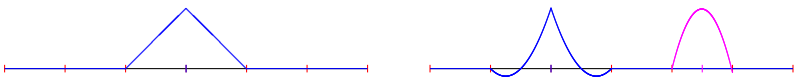


ABBILDUNG I.7.2. Nodale Basisfunktionen für lineare Elemente (links, blau) und quadratische Elemente (rechts, Intervall-Endpunkt blau und Intervall-Mittelpunkt magenta)

Die *nodalen Basisfunktionen* für lineare Elemente sind die Funktionen, die in genau einem Intervall-Endpunkt den Wert 1 annehmen und an allen anderen Intervall-Endpunkten den Wert 0 haben (linker Teil von Abbildung I.7.2).
Die *nodalen Basisfunktionen* für quadratische Elemente sind die Funktionen, die in genau einem Intervall-End- oder -Mittelpunkt den Wert 1 annehmen und an allen anderen Intervall-End- oder Mittelpunkten den Wert 0 haben (rechter Teil von Abbildung I.7.2, Intervall-Endpunkt blau und Intervall-Mittelpunkt magenta).

Bezeichnet man mit $h_\mathcal{T}$ die maximale Länge der Intervalle in $\mathcal{T}$, kann man folgende *a priori Fehlerabschätzung* für die Finite-Element-Diskretisierung des Sturm-Liouville-Porblems zeigen:

Für die Lösungen u des Variationsproblems und $u_\mathcal{T}$ des Finite-Element-Problems gelten die *a priori Fehlerabschätzungen*

$$\|u' - u'_\mathcal{T}\| \leq c_1 h_\mathcal{T},$$
$$\|u - u_\mathcal{T}\| \leq c_2 h_\mathcal{T}^2.$$

Die Konstanten c_1 und c_2 hängen von der unteren Schranke $\underline{p}$ für p, Ableitungen bis zur Ordnung 1 von p, dem Maximum von q und Ableitungen bis zur Ordnung 2 von u ab.

Die obige a priori Fehlerabschätzung hat einige Nachteile:

- Sie trifft nur eine Aussage über das *asymptotische Verhalten* des Fehlers für $h_{\mathcal{T}} \to 0$ d.h. für immer feiner werdende Unterteilungen.
- Sie gibt keine Information über die tatsächliche Größe des Fehlers.
- Sie erlaubt keine Information über die räumliche Verteilung des Fehlers, die benötigt wird, um zusätzliche Gitterpunkte dort zu platzieren, wo der Fehler groß ist.

Diese Nachteile werden durch *a posteriori Fehlerabschätzungen* und darauf beruhende *adaptive Gitterverfeinerungen* behoben (vgl. §V für partielle Differentialgleichungen). Im Gegensatz zu a priori Fehlerabschätzungen benötigen a posteriori Fehlerabschätzungen die explizite Kenntnis der berechneten Lösung des diskreten Problems. Aus diesen und den bekannten Daten der Differentialgleichung wird für jedes Teilintervall ein *Fehlerindikator* berechnet, der bis auf multiplikative Konstanten obere und untere Schranken für den Fehler der diskreten Lösung liefert.

Für jedes Intervall K in $\mathcal{T}$ bezeichne mit h_K seine Länge und definiere den *Fehlerindikator* η_K durch

$$\eta_K = h_K \Big\{ \int_K \big| f + (p u_{\mathcal{T}}')' - q u_{\mathcal{T}} \big|^2 \Big\}^{\frac{1}{2}}.$$

Dann gelten die *a posteriori Fehlerabschätzungen*

$$\|u' - u_{\mathcal{T}}'\| \le \big(\underline{p}\big)^{-1} \Big\{ \sum_{K \in \mathcal{T}} \eta_K^2 \Big\}^{\frac{1}{2}}$$

und für jedes Intervall K in $\mathcal{T}$

$$\eta_K \le c \Big\{ \int_K \big| u' - u_{\mathcal{T}}' \big|^2 \Big\}^{\frac{1}{2}}.$$

Obige a posteriori Fehlerabschätzung hat folgende Eigenschaften:

Die Größe $f + (p u_{\mathcal{T}}')' - q u_{\mathcal{T}}$ ist das intervallweise *Residuum* der Finite-Element-Lösung bzgl. der Differentialgleichung.
Der Aufwand zur Berechnung der Fehlerindikatoren ist vernachlässigbar.
Die a posteriori Fehlerabschätzung gibt zuverlässige Informationen über die tatsächliche Größe des Fehlers und seine räumliche Verteilung.
Die a posteriori Fehlerabschätzung kann für eine *adaptive Gitterverfeinerung* (vgl. Algorithmus I.7.3) genutzt werden, so dass jede vorgegebene Genauigkeit mit (nahezu) minimaler Anzahl an Unbekannten erreicht werden kann.

Die obere Fehlerschranke ist *global*. Dies liegt daran, dass eine lokale Änderung der rechten Seite des Sturm-Liouville-Problems zu einer globalen Änderung der Lösung führt, d.h. lokale Lasten entsprechen globalen Verschiebungen.
Die untere Fehlerschranke ist *lokal*. Dies liegt daran, dass eine lokale Änderung der Lösung des Sturm-Liouville-Problems zu einer lokalen Änderung der rechten Seite führt, d.h. lokale Verschiebungen entsprechen lokalen Lasten.

Algorithmus I.7.3. (Adaptive Gitterverfeinerung)

(0) *Gegeben: eine Toleranz* ε.
Gesucht: eine Finite-Element-Lösung mit Fehler $\leq \varepsilon$.
(1) *Wähle eine grobe Unterteilung* $\mathcal{T}_0$ *und setze* $k = 0$.
(2) *Löse das Finite-Element-Problem zu* $\mathcal{T}_k$.
(3) *Für jedes Intervall* K *in* $\mathcal{T}_k$ *bestimme den Fehlerindikator* η_K *und den Maximalwert* $\eta_k = \max_{K \in \mathcal{T}_k} \eta_K$.
(4) *Falls* $\eta_k \leq \varepsilon$ *ist, stopp.*
(5) *Für jedes* K *in* $\mathcal{T}_k$ *prüfe, ob* $\eta_K \geq \frac{1}{2}\eta_k$ *ist. Falls dies der Fall ist, halbiere* K*, sonst lasse* K *unverändert. Dies bestimmt die neue Unterteilung* $\mathcal{T}_{k+1}$*. Erhöhe* k *um* 1 *und gehe zu Schritt* (2) *zurück.*

KAPITEL II

Partielle Differentialgleichungen

In diesem Kapitel erinnern wir kurz an einige wichtige partielle Differentialgleichungen, ihre Eigenschaften und Klassifizierung. Eine ausführlichere Darstellung und Techniken zur geschlossenen Lösung spezieller partieller Differentialgleichungen finden sich z.B. in [5, Kapitel XIV].

II.1. Beispiele

Beispiel II.1.1 (Poisson-Gleichung)**.** Die vertikale *Auslenkung* $u : \Omega \to \mathbb{R}$ einer elastischen, nicht dehnbaren *Membran* mit Querschnittsfläche Ω in der (x, y)-Ebene unter Einfluss einer vertikalen Last $f : \Omega \to \mathbb{R}$ wird durch die *Membran-* oder *Poisson-Gleichung*

$$-\Delta u = -\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f \quad \text{in } \Omega$$

beschrieben. Auf dem Rand Γ von Ω muss die Auslenkung eine der folgenden *Randbedingungen* erfüllen:

- *Dirichlet-Randbedingung*: $u = 0$ für eine eingespannte Membran,
- *Neumann-Randbedingung*: $\mathbf{n} \cdot \nabla u = \frac{\partial u}{\partial n} = 0$ für eine frei bewegliche Membran.

Die Randbedingungen können in dem Sinne gemischt werden, dass auf einem Teil des Randes die Dirichlet- und auf einem anderen, dazu disjunkten Teil die Neumann-Bedingung gelten soll (vgl. Abbildung II.1.1). In jedem Punkt des Randes muss genau eine der Bedingungen gefordert werden. Die rechten Seiten 0 der Randbedingungen können durch beliebige gegebene Funktionen ersetzt werden. Die Auslenkung u minimiert die *Energiefunktion*

$$\frac{1}{2}\int_\Omega |\nabla u|^2 dx - \int_\Omega f u dx$$

in einer geeigneten Menge zulässiger Auslenkungen.

Beispiel II.1.2 (Biharmonische Gleichung)**.** Die vertikale *Auslenkung* $u : \Omega \to \mathbb{R}$ der Mittelebene Ω einer dünnen, starren *Platte* unter Einfluss einer vertikalen Last $f : \Omega \to \mathbb{R}$ wird durch die *Platten-* oder

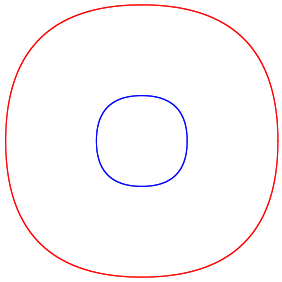


ABBILDUNG II.1.1. Ringförmiges Gebiet mit disjunkten Randkomponenten (rot und blau)

biharmonische Gleichung

$$\Delta^2 u = \frac{\partial^4 u}{\partial x^4} + 2\frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4} = f \quad \text{in } \Omega$$

beschrieben. Auf dem Rand Γ von Ω muss die Auslenkung eine der folgenden *Randbedingungen* erfüllen:

- $u = 0$ *und* $\frac{\partial u}{\partial n} = 0$ für eine eingespannte Platte,
- $u = 0$ *und* $\Delta u = 0$ für eine frei bewegliche Platte.

Die Randbedingungen können in dem Sinne gemischt werden, dass auf einem Teil des Randes die Platte fest und auf einem anderen, dazu disjunkten Teil frei sein soll (vgl. Abbildung II.1.1). In jedem Punkt des Randes muss genau eine der Bedingungen gefordert werden. Die rechten Seiten 0 der Randbedingungen können durch beliebige gegebene Funktionen ersetzt werden. Die Auslenkung u minimiert die *Energiefunktion*

$$\frac{1}{2}\int_\Omega |\Delta u|^2 dx - \int_\Omega f u dx$$

in einer geeigneten Menge zulässiger Auslenkungen.

Beispiel II.1.3 (Lineare Elastizitätstheorie)**.** Die *Verformung* $\mathbf{u} : \Omega \to \mathbb{R}^3$ eines Körpers $\Omega \subset \mathbb{R}^3$ unter dem Einfluss äußerer Kräfte $\mathbf{f} : \Omega \to \mathbb{R}^3$ wird durch die Gleichungen der *linearen Elastizitätstheorie*

$$-\operatorname{div} \sigma = \mathbf{f} \quad \text{in } \Omega$$

beschrieben. Dabei ist

$$\operatorname{div} \sigma = \begin{pmatrix} \frac{\partial \sigma_{11}}{\partial x} + \frac{\partial \sigma_{21}}{\partial y} + \frac{\partial \sigma_{31}}{\partial z} \\ \frac{\partial \sigma_{12}}{\partial x} + \frac{\partial \sigma_{22}}{\partial y} + \frac{\partial \sigma_{32}}{\partial z} \\ \frac{\partial \sigma_{13}}{\partial x} + \frac{\partial \sigma_{23}}{\partial y} + \frac{\partial \sigma_{33}}{\partial z} \end{pmatrix},$$

$$\sigma = C\varepsilon,$$

$$\varepsilon = \frac{1}{2}(\nabla \mathbf{u} + (\nabla \mathbf{u})^t)$$

$$= \begin{pmatrix} \frac{\partial u_1}{\partial x} & \frac{1}{2}\frac{\partial u_1}{\partial y} + \frac{1}{2}\frac{\partial u_2}{\partial x} & \frac{1}{2}\frac{\partial u_1}{\partial z} + \frac{1}{2}\frac{\partial u_3}{\partial x} \\ \frac{1}{2}\frac{\partial u_1}{\partial y} + \frac{1}{2}\frac{\partial u_2}{\partial x} & \frac{\partial u_2}{\partial y} & \frac{1}{2}\frac{\partial u_2}{\partial z} + \frac{1}{2}\frac{\partial u_3}{\partial y} \\ \frac{1}{2}\frac{\partial u_1}{\partial z} + \frac{1}{2}\frac{\partial u_3}{\partial x} & \frac{1}{2}\frac{\partial u_3}{\partial y} + \frac{1}{2}\frac{\partial u_2}{\partial z} & \frac{\partial u_3}{\partial z} \end{pmatrix}.$$

Die Matrix ε heißt *Verzerrung* (engl. *strain*), σ *Spannung* (engl. *stress*). Die Matrix C beschreibt das *Materialgesetz*. Ein wichtiger Spezialfall ist

$$\sigma = 2\mu\varepsilon + \lambda(\varepsilon_{11} + \varepsilon_{22} + \varepsilon_{33})I$$

mit den sog. *Lamé-Parametern* λ und μ. Auf dem Rand Γ von Ω muss eine der *Randbedingungen* $\mathbf{u} = 0$ oder $\mathbf{n} \cdot \sigma = 0$ gefordert werden. Die Poisson- und biharmonische Gleichung entstehen aus den Gleichungen der Elastizitätstheorie durch zusätzliche vereinfachende Modellannahmen. Die Verschiebung $\mathbf{u}$ minimiert die *Energiefunktion*

$$\frac{1}{2}\int_\Omega \varepsilon : \sigma dx - \int_\Omega \mathbf{f} \cdot \mathbf{u} dx$$

in einer geeigneten Menge zulässiger Verformungen.

Beispiel II.1.4 (Minimalflächen)**.** Zu einer gegebenen Menge Ω in der (x, y)-Ebene und einer gegebenen Funktion $u_0 : \Gamma \to \mathbb{R}$ auf dem Rand Γ von Ω wird eine Fläche der Form

$$S = \{(x, y, u(x, y)) : (x, y) \in \Omega,\ u(x, y) = u_0(x, y) \text{ auf } \Gamma\}$$

mit *minimalem Flächeninhalt* gesucht. Dann ist $u : \Omega \to \mathbb{R}$ Lösung der *Minimalflächengleichung*

$$\begin{aligned} -\operatorname{div}\big(\{1 + |\nabla u|^2\}^{-\frac{1}{2}} \nabla u\big) = 0 \quad &\text{in } \Omega, \\ u = u_0 \quad &\text{auf } \Gamma. \end{aligned}$$

Beispiel II.1.5 (Gasgleichung)**.** Betrachte die rotationsfreie Strömung eines idealen, kompressiblen Gases. Dann gibt es ein skalares *Potential* u, so dass für die Geschwindigkeit $\mathbf{v}$ des Gases gilt $\mathbf{v} = \nabla u$. Aus der *Massenerhaltung* folgt $\operatorname{div}(\rho\mathbf{v}) = 0$, wobei $\rho = \rho(\mathbf{v})$ die *Dichte* des Gases ist. Da das Gas ideal ist, gilt die *Zustandsgleichung*

$$\rho(\mathbf{v}) = \big[1 - \frac{\gamma - 1}{2}|\mathbf{v}|^2\big]^{\frac{1}{\gamma - 1}},$$

wobei $\gamma > 1$ der *spezifische Wärmekoeffizient* ist. Daher erfüllt das Potential u die *Gasgleichung*

$$\begin{aligned} -\operatorname{div}\Big(\big[1 - \frac{\gamma - 1}{2}|\nabla u|^2\big]^{\frac{1}{\gamma - 1}} \nabla u\Big) = 0 \quad &\text{in } \Omega, \\ u = u_0 \quad &\text{auf } \Gamma. \end{aligned}$$

Beispiel II.1.6 (Wärmeleitungsgleichung)**.** Die *Temperatur* $u(x, t)$ im Punkt x eines Körpers $\Omega \subset \mathbb{R}^3$ zur Zeit $t > 0$ wird unter dem Einfluss einer äußeren Wärmequelle $f(x, t)$ durch die *Wärmeleitungsgleichung*

$$\frac{\partial u}{\partial t} - \Delta u = f \quad \text{in } \Omega \times (0, \infty)$$

beschrieben. Die anfängliche Temperaturverteilung wird beschrieben durch die *Anfangsbedingung*

$$u(x,0) = u_0(x) \quad \text{für alle } x \in \Omega.$$

Zusätzlich ist auf dem Rand Γ des Körpers eine der folgenden *Randbedingungen* zu erfüllen:

- $u(x,t) = 0$ für alle $x \in \Gamma$ und $t > 0$ *(feste Temperatur)*,
- $\frac{\partial}{\partial n} u(x,t) = 0$ für alle $x \in \Gamma$ und $t > 0$ *(Isolation)*.

Die Randbedingungen können wie üblich gemischt werden. Die rechten Seiten 0 in den Randbedingungen können durch gegebene Funktionen ersetzt werden. Die *Energie*

$$\frac{1}{2} \int_\Omega u(x,t)^2 dx$$

ist eine monoton fallende Funktion der Zeit.

Beispiel II.1.7 (Grundwasserströmung)**.** Die räumliche und zeitliche Verteilung $u(x,t)$ einer Flüssigkeit wie Grundwasser in einem Medium wie Erde wird durch die *Transport-Diffusions-Gleichung*

$$\frac{\partial u}{\partial t} - \operatorname{div}\big(D(x,u)\nabla u\big) + \mathbf{k}(x,u) \cdot \nabla u = f \quad \text{in } \Omega \times (0,\infty)$$

beschrieben. Dabei modelliert der Quellterm f die Zufuhr (Quelle) bzw. Entnahme (Brunnen) von Flüssigkeit. Die *Diffusivität* $D(x,u) \in \mathbb{R}^{3\times 3}$ und die *Konduktivität* $\mathbf{k}(x,u) \in \mathbb{R}^3$ beschreiben spezifische Eigenschaften des Mediums (Ton, Lehm, Sand usw.). Wie bei der Wärmeleitungsgleichung ist die Transport-Diffusions-Gleichung durch Anfangs- und Randbedingungen zu ergänzen. Falls der Quellterm ab einem gewissen Zeitpunkt zeitlich konstant ist, strebt die Lösung u der Transport-Diffusions-Gleichung für $t \to \infty$ gegen einen zeitlich konstanten *stationären Zustand* v. Dieser wird durch die *Konvektions-Diffusions-Gleichung*

$$-\operatorname{div}\big(D(x,v)\nabla v\big) + \mathbf{k}(x,v) \cdot \nabla v = f \quad \text{in } \Omega$$

beschrieben. Diese ist wie die Poisson-Gleichung durch Randbedingungen zu ergänzen.

Beispiel II.1.8 (Wellen-Gleichung)**.** Die zeitlich veränderliche vertikale *Auslenkung* u einer elastischen, nicht dehnbaren *Membran* unter Einfluss einer vertikalen, zeitlich veränderlichen Last f wird durch die *Wellen-Gleichung*

$$\frac{\partial^2 u}{\partial t^2} - \Delta u = f \quad \text{in } \Omega \times (0,\infty)$$

beschrieben. Der Anfangszustand wird durch die *zwei Anfangsbedingungen*

$$u(x,0) = u_0(x)$$
$$\frac{\partial}{\partial t}u(x,0) = u_1(x)$$

für alle x in Ω festgelegt. Die Wellengleichung ist durch Randbedingungen zu ergänzen. Die *Energie*

$$\frac{1}{2}\int_\Omega \Big\{ \big(\frac{\partial u}{\partial t}\big)^2 + |\nabla u|^2 \Big\} dx$$

ist zeitlich konstant.

II.2. Typen

Partielle Differentialgleichungen werden unter verschiedenen Aspekten klassifiziert:

- Ordnung,
- linear oder nichtlinear,
- elliptisch, parabolisch oder hyperblisch.

> Die *Ordnung* einer partiellen Differentialgleichung ist die höchste Differentiationsstufe der in der Gleichung auftretenden partiellen Ableitungen.

Die biharmonische Gleichung hat die Ordnung 4. Alle anderen Beispiele haben die Ordnung 2. Bei einer Differentialgleichung der Ordnung $2k$ sind typischerweise k Randbedingungen zu fordern.

> Eine partielle Differentialgleichung heißt *linear*, wenn eine Überlagerung oder Skalierung der Last zu einer entsprechenden Überlagerung bzw. Skalierung der Lösung führt, d.h. die Zuordnung „Last $\to$ Lösung“ ist eine lineare Funktion.

Die Poisson-Gleichung, die biharmonische Gleichung, die Gleichungen der linearen Elastizitätstheorie, die Wärmeleitungsgleichung und die Wellengleichung sind linear, alle anderen Beispiele sind nichtlinear. Die Transport-Diffusions- und die Konvektions-Diffusions-Gleichung sind linear, wenn die Diffusivität und die Konduktivität nicht von der Lösung u abhängen.

Die allgemeine Form einer linearen partiellen Differentialgleichung zweiter Ordnung ist

$$A(x) : D^2u + \mathbf{a}(x) \cdot \nabla u + \alpha(x)u = f.$$

Wegen der Symmetrie der *Hesse-Matrix* D^2u kann die Matrix $A(x)$ als symmetrisch vorausgesetzt werden.

Eine lineare partielle Differentialgleichung zweiter Ordnung heißt
elliptisch, wenn für alle x alle Eigenwerte von $A(x)$ ungleich Null sind und gleiches Vorzeichen haben,
parabolisch, wenn für alle x genau ein Eigenwert von $A(x)$ gleich Null ist und alle anderen Eigenwerte gleiches Vorzeichen haben,
hyperbolisch, wenn für alle x alle Eigenwerte von $A(x)$ ungleich Null sind und genau ein Eigenwert anderes Vorzeichen hat als die restlichen Eigenwerte.

Die Wellengleichung ist hyperbolisch. Die Wärmeleitungsgleichung und die Transport-Diffusions-Gleichung sind parabolisch. Die Gasgleichung ist elliptisch, falls der Gradient des Potentials hinreichend klein ist (*Unterschallströmung*). Alle anderen Beispiele sind elliptisch. Elliptische Gleichungen beschreiben häufig ein Variations- oder Minimierungsproblem. Parabolische Gleichungen beschreiben häufig ein Dissipationsphänomen, bei dem eine Energie monoton fällt. Hyperbolische Gleichungen beschreiben häufig einen Erhaltungssatz.

II.3. Lösungseigenschaften

Die Lösungseigenschaften partieller Differentialgleichungen sind wesentlich vielfältiger und komplexer als diejenigen gewöhnlicher Differentialgleichungen.

Lineare Differentialgleichungen haben in der Regel eine eindeutige Lösung.
Nichtlineare Differentialgleichungen können mehrere Lösungen zulassen, insbesondere können *Verzweigungen* und *Umkehrpunkte* auftreten (vgl. Abbildung II.3.1).
Anders als bei gewöhnlichen Differentialgleichungen hängt die *Regularität*, d.h Differenzierbarkeit, der Lösung einer partiellen Differentialgleichung von Eigenschaften des Randes Γ ab. *Einspringende Ecken* (engl. *re-entrant corners*) führen zu einem Regularitätsverlust.

Beispiel II.3.1. Betrachte die Poisson-Gleichung $-\Delta u = 0$ auf dem Kreissegment (vgl. Abbildung II.3.2)

$$\Omega = \{(r\cos\varphi, r\sin\varphi) : 0 \le r < 1,\ 0 \le \varphi \le \frac{3\pi}{2}\}$$

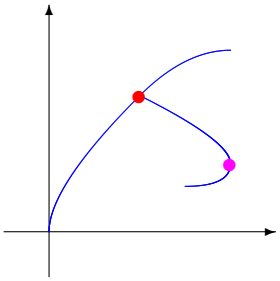


ABBILDUNG II.3.1. Verzweigung (rot) und Umkehrpunkt (magenta)

mit Randbedingung $u = \sin(\frac{2}{3}\varphi)$ auf dem Kreisbogen und $u = 0$ auf den geraden Randstücken. Die Lösung ist

$$u = r^{\frac{2}{3}} \sin(\frac{2}{3}\varphi).$$

Sie hat in der Nähe der einspringenden Ecke $(0, 0)$ keine beschränkte Ableitung.

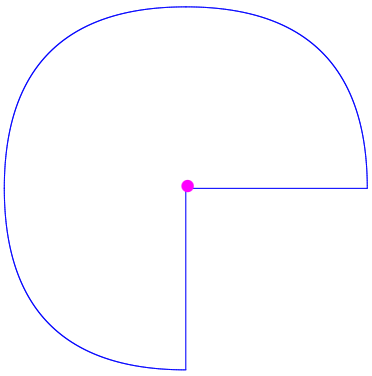


ABBILDUNG II.3.2. Kreissegment mit einspringender Ecke (magenta) im Ursprung

II.4. Überblick über Diskretisierungsmethoden

In den folgenden Kapiteln betrachten wir verschiedene Diskretisierungsmethoden für partielle Differentialgleichungen:

- *Differenzenverfahren* ersetzen Ableitungen durch Differenzenquotienten und fordern die resultierenden Gleichungen nur in den Punkten eines regelmäßigen Gitters.
- *Finite-Element-Methoden* basieren auf einer Variationsformulierung der Differentialgleichung und approximieren die dabei auftretenden Funktionen durch stetige, stückweise Polynome auf einer Unterteilung des Gebietes Ω in einfache Teilgebiete wie Dreiecke oder Vierecke.

- *Finite-Volumen-Methoden* basieren auf einer Erhaltungsgleichung und erfüllen diese für stückweise konstante Funktionen auf einfachen Teilgebieten von Ω, den sog. Kontrollvolumina.

KAPITEL III

Differenzenverfahren für partielle Differentialgleichungen

III.1. Elliptische Differentialgleichungen

In diesem Abschnitt betrachten wir die *Reaktions-Diffusions-Gleichung*

$$\begin{aligned} -\operatorname{div}(A\nabla u) + \alpha u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{auf } \Gamma \end{aligned}$$

als ein Modellproblem für lineare, elliptische Differentialgleichungen zweiter Ordnung. Dabei ist Ω ein Teilmenge des $\mathbb{R}^d$ mit $d = 2$ oder $d = 3$, $A(x)$ für jedes $x \in \Omega$ eine symmetrische, positiv definite, $d \times d$ Matrix und $\alpha(x)$ für jedes $x \in \Omega$ eine nicht-negative Zahl. Physikalisch beschreiben A eine örtlich veränderliche Diffusion und α eine örtlich veränderliche Reaktion.

