Kapitel 4

Finite–Elemente–Methoden (FEM)

4.1 Das Ritzsche Verfahren

Bemerkung 4.1 Grundidee von Finite–Elemente–Methoden, das Ritz[1]sche Verfahren. Sei V ein Hilbert–Raum mit dem Skalarprodukt $a(\cdot,\cdot)$. Wir betrachten das Problem

$$\min_{v\in V} F(v) = \min_{v\in V}\left(\frac{1}{2}a(v,v) - f(v)\right),$$

wobei $f(\cdot) : V \to \mathbb{R}$ ein beschränktes lineares Funktional ist. Wie bereits bewiesen ist, besitzt das Variationsproblem eine eindeutig bestimmte Lösung $u \in V$, die außerdem die Gleichung

$$a(u,v) = f(v) \quad \forall\, v \in V \tag{4.1}$$

löst, Satz 3.36 (Rieszscher Darstellungssatz).

Um die Lösung der obigen Probleme mit einem numerischen Verfahren zu approximieren, setzen wir voraus, dass V ein separabler Hilbert–Raum ist, das heißt V besitzt eine abzählbare Basis. Dann gibt es endlich–dimensionale Teilräume $V_1, V_2, \ldots \subset V$ mit $\dim V_k = k$, die folgende Eigenschaft besitzen: zu jedem $u \in V$ und $\varepsilon > 0$ gibt es ein $K \in \mathbb{N}$ und ein $u_k \in V_k$ mit

$$\|u - u_k\|_V \leq \varepsilon \quad \forall\, k \geq K.$$

Es wird dabei nicht verlangt, dass es eine Inklusion der Form $V_k \subset V_{k+1}$ gibt.

Die Ritz–Approximation von (4.1) ist wie folgt definiert. Gesucht ist $u_k \in V_k$ mit

$$a(u_k, v_k) = f(v_k) \quad \forall\, v_k \in V_k. \tag{4.2}$$

Die wesentliche Idee des Ritzschen Verfahrens besteht also darin, dass man den unendlich–dimensionalen Raum V durch einen endlich–dimensionalen Raum V_k ersetzt. □

Lemma 4.2 Eigenschaften der Ritzschen Approximation.

1. *Der Fehler ist orthogonal zum Raum V_k, das heißt es gilt*

$$a(u - u_k, v_k) = 0 \quad \forall\, v_k \in V_k. \tag{4.3}$$

2. *u_k ist die Bestapproximierende von u in V_k bezüglich der von $a(\cdot,\cdot)$ induzierten Norm.*

[1] Walter Ritz (1878 – 1909)

3. *Die Folge der Ritz–Approximierenden konvergiert gegen die Lösung von (4.1), das heißt $u_k \to u$ für $k \to \infty$.*

Beweis: Da endlich–dimensionale Teilräume von Hilbert–Räumen wiederum Hilbert–Räume sind, besitzt nach dem Rieszschen Darstellungssatz auch die Gleichung der Ritz–Approximation eine eindeutige Lösung, die ebenso ein Minimierungsproblem im Raum V_k löst. Aus der Differenz der Gleichungen (4.1) und (4.2) erhält man die Orthogonalitätsrelation

$$a(u - u_k, v_k) = 0 \quad \forall\, v_k \in V_k.$$

Das besagt, dass der Fehler $u - u_k$ senkrecht zum Raum V_k ist: $u - u_k \perp V_k$. Demnach ist u_k die orthogonale Projektion von u in den Raum V_k bezüglich des Skalarproduktes von V. Das heißt, u_k ist die Bestapproximierende von u in V_k

$$\|u - u_k\|_V = \inf_{v_k \in V_k} \|u - v_k\|_V .$$

Zum Beweis nutzt man die Orthogonalität (4.3) und die Cauchy–Schwarz–Ungleichung. Sei $w_k \in V_k$ beliebig, dann ist

$$\begin{aligned} \|u - u_k\|_V^2 &= a(u - u_k, u - u_k) = a(u - u_k, u - \underbrace{(u_k - w_k)}_{v_k}) = a(u - u_k, u - v_k) \\ &\leq \|u - u_k\|_V \|u - v_k\|_V . \end{aligned}$$

Da $w_k \in V_k$ beliebig ist, ist auch $v_k \in V_k$ beliebig.

Mit der Bestapproximationseigenschaft erhält man

$$\|u - u_k\|_V = \inf_{v_k \in V_k} \|u - v_k\|_V \leq \varepsilon,$$

woraus schließlich die Konvergenz der Ritz–Approximation $u_k \to u$ für $k \to \infty$ folgt. ■

Bemerkung 4.3 Formulierung als lineares Gleichungssystem. Für die Berechnung der u_k kann man eine beliebige Basis $\{\phi_i\}_{i=1}^k$ von V_k verwenden. Zunächst gilt, dass die Gleichung der Ritz–Approximation (4.2) genau dann für alle $v_k \in V_k$ erfüllt ist, wenn sie für jede Basisfunktion ϕ_i erfüllt ist. Das folgt aus der Linearität der Gleichung bezüglich der Testfunktion und daraus, dass man jede Funktion $v_k \in V_k$ als Linearkombination der Basisfunktionen darstellen kann. Man setzt auch die Lösung als Linearkombination der Basisfunktionen an

$$u_k = \sum_{j=1}^{k} u^j \phi_j$$

mit unbekannten Koeffizienten $\mathbf{u} = (u^1, \ldots, u^k)^T$ und erhält, indem man als Testfunktionen jetzt die Basisfunktionen nutzt,

$$\sum_{j=1}^{k} a(u^j \phi_j, \phi_i) = f(\phi_i), \quad i = 1, \ldots, k.$$

Das ist äquivalent zu einem Gleichungssystem $A\mathbf{u} = \mathbf{b}$, wobei

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

Steifigkeitsmatrix genannt wird. Man beachte die unterschiedliche Reihenfolge der Indizes bei den Matrixeinträgen und beim Skalarprodukt. Die rechte Seite ist ein Vektor der Länge k mit den Einträgen $b_i = f(\phi_i)$.

Mit der eineindeutigen Zuordnung zwischen dem Koordinatenvektor $(u^1, \ldots, u^k)^T$ und dem Element $u_k = \sum_{i=1}^k u^i \phi_i$ lässt sich zeigen, dass die Matrix A symmetrisch und positiv definit ist:

$$\begin{aligned} A = A^T &\iff a(v, w) = a(w, v)\ \forall v, w \in V_k, \\ x^T A x > 0 \text{ für } x \neq 0 &\iff a(v, v) > 0\ \forall v \in V_k, v \neq 0. \end{aligned}$$

Übungsaufgabe □

Bemerkung 4.4 Der Fall einer unsymmetrischen Bilinearform. Im nichtvariationellen Fall, also wenn $b(\cdot,\cdot)$ unsymmetrisch, aber äquivalent zum Skalarprodukt $a(\cdot,\cdot)$ ist, kann man: Finde $u \in V$ mit

$$b(u,v) = f(v) \quad \forall\, v \in V \tag{4.4}$$

auch mit dem Ritzschen Verfahren approximieren. Die Eigenschaften von $b(\cdot,\cdot)$ seien Beschränktheit

$$|b(u,v)| \le M \|u\|_V \|v\|_V \quad M \in \mathbb{R},$$

und Koerzitivität

$$m \|v\|_V^2 \le b(v,v), \quad m > 0.$$

Das diskrete Problem lautet: Finde $u_k \in V_k$, so dass

$$b(u_k, v_k) = f(v_k) \quad \forall\, v_k \in V_k. \tag{4.5}$$

Die diskrete Lösung existiert eindeutig nach Satz 3.38 (Lax–Milgram). Sie ist jedoch keine orthogonale Projektion in V_k mehr. Trotzdem kann man die gleiche Fehlerabschätzung wie im variationellen Fall beweisen. □

Lemma 4.5 Lemma von Cea[2]. *Sei die Bilinearform $b(\cdot,\cdot)$ beschränkt und koerzitiv. Dann gilt*

$$\|u - u_k\|_V \le \frac{M}{m} \inf_{v_k \in V_k} \|u - v_k\|_V . \tag{4.6}$$

Beweis: Aus der Differenz der stetigen Gleichung (4.4) und der diskreten Gleichung (4.5)

$$b(u - u_k, v_k) = 0 \quad \forall\, v_k \in V_k$$

und

$$m \|v\|_V^2 \le b(v,v) \text{ und } |b(u,v)| \le M \|u\|_V \|v\|_V$$

folgt sofort

$$\begin{aligned} \|u - u_k\|_V^2 &\le \frac{1}{m} b(u - u_k, u - u_k) = \frac{1}{m} b(u - u_k, u - v_k) \\ &\le \frac{M}{m} \|u - u_k\|_V \|u - v_k\|_V . \end{aligned}$$

■

Bemerkung 4.6 Galerkin[3]–Methode. Im unsymmetrischen Fall wird dieses Verfahren Galerkin–Methode genannt. Das lineare Gleichungssystem wird genauso wie im symmetrischen Fall hergeleitet. Betrachte dazu das Zwei–Punkt–Randwertproblem

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

Die schwache Formulierung lautet: Finde $u \in H_0^1(0,1)$, so dass für alle $v \in H_0^1(0,1)$

$$\int_0^1 \Big(\varepsilon u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x) \Big)\, dx = \int_0^1 f(x)v(x)\, dx$$

gilt. Falls $(\cdot,\cdot)$ das Skalarprodukt in $L^2(0,1)$ bezeichnet, kann die schwache Formulierung übersichtlicher geschrieben werden

$$b(u,v) := \varepsilon(u', v') + (bu', v) + (cu, v) = (f, v).$$

[2] Cea

[3] Boris Grigorievich Galerkin (1871 – 1945)

Sei $\{\phi_i\}_{i=1}^k$ eine beliebige Basis von V_k, dann macht man wieder den Ansatz

$$u_k = \sum_{j=1}^{k} u^j \phi_j$$

mit unbekannten Koeffizienten $\mathbf{u} = (u^1, \ldots, u^k)^T$. Auch im nichtsymmetrischen Fall ist die variationelle Formulierung genau dann erfüllt, wenn sie für alle Basisfunktionen erfüllt ist. Man erhält

$$\sum_{j=1}^{k} \Big[\varepsilon(\phi_j', \phi_i') + (b\phi_j', \phi_i) + (c\phi_j, \phi_i)\Big] u^j = (f, \phi_i), \quad i = 1, \ldots, k,$$

was äquivalent zu einem Gleichungssystem $A\mathbf{u} = \mathbf{b}$ ist. Die Einträge der Steifigkeitsmatrix sind

$$a_{ij} = \varepsilon(\phi_j', \phi_i') + (b\phi_j', \phi_i) + (c\phi_j, \phi_i).$$

Die Systemmatrix ist nicht mehr symmetrisch.

Die Eigenschaften der Bilinearform wurden im Beispiel 3.35 untersucht. Sind $b, c \in L^\infty(0,1)$, so ist die Bilinearform beschränkt und die Konstante M ist in der Größenordnung von $\max\{\|b\|_\infty, \|c\|_\infty\}$. Gilt $-b'(x)/2 + c(x) \geq 0$, so ist sie koerzitiv mit $m = \varepsilon$. Falls beide Bedingungen erfüllt sind, dann ist das Lemma von Cea anwendbar und für den Fehler gilt

$$\|u - u_k\|_{H_0^1} \leq C \frac{\max\{\|b\|_\infty, \|c\|_\infty\}}{\varepsilon} \inf_{v_k \in V_k} \|u - v_k\|_{H_0^1}, \quad C \in \mathbb{R}.$$

Im singulär gestörten Fall $\varepsilon \ll \|b\|_\infty$ ist der erste Faktor in dieser Fehlerabschätzung sehr groß. □

4.2 Finite–Element–Räume in 1D

Bemerkung 4.7 Motivation für die Wahl der Räume beim Ritzschen Verfahren und der Galerkin–Methode. Der wichtigste Punkt beim Ritzschen Verfahren und bei der Galerkin–Methode ist die Wahl der Räume V_k, oder genauer, die Wahl von geeigneten Basen $\{\phi_i\}_{i=1}^k$, die einen Raum V_k aufspannen. In dieser Vorlesung wird nur der Fall betrachtet, dass $V_k \subset V$ gilt. Es gibt auch Finite–Elemente–Methoden, bei denen diese Eigenschaft nicht erfüllt ist.

Vom numerischen Standpunkt aus sollten die Elemente a_{ij} der Steifigkeitsmatrix schnell zu berechnen sein und die Matrix A sollte nur schwach besetzt sein, das heißt sie sollte viele Nulleinträge besitzen. Das führt auf folgende Überlegungen:

- Die Einträge von A berechnen sich mit Hilfe von Integralen, welche die Ansatz– und Testfunktionen, sowie deren Ableitungen enthalten. Funktionen, für die sich solche Integrale besonders einfach berechnen lassen, sind Polynome.
- Falls man Basisfunktionen $\{\phi_i(x)\}_{i=1}^k$ wählt, die im gesamten Intervall $(0,1)$ ungleich Null sein können, so werden im allgemeinen nur sehr wenige Integrale verschwinden und nur wenige Einträge der Matrix A werden Null. Deshalb ist es zweckmäßig Funktionen zu verwenden, die nur auf einem kleinen Teil von $(0,1)$ nicht Null sind.
- Wegen $V_k \subset V (= H_0^1(a,b))$, müssen die Funktionen aus V_k stetig sein, vergleiche Bemerkung 3.22.

Aus diesen Gründen bietet es sich an, als Basis stetige Funktionen zu verwenden, die stückweise polynomial sind.

Analog zu den Finite–Differenzen–Verfahren wird $[0,1]$ mittels eines (zunächst) äquidistanten Gitters mit den Gitterpunkten

$$x_i = ih, \quad i = 0, \ldots, N, \quad h = 1/N,$$

zerlegt. Die Intervalle $K_i = (x_i, x_{i+1})$ werden Gitterzellen genannt. Ihre Vereinigung

$$\mathcal{T}_h = \bigcup_{i=0}^{N-1} \overline{K_i}$$

heißt Triangulierung. □

Bemerkung 4.8 Träger von Finite–Elemente–Funktionen. Ein Kriterium zur Konstruktion von geeigneten Basisfunktionen für Finite–Elemente ist, dass ihr Träger, siehe Definition 3.11, möglichst klein sein soll. Wenn das der Fall ist, dann ist es sehr wahrscheinlich, dass der gemeinsame Träger von zwei verschiedenen Basisfunktionen $\phi_i(x), \phi_j(x)$ das Maß Null hat, zum Beispiel leer ist oder nur ein Punkt ist. In diesem Fall sind alle Integrale für die betreffenden Komponenten a_{ij} und a_{ji} gleich Null, man hat also Nulleinträge in der Matrix.

