Numerik partieller Differentialgleichungen

Mira Schedensack

Vorlesung im Wintersemester 2022/2023 an der Universität Leipzig
Version vom 3. Februar 2023

Inhaltsverzeichnis

1 Einleitung

Partielle Differentialgleichungen spielen in vielen Anwendungen eine wesentliche Rolle. Da diese meist nicht exakt gelöst werden können, müssen numerische Verfahren gefunden werden, die die Lösungen zuverlässig und effizient approximieren. Im ersten Teil der Vorlesung werden (kurz) Finite-Differenzen-Verfahren besprochen, die die klassische Definition der Ableitung ausnutzen, um eine Approximation zu finden. Anschließend werden wir uns den Finite-Elemente-Verfahren zuwenden, die auf der schwachen Formulierung von partiellen Differentialgleichungen basieren.
Wir werden uns in dieser Vorlesung auf lineare partielle Differentialgleichungen zweiter Ordnung beschränken. Nach [Eva98] können diese wie folgt definiert werden.

Definition 1.1 (partielle Differentialgleichung)**.** Es sei $\Omega \subseteq \mathbb{R}^d$ ein offenes Gebiet. Eine lineare partielle Differentialgleichung zweiter Ordnung ist ein Ausdruck der Form

$$\sum_{|\alpha|\leq 2} a_\alpha(x) D^\alpha u = f(x),$$

wobei $a_\alpha : \Omega \to \mathbb{R}$ und $f : \Omega \to \mathbb{R}$ gegebene Funktionen sind und $u : \Omega \to \mathbb{R}$ gesucht ist.
Ein System von linearen partiellen Differentialgleichungen zweiter Ordnung ist ein Ausdruck der Form

$$\sum_{|\alpha|\leq 2} a_{\alpha,j}(x) D^\alpha u_j = f_j(x) \qquad \text{für alle } j = 1, \dots, m,$$

wobei $a_{\alpha,j} : \Omega \to \mathbb{R}$ und $f_j : \Omega \to \mathbb{R}$, $j = 1, \dots, m$ gegebene Funktionen sind und eine vektorwertige Funktion $u : \Omega \to \mathbb{R}^m$ gesucht ist. ♢

Wir werden den Begriff "partielle Differentialgleichung" nach der englischen Übersetzung *partial differential equation* mit PDE abkürzen. Für den Rest der Vorlesung sei $\Omega \subseteq \mathbb{R}^d$ eine beschränkte, zusammenhängende, offene, nicht leere Menge. Wenn wir Finite-Elemente-Methoden betrachten, sei Ω zusätzlich ein Lipschitz-Gebiet, siehe die Definition im späteren Finite-Elemente-Kapitel.
Um eine solche PDE eindeutig lösen zu können werden zusätzlich noch Anfangs- und/oder Randbedingungen gefordert werden. Wir werden uns nun Beispiele ansehen, die wir in der Vorlesung besprechen werden.

Beispiel 1.2 (Wärmeleitungsgleichung)**.** Die Wärmeleitungsgleichung ist ein typisches Beispiel für eine parabolische PDE.
Der Laplace-Operator Δ ist für eine Funktion $u : \Omega \to \mathbb{R}$ definiert als

$$\Delta u(x) = \sum_{k=1}^{d} \partial_k^2 u(x).$$

Es sei $\Gamma_D \subseteq \partial\Omega$ abgeschlossen und $\Gamma_N \subseteq \partial\Omega$ mit $\Gamma_D \cap \Gamma_N = \emptyset$ und $\Gamma_D \cup \Gamma_N = \partial\Omega$. Außerdem seien $f : [0,T] \times \Omega \to \mathbb{R}$, $u_D : [0,T] \times \Gamma_D \to \mathbb{R}$, $g : [0,T] \times \Gamma_N \to \mathbb{R}$ und $u_0 : \Omega \to \mathbb{R}$ gegeben. Die Wärmeleitungsgleichung sucht eine Funktion $u : [0,T] \times \Omega \to \mathbb{R}$ mit

$$\begin{aligned}
\partial_t u - \Delta u &= f && \text{auf } (0,T) \times \Omega,\\
u|_{\Gamma_D} &= u_D && \text{auf } [0,T] \times \Gamma_D,\\
(\nabla u \cdot n)|_{\Gamma_N} &= g && \text{auf } [0,T] \times \Gamma_N,\\
u(0,x) &= u_0(x) && \text{für alle } x \in \Omega.
\end{aligned}$$

Dabei ist n die äußere Normale. Die Gleichung beschreibt die Wärmeverteilung im Gebiet Ω, wobei f eine Wärmequelle ist und u_0 eine Wärmeverteilung zum Zeitpunkt 0. Die

zweite Gleichung heißt Dirichlet-Randbedingung und die dritte Gleichung natürliche oder Neumann-Randbedingung. Statt $\nabla u \cdot n$ schreiben wir auch $\partial_n u$ für die Ableitung in Normalenrichtung. Die Dirichlet-Randbedingung beschreibt dabei eine vorgegebene Temperatur am Rand, während die Neumann-Randbedingung den Wärmefluss aus dem Gebiet beschreibt. Hängen die Daten f, g und u_D nicht von der Zeit ab, dann erwarten wir, dass die Zeitableitung $\partial_t u$ sehr klein wird und u sich der stationären Lösung annähert, die durch das Beispiel 1.5 beschrieben wird. ♢

Zwei weitere zeitabhängige Beispiele, für die wir Finite-Differenzen-Verfahren besprechen werden, sind die Transport- und die Wellengleichung. Lösungen zu diesen Gleichungen haben ganz andere Eigenschaften als Lösungen zur Wärmeleitungsgleichung und auch Finite-Differenzen-Verfahren zeigen für diese Gleichungen andere Eigenschaften.

Beispiel 1.3 (Transportgleichung)**.** Die Transportgleichung ist eine hyperbolische PDE erster Ordnung. Sie beschreibt die Ausbreitung einer Substanz in einer Röhre mit konstantem Profil, die mit einer fließenden Flüssigkeit gefüllt ist. Es bezeichne $a > 0$ die Geschwindigkeit der Flüssigkeit. Gesucht ist $u : [0,T] \times [0,1] \to \mathbb{R}$ mit

$$\begin{aligned} \partial_t u + a\partial_x u &= 0 && \text{auf } (0,T)\times(0,1),\\ u(t,0) &= 0 && \text{auf } (0,T],\\ u(0,x) &= u_0(x) && \text{für alle } x \in (0,1). \end{aligned} \tag{1.1}$$

Die Funktion u_0 beschreibt die Dichte der Substanz zum Zeitpunkt 0. ♢

Beispiel 1.4 (Wellengleichung)**.** Die Wellengleichung ist eine typische hyperbolische PDE zweiter Ordnung. Sie beschreibt die Ausbreitung einer Welle, zum Beispiel eine schwingende Saite.
Die Wellengleichung sucht $u : [0,T] \times \Omega \to \mathbb{R}$ mit

$$\begin{aligned} \partial_t^2 u - c^2\Delta u &= 0 && \text{auf } (0,T)\times\Omega,\\ u|_{\partial\Omega} &= u_D && \text{auf } [0,T]\times\partial\Omega,\\ u(0,x) &= u_0(x) && \text{für alle } x \in \Omega,\\ \partial_t u(0,x) &= v_0(x) && \text{für alle } x \in \Omega. \end{aligned} \tag{1.2}$$

Die Konstante c ist die Ausbreitungsgeschwindigkeit der Welle, die Funktion u_D beschreibt feste Randwerte, u_0 Anfangswerte und v_0 eine Anfangsgeschwindigkeit. ♢

Ein Beispiel, für das wir uns hauptsächlich Finite-Elemente-Methoden ansehen werden, ist das Poisson-Problem.

Beispiel 1.5 (Poisson-Model-Problem)**.** Das Poisson-Problem ist die einfachste elliptische partielle Differentialgleichung und es wird den größten Teil der Vorlesung um die Approximation dieser Lösung gehen.
Es seien Γ_D und Γ_N wie im Beispiel 1.2. Es sei $f : \Omega \to \mathbb{R}$, $u_D : \Gamma_D \to \mathbb{R}$ und $g : \Gamma_N \to \mathbb{R}$. Die Poisson-Gleichung sucht $u : \Omega \to \mathbb{R}$ mit

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega,\\ u|_{\Gamma_D} &= u_D && \text{auf } \Gamma_D,\\ (\nabla u \cdot n)|_{\Gamma_N} &= g && \text{auf } \Gamma_N. \end{aligned}$$

Die Poisson-Gleichung beschreibt beispielsweise stationäre Zustände einer Wärmeverteilung. Die Funktionen u_D und g beschreiben wie oben eine vorgegebene Temperatur am Rand bzw. den Wärmefluss aus dem Gebiet. ♢

Außerdem werden wir uns (voraussichtlich) noch weitere PDEs ansehen, die spezifische Schwierigkeiten in der Approximation aufweisen. Diese geben wir an dieser Stelle nur kurz als Ausblick an.

Beispiel 1.6 (Stokes-Gleichungen)**.** Die Stokes-Gleichungen suchen einen Fluss $u : \Omega \to \mathbb{R}^d$ und einen Druck $p : \Omega \to \mathbb{R}$ mit

$$\begin{aligned} -\Delta u + \nabla p &= f \qquad \text{in } \Omega, \\ \operatorname{div} u &= 0 \qquad \text{in } \Omega. \end{aligned} \tag{1.3}$$

Außerdem werden noch Randbedingungen gefordert. Die Schwierigkeit in der Approximation liegt in der Nebenbedingung $\operatorname{div} u = 0$. ♢

Beispiel 1.7 (Konvektions-Diffusions-Gleichung)**.** Die Konvektions-Diffusions-Gleichung sucht $u : \Omega \to \mathbb{R}$ mit

$$\begin{aligned} -\varepsilon\Delta u + b \cdot \nabla u &= f \qquad \text{in } \Omega, \\ u|_{\partial\Omega} &= 0 \qquad \text{auf } \partial\Omega, \end{aligned} \tag{1.4}$$

wobei $b : \Omega \to \mathbb{R}^d$ mit $\operatorname{div} b = 0$ ein Vektorfeld ist. Interessant ist hier der Konvektions-dominierte Fall $\varepsilon \ll 1$. In diesem Fall ist (1.4) eine singulär gestörte Gleichung, was die Approximation entsprechend schwierig macht. ♢

Die Vorlesung wird größtenteils auf den Büchern [Bra13, BS08, Bar16] aufbauen.

2 Finite Differenzen für zeitabhängige PDEs

In diesem Kapitel betrachten wir Finite-Differenzen-Verfahren für drei zeitabhängige Probleme in einer Raum-Dimension, nämlich die Wärmeleitungsgleichung, die Transportgleichung und die Wellengleichung. Finite-Differenzen-Verfahren sind eine klassische Methode zur Approximation von Lösungen von PDEs. Für zeitabhängige Probleme sind sie immer noch beliebt, für elliptische Probleme sind heute wegen der geringeren Anforderungen an die Regularität die Finite-Elemente-Verfahren gängiger.

2.1 Finite Differenzen für die Wärmeleitungsgleichung

Die Idee von Finite-Differenzen-Verfahren ist es, eine Ableitung $u'(x)$ durch einen Differenzenquotienten zu approximieren. Dafür stehen verschiedene Möglichkeiten zur Verfügung.

Definition 2.1 (Differenzenquotienten)**.** Für einen Vektor $(U_j)_{j=0,\dots,J}$ und eine Schrittweite Δx definieren wir

den Vorwärtsdifferenzenquotienten	$\partial_x^+ U_j := \dfrac{U_{j+1} - U_j}{\Delta x},$
den Rückwärtsdifferenzenquotienten	$\partial_x^- U_j := \dfrac{U_j - U_{j-1}}{\Delta x},$
den symmetrischen Differenzenquotienten erster Ordnung	$\hat\partial_x U_j := \dfrac{U_{j+1} - U_{j-1}}{2\Delta x}$

und den symmetrischen Differenzenquotienten zweiter Ordnung

$$\partial_x^+ \partial_x^- U_j := \frac{U_{j+1} - 2U_j + U_{j-1}}{\Delta x^2}. \qquad \diamondsuit$$

Proposition 2.2. *Es gilt für $(u(x_j))_{j=0,\dots,J}$ mit $x_0 = 0$ und $x_j = j\Delta x$ und $J \le 1/\Delta x$*

$$\begin{aligned}
|\partial^\pm u(x_j) - u'(x_j)| &\le \frac{\Delta x}{2}\|u''\|_{\mathcal{C}([0,1])}, && \text{wenn } u \in \mathcal{C}^2([0,1]),\\
|\hat\partial u(x_j) - u'(x_j)| &\le \frac{\Delta x^2}{6}\|u'''\|_{\mathcal{C}([0,1])}, && \text{wenn } u \in \mathcal{C}^3([0,1]),\\
|\partial^+\partial^- u(x_j) - u''(x_j)| &\le \frac{\Delta x^2}{12}\|u^{(4)}\|_{\mathcal{C}([0,1])}, && \text{wenn } u \in \mathcal{C}^4([0,1]).
\end{aligned}$$

Beweis. Die Aussagen folgen direkt aus der Taylor-Entwicklung. Der Beweis ist eine Übungsaufgabe. □

Wir wollen nun die Lösung der Wärmeleitungsgleichung aus Beispiel 1.2 approximieren. Für die bessere Übersichtlichkeit beschränken wir uns hier auf den eindimensionalen Fall mit homogener rechter Seite und reinen Dirichlet-Randbedingungen, also

$$\begin{aligned}
\partial_t u(t,x) - \partial_x^2 u(t,x) &= 0 && \text{für alle } (t,x) \in (0,T)\times(0,1),\\
u(t,0) = u(t,1) &= 0 && \text{für alle } t \in [0,T],\\
u(0,x) &= u_0(x) && \text{für alle } x \in [0,1].
\end{aligned} \tag{2.1}$$

Wir betrachten zuerst das Verfahren, das $\partial_t u$ mit dem Vorwärtsdifferenzenquotienten approximiert. Wir betrachten Gitterpunkte $(t_k, x_j)_{k=0,\dots,K;j=0,\dots,J}$ mit $t_k = k\Delta t$ und $x_j = j\Delta x$ für $k = 0,\dots,K$ und $j = 0,\dots,J$ mit $T/\Delta t = K \in \mathbb{N}$ und $1/\Delta x = J \in \mathbb{N}$.

Definition 2.3 (explizites Euler-Verfahren für die Wärmeleitungsgleichung)**.** Das explizite Euler-Verfahren berechnet $(U_j^k)_{j=0,\dots,J;\, k=0,\dots,K}$ und ist gegeben durch

$$\begin{aligned} U_j^0 &= u_0(x_j) && \text{für } j = 0,\dots,J,\\ U_0^k = U_J^k &= 0 && \text{für } k = 1,\dots,K,\\ \partial_t^+ U_j^k - \partial_x^+\partial_x^- U_j^k &= 0 && \text{für } j = 1,2,\dots,J-1,\ k = 0,1,\dots,K-1. \end{aligned} \qquad \diamondsuit$$

Das Verfahren heißt explizit, da sich die diskrete Lösung zum Zeitpunkt t_{k+1}, also der Vektor $(U_j^{k+1})_{j=0,\dots,J}$, direkt aus der diskreten Lösung zum Zeitpunkt t_k, also $(U_j^k)_{j=0,\dots,J}$ berechnen lässt, denn

$$U_j^{k+1} = U_j^k + \Delta t\, \partial_x^+\partial_x^- U_j^k = \left(1 - 2\frac{\Delta t}{(\Delta x)^2}\right) U_j^k + \frac{\Delta t}{(\Delta x)^2} U_{j-1}^k + \frac{\Delta t}{(\Delta x)^2} U_{j+1}^k.$$

Proposition 2.4 (Stabilität und Konvergenz des expliziten Euler-Verfahrens)**.** *Wenn* $\lambda := \Delta t/(\Delta x)^2 \le 1/2$ *ist, dann gilt für alle* $k \in \{0,\dots,K\}$

$$\sup_{j=0,\dots,J} |U_j^k| \le \sup_{j=0,\dots,J} |U_j^0|.$$

Gilt außerdem $u \in \mathcal{C}^4([0,T]\times[0,1])$*, dann folgt*

$$\sup_{j=0,\dots,J} |u(t_k,x_j) - U_j^k| \le \frac{t_k}{2}(\Delta t + (\Delta x)^2)\left(\|\partial_x^4 u\|_{\mathcal{C}([0,T]\times[0,1])} + \|\partial_t^2 u\|_{\mathcal{C}([0,T]\times[0,1])}\right).$$

Beweis. Wir zeigen zunächst eine Stabilitätsabschätzung für das etwas allgemeinere Verfahren

$$\begin{aligned} U_j^0 &= u_0(x_j) && \text{für } j = 0,\dots,J,\\ U_0^k = U_J^k &= 0 && \text{für } k = 0,\dots,K,\\ \partial_t^+ U_j^k - \partial_x^+\partial_x^- U_j^k &= F_j^k && \text{für } j = 1,2,\dots,J-1,\ k = 0,1,\dots,K-1, \end{aligned}$$

also für das inhomogene Problem mit rechter Seite $(F_j^k)_{j=0,\dots,J;\, k=0,\dots,K}$.
Da $\lambda \le 1/2$, gilt

$$\begin{aligned} |U_j^{k+1}| &\le \underbrace{(1-2\lambda)}_{\ge 0} \underbrace{|U_j^k|}_{\le \sup_{\ell=0,\dots,J}|U_\ell^k|} + \lambda \underbrace{|U_{j+1}^k|}_{\le \sup_{\ell=0,\dots,J}|U_\ell^k|} + \lambda \underbrace{|U_{j-1}^k|}_{\le \sup_{\ell=0,\dots,J}|U_\ell^k|} + \Delta t |F_j^k|\\ &\le \sup_{\ell=0,\dots,J} |U_\ell^k| + \Delta t \sup_{\ell=0,\dots,J} |F_\ell^k|. \end{aligned}$$

Da die rechte Seite unabhängig von j ist, dürfen wir auf der linken Seite zum Supremum übergehen und es folgt rekursiv

$$\sup_{j=0,\dots,J} |U_j^{k+1}| \le \sup_{j=0,\dots,J} |U_j^0| + \sum_{m=0}^{k} \Delta t \sup_{j=0,\dots,J} |F_j^m|.$$

Für den letzten Term gilt

$$\sum_{m=0}^{k} \Delta t \sup_{j=0,\dots,J} |F_j^m| \le \sup_{\ell=0,\dots,k} \sup_{j=0,\dots,J} |F_j^\ell| \sum_{m=0}^{k} \Delta t = t_k \sup_{\ell=0,\dots,k} \sup_{j=0,\dots,J} |F_j^\ell|,$$

also insgesamt

$$\sup_{j=0,\dots,J} |U_j^{k+1}| \leq \sup_{j=0,\dots,J} |U_j^0| + t_k \sup_{m=0,\dots,k} \sup_{j=0,\dots,J} |F_j^m|. \tag{2.2}$$

Daraus folgt mit $F_j^m = 0$ die behauptete Stabilität.
Der Konsistenzfehler $Z_j^k := \partial_t^+ u(t_k, x_j) - \partial_x^+ \partial_x^- u(t_k, x_j)$ erfüllt nach Proposition 2.2

$$\begin{aligned} |Z_j^k| &\leq |\partial_t^+ u(t_k, x_j) - \partial_t u(t_k, x_j)| + |\partial_x^2 u(t_k, x_j) - \partial_x^+ \partial_x^- u(t_k, x_j)| \\ &\leq \frac{1}{2} \left(\Delta t + (\Delta x)^2\right) \left(\|\partial_x^4 u\|_{\mathcal{C}([0,T]\times[0,1])} + \|\partial_t^2 u\|_{\mathcal{C}([0,T]\times[0,1])}\right). \end{aligned}$$

Der Fehler $E_j^k := u(t_k, x_j) - U_j^k$ erfüllt aber

$$\begin{aligned} \partial_t^+ E_j^k - \partial_x^+ \partial_x^- E_j^k &= Z_j^k \qquad \text{für } j = 1, 2, \dots, J-1,\ k = 0, 1, \dots, K-1, \\ E_j^0 &= 0 \qquad \text{für } j = 0, \dots, J, \\ E_0^k = E_J^k &= 0 \qquad \text{für } k = 0, \dots, K. \end{aligned}$$

Mit (2.2) folgt die Fehlerabschätzung. □

Bemerkung. Die Bedingung $\lambda \leq \frac{1}{2}$ bedeutet, dass die Zeitschrittweite $\Delta t \leq (\Delta x)^2/2$ gewählt werden muss. Dies erfordert einen hohen numerischen Aufwand, weswegen das explizite Verfahren für Wärmeleitungsprobleme in der Praxis kaum Anwendung findet. ♢

Bemerkung. In numerischen Experimenten kann man sehen, dass die Bedingung $\lambda \leq \frac{1}{2}$ scharf ist und die diskrete Lösung "explodiert", wenn die Bedingung nicht erfüllt ist.
Für nicht glatte Anfangsdaten zeigen sich in numerischen Experimenten leichte Oszillationen. ♢

Definition 2.5 (implizites Euler-Verfahren für die Wärmeleitungsgleichung)**.** Das implizite Euler-Verfahren berechnet $(U_j^k)_{j=0,\dots,J;\ k=0,\dots,K}$ und ist gegeben durch

$$\begin{aligned} U_j^0 &= u_0(x_j) \qquad \text{für } j = 0, \dots, J, \\ U_0^k = U_J^k &= 0 \qquad \text{für } k = 1, \dots, K, \\ \partial_t^- U_j^k - \partial_x^+ \partial_x^- U_j^k &= 0 \qquad \text{für } j = 1, 2, \dots, J-1,\ k = 1, \dots, K. \end{aligned}$$ ♢

Dies Verfahren heißt implizit, da U_j^k aus U_j^{k-1} durch

$$U_j^k - \lambda \left(U_{j-1}^k - 2U_j^k + U_{j+1}^k\right) = U_j^{k-1} \tag{2.3}$$

mit $\lambda := \Delta t/(\Delta x)^2$ gegeben ist. Um U_j^k zu berechnen, muss also das lineare Gleichungssystem

$$\underbrace{\begin{pmatrix} 1+2\lambda & -\lambda & & & & \\ -\lambda & 1+2\lambda & -\lambda & & & \\ & -\lambda & 1+2\lambda & -\lambda & & \\ & & \ddots & \ddots & \ddots & \\ & & & -\lambda & 1+2\lambda & -\lambda \\ & & & & -\lambda & 1+2\lambda \end{pmatrix}}_{=:A} \begin{pmatrix} U_1^k \\ U_2^k \\ U_3^k \\ \vdots \\ U_{J-2}^k \\ U_{J-1}^k \end{pmatrix} = \begin{pmatrix} U_1^{k-1} \\ U_2^{k-1} \\ U_3^{k-1} \\ \vdots \\ U_{J-2}^{k-1} \\ U_{J-1}^{k-1} \end{pmatrix}$$

gelöst werden.
Um die Existenz von Lösungen zu zeigen, werden wir den Gerschgorinschen Kreissatz benutzen.

Satz 2.6 (Gerschgorinscher Kreissatz). *Es sei $A \in \mathbb{R}^{n\times n}$ mit Einträgen $(a_{jk})_{j=1,\dots,n;k=1,\dots,n}$ gegeben und $\lambda \in \mathbb{C}$ sei ein Eigenwert von A. Dann gilt*

$$\lambda \in \bigcup_{\ell=1}^{n} K_\ell \quad mit \quad K_\ell := \left\{ z \in \mathbb{C} \;\middle|\; |z - a_{\ell\ell}| \le \sum_{j\in\{1,\dots,n\}\setminus\{\ell\}} |a_{\ell j}| \right\}.$$

Die K_ℓ werden Gerschgorin-Kreise genannt.

Beweis. Es sei $x \in \mathbb{C}^n \setminus \{0\}$ ein Eigenvektor zu λ. Es sei $\ell \in \{1,\dots,n\}$ so, dass der Eintrag x_ℓ den größten Betrag habe, also $|x_j| \le |x_\ell|$ für alle $j = 1,\dots,n$. Insbesondere gilt $x_\ell \neq 0$. Es gilt

$$\lambda x_\ell = (Ax)_\ell = \sum_{j=1}^{n} a_{\ell j} x_j.$$

Da $x_j \neq 0$, ergibt eine Umformung

$$\lambda - a_{\ell\ell} = \sum_{j\in\{1,\dots,n\}\setminus\{\ell\}} a_{\ell j} \frac{x_j}{x_\ell}.$$

Eine Dreiecksungleichung und $|x_j/x_\ell| \le 1$ zeigt die Behauptung. □

Proposition 2.7 (Existenz, Stabilität, Konvergenz des impliziten Euler-Verfahrens). *Es existiert eine eindeutige Lösung zum impliziten Euler-Verfahren und es gilt*

$$\sup_{j=0,\dots,J} |U_j^k| \le \sup_{j=0,\dots,J} |U_j^0| \qquad \text{für alle } k = 0,\dots,K.$$

Wenn außerdem $u \in \mathcal{C}^4([0,T]\times[0,1])$ ist, dann gilt

$$\sup_{j=0,\dots,J} |u(t_k,x_j) - U_j^k| \le \frac{t_k}{2}(\Delta t + (\Delta x)^2)\left(\|\partial_x^4 u\|_{\mathcal{C}([0,T]\times[0,1])} + \|\partial_t^2 u\|_{\mathcal{C}([0,T]\times[0,1])}\right)$$

für $k = 0,\dots,K$.

Beweis. Aus dem Gerschgorinschen Kreissatz folgt, dass A invertierbar ist. Also existiert eine eindeutige Lösung des impliziten Euler-Verfahrens.
Es sei nun $j' \in \{1,2,\dots,J-1\}$ so, dass $|U_{j'}^k| = \sup_{j=0,\dots,J}|U_j^k|$ erfüllt ist. Dann folgt mit (2.3), dass

$$\begin{aligned}(1+2\lambda)\sup_{j=0,\dots,J}|U_j^k| = (1+2\lambda)|U_{j'}^k| &\le |U_{j'}^{k-1}| + \lambda|U_{j'-1}^k| + \lambda|U_{j'+1}^k| \\ &\le \sup_{j=0,\dots,J}|U_j^{k-1}| + 2\lambda \sup_{j=0,\dots,J}|U_j^k|\end{aligned}$$

Rekursiv folgt die Stabilität. Die Fehlerabschätzung ist eine Übungsaufgabe. □

Bemerkung. Die Stabilität und die Fehlerabschätzung für das implizite Euler-Verfahren gelten ohne eine Bedingung an die Zeitschrittweite. Daher wird das implizite Verfahren in der Praxis häufig genutzt. ♢

Bemerkung. In numerischen Experimenten sieht man, dass auch für nicht-glatte Anfangsdaten die diskrete Lösung glatt ist. Durch das Lösen des linearen Gleichungssystems werden die Anfangsdaten sofort geglättet. ♢

Wir werden uns als letztes Verfahren für die Wärmeleitungsgleichung eine Kombination aus dem expliziten und dem impliziten Verfahren ansehen. Dieses Verfahren konvergiert mit höherer Ordnung als das implizite und das explizite Verfahren.

Definition 2.8 (Crank-Nicolson-Verfahren)**.** Das Crank-Nicolson-Verfahren berechnet die diskrete Lösung $(U_j^k)_{j=0,\dots,J;\, k=0,\dots,K}$ durch

$$\begin{aligned} U_j^0 &= u_0(x_j) && \text{für } j=0,\dots,J,\\ U_0^k = U_J^k &= 0 && \text{für } k=1,\dots,K,\\ \partial_t^- U_j^{k+1} - \frac{1}{2}\left(\partial_x^+\partial_x^- U_j^k + \partial_x^+\partial_x^- U_j^{k+1}\right) &= 0 && \text{für } j=1,2,\dots,J-1,\ k=1,\dots,K. \diamondsuit \end{aligned}$$

Dieses Verfahren ist auch ein implizites Verfahren, da in jedem Schritt das Gleichungssystem

$$\begin{pmatrix} 1+\lambda & -\frac{1}{2}\lambda & & & & \\ -\frac{1}{2}\lambda & 1+\lambda & -\frac{1}{2}\lambda & & & \\ & -\frac{1}{2}\lambda & 1+\lambda & -\frac{1}{2}\lambda & & \\ & & \ddots & \ddots & \ddots & \\ & & & -\frac{1}{2}\lambda & 1+\lambda & -\frac{1}{2}\lambda \\ & & & & -\frac{1}{2}\lambda & 1+\lambda \end{pmatrix} \begin{pmatrix} U_1^{k+1} \\ U_2^{k+1} \\ U_3^{k+1} \\ \vdots \\ U_{J-2}^{k+1} \\ U_{J-1}^{k+1} \end{pmatrix}$$

$$= \begin{pmatrix} 1-\lambda & \frac{1}{2}\lambda & & & & \\ \frac{1}{2}\lambda & 1-\lambda & \frac{1}{2}\lambda & & & \\ & \frac{1}{2}\lambda & 1-\lambda & \frac{1}{2}\lambda & & \\ & & \ddots & \ddots & \ddots & \\ & & & \frac{1}{2}\lambda & 1-\lambda & \frac{1}{2}\lambda \\ & & & & \frac{1}{2}\lambda & 1-\lambda \end{pmatrix} \begin{pmatrix} U_1^k \\ U_2^k \\ U_3^k \\ \vdots \\ U_{J-2}^k \\ U_{J-1}^k \end{pmatrix}$$

gelöst werden muss.