Die Idee der Differenzenverfahren kann wie folgt beschrieben werden:

- Ersetze alle partiellen Ableitungen durch *Differenzenquotienten.*
- Fordere die resultierenden Gleichungen nur für die Punkte eines regelmäßigen *Gitters.*

Für die Konstruktion des Gitters wählen wir eine *Gitterweite* $h > 0$ und setzen (vgl. Abbildung III.1.1)

$$\begin{aligned} &G_h = \{\mathbf{i}h : \mathbf{i} \in \mathbb{Z}^d\}, \\ &\overline{\Omega}_h = G_h \cap \overline{\Omega}, \\ &\Gamma_h = \{x \in \overline{\Omega}_h : \min_{y\in\Gamma} |x - y| < h\}, \quad \textit{(diskreter Rand)} \\ &\Omega_h = \overline{\Omega}_h \setminus \Gamma_h. \qquad\qquad\qquad\quad \textit{(diskretes Gebiet)} \end{aligned}$$

Für die Nummerierung der Gitterpunkte bezeichnen wir mit

$$\begin{aligned} N_h &= \sharp\Omega_h \quad \text{die Zahl der Punkte im diskreten Gebiet,} \\ \overline{N}_h &= \sharp\overline{\Omega}_h \quad \text{die Zahl aller Gitterpunkte,} \\ \overline{N}_h - N_h &= \sharp\Gamma_h \quad \text{die Zahl der diskreten Randpunkte.} \end{aligned}$$

Zuerst nummerieren wir die Punkte in Ω_h *lexikographisch*, d.h. zeilenweise von links nach rechts beginnend mit der obersten Zeile. Danach

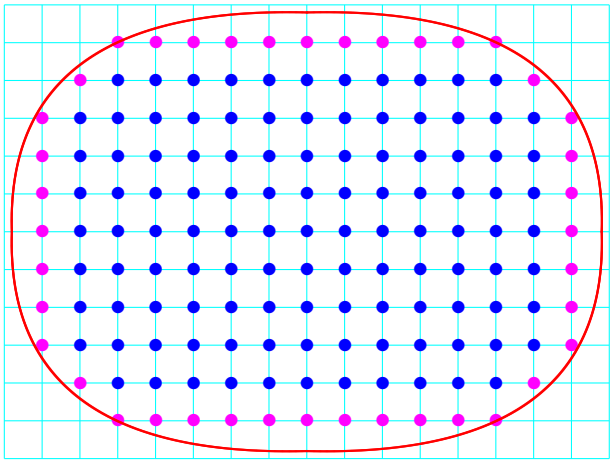


ABBILDUNG III.1.1. Gitter G_h (türkis), Rand Γ (rot), diskreter Rand Γ_h (magenta), diskretes Gebiet Ω_h (blau)

nummerieren wir die Punkte in Γ_h fortlaufend. Es ist $N_h \approx h^{-d}$ und $\overline{N}_h - N_h \approx h^{-(d-1)} \approx N_h^{\frac{d-1}{d}}$.

Für die Diskretisierung benötigen wir die Differenzenquotienten. Dazu bezeichnen wir mit e_i den i-ten Einheitsvektor, dessen Komponenten alle Null sind mit Ausnahme der i-ten Komponente, die 1 ist, und setzen

$$\partial^+_{h,i} u(x) = \frac{1}{h}\big[u(x + he_i) - u(x)\big]$$
(vorwärts Differenzenquotient)

$$\partial^-_{h,i} u(x) = \frac{1}{h}\big[u(x) - u(x - he_i)\big]$$
(rückwärts Differenzenquotient)

Taylor-Entwicklung liefert mit geeignetem $\theta \in (0,1)$:

$$\partial^\pm_{h,i} u(x) = \frac{\partial}{\partial x_i} u(x) \pm \frac{1}{2} h \frac{\partial^2}{\partial x_i^2} u(x \pm \theta h e_i)$$

$$\begin{aligned} \partial^+_{h,i}(\partial^-_{h,i} u)(x) &= \partial^-_{h,i}(\partial^+_{h,i} u)(x) \\ &= \frac{\partial^2}{\partial x_i^2} u(x) + \frac{1}{12} h^2 \frac{\partial^4}{\partial x_i^4} u(x \pm \theta h e_i) \end{aligned}$$

Damit lautet die *Differenzendiskretisierung* der Reaktions-Diffusions-Gleichung:

Bestimme den Vektor $u_h = (u_h(x))_{x \in \overline{\Omega}_h}$ so, dass gilt:

$$u_h(x) = 0$$

für alle Punkte x auf dem diskreten Rand Γ_h und

$$-\sum_{i=1}^{d}\sum_{j=1}^{d}\partial_{h,i}^{-}\big(A_{i,j}\partial_{h,j}^{+}u\big)(x)+\alpha(x)u(x)=f(x)$$

für alle Punkte x im diskreten Gebiet Ω_h.

Die Differenzendiskretisierung hat folgende Eigenschaften:

Sie führt auf ein lineares Gleichungssystem mit N_h Gleichungen für die N_h Unbekannten $u_h(x)$, $x\in\Omega_h$.
Die Matrix L_h des Gleichungssystems ist symmetrisch, positiv definit.
Die Matrix ist *dünn besetzt*, pro Zeile sind höchstens 3^d Elemente von Null verschieden.
Die Diagonalelemente sind positiv.
Die Elemente außerhalb der Diagonalen verschwinden oder sind negativ.
Die Matrix hat *Bandstruktur*, die *Bandbreite* ist $\approx N_h^{1-\frac{1}{d}}$.
Die Lösung des Gleichungssystems mit dem Gaußschen Eliminationsverfahren oder der Cholesky-Zerlegung erfordert $\approx N_h^{3-\frac{2}{d}}$ Operationen und $\approx N_h^{2-\frac{1}{d}}$ Speicherplätze (vgl. Tabelle III.1.1).

Tabelle III.1.1 zeigt, dass das Gaußsche Eliminationsverfahren und die Cholesky-Zerlegung viel zu aufwändig sind in Hinblick auf Speicherbedarf und benötigte arithmetische Operationen. Diese Beobachtung führt auf die effizienten iterativen Löser, die wir in Kapitel VI betrachten werden.

TABELLE III.1.1. Speicherbedarf und Rechenaufwand zur Lösung der Differenzendiskretisierung der Reaktions-Diffusions-Gleichung mit dem Gaußschen Eliminationsverfahren oder der Cholesky-Zerlegung

	$d=2$		$d=3$	
h	Speicher	Aufwand	Speicher	Aufwand
$\frac{1}{16}$	$3.3\cdot10^3$	$7.6\cdot10^5$	$7.6\cdot10^5$	$1.7\cdot10^8$
$\frac{1}{32}$	$2.9\cdot10^4$	$2.8\cdot10^7$	$2.8\cdot10^7$	$2.8\cdot10^{10}$
$\frac{1}{64}$	$2.5\cdot10^5$	$9.9\cdot10^8$	$9.9\cdot10^8$	$3.9\cdot10^{12}$
$\frac{1}{128}$	$2.0\cdot10^6$	$3.3\cdot10^{10}$	$3.3\cdot10^{10}$	$5.3\cdot10^{14}$

Man kann für die Differenzendiskretisierung die folgende *a priori Fehlerabschätzung* beweisen.

Für beliebige Diffusion A gilt

$$\max_{x\in\Omega_h}|u(x)-u_h(x)|\le c_1 h.$$

Die Konstante c_1 hängt von Ableitungen bis zur Ordnung 2 von A und von Ableitungen von u bis zur Ordnung 3 ab. Ist die Diffusion A eine konstante Diagonalmatrix, so gilt

$$\max_{x\in\Omega_h}|u(x)-u_h(x)|\le c_2 h^2.$$

Die Konstante c_2 hängt von Ableitungen von u bis zur Ordnung 4 ab.

In Hinblick auf Beispiel II.3.1 sind die Differenzierbarkeitsannahmen an die Lösung der Reaktions-Diffusions-Gleichung unrealistisch. Ähnlich wie bei gewöhnlichen Randwertproblemen gelten für die Finite-Element-Methoden des nächsten Kapitels Fehlerabschätzungen unter realistischeren Differenzierbarkeitsannahmen.

Die a priori Fehlerabschätzung liefert wieder nur eine Aussage über das asymptotische Verhalten des Fehlers bei immer kleiner werdender Gitterweite h. Sie erlaubt keine Rückschlüsse auf die tatsächliche Größe und räumliche Verteilung des Fehlers. Ähnlich wie bei gewöhnlichen Randwertprobleme kann dieses Manko nur durch eine a posteriori Fehlerkontrolle und darauf basierende adaptive Gitterverfeinerung bei Finite-Element-Methoden behoben werden (vgl. Kapitel V). Differenzenverfahren sind für adaptiv und lokal verfeinerte Gitter ungeeignet.

In der Praxis treten häufig *Neumann-Randbedingungen*

$$\mathbf{n}\cdot A\nabla u = g$$

auf einem Teil des Randes auf. Dies führt auf die Notwendigkeit für einen diskreten Randpunkt x und einen gegebenen Einheitsvektor $\boldsymbol{\nu}$ (hier $\frac{1}{|\mathbf{n}\cdot A|}\mathbf{n}\cdot A$) eine Näherung für die *Richtungsableitung* $\boldsymbol{\nu}\cdot\nabla u(x)$ im Punkt x zu bestimmen. Hierzu gehen wir wie folgt vor (vgl. Abbildung III.1.2):

- Bestimme den Schnittpunkt y (türkis) der Geraden durch x (magenta) in Richtung $\boldsymbol{\nu}$ (rot) mit dem Rand γ_x (türkis) des Quadrates mit Mittelpunkt x und Kantenlänge $2h$.
- Bestimme den am dichtesten an y liegenden Gitterpunkt z (blau) auf γ_x.
- Approximiere $\boldsymbol{\nu}\cdot\nabla u(x)$ durch

$$\partial_{h,\boldsymbol{\nu}}u(x)=\frac{1}{|x-z|}\big[u(x)-u(z)\big].$$

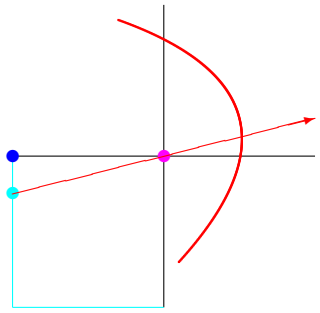


ABBILDUNG III.1.2. Approximation von $\boldsymbol{\nu} \cdot \nabla u(x)$

Eine andere Schwierigkeit ist die Diskretisierung von *Konvektionstermen* der Form

$$\mathbf{a} \cdot \nabla u(x) = a_1 \frac{\partial u}{\partial x_1}(x) + \cdots + a_d \frac{\partial u}{\partial x_d}(x).$$

Dabei ist $\mathbf{a} \neq 0$ ein gegebener Vektor und x ein Punkt im diskreten Gebiet. Für die Diskretisierung von $a_i \dfrac{\partial u}{\partial x_i}(x)$ stehen zur Auswahl:

$$a_i \partial^+_{h,i} u(x) = a_i \frac{1}{h}\big[u(x + he_i) - u(x)\big]$$
(vorwärts Differenzenquotient)

$$a_i \partial^-_{h,i} u(x) = a_i \frac{1}{h}\big[u(x) - u(x - he_i)\big]$$
(rückwärts Differenzenquotient)

$$a_i \frac{1}{2}\big[\partial^+_{h,i} u(x) + \partial^-_{h,i} u(x)\big] = a_i \frac{1}{2h}\big[u(x + he_i) - u(x - he_i)\big]$$
(symmetrischer Differenzenquotient)

Falls die *Péclet-Zahl* $\frac{a_i h}{\nu}$ größer ist als 1, kann jede Wahl zu unphysikalischen Oszillationen der numerischen Lösung führen. Dabei ist ν eine untere Schranke für den kleinsten Eigenwert der Diffusionsmatrix $A(x)$. Die Oszillationen treten auf, wenn die Matrix der Differenzendiskretisierung negative Diagonalelemente und positive Nicht-Diagonalelemente aufweist. Daher ist je nach Vorzeichen von a_i der rückwärts- oder vorwärts Differenzenquotient zu wählen (*upwinding*):

$$a_i \partial^u_{h,i} u(x) = \begin{cases} a_i \frac{1}{h}\big[u(x) - u(x - he_i)\big] & \text{falls } a_i > 0, \\ a_i \frac{1}{h}\big[u(x + he_i) - u(x)\big] & \text{falls } a_i < 0. \end{cases}$$

III.2. Parabolische Differentialgleichungen

In diesem Abschnitt betrachten wir die *Wärmeleitungsgleichung*

$$\begin{aligned} \frac{\partial u}{\partial t} - \operatorname{div}(A\nabla u) + \alpha u &= f && \text{in } \Omega \times (0,\infty) \\ u &= 0 && \text{auf } \Gamma \times (0,\infty) \\ u(x,0) &= u_0(x) && \text{in } \Omega \end{aligned}$$

als ein Modellproblem für lineare, parabolische Differentialgleichungen zweiter Ordnung. Hierbei ist Ω eine Teilmenge des $\mathbb{R}^d$ mit $d = 2$ oder $d = 3$, $A(x)$ für jedes x in Ω eine zeitlich konstante, symmetrische, positiv definite, $d \times d$ Matrix und $\alpha(x)$ für jedes x in Ω eine zeitlich konstante, nicht-negative Zahl. Physikalisch beschreiben wieder A eine örtlich veränderliche Diffusion und α eine örtlich veränderliche Reaktion. Mit etwas größerem notationellen Aufwand kann man genauso eine zeitlich veränderliche Diffusion und Reaktion betrachten.

Die Idee der Differenzendiskretisierung lässt sich wie folgt beschreiben:

- Diskretisiere die Ortsableitungen wie bei der Reaktions-Diffusions-Gleichung.
- Dies führt auf ein System *gewöhnlicher Differentialgleichungen* der Form

$$\begin{aligned} \dot{u}_h(t) &= f_h(t) - L_h u_h(t) \quad \text{für } t > 0 \\ u_h(0) &= u_{0,h}. \end{aligned}$$

- Ersetze die Zeitableitung durch einen rückwärtigen Differenzenquotienten und setze diesen gleich einer Konvexkombination der rechten Seite zu den entsprechenden Zeiten *(θ-Schema)*.

Zur Realisierung dieser Idee wählen wir eine Ortsschrittweite $h > 0$, eine Zeitschrittweite $\tau > 0$ und einen Parameter $\theta \in [0,1]$, bezeichnen mit L_h die Matrix der Differenzendiskretisierung der Reaktions-Diffusions-Gleichung und setzen $f_h^n = f(x, n\tau)$ für alle $x \in \Omega_h$. Damit lautet die *Differenzendiskretisierung* der Wärmeleitungsgleichung:

Setze $u_h^0 = u_0(x)$ für alle $x \in \overline{\Omega}_h$.
Bestimme u_h^n für $n = 1, 2, \ldots$ sukzessive durch

$$\frac{1}{\tau}\big(u_h^n - u_h^{n-1}\big) = \theta\big(f_h^n - L_h u_h^n\big) + (1-\theta)\big(f_h^{n-1} - L_h u_h^{n-1}\big).$$

Die Differenzendiskretisierung hat folgende Eigenschaften:

Die Wahl $\theta = 0$ entspricht dem *expliziten Euler-Verfahren*, $\theta = 1$ dem *impliziten Euler-Verfahren* und $\theta = \frac{1}{2}$ dem *Verfahren von Crank-Nicolson*.
Für $\theta = 0$ erfordert die Berechnung von u_h^n lediglich eine Matrix-Vektor-Multiplikation.
Für $\theta > 0$ erfordert die Berechnung von u_h^n das Lösen eines linearen Gleichungssystems mit symmetrisch, positiv definiter Matrix $\frac{1}{\tau} I + L_h$.
Für $\theta < \frac{1}{2}$ muss für Orts- und Zeitschritt die *CFL-Bedingung* $\tau \lesssim h^2$ erfüllt sein.
Für $\theta \geq \frac{1}{2}$ können Orts- und Zeitschritt unabhängig voneinander gewählt werden.

Um die Fälle $\theta = \frac{1}{2}$ und $\theta \neq \frac{1}{2}$ bei der Fehlerabschätzung unterscheiden zu können, setzen wir

$$\gamma = \begin{cases} 1 & \text{für } \theta \neq \frac{1}{2} \\ 2 & \text{für } \theta = \frac{1}{2} \end{cases}.$$

Damit erhalten wir die folgende *a priori Fehlerabschätzung*. Wie bei elliptischen Problemen stellt sie unrealistische Differenzierbarkeitsforderungen an die Lösung der Differentialgleichung und erlaubt keine Rückschlüsse auf die tatsächliche Größe des Fehlers und seine räumliche und zeitliche Verteilung. Diese Nachteile werden durch die Methoden des Kapitels VII behoben.

Für beliebige Diffusion A gilt

$$\max_{x \in \Omega_h, n \geq 0} |u(x, n\tau) - u_h^n(x)| \leq c_1 \big(\tau^\gamma + h\big).$$

Die Konstante c_1 hängt von Ableitungen bis zur Ordnung 2 von A und von Ableitungen von u bis zur Ordnung 3 ab.
Ist die Diffusion A eine konstante Diagonalmatrix, so gilt

$$\max_{x \in \Omega_h, n \geq 0} |u(x, n\tau) - u_h^n(x)| \leq c_2 \big(\tau^\gamma + h^2\big).$$

Die Konstante c_2 hängt von Ableitungen von u bis zur Ordnung 4 ab.

Bei parabolischen Gleichungen werden zeitabhängige Diffusions- und Reaktionskoeffizienten zu den gleichen diskreten Zeiten ausgewertet wie die entsprechenden u_h-Terme. Konvektionsterme und Neumann-Randbedingungen werden wie bei elliptischen Gleichungen diskretisiert; die Matrix L_h ändert sich entsprechend.

III.3. Hyperbolische Differentialgleichungen

Als ein Modellproblem für lineare, hyperbolische Differentialgleichungen zweiter Ordnung betrachten wir abschließend die *Wellen-Gleichung*

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2} - \operatorname{div}(A\nabla u) + \alpha u &= f && \text{in } \Omega \times (0,\infty) \\ u &= 0 && \text{auf } \Gamma \times (0,\infty) \\ u(x,0) &= u_0(x) && \text{in } \Omega \\ \frac{\partial u}{\partial t}(x,0) &= u_1(x) && \text{in } \Omega \end{aligned}$$

Dabei ist wieder Ω eine Teilmenge des $\mathbb{R}^d$ mit $d = 2$ oder $d = 3$, $A(x)$ für jedes x in Ω eine zeitlich konstante, symmetrische, positiv definite, $d \times d$ Matrix und $\alpha(x)$ für jedes x in Ω eine zeitlich konstante, nicht-negative Zahl. Mit notationellem Mehraufwand können auch eine zeitlich veränderliche Diffusion A und Reaktion α betrachtet werden.

Die Differenzendiskretisierung basiert auf folgender Idee:

- Diskretisiere die Ortsableitungen wie bei der Reaktions-Diffusions-Gleichung.
- Dies führt auf ein System gewöhnlicher Differentialgleichungen der Form

$$\begin{aligned} \ddot{u}_h(t) &= f_h(t) - L_h u_h(t) \quad \text{für } t > 0 \\ u_h(0) &= u_{0,h} \\ \dot{u}_h(0) &= u_{1,h}. \end{aligned}$$

- Ersetze die Zeitableitung durch einen *symmetrischen Differenzenquotienten* und setze diesen gleich der rechten Seite zur mittleren Zeit.

Wie im vorigen Abschnitt wählen wir eine Ortsschrittweite $h > 0$ und eine Zeitschrittweite $\tau > 0$, bezeichnen mit L_h die Matrix der Differenzendiskretisierung der Reaktions-Diffusions-Gleichung und setzen $f_h^n = f(x, n\tau)$ für alle $x \in \Omega_h$. Damit lautet die *Differenzendiskretisierung* der Wellen-Gleichung:

Setze $u_h^0 = u_0(x)$ für alle $x \in \overline{\Omega}_h$.
Setze $u_h^1 = u_h^0(x) + \tau u_1(x)$ für alle $x \in \overline{\Omega}_h$.
Bestimme u_h^n für $n = 2, 3, \ldots$ sukzessive durch

$$\frac{1}{\tau^2}\big(u_h^n - 2u_h^{n-1} + u_h^{n-2}\big) = \big(f_h^{n-1} - L_h u_h^{n-1}\big).$$

Die Differenzendiskretisierung hat folgende Eigenschaften:

Die Berechnung von u_h^n erfordert lediglich eine Matrix-Vektor-Multiplikation.
Für Orts- und Zeitschritt muss die *CFL-Bedingung* $\tau \lesssim h$ erfüllt sein.

Es gilt folgende *a priori Fehlerabschätzung*:

Für beliebige Diffusion A gilt

$$\max_{x\in\Omega_h, n\geq 0} |u(x,n\tau) - u_h^n(x)| \leq c_1\big(\tau + h\big).$$

Die Konstante c_1 hängt von Ableitungen bis zur Ordnung 2 von A und von Ableitungen von u bis zur Ordnung 3 ab.
Ist die Diffusion A eine konstante Diagonalmatrix, so gilt

$$\max_{x\in\Omega_h, n\geq 0} |u(x,n\tau) - u_h^n(x)| \leq c_2\big(\tau^2 + h^2\big).$$

Die Konstante c_2 hängt von Ableitungen von u bis zur Ordnung 4 ab.

Für die Behandlung von zeitabhängigen Diffusions- und Reaktionskoeffizienten sowie von Konvektionstermen und Neumann-Randbedingungen gelten die gleichen Bemerkungen wie bei parabolischen Gleichungen.

KAPITEL IV

Finite-Element-Methoden für elliptische Differentialgleichungen

IV.1. Variationsformulierung

Um die Beschreibung der Finite-Element-Methoden zu erleichtern, betrachten wir zunächst die *Reaktions-Diffusions-Gleichung*

$$\begin{aligned} -\operatorname{div}(A\nabla u) + \alpha u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{auf } \Gamma \end{aligned}$$

als ein Modellproblem für lineare, elliptische Differentialgleichungen zweiter Ordnung. In den Abschnitten IV.4 und IV.5 werden wir andere, komplexere Differentialgleichungen betrachten. Dabei ist Ω ein *Polyeder* in $\mathbb{R}^d$ mit $d = 2$ oder $d = 3$, $A(x)$ für jedes x in Ω eine symmetrische, positiv definite, $d \times d$ Matrix und $\alpha(x)$ für jedes x in Ω eine nicht-negative Zahl. Die Beschränkung auf Polyeder vermeidet unnötige technische Schwierigkeiten bei der Approximation gekrümmter Ränder.

Die Variationsformulierung der Reaktions-Diffusions-Gleichung beruht auf den gleichen Ideen wie die Variationsformulierung des Sturm-Liouville-Problems in Abschnitt I.7. Dazu benötigen wir ein mehrdimensionales Analogon zur partiellen Integration. Dieses beruht auf dem *Integralsatz von Gauß*

Satz von Gauß	$\displaystyle\int_\Omega \operatorname{div} \mathbf{w} dx = \int_\Gamma \mathbf{w} \cdot \mathbf{n} dS$
mit der *Divergenz*	$\displaystyle\operatorname{div} \mathbf{w} = \sum_{i=1}^{d} \frac{\partial w_i}{\partial x_i}$

Wendet man den Satz von Gauß auf $\mathbf{w} = v(A\nabla u)$ an, erhält man das folgende Analogon zur partiellen Integration

$$\begin{aligned} \int_\Omega v \operatorname{div}(A\nabla u) dx + \int_\Omega \nabla v \cdot A\nabla u dx &= \int_\Omega \operatorname{div}(vA\nabla u) dx \\ &= \int_\Gamma v\mathbf{n} \cdot A\nabla u dS. \end{aligned}$$

Ist insbesondere $v = 0$ auf Γ, folgt hieraus

$$\int_\Omega \nabla v \cdot A\nabla u dx = -\int_\Omega v \operatorname{div}(A\nabla u) dx.$$

Mit diesen Vorbereitungen können wir die Idee der Variationsformulierung beschreiben:

- Multipliziere die Differentialgleichung mit einer stetig differenzierbaren Funktion v mit $v = 0$ auf Γ:

  $$-\operatorname{div}(A\nabla u)(x)v(x) + \alpha(x)u(x)v(x) = f(x)v(x)$$

  für alle $x \in \Omega$.
- Integriere das Ergebnis über Ω:

  $$\int_\Omega \big[-\operatorname{div}(A\nabla u)v + \alpha uv\big] dx = \int_\Omega fv dx.$$

- Integriere den Ableitungsterm partiell:

  $$-\int_\Omega \operatorname{div}(A\nabla u)v dx = \int_\Omega \nabla v \cdot A\nabla u dx.$$

Um diese Idee auf ein solides mathematisches Fundament zu stellen, müssen wir ähnlich wie in Abschnitt I.7 die Eigenschaften der Funktionen u und v präziser fassen. Klassische Eigenschaften wie stetige Differenzierbarkeit sind dazu zu restriktiv. Der Begriff der Ableitung muss daher geeignet verallgemeinert werden. In Hinblick auf die Diskretisierung sollten insbesondere stückweise differenzierbare Funktionen im erweiterten Sinn differenzierbar sein.