In 1D muss es Basisfunktionen $\phi_i(x)$ geben, deren Träger aus zwei benachbarten Gitterzellen $[x_{i-1}, x_i] \cup [x_i, x_{i+1}]$ besteht. Ansonsten müssten wegen der Stetigkeit alle Basisfunktionen in den Gitterpunkten $x_0, \ldots, x_N$ verschwinden, womit man nur Funktionen approximieren könnte, die diese Eigenschaft haben. Das wird aber im allgemeinen für die Lösung einer Differentialgleichung nicht gelten. □

Beispiel 4.9 Stückweise lineare Basisfunktionen. Die einfachsten Basisfunktionen besitzen einen Träger aus zwei benachbarten Gitterzellen und sie sind stückweise linear. Sie sind eindeutig durch ihre Werte in den Gitterpunkten bestimmt

$$\phi_i(x_j) = \delta_{ij} = \begin{cases} 1 & \text{für } i = j, \\ 0 & \text{sonst.} \end{cases}, \quad i, j = 1, \ldots, N-1.$$

Die Darstellung als Formel ist

$$\phi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}} & \text{für } x \in [x_{i-1}, x_i], \\ \dfrac{x_{i+1} - x}{x_{i+1} - x_i} & \text{für } x \in [x_i, x_{i+1}], \\ 0 & \text{sonst.} \end{cases}$$

Wegen ihrer charakteristischen Form werden diese Funktionen auch Hütchenfunktionen genannt, siehe Abbildung 4.1. In jeder Gitterzelle $[x_{i-1}, x_i]$ gibt es höchstens zwei (in den inneren Gitterzellen genau zwei) Basisfunktionen, bei welchen diese Gitterzelle eine Teilmenge ihres Trägers ist. Eine der Basisfunktionen nimmt den Wert Eins in x_{i-1} an und Null in x_i, bei der anderen Basisfunktion ist es genau umgekehrt. Somit erhält man für die Testfunktion $\phi_i(x)$ höchsten bei den Ansatzfunktionen $\phi_{i-1}(x), \phi_i(x), \phi_{i+1}(x)$ Nichtnulleinträge. Das bedeutet, in jeder Zeile der Matrix A gibt es höchstens drei Nichtnulleinträge.

Der aufgespannte Finite–Element–Raum $\text{span}\{\phi_i(x)\}_{i=1}^{N-1}$ wird P_1 genannt und er besitzt die Dimension $N-1$. □

Beispiel 4.10 Stückweise quadratische Basisfunktionen. Eine Erweiterung besteht nun darin, stückweise quadratische Basisfunktionen zu betrachten. Auch hier soll der Träger jeder Basisfunktion höchstens die Vereinigung zweier benachbarter Gitterzellen sein. Eine quadratische Funktion ist durch die Vorgabe von drei

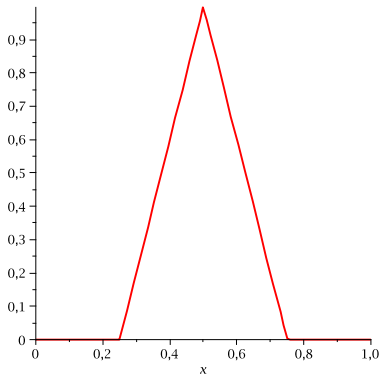

Abbildung 4.1: Hütchenfunktion in $[0.25, 0.5] \cup [0.5, 0.25]$.

Punkten eindeutig festgelegt. In einem Intervall $[x_{i-1}, x_i]$ werden dazu die Werte in den Gitterpunkten x_{i-1} und x_i sowie im Mittelpunkt $(x_{i-1} + x_i)/2$ vorgegeben. Die Gesamtheit der Punkte, in denen man Werte vorgibt, nennt man Knoten. Man hat also $2N - 1$ Knoten ξ_i. Die stückweise quadratische Basis wird so gewählt, dass gilt

$$\phi_i(\xi_j) = \delta_{ij}, \quad i, j = 1, \ldots, 2N - 1.$$

Darstellungsformeln: Übungsaufgabe Damit gibt es zwei Typen von Basisfunktionen, siehe Abbildung 4.2. Ist ξ_j ein Gitterpunkt, dann besteht der Träger der Basisfunktion aus zwei benachbarten Gitterzellen. Für die Testfunktion $\phi_j(x)$ wird man Nichtnulleinträge im allgemeinen dann bekommen, wenn die Ansatzfunktion aus der Menge $\{\phi_{j-2}(x), \phi_{j-1}(x), \phi_j(x), \phi_{j+1}(x), \phi_{j+2}(x)\}$ kommt. Ist ξ_j kein Gitterpunkt, dann ist der Träger sogar nur eine Gitterzelle. Man spricht auch von Blasenfunktionen. Hier wird es Nichtnulleinträge bei Ansatzfunktionen aus der Menge $\{\phi_{j-1}(x), \phi_j(x), \phi_{j+1}(x)\}$ geben.

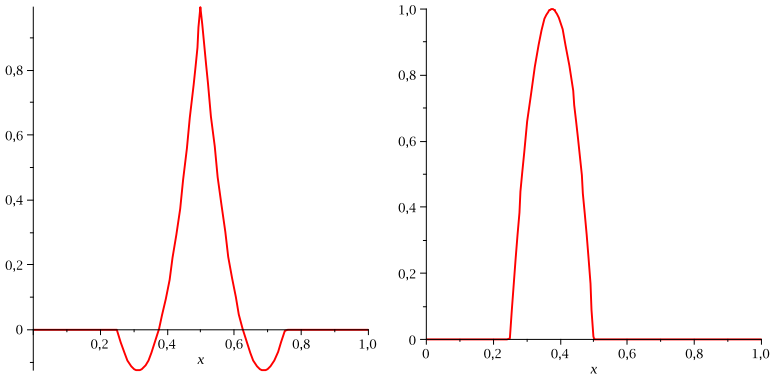

Abbildung 4.2: Quadratische Basisfunktionen in $[0.25, 0.5] \cup [0.5, 0.25]$ beziehungsweise in $[0.25, 0.5]$.

Der aufgespannte Finite–Element–Raum $\text{span}\{\phi_i(x)\}_{i=1}^{2N-1}$ wird P_2 genannt und er besitzt die Dimension $2N-1$. Man hat mit diesem Raum mehr Aufwand als mit P_1 (höhere Dimension des Gleichungssystems, mehr Nichtnulleinträge in der Matrix), aber man wird im allgemeinen auch genauere Ergebnisse erwarten. □

Bemerkung 4.11 Finite–Elemente höherer Ordnung. Diese Konstruktionen lassen sich natürlich fortsetzten. Bei finite Elementen höherer Ordnung gibt es je-

doch unterschiedliche Ansätze, um die Knoten im Inneren der Gitterzelle zu wählen, zumindest in 1D, vergleiche [Sol06]. □

Bemerkung 4.12 Affines Konzept, Referenzzelle, Referenzabbildung. Die Finite–Elemente–Räume in den vorangegangenen Beispielen wurden direkt auf den Zellen des Gitter definiert. Es gibt jedoch noch eine andere Herangehensweise. Bei dieser werden die Basisfunktionen mit ihren Eigenschaften auf einer Referenzzelle $\hat{K}$ definiert. Die Basisfunktionen auf dem Gitter ergeben sich dann durch Referenzabbildungen $F_K : \hat{K} \to K$ auf die Gitterzellen. Falls die Referenzabbildungen affin sind (lineare Abbildung plus konstante Verschiebung), dann sind beide Definitionen oft äquivalent. Die Referenzabbildungen hängen nur von der Gestalt der Gitterzellen ab, aber nicht vom Finite–Element–Raum.

Mit dem affinen Konzept ist zunächst nur definiert, was auf jeder Gitterzelle passiert. Um die Definition eines Finite–Elemente–Raumes zu vervollständigen, man muss noch zusätzliche Eigenschaften für den Übergang zwischen benachbarten Gitterzellen fordern, beispielsweise Stetigkeit für die Räume P_1 und P_2.

Dieses affine Konzept besitzt viele Vorteile bei der Implementierung von Finite–Element–Methoden, da man alle benötigten Informationen (Basisfunktionen, Quadraturformeln) nur auf der Referenzzelle zu programmieren braucht. □

Beispiel 4.13 Affines Konzept für P_1 in 1D. Man nimmt als Referenzgitterzelle beispielsweise $\hat{K} = [-1, 1]$. Die Referenzabbildung auf eine Gitterzelle $K = [x_i, x_{i+1}]$ wird so definiert, dass sie affin ist, das heißt es gilt

$$F_K(\hat{x}) = \alpha\hat{x} + \beta = x,$$

den Punkt -1 bildet man auf x_i sowie den Punkt 1 auf x_{i+1} ab. Das heißt

$$\begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} x_i \\ x_{i+1} \end{pmatrix} \quad \Longrightarrow \quad \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \frac{1}{2} \begin{pmatrix} x_{i+1} - x_i \\ x_{i+1} + x_i \end{pmatrix}.$$

Auf $\hat{K}$ definiert man nun zwei lineare Basisfunktionen

$$\hat{\phi}_1(\hat{x}) = \frac{1}{2}(-\hat{x} + 1), \quad \hat{\phi}_2(\hat{x}) = \frac{1}{2}(\hat{x} + 1).$$

Die erste Funktion ist im Punkt -1 gleich Eins und verschwindet im Punkt 1, bei der zweiten ist es genau umgekehrt.

Die Rücktransformation $F_K^{-1} : K \to \hat{K}$ von der Gitterzelle K auf die Referenzzelle $\hat{K}$ besitzt die Gestalt

$$\hat{x} = \frac{x - \beta}{\alpha} = \frac{2x - (x_{i+1} + x_i)}{x_{i+1} - x_i}.$$

Die Basisfunktionen auf K sind definiert durch

$$\phi_i(x) := \hat{\phi}_i \left(F_K^{-1}(x) \right), \quad i = 1, 2.$$

Damit erhält man

$$\begin{aligned} \phi_1(x) &= \hat{\phi}_1 \left(\frac{2x - (x_{i+1} + x_i)}{x_{i+1} - x_i} \right) = \frac{1}{2} \left(-\frac{2x - (x_{i+1} + x_i)}{x_{i+1} - x_i} + 1 \right) \\ &= -\frac{1}{2} \left(\frac{2x - 2x_{i+1}}{x_{i+1} - x_i} \right) = \frac{x_{i+1} - x}{x_{i+1} - x_i}, \\ \phi_2(x) &= \hat{\phi}_2 \left(\frac{2x - (x_{i+1} + x_i)}{x_{i+1} - x_i} \right) = \frac{x - x_i}{x_{i+1} - x_i}. \end{aligned}$$

Das sind gerade die beiden Basisfunktionen, die man mit direkter Definition auf der Zelle K erhält, Beispiel 4.9. □

Bemerkung 4.14 Assemblierung: Berechnung der Matrixeinträge und der rechten Seite. Die Matrixeinträge des Modellproblems besitzen die Form, siehe Bemerkung 4.6,

$$\begin{aligned} a_{ij} &= \int_0^1 \Big(\varepsilon\phi_j'(x)\phi_i'(x) + b(x)\phi_j'(x)\phi_i(x) + c(x)\phi_j(x)\phi_i(x)\Big)\ dx \\ &= \sum_{k=0}^{N-1} \int_{x_k}^{x_{k+1}} \Big(\varepsilon\phi_j'(x)\phi_i'(x) + b(x)\phi_j'(x)\phi_i(x) + c(x)\phi_j(x)\phi_i(x)\Big)\ dx. \end{aligned}$$

Das heißt, man kann die Integrale auf den einzelnen Gitterzellen berechnen und dann aufsummieren. Beim P_1–Finite–Element hat man für $i = j$ Integrale auf genau zwei Gitterzellen zu berechnen, für $i = j \pm 1$ auf genau einer Gitterzelle und sonst sind alle Integrale Null. Für die Assemblierung der rechten Seite gilt eine analoge Formel.

Es bestehen die Möglichkeiten, die Integrale direkt auf einer Gitterzelle K zu berechnen oder das Integral auf die Referenzzelle $\hat{K}$ zu transformieren. Der zweite Weg ist für die Implementierung von Finite–Element–Methoden günstiger. Nach der Transformation auf $\hat{K}$ kann man eine Quadraturformel anwenden, die für das Referenzelement implementiert ist. Man muss sich jedoch ansehen, wie sich die Terme in den Integralen transformieren.

Die Transformation der Integrale erfolgt natürlich mit der Substitutionsregel unter Verwendung der Referenzabbildung $F_K : [-1,1] \to [x_k, x_{k+1}]$

$$\int_{x_k}^{x_{k+1}} f(x)\ dx = \int_{-1}^{1} f\left(F_K(\hat{x})\right) F_K'(\hat{x})\ d\hat{x}.$$

Es gilt, siehe Beispiel 4.13,

$$F_K(\hat{x}) = \frac{1}{2}\big((x_{k+1} - x_k)\hat{x} + (x_{k+1} + x_k)\big) = x \quad \Longrightarrow \quad F_K'(\hat{x}) = \frac{x_{k+1} - x_k}{2}.$$

Das ist die halbe Länge der Gitterzelle $[x_k, x_{k+1}]$, woraus $F_K'(\hat{x}) > 0$ folgt. Für die Basisfunktion ist die Transformation auf die Referenzzelle durch

$$\phi_i(x) = \hat{\phi}_i\left(F_K^{-1}(x)\right) = \hat{\phi}_i(\hat{x})$$

gegeben. Zur Transformation der Ableitung verwendet man die Kettenregel

$$\phi_i'(x) = \frac{d\phi_i(x)}{dx} = \frac{d\hat{\phi}_i(\hat{x})}{dx} = \frac{d\hat{\phi}_i(\hat{x})}{d\hat{x}}\frac{d\hat{x}}{dx} = \frac{d\hat{\phi}_i(\hat{x})}{d\hat{x}}\frac{2}{x_{k+1} - x_k}.$$

Die Ableitungen der Basisfunktionen auf der Referenzzelle kann man vorher explizit ausrechnen und dann implementieren. Damit erhält man

$$\begin{aligned} \int_{x_k}^{x_{k+1}} \varepsilon\phi_j'(x)\phi_i'(x)\ dx &= \frac{2}{x_{k+1} - x_k}\int_{-1}^{1} \varepsilon\frac{d\hat{\phi}_j(\hat{x})}{d\hat{x}}\frac{d\hat{\phi}_i(\hat{x})}{d\hat{x}}\ d\hat{x}, \\ \int_{x_k}^{x_{k+1}} b(x)\phi_j'(x)\phi_i(x)\ dx &= \int_{-1}^{1} b(F_K(\hat{x}))\frac{d\hat{\phi}_j(\hat{x})}{d\hat{x}}\hat{\phi}_i(\hat{x})\ d\hat{x}, \\ \int_{x_k}^{x_{k+1}} c(x)\phi_j(x)\phi_i(x)dx &= \frac{x_{k+1} - x_k}{2}\int_{-1}^{1} c(F_K(\hat{x}))\hat{\phi}_j(\hat{x})\hat{\phi}_i(\hat{x})\ d\hat{x}, \\ \int_{x_k}^{x_{k+1}} f(x)\phi_i(x)dx &= \frac{x_{k+1} - x_k}{2}\int_{-1}^{1} f(F_K(\hat{x}))\hat{\phi}_i(\hat{x})\ d\hat{x}. \end{aligned}$$

Nun kann man hinreichend genaue Quadraturformeln auf der Referenzzelle zur Approximation der Integrale verwenden. Eine genaue Quadratur ist vor allem für finite Elemente höherer Ordnung wesentlich, damit die Genauigkeit nicht durch Quadraturfehler beeinträchtigt wird. □

Beispiel 4.15 Assemblierung: P_1 in 1D. Betrachte den Fall, dass die Parameterfunktionen konstant sind, $b(x) = b$, $c(x) = c$ und $f(x) = f$. Für das P_1–Finite–Element auf der Referenzzelle $[-1, 1]$ gelten

$$\hat{\phi}_1(\hat{x}) = \frac{1}{2}(-\hat{x} + 1), \quad \hat{\phi}_1'(\hat{x}) = -\frac{1}{2}, \qquad \hat{\phi}_2(\hat{x}) = \frac{1}{2}(\hat{x} + 1), \quad \hat{\phi}_2'(\hat{x}) = \frac{1}{2}.$$