Proposition 2.9 (Existenz, Stabilität, Konvergenz des Crank-Nicolson-Verfahrens)**.** *Es existiert eine eindeutige Lösung zum Crank-Nicolson-Verfahren und sie ist stabil im Sinne, dass*

$$\|U^k\|_{2,\Delta x} \le \|U^0\|_{2,\Delta x} \qquad \textit{für alle } k=0,\dots,K,$$

wobei

$$\|V\|_{2,\Delta x} = \sqrt{\Delta x \sum_{j=0}^{J} V_j^2} \qquad \textit{für alle } V\in\mathbb{R}^{J+1}.$$

Wenn $u\in\mathcal{C}^4([0,T]\times[0,1])$, dann existiert eine Konstante $C>0$ mit

$$\|U^k - u(t_k,\bullet)\|_{2,\Delta x} \le C t_k((\Delta t)^2 + (\Delta x)^2)\|u\|_{\mathcal{C}^4([0,T]\times[0,1])}.$$

Beweis. Die eindeutige Lösbarkeit folgt aus dem Gerschgorinschen Kreissatz.
Die (allgemeinere) Stabilität

$$\|U^k\|_{2,\Delta x} \le \|U^0\|_{2,\Delta x} + \sum_{m=0}^{k} \Delta t \|F^m\|_{2,\Delta x} \tag{2.4}$$

für Lösungen $(U^k)_{k=0,\dots,K}$ vom Crank-Nicolson-Verfahren mit rechter Seite (in der dritten Zeile) F_j^m statt 0 werden wir hier nicht beweisen. Bei Interesse kann dies z.B. in [Bar16] nachgelesen werden.
Wir werden nun die Fehlerabschätzung beweisen. Wir betrachten wieder den Konsistenzfehler

$$Z_j^k := \partial_t^- u(t_{k+1}, x_j) - \frac{1}{2}\partial_x^+\partial_x^-(u(t_{k+1}, x_j) + u(t_k, x_j))$$

und schätzen ab

$$|Z_j^k| \le \underbrace{\left|\partial_t^- u(t_{k+1}, x_j) - \partial_t u(t_{k+\frac{1}{2}}, x_j)\right|}_{=:A} + \underbrace{\left|\frac{1}{2}\partial_x^+\partial_x^-(u(t_{k+1}, x_j) + u(t_k, x_j)) - \partial_x^2 u(t_{k+\frac{1}{2}}, x_j)\right|}_{=:B},$$

wobei $t_{k+\frac{1}{2}} = (k+\frac{1}{2})\Delta t$ ist. Fassen wir $\hat\partial_t u(t_{k+\frac{1}{2}}, x_j)$ bezüglich des Gitters $t_k, t_{k+\frac{1}{2}}, t_{k+1}$ auf, gilt

$$\partial_t^- u(t_{k+1}, x_j) = \hat\partial_t u(t_{k+\frac{1}{2}}, x_j).$$

Mit Proposition 2.2 folgt

$$A \le C(\Delta t)^2 \|u\|_{\mathcal{C}^3([0,T]\times[0,1])}.$$

Taylor-Entwicklungen in der Zeit um $t_{k+\frac{1}{2}}$ führen auf

$$\begin{aligned}
\partial_x^2 u(t_k, x_j) &= \partial_x^2 u(t_{k+\frac{1}{2}}, x_j) - \frac{\Delta t}{2}\partial_x^2\partial_t u(t_{k+\frac{1}{2}}, x_j) + \mathcal{O}((\Delta t)^2)\|u\|_{\mathcal{C}^4([0,T]\times[0,1]},\\
\partial_x^2 u(t_{k+1}, x_j) &= \partial_x^2 u(t_{k+\frac{1}{2}}, x_j) + \frac{\Delta t}{2}\partial_x^2\partial_t u(t_{k+\frac{1}{2}}, x_j) + \mathcal{O}((\Delta t)^2)\|u\|_{\mathcal{C}^4([0,T]\times[0,1]}.
\end{aligned}$$

Andererseits gilt

$$\partial_x^2 u(t_{k+1}, x_j) = \partial_x^+\partial_x^- u(t_{k+1}, x_j) + \mathcal{O}((\Delta x)^2)\|u\|_{\mathcal{C}^4([0,T]\times[0,1]}.$$

Daraus folgt

$$\partial_x^2 u(t_{k+\frac{1}{2}}, x_j) - \frac{1}{2}\partial_x^+\partial_x^-(u(t_k, x_j) + u(t_{k+1}, x_j)) = \big(\mathcal{O}((\Delta t)^2) + \mathcal{O}((\Delta x)^2)\big)\|u\|_{\mathcal{C}^4([0,T]\times[0,1]}$$

Insgesamt gilt also

$$|Z_j^k| \le C\big((\Delta t)^2 + (\Delta x)^2\big)\|u\|_{\mathcal{C}^4([0,T]\times[0,1]}.$$

Mit der Stabilität (2.4) folgt die Behauptung. □

Bemerkung. Das Crank-Nicolson-Verfahren ist stabil bezüglich der $\|\bullet\|_{2,\Delta x}$-Norm ohne Bedingung an die Zeitschrittweite. Im Allgemeinen gilt allerdings *nicht*

$$\sup_{j=0,\dots,J}|U_j^{k+1}| \le \sup_{j=0,\dots,J}|U_j^k|.$$ ♢

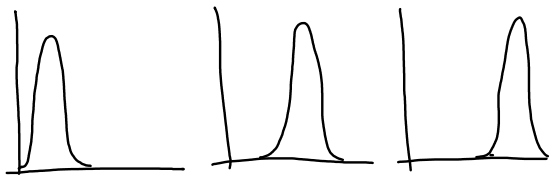

Abbildung 2.1: Lösung einer Transportgleichung.

2.2 Finite Differenzen für die Transportgleichung

In diesem Kapitel werden wir die Lösung der Transportgleichung aus Beispiel 1.3 approximieren, also Lösungen von

$$\begin{aligned} \partial_t u + a\partial_x u &= 0 && \text{auf } [0,T]\times(0,1),\\ u(t,0) &= 0 && \text{auf } (0,T],\\ u(0,x) &= u_0(x) && \text{für alle } x\in(0,1). \end{aligned}$$

Bevor wir das Finite-Differenzen-Verfahren definieren werden, sehen wir uns erst die Lösungen der Transportgleichung an, die in diesem einfachen Beispiel explizit berechnet werden können. Fasst man nämlich die Lösung u als Funktion auf $[0,T]\times(0,1)$ auf, sieht man

$$\nabla u\cdot\begin{pmatrix}1\\ a\end{pmatrix} = \begin{pmatrix}\partial_t u\\ \partial_x u\end{pmatrix}\cdot\begin{pmatrix}1\\ a\end{pmatrix} = \partial_t u + a\partial_x u = 0,$$

d.h. u ist konstant entlang der Geraden $\{(t,x)\mid x-at=c\}$ für eine beliebige feste Konstante c. Diese Geraden, entlang derer die Lösung konstant ist, werden *Charakteristiken* genannt. Die Anfangsbedingungen zeigen dann, dass

$$u(t,x) = \widetilde{u}_0(x-at)$$

gilt, wobei

$$\widetilde{u}_0(x) := \begin{cases} u_0(x) & \text{wenn } x\in[0,1],\\ 0 & \text{sonst.}\end{cases}$$

Abbildung 2.1 zeigt, wie solche Lösungen aussehen.
Wir werden nun ein explizites Finite-Differenzenverfahren für diese Gleichung definieren.

Definition 2.10. Das explizite Finite-Differenzen-Verfahren für die Transportgleichung ist gegeben durch

$$\begin{aligned} U_j^{k+1} &= U_j^k - a\frac{\Delta t}{\Delta x}\left(U_j^k - U_{j-1}^k\right) && \text{für alle } j=1,\dots,J;\ k=1,\dots,K-1,\\ U_0^k &= 0 && \text{für alle } k=1,\dots,K,\\ U_j^0 &= u_0(x_j) && \text{für alle } j=0,\dots,J. \end{aligned}$$ ♢

Wir werden nun untersuchen, wann das Verfahren stabil ist. In diesem Beispiel kann man gut sehen, welche Stabilitätseigenschaften die exakte Lösung hat. Da $u(t,x)=\widetilde{u}_0(x-at)$ ist, gilt

$$\sup_{t\in[0,T]}\sup_{x\in(0,1)}|u(t,x)| = \sup_{t\in[0,T]}\sup_{x\in(0,1)}|\widetilde{u}(x-at)| \le \sup_{x\in(0,1)}|u(x)|.$$

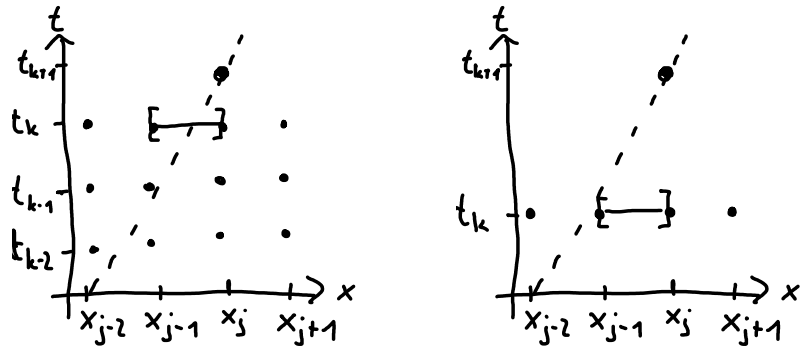

Abbildung 2.2: Veranschaulichung der CFL-Bedingung: Links ist die CFL-Bedingung erfüllt, auf dem rechten Bild ist sie verletzt.

Proposition 2.11 (Stabilität des Finiten-Differenzenverfahrens)**.** *Setze $\mu = a\Delta t/\Delta x$. Wenn $0 \leq \mu \leq 1$, dann ist das Finite-Differenzenverfahren für die Transportgleichung stabil und es gilt*

$$\sup_{k=0,\dots,K} \sup_{j=0,\dots,J} |U_j^k| \leq \sup_{j=0,\dots,J} |U_j^0|.$$

Beweis. Nach Definition gilt $U_j^{k+1} = U_j^k - \mu \left(U_j^k - U_{j-1}^k\right)$. Daraus folgt

$$|U_j^{k+1}| = |U_j^k - \mu \left(U_j^k - U_{j-1}^k\right)| \leq |(1-\mu)U_j^k| + |\mu U_{j-1}^k| = |1-\mu||U_j^k| + |\mu||U_{j-1}^k|.$$

Da nach Voraussetzung $(1-\mu) > 0$ und $\mu > 0$ folgt

$$|U_j^{k+1}| \leq (1-\mu)|U_j^k| + \mu|U_{j-1}^k| \leq (1-\mu) \sup_{j'=0,\dots,J} |U_{j'}^k| + \mu \sup_{j'=0,\dots,J} |U_{j'}^k| = \sup_{j'=0,\dots,J} |U_{j'}^k|.$$

Daraus folgt die Behauptung. □

Wir haben gesehen, dass die exakte Lösung auf den Charakteristiken konstant ist und die Lösung in einem Punkt (t, x) nur von den Werten auf dieser Charakteristik abhängt. Die *CFL-Bedingung* (nach Courant, Friedrichs und Levy) besagt, dass die Charakteristik durch den Punkt (t_{k+1}, x_j) durch die konvexe Hülle der Gitterpunkte $(t_k, x_{j-m_\ell}), \dots, (t_k, x_{j+m_r})$ gehen muss, die in die Berechnung von U_j^{k+1} eingehen, siehe auch Abbildung 2.2. In diesem Beispiel muss die Charakteristik also durch $\{t_k\} \times [x_{j-1}, x_j]$ gehen. Die CFL-Bedingung muss erfüllt sein, damit die Approximation sinnvoll sein kann. Für das explizite Verfahren aus Definition 2.10 ist die CFL-Bedingung genau dann erfüllt, wenn $0 \leq \mu \leq 1$, siehe Übungsaufgaben. Dass die CFL-Bedingung aber nicht hinreichend für die Stabilität ist, wird in einer weiteren Übungsaufgabe gezeigt, da das Finite-Differenzen-Verfahren $\partial_t^+ U_j^k + a\widehat{\partial}_x U_j^k = 0$ die CFL-Bedingung für $|\mu| \leq 1$ erfüllt, aber nicht stabil ist.
Wir werden nun wie im vorigen Kapitel aus der Stabilität und der Konsistenz des Verfahrens die Konvergenz folgern.

Satz 2.12. *Es sei $\mu = a\Delta t/\Delta x$. Wenn $0 \leq \mu \leq 1$ und $u \in C^2([0,T] \times [0,1])$, dann gilt*

$$\sup_{j=0,\dots,J} |u(t_k, x_j) - U_j^k| \leq \frac{t_k}{2}(\Delta t + a\Delta x)\|u\|_{\mathcal{C}^2([0,t]\times[0,1])}$$

für alle $k = 0, \dots, K$.

Beweis. Wir definieren wieder den Konsistenzfehler

$$Z_j^k := \partial_t^+ u(t_k, x_j) + a\partial_x^- u(t_k, x_j).$$

Mit Proposition 2.2 folgt

$$\begin{aligned}|Z_j^k| &= |\partial_t^+ u(t_k, x_j) - \partial_t u(t_k, x_j) + a\partial_x^- u(t_k, x_j) - a\partial_x u(t_k, x_j)| \\ &\le |\partial_t^+ u(t_k, x_j) - \partial_t u(t_k, x_j)| + |a\partial_x^- u(t_k, x_j) - a\partial_x u(t_k, x_j)| \\ &\le \frac{\Delta t}{2} \sup_{t\in[0,T]} |\partial_t^2 u(\bullet, x_j)| + a\frac{\Delta x}{2} \sup_{x\in[0,1]} |\partial_x^2 u(t_k, \bullet)|.\end{aligned}$$

Der Fehler $E_j^k := u(t_k, x_j) - U_j^k$ erfüllt

$$E_j^{k+1} = E_j^k - \mu\left(E_j^k - E_{j-1}^k\right) + \Delta t Z_j^k.$$

Wie im Beweis der Stabilität in Proposition 2.11 gilt

$$|E_j^{k+1}| \le \sup_{j'=0,\dots,J} |E_{j'}^k| + \Delta t |Z_j^k|.$$

Mit der Randbedingung $E_0^{k+1} = 0$ folgt also

$$\sup_{j=0,\dots,J} |E_j^{k+1}| \le \sup_{j=0,\dots,J} |E_j^k| + \Delta t \sup_{j=0,\dots,J} |Z_j^k|.$$

Mit der Abschätzung für den Konsistenzfehler folgt die Behauptung. □

2.3 Finite Differenzen für die Wellengleichung

In diesem Kapitel wollen wir die Lösung der Wellengleichung aus Beispiel 1.4 approximieren. Wir beschränken uns zur besseren Übersichtlichkeit wieder auf den Fall $\Omega = [0, 1]$ und $u_D = 0$, also

$$\begin{aligned}\partial_t^2 u(t, x) - c^2\partial_x^2 u(t, x) &= 0 && \text{für alle } (t, x) \in (0, T) \times (0, 1), \\ u(0, x) &= u_0(x) && \text{für alle } x \in [0, 1], \\ \partial_t u(0, x) &= v_0(x) && \text{für alle } x \in [0, 1] \\ u(t, 0) = u(t, 1) &= 0 && \text{für alle } t \in [0, T].\end{aligned}$$

Die Idee ist es, die beiden zweiten Ableitungen durch einen symmetrischen Differenzenquotienten zweiter Ordnung zu definieren. Dies führt dann auf einen Konsistenzfehler der Ordnung $\mathcal{O}((\Delta t)^2 + (\Delta x)^2)$. Um U_j^1 zu bestimmen, könnten wir $\partial_t u(0, x) = v_0(x)$ benutzen und $U_j^1 = U_j^0 + \Delta t\, v_0(x_j)$ definieren. Dies würde allerdings auf einen Approximationsfehler der Ordnung $\mathcal{O}(\Delta t)$ führen und damit die bessere Konvergenzrate des Verfahrens zerstören. Deshalb führen wir einen zusätzlichen Zeitpunkt $t_{-1} = -\Delta t$ ein und benutzen den zentralen Differenzenquotient zur Approximation von $\partial_t u(0, x_j)$, also

$$\frac{U_j^1 - U_j^{-1}}{2\Delta t} = v_0(x_j) \quad \Leftrightarrow \quad U_j^{-1} = U_j^1 - 2\,\Delta t\, v_0(x_j).$$

Zusammen mit

$$\frac{U_j^1 - 2U_j^0 + U_j^{-1}}{(\Delta t)^2} = \partial_t^+\partial_t^- U_j^0 = c^2\partial_x^+\partial_x^- U_j^0 = c^2\frac{U_{j+1}^0 - 2U_j^0 + U_{j-1}^0}{(\Delta x)^2}$$

ergibt sich mit $\mu = c\Delta t/\Delta x$

$$2U_j^1 = 2(1 - \mu^2)U_j^0 + \mu^2(U_{j+1}^0 + U_{j-1}^0) + 2\Delta t v_0(x_j).$$

Definition 2.13. Das explizite leapfrog-Verfahren ist gegeben durch

$$\begin{aligned}
&\partial_t^+\partial_t^- U_j^k - c^2\partial_x^+\partial_x^- U_j^k = 0 && \text{für alle } j = 1,\dots,J-1;\\
&&& \quad k = 1,\dots,K-1,\\
&U_0^k = U_J^k = 0 && \text{für alle } k = 1,\dots,K,\\
&U_j^0 = u_0(x_j) && \text{für alle } j = 0,\dots,J,\\
&U_j^1 = (1-\mu^2)U_j^0 + \frac{\mu^2}{2}(U_{j+1}^0 + U_{j-1}^0) + \Delta t v_0(x_j) && \text{für alle } j = 1,\dots,J-1. \qquad \diamondsuit
\end{aligned}$$

Bemerkung. Die erste Gleichung ist äquivalent zu

$$U_j^{k+1} = 2\,(1-\mu^2)\,U_j^k - U_j^{k-1} + \mu^2\,(U_{j+1}^k + U_{j-1}^k). \qquad \diamondsuit$$

Für die Stabilität und Fehleranalysis werden wir noch die folgenden Resultate benötigen.

Proposition 2.14 (diskrete inverse Ungleichung)**.** *Es sei* $(V_j)_{j=0,\dots,J} \in \mathbb{R}^{J+1}$*. Dann gilt*

$$\sum_{j=0}^{J-1} \Delta x \left(\frac{V_{j+1}-V_j}{\Delta x}\right)^2 \le \frac{4}{(\Delta x)^2}\sum_{j=0}^{J} \Delta x\, V_j^2.$$

Beweis. Die Ungleichung wird in einer Übungsaufgabe bewiesen. □

Proposition 2.15 (diskrete partielle Integration)**.** *Es seien* $(V_j)_{j=0,\dots,J} \in \mathbb{R}^{J+1}$ *und* $(W_j)_{j=0,\dots,J} \in \mathbb{R}^{J+1}$ *mit* $V_0 = V_J = W_0 = W_J = 0$*. Dann gilt*

$$\sum_{j=1}^{J-1} \Delta x \left(\frac{V_{j+1}-2V_j+V_{j-1}}{(\Delta x)^2}\right) W_j = -\sum_{j=0}^{J-1} \Delta x \left(\frac{V_{j+1}-V_j}{\Delta x}\right)\left(\frac{W_{j+1}-W_j}{\Delta x}\right).$$

Beweis. Die Ungleichung wird in einer Übungsaufgabe bewiesen. □

Wir definieren eine diskrete Energie

$$E^{k+\frac12} := \frac12\sum_{j=0}^{J-1}\Delta x\left(\frac{U_j^{k+1}-U_j^k}{\Delta t}\right)^2 + \frac12 c^2 \sum_{j=0}^{J-1}\Delta x\left(\frac{U_{j+1}^{k+1}-U_j^{k+1}}{\Delta x}\right)\left(\frac{U_{j+1}^k-U_j^k}{\Delta x}\right).$$

Proposition 2.16 (diskrete Energieerhaltung)**.** *Es gilt für alle* $k = 1,\dots,K-1$

$$E^{k+\frac12} = E^{k-\frac12}.$$

Beweis. Wir multiplizieren $\partial_t^+\partial_t^- U_j^k - c^2\partial_x^+\partial_x^- U_j^k = 0$ mit $\Delta x\left(U_j^{k+1} - U_j^{k-1}\right)$ und summieren über $j = 1,\dots,J-1$ und erhalten

$$\begin{aligned}
0 = \sum_{j=1}^{J-1}\Delta x\left(\frac{U_j^{k+1}-2U_j^k+U_j^{k-1}}{(\Delta t)^2}\right)\left(U_j^{k+1}-U_j^{k-1}\right)\\
- c^2\sum_{j=1}^{J-1}\Delta x\left(\frac{U_{j+1}^k-2U_j^k+U_{j-1}^k}{(\Delta x)^2}\right)\left(U_j^{k+1}-U_j^{k-1}\right)
\end{aligned} \tag{2.5}$$

Mit $U_j^{k+1} - U_j^{k-1} = U_j^{k+1} - U_j^k + U_j^k - U_j^{k-1}$ erhalten wir

$$\left(\frac{U_j^{k+1}-2U_j^k+U_j^{k-1}}{(\Delta t)^2}\right)\left(U_j^{k+1}-U_j^{k-1}\right) = \left(\frac{U_j^{k+1}-U_j^k}{\Delta t}\right)^2 - \left(\frac{U_j^k-U_j^{k-1}}{\Delta t}\right)^2$$

Mit der diskreten partiellen Integration aus Proposition 2.15 folgt für den zweiten Term in (2.5)

$$\begin{aligned}
&- c^2 \sum_{j=1}^{J-1} \Delta x \left(\frac{U_{j+1}^k - 2U_j^k + U_{j-1}^k}{(\Delta x)^2} \right) \left(U_j^{k+1} - U_j^{k-1} \right) \\
&= c^2 \sum_{j=0}^{J-1} \Delta x \left(\frac{U_{j+1}^k - U_j^k}{\Delta x} \right) \left(\frac{U_{j+1}^{k+1} - U_j^{k+1}}{\Delta x} \right) - c^2 \sum_{j=0}^{J-1} \Delta x \left(\frac{U_{j+1}^k - U_j^k}{\Delta x} \right) \left(\frac{U_{j+1}^{k-1} - U_j^{k-1}}{\Delta x} \right)
\end{aligned}$$

Also folgt insgesamt mit (2.5), dass

$$0 = E^{k+\frac{1}{2}} - E^{k-\frac{1}{2}}. \qquad \square$$

Wir wollen aus der Energieerhaltung auf die Stabilität des Verfahrens schließen. Dafür werden wir aber die Positivität des zweiten Terms in der Energie benötigen.

Proposition 2.17. *Es gilt*

$$\begin{aligned}
2(1-2\mu^2) \sum_{j=0}^{J-1} \Delta x \left(\frac{U_j^{k+1} - U_j^k}{\Delta t} \right)^2 + c^2 \sum_{j=0}^{J-1} \Delta x \left(\frac{U_{j+1}^{k+1} - U_j^{k+1}}{\Delta x} \right)^2 + c^2 \sum_{j=0}^{J-1} \Delta x \left(\frac{U_{j+1}^k - U_j^k}{\Delta x} \right)^2 \\
\leq 4E^{k+\frac{1}{2}}.
\end{aligned}$$

Wenn $\mu = c\Delta t/\Delta x \leq 1/\sqrt{2}$, dann ist die linke Seite nicht negativ.

Beweis. Es gilt

$$2E^{k+\frac{1}{2}} = \sum_{j=0}^{J-1} \Delta x \left(\frac{U_j^{k+1} - U_j^k}{\Delta t} \right)^2 + c^2 \sum_{j=0}^{J-1} \Delta x \left(\frac{U_{j+1}^{k+1} - U_j^{k+1}}{\Delta x} \right) \left(\frac{U_{j+1}^k - U_j^k}{\Delta x} \right).$$

Für den zweiten Term gilt

$$\begin{aligned}
&c^2 \sum_{j=0}^{J-1} \Delta x \left(\frac{U_{j+1}^{k+1} - U_j^{k+1}}{\Delta x} \right) \left(\frac{U_{j+1}^k - U_j^k}{\Delta x} \right) \\
&\qquad = -\frac{c^2}{2} (\Delta t)^2 \sum_{j=0}^{J-1} \Delta x \left(\frac{U_{j+1}^{k+1} - U_j^{k+1} - (U_{j+1}^k - U_j^k)}{\Delta x \Delta t} \right)^2 \\
&\qquad\qquad + \frac{c^2}{2} \sum_{j=0}^{J-1} \Delta x \left(\frac{U_{j+1}^{k+1} - U_j^{k+1}}{\Delta x} \right)^2 + \frac{c^2}{2} \sum_{j=0}^{J-1} \Delta x \left(\frac{U_{j+1}^k - U_j^k}{\Delta x} \right)^2.
\end{aligned}$$

Mit der diskreten inversen Ungleichung aus Proposition 2.14 gilt für den ersten Term auf der rechten Seite

$$\begin{aligned}
&- \frac{c^2}{2} (\Delta t)^2 \sum_{j=0}^{J-1} \Delta x \left(\frac{U_{j+1}^{k+1} - U_j^{k+1} - (U_{j+1}^k - U_j^k)}{\Delta x \Delta t} \right)^2 \\
&\qquad\qquad \geq -2c^2 \left(\frac{\Delta t}{\Delta x} \right)^2 \sum_{j=0}^{J-1} \Delta x \left(\frac{U_j^{k+1} - U_j^k}{\Delta t} \right)^2.
\end{aligned}$$

Insgesamt folgt also die Behauptung. $\square$

Satz 2.18. *Es existiert eine eindeutige Lösung des expliziten Verfahrens. Gilt $\mu \leq 1/\sqrt{2}$, dann ist sie stabil im Sinne, dass*

$$c^2 \sum_{j=0}^{J-1} \Delta x \left(\frac{U_{j+1}^{k+1} - U_j^{k+1}}{\Delta x} \right)^2 + c^2 \sum_{j=0}^{J-1} \Delta x \left(\frac{U_{j+1}^{k} - U_j^{k}}{\Delta x} \right)^2$$
$$\leq 2 \sum_{j=0}^{J-1} \Delta x \left(\frac{U_j^1 - U_j^0}{\Delta t} \right)^2 + 2 \sum_{j=0}^{J-1} \Delta x \left(\frac{U_{j+1}^1 - U_j^1}{\Delta x} \right) \left(\frac{U_{j+1}^0 - U_j^0}{\Delta x} \right).$$

Außerdem gilt für $e_j^k := u(t_k, x_j) - U_j^k$ mit einer Konstanten $C < \infty$ die Fehlerabschätzung

$$c^2 \sqrt{\sum_{j=0}^{J-1} \Delta x \left(\frac{e_{j+1}^{k+1} - e_j^{k+1}}{\Delta x} \right)^2} \leq C t_k \left((\Delta t)^2 + (\Delta x)^2 \right) \|u\|_{\mathcal{C}^4([0,T]\times[0,1])}.$$

Bemerkung. Mit einer anderen Methode kann man zeigen, dass das explizite Verfahren für die Wellengleichung auch für $\mu \leq 1$ stabil ist, siehe z.B. [Bar16]. ♢

Beweis von Satz 2.18. Die eindeutige Lösbarkeit folgt direkt aus der Definition, weil das Verfahren explizit ist.
Für das Verfahren mit inhomogener rechter Seite, also

$$\partial_t^+ \partial_t^- U_j^k - c^2 \partial_x^+ \partial_x^- U_j^k = F_j^k,$$

gilt die allgemeinere Energieerhaltung

$$E^{k+\frac{1}{2}} = E^{k-\frac{1}{2}} + \frac{1}{2} \sum_{j=0}^{J-1} \Delta x F_j^k \underbrace{\left(U_j^{k+1} - U_j^{k-1} \right)}_{= \left(\frac{U_j^{k+1} - U_j^k}{\Delta t} + \frac{U_j^k - U_j^{k-1}}{\Delta t} \right) \Delta t}.$$

Also folgt mit einer Cauchy-Ungleichung im $\mathbb{R}^J$ und Proposition 2.17

$$E^{k+\frac{1}{2}} - E^{k-\frac{1}{2}}$$
$$\leq \frac{1}{2} \Delta t \sqrt{\sum_{j=0}^{J-1} \Delta x (F_j^k)^2} \Bigg(\underbrace{\sqrt{\sum_{j=0}^{J-1} \Delta x \left(\frac{U_j^{k+1} - U_j^k}{\Delta t} \right)}}_{\leq \frac{2}{\sqrt{2(1-2\mu^2)}} \sqrt{E^{k+\frac{1}{2}}}} + \underbrace{\sqrt{\sum_{j=0}^{J-1} \Delta x \left(\frac{U_j^{k} - U_j^{k-1}}{\Delta t} \right)^2}}_{\leq \frac{2}{\sqrt{2(1-2\mu^2)}} \sqrt{E^{k-\frac{1}{2}}}} \Bigg).$$

Also folgt

$$\frac{E^{k+\frac{1}{2}} - E^{k-\frac{1}{2}}}{\sqrt{E^{k+\frac{1}{2}}} + \sqrt{E^{k-\frac{1}{2}}}} \leq \frac{\Delta t}{\sqrt{2(1-2\mu^2)}} \sqrt{\sum_{j=0}^{J-1} \Delta x (F_j^k)^2}.$$

Andererseits gilt

$$\frac{E^{k+\frac{1}{2}} - E^{k-\frac{1}{2}}}{\sqrt{E^{k+\frac{1}{2}}} + \sqrt{E^{k-\frac{1}{2}}}} = \sqrt{E^{k+\frac{1}{2}}} - \sqrt{E^{k-\frac{1}{2}}}.$$

Aus einer Summation über $\ell = 1, \dots, k$ folgt die Stabilität

$$\sqrt{E^{k+\frac{1}{2}}} \leq \sqrt{E^{\frac{1}{2}}} + \frac{1}{\sqrt{2(1-2\mu^2)}} \sum_{\ell=1}^{k} \Delta t \sqrt{\sum_{j=0}^{J-1} \Delta x (F_j^\ell)^2}. \tag{2.6}$$

Mit Proposition 2.17 folgt die Stabilität in Satz 2.18.
Der Fehler e_j^k erfüllt das leapfrog-Verfahren mit dem Konsistenzfehler

$$\varepsilon_j^k := \partial_t^+ \partial_t^- u(t_k, x_j) - \partial_x^+ \partial_x^- u(t_k, x_j)$$

auf der rechten Seite. Mit

$$|\varepsilon_j^k| \leq \frac{(\Delta t)^2 + (\Delta x)^2}{12} \|u\|_{\mathcal{C}^4([0,T]\times[0,1])}$$

und der Stabilität (2.6) und Proposition 2.17 folgt

$$c^2 \sqrt{\sum_{j=0}^{J-1} \Delta x \left(\frac{e_{j+1}^{k+1} - e_j^{k+1}}{\Delta x} \right)^2}$$
$$\leq \frac{(\Delta t)^2 + (\Delta x)^2}{6\sqrt{2(1-2\mu^2)}} t_k \left((\Delta t)^2 + (\Delta x)^2 \right) \|u\|_{\mathcal{C}^4([0,T]\times[0,1])} + \sqrt{2 \sum_{j=0}^{J-1} \Delta x \left(\frac{e_j^1}{\Delta t} \right)^2}.$$

Mit der Definition von U_j^1 und einer Taylor-Entwicklung von $u(t_1, x_j)$ um $u(0, x_j)$ folgt für ein $R_j \in \mathbb{R}$

$$\begin{aligned} e_j^1 &= U_j^0 + \frac{\mu^2}{2} \left(U_{j+1}^0 - 2U_j^0 + U_{j-1}^0 \right) + \Delta t v_0(x_j) \\ &\qquad - u_0(x_j) - \Delta t v_0(x_j) - \frac{\Delta t^2}{2} \partial_t^2 u(0, x_j) + (\Delta t)^3 R_j \\ &= \frac{c^2 \Delta t^2}{2} \partial_x^+ \partial_x^- U_j^0 - \frac{\Delta t^2}{2} \partial_t^2 u(0, x_j) + (\Delta t)^3 R_j. \end{aligned}$$

Mit der Wellengleichung für die exakte Lösung folgt mit einer Konstanten $\tilde{C} < \infty$

$$|e_j^1| \leq \tilde{C} \left(\Delta t^2 (\Delta x)^2 + (\Delta t)^3 \right) \|u\|_{\mathcal{C}^4([0,T]\times[0,1])}.$$

Damit folgt die Behauptung. □

Korollar 2.19. *Es existiert eine Konstante $\widetilde{C}$ mit*

$$c^2 \sqrt{\sum_{j=0}^{J-1} \Delta x |e_j^{k+1}|^2} \leq c^2 \widetilde{C} \sqrt{\sum_{j=0}^{J-1} \Delta x \left(\frac{e_{j+1}^{k+1} - e_j^{k+1}}{\Delta x} \right)^2}$$
$$\leq \widetilde{C} C t_k \left((\Delta t)^2 + (\Delta x)^2 \right) \|u\|_{\mathcal{C}^4([0,T]\times[0,1])}.$$

Beweis. Die erste Ungleichung ist eine Art diskrete Poincaré-Ungleichung und wird in einer Übungsaufgabe bewiesen. □

Bemerkung. 1. Die Stabilitätsbedingung $c\Delta t \leq \Delta x$ wird auch die CFL-Bedingung (nach Courant, Friedrichs und Levy) genannt. Sie besagt, dass das numerische Abhängigkeitsgebiet das tatsächliche Abhängigkeitsgebiet einschließen muss. Die exakte Lösung u am Punkt (t, x) hängt nämlich nur von einem Kegel ab, dessen Weite durch die Ausbreitungsgeschwindigkeit c bestimmt wird, siehe auch Abbildung 2.3.