Der Satz von Gauß angewandt auf $\mathbf{w} = uv\mathbf{e}_i$ ($\mathbf{e}_i$ i-te Einheitsvektor mit i-ter Komponente 1 und restlichen Komponenten 0) liefert

$$\begin{aligned}\int_\Omega \frac{\partial u}{\partial x_i} v dx + \int_\Omega u \frac{\partial v}{\partial x_i} dx &= \int_\Omega \frac{\partial (uv)}{\partial x_i} dx \\ &= \int_\Gamma uv\mathbf{n}_i dS.\end{aligned}$$

Ist $u = 0$ oder $v = 0$ auf Γ, folgt hieraus

$$\int_\Omega \frac{\partial u}{\partial x_i} v dx = -\int_\Omega u \frac{\partial v}{\partial x_i} dx.$$

Diese Identität wird zur Definition der *schwachen Ableitung* benutzt:

> Die Funktion u heißt *schwach differenzierbar* bzgl. x_i mit *schwacher Ableitung* w_i, wenn für jede stetig differenzierbare Funktion v mit $v = 0$ auf Γ gilt
>
> $$\int_\Omega w_i v dx = -\int_\Omega u \frac{\partial v}{\partial x_i} dx.$$

Ist u bzgl. jeder Variablen $x_1, \ldots, x_d$ schwach differenzierbar, so nennt man u *schwach differenzierbar* und schreibt ∇u für den Vektor $(w_1, \ldots, w_d)$ der schwachen Ableitungen.

Beispiel IV.1.1. Jede stetig differenzierbare Funktion ist schwach differenzierbar und die schwache Ableitung ∇u stimmt mit dem klassischen Gradienten überein.
Eine stückweise stetig differenzierbare Funktion ist genau dann schwach differenzierbar, wenn sie global stetig ist; dann stimmt die schwache Ableitung ∇u mit dem stückweise definierten klassischen Gradienten überein (vgl. Abbildung IV.1.1).

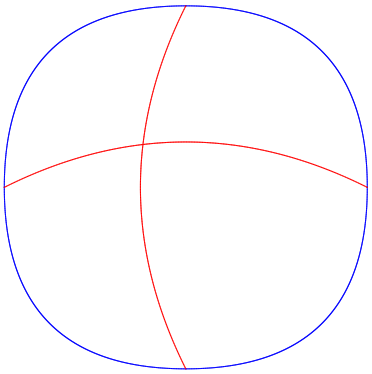


ABBILDUNG IV.1.1. Unterteilung eines Gebietes Ω in Teilgebiete

Wie bei den Sturm-Liouville-Problemen in Abschnitt I.7 beruht die Variationsformulierung der Reaktions-Diffusions-Gleichung auf den Sobolev-Räumen.

Die L^2-*Norm* ist definiert durch

$$\|v\| = \left\{ \int_\Omega |v|^2 dx \right\}^{\frac{1}{2}}.$$

$L^2(\Omega)$ ist der *Lebesgue-Raum* aller Funktionen v mit endlicher L^2-Norm $\|v\|$.
$H^1(\Omega)$ ist der *Sobolev-Raum* aller Funktionen v in $L^2(\Omega)$, deren schwache Ableitung ∇v existiert und deren Euklidische Norm $|\nabla v|$ ebenfalls in $L^2(\Omega)$ ist.
$H_0^1(\Omega)$ ist der *Sobolev-Raum* aller Funktionen v in $H^1(\Omega)$ mit $v = 0$ auf Γ.

Beispiel IV.1.2. Jede beschränkte Funktion ist in $L^2(\Omega)$.
Die Funktion

$$v(x) = \frac{1}{\sqrt{x^2 + y^2}}$$

ist nicht in $L^2(B(0,1))$, da

$$\int_{B(0,1)} |v(x)|^2 dx = 2\pi \int_0^1 \frac{1}{r} dr$$

nicht endlich ist. Dabei bezeichnet $B(0,1)$ den Kreis mit Radius 1 und Mittelpunkt im Ursprung.
Eine stückweise stetig differenzierbare Funktion ist genau dann in $H^1(\Omega)$, wenn sie global stetig ist.
Punktwerte sind für Funktionen in $H^1(\Omega)$ nicht definiert. Die Funktion

$$v(x) = \ln(|\ln(\sqrt{x^2+y^2})|$$

ist in $H^1(B(0,1))$, besitzt aber keinen endlichen Wert im Ursprung.

Mit diesen Notationen lautet die *Variationsformulierung* der Reaktions-Diffusions-Gleichung:

Finde $u \in H_0^1(\Omega)$ so, dass für alle $v \in H_0^1(\Omega)$ gilt

$$\int_\Omega \big[\nabla v \cdot A\nabla u + \alpha uv\big] dx = \int_\Omega fv dx.$$

Sie hat folgende Eigenschaften:

Das Variationsproblem hat eine eindeutige Lösung.
Die Lösung des Variationsproblems ist das eindeutige *Minimum* in $H_0^1(\Omega)$ der *Energiefunktion*

$$\frac{1}{2}\int_\Omega \big[\nabla u \cdot A\nabla u + \alpha u^2\big] dx - \int_\Omega fu dx.$$

IV.2. Finite-Element-Diskretisierung

Die Finite-Element-Diskretisierung beruht auf folgender Idee:

- Zerlege Ω in nicht überlappende, einfache Teilgebiete, sog. *Elemente*, wie Dreiecke, Parallelogramme, Tetraeder, Parallelepipede, ... *(Unterteilung)*.
- Ersetze in der Variationsformulierung den Raum $H_0^1(\Omega)$ durch einen endlich-dimensionalen Unterraum bestehend aus stetigen Funktionen, die stückweise auf den Elementen Polynome sind *(Finite-Element-Raum)*.
- Dies führt auf ein lineares Gleichungssystem für die Approximation $u_\mathcal{T}$ an die Lösung u der partiellen Differentialgleichung.

Im folgenden bezeichnet

$$\mathcal{T} = \{K_i : 1 \le i \le N_{\mathcal{T}}\}$$

stets eine *Unterteilung* von Ω, die folgende Bedingungen erfüllt:

Ω ist die Vereinigung aller Elemente K in $\mathcal{T}$.
Zulässigkeit: Je zwei Elemente K und K' in $\mathcal{T}$ sind entweder disjunkt oder haben einen Eckpunkt oder eine ganze Kante oder, falls $d = 3$ ist, eine ganze Seitenfläche gemeinsam (vgl. Abbildung IV.2.1).
Affine Äquivalenz: Jedes Element K ist ein Dreieck oder Parallelogramm, falls $d = 2$ ist, oder ein Tetraeder oder Parallelepiped, falls $d = 3$ ist.

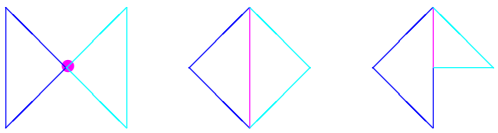


ABBILDUNG IV.2.1. Zulässige (links, Mitte) und nicht zulässige benachbarte Elemente (rechts)

Wenn Ω kein Polyeder ist, kann man es nicht durch geradlinige Elemente wie Dreiecke überdecken, so dass der Rand von Ω zusätzlich approximiert werden muss.

Die Zulässigkeit wird benötigt, damit die Finite-Element-Räume in $H_0^1(\Omega)$ enthalten sind. Bei fehlender Zulässigkeit muss die Inklusion der Finite-Element-Räume explizit erzwungen werden, was die Implementierung erschwert.

Es können auch allgemeine Vierecke und Quader betrachtet werden, was die Implementierung erschwert.

Die zu einer Unterteilung $\mathcal{T}$ gehörigen *Finite-Element-Räume* sind wie folgt definiert:

$$R_k(K) = \begin{cases} \operatorname{span}\{x_1^{\alpha_1} \cdot \ldots \cdot x_d^{\alpha_d} : \alpha_1 + \ldots + \alpha_d \le k\} \\ K \text{ Dreieck oder Tetraeder} \\ \operatorname{span}\{x_1^{\alpha_1} \cdot \ldots \cdot x_d^{\alpha_d} : \max\{\alpha_1, \ldots, \alpha_d\} \le k\} \\ K \text{ Parallelogram oder Parallelepiped} \end{cases}$$

$$S^{k,-1}(\mathcal{T}) = \{v : \Omega \to \mathbb{R} : v\big|_K \in R_k(K) \text{ für alle } K \in \mathcal{T}\}$$

$$S^{k,0}(\mathcal{T}) = S^{k,-1}(\mathcal{T}) \cap C(\overline{\Omega})$$

$$S_0^{k,0}(\mathcal{T}) = S^{k,0}(\mathcal{T}) \cap H_0^1(\Omega)$$

$$= \{v \in S^{k,0}(\mathcal{T}) : v = 0 \text{ auf } \Gamma\}$$

Wegen der globalen Stetigkeit ist $S^{k,0}(\mathcal{T}) \subset H^1(\Omega)$. Der Polynomgrad k kann auch von Element zu Element variieren, was auf einen höheren Aufwand für die Implementierung führt.

Mit diesen Notationen lautet die *Finite-Element-Diskretisierung* der Reaktions-Diffusions-Gleichung:

Finde $u_\mathcal{T} \in S_0^{k,0}(\mathcal{T})$ *(Ansatzfunktion)* so, dass für alle $v_\mathcal{T} \in S_0^{k,0}(\mathcal{T})$ *(Testfunktion)* gilt

$$\int_\Omega \big[\nabla v_\mathcal{T} \cdot A \nabla u_\mathcal{T} + \alpha u_\mathcal{T} v_\mathcal{T}\big] dx = \int_\Omega f v_\mathcal{T} dx.$$

Die Finite-Element-Diskretisierung hat folgende Eigenschaften:

Das diskrete Problem hat eine eindeutige Lösung.
Die Lösung des diskreten Problems ist das eindeutige *Minimum* in $S_0^{k,0}(\mathcal{T})$ der *Energiefunktion*

$$\frac{1}{2}\int_\Omega \big[\nabla u \cdot A \nabla u + \alpha u^2\big] dx - \int_\Omega f u dx.$$

Das diskrete Problem führt nach Wahl einer Basis für $S_0^{k,0}(\mathcal{T})$ auf ein lineares Gleichungssystem.
Die Matrix, genannt *Systemsteifigkeits-* oder *Steifigkeitsmatrix*, ist symmetrisch, positiv definit.
Der Vektor der rechten Seite wird häufig als *Lastvektor* bezeichnet.
Die Zahl der Gleichungen und Unbekannten ist $\approx k^d N_\mathcal{T}$.

Bezeichnet man mit $h_\mathcal{T}$ den maximalen Durchmesser aller Elemente in $\mathcal{T}$, kann man folgende *a priori Fehlerabschätzungen* beweisen:

Für die Lösung u des Varitionsproblems und die Lösung $u_\mathcal{T}$ der Finite-Element-Diskretisierung gilt die *Fehlerabschätzung*

$$\|\nabla u - \nabla u_\mathcal{T}\| \le c_1 h_\mathcal{T}^k.$$

Ist die Menge Ω zusätzlich *konvex*, gilt die verschärfte Abschätzung

$$\|u - u_\mathcal{T}\| \le c_2 h_\mathcal{T}^{k+1}.$$

Die Konstanten c_1 und c_2 hängen von Ω, der Diffusion A, der Reaktion α und den L^2-Normen der Ableitungen bis zur Ordnung $k-1$ von f ab.

Für die praktische Realisierung der Finite-Element-Methoden müssen wir in den folgenden Abschnitten noch folgende Punkte klären:

- Wahl einer Basis für $S_0^{k,0}(\mathcal{T})$,

- Aufstellen der Steifigkeitsmatrix und des Lastvektors,
- Berechnung der dabei auftretenden Integrale,
- Behandlung gekrümmter Ränder,
- Behandlung von Neumann-Randbedingungen,
- Behandlung von Konvektionstermen (s. §IV.4),
- Bestimmung einer optimalen Unterteilung (s. Kap. V),
- Lösung der diskreten Probleme (s. Kap. VI).

IV.3. Praktische Aspekte

Üblicherweise wählt man für die Räume $S^{k,0}(\mathcal{T})$ und $S_0^{k,0}(\mathcal{T})$ eine *nodale Basis* $\lambda_{z,k}$, $z \in \mathcal{N}_{\mathcal{T},k}$, auch *Lagrangesche Basis* genannt. Diese ist dadurch charakterisiert, dass jede Basisfunktion an genau einem Punkt z einer diskreten Punktmenge $\mathcal{N}_{\mathcal{T},k}$, den sog. *globalen Freiheitsgraden* oder *Knoten*, den Wert 1 annimmt und an allen anderen Punkten dieser Menge verschwindet. Die globalen Freiheitsgrade werden durch Vereinigung der Freiheitsgrade $\mathcal{N}_{K,k}$ auf den Elementen, den sog. *Elementfreiheitsgraden*, erzeugt, d.h.

$$\mathcal{N}_{\mathcal{T},k} = \bigcup_{K \in \mathcal{T}} \mathcal{N}_{K,k}.$$

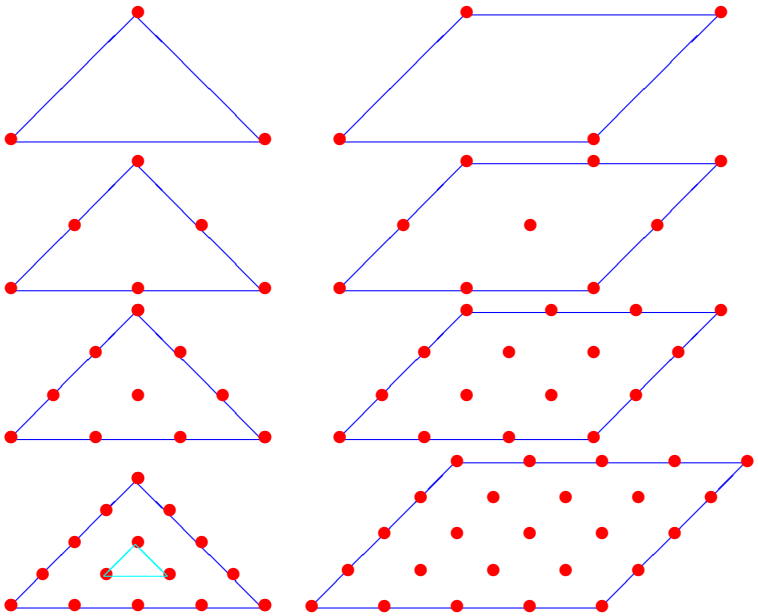


Abbildung IV.3.1. Elementfreiheitsgrade $\mathcal{N}_{K,k}$ für Dreiecke (linke Spalte) und Parallelogramme (rechte Spalte) und Polynomgrade $k = 1, \ldots, 4$ (Zeilen $1-4$)

Abbildung IV.3.1 zeigt die Elementfreiheitsgrade $\mathcal{N}_{K,k}$ für Dreiecke und Parallelogramme K und Polynomgrade $k = 1, \ldots, 4$. Die Menge $\mathcal{N}_{K,k}$ wird wie folgt konstruiert:

- *Parallelogramme und Parallelepipede:* Wähle auf jeder Kante $k+1$ äquidistante Punkte, die die Endpunkte der Kante umfassen, und bilde das entsprechende Gitter.
- *Dreiecke:* Unterteile jede Kante in $k+1$ äquidistante Punkte, die die Endpunkte der Kante umfassen. Dies erzeugt $3k$ Punkte. Falls $3k < \frac{1}{2}(k+1)(k+2)$ ist, wiederhole den Prozess für das Dreieck, das durch die Geraden durch die am weitesten außen liegenden, soeben erzeugten Punkte gebildet wird (türkis farbenes kleines Dreieck in Abbildung IV.3.1 unten links). Wiederhole diesen Prozess bis $\frac{1}{2}(k+1)(k+2)$ Punkte erzeugt sind.
- *Tetraeder:* Wähle die Punkte auf den Seitenflächen wie bei Dreiecken. Falls so weniger als $\frac{1}{6}(k+1)(k+2)(k+3)$ Punkte erzeugt wurden, wiederhole diesen Prozess ggf. mehrmals mit kleineren in das ursprüngliche Tetraeder eingeschriebenen Tetraedern.

Abbildung IV.3.2 zeigt die Menge $\mathcal{N}_{\mathcal{T},k}$ für Unterteilungen, die aus zwei Dreiecken oder Rechtecken bestehen, für die Polynomgrade $k = 1, 2$. Wegen der Zulässigkeit der Unterteilung stimmen die Punktmengen $\mathcal{N}_{K,k}$ für benachbarte Elemente auf den gemeinsamen Kanten und Flächen überein.

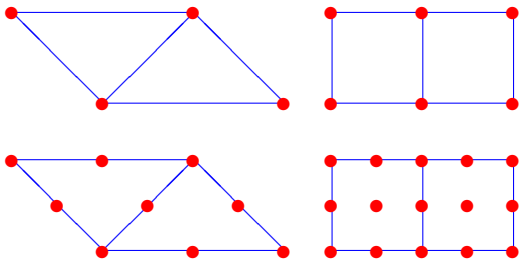


ABBILDUNG IV.3.2. Globale Feiheitsgrade $\mathcal{N}_{\mathcal{T},k}$ für Unterteilungen, die aus zwei Dreiecken (linke Spalte) oder Rechtecken (rechte Spalte) bestehen, für die Polynomgrade $k = 1$ (obere Zeile) und $k = 2$ (untere Zeile)

Die nodale Basisfunktion $\lambda_{z,k}$ zum *Knoten* $z \in \mathcal{N}_{\mathcal{T},k}$ ist dann eindeutig definiert durch die Bedingungen

$$\lambda_{z,k} \in S^{k,0}(\mathcal{T}), \quad \lambda_{z,k}(z) = 1, \quad \lambda_{z,k}(y) = 0 \text{ für alle } y \in \mathcal{N}_{\mathcal{T},k} \setminus \{z\}.$$

Abbildung IV.3.3 zeigt eine typische Funktion $\lambda_{z,1}$.

Die nodalen Basen haben folgende Eigenschaften:

$\{\lambda_{z,k} : z \in \mathcal{N}_{\mathcal{T},k}\}$ ist eine Basis für $S^{k,0}(\mathcal{T})$.
$\{\lambda_{z,k} : z \in \mathcal{N}_{\mathcal{T},k} \setminus \Gamma\}$ ist eine Basis für $S_0^{k,0}(\mathcal{T})$, Freiheitsgrade auf dem Rand Γ werden unterdrückt.

Jede Funktion in $S^{k,0}(\mathcal{T})$ bzw. $S_0^{k,0}(\mathcal{T})$ ist eindeutig bestimmt durch ihre Werte in den Punkten von $\mathcal{N}_{\mathcal{T},k}$ bzw. $S^{k,0}(\mathcal{T}) \setminus \Gamma$. Daher können diese Funktionen mit Vektoren entsprechender Länge identifiziert werden.
$\lambda_{z,k}$ verschwindet außerhalb der Vereinigung aller der Elemente, die den Punkt z enthalten.
Die Steifigkeitsmatrix *dünn besetzt*.

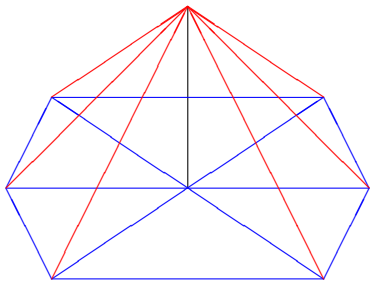


ABBILDUNG IV.3.3. Nodale Basisfunktion $\lambda_{z,1}$

Das Aufstellen der Steifigkeitsmatrix und des Lastvektors erfordert die Auswertung der Funktionen $\lambda_{z,k}$ und ihrer Ableitungen. Dies kann auf zweierlei Weise geschehen:

- Transformation auf ein *Referenzelement* $\widehat{K}$,
- Reduktion auf die nodalen Basisfunktionen $\lambda_{z,1}$ erster Ordnung und direkte Berechnung dieser aus der Elementgeometrie.

Bei dem ersten Ansatz geht man wie folgt vor:

- Bestimme die nodalen Basisfunktionen $\widehat{\lambda}_{\widehat{z},k}$ zum Referenzelement $\widehat{K}$.
- Bestimme eine *affine Transformation*

$$\widehat{K} \ni \widehat{x} \mapsto x = b_K + B_K \widehat{x}$$

  des Referenzelementes $\widehat{K}$ auf das aktuelle Element K.
- Berechne $\lambda_{z,k}$ aus $\widehat{\lambda}_{\widehat{z},k}$ mit Hilfe der affinen Transformation

$$\lambda_{z,k}(x) = \widehat{\lambda}_{\widehat{z},k}(\widehat{x}).$$

Abbildung IV.3.4 zeigt die gebräuchlichen Referenzelemente in zwei und drei Dimensionen.

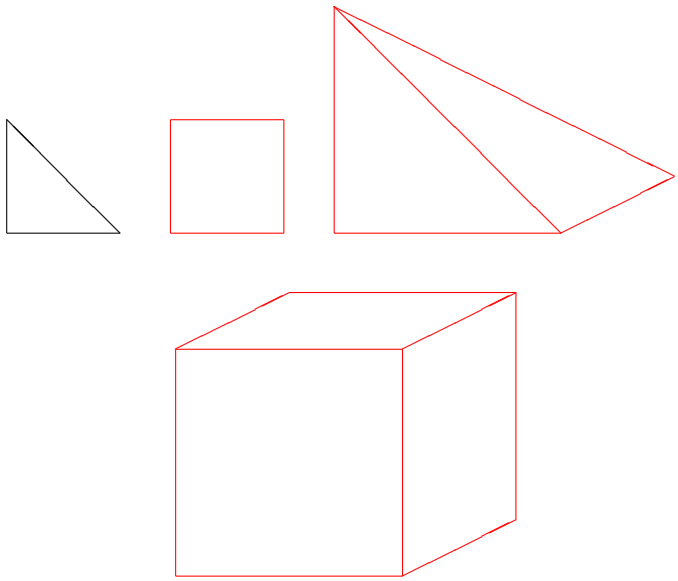


ABBILDUNG IV.3.4. Referenzdreieck, Referenzquadrat, Referenztetraeder und Referenzwürfel (v.l.n.r.)

Beispiel IV.3.1. Für das *Referenzdreieck* lauten die Basisfunktionen $\widehat{\lambda}_{\widehat{z},1}$ für die Eckpunkte

$$1-x-y, \quad x, \quad y$$

und die Basisfunktionen $\widehat{\lambda}_{\widehat{z},2}$ für die Eckpunkte

$$(1-x-y)(1-2x-2y), \quad x(2x-1), \quad y(2y-1)$$

und für die Kantenmittelpunkte

$$4x(1-x-y), \quad 4xy, \quad 4y(1-x-y).$$

Für das *Referenzquadrat* lauten die Basisfunktionen $\widehat{\lambda}_{\widehat{z},1}$ für die Eckpunkte

$$(1-x)(1-y), \quad x(1-y), \quad xy, \quad (1-x)y$$

und die Basisfunktionen $\widehat{\lambda}_{\widehat{z},2}$ für die Eckpunkte

$$(1-2x)(1-x)(1-2y)(1-y), \quad x(2x-1)(1-2y)(1-y),$$
$$x(2x-1)y(2y-1), \quad (1-2x)(1-x)y(2y-1)$$

und für die Kantenmittelpunkte

$$4x(1-x)(1-y)(1-2y), \quad 4x(2x-1)y(1-y),$$
$$4x(1-x)y(2y-1), \quad 4y(1-y)(1-2x)(1-x)$$

sowie für den Elementschwerpunkt

$$16x(1-x)y(1-y).$$

Beispiel IV.3.2. Der Vektor b_K und die Matrix B_K der affinen Transformationen für ein Dreieck, Prallelogramm und Tetraeder lauten (vgl. Abbildung IV.3.5 für die Nummerierung der Eckpunkte)

$$\begin{aligned} b_K &= \mathbf{a}_0, \quad B_K = \big(\mathbf{a}_1 - \mathbf{a}_0\,,\, \mathbf{a}_2 - \mathbf{a}_0\big), \\ b_K &= \mathbf{a}_0, \quad B_K = \big(\mathbf{a}_1 - \mathbf{a}_0\,,\, \mathbf{a}_3 - \mathbf{a}_0\big), \\ b_K &= \mathbf{a}_0, \quad B_K = \big(\mathbf{a}_1 - \mathbf{a}_0\,,\, \mathbf{a}_2 - \mathbf{a}_0\,,\, \mathbf{a}_3 - \mathbf{a}_0\big). \end{aligned}$$

Analoge Formeln gelten für Parallelepipede.