Betrachte nun den Matrixeintrag $a_{i,i+1}$, der mit Hilfe der Testfunktion $\phi_i(x)$, die auf $\hat{\phi}_1(\hat{x})$ transformiert wird, und der Ansatzfunktion $\phi_{i+1}(x)$, die auf $\hat{\phi}_2(\hat{x})$ transformiert wird, berechnet wird. Der gemeinsame Träger ist die Gitterzelle $[x_i, x_{i+1}]$. Bezeichne h die Länge dieser Zelle. Dann folgt

$$\begin{aligned} a_{i,i+1} &= \frac{2\varepsilon}{h} \int_{-1}^{1} \frac{1}{2} \cdot \left(-\frac{1}{2}\right)\, d\hat{x} + b \int_{-1}^{1} \frac{1}{2} \cdot \frac{1}{2}(-\hat{x} + 1)\, d\hat{x} \\ &\quad + \frac{ch}{2} \int_{-1}^{1} \frac{1}{2}(\hat{x} + 1) \frac{1}{2}(-\hat{x} + 1)\, d\hat{x} \\ &= -\frac{\varepsilon}{h} + \frac{b}{2} + \frac{ch}{6}. \end{aligned}$$

Für die i–te Komponente der rechten Seite, erhält, man

$$f_i = f \int_0^1 \phi_i(x)\, dx = hf,$$

da die Fläche unter einer Hütchenfunktion das Maß h besitzt. *andere Einträge als Übungsaufgabe*

Verwendet man zur Approximation der Integrale die Trapezregel, so ergibt sich

$$\frac{ch}{2} \int_{-1}^{1} \frac{1}{2}(-\hat{x} + 1) \frac{1}{2}(\hat{x} + 1)\, d\hat{x} = \frac{ch}{2} 2(0 + 0) = 0.$$

In diesem Fall ist

$$a_{i,i+1} = -\frac{\varepsilon}{h} + \frac{b}{2}.$$

Für $c = 0$ (oder mit Trapezregel) sind das, bis auf den Faktor h, die gleichen Einträge wie beim zentralen Differenzenverfahren, siehe Bemerkung 2.11. Man erhält für $c = 0$

$$-h\varepsilon D^+ D^- u_i + h b_i D^0 u_i = h f_i \quad \Longleftrightarrow \quad -\varepsilon D^+ D^- u_i + b_i D^0 u_i = f_i.$$

Dieser Zusammenhang zwischen Finite–Differenzen–Methoden und Finite–Element–Methoden gilt im allgemeinen nicht mehr, wenn die Koeffizientenfunktionen nicht konstant sind. In höheren Dimensionen unterscheiden sich FDM und FEM im allgemeinen auch bei konstanten Koeffizienten. □

Bemerkung 4.16 Andere Randbedingungen. Hat man inhomogene Dirichlet–Randbedingungen

$$u(0) = a \quad u(1) = b$$

oder Neumann–Randbedingungen

$$\varepsilon u'(0) = \alpha, \quad \varepsilon u'(1) = \beta,$$

so nimmt man zur Gleichungsassemblierung auch die Basisfunktionen hinzu, die in den Randpunkten den Wert Eins haben und in allen anderen Knoten verschwinden. Bei inhomogenen Dirichlet–Randbedingungen ersetzt man dann die entsprechenden

Gleichungen (bei Numerierung von links nach rechts die erste und letzte Gleichung) durch

$$(1,0,\ldots,0)\begin{pmatrix} u^0 \\ \vdots \end{pmatrix} = \begin{pmatrix} a \\ \vdots \end{pmatrix}, \quad (0,0,\ldots,1)\begin{pmatrix} \vdots \\ u^k \end{pmatrix} = \begin{pmatrix} \vdots \\ b \end{pmatrix}.$$

Bei Neumann–Randbedingungen treten in natürlicher Art und Weise, siehe Bemerkung 3.33, zusätzliche Terme auf der rechten Seite der ersten und letzten Gleichung auf. □

Bemerkung 4.17 CSR–Speicherschema von schwach besetzten Matrizen. Von schwach besetzten Matrizen speichert man natürlich nur die Einträge, die nicht Null sind und zugehörige Informationen über die Position des Eintrags. Die am weitesten verbreiteste Herangehensweise ist das CSR–Speicherschema (condensed sparse row). Bei diesem Schema werden die Nichtnulleinträge zeilenweise abgespeichert. Innerhalb einer Zeile brauchen sie nicht bezüglich der Spaltenindizes angeordnet zu werden.

Sei eine schwach besetzte Matrix $A \in \mathbb{R}^{m\times n}$ mit nnz Nichtnullelementen zu speicher. Dann braucht man drei Arrays:

- `double`–Array `entries` der Länge nnz, darin werden die Einträge von A zeilenweise gespeichert,
- `int`–Array `col_ptr` der Länge nnz, darin stehen die Spaltenindizes der zugehörigen Einträge von `entries`.
- `int`–Array `row_ptr` der Länge $m+1$, darin wird abgespeichert, an welcher Stelle im Array `entries` die i–te Zeile beginnt, $i = 1,\ldots,m$; der letzte Eintrag von `row_ptr` verweist auf den ersten Speicherplatz nach dem Ende des Arrays `entries`,

□

Beispiel 4.18 Die Matrix

$$A = \begin{pmatrix} 1 & 0 & 0 & 2 & 0 \\ 3 & 4 & 0 & 5 & 0 \\ 6 & 0 & 7 & 8 & 9 \\ 0 & 0 & 10 & 11 & 0 \\ 0 & 0 & 0 & 0 & 12 \end{pmatrix}$$

kann wie folgt gespeichert werden (Numerierung beginnt bei 0):

`entries`	–	1	2	3	4	5	6	7	8	9	10	11	12
`col_ptr`	–	0	3	0	1	3	0	2	3	4	2	3	4 .
`row_ptr`	–	0	2	5	9	11	12						

Eine andere Möglichkeit ist

`entries`	–	2	1	4	5	3	7	9	8	6	11	10	12
`col_ptr`	–	3	0	1	3	0	2	4	3	0	3	2	4 .
`row_ptr`	–	0	2	5	9	11	12						

□

4.3 Polynominterpolation in Sobolov–Räumen und Konvergenzabschätzungen

Bemerkung 4.19 Motivation. Die variationelle Formulierung partieller Differentialgleichungen benutzt Funktionen aus Sobolev–Räumen. Die Lösung soll mit Hilfe

der Ritzschen Methode und endlich–dimensionaler Finite–Element–Räume approximiert werden. Der Fehler in der durch den Raum V induzierten Norm hängt davon ab, wie gut man Funktionen aus Sobolev–Räumen überhaupt mit Funktionen aus Finite–Element–Räumen annähern kann, siehe zum Beispiel das Lemma von Cea, Abschätzung (4.6). Die Approximationsgüte von Finite–Element–Räumen wird in diesem Abschnitt untersucht. □

4.3.1 Das Bramble–Hilbert–Lemma

Wir beginnen mit grundlegenden Prinzipien der Polynominterpolation in Sobolev–Räumen.

Lemma 4.20 *Sei $(a,b) \subset \mathbb{R}$. Für jeden Index γ mit $0 \leq \gamma \leq m$ sei ein $a_\gamma \in \mathbb{R}$ gegeben. Dann gibt es ein eindeutig bestimmtes Polynom $p \in P_m(a,b)$ mit*

$$\int_a^b p^{(\gamma)}(x)\ dx = a_\gamma,\ 0 \leq \gamma \leq m.$$

Beweis: Jedes Polynom aus $P_m(a,b)$ hat die Gestalt

$$p(x) = \sum_{\mu=0}^{m} b_\mu x^\mu.$$

Einsetzen dieser Darstellung in die Bedingungen ergibt ein lineares Gleichungssystem $M\mathbf{b} = \mathbf{a}$, mit

$$M = (M_{\gamma\mu}),\ M_{\gamma\mu} = \int_a^b (x^\mu)^{(\gamma)}\ dx,\ \mathbf{b} = (b_\mu),\ \mathbf{a} = (a_\gamma),$$

für $0 \leq \gamma, \mu \leq m$. Das ist ein quadratisches Gleichungssystem, welches genau dann eine eindeutige Lösung besitzt, wenn M regulär ist.

Angenommen, M ist singulär. Dann besitzt das zugehörige homogene Gleichungssystem eine nichttriviale Lösung. Dass heißt, es gibt ein Polynom $q \in P_m(a,b) \setminus \{0\}$ mit

$$\int_a^b q^{(\gamma)}(x)\ dx = 0 \text{ für alle } 0 \leq \gamma \leq m.$$

Das Polynom q besitzt die Darstellung $q(x) = \sum_{\mu=0}^m c_\mu x^\mu$. Wähle nun das $c_\mu \neq 0$ mit maximalem μ. Dann gilt $q^{(\mu)}(x) = \mu(\mu-1)\ldots\cdot 2\cdot 1\cdot c_\mu = const \neq 0$, woraus

$$\int_a^b q^{(\mu)}(x)\ dx = \int_a^b const\ dx = (b-a)const \neq 0$$

folgt. Das widerspricht dem Verschwinden des Integrals für $q^{(\mu)}(x)$. Somit ist die Annahme falsch und M ist nicht singulär. ■

Das Lemma besagt, dass ein Polynom eindeutig bestimmt ist, wenn man für jede Ableitung eine Bedingung an das Integral über (a,b) stellt.

Lemma 4.21 Ungleichung vom Poincaré–Typus. *Sei (a,b) mit $R = b-a$. Seien $k,l \in \mathbb{N}$ mit $0 \leq k \leq l$ und sei $p \in \mathbb{R}$ mit $p \in [1,\infty]$. Dann gilt für jedes $v \in W^{l,p}(a,b)$, welches*

$$\int_a^b v^{(\gamma)}(x)\ dx = 0 \text{ für alle } 0 \leq \gamma \leq l-1$$

erfüllt, die Abschätzung

$$\left\|v^{(k)}\right\|_{L^p(a,b)} \leq CR^{l-k}\left\|v^{(l)}\right\|_{L^p(a,b)},$$

wobei die Konstante c nicht von (a,b) und von $v(x)$ abhängt.

Beweis: Im Fall $k = l$ braucht man nichts zu beweisen. Des weiteren genügt es, das Lemma für $k = 0$ und $l = 1$ zu beweisen, da der allgemeine Fall folgt, wenn man das Resultat dann auf die γ–te Ableitung anwendet.

Zu zeigen ist also

$$\|v\|_{L^p(a,b)} \le CR \left\|v'\right\|_{L^p(a,b)} \quad \text{falls} \quad \int_a^b v(x)\, dx = 0. \tag{4.7}$$

Es gilt für $x, y \in (a, b)$

$$\begin{aligned} \int_0^1 v'(tx + (1-t)y)\, dt &= \int_0^1 v'(t(x-y) + y)\, dt \\ &= \frac{1}{x-y}\Big(v(t(x-y)+y)|_{t=1} - v(t(x-y)+y)|_{t=0}\Big) \\ &= \frac{v(x) - v(y)}{x-y}, \end{aligned}$$

was eine Form des Mittelwertsatzes ist. Multiplikation mit $(x - y)$ und anschließende Integration bezüglich y ergibt

$$v(x)\int_a^b dy - \underbrace{\int_a^b v(y)\, dy}_{=0} = \int_a^b \int_0^1 v'(tx + (1-t)y)(x-y)\, dt\, dy,$$

wobei das eine Integral auf der linken Seite nach Voraussetzung an $v(x)$ verschwindet. Damit wurde schon die Voraussetzung von (4.7) verwendet. Es folgt

$$v(x) = \frac{1}{R}\int_a^b \int_0^1 v'(tx + (1-t)y)(x-y)\, dt\, dy.$$

Nun muss man versuchen, die Terme der Behauptung in (4.7) zu bekommen. Man beginnt mit der linken Seite der Ungleichung. Es wird verwendet, dass der Betrag eines Integrals durch das Integral des Betrags abgeschätzt werden kann, sowie dass $|x - y| \le R$ gilt

$$|v(x)| \le \frac{1}{R}\int_a^b \int_0^1 \left|v'(tx + (1-t)y)\right| |x-y|\, dt\, dy \le \frac{R}{R}\int_a^b \int_0^1 \left|v'(tx + (1-t)y)\right|\, dt\, dy. \tag{4.8}$$

Für $p < \infty$ wird diese Abschätzung mit p potenziert und bezüglich x integriert. Man erhält durch Anwendung der Hölderschen Ungleichung mit $p^{-1} + q^{-1} = 1$

$$\begin{aligned} \int_a^b |v(x)|^p\, dx &\le \int_a^b \left(\int_a^b \int_0^1 \left|v'(tx + (1-t)y)\right|\, dt\, dy\right)^p dx \\ &\le \int_a^b \left[\underbrace{\left(\int_a^b \int_0^1 1^q\, dt\, dy\right)^{p/q}}_{R^{p/q}} \left(\int_a^b \int_0^1 \left|v'(tx + (1-t)y)\right|^p\, dt\, dy\right)\right] dx \\ &= R^{p/q} \int_a^b \left(\int_a^b \int_0^1 \left|v'(tx + (1-t)y)\right|^p\, dt\, dy\right) dx. \end{aligned}$$

Damit hat man die p–te Potenz der linken Seite der Ungleichung in (4.7). Nun braucht man noch die p–te Potzen der rechten Seite der Ungleichung. Es werden zunächst die Integrationen vertauscht (Satz von Fubini)

$$\int_a^b |v(x)|^p\, dx \le R^{p/q} \int_0^1 \int_a^b \left(\int_a^b \left|v'(tx + (1-t)y)\right|^p\, dy\right) dx\, dt.$$

Mit dem Mittelwertsatz der Integralrechnung findet man ein $t_0 \in [0, 1]$, so dass

$$\int_a^b |v(x)|^p\, dx \le R^{p/q}(1-0) \int_a^b \left(\int_a^b \left|v'(t_0 x + (1-t_0)y)\right|^p\, dy\right) dx.$$

Man setzt $|v'(x)|^p$ auf $\mathbb{R}$ durch Null fort und nennt die Fortsetzung ebenfalls $|v'(x)|^p$. Dann ist

$$\int_a^b |v(x)|^p \ dx \leq R^{p/q} \int_a^b \left(\int_{\mathbb{R}} |v'(t_0 x + (1-t_0)y)|^p \ dy \right) dx.$$

Sei $t_0 \in [0, 1/2]$. Da das Integrationsgebiet nun der ganze $\mathbb{R}$ ist, ergibt die Variablensubstitution $t_0 x + (1-t_0)y = z$

$$\int_{\mathbb{R}} |v'(t_0 x + (1-t_0)y)|^p \ dy = \frac{1}{1-t_0} \int_{\mathbb{R}} |v'(z)|^p \ dz \leq 2 \left\| v' \right\|^p_{L^p(a,b)},$$

da $1/(1-t_0) \leq 2$. Führt man nun noch die äußere Integration über x aus, erhält man insgesamt

$$\begin{aligned} \int_a^b |v(x)|^p \ dx &\leq R^{p/q} \int_a^b 2 \left\| v' \right\|^p_{L^p(a,b)} \ dx = 2R^{p/q} \left\| v' \right\|^p_{L^p(a,b)} \int_a^b \ dx \\ &= 2R^{p/q+1} \left\| v' \right\|^p_{L^p(a,b)} = 2R^p \left\| v' \right\|^p_{L^p(a,b)}. \end{aligned}$$

Im Fall $t_0 > 1/2$ vertauscht man die Rollen von x und y sowie die Integrationsreihenfolge mit dem Satz von Fubini und argumentiert analog.