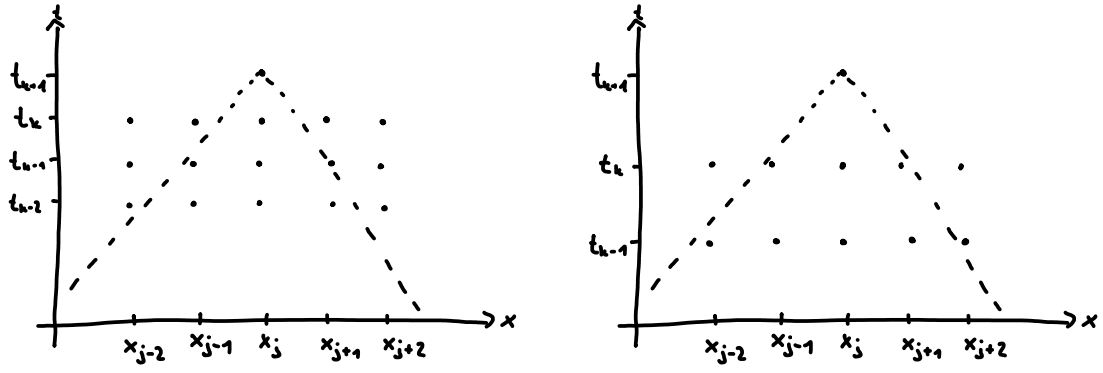

Abbildung 2.3: Numerisches und exaktes Abhängigkeitsgebiet in der Wellengleichung. In der linken Abbildung ist die CFL-Bedingung erfüllt, auf der rechten verletzt.

2. Verfahren, die keine (diskrete) Energieerhaltung erfüllen, führen häufig auf physikalisch nicht sinnvolle diskrete Lösungen, so zum Beispiel das implizite, stabile Verfahren, das durch

$$\partial_t^+\partial_t^- U_j^k - c^2\partial_x^+\partial_x^- U_j^{k+1} = \frac{U_j^{k+1} - 2U_j^k + U_j^{k-1}}{(\Delta t)^2} - c^2\frac{U_{j+1}^{k+1} - 2U_j^{k+1} + U_{j-1}^{k+1}}{(\Delta x)^2} = 0$$

gegeben ist.

3. Ein implizites Verfahren, das eine diskrete Energieerhaltung erfüllt, ist ein Crank-Nicolson-Verfahren, das durch

$$\partial_t^+\partial_t^- U_j^k - \frac{c^2}{4}\partial_x^+\partial_x^-(U_j^{k+1} + 2U_j^k + U_j^{k-1}) = 0$$

gegeben ist. Dieses Verfahren ist im Gegensatz zum expliziten leapfrog-Verfahren für alle Werte von μ stabil. Es hat dieselbe Konvergenzordnung $((\Delta t)^2 + (\Delta x)^2)$ wie das explizite leapfrog-Verfahren.

♢

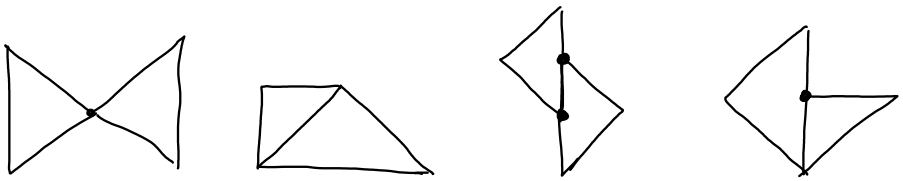

Abbildung 3.1: Reguläre und nicht-reguläre Triangulierungen.

3 Triangulierungen und Gitterverfeinerung

Finite Differenzen wurden auf Gitterpunkten $(k\Delta t, j\Delta x)$ definiert. Wir werden im nächsten Kapitel Finite Elemente definieren, die auf Triangulierungen in Dreiecke definiert sind.

Definition 3.1 (Simplex)**.** Ein Simplex $T \subseteq \mathbb{R}^d$ ist die konvexe Hülle aus $d+1$ Punkten, d.h. $\exists x_1, \dots, x_{d+1} \in \mathbb{R}^d$ mit

$$T = \operatorname{conv}\{x_1, \dots, x_{d+1}\}. \qquad \diamondsuit$$

Für $d = 1$ sind Simplices Intervalle, für $d = 2$ Dreiecke und für $d = 3$ Tetraeder.

Definition 3.2 (reguläre Triangulierung)**.** Eine reguläre (auch konforme) Triangulierung $\mathcal{T}$ von einem beschränkten, polygonalen Lipschitz-Gebiet $\Omega \subseteq \mathbb{R}^d$ ist eine Menge $\mathcal{T} = \{T_1, \dots, T_L\}$ von abgeschlossenen Simplices mit $\overline{\Omega} = \bigcup_{T\in\mathcal{T}} T$ und der Schnitt von zwei Elementen $T_1, T_2 \in \mathcal{T}$ ist entweder leer oder besteht aus einem ganzen Subsimplex beider Elemente. $\diamondsuit$

Wir werden Knoten, Kanten und Flächen in der Triangulierung benötigen.

Definition 3.3. Es bezeichne $\mathcal{N}$ die Menge aller Ecken von Dreiecken in $\mathcal{T}$, $\mathcal{E}$ die Menge aller Kanten und $\mathcal{F}$ die Menge aller Flächen von Simplices., d.h.

$$\begin{aligned}
\mathcal{N} &:= \{x_1 \in \mathbb{R}^d \mid \exists T \in \mathcal{T}, \exists x_2, \dots, x_{d+1} \in \mathbb{R}^d \text{ mit } T = \operatorname{conv}\{x_1, \dots, x_{d+1}\}\},\\
\mathcal{E} &:= \{\operatorname{conv}\{x_1, x_2\} \mid \exists T \in \mathcal{T}, \exists x_3, \dots, x_{d+1} \in \mathbb{R}^d \text{ mit } T = \operatorname{conv}\{x_1, \dots, x_{d+1}\}\},\\
\mathcal{F} &:= \{\operatorname{conv}\{x_1, \dots, x_d\} \mid \exists T \in \mathcal{T}, \exists x_{d+1} \in \mathbb{R}^d \text{ mit } T = \operatorname{conv}\{x_1, \dots, x_{d+1}\}\}.
\end{aligned}$$

Für $d = 1$ setzen wir $\mathcal{E} := \mathcal{N}$. Außerdem gilt für $d = 1$, dass $\mathcal{N} = \mathcal{F}$ und für $d = 2$, dass $\mathcal{F} = \mathcal{E}$. $\diamondsuit$

Bemerkung. Analog lässt sich auch eine reguläre Triangulierung in Vierecke oder andere Formen definieren. $\diamondsuit$

Beispiel 3.4. In Abbildung 3.1 sind auf dem ersten Bild zwei Dreiecke zu sehen, dessen Schnitt aus einer Ecke besteht, d.h. der Schnitt besteht aus einem ganzen Subsimplex. Auf dem zweiten Bild besteht der Schnitt in einer Kante, also auch in einem ganzen Subsimplex. Auf dem dritten und vierten Bild hingegen ist der Schnitt fuer eines der Dreiecke kein ganzer Subsimplex, d.h. eine solche Situation ist in einer regulären Triangulierung nicht zulässig. Die markierten Ecken werden auch *hängende Knoten* genannt. $\diamondsuit$

In der Praxis ist eine Triangulierung $\mathcal{T}$ von $\Omega \subseteq \mathbb{R}^2$ gegeben durch einen Vektor $\texttt{c4n} \in \mathbb{R}^{|\mathcal{N}|\times 2}$ und $\texttt{n4e} \in \mathbb{R}^{|\mathcal{T}|\times 3}$, wobei $|\bullet|$ hier die Anzahl an Elementen bezeichne und

$$\texttt{c4n} = \begin{pmatrix} z_1^\top \\ \vdots \\ z_{|\mathcal{N}|}^\top \end{pmatrix}$$

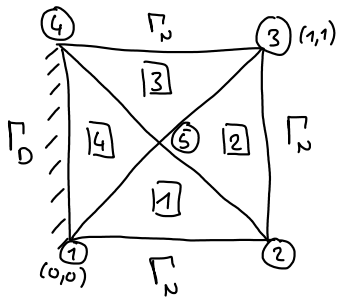

Abbildung 3.2: Ein Beispiel für eine Triangulierung mit nummerierten Dreiecken und Knoten.

und

$$\texttt{n4e}(j,:) = \begin{pmatrix} n_1 & n_2 & n_3 \end{pmatrix},$$

wenn das Dreieck mit Knoten $z_{n_1}, z_{n_2}, z_{n_3}$ in $\mathcal{T}$ ist. Dabei befolgen wir die Konvention, dass die längste Seite zwischen z_{n_1} und z_{n_2} ist und die Nummerierung gegen den Uhrzeigersinn erfolgt. Der Vektor `c4n` definiert eine Nummerierung der Knoten und `n4e` eine Nummerierung der Dreiecke. Die Vektoren $\texttt{n4sDb} \in \mathbb{R}^{|\mathcal{E}_{\Gamma_D}|\times 2}$ und $\texttt{n4sNb} \in \mathbb{R}^{|\mathcal{E}_{\Gamma_N}|\times 2}$ enthalten die Knoten der Kanten, die auf dem Dirichlet- beziehungsweise dem Neumann-Rand liegen. Auf diese Weise können wir per Hand eine grobe (Anfangs-)Triangulierung definieren, auf der wir unser FEM-Programm laufen lassen können.

Beispiel 3.5. Abbildung 3.2 zeigt eine Beispiel-Triangulierung. Für diese Triangulierung gilt

$$\texttt{c4n} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0.5 & 0.5 \end{bmatrix}, \qquad \texttt{n4e} = \begin{bmatrix} 1 & 2 & 5 \\ 2 & 3 & 5 \\ 3 & 4 & 5 \\ 4 & 1 & 5 \end{bmatrix}, \qquad \texttt{n4sDb} = \begin{bmatrix} 4 & 1 \end{bmatrix}, \qquad \texttt{n4sNb} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 4 \end{bmatrix}.$$

♢

Außer der Regularität der Triangulierung werden wir in dieser Vorlesung auch die Form-Regularität der Triangulierung voraussetzen. Für die Definition benötigen wir noch die folgenden Begriffe.

Definition 3.6. Für eine Menge $T \subseteq \mathbb{R}^d$ heißt

$$h_T = \operatorname{diam}(T) = \sup\{|x - y| \mid x, y \in T\}$$

der Durchmesser oder Diameter von T und

$$\rho_T = \sup\{r \mid \exists x \in T \text{ mit } B_r(x) \subseteq T\}$$

bezeichne den Radius des größten Kreises, der in T liegt. ♢

Definition 3.7. Eine Familie von regulären Triangulierungen $(\mathcal{T}_k)_{k\in\mathbb{N}}$ heißt (uniform) form-regulär, wenn eine Konstante $C < \infty$ existiert mit

$$\sup_{k\in\mathbb{N}} \sup_{T\in\mathcal{T}_k} \frac{h_T}{\rho_T} \leq C.$$ ♢

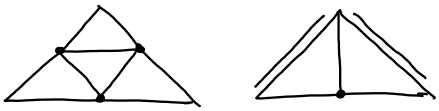

Abbildung 3.3: Rot-Verfeinerung (links) und Bisektion (rechts) eines Dreiecks.

Um genauere Lösungen berechnen zu können, wollen wir ein Gitter verfeinern können. Für $d = 2$ können wir dafür ein Gitter rot verfeinern oder die Dreiecke durch Bisektion teilen.

Definition 3.8 (Rot-Verfeinerung in 2d)**.** Für ein gegebenes Dreieck $T = \operatorname{conv}\{x_1, x_2, x_3\} \subseteq \mathbb{R}^2$, definiere

$$y_j := \frac{x_j + x_{j+1}}{2},$$

wobei Indices hier modulo 3 verstanden werden. Es sei weiterhin

$$\begin{aligned} T_j &:= \operatorname{conv}\{x_j, y_j, y_{j-1}\} \qquad \text{für } j = 1, 2, 3, \\ T_4 &:= \operatorname{conv}\{y_2, y_3, y_1\}. \end{aligned}$$

Dann ist durch $\{T_1, \dots, T_4\}$ die Rot-Verfeinerung von T gegeben. ♢

Definition 3.9 (Bisektion)**.** Es sei $T = \operatorname{conv}\{x_1, x_2, x_3\}$ gegeben mit Referenzkante $E = \operatorname{conv}\{x_1, x_2\}$. Definiere

$$y := \frac{x_1 + x_2}{2}$$

und

$$\begin{aligned} T_1 &:= \operatorname{conv}\{x_3, x_1, y\}, & E_1 &:= \operatorname{conv}\{x_3, x_1\}, \\ T_2 &:= \operatorname{conv}\{x_2, x_3, y\}, & E_2 &:= \operatorname{conv}\{x_2, x_3\}. \end{aligned}$$

Dann ist durch $\{(T_1, E_1), (T_2, E_2)\}$ die Bisektion von T gegeben, wobei E_j die Referenzkante von T_j bezeichne. ♢

Bemerkung. In der Praxis definieren wir die Referenzkante des Dreiecks

$$\operatorname{conv}\{\texttt{c4n(n4e(j,1),:)}, \texttt{c4n(n4e(j,2),:)}, \texttt{c4n(n4e(j,3),:)}\}$$

durch

$$\operatorname{conv}\{\texttt{c4n(n4e(j,1),:)}, \texttt{c4n(n4e(j,2),:)}\}. \qquad ♢$$

Ein rot-verfeinertes Dreieck und eines, bei dem die Bisektion angewandt wurde, sind in Abbildung 3.3 zu sehen.
Bei der Rot-Verfeinerung sind die neu entstehenden Seiten alle parallel zu einer der alten. Die neu entstehenden Dreiecke sind alle ähnlich zum ursprünglichen Dreieck. Eine Familie von Triangulierungen $\{\mathcal{T}_k\}_{k\in\mathbb{N}}$, die durch Rot-Verfeinerung aus $\mathcal{T}_0$ entstanden ist, ist deshalb formregulär (mit Konstante abhängig von $\mathcal{T}_0$.
Bei einer Bisektion können neue Winkel entstehen. Nach zwei Bisektionen kommen aber keine neuen Winkel mehr hinzu. Damit ist die Familie von Triangulierungen, die durch Bisektion entsteht, auch wieder uniform formregulär. Der Vorteil der Bisektion ist, dass Gitter auch lokal verfeinert werden können, siehe Abbildung 3.4. Dies kann sinnvoll sein, um zum Beispiel Singularitäten der Lösung, die Geometrie oder Koeffizienten aufzulösen.

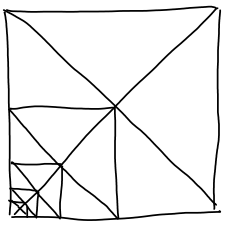

Abbildung 3.4: Lokal verfeinertes Gitter, das durch Bisektion entstanden ist.

4 Klassische Finite-Elemente-Methoden für Elliptische Probleme

In diesem Kapitel werden wir klassische Finite-Elemente-Methoden (FEM) für elliptische Probleme betrachten. Den FEMs liegt die schwache Formulierung der Probleme zugrunde und die funktionalanalytische Theorie dazu.

4.1 Galerkin-Verfahren

Klassische FEMs sind spezielle Galerkin-Verfahren. Diese werden wir in diesem Abschnitt definieren.
In diesem Abschnitt bezeichne V einen Hilbertraum mit Skalarprodukt $(\bullet,\bullet)_V$, d.h. V ist ein $\mathbb{R}$ oder $\mathbb{C}$-Vektorraum und V ist vollständig bezüglich der induzierten Norm $\|\bullet\|_V := \sqrt{(\bullet,\bullet)_V}$. Meistens wird V unendlichdimensional sein. Es bezeichne V' den *Dualraum* von V, d.h. V' ist der Raum aller linearen Abbildungen $F : V \to \mathbb{R}$, die stetig sind. Äquivalent zur Stetigkeit von F ist, dass eine Konstante c existiert mit

$$|F(v)| \leq c\|v\|_V \qquad \text{für alle } v \in V.$$

Dann ist

$$\|F\|_{V'} := \sup_{v\in V\setminus\{0\}} \frac{|F(v)|}{\|v\|_V}$$

eine Norm auf V'. Wir betrachten zu einer Bilinearform $a : V \times V \to \mathbb{R}$ und einem $F \in V'$ das Problem:

Problem 4.1. Finde $u \in V$ mit

$$a(u,v) = F(v) \qquad \text{für alle } v \in V. \tag{4.1}$$

♢

Wir zitieren als nächstes einen Satz aus der Funktionalanalysis, der uns die Lösbarkeit von unserem abstrakten Problem zeigen wird.

Satz 4.2 (Lax-Milgram)**.** *Es sei V ein Hilbert-Raum. Ist die Bilinearform a*

(i) koerzitiv, d.h. es existiert $\alpha > 0$ mit

$$a(u,u) \geq \alpha\|u\|_V^2 \qquad \textit{für alle } u \in V,$$

(ii) beschränkt, d.h. es existiert $k_a < \infty$ mit

$$|a(u,v)| \leq k_a\|u\|_V\|v\|_V \qquad \textit{für alle } u,v \in V,$$

dann existiert für alle $F \in V'$ eine eindeutige Lösung $u \in V$ des Problems 4.1 und es gilt

$$\|u\|_V \leq \frac{1}{\alpha}\|F\|_{V'}.$$

Beweis. Siehe Vorlesung Funktionalanalysis oder [Alt16]. □

Wir nehmen im Folgenden an, dass a koerzitiv und beschränkt ist. Wir interessieren uns für Approximationen von u.

Definition 4.3. Für einen abgeschlossenen Teilraum $V_h \subseteq V$ heißt $u_h \in V_h$ Galerkin-Approximation von $u \in V$, wenn

$$a(u_h, v_h) = F(v_h) \qquad \text{für alle } v_h \in V_h. \tag{4.2}$$

♢

Wir werden als erstes zeigen, dass die Galerkin-Approximation bis auf eine Konstante die beste Approximation ist.

Satz 4.4 (Céas Lemma)**.** *Es existiert eine eindeutige Galerkin-Approximation $u_h \in V_h$ und es gilt die Galerkin-Orthogonalität*

$$a(u - u_h, v_h) = 0 \qquad \textit{für alle } v_h \in V_h$$

und die Quasi-Bestapproximationseigenschaft

$$\|u - u_h\|_V \leq \frac{k_a}{\alpha} \inf_{w_h \in V_h} \|u - w_h\|_V.$$

Beweis. Die Existenz einer eindeutigen Lösung folgt mit dem Satz von Lax-Milgram (da V_h abgeschlossen ist, ist V_h ein Hilbertraum). Aus (4.1) und (4.2) folgt für alle $v_h \in V_h \subseteq V$, dass

$$a(u - u_h, v_h) = a(u, v_h) - a(u_h, v_h) = F(v_h) - F(v_h) = 0.$$

Dies ist die Galerkin-Orthogonalität.
Es gilt wegen der Koerzitivät und Beschränktheit von a und der Galerkin-Orthogonalität für alle $w_h \in V_h$, dass

$$\begin{aligned}\alpha\|u - u_h\|_V^2 &\leq a(u - u_h, u - u_h) = a(u - u_h, u) \\ &= a(u - u_h, u - w_h) \leq k_a\|u - u_h\|_V\|u - w_h\|_V.\end{aligned}$$

Dies ist die Bestapproximationseigenschaft. □

Als nächstes widmen wir uns der Frage, wie die Lösung u_h von (4.2) berechnet werden kann. Ab jetzt sei V_h eindlich-dimensional und es sei $(\varphi_1, \dots, \varphi_n)$ eine Basis von V_h.

Definition 4.5. Die Matrix $A \in \mathbb{R}^{n\times n}$, die durch

$$A_{jk} := a(\varphi_j, \varphi_k) \qquad \text{für alle } 1 \leq j, k \leq n$$

definiert ist, heißt Steifigkeitsmatrix. ♢

Definiere außerdem einen Vektor $b \in \mathbb{R}^n$ durch

$$b_j := F(\varphi_j).$$

Proposition 4.6. *Die Steifigkeitsmatrix A ist positiv definit und für die eindeutige Lösung $U \in \mathbb{R}^n$ von*

$$A^\top U = b$$

gilt

$$u_h = \sum_{j=1}^n U_j \varphi_j.$$

Beweis. Für $W \in \mathbb{R}^n$ definiere $w_h := \sum_{j=1}^n W_j \varphi_j$. Dann gilt

$$\begin{aligned} W^\top A W &= \sum_{j,k=1}^n W_j A_{jk} W_k = \sum_{j,k=1}^n W_j a(\varphi_j, \varphi_k) W_k \\ &= \sum_{j,k=1}^n a(W_j \varphi_j, W_k \varphi_k) = a(w_h, w_h) \geq \alpha \|w_h\|_V^2. \end{aligned}$$

Da $w_h = 0$ genau dann, wenn $W = 0$, folgt, dass A positiv definit ist und damit invertierbar. Es seien nun $w_h \in V_h$ und $W \in \mathbb{R}^n$ mit $w_h = \sum_{j=1}^n W_j \varphi_j$. Dann gilt

$$\begin{aligned} a(u_h, w_h) &= \sum_{j,k=1}^n U_j a(\varphi_j, \varphi_k) W_k = (A^\top U)^\top W, \\ F(w_h) &= \sum_{j=1}^n W_j F(\varphi_j) = \sum_{j=1}^n W_j b_j = b^\top W, \end{aligned}$$

also

$$a(u_h, w_h) = F(w_h) \Leftrightarrow A^\top U = b. \qquad \square$$

4.2 Schwache Formulierung des Poisson Problems

Wir betrachten nun das Poisson-Problem auf einem Gebiet $\Omega \subseteq \mathbb{R}^d$.

Definition 4.7. Ein Gebiet $\Omega \subseteq \mathbb{R}^d$ heißt Lipschitz-Gebiet, falls gilt

(a) Ω ist offen und zusammenhängend

(b) Zu jedem $x \in \partial\Omega$ existiert eine Transformation $\Phi(y) = My + r$ mit Orthogonalmatrix $M \in \mathbb{R}^{d \times d}$ und Vektor $r \in \mathbb{R}^d$, ein Parameter $\delta > 0$, eine offene Menge $Q' \subseteq \mathbb{R}^{d-1}$ und eine Lipschitz-stetige Funktion $h : Q' \to \mathbb{R}$ mit

$$\begin{aligned} \Omega \cap B_\delta(x) &= \Phi(\{(y', y_d) \in Q' \times \mathbb{R} : h(y') < y_d\}) \cap B_\delta(x), \\ \partial\Omega \cap B_\delta(x) &= \Phi(\{(y', y_d) \in Q' \times \mathbb{R} : h(y') = y_d\}) \cap B_\delta(x), \\ (\mathbb{R}^d \setminus \overline{\Omega}) \cap B_\delta(x) &= \Phi(\{(y', y_d) \in Q' \times \mathbb{R} : h(y') > y_d\}) \cap B_\delta(x). \end{aligned}$$ ♢

Lipschitz-Gebiete sind also Gebiete, deren Rand durch eine Lipschitz-Funktion parametrisiert werden kann und die lokal auf einer Seite des Randes liegen.

Beispiel 4.8. Abbildung 4.1 zeigt Beispiele für Lipschitz-Gebiete, Abbildung 4.2 zeigt Beispiele für Gebiete, die keine Lipschitz-Gebiete sind. ♢

Auf dem folgenden Satz von Gauß basiert die schwache Formulierung des Poisson-Problems.

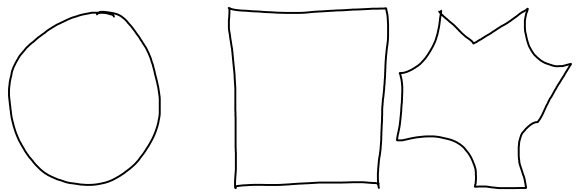

Abbildung 4.1: Diese Gebiete sind Lipschitz-Gebiete

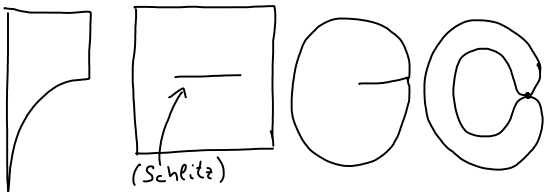

Abbildung 4.2: Diese Gebiete sind keine Lipschitz-Gebiete

Satz 4.9 (Satz von Gauß, partielle Integration)**.** *Ist $\Omega \subseteq \mathbb{R}^d$ ein Lipschitz-Gebiet und $F \in \mathcal{C}^1(\overline{\Omega};\mathbb{R}^d)$, $f, g \in \mathcal{C}^1(\overline{\Omega})$, dann gilt*

$$\begin{aligned}\int_\Omega \operatorname{div}(F)\,dx &= \int_{\partial\Omega} F\cdot n\,ds,\\ \int_\Omega f\partial_j g\,dx &= -\int_\Omega g\partial_j f\,dx + \int_{\partial\Omega} fgn_j\,ds,\\ \int_\Omega F\cdot\nabla g\,dx &= -\int_\Omega g\operatorname{div}(F)\,dx + \int_{\partial\Omega} g\,F\cdot n\,ds,\end{aligned}$$

wobei n die äußere Normal bezeichne.

Beweis. Der Beweis der ersten Aussage kann zum Beispiel in [Alt16] nachgeschlagen werden. Die zweite und dritte Gleichung folgt aus der ersten mit geeigneter Wahl von F. □

Für den Rest der Vorlesung nehmen wir an, dass Ω ein Lipschitz-Gebiet ist, sodass die partielle Integration gilt.
Für den Rand von Ω nehmen wir an, dass wir ihn aufteilen können als $\partial\Omega = \Gamma_D \cup \Gamma_N$, wobei Γ_D abgeschlossen sei und $\Gamma_N = \partial\Omega \setminus \Gamma_D$. Wir möchten für gegebenes $f \in \mathcal{C}(\overline{\Omega})$ und $g \in \mathcal{C}(\Gamma_N)$ und $u_D = 0$ das Poisson-Problem aus Beispiel 1.5 lösen, d.h. es soll

$$\begin{aligned}-\Delta u &= f && \text{in } \Omega,\\ u|_{\Gamma_D} &= 0 && \text{auf } \Gamma_D,\\ (\nabla u\cdot n)|_{\Gamma_N} &= g && \text{auf } \Gamma_N.\end{aligned} \tag{4.3}$$

gelten.
Die schwache Formulierung verallgemeinert das Poisson-Problem auf Situationen, in denen keine Lösung in $\mathcal{C}^1(\overline{\Omega})\cap\mathcal{C}^2(\Omega)$ existiert.

Definition 4.10 (schwache Formulierung des Poisson-Problems)**.** Die schwache Formulierung des Poisson-Problems sucht $u \in V$ (V wird im Folgenden in (4.6) definiert werden) mit $u|_{\Gamma_D} = 0$ und

$$\int_\Omega \nabla u\cdot\nabla v\,dx = \int_\Omega f\,v\,dx + \int_{\Gamma_N} g\,v\,ds \qquad \text{für alle } v \in V \text{mit } v|_{\Gamma_D} = 0. \tag{4.4}$$

♢

Der Raum V wird eine Obermenge von $\mathcal{C}^1(\overline{\Omega}) \cap \mathcal{C}^2(\Omega)$ sein.
Ist $u \in \mathcal{C}^1(\overline{\Omega}) \cap \mathcal{C}^2(\Omega)$, dann sind die starke und die schwache Formulierung äquivalent. Außerdem ist die Lösung der schwachen Formulierung äquivalent zum Minimieren von

$$\int_\Omega |\nabla u|^2 \, dx - \int_\Omega f u \, dx$$

in V. Dies wird in einer Übungsaufgabe gezeigt.
Wir wollen im Folgenden den Satz 4.2 von Lax-Milgram auf die schwache Formulierung anwenden. Dazu werden wir als Norm

$$\|v\|_V := \sqrt{\int_\Omega \nabla v \cdot \nabla v \, dx}$$

betrachten. Der Satz von Lax-Milgram gilt in Hilberträumen, d.h. der Raum V aus Definition 4.10 muss vollständig sein bezüglich dieser Norm. In einer Übungsaufgabe wird gezeigt, dass der Raum $\mathcal{C}^1(\overline{\Omega})$ nicht vollständig ist. Wir werden im Folgenden einen Raum definieren, der vollständig sein wird, er kann auch als Vervollständigung von $\mathcal{C}^1(\overline{\Omega})$ beziehungsweise $\mathcal{C}^\infty(\overline{\Omega})$ definiert werden. Wir werden hier aber anders vorgehen.
Die partielle Integration motiviert die folgende Definition.

Definition 4.11 (schwache Differenzierbarkeit, schwache Ableitung)**.** Es sei $u \in L^1(\Omega)$. Dann ist u schwach differenzierbar, wenn es für $j = 1, \dots, d$ ein $g_j \in L^1(\Omega)$ gibt mit

$$\int_\Omega u \partial_j \varphi \, dx = -\int_\Omega g_j \varphi \, dx \qquad \text{für alle } \varphi \in \mathcal{C}_c^\infty(\Omega).$$

Die Funktion $\partial_j u := g_j$ heißt schwache partielle Ableitung von u bezüglich j und $\nabla u := \begin{pmatrix} \partial_1 u & \dots & \partial_d u \end{pmatrix}^\top$ heißt schwacher Gradient. Die Funktion $u \in L^1(\Omega)$ ist k-mal schwach differenzierbar, wenn für alle Multiindices $\alpha \in \mathbb{N}_0$ mit $|\alpha| = \alpha_1 + \dots + \alpha_d \le k$ eine Funktion $g_\alpha \in L^1(\Omega)$ existiert mit

$$\int_\Omega u \partial^\alpha \varphi \, dx = (-1)^{|\alpha|} \int_\Omega g_\alpha \varphi \, dx \qquad \text{für alle } \varphi \in \mathcal{C}_c^\infty(\Omega).$$

Wir definieren dann $\partial^\alpha u := g_\alpha$ und für $0 \le j \le k$ setzen wir $D^j u = (\partial^\alpha u)_{|\alpha|=j}$. ♢

Bemerkung. Gilt $u \in \mathcal{C}^1(\overline{\Omega})$, dann ist u schwach differenzierbar und die schwache Ableitung und die klassische Ableitung stimmen üeberein. Dies folgt aus der partiellen Integration. ♢

Definition 4.12 (Sobolev-Raum)**.** Für $k \in \mathbb{N}_0$ und $p \in [1, \infty]$ definieren wir den Sobolev-Raum $W^{k,p}(\Omega)$ durch

$$W^{k,p}(\Omega) := \left\{ v \in L^p(\Omega) \;\middle|\; \begin{array}{l} \text{für alle Multiindices } \alpha \in \mathbb{N}_0^d \text{ mit } |\alpha| \le k \text{ existiert} \\ \text{die schwache Ableitung } \partial^\alpha v \text{ und } \partial^\alpha v \in L^p(\Omega) \end{array} \right\}.$$

Der Raum ist ausgestattet mit der Norm

$$\|v\|_{W^{k,p}(\Omega)} := \left(\sum_{|\alpha| \le k} \|\partial^\alpha v\|_{L^p(\Omega)}^p \right)^{\frac{1}{p}} \qquad \text{wenn } 1 \le p < \infty,$$

$$\|v\|_{W^{k,\infty}(\Omega)} := \max_{|\alpha| \le k} \|\partial^\alpha v\|_{L^\infty(\Omega)} \qquad \text{wenn } p = \infty.$$

Wenn $p = 2$, dann schreiben wir $H^k(\Omega) = W^{k,2}(\Omega)$. ♢

Proposition 4.13. *Für $k \in \mathbb{N}_0$ und $p \in [1, \infty]$ sind die Räume $W^{k,p}(\Omega)$ Banach-Räume. Für $p = 2$ ist $H^k(\Omega)$ ein Hilbert-Raum mit Skalarprodukt*

$$\langle u, v\rangle_{H^k(\Omega)} := \sum_{|\alpha|\leq k} \int_\Omega \partial^\alpha u \, \partial^\alpha v \, dx.$$

Beweis. Der Beweis dieser Aussage findet sich zum Beispiel in [Alt16]. □

Da wir Randbedingungen im Poisson-Problem fordern wollen, beschäftigen wir uns in dem nächsten Satz mit der Frage, ob Randwerte für Funktionen in $W^{k,p}(\Omega)$ existieren. Hier soll bemerkt werden, dass Randauswertungen für Funktionen in $L^p(\Omega)$ keinen Sinn machen, da der Rand eine Nullmenge ist.