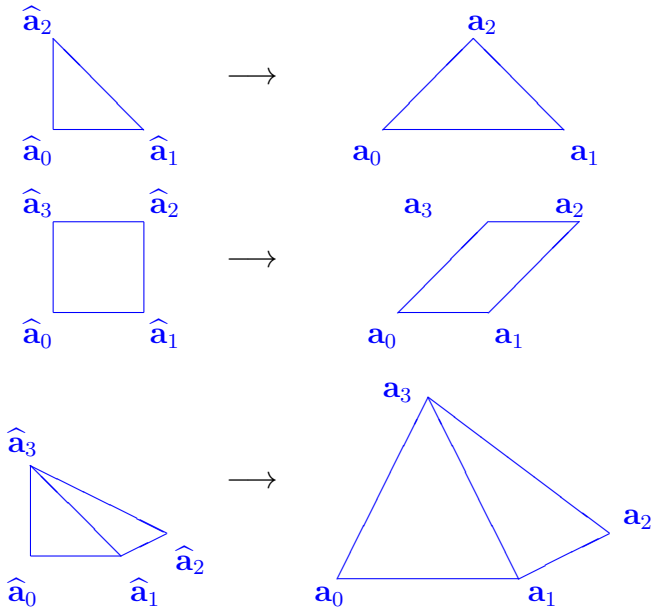


ABBILDUNG IV.3.5. Affine Transformation eines Dreieckes, Parallelogrammes und Tetraeders

Beispiel IV.3.3. Die nodalen Basisfunktionen $\lambda_{\mathbf{a}_i,1}$ erster Ordnung zu dem Eckpunkt $\mathbf{a}_i$ eines Dreieckes, Parallelogrammes oder Tetraeders sind gegeben durch (vgl. Abbildung IV.3.5 für die Nummerierung der Eckpunkte)

$$\begin{aligned} &\frac{\det(x - \mathbf{a}_{i+1}\,,\, \mathbf{a}_{i+2} - \mathbf{a}_{i+1})}{\det(\mathbf{a}_i - \mathbf{a}_{i+1}\,,\, \mathbf{a}_{i+2} - \mathbf{a}_{i+1})}, \\ &\frac{\det(x - \mathbf{a}_{i+2}\,,\, \mathbf{a}_{i+3} - \mathbf{a}_{i+2})}{\det(\mathbf{a}_i - \mathbf{a}_{i+2}\,,\, \mathbf{a}_{i+3} - \mathbf{a}_{i+2})} \cdot \frac{\det(x - \mathbf{a}_{i+2}\,,\, \mathbf{a}_{i+1} - \mathbf{a}_{i+2})}{\det(\mathbf{a}_i - \mathbf{a}_{i+2}\,,\, \mathbf{a}_{i+1} - \mathbf{a}_{i+2})}, \\ &\frac{\det(x - \mathbf{a}_{i+1}, \mathbf{a}_{i+2} - \mathbf{a}_{i+1}, \mathbf{a}_{i+3} - \mathbf{a}_{i+1})}{\det(\mathbf{a}_i - \mathbf{a}_{i+1}, \mathbf{a}_{i+2} - \mathbf{a}_{i+1}, \mathbf{a}_{i+3} - \mathbf{a}_{i+1})}. \end{aligned}$$

Dabei sind alle Indizes modulo der Zahl der Eckpunkte des jeweiligen Elementes zu nehmen. Für Parallelepipede gelten analoge Formeln mit 3 Faktoren entsprechend 3 Tetraedern.

Beispiel IV.3.4. Die nodale Basisfunktion $\lambda_{\mathbf{a}_i,2}$ zweiter Ordnung zum Eckpunkt $\mathbf{a}_i$ eines Dreieckes ist gegeben durch

$$\lambda_{\mathbf{a}_i,2} = \lambda_{\mathbf{a}_i}[\lambda_{\mathbf{a}_i} - \lambda_{\mathbf{a}_{i+1}} - \lambda_{\mathbf{a}_{i+2}}].$$

Analog ist die Funktion $\lambda_{z,2}$ zum Mittelpunkt z der Kante mit den Endpunkten $\mathbf{a}_i$ und $\mathbf{a}_{i+1}$ gegeben durch

$$\lambda_{z,2} = 4\lambda_{\mathbf{a}_i}\lambda_{\mathbf{a}_{i+1}}.$$

Für ein Parallelogramm ergibt sich für den Eckpunkt $\mathbf{a}_i$

$$\lambda_{\mathbf{a}_i,2} = \lambda_{\mathbf{a}_i}[\lambda_{\mathbf{a}_i} - \lambda_{\mathbf{a}_{i+1}} + \lambda_{\mathbf{a}_{i+2}} - \lambda_{\mathbf{a}_{i+3}}],$$

für den Mittelpunkt z der Kante mit Endpunkten $\mathbf{a}_i$ und $\mathbf{a}_{i+1}$

$$\lambda_{z,2} = 4\lambda_{\mathbf{a}_i}[\lambda_{\mathbf{a}_{i+1}} - \lambda_{\mathbf{a}_{i+2}}]$$

und den Schwerpunkt y des Parallelogrammes

$$\lambda_{y,2} = 16\lambda_{\mathbf{a}_0}\lambda_{\mathbf{a}_2}.$$

Die Einträge des Lastvektors und der Steifigkeitsmatrix sind von der Form

$$\int_\Omega f\lambda_{z,k}dx, \quad \int_\Omega \big[\nabla\lambda_{z',k} \cdot A\nabla\lambda_{z,k} + \alpha\lambda_{z',k}\lambda_{z,k}\big]dx,$$

wobei z und z' alle Knoten in $\mathcal{N}_{\mathcal{T},k} \setminus \Gamma$ durchlaufen. Diese Integrale werden kumulativ und elementweise durch Durchlaufen aller Elemente berechnet. Für den *Lastvektor* ergibt sich damit:

> Durchlaufe alle Elemente K:
>
> Durchlaufe alle Elementfreiheitsgrade z von K:
>
> Berechne
>
> $$\int_K f\lambda_{z,k}dx.$$
>
> Addiere das Ergebnis zu dem entsprechenden Eintrag des Lastvektors.

Analog ergibt sich für die *Steifigkeitsmatrix*:

> Durchlaufe alle Elemente K:
>
> Durchlaufe alle *Paare* z, z' von Elementfreiheitsgraden von K:
>
> Berechne
>
> $$\int_K \big[\nabla\lambda_{z',k} \cdot A\nabla\lambda_{z,k} + \alpha\lambda_{z',k}\lambda_{z,k}\big]dx.$$
>
> Addiere das Ergebnis zu dem entsprechenden Eintrag der Steifigkeitsmatrix.

Die exakte Berechnung der beim Aufstellen der Steifigkeitsmatrix und des Lastvektors auftretenden Integrale ist häufig nicht möglich oder zu aufwändig. Daher werden die Integrale in der Regel durch *Quadraturformeln*

$$\int_K \varphi dx \approx Q_k(\varphi) = \sum_{q \in \mathcal{Q}_K} c_q \varphi(q)$$

näherungsweise berechnet. Damit der dadurch verursachte Fehler nicht den Fehler der Finite-Element-Diskretisierung dominiert, muss die Quadraturformel mindestens die *Ordnung* $2k-2$ haben, d.h.

$$\int_K \varphi dx = Q_K(\varphi)$$

für alle $\varphi \in R_{2k-2}(K)$. Für lineare Elemente $S^{1,0}(\mathcal{T})$ reicht also die Ordnung 0; für quadratische Elemente $S^{2,0}(\mathcal{T})$ die Ordnung 2.

Beispiel IV.3.5. Die Daten

$$\begin{aligned} \mathcal{Q}_K &\quad \text{Schwerpunkt von } K, \\ c_q &= |K| \end{aligned}$$

und

$$\begin{aligned} \mathcal{Q}_K &\quad \text{Kantenmittelpunkte von } K, \\ c_q &= \frac{1}{3}|K| \quad \text{für alle } q \end{aligned}$$

liefern Quadraturformeln der Ordnung 1 bzw. 2 für Dreiecke. Dabei ist $|K|$ die Fläche von K.
Die Daten

$$\begin{aligned} \mathcal{Q}_K &\quad \text{Schwerpunkt von } K, \\ c_q &= |K| \end{aligned}$$

und

$$\begin{aligned} \mathcal{Q}_K &\quad \text{Ecken, Kantenmitten und Schwerpunkt von } K, \\ c_q &= \begin{cases} \frac{1}{36}|K| & \text{falls } q \text{ Ecke} \\ \frac{4}{36}|K| & \text{falls } q \text{ Kantenmitte} \\ \frac{16}{36}|K| & \text{falls } q \text{ Schwerpunkt} \end{cases} \end{aligned}$$

liefern Quadraturformeln der Ordnung 1 bzw. 2 für Parallelogramme. Dabei ist wieder $|K|$ die Fläche von K.
Analoge Formeln gelten für Tetraeder und Parallelogramme.

Gebiete mit gekrümmten Rändern können nicht durch geradlinige Elemente wie Dreiecke, Parallelogramme usw. exakt überdeckt werden. Mögliche Auswege sind

- krummlinige (z.B. *isoparametrische*) Elemente in Randnähe,
- Approximation des Gebietes Ω durch ein Polyeder $\Omega_{\mathcal{T}}$, das exakt überdeckt wird.

Die erste Variante macht die Implemetierung erheblich aufwändiger. Außerdem können nur spezielle krummlinige Ränder exakt dargestellt werden.
Die zweite Variante erzeugt Fehler der Ordnung $h_{\mathcal{T}}^2$. Falls Ω konvex ist, ist $\Omega_{\mathcal{T}} \subset \Omega$. Andernfalls ist $\Omega_{\mathcal{T}} \cap \Omega \neq \emptyset$. Dann müssen die Daten f, A, α der Differentialgleichung ggf. auf das größere Gebiet fortgesetzt werden.

Die *Neumann-Randbedingung*

$$\mathbf{n} \cdot A\nabla u = g \quad \text{auf } \Gamma_N \subset \Gamma$$

führt zu einem zusätzlichen Term

$$\int_{\Gamma_N} gv dS$$

auf der rechten Seite des Variationsproblems und einem zusätzlichen Term

$$\int_{\Gamma_N} gv_{\mathcal{T}} dS$$

auf der rechten Seite des diskreten Problems. Die zusätzlichen Beiträge zum Lastvektor werden beim Durchlaufen der Elemente berücksichtigt. Die Elementfreiheitsgrade, die auf dem Neumann-Rand Γ_N liegen, sind zusätzliche Unbekannte.

IV.4. Upwind und Petrov-Galerkin-Verfahren

In diesem Abschnitt betrachten wir die *Konvektions-Diffusions-Gleichung*

$$\begin{aligned} -\operatorname{div}(A\nabla u) + \mathbf{a} \cdot \nabla u + \alpha u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{auf } \Gamma \end{aligned}$$

und die im Vergleich zu den vorigen Abschnitten erforderlichen Modifikationen. Dabei ist wieder Ω ein Polyeder in $\mathbb{R}^d$ mit $d = 2$ oder $d = 3$, $A(x)$ für jedes x in Ω eine symmetrische, positiv definite, $d \times d$ Matrix und $\alpha(x)$ für jedes x in Ω eine nicht-negative Zahl. Zusätzlich ist $\mathbf{a}(x)$ für jedes x in Ω ein Vektor in $\mathbb{R}^d$, der die Bedingung

$$\alpha(x) - \frac{1}{2} \operatorname{div} \mathbf{a}(x) \geq 0$$

für jedes x in Ω erfüllt.

Der zusätzliche *Konvektionsterm* $\mathbf{a} \cdot \nabla u$ führt auf einen zusätzlichen Term $\int_\Omega v\mathbf{a} \cdot \nabla u$ im Variationsproblem. Das Variationsproblem hat nach wie vor eine eindeutige Lösung. *Aber* die Lösung des Variationsproblems ist nicht mehr Minimum einer Energiefunktion.

Bei dem diskreten Problem führt der Konvektionsterm auf einen zusätzlichen Term $\int_\Omega v_{\mathcal{T}}\mathbf{a} \cdot \nabla u_{\mathcal{T}}$. Das diskrete Problem hat nach wie vor eine eindeutige Lösung. *Aber* die Lösung des diskreten Problems ist nicht mehr Minimum einer Energiefunktion und die Steifigkeitsmatrix ist nicht mehr symmetrisch.

Falls die *Péclet-Zahl* $\frac{|\mathbf{a}|h_{\mathcal{T}}}{\nu}$ größer ist als 1, führt die übliche Finite-Element-Diskretisierung zu unphysikalischen Oszillationen der numerischen Lösung. Dabei ist ν eine untere Schranke für den kleinsten Eigenwert der Diffusionsmatrix $A(x)$. Dieses unerwünschte Verhalten rührt daher, dass diese Diskretisierung einer zentralen Differenzendiskretisierung des Konvektionstermes gleicht. Mögliche Auswege sind:

- Upwind-Verfahren,
- Petrov-Galerkin-Verfahren.

Bei den *Upwind-Verfahren* approximiert man beim Aufstellen der Steifigkeitsmatrix die Konvektionsterme durch eine Quadraturformel

$$\int_K \lambda_{z',k}\mathbf{a}\cdot\nabla\lambda_{z,k} \approx \sum_{q\in\mathcal{Q}_K} c_q \lambda_{z',k}(q)\mathbf{a}(q)\cdot\nabla\lambda_{z,k}(q)$$

und ersetzt $\mathbf{a}(q)\cdot\nabla\lambda_{z,k}(q)$ durch einen *upwind-Differenzenquotienten* wie bei Differenzenverfahren. Diese Vorgehensweise hat u.a. folgende Nachteile:

- Der Ansatz führt häufig zu einem Genauigkeitsverlust.
- Wegen der nötigen Interpolation zwischen Elementfreiheitsgraden treten häufig neue unphysikalischen Oszillationen auf.
- Bei nichtlinearen Differentialgleichungen hängt das resultierende diskrete Problem nicht mehr differenzierbar von der Lösung ab, so dass es nicht mit einem Newton-Verfahren gelöst werden kann.

Diese Nachteile werden bei den *Petrov-Galerkin-Verfahren* vermieden. Die Grundidee dieser Verfahren ist es, die Differentialgleichung zusätzlich *elementweise* mit $\mathbf{a}\cdot\nabla v_{\mathcal{T}}$ zu testen. Dies führt auf folgendes diskretes Problem

Finde $u_{\mathcal{T}} \in S_0^{k,0}(\mathcal{T})$ *(Ansatzfunktion)* so, dass für alle $v_{\mathcal{T}} \in S_0^{k,0}(\mathcal{T})$ *(Testfunktion)* gilt

$$\begin{aligned}
&\int_\Omega \big[\nabla v_{\mathcal{T}}\cdot A\nabla u_{\mathcal{T}} + v_{\mathcal{T}}\mathbf{a}\cdot\nabla u_{\mathcal{T}} + \alpha u_{\mathcal{T}} v_{\mathcal{T}}\big]dx \\
&\quad + \sum_{K\in\mathcal{T}} \delta_K h_K \int_K \big[-\operatorname{div}(A\nabla u_{\mathcal{T}}) + \mathbf{a}\cdot\nabla u_{\mathcal{T}} + \alpha u_{\mathcal{T}}\big]\mathbf{a}\cdot\nabla v_{\mathcal{T}} dx \\
&= \int_\Omega f v_{\mathcal{T}} dx + \sum_{K\in\mathcal{T}} \delta_K h_K \int_K f\mathbf{a}\cdot\nabla v_{\mathcal{T}} dx.
\end{aligned}$$

Dabei ist h_K der Durchmesser von K und $\delta_K > 0$ ein Stabilitätsparameter. Eine bewährte Wahl ist

$$\delta_K = \frac{|\mathbf{a}|h_K}{\sqrt{\nu^2 + |\mathbf{a}|^2 h_K^2}}.$$

Die Petrov-Galerkin-Verfahren haben folgende Eigenschaften:

> Das diskrete Problem besitzt eine eindeutige Lösung.
> Unphysikalische Oszillationen werden weitest gehend vermieden.
> Es tritt kein Genauigkeitsverlust ein und es gelten die gleichen Fehlerabschätzungen wie bei der üblichen Diskretisierung.
> Zusätzlich gilt eine analoge Fehlerabschätzung für den Fehler $\|\mathbf{a} \cdot \nabla u - \mathbf{a} \cdot \nabla u_{\mathcal{T}}\|$ der Konvektionsableitung.
> Bei nichtlinearen Differentialgleichungen hängt das diskrete Problem differenzierbar von der Lösung ab, so dass das Newton-Verfahren zur Lösung des diskreten Problems benutzt werden kann.

IV.5. Gemischte Finite-Element-Methoden

Die bisher betrachteten Finite-Element-Methoden beruhen alle auf einem sog. *primalen Variationsproblem* und werden daher auch *primale Methoden* genannt. Sie liefern eine Approximation für die Lösung u der Differentialgleichung; bei Elastizitätsproblemen ist dies die Verschiebung. Häufig sind jedoch abgeleitete Größen wie ∇u physikalisch interessanter; bei Elastizitätsproblemen sind dies die Spannungen. Diese abgeleiteten Größen müssen nachträglich durch numerische Differentiation bestimmt werden und werden dadurch weniger genau approximiert. Bei Elastizitätsproblemen führt der klassische Ansatz, die sog. *Verschiebungsmethode*, zu unphysikalischen Lösungen, dem sog. *locking*. Gesucht sind daher Variationsformulierungen und Finite-Element-Diskretisierungen, die Verschiebungen und Spannungen gleichzeitig mit gleicher Genauigkeit approximieren.

Dies leisten sog. *gemischte Finite-Element-Methoden*, die auf einem sog. *dualen Variationsproblem* oder *Zwei-Energien-Prinzip* beruhen.

Um die Idee mit möglichst geringem notationellem und technischen Aufwand zu erläutern, betrachten wir die Poisson-Gleichung

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{auf } \Gamma \end{aligned}$$

und führen ∇u als zusätzliche Variable ein. Dies liefert folgendes *Differentialgleichungssystem erster Ordnung*

$$\begin{aligned} \sigma - \nabla u &= 0 \quad \text{in } \Omega \\ -\operatorname{div} \sigma &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{auf } \Gamma. \end{aligned}$$

Die Variationsformulierung dieses Systems beruht auf folgender Idee:

- Multipliziere die erste Gleichung mit einem differenzierbaren Vektorfeld $\tau : \Omega \to \mathbb{R}^d$ und integriere das Ergebnis über Ω
$$\int_\Omega \sigma \cdot \tau dx - \int_\Omega \nabla u \cdot \tau dx = 0.$$
- Integriere den ∇u-Term partiell
$$\int_\Omega \nabla u \cdot \tau dx = -\int_\Omega u \operatorname{div} \tau dx.$$
- Multipliziere die zweite Gleichung mit einer Funktion $v : \Omega \to \mathbb{R}$ und integriere das Ergebnis über Ω
$$-\int_\Omega \operatorname{div} \sigma v dx = \int_\Omega f v dx.$$

Damit diese Idee auf ein sinnvolles Variationsproblem führt, müssen wir die Differenzierbarkeitsanforderungen an die Vektorfelder σ und τ geeignet abschwächen. Dies leistet der Raum

$$H(\operatorname{div}, \Omega) = \{\sigma \in L^2(\Omega)^d : \operatorname{div} \sigma \in L^2(\Omega)\}.$$

Er liegt gewissermaßen zwischen den L^2- und H^1-Räumen. Für die spätere Diskretisierung ist folgende Eigenschaft wichtig: Ein stückweise differenzierbares Vektorfeld σ ist genau dann in $H(\operatorname{div}, \Omega)$, wenn seine Normalkomponente $\sigma \cdot \mathbf{n}$ stetig ist über die Grenzen der Teilgebiete (vgl. Abbildung IV.1.1).

Damit erhalten wir das folgende Variationsproblem:

Finde $u \in L^2(\Omega)$ und $\sigma \in H(\operatorname{div}, \Omega)$, so dass für alle $v \in L^2(\Omega)$ und alle $\tau \in H(\operatorname{div}, \Omega)$ gilt
$$\int_\Omega \sigma \cdot \tau dx + \int_\Omega u \operatorname{div} \tau dx = 0$$
$$-\int_\Omega v \operatorname{div} \sigma dx = \int_\Omega f v dx$$

Es hat folgende Eigenschaften:

Das Variationsproblem hat eine eindeutige Lösung (σ, u). Die sog. *primale Variable* σ ist das eindeutige Minimum der *Energiefunktion*
$$\frac{1}{2}\int_\Omega \sigma : \sigma dx$$
unter der *Nebenbedingung* $-\operatorname{div} \sigma = f$ (sog. *Gleichgewichtsbedingung*); die sog. *duale Variable* u ist der zugehörige *Lagrange-Multiplikator*.

Die Lösung (σ, u) ist der eindeutige *Sattelpunkt* der Funktion
$$\mathcal{L}(\tau, v) = \frac{1}{2}\int_\Omega \tau : \tau dx + \int_\Omega v \operatorname{div} \tau dx;$$
σ minimiert $\mathcal{L}$ bei festem v, u maximiert $\mathcal{L}$ bei festem τ.

Für die *gemischte Finite-Element-Diskretisierung* dieses Variationsproblemes wählen wir eine zulässige Unterteilung $\mathcal{T}$ von Ω und endlichdimensionale Teilräume $X(\mathcal{T})$ von $H(\operatorname{div}, \Omega)$ und $Y(\mathcal{T})$ von $L^2(\Omega)$. Dann lautet das diskrete Problem:

Finde $u_\mathcal{T} \in Y(\mathcal{T})$ und $\sigma_\mathcal{T} \in X(\mathcal{T})$, so dass für alle $v_\mathcal{T} \in Y(\mathcal{T})$ und alle $\tau_\mathcal{T} \in X(\mathcal{T})$ gilt
$$\int_\Omega \sigma_\mathcal{T} \cdot \tau_\mathcal{T} dx + \int_\Omega u_\mathcal{T} \operatorname{div} \tau_\mathcal{T} dx = 0$$
$$-\int_\Omega v_\mathcal{T} \operatorname{div} \sigma_\mathcal{T} dx = \int_\Omega f v_\mathcal{T} dx$$

Es hat folgende Eigenschaften:

Die gemischte Finite-Element-Diskretisierung führt nach Wahl von Basen für $X(\mathcal{T})$ und $Y(\mathcal{T})$ auf ein lineares Gleichungssystem; Unbekannte sind die Koeffizienten von $\sigma_\mathcal{T}$ und $u_\mathcal{T}$.
Die Steifigkeitsmatrix hat die Form
$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix}$$
mit einer symmetrischen, positiv definiten Matrix A und einer *rechteckigen* Matrix B.
Die Steifigkeitsmatrix ist *indefinit* und hat positive und negative Eigenwerte.
Die Steifigkeitsmatrix ist genau dann invertierbar, wenn $B^T A^{-1} B$ invertierbar ist.
Um optimale Fehlerabschätzungen zu erhalten, muss der kleinste Eigenwert der Matrix $B^T A^{-1} B$ unabhängig von $\mathcal{T}$ von Null weg beschränkt sein *(inf-sup-Bedingung)*.

Nicht jede auf den ersten Blick plausible Wahl von $X(\mathcal{T})$ und $Y(\mathcal{T})$ erfüllt die inf-sup-Bedingung. Das einfachste Beispiel für eine Diskretisierung, die diese Bedingung erfüllt, ist die *Raviart-Thomas-Diskretisierung*, sog. *Raviart-Thomas-Element*. Bei ihr besteht $\mathcal{T}$ aus Dreiecken

oder Tetraedern. Die Verschiebungen werden stückweise konstant approximiert

$$Y(\mathcal{T}) = S^{0,-1}(\mathcal{T}).$$

Die Spannungen werden stückweise linear approximiert, wobei die Freiheitsgrade die Normalkomponenten in den Kantenmittelpunkten für $d = 2$ und in den Schwerpunkten der Seitenflächen für $d = 3$ sind (vgl. Abbildung IV.5.1)

$$X(\mathcal{T}) = \{\tau : \tau\big|_K \in RT(K) \text{ für alle } K \in \mathcal{T},\\ \tau \cdot \mathbf{n} \text{ ist stetig in den Kanten- bzw. Flächenmitten}\}$$

mit

$$RT(K) = \{\mathbf{a} + bx : \mathbf{a} \in \mathbb{R}^d,\, b \in \mathbb{R}\}.$$

Die Raviart-Thomas-Diskretisierung hat folgende Eigenschaften:

> Die Raviart-Thomas Diskretisierung erfüllt die inf-sup Bedingung.
> Es gelten die Fehlerabschätzungen
>
> $$\|\sigma - \sigma_{\mathcal{T}}\| + \|\operatorname{div}\sigma - \operatorname{div}\sigma_{\mathcal{T}}\| + \|u - u_{\mathcal{T}}\| \le ch_{\mathcal{T}}.$$
>
> Die Konstante c hängt von Ableitungen erster Ordnung von f, u und σ ab.
> Es gibt analoge Diskretisierungen höherer Ordnung und solche für Unterteilungen in Parallelogramme und Parallelepipede.

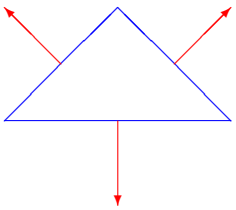


ABBILDUNG IV.5.1. Elementfreiheitsgrade des Raviart-Thomas-Elementes

Bei Anwendung der beschriebenen Ideen auf die Gleichungen der *linearen Elastizitätstheorie* entspricht u dem Verschiebungsvektor und σ dem Spannungstensor. Eine zusätzliche Schwierigkeit bereitet die Symmetrie des Spannungstensors. Für das Variationsproblem ist sie automatisch erfüllt. Wird sie für das diskrete Problem exakt gefordert, treten Stabilitätsprobleme auf. Daher kann die Symmetrie für das diskrete Problem nur in abgeschwächter Form gefordert werden (*PEERS*: *p*lane *e*lasticity *e*lements with *r*educed *s*ymmetry).

In der *Strömungsmechanik* treten vorwiegend Variationsprobleme mit Divergenz-Nebenbedingungen auf, d.h. die Divergenz eines Vektorfeldes wie der Geschwindigkeit muss verschwinden *(Inkompressibilität)*. Diese Nebenbedingung macht die Nutzung gemischter Finiter-Element-Methoden zwingend erforderlich. Bei den resultierenden Variationsproblemen sind die Rollen von u und σ vertauscht; u ist die primale Variable, σ ist der Lagrange-Multiplikator zur Nebenbedingung. Dabei entspricht u der Geschwindigkeit, σ dem Druck.

KAPITEL V

A posteriori Fehlerschätzung und Adaptivität

V.1. Motivation

In den Kapiteln III und IV haben wir stets *a priori Fehlerabschätzungen* betrachtet. Wie dort schon angedeutet haben diese einige Nachteile:

- Sie liefern nur eine *asymptotische* Aussage über den Fehler, d.h. über das Verhalten für immer feiner werdende Unterteilungen.
- Sie geben keine Auskunft über die tatsächliche Größe des Fehlers.
- Sie erlauben keine Rückschlüsse über die räumliche (und zeitliche) Verteilung des Fehlers.