Der Fall $p = \infty$ folgt aus (4.8). *Übungsaufgabe* ■

Bemerkung 4.22 Das Lemma besagt, dass man die $L^p(a,b)$–Norm einer niederen Ableitung von $v(x)$ durch dieselbe Norm einer Ableitung höherer Ordnung abschätzen kann, falls die Integralmittelwerte der niederen Ableitungen verschwinden. Eine wichtige Anwendung dieses Lemmas ist der Beweis des Bramble–Hilbert–Lemmas. Dieses besagt, dass der Wert eines stetigen linearen Funktionals, das auf einem Sobolev–Raum definiert ist und auf einem Polynomraum der Ordnung m verschwindet, durch die Lebesgue–Norm der $m+1$–ten Ableitung der Funktionen aus dem Sobolev–Raum abgeschätzt werden kann. □

Satz 4.23 ***Bramble***[4]***–Hilbert–Lemma.*** *Seien $m \in \mathbb{N}$, $m \geq 0$, $p \in [1, \infty]$ und $F : W^{m+1,p}(a,b) \to \mathbb{R}$ ein stetiges lineares Funktional und seien die Voraussetzungen der Lemmata 4.20 und 4.21 erfüllt. Weiter sei*

$$F(p) = 0 \quad \forall \ p \in P_m(a,b).$$

Dann gibt es eine Konstante $c(a,b)$, die unabhängig von $v(x)$ und F ist, mit

$$|F(v)| \leq c(a,b) \left\| v^{(m+1)} \right\|_{L^p(a,b)} \quad \forall \ v \in W^{m+1,p}(a,b).$$

Beweis: Sei $v \in W^{m+1,p}(a,b)$ mit

$$\int_a^b v^{(\gamma)}(x) \ dx = a_\gamma \text{ für } 0 \leq \gamma \leq m.$$

Wegen Lemma 4.20 gibt es ein Polynom aus $P_m(a,b)$ mit

$$\int_a^b p^{(\gamma)}(x) \ dx = -a_\gamma, \quad 0 \leq \gamma \leq m, \quad \Longrightarrow \quad \int_a^b (v+p)^{(\gamma)}(x) \ dx = 0, \quad 0 \leq \gamma \leq m.$$

Lemma 4.21 liefert, mit $l = m+1$, nun die Abschätzung

$$\begin{aligned} \|v+p\|_{W^{m+1,p}(a,b)} &= \left(\sum_{i=0}^{m+1} \left\| (v+p)^{(i)} \right\|^p_{L^p(a,b)} \right)^{1/p} \\ &\leq \left(\sum_{i=0}^{m+1} c_i(a,b) \left\| (v+p)^{(m+1)} \right\|^p_{L^p(a,b)} \right)^{1/p} \\ &= \left(\sum_{i=0}^{m+1} c_i(a,b) \right)^{1/p} \left\| (v+p)^{(m+1)} \right\|_{L^p(a,b)} \\ &= c(a,b) \left\| (v+p)^{(m+1)} \right\|_{L^p(a,b)} = c(a,b) \left\| v^{(m+1)} \right\|_{L^p(a,b)}. \end{aligned}$$

[4] James Bramble

Aus dem Verschwinden von F für $p \in P_m(a,b)$ und der Stetigkeit von F folgt nun

$$|F(v)| = |F(v) + F(p)| = |F(v+p)| \leq c \|v+p\|_{W^{m+1,p}(a,b)} \leq c(a,b) \left\| v^{(m+1)} \right\|_{L^p(a,b)}.$$

■

4.3.2 Interpolationsfehlerabschätzung

Bemerkung 4.24 Herangehensweise. Der Interpolationsfehler wird nun mit Hilfe des Bramble–Hilbert–Lemmas abgeschätzt. Die Strategie wird darin bestehen, dass man

- zuerst Abschätzungen auf einer Referenzgitterzelle zeigt,
- dann werden alle Abschätzungen über beliebige Gitterzellen K auf Abschätzungen über die Referenzgitterzelle überführt,
- die dort gezeigten Abschätzungen werden verwendet und
- schließlich wird auf K zurücktransformiert.

Dabei muss man auch untersuchen, was bei den beiden Transformationen geschieht.

□

Bemerkung 4.25 Interpolierende. Eine Interpolierende ist eine (vernünftige Approximation einer Funktion aus einem Sobolev–Raum durch eine Funktion aus dem Finite–Elemente–Raum.

Die analytische Formulierung auf einer Referenzgitterzelle $\hat{K}$ ist wie folgt. Seien $\hat{K} \subset \mathbb{R}$, zum Beispiel $\hat{K} = [-1,1]$, $\hat{P}(\hat{K})$ ein Polynomraum der Dimension N und $\hat{\Phi}_1, \ldots, \hat{\Phi}_N : C^s(\hat{K}) \to \mathbb{R}$ stetige lineare Funktionale und $\hat{\phi}_1(\hat{x}), \ldots, \hat{\phi}_N(\hat{x}) \in \hat{P}(\hat{K})$ eine lokale Basis. Das heißt, $\{\hat{\phi}_i(\hat{x})\}_{i=1}^N$ ist eine Basis von $\hat{P}(\hat{K})$ und es gilt

$$\hat{\Phi}_i(\hat{\phi}_j) = \delta_{ij}, \quad i,j = 1, \ldots, N.$$

Für $\hat{v} \in C^s(\hat{K})$ wird die Interpolierende $(I_{\hat{K}}\hat{v})(\hat{x})$ durch

$$I_{\hat{K}}\hat{v}(\hat{x}) = \sum_{i=1}^{N} \hat{\Phi}_i(\hat{v})\hat{\phi}_i(\hat{x})$$

definiert. Der Operator $I_{\hat{K}}$ ist ein stetiger und linearer Operator von $C^s(\hat{K})$ nach $\hat{P}(\hat{K})$. Aus der Linearität folgt, dass $I_{\hat{K}}$ die Identität auf $\hat{P}(\hat{K})$ ist *Übungsaufgabe*

$$\left(I_{\hat{K}}\hat{p}\right)(\hat{x}) = \hat{p}(\hat{x}) \quad \forall\, \hat{p} \in \hat{P}(\hat{K}).$$

□

Beispiel 4.26 Unterschiedliche konstante Interpolierende. Seien $\hat{K} \subset \mathbb{R}$ beliebig, $\hat{P}(\hat{K}) = P_0(\hat{K})$ und

$$\hat{\Phi}(\hat{v}) = \frac{1}{\left|\hat{K}\right|} \int_{\hat{K}} \hat{v}(\hat{x})\, d\hat{x}.$$

Das Funktional $\hat{\Phi}$ ist linear wegen der Linearität der Integration. Es ist stetig auf $C^0(\hat{K})$, da

$$\left|\hat{\Phi}(\hat{v})\right| \leq \frac{1}{\left|\hat{K}\right|} \int_{\hat{K}} |\hat{v}(\hat{x})|\, d\hat{x} \leq \frac{\left|\hat{K}\right|}{\left|\hat{K}\right|} \max_{\hat{x}\in\hat{K}} |\hat{v}(\hat{x})| = \|\hat{v}\|_{C^0(\hat{K})}.$$

Für die konstante Funktion $1 \in P_0(\hat{K})$ gilt $\hat{\Phi}(1) = 1 \neq 0$. Damit ist $\{1\}$ eine lokale Basis. Der Operator

$$I_{\hat{K}}\hat{v}(\hat{x}) = \hat{\Phi}(\hat{v})\hat{\phi}(\hat{x}) = \frac{1}{\left|\hat{K}\right|}\int_{\hat{K}} \hat{v}(\hat{x})\ d\hat{x}$$

ist der Mittelwertoperator, das heißt jede stetige Funktion auf $\hat{K}$ wird durch eine konstante Funktion interpoliert, deren Funktionswert gleich dem Integralmittelwert ist, siehe Abbildung 4.3 für ein konkretes Beispiel.

Man kann auch $\hat{\Phi}(\hat{v}) = \hat{v}(\hat{x}_0)$ für einen beliebigen Punkt $\hat{x}_0 \in \hat{K}$ setzen. Auch dieses Funktional ist linear

$$\hat{\Phi}(\alpha\hat{v} + \beta\hat{w}) = (\alpha\hat{v} + \beta\hat{w})(\hat{x}_0) = \alpha\hat{v}(\hat{x}_0) + \beta\hat{w}(\hat{x}_0)$$

für alle $\alpha, \beta \in \mathbb{R}$ und $\hat{v}, \hat{w} \in C^0(\hat{K})$ und stetig auf $C^0(\hat{K})$

$$\left|\hat{\Phi}(\hat{v})\right| = |\hat{v}(\hat{x}_0)| \leq \max_{\hat{x}\in\hat{K}} |\hat{v}(\hat{x})| = \|\hat{v}\|_{C^0(\hat{K})}\,.$$

Der damit definierte Interpolationsoperator $I_{\hat{K}}$ interpoliert jede stetige Funktion durch eine konstante Funktion, deren Funktionswert gleich dem Funktionswert in $\hat{x}_0$ ist.

Diese Beispiele zeigen, dass der Interpolationsoperator $I_{\hat{K}}$ von $\hat{P}(\hat{K})$ und von den gewählten Funktionalen $\hat{\Phi}_i$ abhängt.

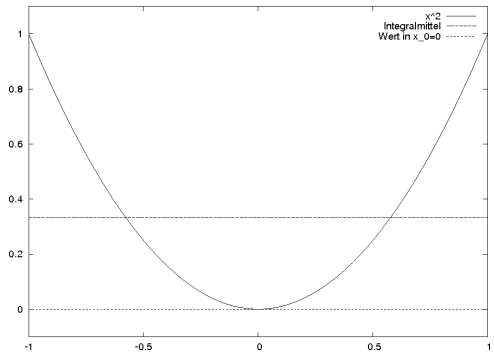

Abbildung 4.3: Interpolation von x^2 im Intervall $[-1, 1]$ in den P_0 mit Integralmittelwert und mit den Funktionswert in $x_0 = 0$.

Übungsaufgabe, Interpolationen für andere FE □

Satz 4.27 Interpolationsfehlerabschätzung auf der Referenzgitterzelle. *Seien $P_m(\hat{K}) \subset \hat{P}(\hat{K})$ und $p \in [1, \infty]$ mit $(m + 1 - s)p > 1$, wobei die Bezeichnungen von Bemerkung 4.25 verwendet werden. Dann gibt es eine von $\hat{v}(\hat{x})$ unabhängige Konstante c mit*

$$\left\|\hat{v} - I_{\hat{K}}\hat{v}\right\|_{W^{m+1,p}(\hat{K})} \leq c\left\|\hat{v}^{(m+1)}\right\|_{L^p(\hat{K})} \quad \forall\ \hat{v} \in W^{m+1,p}(\hat{K}).$$

Beweis: Mit der Bedingung $(m + 1 - s)p > 1$ kann man zeigen, dass

$$W^{m+1,p}(\hat{K}) \subset C^s(\hat{K}), \quad \|\hat{v}\|_{C^s(\hat{K})} \leq c\,\|\hat{v}\|_{W^{m+1,p}(\hat{K})}$$

gelten, siehe Literatur. Damit ist der Interpolationsoperator auf $W^{m+1,p}(\hat{K})$ wohldefiniert.

Aus der Identität des Interpolationsoperators auf $P_m(\hat{K})$, der Beschränktheit des Interpolationsoperators und der obigen Ungleichung erhält man für $\hat{q} \in P_m(\hat{K})$

$$\begin{aligned}\|\hat{v} - I_{\hat{K}}\hat{v}\|_{W^{m+1,p}(\hat{K})} &= \|\hat{v} + \hat{q} - I_{\hat{K}}(\hat{v} + \hat{q})\|_{W^{m+1,p}(\hat{K})} \\ &\leq \|\hat{v} + \hat{q}\|_{W^{m+1,p}(\hat{K})} + \|I_{\hat{K}}(\hat{v} + \hat{q})\|_{W^{m+1,p}(\hat{K})} \\ &\leq \|\hat{v} + \hat{q}\|_{W^{m+1,p}(\hat{K})} + c\,\|\hat{v} + \hat{q}\|_{C^s(\hat{K})} \\ &\leq c\,\|\hat{v} + \hat{q}\|_{W^{m+1,p}(\hat{K})}\,.\end{aligned}$$

In Lemma 4.20 wird $\hat{q}(\hat{x})$ nun so gewählt, dass

$$\int_{\hat{K}} (\hat{v} + \hat{q})^{(\gamma)}(\hat{x})\; d\hat{x} = 0 \quad 0 \leq \gamma \leq m.$$

Damit sind die Voraussetzungen von Lemma 4.21 erfüllt und es gilt

$$\|\hat{v} + \hat{q}\|_{W^{m+1,p}(\hat{K})} \leq c \left\|(\hat{v} + \hat{q})^{(m+1)}\right\|_{L^p(\hat{K})} = c \left\|\hat{v}^{(m+1)}\right\|_{L^p(\hat{K})}.$$

■

Bemerkung 4.28 Eigenschaften der Referenzabbildung. Die Eigenschaften der Referenzabbildung in einer Dimension

$$F_K(\hat{x}) = \alpha\hat{x} + \beta = x$$

lassen sich sehr leicht feststellen. Sie bildet ein Intervall $\hat{K}$ mit fester Länge auf ein Intervall K mit Länge h_K ab. Demzufolge gilt $\alpha = ch_K$ mit $c \in \mathbb{R}$, siehe Beispiel 4.13. Bei der Substitutionsregel der Integration erhält man beim Übergang von K zu $\hat{K}$ den Faktor α und bei der Umkehrsubstitution den Faktor $1/\alpha$.

In höheren Dimensionen ist die Untersuchung der Referenzabbildung weitaus komplizierter. □

Bemerkung 4.29 Transformation des Interpolationsoperators. Als nächstes soll sichergestellt werden, dass der transformierte Interpolationsoperator mit dem natürlichen Interpolationsoperator auf K übereinstimmt. Der letztere ist durch

$$I_K v(x) = \sum_{i=1}^{N} \Phi_{K,i}(v)\phi_{K,i}(x)$$

definiert, wobei $\{\phi_{K,i}(x)\}$ die Basis des Raums

$$P(K) = \{p \;:\; K \to \mathbb{R} \;:\; p = \hat{p} \circ F_K^{-1}, \hat{p} \in \hat{P}(\hat{K})\}$$

ist, die der Beziehung $\Phi_{K,i}(\phi_{K,j}) = \delta_{ij}$ genügt. Die Funktionale waren durch

$$\Phi_{K,i}(v) = \hat{\Phi}_i(v \circ F_K)$$

definiert. Daher folgt aus der Bedingung für die lokale Basis

$$\Phi_{K,i}(\hat{\phi}_j \circ F_K^{-1}) = \hat{\Phi}_{K,i}(\hat{\phi}_j \circ F_K^{-1} \circ F_K) = \hat{\Phi}_i(\hat{\phi}_j) = \delta_{ij},$$

also wegen $\Phi_{K,i}(\phi_{K,j}) = \delta_{ij}$ folgt $\phi_{K,j}(x) = \left(\hat{\phi}_j \circ F_K^{-1}\right)(x)$. Aus

$$\begin{aligned}I_{\hat{K}}\hat{v}(\hat{x}) &= \sum_{i=1}^{N} \hat{\Phi}_i(\hat{v})\hat{\phi}_i(\hat{x}) = \sum_{i=1}^{N} \Phi_{K,i}(\underbrace{\hat{v} \circ F_K^{-1}}_{=v})\,(\phi_{K,i} \circ F_K)\,(\hat{x}) \\ &= \left(\sum_{i=1}^{N} \Phi_{K,i}(v)\phi_{K,i}(x)\right) \circ F_K = (I_K v \circ F_K)\,(x)\end{aligned}$$

ergibt sich, dass $I_{\hat{K}}\hat{v}(\hat{x})$ sich richtig transformiert. □

Satz 4.30 Interpolationsabschätzung für eine beliebige Gitterzelle. *Seien eine Referenzgitterzelle $\hat{K}$, Funktionale $\{\hat{\Phi}_i\}$ und ein Polynomraum $\hat{P}(\hat{K})$ gegeben. Weiter seien alle Bedingungen aus dem Satz 4.27 erfüllt. Dann gibt es eine Konstante c unabhängig von $v \in W^{m+1,p}(K)$ mit*

$$\left\|(v - I_K v)^{(k)}\right\|_{L^p(K)} \leq c h_K^{m+1-k} \left\|v^{(m+1)}\right\|_{L^p(K)}, \quad 0 \leq k \leq m+1, \tag{4.9}$$

für alle $v \in W^{m+1,p}(K)$. Man beachte, dass die Potenz von h_K unabhängig von p ist.