Satz 4.14 (Spuroperator)**.** *Es sei $1 \leq p \leq \infty$. Dann existiert ein eindeutig definierter, beschränkter, linearer Operator $\gamma : W^{1,p}(\Omega) \to L^p(\partial\Omega)$ mit $\gamma(u) = u|_{\partial\Omega}$ für alle $u \in \mathcal{C}^\infty(\overline{\Omega}) \cap W^{1,p}(\Omega)$.*

Beschränkt heißt, dass es eine Konstante $0 < C_\gamma < \infty$ gibt mit

$$\|\gamma(u)\|_{L^p(\partial\Omega)} \leq C_\gamma \|u\|_{W^{1,p}(\Omega)}. \tag{4.5}$$

Wir werden im Folgenden $u|_{\partial\Omega}$ für $u \in W^{1,p}(\Omega)$ schreiben statt $\gamma(u)$.

Definition 4.15. Definiere

$$\begin{aligned} W^{1,p}_D(\Omega) &:= \{v \in W^{1,p}(\Omega) | v|_{\Gamma_D} = 0\}, \\ H^1_D(\Omega) &:= W^{1,2}_D(\Omega), \\ H^1_0(\Omega) &:= H^1_{\Gamma_D}(\Omega) \qquad \text{für } \Gamma_D = \partial\Omega. \end{aligned}$$ ♢

Die schwache Formulierung 4.10 ist mit der folgenden Definition von V nun vollständig,

$$V := H^1_D(\Omega). \tag{4.6}$$

Wir wollen den abstrakten Rahmen aus Abschnitt 4.1 benutzen, um Existenz und Eindeutigkeit von schwachen Lösungen zu zeigen.
Definiere $a : V \times V \to \mathbb{R}$ und $F : V \to \mathbb{R}$ durch

$$\begin{aligned} a(u, v) &:= \int_\Omega \nabla u \cdot \nabla v \, dx \qquad \text{für alle } u, v \in V, \\ F(v) &:= \int_\Omega f \, v \, dx + \int_{\Gamma_N} g \, v \, ds \qquad \text{für alle } v \in V. \end{aligned}$$

Bevor wir den Satz 4.2 von Lax-Milgram anwenden können, benötigen wir noch folgendes Resultat aus der Funktionalanalysis.

Satz 4.16 (Poincaré-Friedrichs-Ungleichung)**.** *Die Menge Γ_D habe ein positives $(d-1)$-dimensionales Maß. Dann existiert eine Konstante $0 < C_P < \infty$ derart, dass für alle $v \in H^1_D(\Omega)$ gilt*

$$\|v\|_{L^2(\Omega)} \leq C_P \|\nabla v\|_{L^2(\Omega)}.$$

Außerdem existiert eine Konstante C derart, dass für alle $v \in H^1(\Omega)$ mit $\int_\Omega v \, dx = 0$

$$\|v\|_{L^2(\Omega)} \leq C_P \|\nabla v\|_{L^2(\Omega)}$$

gilt.

Wir können nun die Existenz und Eindeutigkeit von schwachen Lösungen beweisen.

Satz 4.17 (Existenz und Eindeutigkeit von schwachen Lösungen)**.** *Der Raum V ist ein Hilbert-Raum, a ist eine koerzitive und beschränkte Bilinearform und $F \in V'$. Deshalb gibt es eine eindeutige Lösung $u \in V$ von* (4.4) *und es gilt*

$$\|u\|_{H^1(\Omega)} \leq (1 + C_P^2) \max\{1, C_\gamma\} \left(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma_N)} \right),$$

wobei C_P die Konstante aus Satz 4.16 ist und C_γ die Konstante aus (4.5)*.*

Beweis. Dies ist eine Übungsaufgabe. □

Céas Lemma sagt aus, dass für jeden Teilraum $V_h \subseteq V$ der Fehler der Galerkin-Approximation $u_h \in V_h$ beschränkt ist durch den reinen Approximationsfehler, also

$$\|u - u_h\|_{H^1(\Omega)} \leq (1 + C_P)^2 \inf_{w_h \in V_h} \|u - w_h\|_{H^1(\Omega)}.$$

Wegen der Poincaré-Ungleichung können wir auch $\|\nabla \bullet\|_{L^2(\Omega)}$ als Norm auf V betrachten und erhalten dann

$$\|\nabla(u - u_h)\|_{L^2(\Omega)} \leq \inf_{w_h \in V_h} \|\nabla(u - w_h)\|_{L^2(\Omega)}.$$

Wir werden im Folgenden geeignete Räume V_h definieren und, wenn die Lösung u glatt ist, Konvergenzraten beweisen. Dafür wird folgender Satz nützlich sein, der hier ohne Beweis angegeben wird.

Satz 4.18 (H^2-Regularität des Poisson-Problems)**.** *Es sei Ω konvex (oder Ω habe einen $\mathcal{C}^1$ Rand) und $\Gamma_D = \partial\Omega$. Dann existiert eine Konstante $C < \infty$, sodass für die Lösung $u \in H^1_0(\Omega)$ von* (4.4) *$u \in H^2(\Omega)$ gilt und*

$$\|D^2 u\|_{L^2(\Omega)} \leq C \|f\|_{L^2(\Omega)}.$$

Beweis. Siehe zum Beispiel [Gri11]. □

Wir werden außerdem die folgenden Sätze benötigen.

Satz 4.19 (Sobolevscher Einbettungssatz)**.** *Es sei $1 \leq p < \infty$ und $1 \leq q \leq p^\star$, wobei*

$$p^\star := \begin{cases} dp/(d-p) & \text{für } p < d, \\ 1 \leq q < \infty \text{beliebig} & \text{für } p = d, \\ \infty & \text{für } p > d. \end{cases} \tag{4.7}$$

Dann gilt für alle $u \in W^{1,p}(\Omega)$, dass $u \in L^q(\Omega)$ und es gibt eine Konstante $C < \infty$ mit

$$\|u\|_{L^q(\Omega)} \leq C \|u\|_{W^{1,p}(\Omega)},$$

d.h. die Einbettung $W^{1,p}(\Omega) \to L^q(\Omega)$ ist stetig.
Ist $k > d/p$, dann gilt für alle $u \in W^{k,p}(\Omega)$, dass $u \in \mathcal{C}(\overline{\Omega}) \cap L^\infty(\Omega)$ und

$$\|u\|_{L^\infty(\Omega)} \leq \|u\|_{W^{k,p}(\Omega)},$$

d.h. die Einbettung $W^{k,p}(\Omega) \to \mathcal{C}(\overline{\Omega}) \cap L^\infty(\Omega)$ ist stetig.

Bemerkung. Im Sobolevschen Einbettungssatz ist $u \in \mathcal{C}(\overline{\Omega})$ so zu verstehen, dass es eine Funktion $\widetilde{u}$ gibt mit $u = \widetilde{u}$ fast überall und $\widetilde{u} \in \mathcal{C}(\overline{\Omega})$, d.h. u kann auf einer Nullmenge abgeändert werden, derart, dass u dann stetig ist. Noch präziser kann man sagen, dass es einen Repräsentanten aus der Äquivalenzklasse von u gibt, der in $\mathcal{C}(\overline{\Omega})$ ist. ♢

Satz 4.20 (Rellich-Kandrachov)**.** *Es sei $1 \leq p < \infty$ und $1 \leq q < p^\star$ mit $p^\star$ aus* (4.7)*. Dann ist die Einbettung $W^{1,p}(\Omega) \to L^q(\Omega)$ kompakt, d.h. alle in $W^{1,p}(\Omega)$ beschränkten Folgen haben eine Teilfolge, die in $L^q(\Omega)$ konvergiert.*

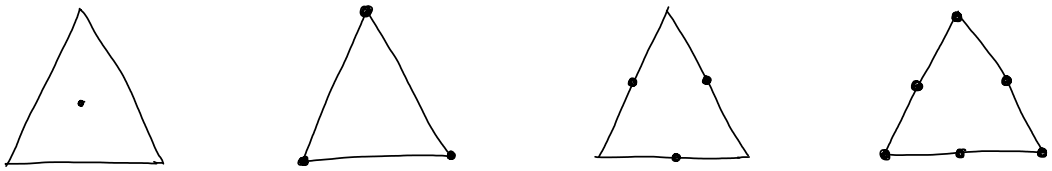

Abbildung 4.3: Schematische Darstellung der Finiten Elemente aus den Beispielen 4.23–4.26.

4.3 Allgemeine P_k-Finite-Elemente-Methoden

Definition 4.21. Für eine abgeschlossene Menge $T \subseteq \mathbb{R}^d$ und $k \in \mathbb{N}_0$ bezeichne

$$P_k(T) := \left\{ v \in \mathcal{C}(T) \,\middle|\, v(x) = \sum_{\alpha \in \mathbb{N}_0^d, |\alpha| \leq k} a_\alpha x^\alpha, a_\alpha \in \mathbb{R} \right\}$$

den Raum der Polynome vom (totalen) Grad $\leq k$ auf T. Es bezeichne

$$Q_k(T) := \left\{ v \in \mathcal{C}(T) \,\middle|\, v(x) = \sum_{\alpha \in \mathbb{N}_0^d, \max\{\alpha_j | j=1,\dots,d\} \leq k} a_\alpha x^\alpha, a_\alpha \in \mathbb{R} \right\}$$

den Raum der Polynome vom (partiellen) Grad $\leq k$ auf T. ♢

Ein Finites Element hat zusätzliche Struktur.

Definition 4.22 (Finites Element)**.** Ein Finites Element (nach Ciarlet) ist gegeben durch $(T, \mathcal{P}, \mathcal{K})$, wobei $T \subset \mathbb{R}^d$ eine abgeschlossene, beschränkte Menge ist, $\mathcal{P} \subseteq P_k(T)$ für ein $k \in \mathbb{N}_0$ (genannt *Ansatzfunktionen*) mit $\dim(\mathcal{P}) = R + 1$ und $\mathcal{K} = \{\chi_0, \dots, \chi_R\}$ eine Menge von Funktionalen $\chi_j : \mathcal{C}^\infty(T) \to \mathbb{R}$ (genannt *lokale Freiheitsgrade*). Das Tripel $(T, \mathcal{P}, \mathcal{K})$ sei derart, dass

(a) Wenn für ein $q \in \mathcal{P}$ gilt, dass $\chi(q) = 0$ für alle $\chi \in \mathcal{K}$, dann ist $q = 0$.

(b) Es existiert ein $m \geq 1$ mit $P_{m-1}(T) \subseteq \mathcal{P}$.

(c) Es existiert $p \in [1, \infty]$, sodass jedes $\chi \in \mathcal{K}$ fortgesetzt werden kann zu einem beschränkten, linearen Operator auf $W^{m,p}(T)$. ♢

Wir werdem später für Funktionen in $W^{m,p}(T)$ eine Abschätzung der für einen Interpolationsoperator der Form

$$|v - \mathcal{I}v|_{W^{k,p}(\Omega)} \leq Ch^{m-k} \|v\|_{W^{m,p}(\Omega)}$$

beweisen. Die Eigenschaft (c) garantiert uns, dass $\mathcal{I}v$ wohldefiniert ist, durch (b) bekommen wir die Rate h^{m-k}.

Beispiel 4.23. Es sei $T \subseteq \mathbb{R}^d$ eine abgeschlossene, beschränkte Menge. Dann ist das Tripel $(T, P_0(T), \chi_T)$ mit $\chi_T(v) = v(x_T)$ für ein beliebiges $x_T \in T$ ein Finites Element für $m = 1$ und $p > d$. Dies folgt aus dem Sobolevschen Einbettungssatz, Satz 4.19. Ist χ_T definiert durch $\chi_T(v) = \int_T v \, dx$, dann ist $(T, P_0(T), \chi_T)$ ein Finites Element für $m = 1$ und $p = 1$. ♢

Beispiel 4.24. Es seien $z_0, \dots, z_d \in \mathbb{R}^d$, $d = 1, 2$ oder 3 so, dass sie nicht auf einer Hyperebene liegen, und $T \subseteq \mathbb{R}^d$ sei das Simplex

$$T = \operatorname{conv}\{z_0, \dots, z_d\} \subseteq \mathbb{R}^d.$$

Definiere $\mathcal{P} := P_1(T)$ und $\mathcal{K} := \{\chi_0, \dots, \chi_d\}$ mit

$$\chi_j(v) = v(z_j) \qquad \text{für } j = 0, \dots, d \text{ und } v \in \mathcal{C}^\infty(T).$$

Gilt für $q \in P_1(T)$, dass $q(z_j) = 0$, dann muss schon $q = 0$ gelten. Für $d \leq 3$ impliziert der Sobolevsche Einbettungssatz, Satz 4.19, dass $W^{2,2}(T) \subseteq \mathcal{C}(T)$. Daher ist $(T, \mathcal{P}, \mathcal{K})$ ein Finites Element mit $m = 2$ und $p = 2$. ♢

Beispiel 4.25. Es seien $z_0, \dots, z_d \in \mathbb{R}^d$ und T wie oben. Es bezeichne $F_j := \operatorname{conv}\{z_k | k \neq j\}$ die Hyperebene, die z_j gegenüber liegt. Es sei außerdem $\mathcal{P} = P_1(T)$ und $\mathcal{K} = \{\chi_0, \dots, \chi_d\}$ mit

$$\chi_j(v) = v(\operatorname{mid}(F_j)) \qquad \text{für } j = 0, \dots, d \text{ und } v \in \mathcal{C}^\infty(T).$$

Dann ist $(T, \mathcal{P}, \mathcal{K})$ ein Finites Element mit $m = 2$ und $p = 2$.
Definiere χ_j nun durch

$$\chi_j(v) = \fint_{F_j} v \, ds := \frac{1}{|F_j|} \int_{F_j} v \, ds,$$

wobei $|F_j|$ den $(d-1)$-dimensionalen Flächeninhalt von F_j bezeichne. Dies ist nach dem Spursatz für alle $v \in W^{1,1}(T)$ ein beschränkter Operator. Also ist $(T, \mathcal{P}, \mathcal{K})$ ein Finites Element für $m = 1$ und $p = 2$ und auch für $m = 2$ und $p = 1$, da nach dem Sobolevschen Einbettungssatz $W^{2,1}(T)$ eingebettet ist in $W^{1,1}(T)$. ♢

Beispiel 4.26. Es sei $d = 2$ oder 3 und T wie oben. Es bezeichne $\mathcal{E}(T)$ die Menge der Kanten in T, d.h. für alle $E \in \mathcal{E}(T)$ existieren $j, k \in \{0, \dots, d\}$ mit $j \neq k$ und $E = \operatorname{conv}\{z_j, z_k\}$. Es sei $\mathcal{P} := P_2(T)$ und $\mathcal{K} = \{\chi_0, \dots, \chi_{k(d)}\}$ mit $k(2) = 6$ und $k(3) = 9$ und

$$\begin{aligned} \chi_j(v) &= v(z_j) && \text{für } j = 0, \dots, d, \\ \chi_{d+j}(v) &= v(\operatorname{mid}(E_j)) && \text{für } j = 1, \dots, k(d) - d, \end{aligned}$$

wobei $\mathcal{E} = \{E_1, \dots, E_{k(d)-d}\}$. Dann ist $(T, \mathcal{P}, \mathcal{K})$ ein Finites Element mit $m = 3$ und $p = 2$ (und auch für $m = 2$ und $p = 2$). ♢

Finite Elemente werden häufig durch Symbole beschrieben. Für die Beispiele 4.23–4.26 finden sich diese schematischen Darstellungen in Abbildung 4.3.
Sind die Funktionale in $\mathcal{K}$ über Punktauswertungen definiert, gilt die Eigenschaft (a) *nicht* automatisch. Beispielsweise erfüllt $(T, \mathcal{P}, \mathcal{K})$ mit einem Dreieck $T \subset \mathbb{R}^2$, $\mathcal{P} = P_2(T)$ und $\mathcal{K} = \{\chi_0, \dots, \chi_5\}$ mit Punktauswertungen an den zwei Gaußpunkten der Kanten E die Eigenschaft (a) *nicht* (siehe Übungsaufgaben).
Die Eigenschaft (a) lässt sich häufig am einfachsten über eine duale Basis zeigen. Diese duale Basis ist wie in folgender Proposition über die Freiheitsgrade definiert.

Proposition 4.27. *Es sei $(T, \mathcal{P}, \mathcal{K})$ ein Finites Element. Dann existiert eine eindeutige Basis $(\varphi_0, \dots, \varphi_R)$ von $\mathcal{P}$ mit*

$$\chi_j(\varphi_k) = \delta_{jk} \qquad \textit{für alle } j, k = 0, \dots, R = \dim(\mathcal{P}) - 1.$$

Bemerkung. Es gilt auch die Umkehrung: Wenn eine duale Basis existiert, dann gilt auch die Eigenschaft (a). ♢

Beweis von Proposition 4.27. Es sei $(q_0, \dots, q_R)$ eine beliebige Basis von $\mathcal{P}$. Definiere $A \in \mathbb{R}^{(R+1)\times(R+1)}$ durch

$$A_{jk} = \chi_j(q_k).$$

Für $y \in \mathbb{R}^{R+1}$ gilt

$$\begin{aligned} Ay = 0 \quad &\Leftrightarrow \quad \forall j \in \{0, \dots, R\} : \sum_{k=0}^{R} A_{jk} y_k = 0 \\ &\Leftrightarrow \quad \forall j \in \{0, \dots, R\} : \chi_j \left(\sum_{k=0}^{R} y_k q_k \right) = 0 \quad \Leftrightarrow \quad \sum_{k=0}^{R} y_k q_k = 0 \quad \Leftrightarrow \quad y = 0, \end{aligned}$$

wobei die dritte Äquivalenz aus Definition 4.22, (a), folgt, und die letzte Äquivalenz gilt, weil $(q_0, \dots, q_R)$ eine Basis ist. Also ist A regulär. Es sei $c_j \in \mathbb{R}^{R+1}$ die Lösung von $Ac_j = e_j$, wobei e_j den j-ten Einheitsvektor bezeichne. Definiere

$$\varphi_j := \sum_{\ell=0}^{R} c_{j\ell} q_\ell.$$

Dann gilt

$$\chi_j(\varphi_k) = \sum_{\ell=0}^{R} c_{k\ell} \underbrace{\chi_j(q_\ell)}_{=A_{j\ell}} = (Ac_k)_j = \delta_{jk}. \qquad \square$$

Mit Hilfe der barycentrischen Koordinaten kann eine duale Basis häufig explizit angegeben werden.

Definition 4.28 (barycentrische Koordinaten)**.** Zu einem Simplex $T = \operatorname{conv}\{z_0, \dots, z_d\}$ existieren die barycentrischen Koordinaten $\lambda_0, \dots, \lambda_d \in P_1(T)$ mit $\lambda_j(z_k) = \delta_{jk}$. ♢

Bemerkung. Die dualen Basisfunktionen zur P_1-FEM aus Beispiel 4.24 erfüllt $\varphi_{z_j}|_T = \lambda_j$. ♢

Bemerkung. Es gilt

$$x = \sum_{j=0}^{d} \lambda_j(x) z_j. \qquad ♢$$

Proposition 4.29. *Es sei $(T, \mathcal{P}, \mathcal{K})$ ein Finites Element und es sei $v \in W^{m,p}(T)$. Dann existiert ein eindeutiger Interpolant $\mathcal{I}_T v \in \mathcal{P}$ mit*

$$\chi(\mathcal{I}_T v) = \chi(v) \qquad \textit{für alle } \chi \in \mathcal{K}. \tag{4.8}$$

Es gilt

$$\mathcal{I}_T v = \sum_{j=0}^{R} \chi_j(v) \varphi_j. \tag{4.9}$$

Beweis. Mit $\mathcal{I}_T$ definiert in (4.9) folgt (4.8). Die Eindeutigkeit folgt aus (a) aus Definition 4.22. $\square$

Wir werden nun zum Bramble-Hilbert-Lemma kommen, aus dem wir später Konvergenzraten folgern werden. Es sei $T \subseteq \mathbb{R}^d$ und $\operatorname{int}(T)$ ein Lipschitz-Gebiet.

Lemma 4.30. *Es sei $v \in W^{m,p}(T)$ und es gelte $\partial^\alpha v = 0$ für alle $\alpha \in \mathbb{N}_0^d$ mit $|\alpha| = m$. Dann existiert ein Polynom $q \in P_{m-1}(T)$ mit $v = q$.*

Beweis. Da $\partial^\beta v = 0$ für alle $\beta \in \mathbb{N}_0^d$ mit $|\beta| \geq m$, gilt $v \in W^{k,p}(T)$ für alle $k \in \mathbb{N}$. Nach dem Sobolevschen Einbettungssatz folgt $v \in \mathcal{C}^\infty(T)$. Daraus folgt die Behauptung mit klassischen Argumenten. □

Lemma 4.31 (Projektion auf Polynome)**.** *Für alle $v \in W^{m,p}(T)$ existiert ein eindeutig definiertes Polynom $q \in P_{m-1}(T)$ mit*

$$\int_T \partial^\alpha (v - q)\, dx = 0 \qquad \text{für alle } \alpha \in \mathbb{N}_0^d \text{ mit } |\alpha| \leq m-1.$$

Beweis. Definiere $N := |\{\alpha \in \mathbb{N}_0^d \mid |\alpha| \leq m-1\}|$. Es gilt

$$P_{m-1}(T) = \left\{ \sum_{\alpha \in \mathbb{N}_0^d, |\alpha| \leq m-1} a_\alpha x^\alpha \;\middle|\; a_\alpha \in \mathbb{R} \right\}.$$

Eine Berechnung der Ableitung für $|\beta| \geq |\alpha|$ ergibt

$$\partial^\beta x^\alpha = \begin{cases} 0 & \text{wenn } \alpha_j < \beta_j \text{ für ein } j \in \{1, \dots, d\}, \\ \alpha! & \text{wenn } \alpha = \beta. \end{cases}$$

Daher ist die Abbildung

$$P_{m-1}(T) \to \mathbb{R}^N, \quad q \mapsto \left(\int_T \partial^\alpha q\, dx \right)_{\alpha \in \mathbb{N}_0^d, |\alpha| \leq m-1}$$

injektiv. Da die Dimensionen von $P_{m-1}(T)$ und $\mathbb{R}^N$ gleich sind, ist sie auch surjektiv. Daraus folgt die Behauptung. □

Lemma 4.32 (verallgemeinerte Poincaré-Ungleichung)**.** *Es existiert eine Konstante C_P' derart, dass gilt: Für alle $v \in W^{m,p}(T)$ mit $\int_T \partial^\alpha v\, dx = 0$ für alle $\alpha \in \mathbb{N}_0^d$ mit $|\alpha| \leq m-1$, gilt*

$$\|v\|_{W^{m,p}(T)} \leq C_P' |v|_{W^{m,p}(T)},$$

wobei

$$|v|_{W^{m,p}(T)} := \left(\sum_{|\alpha|=m} \|\partial^\alpha v\|_{L^p(T)}^p \right)^{1/p}.$$

Beweis. Wir nehmen an, dass die Aussage falsch ist. Dann existiert für alle $k \in \mathbb{N}$ ein $\hat{v}_k \in W^{m,p}(T)$ mit

$$\int_T \partial^\alpha \hat{v}_k\, dx = 0 \qquad \text{für alle } |\alpha| \leq m-1$$
$$\text{und} \qquad \|\hat{v}_k\|_{W^{m,p}(T)} > k |\hat{v}_k|_{W^{m,p}(T)}.$$

Für jedes $k \in \mathbb{N}$ setzen wir $v_k := \hat{v}_k / \|\hat{v}_k\|_{W^{m,p}(T)}$. Dann gilt

$$\|v_k\|_{W^{m,p}(T)} = 1 \qquad \text{und} \qquad |v_k|_{W^{m,p}(T)} < 1/k.$$

Die Folge $(v_k)_{k\in\mathbb{N}}$ ist also beschränkt in $W^{m,p}(T)$, also existiert nach Satz 4.20 ein Grenzwert $v \in W^{m-1,p}(T)$ und eine Teilfolge $(v_{k_j})_{j\in\mathbb{N}}$ mit $v_{k_j} \to v$ in $W^{m-1,p}(T)$ für $j \to \infty$. Da $|v_k|_{W^{m,p}(T)} \to 0$, ist $(v_{k_j})_{j\in\mathbb{N}}$ sogar eine Cauchy-Folge in $W^{m,p}(T)$ mit Grenzwert v. Dann gilt aber $|v|_{W^{m,p}(T)} = 0$ und aus Lemma 4.30 folgt, dass $v = q$ für ein Polynom $q \in P_{m-1}(T)$. Da aber

$$\int_T \partial^\alpha v \, dx = \lim_{j\to\infty} \int_T \partial^\alpha v_{k_j} \, dx = 0 \qquad \text{für alle } \alpha \in \mathbb{N}_0^d \text{ mit } |\alpha| \le m-1$$

gilt, folgt mit Lemma 4.31, dass $q = 0$. Also gilt $v = 0$, was ein Widerspruch zu

$$\|v\|_{W^{m,p}(T)} = \lim_{j\to\infty} \|v_{k_j}\|_{W^{m,p}(T)} = 1$$

ist. □

Satz 4.33 (Bramble-Hilbert-Lemma)**.** *Es sei $1 \le p < \infty$ und es sei ein $F : W^{m,p}(T) \to \mathbb{R}$ gegeben, das beschränkt und quasisublinear ist, d.h. es existieren $c_1, c_2 > 0$ mit*

$$|F(v)| \le c_1 \|v\|_{W^{m,p}(T)},$$
$$|F(v+w)| \le c_2 \left(|F(v)| + |F(w)|\right)$$

für alle $v, w \in W^{m,p}(T)$. Außerdem gelte $F(q) = 0$ für alle $q \in P_{m-1}(T)$. Dann gilt

$$|F(v)| \le C_P' c_1 c_2 \|D^m v\|_{L^p(T)} \qquad \text{für alle } v \in W^{m,p}(T).$$

Beweis. Es sei $v \in W^{m,p}(T)$. Es gilt für alle $q \in P_{m-1}(T)$, dass

$$|F(v)| = |F(v-q+q)| \le c_2 |F(v-q)| \le c_1 c_2 \|v-q\|_{W^{m,p}(T)}.$$

Es sei nun $q \in P_{m-1}(T)$ das Polynom, das nach Lemma 4.31 $\int_T \partial^\alpha (v-q)\, dx = 0$ für alle $\alpha \in \mathbb{N}_0^d$ mit $|\alpha| \le m-1$ erfüllt. Die verallgemeinerte Poincaré-Ungleichung aus Lemma 4.32 beweist dann

$$\|v-q\|_{W^{m,p}(T)} \le C_P' \|D^m(v-q)\|_{L^p(T)} = C_P' \|D^m v\|_{L^p(T)}.$$ □

Korollar 4.34 (Stabilität der Interpolation)**.** *Es sei $(T, \mathcal{P}_T, \mathcal{K}_T)$ ein Finites Element mit $P_{m-1}(T) \subseteq \mathcal{P}_T$ und es sei $|\bullet|_S$ eine Seminorm auf $W^{m,p}(T)$ mit*

$$|v|_S \le C \|v\|_{W^{m,p}(T)} \qquad \text{für alle } v \in W^{m,p}(T)$$

für eine Konstante $C < \infty$. Dann existiert $c_{\mathrm{IS}} < \infty$ mit

$$|v - \mathcal{I}_T v|_S \le c_{\mathrm{IS}} \|D^m v\|_{L^p(T)} \qquad \text{für alle } v \in W^{m,p}(T).$$

Beweis. Es sei $F(v) = |v - \mathcal{I}_T v|_S$. Dann ist F sublinear. Es bezeichne φ_j die duale Basis zu $(\chi_j)_{j=0,\dots,R}$. Die Definition von $\mathcal{I}_T$ in (4.9) und die Definition des Finiten Elements in Definition 4.22 zeigen

$$|\mathcal{I}_T v|_S = \left| \sum_{j=0}^R \chi_j(v) \varphi_j \right|_S \le \sum_{j=0}^R \underbrace{|\chi_j(v)|}_{\le \tilde{C}\|v\|_{W^{m,p}(T)}} |\varphi_j|_S.$$

Es folgt mit einer Dreiecksungleichung

$$|F(v)| \leq |v|_S + |\mathcal{I}_T v|_S \leq \left(C + \tilde{C} R \max_{j=0,\dots,R} |\varphi_j|_S \right) \|v\|_{W^{m,p}(T)}.$$

Also ist F beschränkt. Da $F(q) = 0$ für alle $q \in \mathcal{P}_T$, sind die Bedingungen des Bramble-Hilbert-Lemma erfüllt, woraus die Behauptung folgt. □

Als nächstes widmen wir uns der Frage, wie die Konstante c_{IS} von den Parametern des Gebiets abhängt.

Lemma 4.35. *Es sei $\widehat{T} := \mathrm{conv}\{0, e_1, \dots, e_d\} \subseteq \mathbb{R}^d$ das Referenz-Simplex und es sei $T = \mathrm{conv}\{z_0, \dots, z_d\} \subseteq \mathbb{R}^d$ ein nicht-degeneriertes Simplex. Dann existiert ein eindeutiger affiner Diffeomorphismus $\Phi_T : \widehat{T} \to T$ mit $\Phi_T(0) = z_0$ und $\Phi_T(e_j) = z_j$ für alle $j = 1, \dots, d$. Dann existieren Konstanten C_1 und C_2 derart, dass für jedes $v \in W^{m,p}(T)$ und $\widehat{v} = v \circ \Phi_T \in W^{m,p}(\widehat{T})$ gilt*

$$|v|_{W^{k,p}(T)} \leq C_1 \, \rho_T^{-k} \, |\det B|^{1/p} \, |\widehat{v}|_{W^{k,p}(\widehat{T})},$$
$$|\widehat{v}|_{W^{k,p}(\widehat{T})} \leq C_2 \, h_T^k \, |\det B|^{-1/p} \, |v|_{W^{k,p}(T)},$$

wobei $B = D\Phi_T$ und $k \leq m$.