Viele praktische Probleme weisen aber *lokale* Singularitäten aufgrund von z.B. einspringenden Ecken, inneren Grenzschichten oder Randgrenzschichten auf, die die *globale* Genauigkeit der Diskretisierung reduzieren (vgl. Abbildung V.1.1). Dieser negative Effekt kann nur vermieden werden, indem in der Nähe der Singularitäten das Gitter *lokal* feiner gewählt wird (vgl. Tabellen V.1.1, V.1.2 und Beispiele V.1.2, V.1.3). Leider ist aber die Größe und Lage dieser Singularitäten Teil der zu berechnenden Lösung und nicht a priori bekannt.

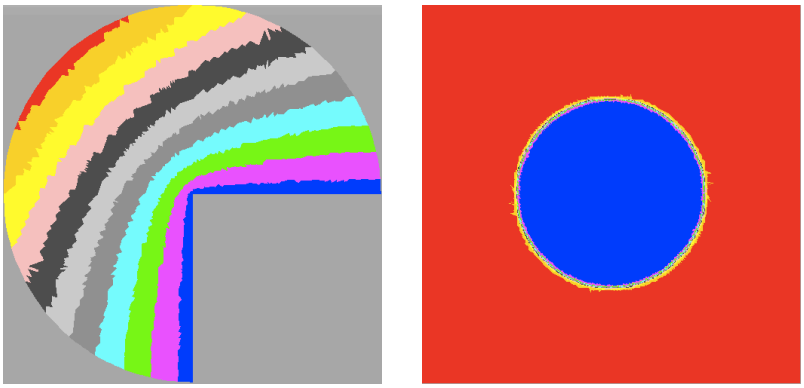


ABBILDUNG V.1.1. Lösung mit Singularität in der Nähe einer einspringenden Ecke (links, vgl. Beispiel V.1.2) und mit innerer Grenzschicht (rechts, vgl. Beispiel V.1.3; rot entspricht dem Wert 1, blau dem Wert -1)

Die Methoden dieses Kapitels haben dementsprechend die folgenden Ziele:

- Liefere eine leicht berechenbare und zuverlässige Information über die tatsächliche Größe des Fehlers und seine räumliche (und zeitliche) Verteilung.
- Bestimme eine Näherungslösung für die gegebene Differentialgleichung mit vorgegebener Toleranz und (nahezu) minimalem Aufwand (z.B. gemessen in der Zahl der Freiheitsgrade).

Grundlage ist der folgende Algorithmus:

Algorithmus V.1.1. (Adaptiver Algorithmus)

(0) *Gegeben: Daten einer partiellen Differentialgleichung und eine Toleranz ε.*
Gesucht: eine numerische Näherungslösung mit einem Fehler kleiner oder gleich der Toleranz.
(1) *Konstruiere eine erste grobe Unterteilung $\mathcal{T}_0$, setze $k = 0$.*
(2) *Stelle das diskrete Problem zu $\mathcal{T}_k$ auf und löse es (näherungsweise).*
(3) *Bestimme für jedes Element K in $\mathcal{T}_k$ einen* Fehlerschätzer.
(4) *Falls der geschätzte Gesamtfehler kleiner oder gleich der Toleranz ε ist,* stopp.
(5) *Entscheide welche Elemente von $\mathcal{T}_k$ unterteilt werden müssen und konstruiere eine entsprechende neue Unterteilung $\mathcal{T}_{k+1}$. Erhöhe k um 1 und gehe zu Schritt* (2) *zurück.*

Wesentliche Ingredienzien von Algorithmus V.1.1 sind

- eine Diskretisierungsmethode (vgl. Kapitel III, IV),
- ein Lösungsverfahren für die diskreten Probleme (vgl. Kapitel VI),
- ein Fehlerschätzer (vgl. §V.2),
- eine Gitteranpassungsstrategie (vgl. §V.3).

Beispiel V.1.2. Betrachte die Poisson-Gleichung $-\Delta u = 0$ auf dem Kreissegment (vgl. Beispiel II.3.1)

$$\Omega = \{(r\cos\varphi, r\sin\varphi) : 0 \le r < 1,\ 0 \le \varphi \le \frac{3\pi}{2}\}$$

mit Randbedingung $u = \sin(\frac{2}{3}\varphi)$ auf dem Kreisbogen und $u = 0$ auf den geraden Randstücken. Die Lösung ist

$$u = r^{\frac{2}{3}}\sin(\frac{2}{3}\varphi).$$

Der linke Teil von Abbildung V.1.1 zeigt einen Farbplot der Lösung. Abbildung V.1.2 zeigt für dieses Problem im linken Teil eine gleichmäßig verfeinerte und im rechten Teil eine mit den Methoden dieses Kapitels erzeugte lokal verfeinerte Unterteilung in Dreiecke. Tabelle V.1.1

gibt für gleichmäßig verfeinerte und adaptiv lokal verfeinerte Unterteilungen in Dreiecke die Zahl der Unbekannten, die Zahl der Elemente und den relativen Fehler $\|\nabla u - \nabla u_{\mathcal{T}}\|/\|\nabla u\|$ der zugehörigen Finite-Element-Approximation in $S_0^{1,0}(\mathcal{T})$ wieder.

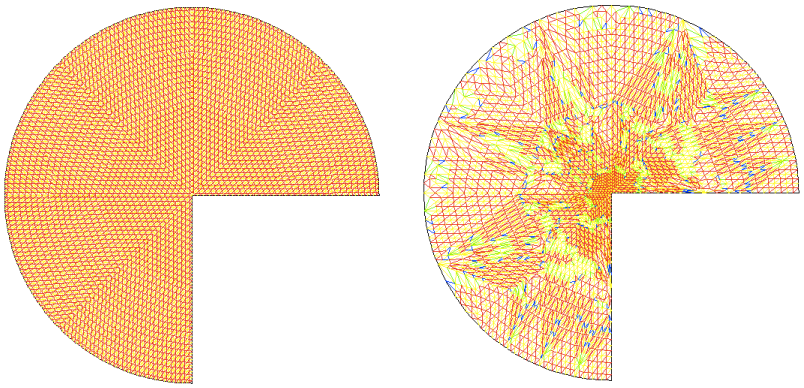


ABBILDUNG V.1.2. Gleichmäßig verfeinerte (links) und lokal verfeinerte (rechts) Unterteilung in Dreiecke für Beispiel V.1.2

TABELLE V.1.1. Zahl der Unbekannten, Zahl der Elemente und relativer Fehler $\|\nabla u - \nabla u_{\mathcal{T}}\|/\|\nabla u\|$ für gleichmßig verfeinerte und lokal verfeinerte Unterteilungen in Dreiecke für Beispiel V.1.2

Verfeinerung	Elemente	Unbekannte	rel. Fehler
gleichmäßig	24576	12033	0.5%
adaptiv	11242	5529	0.5%

Beispiel V.1.3. Betrachte die Reaktions-Diffusions-Gleichung

$$\begin{aligned} -\Delta u + 100u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{auf } \Gamma \end{aligned}$$

in einem Quadrat mit Kantenlänge 2 und Mittelpunkt im Ursprung. Die Funktion f ist so gewählt, dass die Lösung eine innere Grenzschicht der Dicke ≈ 0.01 entlang des Kreises um den Ursprung mit Radius 1 hat. Der rechte Teil von Abbildung V.1.1 zeigt einen Farbplot der Lösung; rot entspricht dem Wert 1, blau dem Wert -1. Abbildung V.1.3 zeigt für dieses Problem mit den Methoden dieses Kapitels erzeugte lokal verfeinerte Unterteilungen in Dreiecke (links) und Vierecke (rechts). Tabelle V.1.2 gibt für gleichmäßig verfeinerte und adaptiv lokal verfeinerte Unterteilungen in Dreiecke und Vierecke die Zahl der Unbekannten, die Zahl der Elemente und den relativen Fehler $|||u - u_{\mathcal{T}}|||/|||u|||$ der zugehörigen Finite-Element-Approximation in $S_0^{1,0}(\mathcal{T})$ wieder. Dabei ist $|||v||| = \left\{\|\nabla v\|^2 + 100\|v\|^2\right\}^{\frac{1}{2}}$.

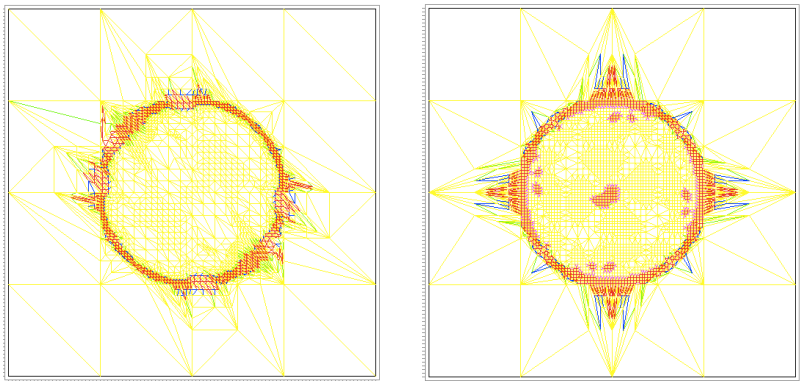


ABBILDUNG V.1.3. Lokal verfeinerte Unterteilung in Dreiecke (links) und Vierecke (rechts) für Beispiel V.1.3

TABELLE V.1.2. Zahl der Unbekannten, Zahl der Elemente und relativer Fehler $|||u - u_{\mathcal{T}}|||/|||u|||$ für gleichmäßig und lokal verfeinerte Unterteilungen in Dreiecke und Vierecke für Beispiel V.1.3

	Dreiecke		Vierecke	
	gleichm.	adaptiv	gleichm.	adaptiv
Unbekannte	16129	2923	16129	4722
Dreiecke	32768	5860	0	3830
Vierecke	0	0	16384	2814
rel. Fehler	3.8%	3.5%	6.1%	4.4%

V.2. A posteriori Fehlerschätzer

Wir betrachten eine allgemeine lineare elliptische Differentialgleichung zweiter Ordnung

$$\begin{aligned} -\operatorname{div}(A\nabla u) + \mathbf{a}\cdot\nabla u + \alpha u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{auf } \Gamma \end{aligned}$$

und bezeichnen mit u die Lösung des zugehörigen Variationsproblems und mit $u_{\mathcal{T}}$ die *berechnete* Lösung des diskreten Problems basierend auf der klassischen Finite-Element-Methode oder einer Petrov-Galerkin-Diskretisierung.

Für den Fehlerschätzer benötigen wir einige zusätzliche Notationen:

- $\mathcal{E}_K$ ist die Menge der Kanten ($d = 2$) bzw. Seitenflächen ($d = 3$) eines Elementes K,
- h_K und h_E bezeichnen den Durchmesser eines Elementes K bzw. einer Kante oder Seitenfläche E von K,
- $\mathbf{n}_E$ ist ein Einheitsvektor senkrecht zu E,

- $[\varphi]_E$ bezeichnet den Sprung einer Funktion φ über E in Richtung $\mathbf{n}_E$.

Man beachte, dass $[\varphi]_E$ von der Orientierung von $\mathbf{n}_E$ abhängt, dass aber Größen wie $[\mathbf{n}_E \cdot \nabla\varphi]_E$ von dieser Orientierung unabhängig sind.

Mit diesen Bezeichnungen definieren wir den *residuellen Fehlerschätzer* η_K durch

$$\eta_K = \Big\{ h_K^2 \int_K |f + \operatorname{div}(A\nabla u_\mathcal{T}) - \mathbf{a}\cdot\nabla u_\mathcal{T} - \alpha u_\mathcal{T}|^2 dx + \sum_{E\in\mathcal{E}_K\setminus\Gamma} h_E \int_E \big|\big[\mathbf{n}_E \cdot A\nabla u_\mathcal{T}\big]_E\big|^2 dS \Big\}^{\frac{1}{2}}.$$

Man kann für diesen Fehlerschätzer folgende *a posteriori Fehlerabschätzungen* beweisen

$$\Big\{\int_\Omega |\nabla u - \nabla u_\mathcal{T}|^2 dx\Big\}^{\frac{1}{2}} \le c^* \Big\{\sum_{K\in\mathcal{T}} \eta_K^2\Big\}^{\frac{1}{2}},$$
$$\eta_K \le c_* \Big\{\int_{\omega_K} |\nabla u - \nabla u_\mathcal{T}|^2 dx\Big\}^{\frac{1}{2}}.$$

Dabei ist ω_K die Vereinigung aller Elemente, die mit K eine Kante ($d = 2$) bzw. Seitenfläche ($d = 3$) gemeinsam haben (vgl. Abbildung V.2.1). Die Konstanten c_* und c^* hängen von der relativen Größe von Diffusion A, Konvektion $\mathbf{a}$ und Reaktion α zueinander, dem Polynomgrad von $u_\mathcal{T}$ und dem Formparameter von $\mathcal{T}$ (maximales Verhältnis von Elementdurchmesser zum Durchmesser des größten eingeschriebenen Balles) ab.

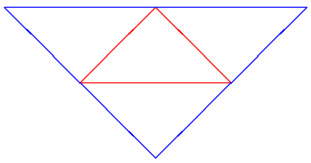


ABBILDUNG V.2.1. Menge ω_K

Der Fehlerschätzer η_K hat folgende Struktur:

Die Größe $f + \operatorname{div}(A\nabla u_\mathcal{T}) - \mathbf{a}\cdot\nabla u_\mathcal{T} - \alpha u_\mathcal{T}$, genannt *Elementresiduum*, ist das Residuum auf K von $u_\mathcal{T}$ bzgl. der Differentialgleichung.
Die Größe $\big[\mathbf{n}_E \cdot A\nabla u_\mathcal{T}\big]_E$, genannt *Kantenresiduum*, ist der Sprung über E des Operators, der starke und schwache

Form der Differentialgleichung verbindet, dies ist der Randterm bei der partiellen Integration.
Für Elemente niedriger Ordnung, z.B. $k = 1$, ist das Kantenresiduum häufig entscheidend.

Obige a posteriori Fehlerabschätzung hat folgende wesentlichen Eigenschaften:

Die obere Schranke ist *global*. Dies liegt daran, dass sie auf Eigenschaften des Lösungsoperators der Differentialgleichung beruht. (Lokale Last impliziert globale Verschiebung.)
Die untere Schranke ist *lokal*. Dies liegt daran, dass sie auf Eigenschaften des Differentialoperators beruht. (Lokale Verschiebung impliziert lokale Last.)

Neben dem vorgestellten residuellen Fehlerschätzer gibt es einen ganzen Zoo weiterer Schätzer. Die wichtigsten Exemplare sind:

- *Lokale Hilfsprobleme:* Löse lokale diskrete Hilfsprobleme höherer Ordnung mit Element- und Kantenresiduen als Lasttermen.
- *Hierarchische Schätzer:* Vergleiche die aktuelle Lösung mit einer Finite Element Lösung höherer Ordnung basierend auf einem Lumping der Steifigkeitsmatrix.
- *ZZ-Schätzer:* Vergleiche den Gradienten der Lösung mit einer Mittelung des Gradienten.

Alle diese Schätzer sind äquivalent in dem Sinne, dass sie bis auf multiplikative Konstanten gegeneinander abgeschätzt werden können.

V.3. Gitteranpassung

Die *Gitteranpassung* beruht auf zwei Säulen:

- *Markierungsstrategien*, die festlegen, welche Elemente unterteilt werden sollen,
- *Verfeinerungsregeln*, die festlegen, wie ein einzelnes Element unterteilt werden soll.

Um die Zulässigkeit der Unterteilung zu gewährleisten und *hängende Knoten* (vgl. Abbildung V.3.2) zu vermeiden, erfolgt die Verfeinerung in zwei Etappen:

- Zuerst werden alle Elemente unterteilt, die einen zu großen Wert des Fehlerschätzers η_K aufweisen *(reguläre Unterteilung)* (vgl. Abbildung V.3.1).
- Danach werden zusätzliche Elemente unterteilt, um hängende Knoten zu beseitigen, die während der ersten Etappe erzeugt wurden *(irreguläre Unterteilung)* (vgl. Abbildung V.3.3).

Die Gitterverfeinerung kann zudem mit einer *Gittervergröberung* und *Gitterglättung* kombiniert werden.

Die einfachste Markierungsstrategie ist die *Maximum-Strategie*.

Algorithmus V.3.1. (Maximum-Strategie)

(0) *Gegeben: Unterteilung* $\mathcal{T}$*, Fehlerschätzer* η_K *für die Elemente* $K \in \mathcal{T}$*, Schwellenwert* $\theta \in (0,1)$*.*
Gesucht: Teilmenge $\widetilde{\mathcal{T}}$ *von markierten Elementen, die unterteilt werden sollen.*

(1) *Berechne*

$$\eta_{\mathcal{T},\max} = \max_{K\in\mathcal{T}} \eta_K.$$

(2) *Falls* $\eta_K \geq \theta\eta_{\mathcal{T},\max}$ *ist, markiere* K *und füge es zu* $\widetilde{\mathcal{T}}$ *hinzu.*

Manchmal verteilen sich die Elemente auf drei Gruppen:

- eine sehr kleine Zahl von Elementen mit einem sehr großen geschätzten Fehler,
- eine sehr große Zahl von Elemente mit einem sehr kleinen geschätzten Fehler,
- eine mittlere Zahl von Elementen mit einem geschätzten Fehler mittlerer Größe.

Dann markiert die Maximum-Strategie nur die Elemente der ersten Gruppe. Dies verschlechtert die Effizienz des adaptiven Algorithmus. Dies kann durch folgende Modifikation vermieden werden:

> Markiere zuerst einen kleinen Prozentsatz ε der Elemente mit dem größten geschätzten Fehler und wende danach die Maximum-Strategie auf die verbleibenden Elemente an.

Für die *reguläre Unterteilung* verbindet man die Kantenmittelpunkte der Elemente (vgl. Abbildung V.3.1). Dies erhält den Formparameter (Verhältnis des Elementdurchmessers zum Durchmesser des größten eingeschriebenen Balles).

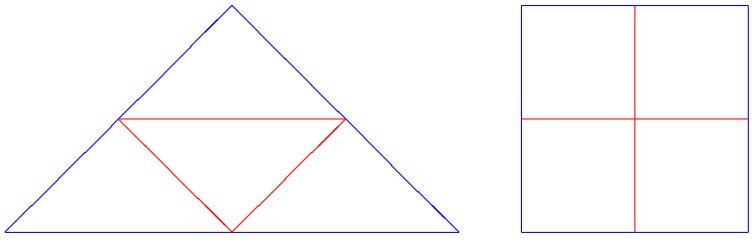


ABBILDUNG V.3.1. Reguläre Unterteilung eines Dreieckes (links) und eines Viereckes (rechts)

Bei der regulären Unterteilung entstehen zwangsläufig *hängende Knoten*. Dies sind Punkte, die gleichzeitig Eckpunkte eines Elementes sind und auf der Kante eines benachbarten Elementes liegen, ohne

Eckpunkte dieses Elementes zu sein (vgl. Abbildung V.3.2). Hängende Knoten zerstören die Zulässigkeit der Unterteilung. Wegen der Stetigkeitsanforderungen an die Finite-Element-Funktionen können sie keine Freiheitsgrade sein. Daher muss man diese Punkte entweder durch zusätzliche irreguläre Unterteilungen auflösen oder die Stetigkeit der Finite-Element-Funktionen künstlich erzwingen. Wie der rechte Teil von Abbildung V.3.2 zeigt, kann letzteres aber der zuvor erfolgten regulären Unterteilung zuwiderlaufen.

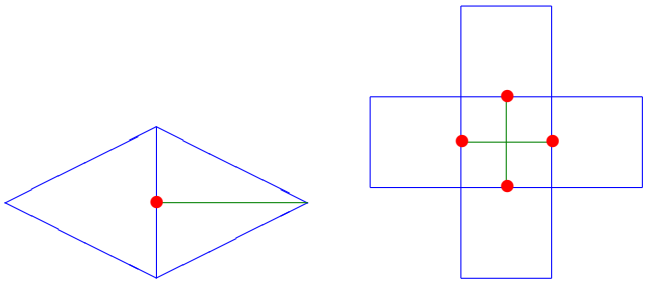


ABBILDUNG V.3.2. Hängende Knoten

Abbildung V.3.3 zeigt die zusätzliche *irreguläre Unterteilung* von Dreiecken und Vierecken zum Auflösen hängender Knoten. Analoge Vorschriften gelten in drei Dimensionen für Tetraeder und Parallelepipede.

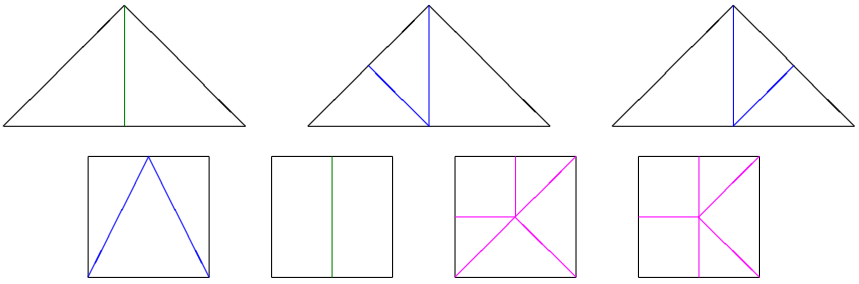


ABBILDUNG V.3.3. Irreguläre Unterteilung von Dreiecken (oben) und Vierecken (unten) zur Auflösung hängender Knoten

Eine Alternative zu der beschriebenen Kombination aus regulärer und irregulärer Unterteilung ist die *Bisektion markierter Kanten* (engl. *marked edge bisection*). Dabei geht man wie folgt vor:

- Konstruiere die erste Unterteilung so, dass die längste Kante eines jeden Elementes auch die längste Kante des angrenzenden Elementes ist.
- Markiere die längsten Kanten in der ersten Unterteilung.

- Unterteile ein Element durch Verbinden des Mittelpunktes seiner markierten Kante mit dem gegenüberliegenden Eckpunkt *(Bisektion)*.
- Bei Unterteilung einer Kante eines Elementes werden die anderen Kanten die markierten Kanten der resultierenden neuen Elemente.

Abbildung V.3.4 zeigt das Ergebnis von vier aufeinander folgenden Bisektionsschritten.

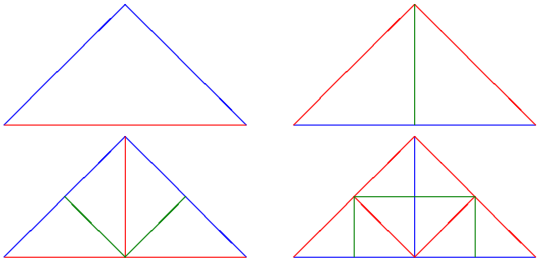


Abbildung V.3.4. Vier Schritte der Bisektion markierter Kanten (jeweils rot)

Eine *Vergröberung* von Unterteilungen ist erforderlich, um die Optimalität des adaptiven Algorithmus zu gewährleisten, d.h. um eine gegebene Toleranz mit einer minimalen Anzahl an Freiheitsgraden zu erreichen, und um zeitlich veränderliche Singularitäten zu erfassen. Sie erfolgt durch das Zusammenfassen von benachbarten Elementen mit zu kleinem Fehler.

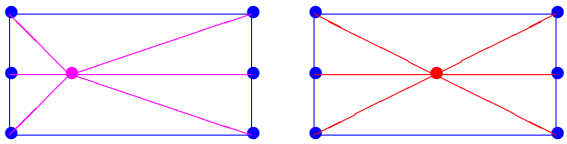


Abbildung V.3.5. Ergebnis einer Gitterglättung

Wenn eine weitere Gitterverfeinerung nicht mehr möglich ist, weil z.B. der verfügbare Speicherplatz ausgeschöpft ist, kann die Qualität der Diskretisierung u.U. noch durch eine *Gitterglättung* verbessert werden. Dabei wird die *Qualität* einer Unterteilung $\mathcal{T}$ durch *Verschieben* der Elementeckpunkte unter Beibehaltung der Nachbarschaftsbeziehungen verbessert. Die Qualität wird dabei durch eine *Qualitätsfunktion* q gemessen, wobei ein größerer Wert von q einer besseren Qualität entspricht. Die Qualität wird durch einen Gauß-Seidel artigen *Glättungsprozess* verbessert:

Durchlaufe alle Elementeckpunkte z in $\mathcal{T}$, fixiere alle durch eine Kante mit z verbundenen Elementeckpunkte und suche einen neuen Punkt $\widetilde{z}$ mit

$$\min_{\widetilde{z}\in\widetilde{K}} q(\widetilde{K}) > \min_{z\in K} q(K).$$

Abbildung V.3.5 zeigt das Ergebnis dieses Prozesses für einen Punkt z.

KAPITEL VI

Effiziente Löser

VI.1. Motivation

Als typisches Beispiel für die linearen Gleichungsssysteme, die bei der Diskretisierung partieller Differentialgleichungen zu lösen sind, betrachten wir die Diskretisierung der Poisson-Gleichung

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{auf } \Gamma \end{aligned}$$

im Einheitsquadrat $\Omega = (0,1)^2$ durch lineare Dreieckselemente zu einer *Courant-Triangulierung*, die aus $2n^2$ rechtwinklig gleichschenkligen Dreiecken mit Katheten der Länge $h = n^{-1}$ besteht (vgl. Abbildung VI.1.1).