Beweis: Seien $\hat{v}(\hat{x}) = v\left(F_K(\hat{x})\right)$ beziehungsweise $v(x) = \hat{v}\left(F_K^{-1}(x)\right)$. Mit der Kettenregel folgen

$$\frac{d\hat{v}(x)}{d\hat{x}} = \frac{dv(x)}{dx}\frac{dx}{d\hat{x}} = \alpha\frac{dv(x)}{dx} = c h_K \frac{dv(x)}{dx}, \quad \frac{dv(x)}{dx} = \frac{d\hat{v}(x)}{d\hat{x}}\frac{d\hat{x}}{dx} = \frac{1}{\alpha}\frac{d\hat{v}(x)}{d\hat{x}} = \frac{c}{h_K}\frac{d\hat{v}(x)}{d\hat{x}}.$$

Die Konstante c kann an unterschiedlichen Stellen verschiedene Werte annehmen. Daraus ergeben sich, mit jeder Ableitung erhält man einen weiteren Faktor ch_K beziehungsweise c/h_K,

$$\left|v^{(k)}(x)\right| \leq c h_K^{-k} \left|\hat{v}^{(k)}(\hat{x})\right|, \quad \left|\hat{v}^{(k)}(\hat{x})\right| \leq c h_K^{k} \left|v^{(k)}(x)\right|.$$

Man erhält mit Substitutionsregel für jede Funktion $v \in W^{k,p}(K)$

$$\begin{aligned}\int_K \left|v^{(k)}(x)\right|^p \, dx &\leq c h_K^{-kp} \int_{\hat{K}} \left|\hat{v}^{(k)}(\hat{x})\right|^p h_K \, d\hat{x} = c h_K^{-kp+1} \int_{\hat{K}} \left|\hat{v}^{(k)}(\hat{x})\right|^p \, d\hat{x} \\ &= c h_K^{-kp+1} \left\|\hat{v}^{(k)}\right\|^p_{L^p(\hat{K})}\end{aligned}$$

und

$$\begin{aligned}\int_{\hat{K}} \left|\hat{v}^{(k)}(\hat{x})\right|^p \, d\hat{x} &\leq c h_K^{kp} \int_K \left|v^{(k)}(x)\right|^p h_K^{-1} \, dx = c h_K^{kp-1} \int_K \left|v^{(k)}(x)\right|^p \, dx \\ &= c h_K^{kp-1} \left\|v^{(k)}\right\|^p_{L^p(K)}.\end{aligned}$$

Aus der Interpolationsfehlerabschätzung auf der Referenzzelle folgt

$$\left\|(\hat{v} - I_{\hat{K}}\hat{v})^{(k)}\right\|^p_{L^p(\hat{K})} \leq c \left\|\hat{v}^{(k)}\right\|^p_{L^p(\hat{K})}, \quad 0 \leq k \leq m+1.$$

Fasst man alle Abschätzungen zusammen, dann erhält man für den Interpolationsfehler

$$\begin{aligned}\left\|(v - I_K v)^{(k)}\right\|^p_{L^p(K)} &\leq c h_K^{-kp+1} \left\|(\hat{v} - I_{\hat{K}}\hat{v})^{(k)}\right\|^p_{L^p(\hat{K})} \\ &\leq c h_K^{-kp+1} \left\|\hat{v}^{(m+1)}\right\|^p_{L^p(\hat{K})} \\ &\leq c h_K^{(m+1-k)p} \left\|v^{(m+1)}\right\|^p_{L^p(K)}.\end{aligned}$$

Damit ist die Interpolationsfehlerabschätzung für eine beliebige Gitterzelle gezeigt. ■

Bemerkung 4.31 Uniforme Triangulierung. Sei eine uniforme Triangulierung mit $h_K = h$ für alle Gitterzellen gegeben. Dann erhält man durch Summation über die Gitterzellen die Interpolationsfehlerabschätzung für den globalen Finite–Element–Raum

$$\begin{aligned}\left\|(v - I_h v)^{(k)}\right\|_{L^p(a,b)} &= \left(\sum_{K \in \mathcal{T}_h} \left\|(v - I_K v)^{(k)}\right\|^p_{L^p(K)}\right)^{1/p} \\ &\leq \left(\sum_{K \in \mathcal{T}_h} c h_K^{(m+1-k)p} \left\|v^{(m+1)}\right\|^p_{L^p(K)}\right)^{1/p} \\ &\leq c h^{(m+1-k)} \left\|v^{(m+1)}\right\|_{L^p(a,b)}.\end{aligned}$$

Für lineare Finite–Elemente P_1 ($m = 1$) hat man beispielsweise die Abschätzungen

$$\|v - I_h v\|_{L^p(a,b)} \le ch^2 \|v''\|_{L^p(a,b)}, \quad \|(v - I_h v)'\|_{L^p(a,b)} \le ch \|v''\|_{L^p(a,b)},$$

für alle $v \in W^{2,p}(a,b)$. □

Bemerkung 4.32 Im singulär gestörten Fall ist die Interpolationsfehlerabschätzung noch nicht hinreichend um für große Gitterweiten gute Ergebnisse zu erhalten, weil der Faktor im Lemma von Cea sehr groß ist, siehe Bemerkung 4.6. Auf groben Gitter wird man deswegen große Fehler bekommen. □

Beispiel 4.33 Konvergenzordnung. Betrachte das Beispiel

$$-\varepsilon u'' + u' = 1 \quad \text{auf } (0,1), \quad u(0) = u(1) = 0.$$

Mit linearen Finite-Elementen erhält man folgende Fehler und Konvergenzordnungen im Fall $\varepsilon = 0.1$:

Int.	$\|(u-u_h)'\|_{L^2}$	Ord.	$\|u-u_h\|_{L^2}$	Ord.	$\|u-u_h\|_{L^\infty}$	Ord.
2	3.8323		0.53506		0.75669	
4	2.399	0.67576	0.096706	2.468	0.19332	1.9687
8	1.3309	0.8501	0.022568	2.0993	0.055709	1.795
16	0.68964	0.94844	0.0053014	2.0899	0.012119	2.2006
32	0.34823	0.98581	0.0012958	2.0325	0.0030185	2.0054
64	0.17455	0.99636	0.00032195	2.009	0.00074843	2.0119
128	0.087332	0.99908	8.0358e-5	2.0023	0.00018708	2.0003
256	0.043673	0.99977	2.0082e-5	2.0006	4.6746e-5	2.0007
512	0.021837	0.99994	5.0199e-6	2.0001	1.1686e-5	2
1024	0.010919	0.99999	1.2549e-6	2	2.9215e-6	2
2048	0.0054594	1	3.1373e-7	2	7.3038e-7	2
4096	0.0027297	1	7.8426e-8	2.0001	1.8259e-7	2

Für $\|u - u_h\|_{L^2}$ und $\|(u-u_h)'\|_{L^2}$ sind das genau die Ordnungen, die von der Theorie vorhergesagt werden. Im singulär gestörten Fall, $\varepsilon = 10^{-6}$, erhält man folgende Ergebnisse:

Int.	$\|(u-u_h)'\|_{L^2}$	Ord.	$\|u-u_h\|_{L^2}$	Ord.	$\|u-u_h\|_{L^\infty}$	Ord.
2	4.3301e+5		88388		1.25e+5	
4	3.3072e+5	0.38881	22097	2	31250	2
8	2.4206e+5	0.45024	5523.9	2.0001	7812.4	2
16	1.7399e+5	0.47636	1380.7	2.0003	1953.1	2
32	1.2402e+5	0.48848	344.91	2.0011	488.25	2
64	88037	0.49434	85.965	2.0044	122.06	2.0001
128	62370	0.49726	21.234	2.0174	30.52	1.9997
256	44139	0.49878	5.076	2.0646	7.6686	1.9927
512	31217	0.49972	1.132	2.1648	2.0758	1.8853
1024	22063	0.50073	0.35015	1.6928	1.0265	1.016
2048	15577	0.50219	0.17193	1.0262	0.99184	0.049515
4096	10981	0.5044	0.08561	1.0059	0.98375	0.011819

Diese schlechten Ergebnisse sind wegen des Zusammenhanges mit dem zentralen Differenzenverfahren, siehe Beispiel 4.15, zu erwarten gewesen. □

Bemerkung 4.34 Inverse Ungleichung. In diesem Abschnitt wird die Methode zum Beweis der Interpolationsfehlerabschätzung dazu verwendet, um sogenannte inverse Abschätzungen zu zeigen. Im Gegensatz zu Interpolationsfehlerabschätzungen wird dabei eine Norm einer höheren Ableitung einer Finite–Element–Funktion durch die Norm einer niederen Ableitung abgeschätzt. Man erhält als Faktor dann negative Potenzen des Durchmessers der Gitterzelle. □

Satz 4.35 Inverse Ungleichung. *Seien $0 \leq k \leq l$ natürliche Zahlen und $p, q \in [1, \infty]$. Dann gibt es eine Konstante c, die nur von $k, l, p, q, \hat{K}, \hat{P}(\hat{K})$ abhängt, mit*

$$\left\| v_h^{(l)} \right\|_{L^q(K)} \leq c h_K^{(k-l)-(p^{-1}-q^{-1})} \left\| v_h^{(k)} \right\|_{L^p(K)} \quad \forall\ v_h \in P(K).$$

Beweis: Zunächst wird die Abschätzung für $h_{\hat{K}} = c$ und $k = 0$ auf der Referenzzelle gezeigt. Da in einem endlichdimensionalen Raum alle Normen äquivalent sind, erhält man

$$\left\| \hat{v}_h^{(l)} \right\|_{L^q(\hat{K})} \leq \|\hat{v}_h\|_{W^{l,q}(\hat{K})} \leq c \|\hat{v}_h\|_{L^p(\hat{K})} \quad \forall\ \hat{v}_h \in \hat{P}(\hat{K}).$$

Im Falle $k > 0$ setzt man

$$\tilde{P}(\hat{K}) = \left\{ \hat{v}_h^{(k)} \ :\ \hat{v}_h \in \hat{P}(\hat{K}) \right\},$$

was gleichfalls ein Polynomraum ist. Wendet man die obige Abschätzung auf $\tilde{P}(\hat{K})$ an, erhält man

$$\left\| \hat{v}_h^{(l)} \right\|_{L^q(\hat{K})} = \left\| \left(\hat{v}_h^{(k)} \right)^{(l-k)} \right\|_{L^q(\hat{K})} \overset{\text{Normäquivalenz}}{\leq} c \left\| \hat{v}_h^{(k)} \right\|_{L^p(\hat{K})}.$$

Diese Abschätzung wird genauso wie in der Interpolationsfehlerabschätzung auf die Gitterzelle K transformiert. Aus den Abschätzungen für die Transformationen erhält man

$$\begin{aligned} \left\| v_h^{(l)} \right\|_{L^q(K)} &\leq c h_K^{-l+1/q} \left\| \hat{v}_h^{(l)} \right\|_{L^q(\hat{K})} \leq c h_K^{-l+1/q} \left\| \hat{v}_h^{(k)} \right\|_{L^p(\hat{K})} \\ &\leq c h_K^{k-l+1/q-1/p} \left\| v_h^{(k)} \right\|_{L^p(K)}. \end{aligned}$$

■

Übungsaufgabe: per Hand für gewisse FE nachrechnen.

Bemerkung 4.36

- Der springende Punkt im Beweis war die Äquivalenz aller Normen, eine Eigenschaft die bekanntlich bei unendlich–dimensionalen Räumen nicht gilt.
- Für $p = q$ überträgt sich die Abschätzung auf den globalen Finite–Element–Raum, sofern eine uniforme Triangulierung von (a, b) verwendet wird

$$\left\| v_h^{(l)} \right\|_{L_h^p(a,b)} \leq c h^{k-l} \left\| v_h^{(k)} \right\|_{L_h^p(a,b))}, \tag{4.10}$$

mit

$$\|\cdot\|_{L_h^p(a,b)} = \left(\sum_{K \in \mathcal{T}_h} \|\cdot\|_{L^p(K)}^p \right)^{1/p}.$$

Die zellenweise Normdefinition ist wichtig für $l \geq 2$, da dann die Finite–Element–Funktionen im allgemeinen nicht mehr die nötige Regularität für die globale Norm besitzen.