Beweis. Definiere $B \in \mathbb{R}^{d \times d}$ und $b \in \mathbb{R}^d$ durch

$$B = \begin{pmatrix} z_1 - z_0 & z_2 - z_0 & \dots & z_d - z_0 \end{pmatrix} \quad \text{und} \quad b = z_0.$$

Es gilt $Be_j = z_j - z_0$. Es gilt außerdem, dass T nicht degeneriert ist genau dann, wenn $z_1 - z_0, \dots, z_d - z_0$ linear unabhängig sind. Also ist B regulär. Definiere $\Phi_T : \widehat{T} \to \mathbb{R}^d$ durch $\Phi_T(\widehat{x}) = B\widehat{x} + b$. Dann gilt $\Phi_T(0) = z_0$ und $\Phi_T(e_j) = z_j$. Daher gilt auch $\Phi_T(\widehat{T}) = T$ und es gilt $\Phi_T^{-1} : T \to \widehat{T}$, $\Phi_T^{-1}(x) = B^{-1}x - B^{-1}b$. Also ist Φ_T ein Diffeomorphismus mit $D\Phi_T = B$. Wir wollen die Aussage der Proposition aus dem Transformationssatz folgern. Dafür sei $\alpha \in \mathbb{N}_0^d$ mit $|\alpha| = k$. Dann gilt

$$\int_T |\partial^\alpha w|^p \, dx = \int_{\widehat{T}} |(\partial^\alpha w) \circ \Phi_T|^p \underbrace{|\det D\Phi_T|}_{=|\det B|} \, d\widehat{x}.$$

Die Kettenregel beweist

$$\partial_j w = \partial_j(\widehat{w} \circ \Phi_T^{-1}) = \sum_{k=1}^d (\partial_k \widehat{w}) \circ \Phi_T^{-1} \underbrace{\partial_j (\Phi_T^{-1})_k}_{=(B^{-1})_{kj}} = (B^{-\top}((\widehat{D}\widehat{w}) \circ \Phi_T^{-1}))_j.$$

Da B konstant ist, können wir wiederum die Kettenregel anwenden und erhalten

$$\partial_{j_1}\partial_{j_2} w = \sum_{k_1=1}^d \sum_{k_2=1}^d (\partial_{k_1}\partial_{k_2}\widehat{w}) \circ \Phi_T^{-1} (B^{-1})_{k_2 j_2} (B^{-1})_{k_1 j_1} = \left(B^{-\top}\left(B^{-\top}((\widehat{D}\widehat{w}) \circ \Phi_T^{-1})\right)_{j_2}\right)_{j_1}.$$

Führen wir dieses Argument fort, erhalten wir mit einer Konstanten $\widetilde{C}$ aus der Normäquivalenz von $|\bullet|$ und der Maximumsnorm

$$|\partial^\alpha w| \leq \widetilde{C} \, \|B^{-\top}\|^{|\alpha|} \max_{|\beta|=|\alpha|} |\widehat{\partial}^\beta \widehat{w}|,$$

wobei $\|A\| := \sup_{x\in\mathbb{R}^d\setminus\{0\}} \frac{|Ax|}{|x|}$. Also folgt

$$\int_T |\partial^\alpha w|^p \,dx \le \widetilde{C}^p \|B^{-\top}\|^{p|\alpha|} \,|\det B| \underbrace{\max_{|\beta|=|\alpha|} \int_{\widehat{T}} |\widehat{\partial}^\beta \widehat{w}|^p \,d\widehat{x}}_{\le |\widehat{w}|^p_{W^{k,p}(\widehat{T})}}.$$

Es sei $z \in \mathbb{R}^d$ mit $|z| = \rho_T$ beliebig. Dann existieren $\xi, \eta \in T$ mit $z = \xi - \eta$, also $B^{-1}z = B^{-1}\xi - B^{-1}\eta = \Phi_T^{-1}(\xi) - \Phi_T^{-1}(\eta)$. Daraus folgt

$$\|B^{-1}\| = \sup_{z\in R^d, |z|=\rho_T} \frac{|B^{-1}z|}{\rho_T} \le \frac{1}{\rho_T} \sup_{\xi,\eta\in T} |\underbrace{\Phi_T^{-1}(\xi)}_{\in\widehat{T}} - \underbrace{\Phi_T^{-1}(\eta)}_{\in\widehat{T}}| = \frac{h_{\widehat{T}}}{\rho_T}.$$

Es gilt für alle $x \in \mathbb{R}^d$

$$|x| = \sup_{y\in\mathbb{R}^d\setminus\{0\}} \frac{x\cdot y}{|y|}$$

und deshalb

$$\begin{aligned}\|B^{-\top}\| &= \sup_{y\in\mathbb{R}^d\setminus\{0\}} \frac{|B^{-\top}y|}{|y|} = \sup_{y\in\mathbb{R}^d\setminus\{0\}} \frac{|y^\top B^{-1}|}{|y|} = \sup_{y\in\mathbb{R}^d\setminus\{0\}} \sup_{x\in\mathbb{R}^d\setminus\{0\}} \frac{|y^\top B^{-1}x|}{|x|\,|y|} \\ &= \sup_{x\in\mathbb{R}^d\setminus\{0\}} \frac{|B^{-1}x|}{|x|} = \|B^{-1}\| \le \frac{h_{\widehat{T}}}{\rho_T}.\end{aligned}$$

Insgesamt folgt

$$\int_T |\partial^\alpha w|^p \,dx \le \widetilde{C}^p \left(\frac{h_{\widehat{T}}}{\rho_T}\right)^{p|\alpha|} |\det B|\, |\widehat{w}|^p_{W^{k,p}(\widehat{T})}.$$

Da $h_{\widehat{T}} \le 2$, beweist dies die erste Aussage. Die zweite folgt durch die Vertauschung von T und $\widehat{T}$ und da $\rho_{\widehat{T}} \ge c$ für eine Konstante $c > 0$. □

Satz 4.36 (Interpolationsfehler)**.** *Es sei $(\widehat{T}, \widehat{\mathcal{P}}, \widehat{\mathcal{K}})$ ein Referenz-Finites-Element mit $P_{m-1}(\widehat{T}) \subseteq \widehat{\mathcal{P}}$ und $\widehat{T} \subseteq \mathbb{R}^d$. Es sei $(T, \mathcal{P}, \mathcal{K})$ ein Finites Element, das durch eine affine Transformation $\Phi_T : \widehat{T} \to T$ erzeugt wird, d.h. es gilt*

$$\begin{aligned}&T = \Phi_T(\widehat{T}), \qquad \mathcal{P} = \{\widehat{q}\circ\Phi_T^{-1} \mid \widehat{q} \in \widehat{\mathcal{P}}\}, \\ &\mathcal{K} = \{\chi \mid \chi(v) = \widehat{\chi}(v\circ\Phi_T^{-1}) \text{ für alle } v \in \mathcal{C}^\infty(T),\ \widehat{\chi} \in \widehat{\mathcal{K}}\}.\end{aligned}$$

Dann gilt für jedes $v \in W^{m,p}(T)$, dass

$$|v - \mathcal{I}_T v|_{W^{k,p}(T)} \le c_{\mathcal{I}}\, h_T^m\, \rho_T^{-k}\, |v|_{W^{m,p}(T)}$$

mit einer Konstanten $c_{\mathcal{I}} = c_{\mathcal{I}}(d, m, \widehat{T})$ für alle $0 \le k \le m$.

Beweis. Wir benutzen Folgerung 4.34 auf dem Referenz-Element für $\widehat{v} = v\circ\Phi_T \in W^{k,m}(\widehat{T})$,

$$\begin{aligned}|v - \mathcal{I}_T v|_{W^{k,p}(T)} &\overset{\text{Lemma 4.35}}{\le} C_1\, \rho_T^{-k}\, |\det B|^{1/p}\, |\widehat{v} - \mathcal{I}_{\widehat{T}}\widehat{v}|_{W^{k,p}(\widehat{T})} \\ &\overset{\text{Folgerung 4.34}}{\le} C_1 c_{\mathrm{IS}}\, \rho_T^{-k}\, |\det B|^{1/p}\, |\widehat{v}|_{W^{m,p}(\widehat{T})} \\ &\overset{\text{Lemma 4.35}}{\le} C_1 C_2 c_{\mathrm{IS}}\, h_T^m\, \rho_T^{-k}\, |\det B|^{1/p}\, |\det B|^{-1/p}\, |v|_{W^{m,p}(T)}.\end{aligned}$$ □

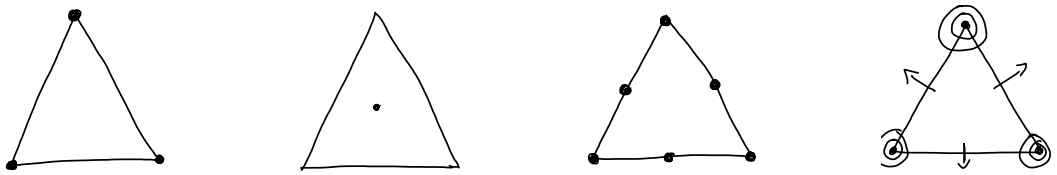

Abbildung 4.4: Globale Finite Elemente aus den Beispielen 4.39–4.43.

Als nächstes setzen wir Finite Elemente zusammen, um globale Approximationen zu definieren. Dafür betrachten wir Finite Elemente auf Triangulierungen, siehe Definition 3.2.

Definition 4.37 (affine Familie von Finiten Elementen)**.** Es sei $\mathcal{T}$ eine reguläre Triangulierung von Ω. Eine affine Familie (von Finiten Elementen) ist eine Familie $(T, \mathcal{P}_T, \mathcal{K}_T)_{T\in\mathcal{T}}$ derart, dass es ein Referenz-Finites-Element $(\widehat{T}, \widehat{\mathcal{P}}, \widehat{\mathcal{K}})$ gibt und dass jedes Finite Element $(T, \mathcal{P}_T, \mathcal{K}_T)$ durch eine affine Transformation $\Phi_T : \widehat{T} \to T$ erzeugt ist (siehe auch Satz 4.36). ♢

Definition 4.38 (globaler Interpolant, globale Freiheitsgrade)**.** Es sei $\mathcal{T}$ eine Triangulierung von Ω und $(T, \mathcal{P}_T, \mathcal{K}_T)_{T\in\mathcal{T}}$ eine affine Familie. Der globale Interpolant $\mathcal{I}_\mathcal{T} : W^{m,p}(\Omega) \to L^\infty(\Omega)$ ist definiert durch

$$(\mathcal{I}_\mathcal{T} v)|_T = \mathcal{I}_T(v|_T) \qquad \text{für alle } T \in \mathcal{T}.$$

Die affine Familie wird ein $\mathcal{C}^r$-Element genannt, wenn $\mathcal{I}_\mathcal{T} v \in \mathcal{C}^r(\overline{\Omega})$ gilt für alle $v \in \mathcal{C}^r(\overline{\Omega}) \cap W^{m,p}(\Omega)$.
Die globalen Freiheitsgrade sind alle Funktionale $\chi : W^{m,p}(\Omega) \to \mathbb{R}$, für die ein $T \in \mathcal{T}$ existiert und ein $\chi_T \in \mathcal{K}_T$ mit $\chi(v) = \chi_T(v|_T)$ für alle $v \in W^{m,p}(\Omega)$. ♢

Beispiel 4.39. Es sei $\mathcal{T}$ eine reguläre Triangulierung und für jedes $T \in \mathcal{T}$ sei $(T, P_1(T), \mathcal{K}_T)$ das Finite Element, für das $\mathcal{K}_T$ die Knotenauswertungen sind (siehe Beispiel 4.24). Dann ist $\mathcal{I}_\mathcal{T} v$ die eindeutige Funktion, die auf jedem Dreieck affin ist und in den Ecken der Dreiecke gleich v ist. Dann ist $\mathcal{I}_\mathcal{T} v$ stetig, also ist diese affine Familie ein $\mathcal{C}^0$-Element. Die globalen Freiheitsgrade sind

$$\{\chi_p : \mathcal{C}^\infty(\overline{\Omega}) \to \mathbb{R} \mid p \in \mathcal{N},\ \chi_p(v) = v(p)\}.$$

Die Anzahl der globalen Freiheitsgrade ist $\#\mathcal{N}$. ♢

Beispiel 4.40. Es sei $\mathcal{T}$ wie oben und für jedes $T \in \mathcal{T}$ sei $(T, P_1(T), \mathcal{K}_T)$ das Finite Element, für das $\mathcal{K}_T$ die Auswertung am Mittelpunkt der Flächen sind (siehe Beispiel 4.25). Dann ist $\mathcal{I}_\mathcal{T} v$ die eindeutige Funktion, die auf jedem Dreieck affin ist und in den Kantenmittelpunkten gleich v ist. Dies ist im Allgemeinen keine global stetige Funktion, also ist dies kein $\mathcal{C}^0$-Element. Die globalen Freiheitsgrade sind

$$\{\chi_F : \mathcal{C}^\infty(\overline{\Omega}) \to \mathbb{R} \mid F \in \mathcal{F},\ \chi_F(v) = v(\mathrm{mid}(F))\}.$$

Die Anzahl der globalen Freiheitsgrade ist $\#\mathcal{F}$. ♢

Beispiel 4.41. Für das P_0-Element mit $\chi_T(v) = \fint_T v\,dx$ für alle $v \in \mathcal{C}^\infty(\Omega)$ ist $\mathcal{I}_\mathcal{T} v$ eine stückweise konstante Funktion. Dies Finite Element ist kein $\mathcal{C}^0$-Element. Die Anzahl der globalen Freiheitsgrade ist $\#\mathcal{T}$. ♢

Beispiel 4.42. Es sei $(T, P_2(T), \mathcal{K}_T)$ für alle $T \in \mathcal{T}$ das P_2 Finite Element aus Beispiel 4.26. Dann ist $\mathcal{I}_\mathcal{T} v$ die eindeutige Funktion, die auf jedem $T \in \mathcal{T}$ quadratisch ist und $(\mathcal{I}_\mathcal{T} v)(p) = v(p)$ erfüllt, wenn $p \in \mathcal{N}$ oder wenn $p = \mathrm{mid}(E)$ für ein $E \in \mathcal{E}$. Dies Finite Element ist ein $\mathcal{C}^0$ Finites Element, aber kein $\mathcal{C}^1$ Finites Element (siehe Übungsaufgabe). ♢

Beispiel 4.43. Das $\mathcal{C}^1$ Finite Element mit dem niedrigsten Polynomgrad ist das Finite Element von Argyris mit $\mathcal{P} = P_5(T)$ mit 21 lokalen Freiheitsgraden abgebildet in Abbildung 4.4. Dabei stehen Punkte für Auswertungen der Funktion, der erste Kreis für Auswertungen des Gradienten, der zweite Kreis für Auswertungen der Hesse Matrix und die Pfeile für die Auswertung der Ableitung in Normalen-Richtung. ♢

Siehe Abbildung 4.4 für Abbildungen zu den oben besprochenen Finiten Elementen.

Proposition 4.44 (globale Interpolationsabschätzung)**.** *Es sei $(T, \mathcal{P}_T, \mathcal{K}_T)_{T\in\mathcal{T}}$ eine affine Familie, die ein $\mathcal{C}^{\ell-1}$ Finites Element ist. Für jedes $v \in W^{m,p}(\Omega)$ und $0 \le k \le \ell$ gilt*

$$|v - \mathcal{I}_\mathcal{T} v|_{W^{k,p}(\Omega)} \le C \max_{T\in\mathcal{T}} h_T^m \rho_T^{-k} |v|_{W^{m,p}(\Omega)}.$$

Beweis. Die Abschätzung folgt aus Satz 4.36, indem über alle $T \in \mathcal{T}$ summiert wird. □

Korollar 4.45. *Für die affine Familie des P_1-Finiten-Elements existieren auf formregulären Triangulierungen Konstanten C_1 und C_2 derart, dass für alle $v \in H^2(\Omega)$*

$$\|v - \mathcal{I}_\mathcal{T} v\|_{L^2(\Omega)} \le C_1 \left(\max_{T\in\mathcal{T}} h_T^2\right) |v|_{H^2(\Omega)}$$
$$\text{und} \quad \|\nabla(v - \mathcal{I}_\mathcal{T} v)\|_{L^2(\Omega)} \le C_2 \left(\max_{T\in\mathcal{T}} h_T\right) |v|_{H^2(\Omega)}.$$

Korollar 4.46. *Für die affine Familie des P_2-Finiten-Elements gilt auf formregulären Triangulierungen $\mathcal{T}$, dass Konstanten C_1, C_2, C_3 und C_4 existieren so, dass für alle $v \in H^3(\Omega)$*

$$\|v - \mathcal{I}_\mathcal{T} v\|_{L^2(\Omega)} \le C_1 \left(\max_{T\in\mathcal{T}} h_T^3\right) |v|_{H^3(\Omega)}$$
$$\text{und} \quad \|\nabla(v - \mathcal{I}_\mathcal{T} v)\|_{L^2(\Omega)} \le C_2 \left(\max_{T\in\mathcal{T}} h_T^2\right) |v|_{H^3(\Omega)}$$

und für alle $v \in H^2(\Omega)$

$$\|v - \mathcal{I}_\mathcal{T} v\|_{L^2(\Omega)} \le C_1 \left(\max_{T\in\mathcal{T}} h_T^2\right) |v|_{H^2(\Omega)}$$
$$\text{und} \quad \|\nabla(v - \mathcal{I}_\mathcal{T} v)\|_{L^2(\Omega)} \le C_2 \left(\max_{T\in\mathcal{T}} h_T\right) |v|_{H^2(\Omega)}$$

gilt.

Bemerkung. *Achtung:* Für die P_1 und die P_2-FEM gilt *nicht*, dass

$$\|v - \mathcal{I}_\mathcal{T} v\|_{L^2(\Omega)}$$

durch

$$\left(\max_{T\in\mathcal{T}} h_T\right) |v|_{H^1(\Omega)}$$

beschränkt ist, da $\mathcal{I}_\mathcal{T}$ für allgemeine $v \in H^1(\Omega)$ nicht definiert ist! ♢

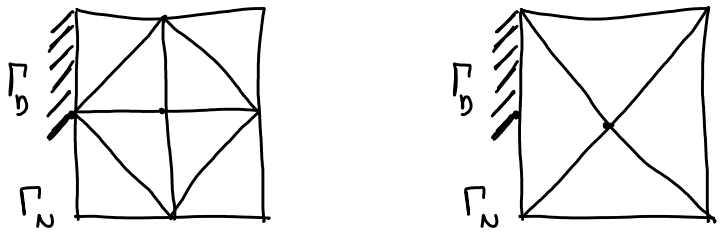

Abbildung 4.5: Zulässige und nicht zulässige Konfiguration von Kanten auf dem Rand.

4.4 Die P_k-FEM für das Poisson-Problem

Wir betrachten nun wieder das Poisson-Problem

$$-\Delta u = f \quad \text{in } \Omega, \qquad u|_{\Gamma_D} = 0, \qquad (\nabla u \cdot n)|_{\Gamma_N} = g,$$

wobei $\Omega \subseteq \mathbb{R}^d$ ein beschränktes Lipschitz-Gebiet mit polygonalem Rand $\partial\Omega = \Gamma_D \cup \Gamma_N$ ist mit $\Gamma_D \cap \Gamma_N = \emptyset$ und Γ_D abgeschlossen mit positivem Oberflächenmaß ist. Nach Definition 4.10 erfüllt die (schwache) Lösung $u \in H^1_D(\Omega)$

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f\, v \, dx + \int_{\Gamma_N} g\, v \, ds \qquad \text{für alle } v \in H^1_D(\Omega).$$

Es sei $\mathcal{T}$ eine formreguläre Triangulierung von Ω derart, dass $\Gamma_D = \bigcup_{F\in\mathcal{F},\, F\subseteq\Gamma_D} F$, siehe auch Abbildung 4.5.
Wir werden nun die Finite-Elemente-Methode (FEM) zum P_k-Finiten-Element, insbesondere die P_1-FEM aus den Beispielen 4.24 und 4.39 näher betrachten. Wir werden ab jetzt

$$\|\bullet\| := \|\bullet\|_\Omega := \|\bullet\|_{L^2(\Omega)}$$

abkürzen und wir werden

$$A \lesssim B$$

schreiben, wenn eine Konstante $C < \infty$ existiert, die nicht von der Gitterweite $(h_T)_{T\in\mathcal{T}}$ abhängt, für die

$$A \le C\, B$$

gilt. Die Konstante C darf von der Konstante aus der Formregularität aus Definition 3.7 abhängen. Außerdem verwenden wir

$$A \approx B \Leftrightarrow A \lesssim B \lesssim A.$$

Definition 4.47. Für eine Triangulierung $\mathcal{T}$ von Ω ist der konforme P_k-Finite-Elemente-Raum definiert als

$$S^k(\mathcal{T}) := \{v_h \in \mathcal{C}(\overline{\Omega}) \mid v_h|_T \in P_k(T) \text{ für alle } T \in \mathcal{T}\}$$

und mit Randbedingungen als

$$S^k_D(\mathcal{T}) := \{v_h \in S^k(\mathcal{T}) \mid v_h|_{\Gamma_D} = 0\},$$
$$S^k_0(\mathcal{T}) := S^k_D(\mathcal{T}), \quad \text{wenn } \Gamma_D = \partial\Omega. \qquad \diamondsuit$$

Definition 4.48. Die P_k-Finite-Elemente-Approximation des Poisson-Problems $u_h \in S^k_D(\mathcal{T})$ ist gegeben durch

$$\int_\Omega \nabla u_h \cdot \nabla v_h \, dx = \int_\Omega f\, v_h \, dx + \int_{\Gamma_N} g\, v_h \, ds \qquad \text{für alle } v_h \in S^k_D(\mathcal{T}). \qquad \diamondsuit$$

Korollar 4.49. *Wenn $u \in H^m(\Omega)$ für ein $2 \le m \le k+1$, dann gilt*

$$\|\nabla(u-u_h)\| \le \inf_{v_h \in S^k_D(\mathcal{T})} \|\nabla(u-v_h)\| \le \|\nabla(u-\mathcal{I}_\mathcal{T} u)\| \lesssim h^{m-1}\|D^m u\|.$$

Beweis. Die erste Ungleichung folgt aus Céas Lemma, da $u_h \in S^k_D(\mathcal{T})$ eine Galerkin-Approximation von u ist, die zweite, da $\mathcal{I}_\mathcal{T} u \in S^k_D(\mathcal{T})$, und die dritte aus den Interpolationsabschätzungen aus Proposition 4.44. □

Bemerkung. Wegen der Poincaré-Ungleichung ist $\|\nabla \bullet\|$ eine Norm auf $H^1_D(\Omega)$. $\diamondsuit$

Als nächstes wollen wir eine Abschätzung in der L^2-Norm herleiten. Aus der Poincaré-Ungleichung folgt direkt

$$\|u-u_h\| \lesssim \|\nabla(u-u_h)\| \lesssim h^{m-1}\|D^m u\|.$$

Aus den Interpolationsabschätzungen wissen wir aber, dass $\|u-\mathcal{I}_\mathcal{T} u\| \lesssim h^m \|D^m u\|$, also wäre die Konvergenzrate h^m prinzipiell in $L^2(\Omega)$ möglich. Um zu zeigen, dass diese Konvergenzrate tatsächlich angenommen wird, müssen wir die folgende Annahme machen.

Annahme 4.50 (H^2-Regularität des Poisson-Problem)**.** Für alle $q \in L^2(\Omega)$ gelte für die Lösung $w \in H^1_D(\Omega)$ des Poisson-Problems

$$\int_\Omega \nabla w \cdot \nabla v \, dx = \int_\Omega q\, v \, dx \qquad \text{für alle } v \in H^1_D(\Omega),$$

dass $w \in H^2(\Omega)$ und

$$\|D^2 w\| \lesssim \|q\|. \qquad \diamondsuit$$

Bemerkung. Satz 4.18 sagt, dass, wenn Ω konvex ist und $\Gamma_D = \partial\Omega$, das Poisson-Problem H^2 regulär ist. $\diamondsuit$

Satz 4.51 (Aubin-Nitsche-Lemma)**.** *Ist das Poisson-Problem H^2 regulär im Sinne von Annahme 4.50, dann gilt*

$$\|u-u_h\| \lesssim h\|\nabla(u-u_h)\|.$$

Gilt zusätzlich $u \in H^m(\Omega)$ für ein $2 \le m \le k+1$, dann gilt

$$\|u-u_h\| \lesssim h^m \|D^m u\|.$$

Beweis. Es sei $e := u - u_h$ und $z \in H^1_D(\Omega)$ die Lösung zu

$$\int_\Omega \nabla z \cdot \nabla v \, dx = \int_\Omega e\, v \, dx \qquad \text{für alle } v \in H^1_D(\Omega).$$

Dann gilt

$$\|e\|^2 = \int_\Omega e\,(u-u_h)\, dx = \int_\Omega \nabla z \cdot \nabla(u-u_h)\, dx.$$

Setze $z_h = \mathcal{I}_{\mathcal{T}} z \in S^k_D(\mathcal{T})$. Dann folgt aus der Symmetrie des Poisson-Problems und der Galerkin-Orthogonalität, dass

$$\int_\Omega \nabla z \cdot \nabla(u - u_h)\,dx = \int_\Omega \nabla(z - z_h) \cdot \nabla(u - u_h)\,dx \leq \|\nabla(z - z_h)\|\,\|\nabla e\|.$$

Nach der Annahme, dass das Poisson-Problem H^2-regulär ist, folgt aus den Interpolationsabschätzungen, dass

$$\|\nabla(z - z_h)\| \lesssim h\|D^2 z\| \lesssim h\|e\|.$$

Setzen wir diese Ungleichung oben ein, erhalten wir

$$\|e\| \lesssim h\|\nabla e\|.$$

Dies ist die erste Abschätzung aus Satz 4.51. Die zweite folgt dann aus Korollar 4.49. □

Man kann sich überlegen, dass der H^1- und der L^2-Fehler im Allgemeinen auch nicht besser konvergieren können als mit Rate h^k beziehungsweise h^{k+1}.
Wir wenden uns nun der Implementation der P_1-FEM zu. Für $k \geq 1$ definiert die duale Basis auch eine globale. Eine Basis von $S^1(\mathcal{T})$ ist gegeben durch die Hutfunktionen $(\varphi_z)_{z\in\mathcal{N}}$, wobei $\varphi_z \in S^1(\mathcal{T})$ definiert ist durch $\varphi_z(y) = \delta_{zy}$ für alle $z, y \in \mathcal{N}$. Für $S^1_D(\mathcal{T})$ ist $(\varphi_z)_{z\in\mathcal{N},\, z\notin\Gamma_D}$ eine Basis. Dafür sei

$$u_h = \sum_{y\in\mathcal{N}\setminus\Gamma_D} U_y \varphi_y$$

mit $U_y \in \mathbb{R}$ und $(\varphi_y)_{y\in\mathcal{N}\setminus\Gamma_D}$ die nodale Basis. Dann ist nach Proposition 4.6 U die Lösung zu

$$A^\top U = b,$$

wobei die Steifigkeitsmatrix und die rechte Seite gegeben sind durch

$$A_{jk} = \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_k\,dx,$$
$$b_j = \int_\Omega f\varphi_j\,dx.$$

Um A und b zu berechnen, zerlegen wir Ω in die Elemente $T \in \mathcal{T}$

$$A_{jk} = \sum_{T\in\mathcal{T}} \int_T \nabla\varphi_j \cdot \nabla\varphi_k\,dx,$$
$$b_j = \sum_{T\in\mathcal{T}} \int_T f\,\varphi_j\,dx.$$

Der Eintrag b_j kann zum Beispiel durch eine Mittelpunktsquadratur approximiert werden

$$b_j \approx \sum_{T\in\mathcal{T}} |T|\; f(\mathrm{mid}(T)) \underbrace{\varphi_j(\mathrm{mid}(T))}_{=1/(d+1)} = \sum_{T\in\mathcal{T}} \frac{|T|}{d+1} f(\mathrm{mid}(T)),$$

wobei $|T|$ das Maß von T bezeichne. Um A_{jk} zu berechnen, benutzen wir das folgende Lemma.

Lemma 4.52. *Es sei $T = \operatorname{conv}\{z_0, \dots, z_d\}$ ein Simplex mit Ecken $z_0, \dots, z_d \in \mathbb{R}^d$ und es sei*

$$X_T = \begin{pmatrix} 1 & 1 & \dots & 1 \\ z_0 & z_1 & \dots & z_d \end{pmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}.$$

Dann gilt, dass das Volumen von T gegeben ist durch

$$|T| = \frac{1}{d!} \det X_T$$

$$\text{und} \qquad \underbrace{X_T \begin{pmatrix} \nabla\varphi_{z_0}|_T \\ \nabla\varphi_{z_1}|_T \\ \vdots \\ \nabla\varphi_{z_d}|_T \end{pmatrix}}_{\in\mathbb{R}^{(d+1)\times d}} = \begin{pmatrix} 0 & \dots & 0 \\ 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}. \tag{4.10}$$

Beweis. Der Übersichtlichkeit halber wird hier $|T| = (1/d!) \det X_T$ nur für $d = 2$ gezeigt. Mit der Entwicklung nach der ersten Zeile gilt

$$\begin{aligned} \det X_T &= \det \begin{pmatrix} z_1 & z_2 \end{pmatrix} \underbrace{- \det \begin{pmatrix} z_0 & z_2 \end{pmatrix} + \det \begin{pmatrix} z_0 & z_1 \end{pmatrix}}_{= \det \begin{pmatrix} z_0 & z_1 - z_2 \end{pmatrix}} \\ &= \det \begin{pmatrix} z_1, z_2 - z_1 \end{pmatrix} + \det \begin{pmatrix} z_0, z_1 - z_2 \end{pmatrix} = \det \begin{pmatrix} z_0 - z_1, z_2 - z_1 \end{pmatrix}. \end{aligned}$$

Dies ist der Flächeninhalt des Parallelogramms, das durch $z_0 - z_1$ und $z_2 - z_1$ aufgespannt wird, also $2|T|$.
Für den Beweis der zweiten Aussage, bemerken wir, dass $\varphi_{z_0} + \dots + \varphi_{z_d} = 1$ auf T gilt, also $\nabla\varphi_{z_0} + \dots + \nabla\varphi_{z_d} = 0$. Dies ist die erste Zeile von (4.10). Außerdem gilt, dass

$$\varphi_{z_0}(x) z_0 + \dots + \varphi_{z_d}(x) z_d = x \qquad \text{für alle } x \in T,$$

da die rechte und die linke Seite affine Funktionen auf T sind, die in den Punkten $z_0, \dots, z_d$ übereinstimmen. Bildet man auf beiden Seiten die Ableitung, folgt (4.10). □

Bemerkung. Eine alternative Berechnung von $\int_T \nabla\varphi_j \cdot \nabla\varphi_k \, dx$ kann durch eine Transformation erfolgen

$$\int_T \nabla\varphi_j \cdot \nabla\varphi_k \, dx = \det(D\Phi_T) \int_{\widehat{T}} (\nabla\widehat{\varphi}_j)^\top D\Phi_T D\Phi_T^\top (\nabla\widehat{\varphi}_k) \, d\widehat{x},$$

wobei $\Phi_T : \widehat{T} \to T$ der Diffeomorphismus aus Lemma 4.35 ist. Auf dem Referenzsimplex $\widehat{T}$ gilt

$$\begin{aligned} \widehat{\varphi}_0(x) &= 1 - \sum_{j=1}^d x_j, \\ \widehat{\varphi}_k(x) &= x_k. \end{aligned}$$ ♢

Proposition 4.53 (Integrationsformel für barycentrische Koordinaten)**.** *Es gilt für ein Dreieck $T \subseteq \mathbb{R}^2$*

$$\int_T \lambda_1^\alpha \lambda_2^\beta \lambda_3^\gamma \, dx = 2|T| \frac{\alpha!\beta!\gamma!}{(\alpha+\beta+\gamma+2)!}.$$

Beweis. Der Beweis ist eine Übungsaufgabe. □

Ein FEM-Program sieht zum Beispiel wie folgt aus:

Input: $\mathcal{T}$, Γ_D (zum Beispiel in Form von `c4n, n4e, n4sDb, n4sNb`, siehe Abschnitt 3), f, g

Berechne lokale Steifigkeitsmatrizen $A_T := \left(\int_T \nabla\varphi_{z_j} \cdot \nabla\varphi_{z_k}\, dx\right)_{1\leq j,k\leq 3}$ für die lokale Basis $(\varphi_{z_1}, \varphi_{z_2}, \varphi_{z_3})$ auf T.
Assembliere globale Steifigkeitsmatrix A durch $A_{jk} = \sum_{T\in\mathcal{T}} \int_T \nabla\varphi_j \cdot \nabla\varphi_k\, dx \in \mathbb{R}^{|\mathcal{N}|\times|\mathcal{N}|}$.
Berechne $b \in \mathbb{R}^{|\mathcal{N}|}$ durch $b_j := \int_\Omega f\,\varphi_j\, dx + \int_{\Gamma_N} g\,\varphi_j\, ds$.
Berechne $\texttt{dof} = \mathcal{N} \setminus \Gamma_D$.
Löse $A(\texttt{dof}, \texttt{dof})x(\texttt{dof}) = b(\texttt{dof})$.