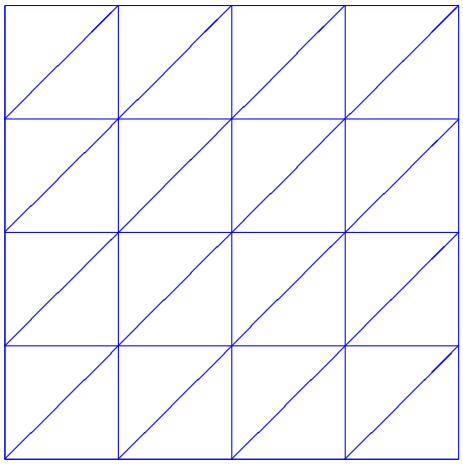


Abbildung VI.1.1. Courant-Triangulierung des Einheitsquadrates

Die Zahl N der Unbekannten ist $\approx n^2 = h^{-2}$. Die Steifigkeitsmatrix ist symmetrisch, positiv definit und hat höchstens 5 von Null verschiedene Elemente pro Zeile, d.h. sie ist *dünn besetzt*. Zudem hat sie eine *Bandstruktur* mit einer *Bandbreite* $\approx h^{-1} \approx N^{\frac{1}{2}}$. Daher erfordert das Gaußsche Eliminationsverfahren $\approx N^2$ Operationen. Im Vergleich dazu benötigt eine Matrix-Vektor-Multiplikation nur $5N$ Operationen. In Hinblick auf die iterativen Löser dieses Kapitels halten wir schließlich fest, dass der kleinste und der größte Eigenwert ≈ 1 bzw. $\approx h^{-2} \approx N$ sind. Damit ist die *Kondition* der Steifigkeitsmatrix $\approx h^{-2} \approx N$.

Dieses Beispiel zeigt die typischen Eigenschaften direkter Löser zur Lösung der bei der Diskretisierung partieller Differentialgeichung auftretenden linearen Gleichungssysteme (vgl. Tabelle III.1.1 (S. 43)):

- Für ein diskretes Problem mit N Unbekannten in d Raumdimensionen benötigen sie $\approx N^{2-\frac{1}{d}}$ Speicherplätze.
- Sie erfordern $\approx N^{3-\frac{2}{d}}$ arithmetische Operationen.
- Bis auf Rundungsfehler liefern sie die exakte Lösung des diskreten Problems.
- Sie liefern eine Approximation für die Lösung der Differentialgleichung mit einem Fehler $\approx h^\alpha \approx N^{-\frac{\alpha}{d}}$. Dabei ist typischerweise $\alpha \in \{1, 2\}$.

Im Vergleich hierzu haben klassische iterativer Löser folgende typischen Eigenschaften:

- Sie benötigen $\approx N$ Speicherplätze.
- Sie erfordern $\approx N$ arithmetische Operationen *pro Iteration*.
- Ihre *Konvergenzrate*, d.h. der Faktor um den der Fehler pro Iteration reduziert wird, verschlechtert sich mit wachsender Kondition des diskreten Problems, die typischerweise $\approx h^{-2} \approx N^{\frac{2}{d}}$ ist.
- Um einen Anfangsfehler um den Faktor 0.1 zu reduzieren benötigt man typischerweise folgende Anzahl arithmetischer Operationen:
  - $\approx N^{1+\frac{2}{d}}$ für den *Gauß-Seidel-Algorithmus*,
  - $\approx N^{1+\frac{1}{d}}$ für das *CG-Verfahren*,
  - $\approx N^{1+\frac{1}{2d}}$ für das *CG-Verfahren mit SSOR-Vorkonditionierung*.

Ein Vergleich dieser Eigenschaften führt zu folgenden Schlussfolgerungen:

- Direkte Verfahren erfordern zu viel Speicherplatz und Rechenzeit.
- Es ist vollkommen ausreichend, eine Näherungslösung des diskreten Problems zu bestimmen, die verglichen mit der exakten Lösung der Differentialgleichung eine ähnliche Genauigkeit hat wie die exakte Lösung des diskreten Problems.
- Iterative Löser sind überlegen, wenn es gelingt, ihre Konvergenzrate zu verbessern und eine gute Startnäherung zu finden.

VI.2. Geschachtelte Iteration

Häufig muss man eine *Folge diskreter Probleme*

$$L_k u_k = f_k$$

lösen, die zunehmend genaueren Diskretisierungen entsprechen. In der Regel kennt man dabei einen *Interpolationsoperator* $I_{k-1,k}$, der Funktionen der $(k-1)$-ten Diskretisierung in solche der k-ten Diskretisierung

abbildet. Die Interpolierende einer vernünftigen Näherungslösung des $(k-1)$-ten diskreten Problem ist dann ein guter Startwert für jeden iterativen Löser für das k-te diskrete Problem. Häufig muss dabei der Anfangsfehler nur um einen Faktor 0.1 reduziert werden.

Diese Beobachtungen sind Grundlage des folgenden Algorithmus:

Algorithmus VI.2.1. (Geschachtelte Iteration)

(1) *Berechne*
$$\widetilde{u}_0 = u_0 = L_0^{-1} f_0.$$
(2) *Für $k = 1, 2, \ldots$ berechne sukzessive eine Näherungslösung $\widetilde{u}_k$ für $u_k = L_k^{-1} f_k$ durch Anwenden von m_k Iterationen eines iterativen Lösers auf das Problem*
$$L_k u_k = f_k$$
mit Startnäherung $I_{k-1,k}\widetilde{u}_{k-1}$.

Dabei wird in Algorithmus VI.2.1 die Zahl m_k der Iterationen implizit durch das *Abbruchkriterium*
$$\|f_k - L_k\widetilde{u}_k\| \leq \varepsilon \|f_k - L_k(I_{k-1,k}\widetilde{u}_{k-1})\|$$
bestimmt.

VI.3. Klassische iterative Verfahren

In diesem und dem nächsten Abschnitt betrachten wir folgende Problemstellung: Zu lösen ist ein lineares Gleichungssystem
$$Lu = f$$
mit N Unbekannten und einer symmetrisch, positiv definiten Matrix L. Wir bezeichnen mit κ die *Kondition* von L, d.h. das Verhältnis des größten Eigenwertes von L zum kleinsten Eigenwert, und gehen davon aus, dass $\kappa \approx N^{\frac{2}{d}}$ ist.

Alle Verfahren dieses Abschnittes sind sog. *stationäre Iterationsverfahren* und haben folgende allgemeine Struktur:

Algorithmus VI.3.1. (Stationäres Iterationsverfahren)

(0) *Gegeben: Matrix L, rechte Seite f, Startnäherung u_0 und Toleranz ε.*
Gesucht: eine Näherungslösung für das lineare Gleichungssystem $Lu = f$.
(1) *Setze $i = 0$.*
(2) *Falls*
$$|Lu_i - f| < \varepsilon$$
ist, gebe u_i als Näherungslösung aus; stopp.
(3) *Berechne*
$$u_{i+1} = F(u_i; L, f)$$
erhöhe i um 1 und gehe zu Schritt (2) zurück.

Dabei ist $u \mapsto F(u; L, f)$ eine affine Abbildung, die sog. *Iterationsvorschrift*, die das jeweilige Verfahren charakterisiert. $|\cdot|$ ist irgendeine Norm auf auf dem $\mathbb{R}^N$ wie z.B. die Euklidische Norm.

Das einfachste klassische Iterationsverfahren zur Lösung linearer Gleichungssysteme ist die *Richardson-Iteration*. Die Iterationsvorschrift lautet

$$\boxed{u \mapsto u + \frac{1}{\omega}(f - Lu).}$$

Dabei ist ω ein *Dämpfungsparameter*, der die gleiche Größenordnung wie der größte Eigenwert von L haben muss. Die Konvergenzrate der Richardson-Iteration ist $\frac{\kappa-1}{\kappa+1} \approx 1 - N^{-\frac{2}{d}}$.

Eng verwandt zur Richardson-Iteration ist die *Jacobi-Iteration*. Ihre Iterationsvorschrift lautet

$$\boxed{u \mapsto u + D^{-1}(f - Lu).}$$

Dabei ist D ist die Diagonale von L. Die Konvergenzrate der Jacobi-Iteration ist wieder $\frac{\kappa-1}{\kappa+1} \approx 1 - N^{-\frac{2}{d}}$. Man beachte, dass bei der Jacobi-Iteration in jedem Iterationsschritt sukzessive alle Gleichungen durchlaufen werden und die i-te Gleichung exakt nach der i-ten Unbekannten aufgelöst wird, ohne dabei nachfolgende Gleichungen entsprechend anzupassen.

Die *Gauß-Seidel-Iteration* ist eng verwandt zum Jacobi-Verfahren. Sie beruht auf folgender Idee: In jedem Iterationsschritt werden sukzessive alle Gleichungen durchlaufen, die i-te Gleichung exakt nach der i-ten Unbekannten aufgelöst und – anders als beim Jacobi-Verfahren – das Ergebnis unmittelbar in alle nachfolgenden Gleichungen eingesetzt. Dies führt auf folgende Iterationsvorschrift:

$$\boxed{u \mapsto u + \mathcal{L}^{-1}(f - Lu).}$$

Dabei ist $\mathcal{L}$ ist der Teil von L unterhalb der Diagonalen einschließlich der Diagonalen. Die Konvergenzrate der Gauß-Seidel-Iteration ist wie bei den vorigen Verfahren $\frac{\kappa-1}{\kappa+1} \approx 1 - N^{-\frac{2}{d}}$.

VI.4. CG-Verfahren

Das *Konjugierte-Gradienten-Verfahren* oder *CG-Verfahren* beruht auf folgenden Ideen:

- Die Lösung des *symmetrischen, positiv definiten* Gleichungssystems

$$Lu = f$$

ist das eindeutige *Minimum* der quadratischen Funktion

$$J(u) = \frac{1}{2} u \cdot (Lu) - f \cdot u.$$

- Der negative Gradient
$$-\nabla J(v) = f - Lv$$
von J an der Stelle v ist die Richtung des steilsten Abstieges.
- J nimmt auf der Geraden $t \mapsto v + td$ sein eindeutiges Minimum an der Stelle
$$t^* = \frac{(f - Lv) \cdot d}{d \cdot (Ld)}$$
an.
- Bei sukzessiver Minimierung in Richtung der negativen Gradienten verlangsamt sich das Verfahren zunehmend, da die Suchrichtungen nahezu parallel werden.
- Das Verfahren wird durch Wahl *L-orthogonaler* Suchrichtungen mit
$$d_i \cdot (Ld_j) = 0 \quad \text{für } i \neq j$$
beschleunigt.
- L-orthogonale Suchrichtungen können während des Verfahrens rekursiv mit bestimmt werden.

Algorithmus VI.4.1. (CG-Verfahren)

(0) *Gegeben: Matrix L, rechte Seite f, Startnäherung u_0 und Toleranz ε.*
Gesucht: eine Näherungslösung für das lineare Gleichungssystem $Lu = f$.

(1) *Berechne*
$$\begin{aligned} r_0 &= f - Lu_0, \\ d_0 &= r_0, \\ \gamma_0 &= r_0 \cdot r_0. \end{aligned}$$
Setze $i = 0$.

(2) *Falls*
$$\gamma_i < \varepsilon^2$$
ist, gebe u_i als Näherungslösung aus; stopp.

(3) *Berechne*
$$\begin{aligned} s_i &= Ld_i, \\ \alpha_i &= \frac{\gamma_i}{d_i \cdot s_i}, \\ u_{i+1} &= u_i + \alpha_i d_i, \\ r_{i+1} &= r_i - \alpha_i s_i, \\ \gamma_{i+1} &= r_{i+1} \cdot r_{i+1}, \end{aligned}$$

$$\beta_i = \frac{\gamma_{i+1}}{\gamma_i},$$
$$d_{i+1} = r_{i+1} + \beta_i d_i.$$

Erhöhe i um 1 und gehe zu Schritt (2) zurück.

Das CG-Verfahren benötigt nur Skalarprodukte und Matrix-Vektor-Multiplikationen. Seine Konvergenzrate ist $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \approx 1 - N^{-\frac{1}{d}}$. Man beachte, dass das CG-Verfahren nur auf symmetrische, positiv definite Gleichungssystem angewendet werden kann und dass es für unsymmetrische oder indefinite Systeme in der Regel zusammenbricht (vgl. §VI.7).

Das CG-Verfahren kann durch eine geschickte *Vorkonditionierung* (engl. *pre-conditioning*) erheblich beschleunigt werden. Die Grundidee ist dabei folgende:

- Löse statt des ursprünglichen Systems
$$Lu = f$$
das äquivalente System
$$\widehat{L}\widehat{u} = \widehat{f}$$
mit
$$\widehat{L} = H^{-1}LH^{-t}, \quad \widehat{f} = H^{-1}f, \quad \widehat{u} = H^t u$$
und einer invertierbaren Matrix H.
- Wähle H so, dass:
  - die Kondition von $\widehat{L}$ wesentlich kleiner ist als diejenige von L,
  - Gleichungssysteme der Form $Cv = d$ mit $C = HH^t$ wesentlich leichter zu lösen sind als das ursprüngliche Problem $Lu = f$.
- Wende das CG-Verfahren auf das neue Gleichungssystem
$$\widehat{L}\widehat{u} = \widehat{f}$$
an und drücke alle Größen durch die ursprünglichen Daten L, f und u aus.

Man beachte, dass obige Bedingungen an H widersprüchlich sind. Die Wahl $H = L$ wäre in Hinblick auf die erste Bedingung optimal, aber unter dem Aspekt der zweiten Bedingung katastrophal. Umgekehrt wäre die Wahl $H = I$ mit der Einheitsmatrix I in Hinblick auf die zweite Bedingung optimal, würde aber für die erste Bedingung keinen Gewinn bringen.

Algorithmus VI.4.2. (PCG- *oder* vorkonditioniertes konjugiertes Gradienten-Verfahren)

(0) *Gegeben: Matrizen L und C, rechte Seite f, Startnäherung u_0 und Toleranz ε.*
Gesucht: eine Näherungslösung für das lineare Gleichungssystem $Lu = f$.

(1) *Berechne*

$$r_0 = f - Lu_0,$$

löse

$$Cz_0 = r_0$$

und berechne

$$d_0 = z_0,$$
$$\gamma_0 = r_0 \cdot z_0.$$

Setze $i = 0$.

(2) *Falls*

$$\gamma_i < \varepsilon^2$$

ist, gebe u_i als Näherungslösung aus; stopp.

(3) *Berechne*

$$s_i = Ld_i,$$
$$\alpha_i = \frac{\gamma_i}{d_i \cdot s_i},$$
$$u_{i+1} = u_i + \alpha_i d_i,$$
$$r_{i+1} = r_i - \alpha_i s_i,$$

löse

$$Cz_{i+1} = r_{i+1}$$

und berechne

$$\gamma_{i+1} = r_{i+1} \cdot z_{i+1},$$
$$\beta_i = \frac{\gamma_{i+1}}{\gamma_i},$$
$$d_{i+1} = z_{i+1} + \beta_i d_i.$$

Erhöhe i um 1 und gehe zu Schritt (2) zurück.

Die Konvergenzrate des PCG-Verfahrens ist $\frac{\sqrt{\widehat{\kappa}}-1}{\sqrt{\widehat{\kappa}}+1}$, wobei $\widehat{\kappa}$ die Kondition von $\widehat{L}$ ist. Bei geschickter Wahl von C, z.B. bei der *SSOR-Vorkonditionierung*, ist $\widehat{\kappa} = N^{\frac{1}{d}}$, was die Konvergenzrate $1 - N^{-\frac{1}{2d}}$ ergibt.

Algorithmus VI.4.3. (SSOR-Vorkonditionierung)

(0) *Gegeben: r und ein Relaxationsparameter $\omega \in (0, 2)$.*
Gesucht: $z = C^{-1}r$.

(1) *Setze*

$$z = 0.$$

(2) *Berechne für* $i = 1, \dots, N$

$$z_i = z_i + \omega L_{ii}^{-1} \Big\{ r_i - \sum_{j=1}^{N} L_{ij} z_j \Big\}.$$

(3) *Berechne für* $i = N, \dots, 1$

$$z_i = z_i + \omega L_{ii}^{-1} \Big\{ r_i - \sum_{j=1}^{N} L_{ij} z_j \Big\}.$$

Bis auf den Dämpfungsparameter entspricht die SSOR-Vorkonditionierung zwei Schritten der Gauß-Seidel-Iteration für das Gleichungssystem $Cz = r$ mit Startwert 0. Dabei werden im ersten Schritt die Gleichungen des Gleichungssystems in aufsteigender Reihenfolge und im zweiten Schritt in absteigender Reihenfolge durchlaufen.

VI.5. Mehrgitterverfahren

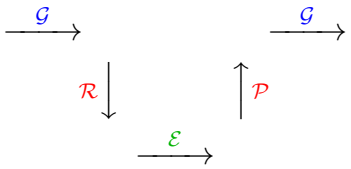
Eine genauere Analyse der klassischen iterativen Löser aus §VI.3 zeigt, dass sie schnell schwingende Fehlerkomponenten stark, langsam schwingende Fehlerkomponenten hingegen nur sehr schlecht dämpfen. Andererseits können langsam schwingende Fehlerkomponenten auf einem gröberen Gitter mit weniger Unbekannten gut approximiert werden.

Diese Beobachtung führt auf folgende Idee:

- Führe auf dem aktuellen Gitter mehrere Schritte eines klassischen iterativen Verfahrens durch.
- Korrigiere die aktuelle Näherungslösung wie folgt:
  - Berechne das aktuelle Residuum.
  - Schränke das aktuelle Residuum auf das nächst gröbere Gitter ein.
  - Löse das resultierende Problem auf dem gröberen Gitter exakt.
  - Setze die Grobgitter-Lösung durch Interpolation auf das nächst feinere Gitter fort.
- Führe auf dem aktuellen Gitter mehrere Schritte eines klassischen iterativen Verfahrens durch.

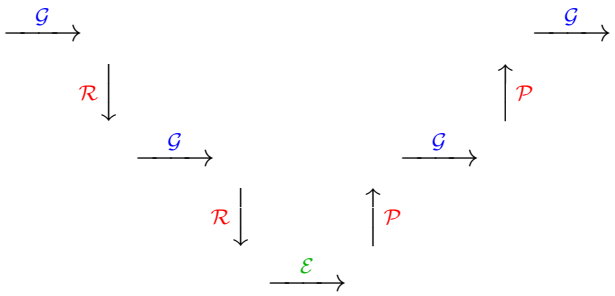
Dabei soll der letzte Schritt schnell schwingende Fehlerkomponenten dämpfen, die eventuell beim Übergang vom gröberen zum feineren Gitter erzeugt wurden.

Diese Idee lässt sich schematisch wie folgt darstellen:



Dabei steht $\mathcal{E}$ für das exakte Lösen auf dem gröberen Gitter.

In dieser primitiven Form ist die beschriebene Idee nicht konkurrenzfähig, da das Lösen auf dem gröberen Gitter nach wie vor zu aufwändig ist. Der entscheidende Schritt ist die Idee, den $\mathcal{E}$-Blcok rekursiv durch den gleichen Algorithmus zu ersetzen, schematisch:



Für die Formulierung des resultierenden Algorithmus benötigen wir folgende Ingredienzien:

- Eine Folge $\mathcal{T}_k$ zunehmend (gleichmäßig oder adaptiv) verfeinerter Unterteilungen mit zugehörigen diskreten Problemen $L_k u_k = f_k$.
- Einen *Glättungsoperator* M_k, der leicht auswertbar ist und der gleichzeitig eine passable Näherung für L_k^{-1} liefert.
- Einen *Restriktionsoperator* $R_{k,k-1}$, der Funktionen zur Unterteilung $\mathcal{T}_k$ in solche zur nächst gröberen Unterteilung $\mathcal{T}_{k-1}$ abbildet.
- Einen *Prolongationsoperator* $I_{k-1,k}$, der Funktionen zur Unterteilung $\mathcal{T}_{k-1}$ in solche zur nächst feineren Unterteilung $\mathcal{T}_k$ abbildet.

Damit lautet der *Mehrgitter-Algorithmus*:

Algorithmus VI.5.1. (Mehrgitter-Algorithmus)

(0) *Gegeben: das aktuelle Niveau k, Parameter μ, ν_1 und ν_2, die Matrix L_k, die rechte Seite f_k und eine Startnäherung u_k. Gesucht: eine verbesserte Näherungslösung u_k.*

(1) *Falls $k = 0$ ist, berechne*

$$u_0 = L_0^{-1} f_0;$$

stopp.

(2) (Vor-Glättung) *Führe ν_1 Schritte des stationären Iterationsverfahrens*

$$u_k \mapsto u_k + M_k(f_k - L_k u_k)$$

durch.

(3) (Grobgitter-Korrektur)

(a) *Berechne*

$$f_{k-1} = R_{k,k-1}(f_k - L_k u_k)$$

und setze

$$u_{k-1} = 0.$$

(b) *Führe μ Iterationen des Mehrgitter-Algorithmus mit Parametern $k-1$, μ, ν_1, ν_2, L_{k-1}, f_{k-1}, u_{k-1} durch und bezeichne das Ergebnis mit u_{k-1}.*

(c) *Ersetze u_k durch $u_k + I_{k-1,k}u_{k-1}$.*

(4) (Nach-Glättung) *Führe ν_2 Schritte des stationären Iterationsverfahrens*

$$u_k \mapsto u_k + M_k(f_k - L_k u_k)$$

durch.

Typische Werte der Parameter sind $\mu = 1$ (sog. *V-Zyklus*) oder $\mu = 2$ (sog. *W-Zyklus*), $\nu_1 = \nu_2 = \nu$ oder $\nu_1 = \nu$, $\nu_2 = 0$ oder $\nu_1 = 0$, $\nu_2 = \nu$ mit $1 \leq \nu \leq 4$.

Die Prolongation ist typischerweise bestimmt durch die natürliche Inklusion der Finite-Element-Räume, d.h. eine Finite-Element-Funktion zu einer gröberen Unterteilung kann durch die Basisfunktionen der aktuellen, feineren Unterteilung ausgedrückt werden (vgl. Abbildung VI.5.1).

Die Restriktion wird üblicherweise dadurch berechnet, dass Basisfunktionen zur gröberen Unterteilung in das diskrete Problem zur feineren Unterteilung als Testfunktionen eingesetzt werden.

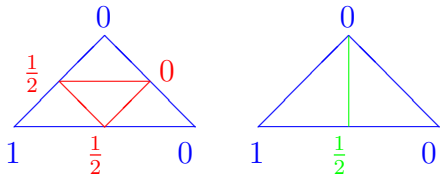


ABBILDUNG VI.5.1. Darstellung einer nodalen Basisfunktion eines Elementes durch die nodalen Basisfunktionen zu kleineren, durch reguläre oder irreguläre Unterteilung gewonnenen Elementen

Für die Glättung haben sich u.a. die Gauß-Seidel-Iteration, die SSOR-Iteration und die ILU-Iteration bewährt.
Bei der SSOR-Iteration werden zunächst die Gleichungen in aufsteigender Nummerierung mit einer Gauß-Seidel-Iteration als Vor-Glättung durchlaufen und danach die Gleichungen in absteigender Nummerierung mit einer Gauß-Seidel-Iteration als Nach-Glättung.
Bei der *ILU-Iteration* wird zunächst eine *unvollständige LR-Zerlegung* von L_k bestimmt, indem bei der üblichen LR-Zerlegung [6, Algorithmus I.8] alle Nullelemente der Matrix unterdrückt werden. Das Ergebnis ist eine näherungsweise Zerlegung $\mathcal{L}_k\mathcal{U}_k \approx L_k$. Bei der Glättung wird dann $v_k = M_k u_k$ durch Lösen des Systems $\mathcal{L}_k\mathcal{U}_k v_k = u_k$ bestimmt.

Falls $\mu \leq 2$ und $N_k > \mu N_{k-1}$ ist (was in der Praxis praktisch immer der Fall ist), benötigt eine Iteration des Mehrgitter-Algorithmus $O(N_k)$ Operationen.

Man kann zeigen, dass die Konvergenzrate des Mehrgitter-Algorithmus kleiner ist als 1 *gleichmäßig* für alle Unterteilungen und dass sie durch $\frac{c}{c+\nu_1+\nu_2}$ beschränkt ist, wobei c nur von den Formparametern der Unterteilungen abhängt. In der Praxis werden typischerweise Konvergenzraten zwischen 0.1 und 0.5 erzielt.

VI.6. Verfahrensvergleiche

Tabellen VI.6.1, VI.6.2 und VI.6.3 erlauben den Vergleich des Gaußschen Eliminationsverfahrens (Gauß-El.), der Gauß-Seidel-Iteration (GS), des CG-Verfahrens (CG), des PCG-Verfahrens mit SSOR-Vorkonditionierung (PCG) und des Mehrgitter-Algorithmus (MG) unter verschiedenen Aspekten. Dabei wird bei den iterativen Verfahren der Anfangsfehler stets um den Faktor 0.05 reduziert. Die Tabellen belegen eindrücklich die Überlegenheit des PCG-Verfahrens für kleine und mittlere Probleme und des Mehrgitter-Algorithmus für größere Probleme.