□

4.4 Stabilisierte Finite–Element–Methoden

Bemerkung 4.37 Zum Lemma von Lax–Milgram für singulär gestörte Probleme. Betrachte das Modellproblem: Finde $u \in V = H_0^1(0,1)$ so dass

$$a(u,v) = f(v) \quad \forall\, v \in V$$

mit

$$\begin{aligned} a(u,v) &:= \int_0^1 \Big(\varepsilon u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x)\Big)\, dx, \\ f(v) &:= \int_0^1 f(x)v(x)\, dx. \end{aligned}$$

Sei

$$c(x) - \frac{b'(x)}{2} \geq \omega > 0 \quad \text{für alle } x \in [0,1].$$

Mit einer analogen Rechnung wie in Bemerkung 3.35 zeigt man, dass $a(\cdot,\cdot)$ koerzitiv bezüglich der von ε abhängigen Norm

$$\|v\|_\varepsilon^2 := \varepsilon\, |v|_{1,2}^2 + \|v\|_0^2 = \varepsilon\, \|v'\|_{L^2(0,1)}^2 + \|v\|_{L^2(0,1)}^2$$

ist. Das heißt, es existiert eine von ε unabhängige Konstante μ, so dass

$$a(v,v) \geq \mu\, \|v\|_\varepsilon^2 \quad \forall\, v \in V$$

gilt. Mit partieller Integration (*Übungsaufgabe*) zeigt man, dass es eine von ε unabhängige Konstante β gibt, so dass

$$|a(v,w)| \leq \beta\, \|v\|_\varepsilon\, \|w\|_{H^1} \quad \forall\, (v,w) \in V \times V.$$

Es gibt jedoch keine von ε unabhängige Konstante γ mit

$$|a(v,w)| \leq \gamma\, \|v\|_\varepsilon\, \|w\|_\varepsilon \quad \forall\, (v,w) \in V \times V.$$

Nutzt man die Abschätzungen mit Konstanten die unabhängig von ε sind, erhält man analog zum Beweis des Lemmas von Cea

$$\|u - u_h\|_\varepsilon \leq C \inf_{v_h \in V_h} \|u - v_h\|_{H^1}$$

mit C unabhängig von ε. Ist V^h ein Standard–Finite–Elemente–Raum (stückweise polynomial), dann kann man zeigen, dass in Grenzschichten

$$\inf_{v_h \in V_h} \|u - v_h\|_{H^1} \to \infty \quad \text{für} \quad \varepsilon \to 0$$

für festes h gilt. Deswegen hat man keine gleichmäßige Konvergenz $\|u - u_h\|_\varepsilon \to 0$ für $h \to 0$. □

4.4.1 Petrov–Galerkin–Methoden und Upwind–Verfahren

Bemerkung 4.38 Petrov[5]–Galerkin–Methode. Eine Finite–Element–Methode, bei welcher Ansatz– und Testraum unterschiedlich sind, wird Petrov–Galerkin–Methode genannt. Seien S_h der Ansatzraum und T_h der Testraum, mit $\dim(S_h) = \dim(T_h)$, dann lautet eine Petrov–Galerkin–Methode: Finde $u_h \in S_h$, so dass

$$a(u_h, v_h) = f(v_h) \quad \forall\, v_h \in T_h.$$

□

[5] Petrov

Beispiel 4.39 Petrov–Galerkin–Methode und Upwind–Verfahren. Betrachte

$$-\varepsilon u''(x) + bu'(x) = 0$$

mit $b \in \mathbb{R} \setminus \{0\}$. Nutze als Ansatzfunktionen stückweise lineare Funktionen

$$\phi_i(x) = \begin{cases} (x - x_{i-1})/h & \text{für } x \in [x_{i-1}, x_i], \\ (x_{i+1} - x)/h & \text{für } x \in [x_i, x_{i+1}], \\ 0 & \text{sonst,} \end{cases} \quad , \quad i = 1, \ldots, N-1.$$

Definiere die Blasenfunktion

$$\sigma_{i-1/2}(x) = \begin{cases} 4(x - x_{i-1})(x_i - x)/h^2 & \text{für } x \in [x_{i-1}, x_i], \\ 0 & \text{sonst.} \end{cases}$$

Die Testfunktionen werden nun als stückweise quadratische Funktionen definiert

$$\psi_i(x) = \phi_i(x) + \frac{3}{2}\kappa \left(\sigma_{i-1/2}(x) - \sigma_{i+1/2}(x)\right), \quad i = 1, \ldots, N-1,$$

wobei κ ein zu wählender Upwind–Parameter ist. Direktes Nachrechnen (*Übungsaufgabe*) zeigt, dass man damit das folgende Schema erhält

$$-\varepsilon D^+ D^- u_i + b\left[\left(\frac{1}{2} - \kappa\right) D^+ u_i + \left(\frac{1}{2} + \kappa\right) D^- u_i\right] = 0.$$

Wählt man $\kappa = \operatorname{sgn}(b)/2$, so erhält man das einfache Upwind–Finite–Differenzen–Verfahren, siehe Definition 2.32.

Eine Testfunktion $\psi_i(x)$, definiert in den Knoten $\{0, 0.5, 1\}$ für $\kappa = 1/2$ ist in Abbildung 4.4 dargestellt.

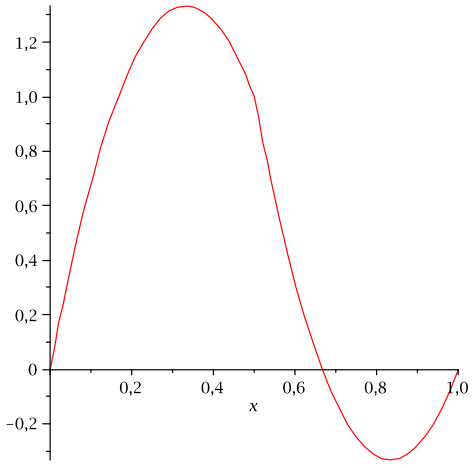

Abbildung 4.4: Stückweise quadratische Testfunktion.

□

Bemerkung 4.40 Man kann auch einige angepasste Upwind–Verfahren mit Hilfe von Petrov–Galerkin–Methoden gewinnen. Das funktioniert auch für nichtkonstante Koeffizientenfunktionen. Bessere Ergebnisse als etwa beim Iljin–Allen–Southwell–Verfahren, Satz 2.54, das heißt lineare Konvergenz, sind aber nicht zu erreichen. □

4.4.2 Die Stromlinien–Diffusions–Finite–Elemente–Methode

Bemerkung 4.41 Ziele. Das Ziel besteht darin, ein Verfahren zu konstruieren, welches stabiler als das Galerkin–Verfahren ist und welches für Finite–Elemente beliebiger Ordnung genutzt werden kann. Die Konvergenz dieses Verfahrens soll zudem von höherer Ordnung sein.

Es wird das Modellproblem

$$-\varepsilon u''(x) + b(x)u'(x) + c(x)u(x) = f(x) \text{ in } (0,1), \quad u(0) = u(1) = 0, \tag{4.11}$$

unter der Bedingung

$$c(x) - \frac{b'(x)}{2} \geq \omega > 0 \quad \text{für alle } x \in [0,1].$$

betrachtet. □

Bemerkung 4.42 Idee. Eine Idee zur Konstruktion eines stabileren Verfahrens besteht darin, zur Galerkin–Finite–Element–Methode gewichtete Residuen der starken Formulierung der Differentialgleichung (4.11) zu addieren. Dazu wird (4.11) mit $(bv')(x)$ multipliziert, über jedes Teilintervall (x_{i-1}, x_i), $i = 1, \ldots, N$, mit einem Gewicht versehen und integriert, und dann zur Galerkin–Methode addiert. □

Definition 4.43 Stromlinien–Diffusions–Finite–Elemente–Methode, SDFEM, Stromlinien–Upwind–Petrov–Galerkin FEM. Die Stromlinien–Diffusions–Finite–Elemente–Methode (SDFEM) oder Stromlinien–Upwind–Petrov–Galerkin (SUPG) FEM ist wie folgt definiert: Finde $u_h \in V_h$, so dass

$$a_h(u_h, v_h) = f_h(v_h) \quad \forall\ v_h \in V_h \tag{4.12}$$

gilt, mit

$$\begin{aligned} a_h(v,w) &:= \varepsilon(v', w') + (bv' + cv, w) \\ &\quad + \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} \delta_i \Big(-\varepsilon v''(x) + b(x)v'(x) + c(x)v(x) \Big) \Big(b(x)w'(x) \Big)\, dx, \\ f_h(w) &:= (f,w) + \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} \delta_i f(x) \Big(b(x)w'(x) \Big)\, dx. \end{aligned} \tag{4.13}$$

Dabei sind $\{\delta_i\}_{i=1}^N$ geeignet zu wählende Gewichte, welche SD–Parameter genannt werden. □

Bemerkung 4.44 Zur SDFEM.

- Der Name Stromlinien–Diffusion–FEM wird erst in höheren Dimensionen klar. Dort wird als Testfunktion für das Residuum die Ableitung in Konvektionsrichtung gewählt. Das ist die sogenannte Stromlinienrichtung.
- Für die Lösung der Galerkin–FEM wird sich das Residuum der starken Formulierung der Gleichung im allgemeinen stark von Null unterscheiden. Bei der SDFEM wird verlangt, dass dieses Residuum (in einem schwachen Sinne) nicht zu groß sein darf. Das Gewicht dieses Residuum in der SDFEM wird durch die SD–Parameter bestimmt.
- Für eine Finite–Elemente–Funktion ist die zweite Ableitung im allgemeinen nur stückweise definiert, und zwar innerhalb der Gitterzellen.
- Die SD–Parameter werden oft in jedem Intervall (x_{i-1}, x_i) konstant gewählt. Das Ziel der Analysis besteht darin, eine möglichst günstige Wahl dieser Parameter aufzuzeigen.

□

Beispiel 4.45 SDFEM für P_1. Betrachte $V_h = P_1$ auf einem äquidistanten Gitter mit $h_i = h$, $i = 1, \ldots, N$. Sind alle Koeffizientenfunktionen konstant, $c = 0$, und wählt man die SD–Parameter auch konstant, so reduziert sich die rechte Seite der SDFEM zu

$$\begin{aligned}\varepsilon(u_h', v_h') + (bu_h', v_h) + \sum_{i=1}^{N} \delta \int_{x_{i-1}}^{x_i} \Big(-\varepsilon \cdot 0 + bu_h'(x)\Big)\Big(bv_h'(x)\Big)\, dx \\ = \quad \varepsilon(u_h', v_h') + b(u_h', v_h) + \delta b^2 (u_h', v_h').\end{aligned}$$

Das entspricht der Galerkin FEM einer Gleichung mit rechter Seite

$$-\left(\varepsilon + \delta b^2\right) u''(x) + bu'(x).$$

Aus Beispiel 4.15 ist bekannt, dass die Galerkin FEM wiederum einem zentralen Differenzenverfahren entspricht. Die rechte Seite ist

$$(f, v_h) + \sum_{i=1}^{N} \delta \int_{x_{i-1}}^{x_i} fbv_h'(x)\, dx = (f, v_h) + \delta f b \underbrace{\sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} v_h'(x)\, dx}_{=0} = (f, v_h).$$

Die Summe verschwindet, da sich jede Testfunktion $v_h'(x)$ als Linearkombination der Basisfunktionen $\{\phi_i(x)\}$ von P_1 schreiben lässt und das Integral über die Ableitung jeder Basisfunkion gleich Null ist.

Insgesamt entspricht die SDFEM unter den obigen Voraussetzungen dem angepassten Finite–Differenzen–Upwind–Verfahren (2.7)

$$-\varepsilon \left(1 + \delta \frac{b^2}{\varepsilon}\right) D^+ D^- u_i + bD^0 u_i, = f_i,$$

das heißt $\sigma(q) = 1 + \delta b^2/\varepsilon$, $q = bh/(2\varepsilon)$. Wählt man den SD–Parameter als

$$\delta(q) = \frac{h}{2b}\left(\coth(q) - \frac{1}{q}\right),$$

so ist

$$\sigma(q) = 1 + \frac{hb^2}{2b\varepsilon}\left(\coth(q) - \frac{1}{q}\right) = 1 + q\left(\coth(q) - \frac{1}{q}\right) = q\coth(q).$$

Damit erhält man das Iljin–Allen–Southwell–Verfahren.

Mit $\delta = h/(2b)$ erhält man das einfache Upwind–Verfahren.

Achtung: diese einfachen Zusammenhänge gelten in höheren Dimensionen nicht mehr! □

Definition 4.46 Konsistente Finite–Element–Methode. Sei $u(x)$ eine hinreichend glatte Lösung von (4.11). Eine Finite–Element–Methode: Finde $u_h \in V_h$, so dass

$$a_h(u_h, v_h) = f_h(v_h) \quad \forall\ v_h \in V_h,$$

wird konsistent genannt, wenn gilt

$$a_h(u, v_h) = f_h(v_h) \quad \forall\ v_h \in V_h. \tag{4.14}$$

□

Bemerkung 4.47 Konsistenz einer Finite–Element–Methode ist nicht das gleiche wie Konsistenz einer Finiten–Differenzen–Methode, siehe Definition 2.6. Für Finite–Element–Methoden bedeutet Konsistenz, dass eine hinreichend glatte Lösung auch die diskrete Gleichung erfüllt. □

Lemma 4.48 Galerkin–Orthogonalität. *Eine konsistente Finite–Element–Methode besitzt die Eigenschaft der Galerkin–Orthogonalität*

$$a_h(u - u_h, v_h) = 0 \quad \forall\ v_h \in V_h. \tag{4.15}$$

Man sagt auch, dass der Fehler „senkrecht“ auf dem Finite–Element–Raum steht.

Beweis: Die Aussage folgt sofort aus der Gültigkeit von (4.12) und (4.14) durch Subtraktion dieser beiden Gleichungen. ■

Lemma 4.49 Konsistenz der SDFEM. *Die SDFEM (4.12) – (4.13) ist konsistent.*

Beweis: Für eine hinreichend glatte Lösung $u(x)$ von (4.11) ist das Residuum der starken Form der Gleichung gleich Null. Damit verschwinden die SDFEM–Terme in (4.13). Durch partielle Integration erhält man aus den übrigen Termen, dass

$$\int_0^1 \Big(-\varepsilon u''(x) + b(x)u'(x) + c(x)u(x) - f(x)\Big) v_h(x)\ dx = 0 \quad \forall\ v_h \in V_h$$

gilt. Für eine hinreichend glatte Lösung verschwindet der Ausdruck in der Klammer und diese Aussage ist wahr. Damit ist die SDFEM konsistent. ■

Bei der Analysis stabilisierter FEM ist es wichtig, dass man geeignete Normen verwendet.

Definition 4.50 Stromlinien–Diffusions–Norm, SD–Norm. Auf V_h wird die Stromlinien–Diffusions–Norm

$$|||v_h|||_{SD} := \left(\varepsilon\, |v_h|_1^2 + \omega\, \|v_h\|_0^2 + \sum_{i=1}^{N} \left\| \sqrt{\delta_i} b v_h' \right\|_{0,I_i}^2 \right)^{1/2}$$

definiert. Hierbei ist $I_i := (x_{i-1}, x_i)$ und $\|\cdot\|_{0,I_i}$ ist die Norm in $L^2(I_i)$. □

Satz 4.51 Koerzitivität der SD–Bilinearform. *Sei*

$$0 < \delta_i \le \frac{1}{2} \min \left\{ \frac{h_i^2}{\varepsilon c_{\text{inv}}^2}, \frac{\omega}{\|c\|_{L^\infty(I_i)}^2} \right\}, \tag{4.16}$$

wobei c_{inv} die Konstante der inversen Ungleichung

$$\|v_h''\|_{0,I_i} \le c_{\text{inv}} h_i^{-1} \|v_h'\|_{0,I_i} \tag{4.17}$$

ist. Dann ist die SD–Bilinearform (4.13) koerzitiv bezüglich der SD–Norm, das heißt es gilt

$$a_h(v_h, v_h) \ge \frac{1}{2} |||v_h|||_{SD}^2 \quad \forall\ v_h \in V_h.$$

Beweis: Mit partieller Integration folgt, siehe Beispiel 3.35,

$$(bv_h' + cv_h, v_h) = \left(\left(-\frac{b'}{2} + c \right) v_h, v_h \right) \quad \forall\ v_h \in V_h.$$

Mit der Definition von ω ergibt sich

$$\begin{aligned} a_h(v_h,v_h) &= \varepsilon\,|v_h|_1^2 + \int_0^1 \underbrace{\left(c(x) - \frac{b'(x)}{2}\right)}_{\geq\omega>0} v_h^2(x)\;dx + \sum_{i=1}^N \left\|\sqrt{\delta_i}bv_h'\right\|_{0,I_i}^2 \\ &\quad + \sum_{i=1}^N \int_{x_{i-1}}^{x_i} \delta_i\left(-\varepsilon v_h''(x) + c(x)v_h(x)\right)\left(b(x)v_h'(x)\right)\;dx \\ &\geq |||v_h|||_{SD}^2 + \sum_{i=1}^N \int_{x_{i-1}}^{x_i} \delta_i\left(-\varepsilon v_h''(x) + c(x)v_h(x)\right)\left(b(x)v_h'(x)\right)\;dx. \end{aligned}$$