Output: x

Bemerkung. Die Matrix A ist keine voll besetzte Matrix, da $\int_\Omega \nabla\varphi_j \cdot \nabla\varphi_k\, dx = 0$ gilt, falls z_j und z_k nicht Ecken des selben Dreiecks sind. Die Steifigkeitsmatrix A sollte deshalb immer als *sparse*-Matrix definiert werden! ♢

Bemerkung. Um A zu assemblieren, könnte man eine Schleife über alle Dreiecke laufen lassen und dann für entsprechende Indizes `ind` die lokale Steifigkeitsmatrizen A_T in die globale Steifigkeitsmatrix mit $A(\texttt{ind}, \texttt{ind}) = A(\texttt{ind}, \texttt{ind}) + A_T$ eintragen. Diese Aufrufe von Einträgen einer sparse-Matrix sind zumindest in Matlab allerdings langsam. Eine Alternative in Matlab ist es, A über den Befehl

$$A = \texttt{sparse(I,J,Aloc,dim,dim)}$$

zu assemblieren, wobei `I` und `J` Index-Mengen sind, $\texttt{dim} = |\mathcal{N}|$ und

$$A_{jk} = \sum_{\texttt{I}(\ell)=j,\ \texttt{J}(\ell)=k} \texttt{Aloc}(\ell). \qquad ♢$$

Bemerkung. Wenn man Konvergenzraten numerisch überprüfen will, fängt man mit einer groben Triangulierung an, berechnet die FEM-Lösung, berechnet den Fehler, verfeinert die Triangulierung, berechnet die FEM-Lösung usw. Für die Verfeinerung kann zum Beispiel die rot-Verfeinerung oder die Bisektion aus Abbildung 3.3 benutzt werden. ♢

5 Sattelpunktprobleme

5.1 Abstrakte Sattelpunktprobleme

In diesem Kapitel werden wir Sattelpunktprobleme der Form

$$\begin{aligned} a(u,v)+b(v,p)&=F(v) \qquad \text{für alle } v\in V,\\ b(u,q)&=G(q) \qquad \text{für alle } q\in Q \end{aligned} \tag{5.1}$$

betrachten, wie sie in gemischten Methoden oder auch bei der Minimierung unter Nebenbedingungen auftreten. Definieren wir $\mathcal{B}:(V\times Q)\times(V\times Q)\to\mathbb{R}$ durch

$$\mathcal{B}((u,p),(v,q)):=a(u,v)+b(v,p)+b(u,q),$$

dann ist (5.1) äquivalent zu

$$\mathcal{B}((u,p),(v,q))=F(v)+G(q) \qquad \text{für alle } (v,q)\in V\times Q.$$

Allerdings ist $\mathcal{B}$ nicht koerzitiv, denn für alle $p\in Q$ gilt

$$\mathcal{B}((0,p),(0,p))=0.$$

Deswegen ist das Lax-Milgram-Lemma nicht anwendbar. Wir werden nun eine allgemeine Theorie für solche Sattelpunktprobleme betrachten.
Es seien X und Z Banach-Räume und X',Z' ihre Dualräume. Für $\varphi\in X'$ und $\psi\in Z'$ schreiben wir

$$\langle\varphi,x\rangle=\varphi(x) \quad \text{für alle } x\in X \qquad \text{und} \qquad \langle\psi,z\rangle=\psi(z) \quad \text{für alle } z\in Z.$$

Es sei $L:X\to Z$ ein beschränkter, linearer Operator, d.h. L ist linear und es existiert $C<\infty$ mit

$$\|Lx\|_Z\le C\|x\|_X \qquad \text{für alle } x\in X.$$

Für jedes feste $\psi\in Z'$ definiert die Abbildung

$$x\mapsto\langle\psi,Lx\rangle\in\mathbb{R}$$

ein Element in X', denn die Abbildung ist linear und es gilt

$$|\langle\psi,Lx\rangle|=|\psi(Lx)|\le\|\psi\|_{Z'}\,\|Lx\|_Z\le C\|\psi\|_{Z'}\,\|x\|_X.$$

Definition 5.1 (adjungierter Operator)**.** Der adjungierte Operator $L':Z'\to X'$ ist definiert durch

$$\langle L'\psi,x\rangle=\langle\psi,Lx\rangle \qquad \text{für alle } \psi\in Z' \text{ und } x\in X. \qquad \diamondsuit$$

Dies ist eine Verallgemeinerung der Transposition von Matrizen.

Definition 5.2 (Polare)**.** Es sei $W\subseteq Z'$. Dann definieren wir die Polare $W^\circ\subseteq Z$ von W durch

$$W^\circ=\{z\in Z\mid\langle\psi,z\rangle=0 \text{ für alle } \psi\in W\}. \qquad \diamondsuit$$

Ist Z ein Hilbertraum, dann können wir ihn mit seinem Dualraum identifizieren über die Abbildung $J:Z\to Z'$, $\langle Jz,y\rangle:=(z,y)_Z$ für alle $z,y\in Z$. Diese Abbildung wird auch Riesz-Abbildung genannt und ist ein Isomorphismus und eine Isometrie (d.h. $\|Jz\|_{Z'}=\|z\|_Z$ für alle $z\in Z$). In dieser Situation gilt $W^\circ=W^\perp$.

Satz 5.3 (Closed Range Theorem)**.** *Es sei $L : X \to Z$ ein beschränkter, linearer Operator. Dann ist*

$$\mathrm{Im}(L) \text{ abgeschlossen in } Z \qquad \Leftrightarrow \qquad \mathrm{Im}(L) = (\ker L')^\circ.$$

Für den Beweis benötigen wir den folgenden Satz aus der Funktionalanalysis.

Satz 5.4 (trennende Hyperebenen)**.** *Es sei V ein Banach-Raum und es seien $C_1, C_2 \subseteq V$ konvex mit $C_1 \cap C_2 = \emptyset$ und C_1 sei offen. Dann existiert ein $b \in V'$ und ein $m \in \mathbb{R}$ mit*

$$b(w_1) > m \geq b(w_2) \qquad \text{für alle } w_1 \in C_1, w_2 \in C_2.$$

Die Menge $\{v \in V \mid b(v) = m\}$ heißt trennende Hyperebene.

Bemerkung. Dieser Satz folgt aus dem Satz von Hahn-Banach. Der Satz von Hahn-Banach ist äquivalent zum Lemma von Zorn und zum Auswahlaxiom. ♢

Beweis von Satz 5.3. Wir beweisen zuerst $\Rightarrow$. Es sei $y = Lx \in \mathrm{Im}(L)$. Dann gilt für alle $\psi \in \ker L'$, dass

$$\langle \psi, y\rangle = \langle \psi, Lx\rangle = \langle L'\psi, x\rangle = 0,$$

also $y \in (\ker L')^\circ$.
Angenommen, es existiert ein $z_0 \in (\ker L')^\circ \setminus \mathrm{Im}(L)$. Da $\mathrm{Im}(L)$ abgeschlossen ist, existiert ein $\varepsilon > 0$ mit $B_\varepsilon(z_0) \cap \mathrm{Im}(L) = \emptyset$. Mit $C_1 = B_\varepsilon(z_0)$ und $C_2 = \mathrm{Im}(L)$ existiert nach Satz 5.4 dann ein $\psi \in Z'$ und ein $m \in \mathbb{R}$ mit

$$\langle \psi, z_0\rangle > m \geq \langle \psi, Lx\rangle \qquad \text{für alle } x \in X.$$

Dann gilt aber $\langle \psi, Lx\rangle = 0$ für alle $x \in X$ (sonst führt $\tilde{x} = \pm\alpha x$ auf $\langle \psi, L\tilde{x}\rangle = \pm\alpha\langle \psi, Lx\rangle$). Also $m \geq 0$. Außerdem gilt

$$\langle L'\psi, x\rangle = \langle \psi, Lx\rangle = 0 \qquad \text{für alle } x \in X,$$

also $\psi \in \ker L'$. Aber $z_0 \in (\ker L')^\circ$, also $\langle \psi, z_0\rangle = 0$ für alle $\psi \in \ker L'$, was ein Widerspruch ist.
Wir zeigen nun $\Leftarrow$. Es gilt

$$(\ker L')^\circ = \{z \in Z \mid \langle \psi, z\rangle = 0 \text{ für alle } \psi \in \ker L'\} = \bigcap_{\psi \in \ker L'} \ker \psi.$$

Da $\ker \psi$ abgeschlossen ist, ist $(\ker L')^\circ$ abgeschlossen, also auch $\mathrm{Im}(L)$. □

Lemma 5.5 (inf-sup-Bedingung für Operatoren)**.** *Es seien X und Y Banach-Räume und es sei $L : X \to Y'$ ein beschränkter, linearer Operator. Dann gilt*

$$L : X \to \mathrm{Im}(L) \text{ ist ein Isomorphismus} \quad \Leftrightarrow \quad \exists \beta > 0 \text{ mit } \inf_{x \in X\setminus\{0\}} \sup_{y \in Y\setminus\{0\}} \frac{\langle Lx, y\rangle}{\|x\|_X \, \|y\|_Y} \geq \beta,$$

wobei Isomorphismus heißt, dass L beschränkt ist, L^{-1} existiert und L^{-1} beschränkt ist.

Bemerkung. Mit der Definition der Operatornorm ist die inf-sup-Bedingung äquivalent zu

$$\|x\|_X \leq \beta^{-1} \|Lx\|_{Y'}. \qquad ♢$$

Beweis von Lemma 5.5. Wir beweisen zuerst $\Rightarrow$. Es sei $\psi \in \mathrm{Im}(L) \subseteq Y'$, $\psi = Lx$. Da L ein Isomorphismus ist, gilt

$$\|x\|_X \leq C_L \|\psi\|_{Y'}$$

mit der Stetigkeitskonstanten C_L von L^{-1}.
Wir zeigen nun $\Leftarrow$. Da $\beta\|x\|_X \leq \|Lx\|_{Y'}$, ist L injektiv, also $L : X \to \mathrm{Im}(L)$ bijektiv. Insbesondere existiert für alle $\psi \in \mathrm{Im}(L)$ ein eindeutiges $x \in X$ mit $Lx = \psi$. Dann gilt

$$\|x\|_X \leq \beta^{-1}\|Lx\|_{Y'} = \beta^{-1}\|\psi\|_{Y'},$$

also ist $L^{-1} : \mathrm{Im}(L) \to X$ beschränkt. □

Wir betrachten nun eine (nicht notwendig symmetrische) Bilinearform $\mathcal{B} : U \times V \to \mathbb{R}$, wobei U und V Hilbert-Räume sind. Wir betrachten den Operator $L : U \to V'$, der durch

$$\langle Lu, v\rangle = \mathcal{B}(u, v) \qquad \text{für alle } u \in U, v \in V$$

definiert ist. Wir werden die inf-sup-Bedingung für Operatoren nun durch Bedingungen an $\mathcal{B}$ ausdrücken.

Satz 5.6. *Die Abbildung $L : U \to V'$ ist genau dann ein Isomorphismus, wenn $\mathcal{B}$ folgende Bedingungen erfüllt*

1. *(Stetigkeit)* $\exists 0 < C < \infty$ *mit*

$$|\mathcal{B}(u, v)| \leq C\|u\|_U \, \|v\|_V \qquad \textit{für alle } u \in U, v \in V.$$

2. *(inf-sup-Bedingung)* $\exists 0 < \alpha < \infty$ *mit*

$$\alpha\|u\|_U \leq \sup_{v \in V\setminus\{0\}} \frac{\mathcal{B}(u, v)}{\|v\|_V} \qquad \textit{für alle } u \in U.$$

3. $\forall v \in V \setminus \{0\}$ $\exists u \in U$ *mit*

$$\mathcal{B}(u, v) \neq 0.$$

Beweis. Wir werden die Bedingung aus Lemma 5.5 zeigen. Die erste Bedingung aus Satz 5.6 ist äquivalent zu

$$\|Lu\|_{V'} \lesssim \|u\|_U,$$

was die Beschränktheit von L ist.
Die zweite Bedingung aus Satz 5.6 ist äquivalent dazu, dass L die inf-sup-Bedingung mit Konstante $\alpha > 0$ erfüllt.
Also ist $L : U \to \mathrm{Im}(L)$ ein Isomorphismus und es bleibt die Surjektivität von L auf V' zu zeigen. Dafür zeigen wir als erstes, dass $\mathrm{Im}(L)$ abgeschlossen ist:
Es sei $(v_j)_{j\in\mathbb{N}}$ eine Cauchy-Folge in $\mathrm{Im}(L) \subseteq V'$. Aus der Funktionalanalysis folgt, dass V' vollständig ist, da V ein Hilbertraum ist. Also existiert ein $v \in V'$ mit $v_j \to v$ in V'. Da $v_j \in \mathrm{Im}(L)$, existiert $u_j \in U$ mit $v_j = Lu_j$. Es gilt

$$\|u_j - u_k\|_U = \|L^{-1}(v_j - v_k)\|_U \overset{L^{-1} \text{ stetig}}{\lesssim} \|v_j - v_k\|_{V'} \to 0.$$

Also ist $(u_j)_{j\in\mathbb{N}}$ eine Cauchy-Folge in U und da U vollständig ist, existiert ein $u \in U$ mit $u_j \to u$ in U. Daraus folgt

$$\|v - Lu\|_{V'} \leq \|v - v_j\|_{V'} + \|L(u_j - u)\|_{V'} \lesssim \|v - v_j\|_{V'} + \|u_j - u\|_U \to 0,$$

also $v \in \mathrm{Im}(L)$, also ist $\mathrm{Im}(L)$ abgeschlossen. Nach dem Closed-Range-Theorem, Satz 5.3, gilt dann $\mathrm{Im}(L) = (\ker L')^\circ$, also ist L surjektiv auf

$$\mathrm{Im}(L) = (\ker L')^\circ = \{\psi \in V' \mid \langle v, \psi\rangle = 0 \text{ für alle } v \in \ker(L')\}.$$

Es gilt $v \in \ker(L')$ genau dann, wenn für alle $u \in U$ gilt $0 = \langle L'v, u\rangle = \langle v, Lu\rangle$. Mit der Identifikation von V mit V'' gilt dann $\langle v, Lu\rangle = \langle Lu, v\rangle = \mathcal{B}(u, v)$. Nach der dritten Bedingung in Satz 5.6 ist also $\ker(L') = \{0\}$, also $\mathrm{Im}(L) = V'$. □

Bemerkung. Satz 5.6 ist eine Verallgemeinerung des Satzes von Lax-Milgram, da die Bedingungen 2 und 3 aus der Koerzitivität folgen. ♢

Für die Approximation der exakten Lösung $u \in U$ betrachten wir nun das Galerkin-Verfahren für endlich-dimensionale Unterräume $U_h \subseteq U$ und $V_h \subseteq V$. Zu $f \in V'$ approximieren wir die Lösung $u \in U$ zu

$$\mathcal{B}(u, v) = \langle f, v\rangle \qquad \text{für alle } v \in V$$

durch $u_h \in U_h$ mit

$$\mathcal{B}(u_h, v_h) = \langle f, v_h\rangle \qquad \text{für alle } v_h \in V_h. \tag{5.2}$$

Satz 5.7. *Die Bilinearform $\mathcal{B} : U \times V \to \mathbb{R}$ erfülle die Bedingungen aus Satz 5.6. Die Unterräume $U_h \subseteq U$ und $V_h \subseteq V$ seien so gewählt, dass sie*

$$\alpha_h \leq \inf_{u_h \in U_h} \sup_{v_h \in V_h} \frac{\mathcal{B}(u_h, v_h)}{\|u_h\|_U \|v_h\|_V}$$

erfüllen und dass gilt: $\forall v_h \in V_h \setminus \{0\}\ \exists u_h \in U_h$ mit

$$\mathcal{B}(u_h, v_h) \neq 0.$$

Dann existiert ein $u_h \in U_h$, das (5.2) löst und es gilt

$$\|u - u_h\|_U \leq \left(1 + \frac{C}{\alpha_h}\right) \inf_{w_h \in U_h} \|u - w_h\|_U.$$

Beweis. Aus den Voraussetzungen an $\mathcal{B}$, U_h und V_h folgt, dass es eine eindeutige Lösung $u_h \in U_h$ von (5.2) gibt. Es sei $w_h \in U_h$ beliebig. Dann existiert wegen der diskreten inf-sup-Bedingung ein $v_h \in V_h$ mit $\|v_h\|_V = 1$ und

$$\begin{aligned}\|u_h - w_h\|_U &\leq \alpha_h^{-1} \mathcal{B}(u_h - w_h, v_h) = \alpha_h^{-1}(\langle f, v_h\rangle - \mathcal{B}(w_h, v_h)) \\ &= \alpha_h^{-1} \mathcal{B}(u - w_h, v_h) \leq \alpha_h^{-1} C \|u - w_h\|_U.\end{aligned}$$

Mit der Dreiecksungleichung folgt die Behauptung. □

Wir betrachten als Motivation im Folgenden das folgende Minimierungsproblem unter Nebenbedingungen.

Problem 5.8 (Minimierung unter Nebenbedingung)**.** Minimiere

$$J(v) = \frac{1}{2}a(v,v) - \langle f, v\rangle \qquad \text{in } X$$

unter der Nebenbedingung

$$b(v,\mu) = \langle g, \mu\rangle \qquad \text{für alle } \mu \in M. \qquad \diamondsuit$$

Dabei seien X und M Hilbert-Räume, $a : X \times X \to \mathbb{R}$ und $b : X \times M \to \mathbb{R}$ stetige Bilinearformen. Es sei a symmetrisch und $a(v,v) \geq 0$ für alle $v \in X$. Außerdem sei $f \in X'$ und $g \in M'$. Wir betrachten weiterhin das Lagrange-Funktional $\mathcal{L} : X \times M \to \mathbb{R}$,

$$\mathcal{L}(v,\mu) := J(v) + b(v,\mu) - \langle g, \mu\rangle$$

und das Problem

Problem 5.9 (Sattelpunktproblem)**.** Finde $(u,\lambda) \in X \times M$ mit

$$\begin{aligned} a(u,v) + b(v,\lambda) &= \langle f, v\rangle \qquad \text{für alle } v \in X, \\ b(u,\mu) &= \langle g, \mu\rangle \qquad \text{für alle } \mu \in M. \end{aligned} \qquad \diamondsuit$$

Für die Lösung des Problems 5.9 gilt die Sattelpunktseigenschaft

$$\mathcal{L}(u,\mu) \leq \mathcal{L}(u,\lambda) \leq \mathcal{L}(v,\lambda) \qquad \text{für alle } (v,\mu) \in X \times M,$$

denn:

$$\begin{aligned} &\mathcal{L}(u,\lambda) - \mathcal{L}(v,\lambda) \\ &\quad = \frac{1}{2}a(u,u) - \langle f, u\rangle + \underbrace{b(u,\lambda) - \langle g,\lambda\rangle}_{=0} - \frac{1}{2}a(v,v) + \langle f, v\rangle \underbrace{-b(v,\lambda)}_{a(u,v)-\langle f,v\rangle} \underbrace{+\langle g,\lambda\rangle}_{=b(u,\lambda)} \end{aligned}$$

und wegen $b(u,\lambda) = \langle f, u\rangle - a(u,u)$ folgt

$$\mathcal{L}(u,\lambda) - \mathcal{L}(v,\lambda) = -\frac{1}{2}a(u-v,u-v) \leq 0.$$

Außerdem gilt

$$\mathcal{L}(u,\lambda) - \mathcal{L}(u,\mu) = b(u,\lambda) - \langle g,\lambda\rangle - b(u,\mu) + \langle g,\mu\rangle = 0.$$

Proposition 5.10. *Ist $(u,\lambda) \in X \times M$ eine Lösung von Problem 5.9, dann ist u eine Lösung von Problem 5.8.*

Beweis. Nach Definition erfüllt u die Nebenbedingung.
Definiere $V := \{v \in X \mid b(v,\mu) = 0 \text{ für alle } \mu \in M\}$. Dann sucht Problem 5.8 den Minimierer von J in $u_g + V$, wobei $u_g \in X$ eine beliebige, aber feste Funktion sei, die $b(u_g,\mu) = \langle g,\mu\rangle$ für alle $\mu \in M$ erfüllt. Mit $w = u_0 + u_g$ ist dies äquivalent zum Minimierer $u_0 \in V$ von

$$J(v) + a(u_g, v) \qquad \text{in } V.$$

Nach einer Übungsaufgabe ist dieses Minimierungsproblem äquivalent dazu, dass

$$a(u_0,v) = \langle f, v\rangle - a(u_g,v) \qquad \text{für alle } v \in V,$$

d.h. $w = u_0 + u_g$ erfüllt

$$a(w,v) = \langle f, v\rangle \qquad \text{für alle } v \in V.$$

Da diese Gleichung von u erfüllt ist, ist u die Lösung von Problem 5.8. $\square$

Wir wenden uns jetzt der Frage zu, wann eine Lösung vom Sattelpunktproblem 5.9 existiert. Dafür sei V wie im Beweis und

$$V(g) := \{v \in X \mid b(v,\mu) = \langle g, \mu\rangle \text{ für alle } \mu \in M\}.$$

Aus der Stetigkeit von b folgt, dass V abgeschlossen ist. Wir definieren die folgenden Abbildungen

$$\begin{aligned} A : X \to X', \quad \langle Au, v\rangle &= a(u,v) \qquad \text{für alle } v \in X,\\ B : X \to M', \quad \langle Bu, \mu\rangle &= b(u,\mu) \qquad \text{für alle } \mu \in M \end{aligned}$$

und es sei $B' : M \to X'$ die adjungierte Abbildung, wobei wir M mit M'' identifizieren, d.h.

$$\langle B'\lambda, v\rangle = b(v,\lambda) \qquad \text{für alle } v \in X.$$

Dann ist das Sattelpunktproblem 5.9 äquivalent zu

$$\begin{aligned} Au + B'\lambda &= f \qquad \text{in } X',\\ Bu &= g \qquad \text{in } M'. \end{aligned}$$

Lemma 5.11. *Die folgenden Aussagen sind äquivalent:*

1. $\exists \beta > 0$ *mit*

$$\inf_{\mu\in M\setminus\{0\}} \sup_{v\in X\setminus\{0\}} \frac{b(v,\mu)}{\|v\|_X\,\|\mu\|_M} \geq \beta.$$

2. $B : V^\perp \to M'$ *ist ein Isomorphismus und*

$$\|Bv\|_{M'} \geq \beta\|v\|_X \qquad \textit{für alle } v \in V^\perp$$

   (dabei ist $V^\perp = \{x \in X \mid (x,v)_X = 0$ *für alle* $v \in V\}$*).*

3. $B' : M \to V^\circ \subseteq X'$ *ist ein Isomorphismus und*

$$\|B'\mu\|_{X'} \geq \beta\|\mu\|_M \qquad \textit{für alle } \mu \in M.$$

Beweis. Wir zeigen als erstes (1) $\Leftrightarrow$ (3): Da $\sup_{v\in X\setminus\{0\}} \frac{b(v,\mu)}{\|v\|_X} = \|B'\mu\|_{X'}$, folgt (3) $\Rightarrow$ (1). Außerdem folgt damit aus (1) die Ungleichung in (3) und dass $B' : M \to \mathrm{Im}(B')$ ein Isomorphismus ist. Wie im Beweis von Satz 5.6 folgt, dass $\mathrm{Im}(B')$ abgeschlossen ist. Nach dem Closed-Range-Theorem, Satz 5.3, gilt also $\mathrm{Im}(B') = (\ker(B''))^\circ = (\ker(B))^\circ = V^\circ$. Daraus folgt (3).
Es sei nun (3) erfüllt. Zu $v \in V^\perp$ definiere $(v,\bullet)_X \in X'$. Es gilt $(v,w)_X = 0$ für alle $w \in V$, also $(v,\bullet)_X \in V^\circ$ nach Definition. Da $B' : M \to V^\circ$ ein Isomorphismus ist, existiert $\lambda \in M$ mit

$$\langle B'\lambda, w\rangle = b(w,\lambda) = (v,w)_X \qquad \text{für alle } w \in X.$$

Es gilt mit der Ungleichung in (3), dass $\|v\|_X = \|(v,\bullet)_X\|_{X'} = \|B'\lambda\|_{X'} \geq \beta\|\lambda\|_M$. Also

$$\|Bv\|_{M'} = \sup_{\mu\in M\setminus\{0\}} \frac{b(v,\mu)}{\|\mu\|_M} \geq \frac{b(v,\lambda)}{\|\lambda\|_M} = \frac{(v,v)_X}{\|\lambda\|_M} \geq \beta\|v\|_X.$$

Für $B : V^\perp \to M'$ gilt also die inf-sup-Bedingung. Außerdem gilt die Stetigkeit und aus (1) folgt, dass es für alle $\mu \in M \setminus \{0\}$ ein $v \in V^\perp$ gibt mit $b(v,\mu) \neq 0$. Damit sind die Voraussetzungen aus Satz 5.6 erfüllt und B ist ein Isomorphismus.
Es gelte nun (2). Für $\mu \in M$ gilt dann

$$\|\mu\|_M = \sup_{g\in M'\setminus\{0\}} \frac{\langle g,\mu\rangle}{\|g\|_{M'}} \overset{(2)}{=} \sup_{v\in V^\perp\setminus\{0\}} \frac{\langle Bv,\mu\rangle}{\|Bv\|_{M'}} = \sup_{v\in V^\perp\setminus\{0\}} \frac{b(v,\mu)}{\|Bv\|_{M'}} \overset{(2)}{\leq} \sup_{v\in V^\perp\setminus\{0\}} \frac{b(v,\mu)}{\beta\|v\|_X}.$$

Daraus folgt (1). □

Satz 5.12 (Brezzi's Splitting Theorem)**.** *Durch Problem 5.9 wird genau dann ein Isomorphismus $L : X \times M \to X' \times M'$ erklärt, wenn gilt*

(a) a ist V-elliptisch, d.h. es existiert ein $\alpha > 0$ mit

$$a(v,v) \geq \alpha\|v\|_X^2 \qquad \text{für alle } v \in V,$$

(b) b erfüllt die inf-sup-Bedingung

$$\inf_{\mu\in M\setminus\{0\}} \sup_{v\in X\setminus\{0\}} \frac{b(v,\mu)}{\|v\|_X\ \|\mu\|_M} \geq \beta.$$

Bemerkung. Satz 5.12 wird auch LBB-Bedingung genannt nach Ladyženskaya, Babuška und Brezzi. ♢

Bemerkung. Durch Problem 5.9 wird eine Abbildung $L : X \times M \to X' \times M'$ erklärt im Sinne, dass $L(u,\lambda) = (f,g)$ genau dann, wenn

$$\begin{aligned} a(u,v) + b(v,\lambda) &= \langle f, v\rangle \qquad \text{für alle } v \in X \\ b(u,\mu) &= \langle g,\mu\rangle \qquad \text{für alle } \mu \in M. \end{aligned}$$

Die Norm auf $X \times M$ ist gegeben durch

$$\|(v,\mu)\|_{X\times M} := \|v\|_X + \|\mu\|_M \qquad \text{für alle } (v,\mu) \in X \times M$$

und auf $X' \times M'$ durch

$$\|(f,g)\|_{X'\times M'} := \|f\|_{X'} + \|g\|_{M'} \qquad \text{für alle } (f,g) \in X' \times M'. \qquad ♢$$

Beweis von Satz 5.12. Es gelte (a) und (b) und es sei $(f,g) \in X'\times M'$. Die inf-sup-Bedingung (b) und Lemma 5.11 implizieren, dass $u_0 \in V^\perp$ existiert mit $Bu_0 = g$, also $V(g) \neq \emptyset$. Es gilt

$$\|u_0\|_X \leq \beta^{-1}\|g\|_{M'}.$$

Mit $w := u - u_0$ ist Problem 5.9 äquivalent zu

$$\begin{aligned} a(w,v) + b(v,\lambda) &= \langle f,v\rangle - a(u_0,v) \qquad \text{für alle } v \in X, \\ b(w,\mu) &= 0 \qquad \text{für alle } \mu \in M. \end{aligned}$$

Nach dem Satz von Lax-Milgram gilt wegen der V-Elliptizität von a, dass es eine eindeutige Lösung $w \in V$ gibt mit

$$a(w,v) = \langle f,v\rangle - a(u_0,v) \qquad \text{für alle } v \in V$$

und es gilt

$$\|w\|_X \leq \alpha^{-1}(\|f\|_{X'} + C\|u_0\|_X),$$

wobei C die Stetigkeitskonstante von a ist. Das Funktional

$$\langle f, \bullet\rangle - a(u_0 + w, \bullet) \in X'$$

erfüllt

$$\langle f, v\rangle - a(u_0 + w, v) = 0 \qquad \text{für alle } v \in V,$$

also ist $h := \langle f, \bullet\rangle - a(u_0 + w, \bullet) \in V^\circ$. Nach Lemma 5.11, (3) existiert also ein $\lambda \in M$ mit $B'\lambda = h$, also

$$b(v, \lambda) = \langle B'\lambda, v\rangle = \langle f, v\rangle - a(u_0 + w, v) \qquad \text{für alle } v \in X$$

und

$$\|\lambda\|_M \leq \beta^{-1}\|B'\lambda\|_{X'} \leq \beta^{-1}(\|f\|_{X'} + C\|u_0 + w\|_X).$$

Also ist mit $u = u_0 + w$ das Paar $(u, \lambda) \in X \times M$ eine Lösung von Problem 5.9 (also $L(u, \lambda) = (f, g)$). Dies zeigt die Surjektivität von L. Die Stetigkeit von L^{-1} folgt wegen

$$\begin{aligned}\|u\|_X \leq \|u_0\|_X + \|w\|_X &\leq \beta^{-1}\|g\|_{M'} + \alpha^{-1}\|f\|_{X'} + \alpha^{-1}C\beta^{-1}\|g\|_{M'} \\ &= \alpha^{-1}\|f\|_{X'} + \beta^{-1}(1 + C\alpha^{-1})\|g\|_{M'}\end{aligned}$$

und

$$\|\lambda\|_M \leq \beta^{-1}\|f\|_{X'} + \beta^{-1}C\|u\|_X \leq \beta^{-1}(1 + C\alpha^{-1})\|f\|_{X'} + \beta^{-2}C(1 + C\alpha^{-1})\|g\|_{M'}.$$

Für die Injektivität von L sei $f = 0$ und $g = 0$. Dann gilt mit der Testfunktion u, dass

$$\alpha\|u\|_X^2 \leq a(u, u) = 0,$$

also $u = 0$. Außerdem folgt wegen $b(v, \lambda) = 0$ für alle $v \in X$ aus der inf-sup-Bedingung von b, dass $\lambda = 0$.
Außerdem ist L stetig mit Stetigkeitskonstante $\max\{C, C_b\}$, wobei C_b die Stetigkeitskonstante von b ist. Wir haben also gezeigt, dass L ein Isomorphismus ist.
Für die andere Richtung sei L nun ein Isomorphismus. Es sei $u \in V$ und $f \in X'$ sei definiert durch $\langle f, v\rangle = a(u, v)$ für alle $v \in X$. Dann erfüllt $\lambda = 0 \in M$, dass $(u, \lambda) = L^{-1}(f, 0)$, also

$$\begin{aligned} a(u, v) + b(v, \lambda) = \langle f, v\rangle \qquad &\text{für alle } v \in X, \\ b(u, \mu) = 0 \qquad &\text{für alle } \mu \in M \quad (\text{da } u \in V).\end{aligned}$$

Es gilt

$$\|f\|_{X'} = \sup_{v \in X\setminus\{0\}} \frac{a(u, v)}{\|v\|_X} \leq \sup_{v \in X\setminus\{0\}} \frac{\sqrt{a(u, u)}\sqrt{a(v, v)}}{\|v\|_X} \leq C_a\sqrt{a(u, u)}.$$

Da L^{-1} nach Voraussetzung beschränkt ist, also

$$\|u\|_X \leq \|(u, \lambda)\|_{X\times M} \leq C_{L^{-1}}\,\|(f, 0)\|_{X'\times M'} = C_{L^{-1}}\|f\|_{X'} \leq C_{L^{-1}}C_a\sqrt{a(u, u)},$$

wobei $C_{L^{-1}}$ die Stetigkeitskonstante von L^{-1} ist. Da $u \in V$ beliebig war, folgt die V-Elliptizität von a.
Wir betrachten nun $B : X \to M'$, definiert durch $\langle Bu, \mu\rangle = b(u, \mu)$ für alle $\mu \in M$. Wenn wir zeigen, dass $B : V^\perp \to M'$ ein Isomorphismus ist und $\|Bv\|_{M'} \geq \beta\|v\|_X$ für alle $v \in V^\perp$, dann folgt mit Lemma 5.11 die inf-sup-Bedingung von b. Da $V = \ker B$ nach Definition gilt, ist $B : V^\perp \to M'$ injektiv. Die Beschränktheit von B folgt aus der Beschränktheit von b. Es sei nun $g \in M'$. Da L ein Isomorphismus ist, existiert $(u, \lambda) \in X \times M$ mit $(u, \lambda) = L^{-1}(0, g)$ und $\|u\|_X + \|\lambda\|_M \leq C_{L^{-1}}\|g\|_{M'}$. Es gilt also $b(u, \mu) = g(\mu)$ für alle $\mu \in M$, also $Bu = g$. Es sei $u^\perp$ die Orthogonalprojektion von u auf $V^\perp$, d.h.

$$(u^\perp, v)_X = (u, v)_X \qquad \text{für alle } v \in V^\perp.$$

Dann gilt $Bu^\perp = Bu + B(u^\perp - u) = Bu = g$. Also ist $B : V^\perp \to M'$ surjektiv. Es gilt $\|u^\perp\|_X \leq \|u\|_X \leq C_{L^{-1}}\|g\|_{M'}$, also ist $(B|_{V^\perp})^{-1}$ beschränkt, also $B : V^\perp \to M'$ ein Isomorphismus. Außerdem zeigt die Ungleichung $\|u^\perp\|_X \leq C_{L^{-1}}\|g\|_{M'}$ die Ungleichung in der Eigenschaft (2) in Lemma 5.11. □

Es seien nun $X_h \subseteq X$ und $M_h \subseteq M$ endlich-dimensionale Teilräume. Wir wollen die Lösung $(u, \lambda) \in X \times M$ von

$$\begin{aligned} a(u, v) + b(v, \lambda) &= \langle f, v\rangle \qquad \text{für alle } v \in X, \\ b(u, \mu) &= \langle g, \mu\rangle \qquad \text{für alle } \mu \in M \end{aligned} \tag{5.3}$$

approximieren durch eine diskrete Lösung $(u_h, \lambda_h) \in X_h \times M_h$ von

$$\begin{aligned} a(u_h, v_h) + b(v_h, \lambda_h) &= \langle f, v_h\rangle \qquad \text{für alle } v_h \in X_h, \\ b(u_h, \mu_h) &= \langle g, \mu_h\rangle \qquad \text{für alle } \mu_h \in M_h. \end{aligned} \tag{5.4}$$

Definiere $V_h := \{v_h \in X_h \mid b(v_h, \mu_h) = 0 \text{ für alle } \mu_h \in M_h\}$.