TABELLE VI.6.1. Arithmetische Operationen zur Bestimmung einer linearen Finite-Element-Diskretisierung auf einer Courant-Triangulierung der Poisson-Gleichung im Einheitsquadrat

h	Gauß-El.	GS	CG	PCG	MG
$\frac{1}{16}$	$7.6 \cdot 10^5$	$2.6 \cdot 10^5$	$2.7 \cdot 10^4$	$1.6 \cdot 10^4$	$1.2 \cdot 10^4$
$\frac{1}{32}$	$2.8 \cdot 10^7$	$4.5 \cdot 10^6$	$2.2 \cdot 10^5$	$8.6 \cdot 10^4$	$4.9 \cdot 10^4$
$\frac{1}{64}$	$9.9 \cdot 10^8$	$7.6 \cdot 10^7$	$1.9 \cdot 10^6$	$5.0 \cdot 10^5$	$2.1 \cdot 10^5$
$\frac{1}{128}$	$3.3 \cdot 10^{10}$	$1.2 \cdot 10^9$	$1.5 \cdot 10^7$	$3.2 \cdot 10^6$	$8.4 \cdot 10^5$

TABELLE VI.6.2. Iterationen zur Bestimmung einer linearen Finite-Element-Diskretisierung auf einer Courant-Triangulierung der Poisson-Gleichung im Einheitsquadrat

h	GS	CG	PCG	MG
$\frac{1}{16}$	236	12	4	1
$\frac{1}{32}$	954	23	5	2
$\frac{1}{64}$	3820	47	7	2
$\frac{1}{128}$	15287	94	11	1

TABELLE VI.6.3. Iterationen (It.) und Konvergenzraten (κ) zur Bestimmung einer linearen Finite-Element-Diskretisierung mit adaptiver Verfeinerung und DOF Freiheitsgraden einer Reaktions-Diffusions-Gleichung im Einheitsquadrat mit innerer Grenzschicht

	CG		PCG		MG	
DOF	It.	κ	It.	κ	It.	κ
9	4	0.10	3	0.2	4	0.3
47	10	0.60	7	0.5	3	0.3
185	24	0.80	12	0.7	5	0.2
749	49	0.90	21	0.8	5	0.4
2615	94	0.95	37	0.9	6	0.4
5247	130	0.96	55	0.9	5	0.4

VI.7. Unsymmetrische, indefinite und nichtlineare Probleme

Das CG-Verfahren bricht für unsymmetrische oder indefinite Probleme, bei denen die Steifigkeitsmatrix Eigenwerte mit positivem und mit negativem Realteil hat, in der Regel zusammen. Ein naiver Ausweg ist das Anwenden des CG-Verfahrens auf die symmetrischen, positiv definiten *Normalengleichungen*

$$L^T L u = L^T f.$$

Dadurch verdoppelt sich aber der Aufwand, weil sich die Kondition bei Übergang zu den Normalengleichungen quadriert. Einen besseren Ausweg bieten spezielle Varianten des CG-Verfahrens wie das *stabilisierte bi-konjugierte Gradienten Verfahren (Bi-CG-Stab-Verfahren).*

Algorithmus VI.7.1. (Bi-CG-Stab-Verfahren)

(0) *Gegeben: Matrix L, rechte Seite f, Startnäherung u_0 und Toleranz $\varepsilon > 0$.*
Gesucht: verbesserte Näherungslösung für $Lu = f$.

(1) *Berechne*

$$r_0 = b - Lu_0$$

und setze

$$\begin{aligned} \overline{r}_0 &= r_0, \\ v_{-1} &= 0, \\ p_{-1} &= 0, \\ \alpha_{-1} &= 1, \\ \rho_{-1} &= 1, \\ \omega_{-1} &= 1, \end{aligned}$$

sowie $i = 0$.

(2) *Falls*

$$r_i \cdot r_i < \varepsilon^2$$

ist, gebe u_i *als Näherungslösung aus;* stopp.

(3) *Berechne*

$$\rho_i = \overline{r}_i \cdot r_i,$$
$$\beta_{i-1} = \frac{\rho_i \alpha_{i-1}}{\rho_{i-1}\omega_{i-1}}.$$

Falls

$$|\beta_{i-1}| < \varepsilon$$

ist, liegt ein möglicher Abbruch vor; stopp. *Andernfalls berechne*

$$p_i = r_i + \beta_{i-1}\{p_{i-1} - \omega_{i-1}v_{i-1}\},$$
$$v_i = Lp_i,$$
$$\alpha_i = \frac{\rho_i}{\overline{r}_0 \cdot v_i}.$$

Falls

$$|\alpha_i| < \varepsilon$$

ist, liegt ein möglicher Abbruch vor; stopp. *Andernfalls berechne*

$$s_i = r_i - \alpha_i v_i,$$
$$t_i = Ls_i,$$
$$\omega_i = \frac{t_i \cdot s_i}{t_i \cdot t_i},$$
$$u_{i+1} = u_i + \alpha_i p_i + \omega_i s_i,$$
$$r_{i+1} = s_i - \omega_i t_i.$$

Erhöhe i *um* 1 *und gehe zu Schritt* (2) *zurück.*

Das Bi-CG-Stab-Verfahren versucht, simultan das ursprüngliche Problem $Lu = f$ und das *adjungierte Problem* $L^T v = f$ zu lösen. Es benötigt aber nur die Steifigkeitsmatrix L des ursprünglichen Problems. Es erfordert nur Skalarprodukte und Matrix-Vektor-Multiplikationen. Das Bi-CG-Stab-Verfahren kann auch vorkonditioniert werden; mögliche Vorkonditionierer sind das SSOR-Verfahren oder die ILU-Zerlegung angewandt auf den symmetrischen Anteil $\frac{1}{2}(L + L^T)$ von L.

Mehrgitterverfahren können direkt auf unsymmetrische oder indefinite Probleme angewendet werden. Unter Umständen muss man aber spezielle Glätter verwenden. Die Richardson-Relaxation angewandt auf die Normalengleichungen ist ein robuster Glätter, der allerdings zu Konvergenzraten von etwa 0.8 für das Mehrgitterverfahren führt. Die ILU-Zerlegung ist ebenfalls robust, aber aufwändiger und führt zu Konvergenzraten von etwa 0.5.

Nichtlineare Probleme werden typischerweise mit einem (gedämpften) Newton-Verfahren gelöst. In jeder Iteration des Newton-Verfahrens ist ein lineares Problem zu lösen. Dies kann mit iterativen Lösern geschehen, das Ergebnis der vorhergehenden Newton-Iteration ist dann meist ein guter Startwert für die innere Iteration. Bei Mehrgitterverfahren kann auch die Rolle von äußerer und innerer Iteration vertauscht werden, dann werden wenige Newton-Iterationen verbunden mit einer wenig genauen Lösung der linearen Hilfsprobleme als Glätter verwendet.

KAPITEL VII

Parabolische Differentialgleichungen

VII.1. Diskretisierungsmethoden

In diesem Kapitel betrachten wir lineare parabolische Differentialgleichungen zweiter Ordnung

$$\begin{aligned} \frac{\partial u}{\partial t} - \operatorname{div}(A\nabla u) + \mathbf{a}\cdot\nabla u + \alpha u &= f \quad \text{in } \Omega\times(0,T] \\ u &= 0 \quad \text{auf } \Gamma\times(0,T] \\ u(\cdot,0) &= u_0 \quad \text{in } \Omega \end{aligned}$$

in einem Raum-Zeit-Zylinder $\Omega\times(0,T]$. Dabei ist Ω ein *Polyeder* in $\mathbb{R}^d$ mit $d=2$ oder $d=3$, $A(x,t)$ für jedes x in Ω und t in $(0,T]$ eine symmetrische, positiv definite, $d\times d$ Matrix, $\mathbf{a}(x,t)$ für jedes x in Ω und t in $(0,T]$ ein Vektor in $\mathbb{R}^d$ und $\alpha(x,t)$ für jedes x in Ω und t in $(0,T]$ eine nicht-negative Zahl mit

$$\alpha(x,t) - \frac{1}{2}\operatorname{div}\mathbf{a}(x,t) \geq 0$$

für jedes x in Ω und t in $(0,T]$. Die Einschränkung auf polyhedrale Gebiete Ω erleichtert die Diskretisierung. Gebiete mit gekrümmten Rändern werden wie in §IV.3 beschrieben behandelt. Obige Bedingung an $\mathbf{a}$ und α garantiert die eindeutige Lösbarkeit der Differentialgleichung.

Für die Diskretisierung parabolischer Differentialgleichungen gibt es im wesentlichen drei Zugänge:

- die *Linien-Methode*,
- das *Rothe-Verfahren*,
- *Raum-Zeit Finite-Elemente*.

Für klassische, nicht adaptive Unterteilungen liefern sie häufig die selben diskreten Lösungen. Die Linien-Methode ist unflexibel und für Adaptivität nicht geeignet. Die Analyse des Rothe-Verfahrens ist knifflig, da sie Differenzierbarkeitseigenschaften bzgl. der Zeitvariablen benötigt, die häufig nicht zur Verfügung stehen. Raum-Zeit Finite-Elemente hingegen erlauben a posteriori Fehlerabschätzungen und sind für Raum-Zeit-Adaptivität gut geeignet.

VII.2. Linien-Methode

Die *Linien-Methode* kann wie folgt beschrieben werden. Wähle eine *feste* Unterteilung $\mathcal{T}$ von Ω und einen *festen*, zugehörigen Finite-Element-Raum $X(\mathcal{T})$ *(Orts-Diskretisierung)*. Bezeichne mit $A_\mathcal{T}$ und $f_\mathcal{T}$ die zugehörige Steifigkeitsmatrix und den Lastvektor. Dann liefert die Orts-Diskretisierung das folgende System gewöhnlicher Differentialgleichungen:

$$\frac{du_\mathcal{T}}{dt} = f_\mathcal{T} - A_\mathcal{T} u_\mathcal{T}.$$

Wende hierauf ein Standardverfahren für gewöhnliche Differentialgleichungen an wie z.B. das implizite Euler-Verfahren, das Crank-Nicolson-Verfahren oder ein Runge-Kutta-Verfahren *(Zeit-Diskretisierung)*. Das Crank-Nicolson-Verfahren ergibt so z.B. die Vorschrift

$$\frac{u_\mathcal{T}^n - u_\mathcal{T}^{n-1}}{\tau} = \frac{1}{2}\big(f_\mathcal{T}^n - A_\mathcal{T} u_\mathcal{T}^n + f_\mathcal{T}^{n-1} - A_\mathcal{T} u_\mathcal{T}^{n-1}\big)$$

für die sukzessive Bestimmung der Approximation $u_\mathcal{T}^n$ für die Lösung u der Differentialgleichung zur Zeit $t_0 + n\tau$.

VII.3. Rothe-Verfahren

Bei dem *Rothe-Verfahren* ist die Reihenfolge von Orts- und Zeit-Diskretisierung vertauscht. Dazu interpretiert man das parabolische Problem als eine gewöhnliche Differentialgleichung und wendet hierauf ein Standardverfahren an wie z.B. das implizite Euler-Verfahren, das Crank-Nicolson-Verfahren oder ein Runge-Kutta-Verfahren. *(Zeit-Diskretisierung)*. Jeder Zeitschritt erfordert dann die Lösung einer stationären elliptischen Differentialgleichung, die mit einem üblichen Finite-Element-Verfahren diskretisiert wird *(Orts-Diskretisierung)*. Das Crank-Nicolson-Verfahren ergibt so z.B. die elliptischen Differentialgleichungen

$$\begin{aligned}\frac{u^n - u^{n-1}}{\tau} + \frac{1}{2}\big(&- \operatorname{div}(A\nabla u^n) + \mathbf{a}\cdot\nabla u^n + \alpha u^n \\ &- \operatorname{div}(A\nabla u^{n-1}) + \mathbf{a}\cdot\nabla u^{n-1} + \alpha u^{n-1}\big) \\ &= \frac{1}{2}\big(f^n + f^{n-1}\big).\end{aligned}$$

Werden diese mit einem üblichen Finite-Element-Verfahren zu einer *festen* Unterteilung diskretisiert, erhält man die gleichen diskreten Probleme wie bei der Linienmethode.

Die Linienmethode und das Rothe-Verfahren unterscheiden sich in der mathematischen Analyse und ihrer Eignung für Adaptivität. Die Linienmethode erlaubt variable Zeitschritte, aber keine Ortsadaptivität. Das Rothe-Verfahren hingegen erlaubt verschiedene Orts-Diskretisierungen zu unterschiedlichen Zeiten, nicht jedoch variable Zeitschritte.

VII.4. Raum-Zeit Finite-Elemente

Raum-Zeit Finite-Elemente vermeiden die geschilderten Probleme. Für ihre Beschreibung nimmt man am besten an, man habe eine Unterteilung $\mathcal{I} = \{[t_{n-1}, t_n] : 1 \leq n \leq N_\mathcal{I}\}$ von $[0, T]$ mit $0 = t_0 < \ldots < t_{N_\mathcal{I}} = T$ gewählt. In der Praxis werden die Zeitpunkte t_n jedoch sukzessive im Laufe des Verfahrens bestimmt (vgl. Algorithmus VII.6.1). Setze

$$\tau_n = t_n - t_{n-1}.$$

Zu jedem Zeitpunkt t_n wähle man anschließend eine Unterteilung $\mathcal{T}_n$ von Ω und einen zugehörigen Finite-Element-Raum $X_n = X(\mathcal{T}_n)$. Dabei können die Unterteilungen $\mathcal{T}_n$ und die Räume X_n von Zeitschritt zu Zeischritt beliebig varieren (vgl. Abbildung VII.4.1).

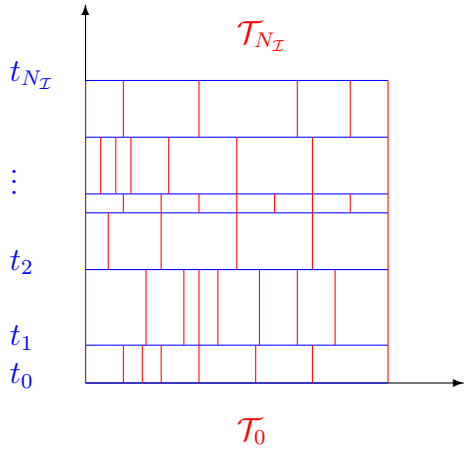


Abbildung VII.4.1. Raum-Zeit-Gitter

Damit lautet die *Raum-Zeit Finite-Element-Diskretisierung*

Berechne eine Interpolierende $u^0_{\mathcal{T}_0} \in X_0$ von u_0 und bestimme für $n = 1, 2, \ldots$ sukzessive $u^n_{\mathcal{T}_n} \in X_n$ *(Ansatzfunktion)* so, dass mit

$$u^{n\theta} = \theta u^n_{\mathcal{T}_n} + (1-\theta) u^{n-1}_{\mathcal{T}_{n-1}}$$

für alle $v^n_{\mathcal{T}_n} \in X_n$ *(Testfunktion)* gilt

$$\begin{aligned} &\int_\Omega \frac{1}{\tau_n}(u^n_{\mathcal{T}_n} - u^{n-1}_{\mathcal{T}_{n-1}}) v_{\mathcal{T}_n} dx + \int_\Omega \nabla u^{n\theta} \cdot A \nabla v_{\mathcal{T}_n} dx \\ &+ \int_\Omega \mathbf{a} \cdot \nabla u^{n\theta} v_{\mathcal{T}_n} dx + \int_\Omega \alpha u^{n\theta} v_{\mathcal{T}_n} dx \\ &= \int_\Omega f v_{\mathcal{T}_n} dx. \end{aligned}$$

Die Wahl des Parameters θ bestimmt den Typ der Zeit-Diskretisierung:

- $\theta = \frac{1}{2}$ entspricht dem Crank-Nicolson-Verfahren.
- $\theta = 1$ entspricht dem impliziten Euler-Verfahren.
- $\theta = 0$ entspricht dem expliziten Euler-Verfahren.

Aus Stabilitätsgründen sollte $\theta \geq \frac{1}{2}$ gewählt werden.

Die Raum-Zeit Finite-Element-Diskretisierung hat folgende Eigenschaften:

> Für $\theta > 0$ ist in jedem Zeitschritt ein lineares Gleichungssystem zu lösen, das der Finite-Element-Diskretisierung einer elliptischen Differentialgleichung entspricht.
> Für $\mathbf{a} \neq 0$ ist die Steifigkeitsmatrix unsymmetrisch und indefinit.
> Bei Verwenden eines iterativen Lösers ist $u^{n-1}_{\mathcal{T}_{n-1}}$ ein guter Startwert für die Berechnung von $u^n_{\mathcal{T}_n}$.
> Der Fehler der Diskretisierung verhält sich wie $h^2 + \tau^\gamma$ mit $\gamma = 2$ für $\theta = \frac{1}{2}$ und $\gamma = 1$ für $\theta \neq \frac{1}{2}$. Dabei ist h die maximale Orts-Gitterweite und τ die maximale Zeitschrittweite.

VII.5. Charakteristiken-Methode

Die im vorigen Abschnitt beschriebenen Verfahren haben alle Probleme, wenn der Konvektionsterm $\mathbf{a}$ groß wird im Verhältnis zum Diffusionsterm A. Dann treten unphysikalische Oszillationen auf, die mit Petrov-Galerkin-Verfahren wie in §IV.4 geschildert stabilisiert werden müssen. Außerdem ist die Steifigkeitsmatrix für $\mathbf{a} \neq 0$ immer unsymmetrisch und indefinit, so dass spezielle Löser wie in §VI.7 beschrieben benutzt werden müssen.

Diese Schwierigkeiten vermeidet die *Charakteristiken-Methode*, alias *Transport-Diffusions-Algorithmus*. Sie beruht auf folgender Idee:

- Für jeden Punkt $(x^*, t^*) \in \Omega \times (0, T]$ besitzt die *Charakteristiken-Gleichung*
$$\begin{aligned} \frac{d}{dt} x(t; x^*, t^*) &= \mathbf{a}(x(t; x^*, t^*), t) \\ x(t^*; x^*, t^*) &= x^* \end{aligned}$$
eine eindeutige Lösung auf dem Intervall $(0, t^*)$.
- Für
$$U(x^*, t) = u(x(t; x^*, t^*), t)$$
gilt
$$\frac{dU}{dt} = \frac{\partial u}{\partial t} + \mathbf{a} \cdot \nabla u.$$
- Daher kann die parabolische Differentialgleichung in der Form
$$\frac{dU}{dt} - \operatorname{div}(A \nabla u) + \alpha u = f$$
geschrieben werden.

- Der Diffusions-Reaktions-Term $-\operatorname{div}(A\nabla u)+\alpha u$ und die Materialableitung $\frac{dU}{dt}$ werden separat diskretisiert.

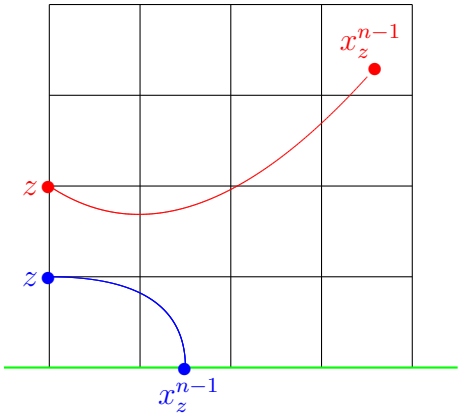


ABBILDUNG VII.5.1. Näherungslösungen x_z^{n-1} der Charakteristiken-Gleichung zu zwei verschiedenen Punkten $z \in \mathcal{N}_n$

Für die Realisierung dieser Idee bezeichnen wir mit $\mathcal{N}_n$ die globalen Freiheitsgrade von X_n. Für jedes n und jedes $z \in \mathcal{N}_n$ wenden wir dann ein klassisches Verfahren wie z.B. das implizite Euler-Verfahren, oder das Crank-Nicolson-Verfahren oder ein Runge-Kutta-Verfahren auf die Charakteristiken-Gleichung zu $(x^*, t^*) = (z, t_n)$ an und bezeichnen mit x_z^{n-1} die resultierende Näherung für $x(t_{n-1}; z, t_n)$ (vgl. Abbildung VII.5.1). Damit lautet der n-te Zeitschritt der *Charakteristiken-Methode*

Bestimme zuerst $\widetilde{u}_{\mathcal{T}_n}^{n-1} \in X_n$ so, dass

$$\widetilde{u}_{\mathcal{T}_n}^{n-1}(z) = u_{\mathcal{T}_{n-1}}^{n-1}(x_z^{n-1})$$

ist für alle $z \in \mathcal{N}_n$.
Bestimme danach $u_{\mathcal{T}_n}^n \in X_n$ so, dass für alle $v_{\mathcal{T}_n}^n \in X_n$ gilt

$$\frac{1}{\tau_n}\int_\Omega u_{\mathcal{T}_n}^n v_{\mathcal{T}_n}^n dx + \int_\Omega \nabla u_{\mathcal{T}_n}^n \cdot A\nabla v_{\mathcal{T}_n}^n dx + \int_\Omega \alpha u_{\mathcal{T}_n}^n v_{\mathcal{T}_n}^n dx = \frac{1}{\tau_n}\int_\Omega \widetilde{u}_{\mathcal{T}_n}^{n-1} v_{\mathcal{T}_n}^n dx + \int_\Omega f v_{\mathcal{T}_n}^n dx.$$

Die Charakteristiken-Methode hat folgende Eigenschaften:

Sie ist besonders geeignet für die Diskretisierung parabolischer Gleichungen mit einem großen Konvektionsterm.

Sie entkoppelt die Diskretisierung der Zeit- und Konvektionsableitungen von der Diskretisierung der restlichen Terme.
Sie erfordert die Lösung einer Folge gewöhnlicher Differentialgleichungen und von Reaktions-Diffusions-Gleichungen mit *symmetrisch, positiv definiter Steifigkeitsmatrix*.

VII.6. Adaptivität

Bezeichne mit $\mathcal{E}_n$ die Kanten ($d = 2$) bzw. Seitenflächen ($d = 3$) der Elemente in $\mathcal{T}_n$. Definiere einen *Ortsindikator* η_h^n durch

$$\eta_h^n = \Big\{ \sum_{K\in\mathcal{T}_n} h_K^2 \int_K \big| f(x,t_n) - \frac{1}{\tau_n}(u^n_{\mathcal{T}_n} - u^{n-1}_{\mathcal{T}_{n-1}}) + \operatorname{div}(A\nabla u^n_{\mathcal{T}_n}) - \mathbf{a}\cdot\nabla u^n_{\mathcal{T}_n} - \alpha u^n_{\mathcal{T}_n}\big|^2 dx + \sum_{E\in\mathcal{E}_n} h_E \int_E \big|\big[\mathbf{n}_E \cdot A\nabla u^n_{\mathcal{T}_n}\big]_E\big|^2 dS \Big\}^{\frac{1}{2}}$$

und einen *Zeitindikator* η_τ^n durch

$$\eta_\tau^n = \Big\{ \int_\Omega |\nabla u^n_{\mathcal{T}_n} - \nabla u^{n-1}_{\mathcal{T}_{n-1}}|^2 dx \Big\}^{\frac{1}{2}}.$$

Dann gilt für die stetige, bezüglich der Zeit stückweise lineare Funktion $u_{\mathcal{I}}$, die zur Zeit t_n mit $u^n_{\mathcal{T}_n}$ übereinstimmt, die *a posteriori Fehlerabschätzung*

$$\Big\{ \max_{0\le t\le T} \int_\Omega |u - u_{\mathcal{I}}|^2 dx + \int_0^T \int_\Omega |\nabla u - \nabla u_{\mathcal{I}}|^2 dx dt \Big\}^{\frac{1}{2}} \approx \Big\{ \int_\Omega |u^0_{\mathcal{T}_0} - u_0|^2 dx + \sum_{n=1}^{N_{\mathcal{I}}} \tau_n \big[\big(\eta_h^n\big)^2 + \big(\eta_\tau^n\big)^2 \big] \Big\}^{\frac{1}{2}}.$$

Dabei bedeutet „≈“ obere und untere Schranken modulo multiplikativer Faktoren. Diese Faktoren hängen von dem Polynomgrad der Finite-Element-Funktionen und den Formparametern der Unterteilungen ab.

Die a posteriori Fehlerabschätzung hat folgende Eigenschaften:

Die oberen Schranken sind global in Ort und Zeit.
Die unteren Schranken sind global im Ort und lokal in der

Zeit.
Der Ortindikator η_h^n kontrolliert den Fehler der Orts-Diskretisierung.
Der Zeitindikator η_τ^n kontrolliert den Fehler der Zeit-Diskretisierung.
Der Ortsindikator η_h^n besteht aus den Elementresiduen der diskreten Lösung und aus Sprungtermen über die Elementgrenzen.
Bei den Elementresiduen wird die diskrete Lösung elementweise in die Differentialgleichung eingesetzt.
Die Sprungterme sind die gleichen wie bei der entsprechenden elliptischen Differentialgleichung ohne Zeitableitung.
Der Zeitindikator η_τ^n beschreibt einen Sprungterm bezüglich der Zeitvariablen.

Die a posteriori Fehlerabschätzung führt auf den folgenden Algorithmus zur Anpassung der Zeitschrittweiten τ_n und der Unterteilungen $\mathcal{T}_n$.

Algorithmus VII.6.1. (Raum-Zeit Adaptivität)

(0) *Gegeben: Toleranz ε, Unterteilung $\mathcal{T}_0$, Zeitschritt τ_1*

(1) *Passe $\mathcal{T}_0$ so an, dass*

$$\int_\Omega |u^0_{\mathcal{T}_0} - u_0|^2 dx \le \frac{1}{4}\varepsilon^2$$

ist. Setze $n = 1$, $t_1 = \tau_1$.

(2) *Löse das diskrete Problem zur Zeit t_n und bestimme die Indikatoren η_h^n und η_τ^n.*

(3) *Falls*

$$\eta_\tau^n > \frac{\varepsilon}{2\sqrt{T}}$$

ist, halbiere τ_n und ersetze t_n durch $\frac{1}{2}(t_{n-1}+t_n)$; gehe zu Schritt (2) zurück.

(4) *Passe $\mathcal{T}_n$ so an, dass*

$$\eta_h^n \le \frac{\varepsilon}{2\sqrt{T}}$$

ist. Falls

$$\eta_\tau^n < \frac{\varepsilon}{4\sqrt{T}}$$

ist, verdoppele τ_n.

(5) *Falls $t_n = T$ ist,* stopp. *Andernfalls setze*

$$t_{n+1} = \min\{T, t_n + \tau_n\},$$

erhöhe n um 1 und gehe zu Schritt (2) zurück.