Nun wird der zweite Term nach oben abgeschätzt, womit man insgesamt eine Abschätzung nach unten erhält, wenn man die Abschätzung des zweiten Terms vom ersten Term subtrahiert. In der Abschätzung wird die Definition des SD–Parameters verwendet. Es ist

$$\begin{aligned} &\left|\int_{x_{i-1}}^{x_i} \delta_i\left(-\varepsilon v_h''(x) + c(x)v_h(x)\right)\left(b(x)v_h'(x)\right)\;dx\right| \\ &\leq \int_{x_{i-1}}^{x_i} \left(\delta_i^{1/2}\varepsilon\left|v_h''(x)\right|\right)\left(\delta_i^{1/2}\left|b(x)v_h'(x)\right|\right)\;dx \\ &\quad + \int_{x_{i-1}}^{x_i} \left(\delta_i^{1/2}\left|c(x)\right|\left|v_h(x)\right|\right)\left(\delta_i^{1/2}\left|b(x)v_h'(x)\right|\right)\;dx \\ &\overset{\text{CSU}}{\leq} \left(\delta_i^{1/2}\varepsilon\left\|v_h''\right\|_{0,I_i} + \delta_i^{1/2}\left\|c\right\|_{L^\infty(I_i)}\left\|v_h\right\|_{0,I_i}\right)\left\|\sqrt{\delta_i}bv_h'\right\|_{0,I_i} \\ &\overset{(4.17)}{\leq} \left(\delta_i^{1/2}\frac{\varepsilon c_{\text{inv}}}{h_i}\left\|v_h'\right\|_{0,I_i} + \delta_i^{1/2}\left\|c\right\|_{L^\infty(I_i)}\left\|v_h\right\|_{0,I_i}\right)\left\|\sqrt{\delta_i}bv_h'\right\|_{0,I_i} \\ &\overset{(4.16)}{\leq} \left(\frac{h_i}{\sqrt{2\varepsilon}c_{\text{inv}}}\frac{\varepsilon c_{\text{inv}}}{h_i}\left\|v_h'\right\|_{0,I_i} + \frac{\sqrt{\omega}}{\sqrt{2}\left\|c\right\|_{L^\infty(I_i)}}\left\|c\right\|_{L^\infty(I_i)}\left\|v_h\right\|_{0,I_i}\right)\left\|\sqrt{\delta_i}bv_h'\right\|_{0,I_i} \\ &= \left(\sqrt{\frac{\varepsilon}{2}}\left\|v_h'\right\|_{0,I_i} + \sqrt{\frac{\omega}{2}}\left\|v_h\right\|_{0,I_i}\right)\left\|\sqrt{\delta_i}bv_h'\right\|_{0,I_i} \\ &\overset{\text{Young Ugl.}}{\leq} \frac{\varepsilon}{2}\left\|v_h'\right\|_{0,I_i}^2 + \frac{1}{4}\left\|\sqrt{\delta_i}bv_h'\right\|_{0,I_i}^2 + \frac{\omega}{2}\left\|v_h\right\|_{0,I_i}^2 + \frac{1}{4}\left\|\sqrt{\delta_i}bv_h'\right\|_{0,I_i}^2 \\ &= \frac{1}{2}\left|||v_h|||_{SD,I_i}\right.. \end{aligned}$$

Summation über alle Gitterzellen und Einsetzen in die erste Abschätzung ergibt die Aussage des Satzes ■

Folgerung 4.52 Koerzitivität der SD–Bilinearform für lineare finite Elemente. *Für stückweise lineare finite Elemente ist die SD–Bilinearform (4.13) koerzitiv bezüglich der SD–Norm mit der Parameterwahl*

$$0 < \delta_i \leq \frac{\omega}{\left\|c\right\|_{L^\infty(I_i)}^2}. \tag{4.18}$$

Beweis: Der Beweis ist wie für Satz 4.51, wobei man ausnutzt, dass für stückweise lineare finite Elemente $v_h''(x) = 0$ in I_i, $i = 1,\ldots,N$, ist und die entsprechenden Terme im Beweis entfallen. ■

Bemerkung 4.53 Zur Koerzitivität.

- Der Beweis von Satz 4.51 ist typisch für die Untersuchung stabilisierter Finite–Element–Methoden. Man versucht die störenden Terme irgendwie mit der verwendeten Norm abzuschätzen. Das geht im allgemeinen nur, wenn man eine geeignete Norm verwendet. Insbesondere muss die Stabilisierung in dieser Norm irgendwie auftauchen.

- Aus Satz 4.51 folgt die Stabilität der SDFEM bezüglich der SD–Norm. Alle $v_h \in V_h$ erfüllen
$$|||v_h|||_{SD} \geq \min\{1, \omega\} \|v_h\|_\varepsilon .$$
Damit folgt, dass die SDFEM auch bezüglich der Norm $\|\cdot\|_\varepsilon$ stabil ist. Bezüglich $\|\cdot\|_\varepsilon$ ist auch die Galerkin–FEM stabil, jedoch nicht bezüglich $|||\cdot|||_{SD}$. Damit ist die Stabilitätsaussage von Satz 4.51 stärker als die Stabilitätsaussage für die Galerkin–FEM.

□

Beispiel 4.54 Fortsetzung: SDFEM für P_1. Im Beispiel 4.45 wurde gezeigt, dass man unter gewissen Bedingungen mit
$$\delta(q) = \frac{h}{2b}\left(\coth(q) - \frac{1}{q}\right), \quad q = \frac{bh}{2\varepsilon},$$
das Iljin–Allen–Southwell–Verfahren, also ein gleichmäßig konvergentes Verfahren, erhält. Man braucht aber auch Parameter im Falle von nichtkonstanten Koeffizientenfunktionen, Finite–Elementen höherer Ordnung und für Probleme in höheren Dimensionen. Dabei kann man versuchen, den Spezialfall zu verallgemeinern. Mit Taylor–Entwicklung, Übungsaufgabe, sieht man, dass
$$\begin{aligned} \coth q - \frac{1}{q} &= \frac{q}{3} + \mathcal{O}\left(q^3\right) \quad \text{für } q \to 0, \\ \coth q - \frac{1}{q} &= 1 + \mathcal{O}\left(\frac{1}{q}\right) \quad \text{für } q \to \infty. \end{aligned}$$
Ist ε konstant und geht $h \to 0$, so folgt $q \to 0$ und es ist $\delta(q) \approx hq/(6b)$. Für festes h und $\varepsilon \to 0$ folgt $q \to \infty$ und es ist $\delta(q) \approx h/(2b)$. Damit sind ist die folgende Wahl des SD–Parameters motiviert
$$\delta(q) = \begin{cases} \dfrac{h^2}{12\varepsilon} & \text{für } 0 < q \ll 1, \\ \dfrac{h}{2b} & \text{für } q \gg 1. \end{cases}$$
Falls das Gitter sehr grob im Vergleich zu ε ist, also $q \gg 1$, dann geht die SDFEM in 1D in das einfache Upwind–Verfahren über. □

Satz 4.55 Konvergenz des SDFEM. *Gelte für die Lösung von (4.11) $u \in H^{k+1}(0,1)$ und betrachte die SDFEM mit P_k–Finite–Elementen. Die SD–Parameter seien wie folgt gegeben*
$$\delta_i = \begin{cases} C_0 \dfrac{h_i^2}{\varepsilon} & \text{für } h_i < \varepsilon, \\ C_0 h_i & \text{für } \varepsilon \leq h_i, \end{cases} \tag{4.19}$$
wobei die Konstante $C_0 > 0$ klein genug ist, um (4.16) für $k \geq 2$ beziehungsweise (4.18) für $k = 1$ zu erfüllen. Dann erfüllt die Lösung $u_h \in P_k$ die Fehlerabschätzung
$$|||u - u_h|||_{SD} \leq C\left(\varepsilon^{1/2} h^k + h^{k+1/2}\right) |u|_{k+1}$$
mit einer von ε unabhängigen Konstanten C und $h = \max_{i=1,\ldots,N} h_i$.

Beweis: Sei $u^I \in V_h$ die Knoteninterpolierende von $u(x)$. Mit Dreiecksungleichung erhält man
$$|||u - u_h|||_{SD} \leq \left|\left|\left|u - u^I\right|\right|\right|_{SD} + \left|\left|\left|u^I - u_h\right|\right|\right|_{SD}.$$

Der erste Term auf der rechten Seite ist der Interpolationsfehler. Mit Hilfe der Interpolationsfehlerabschätzung (4.9), die man für jeden Term der SD–Norm anwendet, erhält man

$$\left|\left|\left|u-u^I\right|\right|\right|_{SD} \leq C\left(\varepsilon^{1/2}h^k + h^{k+1/2}\right)|u|_{k+1}\,.$$

Betrachte nun den zweiten Term auf der rechten Seite. Die Koerzitivität, Satz 4.51 und die Galerkin–Orthogonalität (4.15) ergeben

$$\frac{1}{2}\left|\left|\left|u^I-u_h\right|\right|\right|_{SD}^2 \leq a_h(u^I-u_h,u^I-u_h) = a_h(u^I-u,u^I-u_h).$$

Nun wird die Dreiecksungleichung auf $a_h(u^I-u,u^I-u_h)$ angewandt und dann jeder Term einzeln abgeschätzt. Wesentlich dabei ist die Interpolationsabschätzung (4.9). Sei $w_h = u^I - u_h$. Für den Diffusionsterm gilt

$$\begin{aligned}\left|\varepsilon\left((u^I-u)',w_h'\right)\right| &\overset{\text{CSU}}{\leq} \varepsilon\left\|(u^I-u)'\right\|_0\left\|w_h'\right\|_0 = \varepsilon^{1/2}\left\|(u^I-u)'\right\|_0\varepsilon^{1/2}\left\|w_h'\right\|_0\\ &\overset{(4.9)}{\leq} C\varepsilon^{1/2}h^k|u|_{k+1}\,ep^{1/2}\left\|w_h'\right\|_0 \leq C\varepsilon^{1/2}h^k|u|_{k+1}\,|||w_h|||_{SD}\,.\end{aligned}$$

Für den reaktiven Term erhält man auf ähnliche Art und Weise

$$\begin{aligned}\left|\left(c(u^I-u),w_h\right)\right| &\overset{\text{CSU}}{\leq} \|c\|_\infty\left\|u^I-u\right\|_0\|w_h\|_0 = \omega^{-1/2}\|c\|_\infty\left\|u^I-u\right\|_0\omega^{1/2}\|w_h\|_0\\ &\overset{(4.9)}{\leq} Ch^{k+1}|u|_{k+1}\,|||w_h|||_{SD}\,.\end{aligned}$$

Als nächstes werden die Terme betrachtet, die man bei der SDFEM–Stabilisierung erhält. Wegen $\varepsilon\delta_i \leq C_0h_i^2$ folgt

$$\begin{aligned}&\left|\sum_{i=1}^N\left(-\varepsilon(u^I-u)'',\delta_ibw_h'\right)\right|\\ &\overset{\text{CSU}}{\leq} \sum_{i=1}^N\varepsilon^{1/2}\left\|(u^I-u)''\right\|_{0,I_i}\varepsilon^{1/2}\delta_i^{1/2}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}\\ &\leq C_0^{1/2}\sum_{i=1}^N h_i\varepsilon^{1/2}\left\|(u^I-u)''\right\|_{0,I_i}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}\\ &\overset{\text{CSU}}{\leq} C_0^{1/2}\varepsilon^{1/2}h\left(\sum_{i=1}^N\left\|(u^I-u)''\right\|_{0,I_i}^2\right)^{1/2}\left(\sum_{i=1}^N\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}^2\right)^{1/2}\\ &\overset{(4.9)}{\leq} C\varepsilon^{1/2}h\left(\sum_{i=1}^N h_i^{2(k-1)}|u|_{k+1,I_i}^2\right)^{1/2}\left(\sum_{i=1}^N\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}^2\right)^{1/2}\\ &\leq C\varepsilon^{1/2}h^k|u|_{k+1}\,|||w_h|||_{SD}\,.\end{aligned}$$

Für die anderen Terme erhält man unter Nutzung von $\delta_i \leq C_0 h_i$

$$\begin{aligned}
&\left|\sum_{i=1}^{N}\left(b(u^I-u)'+c(u^I-u),\delta_i b w_h'\right)\right| \\
&\overset{\text{CSU}}{\leq} \quad \sum_{i=1}^{N}\|b\|_\infty \left\|(u^I-u)'\right\|_{0,I_i}\delta_i^{1/2}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i} ep^{1/2}\left\|w_h'\right\|_0 \\
&\qquad +\sum_{i=1}^{N}\|c\|_\infty \left\|(u^I-u)\right\|_{0,I_i}\delta_i^{1/2}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i} \\
&\leq \quad C\left(\sum_{i=1}^{N}h_i^{1/2}\left\|(u^I-u)'\right\|_{0,I_i}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}\right. \\
&\qquad \left.+\sum_{i=1}^{N}h_i^{1/2}\left\|(u^I-u)\right\|_{0,I_i}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}\right) \\
&\leq \quad Ch_i^{1/2}\left[\left(\sum_{i=1}^{N}\left\|(u^I-u)'\right\|_{0,I_i}^2\right)^{1/2}+\left(\sum_{i=1}^{N}\left\|(u^I-u)\right\|_{0,I_i}^2\right)^{1/2}\right]\left(\sum_{i=1}^{N}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}^2\right)^{1/2} \\
&\overset{(4.9)}{\leq} \quad C\left(h^{k+1/2}+h^{k+3/2}\right)|u|_{k+1}\,|||w_h|||_{SD}\,.
\end{aligned}$$

Für eine optimale Abschätzung des konvektiven Terms muss man diesen erst partiell integrieren

$$\begin{aligned}
\left(b(u^I-u)',w_h\right) &= \left((u^I-u)',bw_h\right) = -\left((u^I-u),(bw_h)'\right) \\
&= -\left((u^I-u),b'w_h\right)-\left((u^I-u),bw_h'\right).
\end{aligned}$$

Nun schätzt man die letzten beiden Terme einzeln ab. Mit den gleichen Techniken wie bei den bisherigen Abschätzungen erhält man

$$\begin{aligned}
\left|\left((u^I-u),b'w_h\right)\right| &\leq \omega^{-1/2}\left\|b'\right\|_\infty\left(\sum_{i=1}^{N}\left\|u^I-u\right\|_{0,I_i}^2\right)^{1/2}\omega^{1/2}\left\|w_h\right\|_0 \\
&\leq Ch^{k+1}\,|u|_{k+1}\,|||w_h|||_{SD}\,.
\end{aligned}$$

Bei der Abschätzung des anderen Terms muss man unterscheiden, ob im Intervall I_i gilt $\varepsilon \leq h_i$ oder $\varepsilon > h_i$. Man erhält

$$\begin{aligned}
&\left|\left((u^I-u),bw_h'\right)\right| \\
&\overset{\text{CSU}}{\leq} \quad \sum_{\varepsilon\leq h_i}\delta_i^{-1/2}\left\|u^I-u\right\|_{0,I_i}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}+\sum_{\varepsilon>h_i}\|b\|_\infty\left\|u^I-u\right\|_{0,I_i}\left\|w_h'\right\|_{0,I_i} \\
&\overset{(4.9)}{\leq} \quad C\left(\sum_{\varepsilon\leq h_i}\delta_i^{-1/2}h_i^{k+1}\,|u|_{k+1,I_i}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}+\sum_{\varepsilon>h_i}h_i^{k+1}\,|u|_{k+1,I_i}\left\|w_h'\right\|_{0,I_i}\right) \\
&\overset{\delta_i\leq\ldots,\varepsilon>h_i}{\leq} \quad C\left(\sum_{\varepsilon\leq h_i}C_0^{-1/2}h_i^{-1/2}h_i^{k+1}\,|u|_{k+1,I_i}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}+\sum_{\varepsilon>h_i}h_i^{k+1/2}\,|u|_{k+1,I_i}\,\varepsilon^{1/2}\left\|w_h'\right\|_{0,I_i}\right) \\
&\overset{\text{CSU}}{\leq} \quad Ch^{k+1/2}\,|u|_{k+1}\left[\left(\sum_{i=1}^{N}\left\|\sqrt{\delta_i}bw_h'\right\|_{0,I_i}^2\right)^{1/2}+\varepsilon\,|w_h|_1\right] \\
&\leq \quad Ch^{k+1/2}\,|u|_{k+1}\,|||w_h|||_{SD}\,.
\end{aligned}$$

Fasst man nun alle Abschätzungen zusammen, so erhält man die Aussage des Satzes. ■

Bemerkung 4.56 Zur Konvergenzabschätzung. Wesentlich für die Abschätzung mit einer von ε unabhängigen Konstanten C ist, dass der Term

$$\left(\sum_{i=1}^{N} \left\| \sqrt{\delta_i} b w_h' \right\|_{0,I_i}^2 \right)^{1/2}$$

Bestandteil der Norm ist, in der man den Fehler abschätzt. Eine solche Abschätzung gilt für die von ε abhängigen Norm $\|\cdot\|_\varepsilon$ nicht. □

Beispiel 4.57 SDFEM. Das Standardbeispiel

$$-\varepsilon u'' + u' = 1 \quad \text{auf } (0,1), \quad u(0) = u(1) = 0,$$

passt nicht in die Theorie der SDFEM, da $c(x) - \frac{b'(x)}{2} = 0$ ist. Trotzdem kann man auch auf dieses Beispiel die SDFEM–Stabilisierung anwenden. Man hat allerdings nicht die in Satz 4.55 bewiesenen Konvergenzordnungen.