Bemerkung. Im allgemeinen gilt *nicht* $V_h \subseteq V$. Fasst man die beiden Probleme aus (5.3) und (5.4) als

$$\begin{aligned} a(u, v) &= \langle f, v\rangle \qquad \text{für alle } v \in V, \\ a(u_h, v_h) &= \langle f, v_h\rangle \qquad \text{für alle } v_h \in V_h \end{aligned}$$

auf, dann folgt durch die Theorie der elliptischen Probleme *keine* Fehlerabschätzung der Art

$$\|u - u_h\|_X \lesssim \inf_{v_h \in V_h} \|u - v_h\|_X.$$

Die diskrete Lösung u_h kann man als nicht-konforme Approximation von u auffassen. ♢

Die Räume X_h und M_h seien so gewählt, dass

(i) a ist V_h elliptisch mit Elliptizitätskonstante $\alpha > 0$, die unabhängig von der Gitterweite sei

(ii) Es gelte

$$\beta\|\lambda_h\|_M \leq \sup_{v_h \in X_h} \frac{b(v_h, \lambda_h)}{\|v_h\|_X} \qquad \text{für alle } \lambda_h \in M_h,$$

wobei $\beta > 0$ unabhängig von der Gitterweite sei.

Korollar 5.13. *Es gelten die Voraussetzung aus Brezzi's Splitting Theorem, Satz 5.12, und es gelte (i) und (ii). Dann existieren die Lösungen $(u, \lambda) \in X \times M$ und $(u_h, \lambda_h) \in X_h \times M_h$ und es gilt*

$$\|u - u_h\|_X + \|\lambda - \lambda_h\|_M \lesssim \inf_{v_h \in X_h} \|u - v_h\|_X + \inf_{\mu_h \in M_h} \|\lambda - \mu_h\|_M$$

Beweis. Es sind durch $\langle f, \bullet \rangle$ und $\langle g, \bullet \rangle$ Funktionale in X_h' bzw. M_h' definiert und die V_h-Elliptizität und die diskrete inf-sup-Bedingung zeigen nach Brezzi's Splitting Theorem, dass das zugehörige $L_h : X_h \times M_h \to X_h' \times M_h'$ ein Isomorphismus ist, also existiert eine eindeutige diskrete Lösung. Außerdem erfüllt die zugehörige Bilinearform $\mathcal{B}$ nach Satz 5.6 die inf-sup-Bedingung bezüglich X_h und M_h und die Eigenschaft 3 aus Satz 5.6. Damit sind die Voraussetzungen von Satz 5.7 erfüllt, also folgt die Fehlerabschätzung folgt aus Satz 5.7. □

Bemerkung. Falls $V_h \subseteq V$ gilt, dann folgt aus Céas Lemma, dass

$$\|u - u_u\|_X \lesssim \inf_{v_h \in V_h} \|u - v_h\|_X. \qquad \diamondsuit$$

Nun wollen wir noch die Fortin-Interpolation einführen, die benutzt werden kann zum Nachweis der diskreten inf-sup-Bedingung.

Satz 5.14 (Fortin-Interpolation)**.** *Die Bilinearform $b : X \times M \to \mathbb{R}$ erfülle die inf-sup-Bedingung bezüglich X und M. Dann gilt, dass X_h und M_h die (diskrete) inf-sup-Bedingung mit Konstante unabhängig von h erfüllen genau dann, wenn ein $\Pi_h : X \to X_h$ existiert mit $b(v - \Pi_h v, \mu_h) = 0$ für alle $\mu_h \in M_h$ und*

$$\|\Pi_h\|_{L(X,X)} := \sup_{v \in X} \frac{\|\Pi_h v\|_X}{\|v\|_X} \leq C$$

unabhängig von h gilt.

Beweis. Wir zeigen zuerst $\Leftarrow$. Es sei $\mu_h \in M_h \subseteq M$. Dann gilt

$$\beta \|\mu_h\|_M \leq \sup_{v \in X} \frac{b(v, \mu_h)}{\|v\|_X} = \sup_{v \in X} \frac{b(\Pi_h v, \mu_h)}{\|v\|_X} \leq C \sup_{v \in X} \frac{b(\Pi_h v, \mu_h)}{\|\Pi_h v\|_X} \leq C \sup_{v_h \in X_h} \frac{b(v_h, \mu_h)}{\|v_h\|_X}.$$

Also ist die diskrete inf-sup-Bedingung erfüllt.
Wir zeigen nun $\Rightarrow$. Zu $v \in X$ definiere $w_h \in X_h$ (und $\lambda_h \in M_h$) als Lösung von

$$\begin{aligned} (w_h, v_h)_X + b(v_h, \lambda_h) &= (v, v_h)_X \qquad \text{für alle } v_h \in X_h, \\ b(w_h, \mu_h) &= b(v, \mu_h) \qquad \text{für alle } \mu_h \in M_h. \end{aligned}$$

Die Existenz einer solchen Lösung folgt aus der Koerzitivität des Skalarprodukts und der diskreten inf-sup-Bedingung von b aus Brezzi's Splitting Theorem. Dann gilt wegen der zweiten Gleichung

$$b(v - w_h, \mu_h) = 0 \qquad \text{für alle } \mu_h \in M_h,$$

und aus der Stabilität des Systems folgt

$$\|w_h\|_X \lesssim \sup_{v_h \in X_h} \frac{(v, v_h)_X}{\|v_h\|_X} + \sup_{\mu_h \in M_h} \frac{b(v, \mu_h)}{\|\mu_h\|_M} \lesssim \|v\|_X.$$

Damit erfüllt $\Pi_h v := w_h$ die geforderten Bedingungen. □

Bemerkung. Wir haben für den Beweis $M_h \subseteq M$ benutzt. Es muss allerdings nicht $X_h \subseteq X$ gelten. $\diamondsuit$

5.2 Gemischte FEM für das Poisson-Problem

Die gemischte Formulierung des Poisson-Problems (PMP) basiert auf der Aufsplittung des PMP $-\Delta u = f$ in

$$\begin{aligned} p &= \nabla u, \\ -\operatorname{div} p &= f. \end{aligned}$$

Für die schwache Formulierung müssen wir noch den Raum definieren, in dem p gesucht werden wird.

Definition 5.15. Definiere

$$H(\operatorname{div},\Omega) := \left\{ q \in L^2(\Omega;\mathbb{R}^d) \;\middle|\; \exists g \in L^2(\Omega) \text{ mit } \int_\Omega q\cdot\nabla v\,dx = -\int_\Omega v\,g\,dx \text{ für alle } v\in\mathcal{C}_c^\infty(\Omega) \right\}.$$

Wir schreiben dann auch $\operatorname{div} q = g$ und bezeichnen $\operatorname{div} q$ als schwache Divergenz von q. Wir statten den Raum $H(\operatorname{div},\Omega)$ mit der Norm

$$\|q\|_{H(\operatorname{div},\Omega)} := \sqrt{\|q\|^2 + \|\operatorname{div} q\|^2}$$

aus. ♢

Bemerkung. Es gilt $(H^1(\Omega))^d \subset H(\operatorname{div},\Omega)$, aber wie wir später sehen werden $H(\operatorname{div},\Omega) \not\subset (H^1(\Omega))^d$. ♢

Definition 5.16 (gemischte Formulierung des Poisson-Problems)**.** Die gemischte Formulierung des Poisson-Problems sucht $(p,u)\in H(\operatorname{div},\Omega)\times L^2(\Omega)$ mit

$$\begin{aligned} \int_\Omega p\cdot q\,dx + \int_\Omega u\operatorname{div} q\,dx &= 0 && \text{für alle } q\in H(\operatorname{div},\Omega), \\ \int_\Omega v\operatorname{div} p\,dx &= -\int_\Omega f\,v\,dx && \text{für alle } v\in L^2(\Omega). \end{aligned}$$ ♢

Satz 5.17. *Es existiert eine eindeutige Lösung $(p,u)\in H(\operatorname{div},\Omega)\times L^2(\Omega)$ zur gemischten Formulierung und es gilt $u\in H_0^1(\Omega)$ und $p=\nabla u$ in $L^2(\Omega)$ und u ist die Lösung des Poisson-Problems*

$$\int_\Omega \nabla u\cdot\nabla v\,dx = \int_\Omega f\,v\,dx \qquad \textit{für alle } v\in H_0^1(\Omega).$$

Beweis. Setze $X = H(\operatorname{div},\Omega)$ mit Norm $\|\cdot\|_{H(\operatorname{div},\Omega)}$, $M = L^2(\Omega)$ mit Norm $\|\bullet\|$, $a(p,q) = \int_\Omega p\cdot q\,dx$ für alle $p,q\in X$ und $b(q,v) = \int_\Omega v\operatorname{div} q\,dx$ für alle $v\in M, q\in X$. Dann sind a und b stetig. Es sei nun $q\in X$ mit $b(q,v)=0$ für alle $v\in M$. Dann gilt $\operatorname{div} q = 0$ in $L^2(\Omega)$, also

$$a(q,q) = \|q\|^2 = \|q\|^2 + \|\operatorname{div} q\|^2 = \|q\|^2_{H(\operatorname{div},\Omega)},$$

also ist a auch elliptisch auf dem Kern V von b.
Es sei nun $w\in M$ beliebig. Es sei $u\in H_0^1(\Omega)$ die Lösung zu

$$\int_\Omega \nabla u\cdot\nabla v\,dx = \int_\Omega w\,v\,dx \qquad \text{für alle } v\in H_0^1(\Omega).$$

Setze $p := -\nabla u$. Dann gilt $p\in H(\operatorname{div},\Omega)$ mit $\operatorname{div} p = w$, also

$$b(w,p) = \int_\Omega w\operatorname{div} p\,dx = \|w\|^2.$$

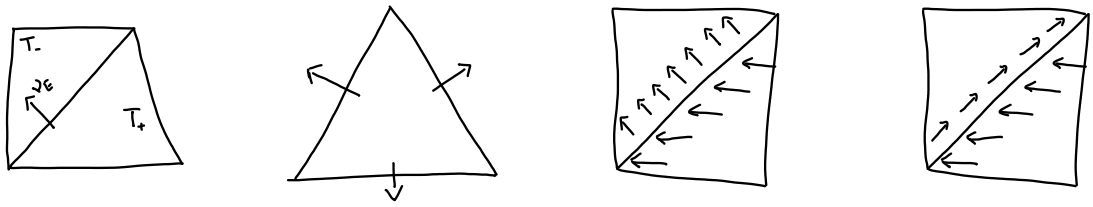

Abbildung 5.1: Situation in Definition 5.18, schematische Darstellung des Raviart-Thomas Finiten Elements und Skizze von zulässigen und nicht zulässigen Sprüngen.

Außerdem gilt

$$\|p\|^2_{H(\mathrm{div},\Omega)} = \underbrace{\|p\|^2}_{=\|\nabla u\|^2} + \underbrace{\|\mathrm{div}\, p\|^2}_{=\|w\|^2},$$
$$\|\nabla u\|^2 = \int_\Omega w\, u\, dx \le \|w\|\, \|u\| \lesssim \|w\|\, \|\nabla u\|,$$

also

$$\|p\|_{H(\mathrm{div},\Omega)} \lesssim \|w\|.$$

Also gilt

$$\frac{b(p,w)}{\|p\|_{H(\mathrm{div},\Omega)}} \gtrsim \frac{\|w\|^2}{\|w\|} = \|w\|.$$

Also gilt die inf-sup-Bedingung für b und es existiert eine eindeutige Lösung.
Mit Testfunktionen $v \in C_c^\infty(\Omega)$ und $q \in H(\mathrm{div},\Omega)$ definiert durch $q_j = v$, $q_k = 0$ für $k \neq j$, folgt aus $\int_\Omega u\, \mathrm{div}\, q\, dx = -\int_\Omega p \cdot q\, dx$, dass $u \in H^1(\Omega)$ und $p = \nabla u$. Benutzen wir diese Gleichheit und testen nun mit Funktionen $v \in C^\infty(\Omega)$, die nicht auf dem Rand von Ω verschwinden müssen, folgt

$$\int_{\partial\Omega} uvn_j\, ds = 0.$$

Daraus folgt $u|_{\partial\Omega} = 0$, also $u \in H_0^1(\Omega)$. Außerdem gilt für alle $v \in H_0^1(\Omega)$

$$\int_\Omega \nabla u \cdot \nabla v\, dx = \int_\Omega p \cdot \nabla v\, dx = -\int_\Omega v\, \mathrm{div}\, p\, dx = \int_\Omega f\, v\, dx.$$

Also ist u die Lösung des Poisson-Problems. □

Aus Proposition 5.10 folgt, dass p die Lösung zum dualen Problem

$$\text{Maximiere} \quad -\frac{1}{2}\int_\Omega p \cdot p\, dx$$
$$\text{unter der Nebenbedingung} \quad \int_\Omega v\, \mathrm{div}\, p\, dx = -\int_\Omega f\, v\, dx \qquad \text{für alle } v \in L^2(\Omega)$$

ist. Für eine Galerkin-Approximation definieren wir als nächstes endlich-dimensionale Teilräume von $H(\mathrm{div},\Omega)$ und $L^2(\Omega)$.

Definition 5.18 (Raviart-Thomas Finites Element, 1977)**.** Für eine formreguläre Triangulierung $\mathcal{T}$ eines Lipschitz-Gebiets $\Omega \subseteq \mathbb{R}^2$ ist der Finite-Elemente-Raum von Raviart und Thomas definiert als

$$\mathrm{RT}_k(\mathcal{T}) := \left\{ q \in L^2(\Omega;\mathbb{R}^2) \,\middle|\, \begin{array}{l} \forall T \in \mathcal{T} \exists a_T, b_T, c_T \in P_k(T) \text{ mit } q|_T(x) = \begin{pmatrix} a_T \\ b_T \end{pmatrix} + c_T x \\ \qquad \text{und } \forall E \in \mathcal{E}(\Omega) \text{ mit } E = T_+ \cap T_- \text{ gilt } [p \cdot \nu_E]_E = 0 \end{array} \right\}.$$

Hierbei bezeichne $\mathcal{E}(\Omega)$ die Menge der inneren Kanten, ν_E die äußere Normale zu T_+ entlang E und $[\bullet]_E$ den Sprung entlang E, d.h. $[v]_E := v|_{T_+} - v|_{T_-}$.
Es sei $\mathcal{E}(T) = \{E_1, E_2, E_3\}$ die Menge der Kanten von T. Das RT-Finite-Element niedrigster Ordnung ist gegeben durch $(T, \mathrm{RT}_0(\{T\}), \mathcal{K})$ mit

$$\mathcal{K} := \left\{ \chi : \mathcal{C}^\infty(T;\mathbb{R}^2) \to \mathbb{R} \;\middle|\; \chi(p) = \fint_{E_j} p \cdot \nu_{E_j} \, ds \text{ für } j = 1,2,3 \right\}. \qquad \diamondsuit$$

Wir zeigen, dass dies in der Tat ein Finites Element ist, indem wir eine duale Basis angeben. Zu einer Ecke $z_j \in \mathcal{N}(T)$, wobei $\mathcal{N}(T)$ die Menge der drei Ecken von T bezeichne, mit gegenüberliegender Seite E_j definiere

$$\psi_j(x) := \frac{|E_j|}{2|T|}(x - z_j),$$

wobei $|E_j|$ das eindimensionale Maß von E_j bezeichne und $|T|$ das zwei-dimensionale Maß von T. Dann gilt für $E \in \mathcal{E}(T) \setminus \{E_j\}$, dass $\psi_j|_E(x) \cdot \nu_E = 0$ für alle $x \in E$, da $x - z_j$ tangential ist auf E. Außerdem gilt $\psi_j|_{E_j}(x) \cdot \nu_{E_j} = \frac{|E_j|}{2|T|} \cdot$ Höhe des Dreiecks $= 1$, also $\fint_{E_j} \psi_j \cdot \nu_{E_j} \, ds = 1$. Außerdem sind die drei Funktionen linear unabhängig und in $\mathrm{RT}_0(\{T\})$, bilden also eine duale Basis und das RT-Finite-Element ist tatsächlich ein Finites Element.

Aus der lokalen Basis können wir auch eine globale bilden. Dafür müssen wir uns für eine gegebene Triangulierung $\mathcal{T}$ auf eine Orientierung der Kanten festlegen. Für eine Innenkante $E \in \mathcal{E}(\Omega)$ seien $T_+, T_- \in \mathcal{T}$ die beiden Dreiecke mit $E = T_+ \cap T_-$. Wir legen uns nun durch die Wahl, welches der Dreiecke T_+ und welches T_- ist, auf eine Orientierung fest und definieren $\nu_E := (\nu_{T_+})|_E = -(\nu_{T_-})|_E$. Außerdem definieren wir für ein $v \in L^2(\Omega)$ mit $v|_T \in H^1(T)$ für alle $T \in \mathcal{T}$ den Sprung $[v]_E := v|_{T_+} - v|_{T_-}$. Für eine Randkante $E \in \mathcal{E}(\partial\Omega)$ sei $T_+ \in \mathcal{T}$ das eindeutige Dreieck mit $E \subseteq T_+$. Setze $\nu_E := (\nu_{T_+})|_E$ und $[v]_E := v|_{T_+}$.
Für die Definition der globalen Basis definiere $\sigma_{T,E} := 1$, wenn $T = T_+$ bezüglich E ist und $\sigma_{T,E} := -1$, wenn $T = T_-$ bezüglich E ist. Definiere dann

$$(\psi_E)|_T(x) := \begin{cases} \sigma_{T,E} \frac{|E|}{2|T|}(x - z_{T,E}) & \text{wenn } T \in \{T_+, T_-\}, \\ 0 & \text{sonst,} \end{cases}$$

wobei $z_{T,E} \in \mathcal{N}$ der Knoten in T ist, der gegenüber von E liegt. Obige Rechnungen ergeben dann, dass $\psi_E \cdot \nu_F = \delta_{EF}$ gilt für alle Kanten $E, F \in \mathcal{E}$. Insbesondere gilt also $[\psi_E \cdot \nu_F]_F = \psi_E|_{T_+} \cdot \nu_F - \psi_E|_{T_-} \cdot \nu_F = \psi_E|_{T_+} \cdot \nu_{T_+}|_E + \psi_E|_{T_-} \cdot \nu_{T_-}|_E = 0$ für alle $F \in \mathcal{E}(\Omega)$. Damit bildet $\{\psi_E\}_{E \in \mathcal{E}}$ eine globale, duale Basis von $\mathrm{RT}_0(\mathcal{T})$.
Die folgende Proposition definiert den globalen Interpolationsoperator und zeigt eine Stabilität und dass der Interpolationsoperator und die Divergenz kommutieren.

Proposition 5.19. *Es sei $p \in (H^1(\Omega))^2$ gegeben. Dann ist der globale Interpolationsoperator $I_{\mathrm{RT}} : (H^1(\Omega))^2 \to \mathrm{RT}_0(\mathcal{T})$ definiert durch*

$$I_{\mathrm{RT}} p := \sum_{E \in \mathcal{E}} \alpha_E \psi_E \qquad \textit{mit} \qquad \alpha_E = \fint_E p \cdot \nu_E \, ds.$$

Dann gilt

$$\operatorname{div} I_{\mathrm{RT}} p = \fint_T \operatorname{div} p \, dx,$$

d.h. das Diagramm

$$\begin{array}{ccc} (H^1_0(\Omega))^2 & \xrightarrow{\operatorname{div}} & L^2(\Omega) \\ I_{\mathrm{RT}} \big\downarrow & & \big\downarrow \Pi_0 \\ \mathrm{RT}_0(\mathcal{T}) & \xrightarrow{\operatorname{div}} & P_0(\mathcal{T}) \end{array}$$

kommutiert. Außerdem gilt

$$\|I_{\mathrm{RT}} p\|_{H(\operatorname{div},\Omega)} \lesssim \|p\|_{H^1(\Omega)}.$$

Bemerkung. Beachte, dass in der Definition $p \in (H^1(\Omega))^2$ vorausgesetzt ist und deswegen nach dem Spursatz $p|_E \in L^2(E)$ gilt. Damit ist α_E wohldefiniert. Für allgemeinere Funktionen $p \in H(\operatorname{div}, \Omega)$ ist $\fint_E p \cdot \nu_E \, ds$ hingegen *nicht* definiert. Dazu folgt noch ein kleiner Exkurs zu Spurräumen nach dem Beweis der Proposition. ♢

Beweis. Setze $p_{\mathrm{RT}} = I_{\mathrm{RT}} p$. Eine partielle Integration zeigt

$$\begin{aligned} \underbrace{\operatorname{div} p_{\mathrm{RT}}|_T}_{\in P_0(\mathcal{T})} &= \frac{1}{|T|} \int_T \operatorname{div} p_{\mathrm{RT}} \, dx = \frac{1}{|T|} \int_{\partial T} p_{\mathrm{RT}} \cdot \nu_T \, ds \\ &= \frac{1}{|T|} \int_{\partial T} p \cdot \nu_T \, ds = \frac{1}{|T|} \int_T \operatorname{div} p \, dx. \end{aligned}$$

Dies beweist die Kommutativität des Diagramms.
Außerdem gilt

$$\|p_{\mathrm{RT}}\|^2_{L^2(T)} \leq \sum_{E,F \in \mathcal{E}(T)} |\alpha_E||\alpha_F| \left| \int_\Omega \psi_E \cdot \psi_F \, dx \right|.$$

Es gilt mit einer Cauchy-Ungleichung

$$\begin{aligned} |\alpha_E| \leq \frac{1}{|E|} \int_E |p| \, ds &\leq |E|^{-1} \|p\|_{L^2(E)} \underbrace{\|1\|_{L^2(E)}}_{=|E|^{1/2}} \\ &\overset{\text{Spurungleichung}}{\leq} |E|^{-1/2} \left(h_T^{-1/2} \|p\|_{L^2(T)} + h_T^{1/2} \|\nabla p\|_{L^2(T)} \right) \\ &\overset{\mathcal{T} \text{ formregulär}}{\lesssim} h_T^{-1} \|p\|_{L^2(T)} + \|\nabla p\|_{L^2(T)}, \end{aligned}$$

wobei $T \in \mathcal{T}$ beliebig ist mit $E \in \mathcal{E}(T)$ und die Spurungleichung in einer Übungsaufgabe bewiesen wird. Wenn es für $E, F \in \mathcal{E}$ kein Dreieck gibt mit $E, F \in \mathcal{E}(T)$, dann ist $\int_\Omega \psi_E \cdot \psi_F \, dx = 0$. Ansonsten gilt für $E, F \in \mathcal{E}(T)$

$$\begin{aligned} |\psi_E(x)| \leq \frac{|E|}{2|T|} \underbrace{|x - z_{T,E}|}_{\leq \operatorname{diam}(T)} &\overset{\mathcal{T} \text{ formregulär}}{\lesssim} \frac{h_T^2}{|T|} \lesssim 1 \\ \Rightarrow \int_\Omega \psi_E \cdot \psi_F \, dx = \int_T \psi_E \cdot \psi_F \, dx &\lesssim |T|. \end{aligned}$$

Also folgt

$$\|p_{\mathrm{RT}}\|_{L^2(T)}^2 \lesssim |T|(h_T^{-1}\|p\|_{L^2(T)}^2 + \|\nabla p\|_{L^2(T)}^2) \lesssim \|p\|_{L^2(T)}^2 + h_T^2\|\nabla p\|_{L^2(T)}^2$$
$$\Rightarrow \|p_{\mathrm{RT}}\|_{L^2(\Omega)}^2 \lesssim \|p\|_{L^2(\Omega)}^2 + \sum_{T\in\mathcal{T}} h_T^2\|\nabla p\|_{L^2(T)}^2 \lesssim \|p\|_{H^1(\Omega)}^2.$$

Mit $\operatorname{div} p_{\mathrm{RT}} = \Pi_0 \operatorname{div} p$ folgt außerdem, dass

$$\|\operatorname{div} p_{\mathrm{RT}}\|_{L^2(\Omega)} \le \|\operatorname{div} p\|_{L^2(\Omega)}.$$

Also folgt insgesamt die behauptete Stabilität. □

Kleiner Exkurs zu Spurräumen
Es sei $T \in \mathcal{T}$. Definiere $H^{1/2}(\partial T) := \{v : \partial T \to \mathbb{R} \mid \exists \widetilde{v} \in H^1(T) \text{ mit } \widetilde{v}|_{\partial T} = v\}$. Der Spursatz zeigt, dass $H^{1/2}(\partial T) \subseteq L^2(\partial T)$ gilt. Auf $H^{1/2}(\partial T)$ ist durch

$$\|v\|_{H^{1/2}(\partial T)} := \inf_{\widetilde{v}\in H^1(T),\, \widetilde{v}|_{\partial T}=v} \|\widetilde{v}\|_{H^1(T)}$$

definiert. Für eine beliebige Funktion $p \in H(\operatorname{div},\Omega)$ kann ein Funktional $F_p \in H^{-1/2}(\partial T) := (H^{1/2}(\partial T))'$ definiert werden durch

$$\langle F_p, v\rangle := \int_T \widetilde{v} \operatorname{div} p \, dx + \int_T p\cdot\nabla\widetilde{v}\, dx \qquad \text{für alle } v \in H^{1/2}(\partial T) \text{ mit } \widetilde{v}|_{\partial T} = v.$$

Die Wohldefiniertheit folgt, denn wenn $\widetilde{v}, \widehat{v} \in H^1(T)$ mit $\widetilde{v}|_{\partial T} = \widehat{v}|_{\partial T} = v$, dann gilt $(\widetilde{v} - \widehat{v})|_{\partial T} = 0$ und

$$\int_T (\widetilde{v}-\widehat{v}) \operatorname{div} p\, dx = -\int_T p\cdot\nabla(\widetilde{v}-\widehat{v})\, dx \qquad \text{für alle } p \in H(\operatorname{div},\Omega),$$
$$\text{also} \quad \int_T \widetilde{v}\operatorname{div} p\, dx + \int_T p\cdot\nabla\widetilde{v}\, dx = \int_T \widehat{v}\operatorname{div} p\, dx + \int_T p\cdot\nabla\widehat{v}\, dx,$$

also hängt $\langle F_p, v\rangle$ nicht von der Wahl der Fortsetzung ab. Die Cauchy-Ungleichung zeigt die Beschränktheit von F_p.
Für eine glatte Funktion $p \in \mathcal{C}^1(\Omega)$ gilt

$$\int_{\partial T} v\, p\cdot\nu\, ds = \int_T v \operatorname{div} p\, dx + \int_T p\cdot\nabla v\, dx,$$

deshalb schreiben wir auch $p\cdot\nu := F_p$, wobei $p\cdot n$ als Funktional in $H^{-1/2}(\partial T)$ verstanden werden muss, $(p\cdot\nu)|_E$ ist *nicht* erklärt! Also ist $\int_{\partial T} p\cdot\nu\, ds := \int_{\partial T} 1p\cdot\nu\, ds = \int_T \operatorname{div} p\, dx$ erklärt, aber *nicht* $\int_E p\cdot\nu\, ds$. Damit ist der Interpolationsoperator $I_{\mathrm{RT}}p$ nicht auf ganz $H(\operatorname{div},\Omega)$ definiert!