Am Ende von Algorithmus VII.6.1 ist

$$\Big\{\int_\Omega |u^0_{\mathcal{T}_0} - u_0|^2 dx + \sum_{n=1}^{N_\mathcal{I}} \tau_n \big[\big(\eta_h^n\big)^2 + \big(\eta_\tau^n\big)^2\big]\Big\}^{\frac{1}{2}} \leq \varepsilon.$$

Beim Anpassen von $\mathcal{T}_n$ werden t_n, τ_n und η_τ^n fest gehalten. Die Anpassung von $\mathcal{T}_n$ erfordert ggf. das wiederholte Lösen diskreter Probleme und die Neuberechnung von η_h^n.

KAPITEL VIII

Finite-Volumen-Methoden

VIII.1. Systeme in Divergenzform

Finite-Volumen-Methoden sind maßgeschneidert für *Systeme in Divergenzform*, bei denen ein Vektorfeld $\mathbf{U}$ auf einer Teilmenge Ω des $\mathbb{R}^d$ mit Werten in $\mathbb{R}^m$ gesucht ist, das die Differentialgleichung

$$\begin{aligned} \frac{\partial \mathbf{M}(\mathbf{U})}{\partial t} + \operatorname{div} \underline{\mathbf{F}}(\mathbf{U}) &= \mathbf{g}(\mathbf{U}, x, t) \quad \text{in } \Omega \times (0, \infty) \\ \mathbf{U}(\cdot, 0) &= \mathbf{U}_0 \qquad\quad \text{in } \Omega \end{aligned}$$

erfüllt. Dabei ist $\mathbf{g}$, die *Quelle*, ein Vektorfeld auf dem $\mathbb{R}^m \times \Omega \times (0, \infty)$ mit Werten in $\mathbb{R}^m$, $\mathbf{M}$, die *Masse*, ein Vektorfeld auf dem $\mathbb{R}^m$ mit Werten in $\mathbb{R}^m$, $\underline{\mathbf{F}}$ der *Fluss* eine Matrix-wertige Funktion auf dem $\mathbb{R}^m$ mit Werten in $\mathbb{R}^{m \times d}$ und $\mathbf{U}_0$, der *Anfangswert*, ein Vektorfeld auf Ω mit Werten in $\mathbb{R}^m$. Obige Differentialgleichung ist mit geeigneten Randbedingungen zu versehen, auf die wir in diesem Kapitel aber nicht eingehen. Man beachte, dass die Divergenz zeilenweise zu nehmen ist

$$\operatorname{div} \underline{\mathbf{F}}(\mathbf{U}) = \Big(\sum_{j=1}^{d} \frac{\partial \underline{\mathbf{F}}(\mathbf{U})_{i,j}}{\partial x_j} \Big)_{1 \le i \le m}.$$

Der Fluss $\underline{\mathbf{F}}$ kann in zwei Beiträge aufgespalten werden

$$\underline{\mathbf{F}} = \underline{\mathbf{F}}_{\text{adv}} + \underline{\mathbf{F}}_{\text{visc}}.$$

$\underline{\mathbf{F}}_{\text{adv}}$ heißt *advektiver Fluss* und enthält keine Ableitungen. $\underline{\mathbf{F}}_{\text{visc}}$ heißt *viskoser Fluss* und enthält räumliche Ableitungen. Der advektive Fluss modelliert Transport- oder Konvektionsphänomene, der viskose Fluss Diffusionsphänomene.

Beispiel VIII.1.1. Eine lineare parabolische Differentialgleichung zweiter Ordnung

$$\frac{\partial u}{\partial t} - \operatorname{div}(A \nabla u) + \mathbf{a} \cdot \nabla u + \alpha u = f,$$

wie wir sie im vorigen Kapitel betrachtet haben, ist ein System in Divergenzform mit

$$\begin{aligned} m = 1, \qquad \mathbf{U} = u, \qquad \mathbf{M}(\mathbf{U}) &= u, \\ \underline{\mathbf{F}}_{\text{adv}}(\mathbf{U}) = \mathbf{a} u, \quad \underline{\mathbf{F}}_{\text{visc}}(\mathbf{U}) &= -A \nabla u, \quad \mathbf{g}(\mathbf{U}) = f - \alpha u + (\operatorname{div} \mathbf{a}) u. \end{aligned}$$

Beispiel VIII.1.2. Die *Burger-Gleichung*

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0$$

ist ein System in Divergenzform mit

$$m = d = 1, \qquad \mathbf{u} = u, \qquad \mathbf{M}(\mathbf{U}) = u,$$
$$\underline{\mathbf{F}}_{\text{adv}}(u) = \frac{1}{2}u^2, \qquad \underline{\mathbf{F}}_{\text{visc}}(\mathbf{U}) = 0, \qquad \mathbf{g}(\mathbf{U}) = 0.$$

Andere wichtige Beispiel für Systeme in Divergenzform sind die *Euler-* und *Navier-Stokes-Gleichungen* für reibungsfreie und viskose Strömungen. Bei ihnen ist $d = 2$ oder $d = 3$ und $m = d + 2$. Der Vektor $\mathbf{U}$ besteht aus der Massendichte, dem Geschwindigkeitsfeld und der internen Energie der Strömung.

VIII.2. Grundidee der Finite Volumen Verfahren

Wir wählen einen Zeitschritt $\tau > 0$ und eine Unterteilung $\mathcal{T}$ von Ω in beliebige, nicht überlappende Polyeder. Dabei können die Elemente K in $\mathcal{T}$ kompliziertere Formen haben als Dreiecke, Vierecke usw. (vgl. Abbidldung VIII.3.1), und hängende Knoten sind erlaubt.

Nun wählen wir eine Zahl $n \geq 1$ und ein Element $K \in \mathcal{T}$ und halten beide im Folgenden fest. Als erstes integrieren wir die Differentialgleichung über $K \times [(n-1)\tau, n\tau]$

$$\int_{(n-1)\tau}^{n\tau}\int_K \frac{\partial \mathbf{M}(\mathbf{U})}{\partial t}dxdt + \int_{(n-1)\tau}^{n\tau}\int_K \operatorname{div}\underline{\mathbf{F}}(\mathbf{U})dxdt$$
$$= \int_{(n-1)\tau}^{n\tau}\int_K \mathbf{g}(\mathbf{U}, x, t)dxdt.$$

Im zweiten Schritt integrieren wir die Terme auf der linken Seite partiell

$$\int_{(n-1)\tau}^{n\tau}\int_K \frac{\partial \mathbf{M}(\mathbf{U})}{\partial t}dxdt = \int_K \mathbf{M}(\mathbf{U}(x, n\tau))dx - \int_K \mathbf{M}(\mathbf{U}(x, (n-1)\tau))dx,$$
$$\int_{(n-1)\tau}^{n\tau}\int_K \operatorname{div}\underline{\mathbf{F}}(\mathbf{U})dxdt = \int_{(n-1)\tau}^{n\tau}\int_{\partial K} \underline{\mathbf{F}}(\mathbf{U})\cdot\mathbf{n}_K dSdt.$$

Für das Folgende nehmen wir an, dass $\mathbf{U}$ bezüglich Ort und Zeit stückweise konstant ist, und bezeichnen mit $\mathbf{U}_K^n$ und $\mathbf{U}_K^{n-1}$ den Wert von $\mathbf{U}$

auf K zu den Zeiten $n\tau$ und $(n-1)\tau$. Dann ist

$$\begin{aligned}
\int_K \mathbf{M}(\mathbf{U}(x,n\tau))dx &\approx |K|\mathbf{M}(\mathbf{U}_K^n)\\
\int_K \mathbf{M}(\mathbf{U}(x,(n-1)\tau))dx &\approx |K|\mathbf{M}(\mathbf{U}_K^{n-1})\\
\int_{(n-1)\tau}^{n\tau}\int_{\partial K} \underline{\mathbf{F}}(\mathbf{U})\cdot\mathbf{n}_K dSdt &\approx \tau\int_{\partial K}\underline{\mathbf{F}}(\mathbf{U}_K^{n-1})\cdot\mathbf{n}_K dS\\
\int_{(n-1)\tau}^{n\tau}\int_K \mathbf{g}(\mathbf{U},x,t)dxdt &\approx \tau|K|\mathbf{g}(\mathbf{U}_K^{n-1},x_K,(n-1)\tau).
\end{aligned}$$

Dabei ist $|K|$ die Fläche von K, falls $d=2$ ist, und das Volumen von K, falls $d=3$ ist.
Im letzten Schritt approximieren wir das Randintegral für den Fluss durch einen *numerischen Fluss*

$$\begin{aligned}
&\tau\int_{\partial K}\underline{\mathbf{F}}(\mathbf{U}_K^{n-1})\cdot\mathbf{n}_K d\\
&\approx \tau\sum_{\substack{K'\in\mathcal{T}\\ \partial K\cap\partial K'\in\mathcal{E}}} |\partial K\cap\partial K'|\mathbf{F}_{\mathcal{T}}(\mathbf{U}_K^{n-1},\mathbf{U}_{K'}^{n-1}).
\end{aligned}$$

Insgesamt erhalten wir so das folgende *Finite-Volumen-Verfahren*

Berechne zuerst für jedes Element $K\in\mathcal{T}$

$$\mathbf{U}_K^0 = \frac{1}{|K|}\int_K \mathbf{U}_0(x).$$

Berechne danach für $n=1,2,\ldots$ sukzessive für jedes Element $K\in\mathcal{T}$

$$\begin{aligned}
\mathbf{M}(\mathbf{U}_K^n) = \mathbf{M}(\mathbf{U}_K^{n-1})&\\
&- \tau\sum_{\substack{K'\in\mathcal{T}\\ \partial K\cap\partial K'\in\mathcal{E}}} \frac{|\partial K\cap\partial K'|}{|K|}\mathbf{F}_{\mathcal{T}}(\mathbf{U}_K^{n-1},\mathbf{U}_{K'}^{n-1})\\
&+ \tau\mathbf{g}(\mathbf{U}_K^{n-1},x_K,(n-1)\tau).
\end{aligned}$$

Dabei bezeichnet $|\partial K\cap\partial K'|$ die Länge des gemeinsamen Randes von $K\cap K'$, falls $d=2$ ist, bzw. seine Fläche, falls $d=3$ ist.

Obiges Verfahren kann leicht wie folgt modifiziert werden:

- Die Zeitschrittweite kann variabel sein.
- Die Unterteilung von Ω kann sich von Zeitschritt zu Zeitschritt ändern.
- Die Näherung $\mathbf{U}_K^n$ muss nicht stückweise konstant sein.

Damit das Verfahren praktisch durchführbar ist, müssen wir noch folgende Aufgaben erledigen:

- Konstruktion der Unterteilung $\mathcal{T}$,
- Bestimmung des numerischen Flusses $\underline{\mathbf{F}}_{\mathcal{T}}$.

Außerdem müssten noch Randbedingungen berücksichtigt werden. Auf diesen Punkt gehen wir aber nicht ein.

VIII.3. Konstruktion der Gitter

Die Unterteilung $\mathcal{T}$ wird häufig als *duales Gitter* zu einer zulässigen *primalen Finite-Element-Unterteilung* $\widetilde{\mathcal{T}}$ konstruriert. Für Probleme in zwei Raumdimensionen, $d = 2$, gibt es hierfür im wesentlichen zwei verschiedene Ansätze (vgl. Abbildung VIII.3.1):

- Konstruiere für jedes Element $\widetilde{K} \in \widetilde{\mathcal{T}}$ die Mittelsenkrechten.
- Verbinde für jedes Element $\widetilde{K} \in \widetilde{\mathcal{T}}$ seinen Schwerpunkt mit den Kantenmittelpunkten.

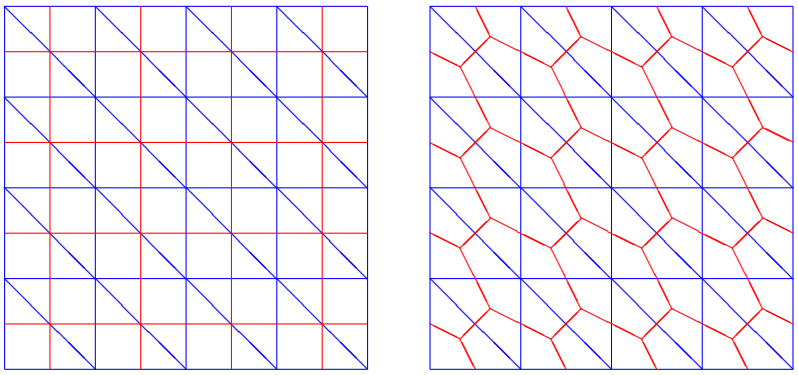


Abbildung VIII.3.1. Primale Finite-Element-Unterteilung (blau) und resultierendes duales Gitter (rot) basierend auf Mittelsenkrechten (links) und Schwerpunkten (rechts)

Duale Gitter haben zwei sehr nützliche Eigenschaften:

- Jedes Element in $K \in \mathcal{T}$ entspricht einem Elementeckpunkt x_K von $\widetilde{\mathcal{T}}$ und umgekehrt (vgl. Abbildung VIII.3.2 linker Teil).
- Zu jeder Kante E von $\mathcal{T}$ gibt es zwei Elementeckpunkte $x_{E,1}$, $x_{E,2}$ von $\widetilde{\mathcal{T}}$ so, dass die Strecke $\overline{x_{E,1}\, x_{E,2}}$ die Kante E schneidet (vgl. Abbildung VIII.3.2 rechter Teil).

Die Konstruktion über die Mittelsenkrechten hat den Vorteil, dass die Strecke $\overline{x_{E,1}\, x_{E,2}}$ und die Kante E sich in einem rechten Winkel schneiden. Andererseits hat sie auch erhebliche Nachteile:

- Die Mittelsenkrechten eines Dreiecks können sich in einem Punkt außerhalb des Dreieckes schneiden. Der Schnittpunkt der Mittelsenkrechten liegt genau dann im Dreieck, wenn das Dreieck spitzwinklig ist.

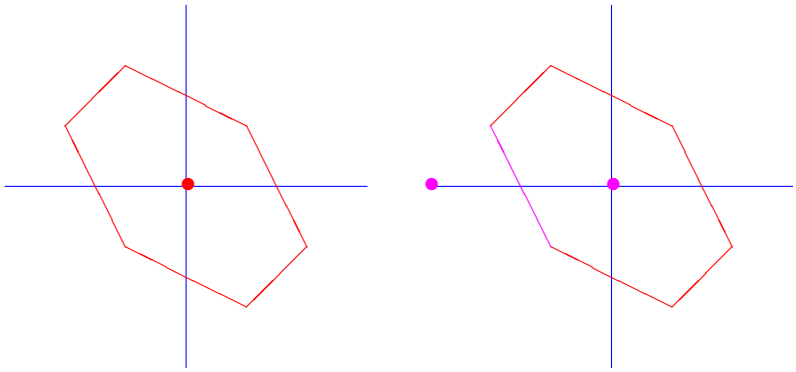


ABBILDUNG VIII.3.2. Duales Element K (rot) mit zugehörigem Eckpunkt x_K (rot) im primalen Gitter (links) und Kante E (magenta) im dualen Gitter mit zugehörigen Endpunkten $x_{E,1}$, $x_{E,2}$ (magenta) im primalen Gitter (rechts)

- Die Mittelsenkrechten eines Vierecks brauchen sich nicht zu schneiden. Sie schneiden sich genau dann in einem Punkt, wenn das Viereck ein Rechteck ist.
- Die Konstruktion mit Mittelsenkrechten hat kein dreidimensionales Analogon.

VIII.4. Konstruktion der numerischen Flüsse

Für die Konstruktion der numerischen Flüsse nehmen wir an, dass $\mathcal{T}$ ein duales Gitter zu einer primalen Finite-Element-Unterteilung $\widetilde{\mathcal{T}}$ ist. Für jede Kante oder Fläche E von $\mathcal{T}$ seien K_1 und K_2 die angrenzenden Volumina, $\mathbf{U}_1$ und $\mathbf{U}_2$ die Werte $\mathbf{U}_{K_1}^{n-1}$ und $\mathbf{U}_{K_2}^{n-1}$ und x_1, x_2 die Elementeckpunkte von $\widetilde{\mathcal{T}}$, für die die Strecke $\overline{x_1\, x_2}$ die Kante oder Fläche E schneidet.

Analog zum analytischen Fall spalten wir den numerischen Fluss $\underline{\mathbf{F}}_{\mathcal{T}}(\mathbf{U}_1, \mathbf{U}_2)$ in einen *viskosen numerischen Fluss* $\underline{\mathbf{F}}_{\mathcal{T},\mathrm{visc}}(\mathbf{U}_1, \mathbf{U}_2)$ und einen *advektiven numerischen Fluss* $\underline{\mathbf{F}}_{\mathcal{T},\mathrm{adv}}(\mathbf{U}_1, \mathbf{U}_2)$ auf, die wir separat konstruieren.

Für die Approximation der *viskosen Flüsse* führen wir ein lokales Koordinatensystem $\eta_1, \ldots, \eta_d$ so ein, dass die Richtung η_1 parallel ist zur Richtung von $\overline{x_1\, x_2}$ und die restlichen Richtungen tangential sind zu E (vgl. Abbildung VIII.4.1). Anschließend drücken wir alle Ableitungen in $\underline{\mathbf{F}}_{\mathrm{visc}}$ durch partielle Ableitungen bezüglich des neuen Koordinatensystems aus, unterdrücken alle partiellen Ableitungen, die nicht η_1 betreffen und approximiere partielle Ableitungen bezüglich η_1 durch Differenzenquotienten der Form $\frac{\varphi_1 - \varphi_2}{|x_1 - x_2|}$.

Für die Approximation der *advektiven Flüsse* bezeichnen wir mit

$$C(\mathbf{V}) = D(\underline{\mathbf{F}}_{\mathrm{adv}}(\mathbf{V}) \cdot \mathbf{n}_{K_1}) \in \mathbb{R}^{m \times m}$$

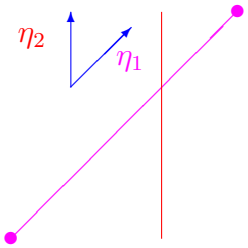


ABBILDUNG VIII.4.1. Lokales Koordinatensystem für die Approximation der viskosen Flüsse

die Ableitung von $\underline{\mathbf{F}}_{\text{adv}}(\mathbf{V}) \cdot \mathbf{n}_{K_1}$ nach $\mathbf{V}$ und nehmen an, dass diese Matrix diagonalisierbar ist, d.h.

$$Q(\mathbf{V})^{-1} C(\mathbf{V}) Q(\mathbf{V}) = \Delta(\mathbf{V})$$

mit einer invertierbaren Matrix $Q(\mathbf{V}) \in \mathbb{R}^{m\times m}$ und einer Diagonalmatrix $\Delta(\mathbf{V}) \in \mathbb{R}^{m\times m}$. Für die Euler- und Navier-Stokes-Gleichungen ist diese Annahme z.B. erfüllt. Jeder reellen Zahl z ordnen wir dann durch

$$z^+ = \max\{z, 0\}, \quad z^- = \min\{z, 0\}$$

ihren positiven und negativen Teil zu und definieren damit

$$\begin{aligned}\Delta(\mathbf{V})^\pm &= \operatorname{diag}\big(\Delta(\mathbf{V})_{11}^\pm, \dots, \Delta(\mathbf{V})_{mm}^\pm\big),\\ C(\mathbf{V})^\pm &= Q(\mathbf{V})\Delta(\mathbf{V})^\pm Q(\mathbf{V})^{-1}.\end{aligned}$$

Mit diesen Bezeichnungen ist das *Steger-Warming-Schema* zur Approximation der advektiven Flüsse gegeben durch

$$\mathbf{F}_{\mathcal{T},\text{adv}}(\mathbf{U}_1, \mathbf{U}_2) = C(\mathbf{U}_1)^+ \mathbf{U}_1 + C(\mathbf{U}_2)^- \mathbf{U}_2.$$

Eine andere, bessere Approximation ist das *van Leer-Schema*

$$\begin{aligned}&\mathbf{F}_{\mathcal{T},\text{adv}}(\mathbf{U}_1, \mathbf{U}_2)\\ &= \Big[C(\mathbf{U}_1) + C(\tfrac{1}{2}(\mathbf{U}_1 + \mathbf{U}_2))^+ - C(\tfrac{1}{2}(\mathbf{U}_1 + \mathbf{U}_2))^-\Big]\mathbf{U}_1\\ &\quad + \Big[C(\mathbf{U}_2) - C(\tfrac{1}{2}(\mathbf{U}_1 + \mathbf{U}_2))^+ + C(\tfrac{1}{2}(\mathbf{U}_1 + \mathbf{U}_2))^-\Big]\mathbf{U}_2.\end{aligned}$$

Beide Ansätze erfordern die Berechnung von $D\underline{\mathbf{F}}_{\text{adv}}(\mathbf{V}) \cdot \mathbf{n}_{K_1}$ und der entsprechenden Eigenwerte und Eigenvektoren für geeignete Werte von $\mathbf{V}$. Im allgemeinen ist der Ansatz von van Leer aufwändiger als der von Steger-Warming, da er drei statt zwei Auswertungen von $C(\mathbf{V})$ erfordert. Für die Navier-Stokes- und Euler-Gleichungen kann dieser Mehraufwand vermieden werden, da diese Gleichungen die spezielle Struktur $\underline{\mathbf{F}}_{\text{adv}}(\mathbf{V}) \cdot \mathbf{n}_{K_1} = C(\mathbf{V})\mathbf{V}$ haben.

Beispiel VIII.4.1. Für die Burger-Gleichung aus Beispiel VIII.1.2 lautet die Steeger-Warming-Schema

$$\underline{\mathbf{F}}_{\mathcal{T},\mathrm{adv}}(u_1,u_2) = \begin{cases} u_1^2 & \text{if } u_1 \geq 0, u_2 \geq 0 \\ u_1^2 + u_2^2 & \text{if } u_1 \geq 0, u_2 \leq 0 \\ u_2^2 & \text{if } u_1 \leq 0, u_2 \leq 0 \\ 0 & \text{if } u_1 \leq 0, u_2 \geq 0 \end{cases}$$

und das van Leer-Schema

$$\underline{\mathbf{F}}_{\mathcal{T},\mathrm{adv}}(u_1,u_2) = \begin{cases} u_1^2 & \text{if } u_1 \geq -u_2 \\ u_2^2 & \text{if } u_1 \leq -u_2. \end{cases}$$

VIII.5. Zusammenhang mit Finite-Element-Methoden

Es sei $\mathcal{T}$ ein duales Gitter zu einer primalen Finite-Element-Unterteilung $\widetilde{\mathcal{T}}$. Dann gibt es eine einfache Eins-zu-Eins-Beziehung zwischen stückweise konstanten Funktionen zu $\mathcal{T}$ und stetigen stückweise linearen Funktionen zu $\widetilde{\mathcal{T}}$

$$\begin{aligned} S^{0,-1}(\mathcal{T})^m \ni \mathbf{U}_{\mathcal{T}} &\leftrightarrow \widetilde{\mathbf{U}}_{\widetilde{\mathcal{T}}} \in S^{1,0}(\widetilde{\mathcal{T}})^m \\ \mathbf{U}_{\mathcal{T}}|_K &= \widetilde{\mathbf{U}}_{\widetilde{\mathcal{T}}}(x_K) \qquad \text{für alle } K \in \mathcal{T} \end{aligned}$$

d.h. $\mathbf{U}_{\widetilde{\mathcal{T}}}$ ist diejenige stetige stückweise lineare Funktion zu $\widetilde{\mathcal{T}}$, die im Punkt x_K den gleichen Wert hat wie die stückweise konstante Funktion $\mathbf{U}_{\mathcal{T}}$ auf dem dualen Element K. Diese Beziehung erlaubt folgenden einfachen adaptiven Algorithmus für Finite-Volumen-Methoden:

> Zu gegebener Lösung $\mathbf{U}_{\mathcal{T}}$ der Finite-Volumen-Diskretisierung berechne die zugehörige Finite-Element-Funktion $\widetilde{\mathbf{U}}_{\widetilde{\mathcal{T}}}$.
> Wende einen gebräuchlichen Fehlerschätzer auf $\widetilde{\mathbf{U}}_{\widetilde{\mathcal{T}}}$ an.
> Basierend auf diesem Fehlerschätzer wende eine gebräuchliche Verfeinerungsstrategie auf $\widetilde{\mathcal{T}}$ an und konstruiere so eine neue lokal verfeinerte Unterteilung $\widehat{\mathcal{T}}$.
> Benutze $\widehat{\mathcal{T}}$ als primales Gitter zur Konstruktion eines neuen dualen Gitters $\mathcal{T}'$. Dies ist die Verfeinerung von $\mathcal{T}$.

Literaturverzeichnis

[1] Dietrich Braess, *Finite Elemente. Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie.) 4th revised and extended ed.*, Springer, Berlin, 2007.

[2] Wolfgang Dahmen and Arnold Reusken, *Numerik für Ingenieure und Naturwissenschaftler*, Springer, Berlin, 2006.

[3] Peter Deuflhard and Folkmar Bornemann, *Numerische Mathematik 2*, revised ed., de Gruyter Lehrbuch., Walter de Gruyter & Co., Berlin, 2008, Gewöhnliche Differentialgleichungen.

[4] Hans Rudolf Schwarz and Norbert Köckler, *Numerische Mathematik*, 7th ed., Vieweg+Teubner, Wiesbaden, 2009.

[5] Rüdiger Verfürth, *Mathematik für Maschinenbauer, Bauingenieure und Umwelttechniker I-III*, Vorlesungsskriptum, Ruhr-Universität Bochum, Bochum, November 2006, 543 pages.

[6] ———, *Numerische Mathematik für Maschinenbauer, Bauingenieure und Umwelttechniker*, Vorlesungsskriptum, Ruhr-Universität Bochum, Bochum, Oktober 2006, 108 pages.

Index