Ein grundlegendes Problem der Anwendung der SDFEM ist die freie Konstante C_0 in der Parameterdefinition (4.19). Am Standardbeispiel sieht man sehr gut, dass man für verschiedene Konstanten stark unterschiedliche Ergebnisse erhält, siehe Abbildung 4.5. Ist C_0 zu groß, dann ist die Grenzschicht verschmiert, für ein geeignetes C_0 findet man eine Lösung die (fast) knotenexakt ist, und ist C_0 zu klein, dann entstehen an der Grenzschicht unphysikalische Oszillationen.

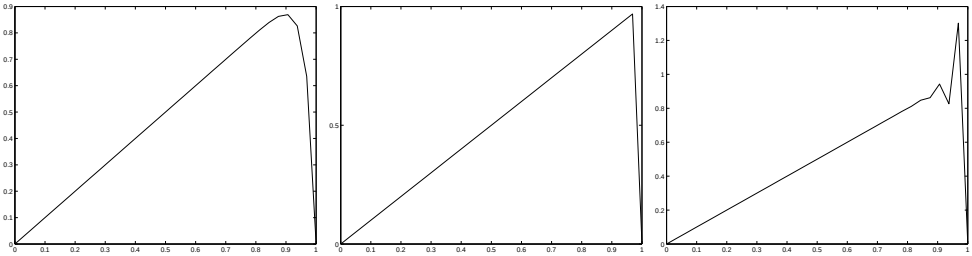

Abbildung 4.5: Mit der SDFEM berechnete Ergebnisse für das Standardbeispiel, $C_0 = 1$, $C_0 = 0.5$, $C_0 = 0.25$ von links nach rechts, $h = 1/32$, P_1 Finites–Element.

Für allgemeine Probleme ist es schwierig, C_0 geeignet zu wählen. In höheren Dimensionen wird man im allgemeinen auch kein C_0 mehr finden, so dass man eine (fast) knotenexakte Lösung erhält. Dafür besitzen mit der SDFEM berechnete Lösungen in höheren Dimensionen im allgemeinen unphysikalische Oszillationen an Grenzschichten. □

Bemerkung 4.58 Andere Wahl des SDFEM–Parameters. Man nimmt auch statt (4.19) den Parameter aus Beispiel 4.54

$$\delta_i = \frac{h_i}{2 \|b\|_{L^\infty(I_i)}} \left(\coth(\mathrm{Pe}_i) - \frac{1}{\mathrm{Pe}_i} \right), \quad \mathrm{Pe}_i = \frac{\|b\|_{L^\infty(I_i)} h_i}{2\varepsilon},$$

wobei Pe_K die lokale Péclet–Zahl ist. In dieser Definition hat man zwar keinen freien Parameter mehr, aber man stellt fest, dass bei Gleichungen in höheren Dimensionen unphysikalische Oszillationen in den berechneten Lösungen auftreten. □

Kapitel 5

Finite–Volumen–Methoden

Bemerkung 5.1 Grundlegende Idee. Finite–Volumen–Methoden (FVM) basieren auf Integralbilanzen über sogenannten Kontrollvolumen. Dazu wird das Intervall zunächst in kleine Gebiete, eben diese Kontrollvolumen, zerlegt und die Differentialgleichung wird über jedem Kontrollvolumen integriert. Im Anschluss wird partielle Integration (Gaußscher Satz in mehreren Dimensionen) angewandt, um die Integrale über den Kontrollvolumen, die Ableitungen enthalten, in Integrale auf dem Rand der Kontrollvolumen zu überführen. In einer Dimension, sind das Punktwerte in den Randpunkten. Dann verwendet man geeignete Approximationen für die Randintegrale, womit man ein Differenzenverfahren erhält.

Die Integralbilanzen können oft als Erhaltungsgesetze für physikalische Größen interpretiert werden. Deshalb werden Finite–Volumen–Methoden vor allem bei solchen Problemen mit Erfolg verwendet, bei denen die Erhaltung von Größen sehr wichtig ist, da diese Verfahren die Erhaltungseigenschaft bei der Approximation bewahren. Ein Beispiel sind inkompressible Strömungen, bei denen die Masse des Fluids in einem festen Strömungsgebiet konstant ist. □

Beispiel 5.2 Betrachte

$$-\varepsilon u''(x) + (b(x)u(x))' + c(x)u(x) = f(x) \text{ für } x \in (0,1), \quad u(0) = u(1) = 0,$$

mit $b(x) \geq \beta > 0$ und $c(x) \geq 0$. Das Intervall wird mit Hilfe eines Gitters mit den Gitterpunkten $0 = x_0, \ldots, x_N = 1$ zerlegt. Der Einfachheit halber sei das Gitter äquidistant mit Gitterweite h.

Finite–Volumen–Methoden benötigen ein Zweit–Gitter (secondary grid). Dieses wird in einer Dimension mit Hilfe der Mittelpunkte der Teilintervalle definiert. Setze

$$x_{i+1/2} := \frac{x_i + x_{i-1}}{2}, \quad i = 0, \ldots, N-1.$$

Die Kontrollvolumen werden mit Hilfe des Zweit–Gitters definiert

$$(0, x_{1/2}), (x_{1/2}, x_{3/2}), \ldots, (x_{N-1/2}, 1).$$

Integration der Gleichung über ein Kontrollvolumen ergibt

$$\begin{aligned}
&\int_{x_{i-1/2}}^{x_{i+1/2}} \Big(-\varepsilon u''(x) + (b(x)u(x))' + c(x)u(x) \Big)\, dx \\
&= \quad -\varepsilon u'(x)|_{x_{i+1/2}}^{x_{i-1/2}} + (bu)(x_{i+1/2}) - (bu)(x_{i-1/2}) + \int_{x_{i-1/2}}^{x_{i+1/2}} c(x)u(x)\, dx \quad (5.1) \\
&= \quad \int_{x_{i-1/2}}^{x_{i+1/2}} f(x)\, dx.
\end{aligned}$$

Die Terme werden nun durch Werte auf dem Originalgitter approximiert: die Ableitungen im ersten Term auf der linken Seite durch Differenzenquotienten, die Funktionswerte durch Mittelwerte und die Integrale durch Quadraturformeln. Mögliche Varianten sind

$$\begin{aligned} u'(x_{i+1/2}) &\approx \frac{u_{i+1}^N - u_i^N}{h}, & u'(x_{i-1/2}) &\approx \frac{u_i^N - u_{i-1}^N}{h}, \\ g(x_{i\pm 1/2}) &\approx \frac{g(x_i) + g(x_{i\pm 1})}{2} & \int_{x_{i-1/2}}^{x_{i+1/2}} g(x)\, dx &\approx g(x_i)h. \end{aligned}$$

Dafür braucht man vernünftige Approximationen für $u'(0)$ und $u'(1)$.

Für konstantes $b(x)$ erhält man mit diesen Approximationen

$$-\varepsilon \left(\frac{u_{i+1}^N - u_i^N}{h} - \frac{u_i^N - u_{i-1}^N}{h} \right) + b \left(\frac{u_i^N + u_{i+1}^N}{2} - \frac{u_i^N + u_{i-1}^N}{2} \right) + c_i h u_i^N = f_i h$$

mit $c_i = c(x_i), f_i = f(x_i)$. Das ist äquivalent zum zentralen Differenzenschema

$$-\varepsilon \frac{u_{i+1}^N - 2u_i^N + u_{i-1}^N}{h^2} + b \frac{u_{i+1}^N - u_{i-1}^N}{2h} + c_i u_i^N = f_i.$$

□

Bemerkung 5.3 Cell–centered Finite–Volumen–Methoden. Finite–Volumen–Methoden, welche ein Zweit–Grid nutzen um die Kontrollvolumen zu definieren, werden cell–centered Finite–Volumen–Methoden genannt. Man kann auch das Originalgitter zur Definition der Kontrollvolumen verwenden. Diese Methoden heißen dann cell–vertex Finite–Volumen–Methoden. Die letzteren Methoden sind aber nicht besonders populär, da sie instabil sind. □

Bemerkung 5.4 Finite–Volumen–Methoden für singulär gestörte Probleme. Um für singulär gestörte Probleme eine stabile Finite–Volumen–Methode zu erhalten, muss man den Konvektionsterm $(bu)(x_{i\pm 1/2})$ in (5.1) durch einen Upwind–Term approximieren, zum Beispiel durch

$$(bu)(x_{i+1/2}) \approx b(x_{i+1/2}) \left(\lambda_i u_{i+1}^N + (1-\lambda_i) u_i^N \right),$$

mit $\lambda_i \in [0, 1/2]$. Für $\lambda_i = 1/2$ erhält man das zentrale Differenzenschema und für $\lambda_i = 0$ das einfache Upwind–Verfahren aus Definition 2.32. Mit Werten zwischen 0 und 1/2 kann man die Größe des Upwindings variieren.

Seien $b(x)$ und $\lambda_i = \lambda$ konstant. Dann erhält man mit der Upwind–Approximation

$$\begin{aligned} &(bu)(x_{i+1/2}) - (bu)(x_{i-1/2}) \\ &\quad\approx b\left(\left(\lambda u_{i+1}^N + (1-\lambda) u_i^N \right) - \left(\lambda u_i^N + (1-\lambda) u_{i-1}^N \right) \right) \\ &\quad= b\left(\lambda u_{i+1}^N + (1-2\lambda) u_i^N - (1-\lambda) u_{i-1}^N \right) \\ &\quad= b\left(\frac{u_{i+1}^N - u_{i-1}^N}{2} + \left(\lambda - \frac{1}{2} \right) u_{i+1}^N + (1-2\lambda) u_i^N - \left(\frac{1}{2} - \lambda \right) u_{i-1}^N \right) \\ &\quad= b\left(\frac{u_{i+1}^N - u_{i-1}^N}{2} \right) - \frac{bh\,(1-2\lambda)}{2} \left(\frac{u_{i+1}^N - 2u_i^N + u_{i-1}^N}{h} \right). \end{aligned}$$

Nun kann das stabilisierte Finite–Volumen–Verfahren als angepasstes Upwind–Verfahren (2.7) mit

$$\sigma(q) = 1 + q\,(1-2\lambda), \quad q = \frac{bh}{2\varepsilon},$$

interpretiert werden. Damit übertragen sich auch alle Eigenschaften von angepassten Upwind–Verfahren auf diese stabilisierte Finite–Volumen–Methode.

Insbesondere ist es auch möglich, dass Iljin–Allen–Southwell–Verfahren aus Definition 2.53 mit Hilfe einer Finiten–Volumen–Methode zu generieren, siehe [RST08]. □

Bemerkung 5.5 Finite–Volumen–Methoden in höheren Dimensionen. Anders als in einer Dimension, sind Finite–Volumen–Methoden in höheren Dimensionen grundsätzlich von Finite–Differenzen–Methoden und Finite–Element–Methoden verschieden ! □

Kapitel 6

Zusammenfassung und Ausblick

Bemerkung 6.1 Verfahren. Zur Diskretisierung von partiellen Differentialgleichungen gibt es im wesentlichen drei Verfahren:

- Finite–Differenzen–Methoden:
  - approximieren die Ableitungen der starken Form der Gleichung mit Hilfe von Differenzenquotienten,
  - einfach zu verstehen und zu implementieren,
  - Taylor–Entwicklung wesentlich in der Analysis,
- Finite–Element–Methoden:
  - basieren auf der schwachen (variationellen) Formulierung der zu Grunde liegenden Gleichung in Sobolev–Räumen,
  - approximieren den unendlich–dimensionalen Sobolev–Raum durch einen endlich–dimensionalen Raum,
  - Analyis basiert auf Konzepten der Funktionalanalysis,
- Finite–Volumen–Methoden:
  - basiert auf Integration der zu Grunde liegenden Gleichung,
  - sichert die Erhaltung von Größen in Kontrollvolumen.

□

Bemerkung 6.2 Dimension.

- In einer Dimension lassen sich die Verfahren oft ineinander überführen.
- In höheren Dimensionen sind die drei Herangehensweisen grundsätzlich verschieden. Alle Verfahren besitzen Vor– und Nachteile, zum Beispiel:
  - in komplizierten Gebieten sind Finite–Elemente und Finite–Volumen flexibler als Finite–Differenzen,
  - die Implementierung von Finite–Differenzen ist wesentlich aufwändiger als die von Finite–Differenzen und Finite–Volumen,
  - Finite–Volumen–Methoden sind dort erfolgreich, wo man Erhaltungssätze erfüllen muss,
  - für Finite–Element–Methoden ist die Theorie am weitesten entwickelt.
- Ein neues Problem in höheren Dimensionen ist, dass komplizierte Gebiete auftreten können. Das hat sowohl Auswirkungen in der Analysis (Regularität der Lösung) als auch in der Praxis (Gittergenerierung).
- Die Gitterzellen in d Dimensionen sind d–dimensional. Diese Gitterzellen müssen geeignet angeordnet werden, damit ein vernünftiges Gitter entsteht. Das ist insbesondere bei komplizierten Gebieten nicht trivial. Gittergenerie-

rung, insbesondere in drei Dimensionen, ist ein wichtiges Forschungsgebiet.

□

Bemerkung 6.3 Singulär gestörte Probleme. Standard–Diskretisierungen berechnen nutzlose Lösungen für singulär gestörte Probleme, schon bei konstanten Koeffizienten. Man benötigt geeignete Stabilisierungen.

- In einer Dimension findet man Verfahren, um sehr gute Lösungen zu erhalten, zum Beispiel das Iljin–Allen–Southwell–Verfahren.
- In höheren Dimensionen ist die Entwicklung geeigneter stabiler Verfahren ein aktueller Forschungsgegenstand. Die bisher entwickelten Verfahren führen oft zu nicht zufriedenstellenden Lösungen (Grenzschichten zu stark verschmiert, unphysikalische Oszillationen).

□