Definition 5.20 (RT-FEM)**.** Die gemischte Raviart-Thomas-FEM niedrigster Ordnung sucht $(p_{\mathrm{RT}}, u_h) \in \mathrm{RT}_0(\mathcal{T}) \times P_0(\mathcal{T})$ mit

$$\int_\Omega p_{\mathrm{RT}}\cdot q_{\mathrm{RT}}\, dx + \int_\Omega u_h \operatorname{div} q_{\mathrm{RT}} = 0 \qquad \text{für alle } q_{\mathrm{RT}} \in \mathrm{RT}_0(\mathcal{T}),$$
$$\int_\Omega v_h \operatorname{div} p_{\mathrm{RT}}\, dx = -\int_\Omega f\, v_h\, dx \qquad \text{für alle } v_h \in P_0(\mathcal{T}). \qquad \diamondsuit$$

Bemerkung. Eine Übungsaufgabe zeigt, dass $\mathrm{RT}_k(\mathcal{T}) \subseteq H(\operatorname{div},\Omega)$ wegen der Stetigkeit in Normalen-Richtung. Also ist die RT-FEM eine konforme Methode. $\diamondsuit$

Bemerkung. Da die zweite Gleichung für alle Funktionen in $P_0(\mathcal{T})$ erfüllt ist, folgt punktweise $-\operatorname{div} p_{\mathrm{RT}} = \Pi_0 f$, wobei $\Pi_0 : L^2(\Omega) \to P_0(\mathcal{T})$ die L^2-Projektion auf stückweise konstante Funktionen bezeichnet. Für die exakte Lösung $(p, u) \in H(\operatorname{div}, \Omega) \times L^2(\Omega)$ aus Definition 5.16 wird die zweite Gleichung mit ganz $L^2(\Omega)$ getestet, weswegen hier auch $-\operatorname{div} p = f$ punktweise in $L^2(\Omega)$ gilt. ♢

Satz 5.21. *Es existiert eine eindeutige Lösung (p_{RT}, u_h) der RT-FEM und es gilt*

$$\|p - p_{\mathrm{RT}}\|_{H(\operatorname{div},\Omega)} + \|u - u_h\| \lesssim \inf_{q_{\mathrm{RT}} \in \mathrm{RT}_0(\mathcal{T})} \|p - q_{\mathrm{RT}}\|_{H(\operatorname{div},\Omega)} + \inf_{v_h \in P_0(\mathcal{T})} \|u - v_h\|.$$

Beweis. Es gilt $\mathrm{RT}_0(\mathcal{T}) \subseteq H(\operatorname{div}, \Omega)$ und $P_0(\mathcal{T}) \subseteq L^2(\Omega)$. Außerdem gilt für $q_{\mathrm{RT}}|_T = \begin{pmatrix} a_T \\ b_T \end{pmatrix} + c_T x$, dass

$$(\operatorname{div} q_{\mathrm{RT}})|_T = 2c_T \in P_0(T).$$

Also folgt, wenn $b(q_{\mathrm{RT}}, v_h) = 0$ für alle $v_h \in P_0(\mathcal{T})$, dass $\operatorname{div} q_{\mathrm{RT}} = 0$. Also ist a elliptisch auf dem Kern von b. Die eindeutige Lösbarkeit folgt aus Brezzi's Splitting Theorem, wenn wir die inf-sup-Bedingung

$$\|v_h\| \lesssim \sup_{q_{\mathrm{RT}} \in \mathrm{RT}_0(\mathcal{T})} \frac{\int_\Omega v_h \operatorname{div} q_{\mathrm{RT}}\, dx}{\|q_{\mathrm{RT}}\|_{H(\operatorname{div},\Omega)}} \qquad \text{für alle } v_h \in P_0(\mathcal{T})$$

gezeigt haben.
Dafür gehen wir ähnlich vor wie im Beweis der kontinuierlichen inf-sup-Bedingung. Es sei $w_h \in P_0(\mathcal{T})$. Als erstes vergrößern wir unser Gebiet Ω zu $\widetilde{\Omega}$ durch hinzufügen von Dreiecken derart, dass $\widetilde{\Omega}$ konvex ist, und setzen w_h durch 0 auf $\widetilde{\Omega}$ fort. Dann existiert nach Satz 4.18 eine Lösung $u \in H_0^1(\widetilde{\Omega}) \cap H^2(\widetilde{\Omega})$ mit

$$\begin{aligned} \int_{\widetilde{\Omega}} \nabla u \cdot \nabla v\, dx &= \int_\Omega w_h\, v\, dx \qquad \text{für alle } v \in H_0^1(\widetilde{\Omega}), \\ \text{und} \quad \|D^2 u\|_{L^2(\widetilde{\Omega})} &\lesssim \|w_h\|_{L^2(\widetilde{\Omega})} = \|w_h\|. \end{aligned}$$

Setze $p := -\nabla u$. Dann ist $p \in H^1(\widetilde{\Omega})$ und es gilt $\operatorname{div} p = w_h$. Setze nun $p_{\mathrm{RT}} = I_{\mathrm{RT}} p$. Dann gilt nach Proposition 5.19, dass $\operatorname{div} p_{\mathrm{RT}} = \Pi_0 \operatorname{div} p = \Pi_0 w_h = w_h$. Also gilt

$$\int_\Omega w_h \operatorname{div} p_{\mathrm{RT}}\, dx = \|w_h\|^2.$$

Außerdem gilt mit der Stabilität aus Proposition 5.19

$$\|p_{\mathrm{RT}}\|_{H(\operatorname{div},\Omega)} \lesssim \|p\|_{H^1(\Omega)}.$$

Da $p = -\nabla u$ und $\|D^2 u\| + \|\nabla u\| \lesssim \|w_h\|$, folgt

$$\frac{\int_\Omega w_h \operatorname{div} p_{\mathrm{RT}}\, dx}{\|p_{\mathrm{RT}}\|_{H(\operatorname{div},\Omega)}} \gtrsim \frac{\|w_h\|^2}{\|w_h\|} = \|w_h\|.$$

Also gilt die inf-sup-Bedingung (mit Konstante unabhängig von der Gitterweite) und damit existiert eine eindeutige Lösung und nach Korollar 5.13 gilt die Fehlerabschätzung. □

Nach der Poincaré-Ungleichung (bzw. dem Bramble-Hilbert-Lemma) und einer Transformation aufs Referenzelement gilt

$$\inf_{v_h\in P_0(\mathcal{T})}\|u-v_h\|\lesssim h\|u\|_{H^1(\Omega)}.$$

Für den Interpolationsoperator $I_{\mathrm{RT}}p$ für $p\in H^1(\Omega;\mathbb{R}^2)$ haben wir gezeigt, dass $\operatorname{div}(I_{\mathrm{RT}}p)|_T = \fint_T \operatorname{div} p\,dx$ gilt, also folgt mit $\operatorname{div} p = f$

$$\inf_{q_{\mathrm{RT}}\in\mathrm{RT}_0(\mathcal{T})}\|p-q_{\mathrm{RT}}\|_{H(\operatorname{div},\Omega)}\le\|p-I_{\mathrm{RT}}p\|+\underbrace{\inf_{f_h\in P_0(\mathcal{T})}\|f-f_h\|}_{\lesssim h\|f\|_{H^1(\Omega)}},$$

falls $f\in H^1(\Omega)$. Wir haben gezeigt, dass $\|p-I_{\mathrm{RT}}p\|_{L^2(T)}\lesssim\|p\|_{H^1(T)}$ und es gilt $I_{\mathrm{RT}}p|_T=p|_T$, falls $p|_T\in P_0(T;\mathbb{R}^2)\cap\mathrm{RT}_0(\{T\})$. Mit dem Bramble-Hilbert-Lemma und der Transformation aufs Referenzelement folgt

$$\|p-I_{\mathrm{RT}}p\|\lesssim h\|p\|_{H^1(\Omega)}\qquad\text{für } p\in H^1(\Omega;\mathbb{R}^2).$$

Insgesamt folgt

$$\|p-p_{\mathrm{RT}}\|_{H(\operatorname{div},\Omega)}+\|u-u_h\|\lesssim h(\|u\|_{H^2(\Omega)}+\|f\|_{H^1(\Omega)}).$$

Als nächstes wollen wir noch den folgenden Satz beweisen, der der Ausgangspunkt ist für so genannte äquilibrierte Fehlerschätzer.

Satz 5.22 (Zwei-Energien-Prinzip, Satz von Prager und Synge, Hyperkreis-Identität)**.** *Es sei $v\in H^1_0(\Omega)$ und $q\in H(\operatorname{div},\Omega)$ erfülle $\operatorname{div} q+f=0$. Dann gilt*

$$\|\nabla(u-v)\|^2+\|\nabla u-q\|^2=\|\nabla v-q\|^2.$$

Beweis. Da u die Lösung des Poisson-Problems ist und $\operatorname{div} q=f$, gilt

$$\int_\Omega\nabla(u-v)\cdot(\nabla u-q)\,dx=0.$$

Mit dem Satz von Pythagoras folgt die Behauptung. □

Bemerkung. Die Bedeutung des Satzes für die Fehlerschätzung ist die folgende: Ist $(p_{\mathrm{RT}},u_h)\in\mathrm{RT}_0(\mathcal{T})\times P_0(\mathcal{T})$ Lösung zur RT-FEM mit rechter Seite $f\in P_0(\mathcal{T})$, dann gilt $-\operatorname{div} p_{\mathrm{RT}}=f$. Es sei $v_h\in S^1_0(\mathcal{T})$ die Lösung zur konformen P_1-FEM, dann gilt

$$\underbrace{\|\nabla(u-v_h)\|^2}_{\text{Fehler zur }P_1\text{-FEM}}+\underbrace{\|\nabla u-p_{\mathrm{RT}}\|^2}_{\text{Fehler zur RT-FEM}}=\underbrace{\|p_{\mathrm{RT}}-\nabla v_h\|^2}_{\text{berechenbar}}.$$

Nachteil dieser Methode ist, dass für ein festes Gitter zwei diskrete Probleme gelöst werden müssen.
Ein anderer Ansatz ist, dass v_h aus (p_{RT},u_h) durch ein sogenanntes Post-Processing gewonnen wird, z.B. durch geeignete Mittelungen. Dabei sollte dann aber sichergestellt werden, dass $\|\nabla(u-v_h)\|\lesssim\|\nabla u-p_{\mathrm{RT}}\|$ gilt, da sonst der Fehler der RT-FEM überschätzt wird.
Der Vorteil dieser Art der Fehlerschätzung ist, dass es in der Hyperkreis-Identität keine nicht-bekannten Konstanten gibt. ♢

Das folgende Lemma zeigt, wie die lokalen Steifigkeitsmatrizen berechnet werden können.

Lemma 5.23. *Es seien $B_T, C_T \in \mathbb{R}^{3\times 3}$,*

$$(B_T)_{jk} := \int_T \psi_j \cdot \psi_k \, dx, \qquad (C_T)_{jk} := \begin{cases} \int_T \operatorname{div} \psi_j \, dx, & \text{wenn } j = k \\ 0 & \text{sonst} \end{cases}$$

die lokalen Steifigkeitsmatrizen. Definiere die Matrizen

$$M := \begin{pmatrix} 2 & 0 & 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 & 0 & 1 \\ 1 & 0 & 2 & 0 & 1 & 0 \\ 0 & 1 & 0 & 2 & 0 & 1 \\ 1 & 0 & 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 1 & 0 & 2 \end{pmatrix}, \qquad N := \begin{pmatrix} 0 & z_1 - z_2 & z_1 - z_3 \\ z_2 - z_1 & 0 & z_2 - z_3 \\ z_3 - z_1 & z_3 - z_2 & 0 \end{pmatrix}.$$

Dann gilt

$$B_T = \frac{1}{48\,|T|} C_T^\top N^\top M N C_T.$$

Der Beweis des Lemmas soll hier nur skizziert werden. Dafür benutzen wir noch die folgende Quadraturregel.

Lemma 5.24. *Es sei T ein Dreieck mit Seitenmittelpunkten E_j. Dann ist die Quadraturregel*

$$Q(p) := \frac{1}{3} \sum_{j=1}^{3} p(\operatorname{mid}(E_j))$$

exakt für Polynome zweiten Grades, das heißt, es gilt

$$Q(p) = \int_T p \, dx \qquad \text{für alle } p \in P_2(T).$$

Beweisidee. Dies kann für Basisfunktionen von $P_2(T)$ nachgerechnet werden. □

Beweisskizze von Lemma 5.23. Die Basisfunktionen sind von der Form

$$\psi_j = c_j(x - z_j).$$

Daher gilt

$$\operatorname{div} \psi_j = 2c_j.$$

Damit folgt

$$\int_T \psi_j \cdot \psi_k \, dx = \frac{\operatorname{div} \psi_j}{2} \frac{\operatorname{div} \psi_k}{2} \int_T (x - z_j) \cdot (x - z_k) \, dx.$$

Mit der Quadraturformel aus Lemma 5.24 und $\operatorname{mid}(E_j) = (z_{j-1} + z_{j+1})/2$ (wobei hier z_{j-1} und z_{j+1} modulo 3 verstanden werden soll) folgt

$$\begin{aligned} \int_T (x - z_j) \cdot (x - z_k) \, dx &= \frac{1}{3} \sum_{\ell=1}^{3} \left(\frac{z_\ell + z_{\ell+1}}{2} - z_j \right) \cdot \left(\frac{z_\ell + z_{\ell+1}}{2} - z_k \right) \\ &= \frac{1}{12} \sum_{\ell=1}^{3} (z_\ell - z_j + z_{\ell+1} - z_j) \cdot (z_\ell - z_k + z_{\ell+1} - z_k) \end{aligned}$$

Ausmultiplizieren ergibt

$$\int_T (x - z_j) \cdot (x - z_k) \, dx = \frac{1}{12} \sum_{\ell,m=1}^{3} (z_\ell - z_j) \cdot (z_m - z_k) + \frac{1}{12} \sum_{\ell=1}^{3} (z_\ell - z_j) \cdot (z_\ell - z_k).$$

Dies kann schließlich mit Matrizen in der Form wie in Lemma 5.23 geschrieben werden. □

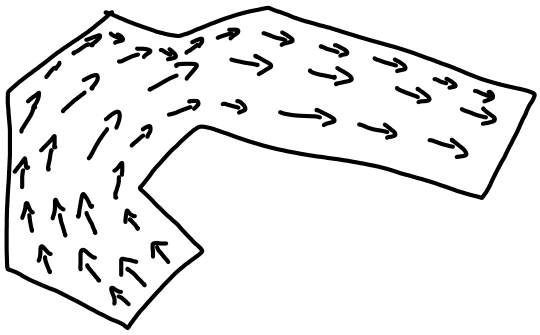

Abbildung 5.2: Ein Fluss durch ein Gebiet Ω.

5.3 Das Stokes-Problem

In diesem Kapitel soll es um die Lösung der stationären, linearen, inkompressiblen Stokes-Gleichungen gehen, die beschreiben, wie eine inkompressible Flüssigkeit durch ein Gebiet Ω strömt.
Wir betrachten ein beschränktes Lipschitz-Gebiet $\Omega \subseteq \mathbb{R}^d$, ein $f \in L^2(\Omega;\mathbb{R}^d)$ und $u_D \in H^{1/2}(\partial\Omega;\mathbb{R}^d)$. Gesucht ist ein Geschwindigkeitsfeld u und ein Druck p mit

$$\begin{aligned} -\Delta u - \nabla p &= f && \text{in } \Omega \\ \operatorname{div} u &= 0 && \text{in } \Omega \qquad \text{(Inkompressibilität)}, \\ u|_{\partial\Omega} &= u_D. \end{aligned}$$

Damit überhaupt Lösungen existieren können, muss gelten

$$\int_{\partial\Omega} u_D \cdot \nu \, ds = \int_{\partial\Omega} u \cdot \nu \, ds = \int_\Omega \operatorname{div} u \, dx = 0,$$

d.h. es muss so viel in das Gebiet hineinfließen, wie auch hinausfließt. Andererseits ist p nur bis auf Konstanten eindeutig. Die schwache Formulierung ist gegeben durch

Definition 5.25 (Stokes-Gleichungen)**.** Finde $u \in H^1(\Omega;\mathbb{R}^d)$ mit $u|_{\partial\Omega} = u_D$ und $p \in L^2(\Omega)/\mathbb{R} := \{q \in L^2(\Omega) \mid \int_\Omega q \, dx = 0\}$ mit

$$\begin{aligned} a(u,v) + b(v,p) &= \int_\Omega f \cdot v \, dx && \text{für alle } v \in H^1_0(\Omega;\mathbb{R}^d), \\ b(u,q) &= 0 && \text{für alle } q \in L^2(\Omega)/\mathbb{R}, \end{aligned}$$

wobei

$$\begin{aligned} a(u,v) &:= \int_\Omega \nabla u \cdot \nabla v \, dx, \\ b(v,q) &:= -\int_\Omega q \operatorname{div} v \, dx. \end{aligned}$$ ♢

Wie wir in der abstrakten Situation gesehen haben, ist dies äquivalent zur Minimierung von $a(u,u) - \int_\Omega f \cdot v \, dx$ unter der Nebenbedingung $\operatorname{div} u = 0$.

Bemerkung (lineare Elastizität)**.** In der linearen Elastizität wird eine Verschiebung $u \in H^1(\Omega;\mathbb{R}^d)$ gesucht mit

$$\begin{aligned} -\operatorname{div}\sigma &= f && \text{in } \Omega,\\ \sigma &= \mathbb{C}\varepsilon(u) && \text{in } \Omega,\\ u|_{\Gamma_D} &= 0,\\ \sigma\cdot\nu|_{\Gamma_N} &= g, \end{aligned}$$

wobei $\mathbb{C}$ der Elastizitätstensor ist, der für isotrope Materialien gegeben ist durch $\mathbb{C}A = 2\mu A + \lambda\operatorname{tr}(A)I_{d\times d}$. Dabei sind $\mu > 0$ und $\lambda > 0$ Parameter. Der linearisierte Greensche Verzerrungstensor $\varepsilon(u)$ ist der symmetrische Anteil des Gradienten, also $\varepsilon(u) = (\nabla u + \nabla u^\top)/2$. Der Parameter λ beschreibt die Inkompressibilität des Materials. Fassen wir die ersten beiden Gleichungen zusammen erhalten wir

$$f = -\operatorname{div}(2\mu\varepsilon(u) + \lambda\operatorname{div} uI_{d\times d}) = -2\mu\operatorname{div}\varepsilon(u) - \lambda\nabla\operatorname{div} u.$$

Für festes f und $\lambda\to\infty$ (d.h. dass das Material fast inkompressibel ist) muss also $\operatorname{div} u$ klein werden. Das Stokes-Problem kann auf diese Weise als das Grenzproblem zur linearen Elastizität für fast inkompressible Materialien interpretiert werden. ♢

Um die Existenz von Lösungen zum Stokes-Problem zu zeigen, müssen wir die folgende inf-sup-Bedingung zeigen.

Satz 5.26. *Es sei $\Omega\subseteq\mathbb{R}^d$ ein beschränktes, zusammenhängendes Lipschitz-Gebiet. Dann gilt für alle $q\in L^2(\Omega)/\mathbb{R}$*

$$\|q\| \lesssim \sup_{v\in H^1_0(\Omega;\mathbb{R}^d)} \frac{\int_\Omega q\operatorname{div} v\,dx}{\|\nabla v\|}.$$

Beweisskizze. Wir skizzieren hier nur den Beweis des Spezialfalls, dass $\Omega\subseteq\mathbb{R}^2$ konvex ist, siehe auch [BS08]. Für den allgemeinen Fall, siehe zum Beispiel [GR86].
Es sei $w\in H^1(\Omega)\cap(L^2(\Omega)/\mathbb{R})\cap H^2(\Omega)$ die eindeutige (schwache) Lösung zu $-\Delta w = q$ in Ω und $\partial w/\partial\nu = 0$ auf $\partial\Omega$ (homogenes Neumannproblem). Die Existenz von Lösungen zu diesem Problem folgt aus dem Satz von Lax-Milgram, da die Poincaré-Ungleichung für Funktionen aus $H^1(\Omega)\cap(L^2(\Omega)/\mathbb{R})$ gilt. Außerdem gilt (was wir hier nicht beweisen)

$$\|w\|_{H^2(\Omega)} \lesssim \|q\|.$$

Es sei $v_1 = -\nabla w$. Dann gilt $v_1\in H^1(\Omega;\mathbb{R}^d)$,

$$\operatorname{div} v_1 = q \quad\text{und}\quad \|v_1\|_{H^1(\Omega)} \lesssim \|q\|.$$

Außerdem gilt $v_1\cdot\nu = \nabla w\cdot\nu = 0$ auf $\partial\Omega$. Es sei τ der Tangentialvektor auf $\partial\Omega$. Dann kann man zeigen, dass ein $\psi\in H^2(\Omega)$ existiert mit

$$\begin{aligned} &\psi|_{\partial\Omega} = 0 \quad\text{und}\quad \partial\psi/\partial\nu|_{\partial\Omega} = -v_1|_{\partial\Omega}\cdot\tau\\ \text{und}\qquad &\|\psi\|_{H^2(\Omega)} \lesssim \|v_1\|_{H^1(\Omega)}. \end{aligned}$$

Setze $v_2 := \operatorname{Curl}\psi := \begin{pmatrix}-\partial\psi/\partial y & \partial\psi/\partial x\end{pmatrix}^\top = \begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}\nabla\psi$. Dann gilt

$$\begin{aligned} &v_2|_{\partial\Omega}\cdot\nu = \nabla\psi|_{\partial\Omega}\cdot\tau = 0 = -v_1|_{\partial\Omega}\cdot\nu,\\ &v_2|_{\partial\Omega}\cdot\tau = \nabla\psi|_{\partial\Omega}\cdot\nu = -v_1|_{\partial\Omega}\cdot\tau\\ \Rightarrow\quad &v_2|_{\partial\Omega} = -v_1|_{\partial\Omega}. \end{aligned}$$

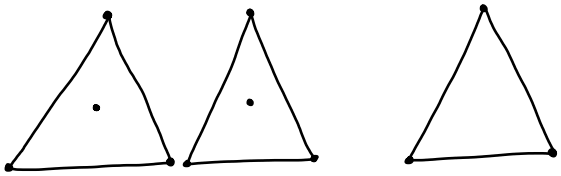

Abbildung 5.3: Schematische Darstellung der Mini-FEM.

Setze $u := v_1 + v_2$. Dann gilt $\operatorname{div} u = \operatorname{div} v_1 = q$ und $u|_{\partial\Omega} = 0$ und

$$\|u\|_{H^1(\Omega)} \leq \|v_1\|_{H^1(\Omega)} + \underbrace{\|v_2\|_{H^1(\Omega)}}_{\leq \|\psi\|_{H^2(\Omega)} \lesssim \|v_1\|_{H^1(\Omega)}} \lesssim \|v_1\|_{H^1(\Omega)} \lesssim \|q\|.$$

Also folgt

$$\frac{\int_\Omega q \operatorname{div} u \, dx}{\|u\|_{H^1(\Omega)}} = \frac{\|q\|^2}{\|u\|_{H^1(\Omega)}} \gtrsim \|q\|. \qquad \square$$

Korollar 5.27. *Es existiert eine eindeutige Lösung $(u,p) \in H^1(\Omega;\mathbb{R}^d) \times L^2(\Omega)/\mathbb{R}$ zum Stokes-Problem.*

Beweis. Die Bilinearform a ist elliptisch auf ganz $H^1_0(\Omega;\mathbb{R}^d)$. Also folgt aus Brezzi's Splitting Theorem mit der inf-sup-Bedingung aus Satz 5.26 die Behauptung. $\square$

Für eine Diskretisierung wählen wir Unterräume $V_h(\mathcal{T}) \subseteq H^1_0(\Omega;\mathbb{R}^d)$ und $Q_h(\mathcal{T}) \subseteq L^2(\Omega)/\mathbb{R}$. Eine natürliche Wahl für $Q_h(\mathcal{T})$ wäre ein Raum, der keine Stetigkeit zwischen Elementen fordert, z.B. $P_0(\mathcal{T})/\mathbb{R}$. Außerdem würde dies bedeuten, dass aus $0 = b(u_h, q_h) = \int_\Omega q_h \operatorname{div} u_h \, dx$ folgt, dass $\int_T \operatorname{div} u_h \, dx = 0$ gilt für alle $T \in \mathcal{T}$, die Divergenzfreiheit wäre also in einem lokalen Sinne erfüllt. Allerdings gilt

$$\left\{ v_h \in S^1_0(\mathcal{T};\mathbb{R}^2) \;\middle|\; \operatorname{div} v_h|_T = \fint_T \operatorname{div} v_h \, dx = 0 \right\} = \{0\}$$

auf einigen Triangulierungen (Beweis siehe Übungsaufgabe). Deshalb liefert die Wahl von $V_h(\mathcal{T}) = S^1_0(\mathcal{T};\mathbb{R}^d)$ und $Q_h(\mathcal{T}) = P_0(\mathcal{T})/\mathbb{R}$ *keine* sinnvolle Approximation.
Wir werden hier stattdessen das Mini-Element für $\Omega \subseteq \mathbb{R}^2$ betrachten. Definiere den Raum der Volumen-Bubbles

$$\mathcal{B}(\mathcal{T}) := \{\varphi \in H^1(\Omega) \mid \forall T = \operatorname{conv}\{a,b,c\} \in \mathcal{T} \, \exists \alpha_T \in \mathbb{R} \text{ mit } \varphi|_T = \alpha_T \lambda_a \lambda_b \lambda_c\}.$$

Setze

$$\begin{aligned} V_h(\mathcal{T}) &:= (S^1_0(\mathcal{T}) + \mathcal{B}(\mathcal{T}))^2 \subseteq H^1_0(\Omega;\mathbb{R}^2), \\ Q_h(\mathcal{T}) &:= S^1(\mathcal{T})/\mathbb{R}. \end{aligned}$$

Siehe auch Abbildung 5.3 für die schematische Darstellung.

Definition 5.28. Die Mini-FEM sucht $(u_h, p_h) \in V_h(\mathcal{T}) \times Q_h(\mathcal{T})$ mit

$$\begin{aligned} a(u_h, v_h) + b(v_h, p_h) &= \int_\Omega f \cdot v_h \, dx \qquad \text{für alle } v_h \in V_h(\mathcal{T}), \\ b(u_h, q_h) &= 0 \qquad\qquad\qquad\quad \text{für alle } q_h \in Q_h(\mathcal{T}). \end{aligned} \qquad \diamondsuit$$

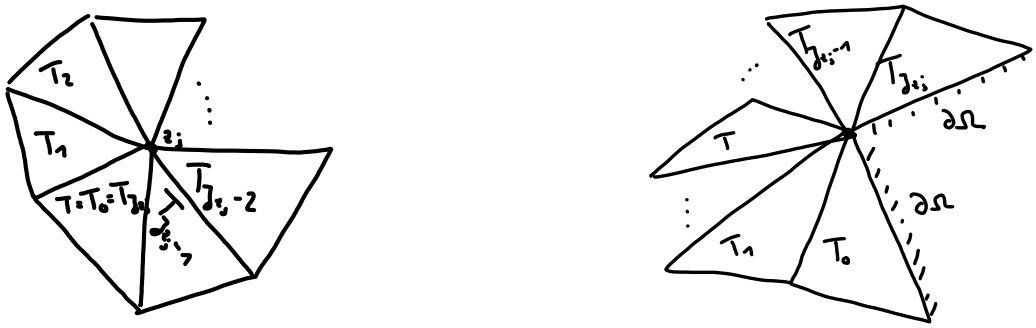

Abbildung 5.4: Situationen im Beweis von Proposition 5.30.

Lemma 5.29 (diskrete inf-sup-Bedingung für Mini-FEM)**.** *Es gilt*

$$\|q_h\| \lesssim \sup_{v_h \in V_h(\mathcal{T})} \frac{\int_\Omega q_h \operatorname{div} v_h \, dx}{\|\nabla v_h\|} \qquad \text{für alle } q_h \in Q_h(\mathcal{T}).$$

Bevor wir dieses Lemma beweisen, beweisen wir zunächst ein Hilfsresultat. Für eine Funktion $g \in P_1(\mathcal{T})$ definieren wir einen Mittelungsoperator $J_1 : P_1(\mathcal{T}) \to S^1_0(\mathcal{T})$ durch

$$(J_1 g)(z) = \sum_{T \in \mathcal{T}(z)} (g|_T)(z)/|\mathcal{T}(z)| \qquad \text{für alle } z \in \mathcal{N}(\Omega),$$

wobei $\mathcal{N}(\Omega) = \mathcal{N} \setminus \partial\Omega$ die Menge der inneren Knoten bezeichne und

$$\mathcal{T}(z) := \{T \in \mathcal{T} \mid z \in T\}.$$

Außerdem bezeichne $\mathcal{E}(\Omega)$ die Innenkanten und $\mathcal{E}(\partial\Omega)$ die Kanten auf dem Rand.

Proposition 5.30. *Sei $v_h \in P_1(\mathcal{T})$. Es gilt für alle $T \in \mathcal{T}$*

$$\|h_T^{-1}(v_h - J_1 v_h)\|^2_{L^2(T)} + \|\nabla(v_h - J_1 v_h)\|^2_{L^2(T)} \lesssim \sum_{z \in \mathcal{N} \cap T} \sum_{E \in \mathcal{E}(z) \setminus \mathcal{E}(\partial\Omega)} |E|^{-1} \|[v_h]_E\|^2_{L^2(E)},$$

wobei

$$\mathcal{E}(z) := \{E \in \mathcal{E} \mid z \in E\}.$$

Beweis. Es seien $\lambda_1, \lambda_2, \lambda_3$ die barycentrischen Koordinaten zu z_1, z_2, z_3 auf T. Dann gilt, da $(v_h - J_1 v_h)|_T = \sum_{j=1}^3 (v_h - J_1 v_h)(z_j)\lambda_j$, dass

$$\|\nabla(v_h - J_1 v_h)\|^2_{L^2(T)} = \sum_{j,k=1}^3 (v_h|_T - J_1 v_h)(z_j) \underbrace{\int_T \nabla\lambda_j \cdot \nabla\lambda_k \, dx}_{\lesssim 1} (v_h|_T - J_1 v_h)(z_k)$$

$$\overset{\text{Young-Ungleichung}}{\lesssim} \sum_{j=1}^3 |(v_h|_T - J_1 v_h)(z_j)|^2,$$

wobei die Young-Ungleichung $ab \le (a^2 + b^2)/2$ besagt und zu $0 \le (a-b)^2$ äquivalent ist. Es gilt

$$|(v_h - J_1 v_h)(z_j)|^2 = |\mathcal{T}(z_j)|^{-2} \left| \sum_{K \in \mathcal{T}(z)} (v_h|_T - v_h|_K)(z_j) \right|^2.$$

Wenn $z_j \in \mathcal{N}(\Omega)$, dann seien $T_0, \dots, T_{J_{z_j}} \in \mathcal{T}$ mit $T_0 = T_{J_{z_j}} = T$ und $T_k \cap T_{k+1} \in \mathcal{E}(z)$ für alle $k = 0, \dots, J_k - 1$, siehe auch Abbildung 5.4. Falls $z_j \in \partial\Omega$, dann seien $T_0, \dots, T_{J_{z_j}}$ wie in Abbildung 5.4. Dann gilt mit einer Teleskopsumme und einer Young-Ungleichung

$$\left| \sum_{K \in \mathcal{T}(z)} (v_h|_T - v_h|_K)(z_j) \right|^2 \lesssim \sum_{k=0}^{J_{z_j}-1} \left| (v_h|_{T_k} - v_h|_{T_{k+1}})(z_j) \right|^2.$$

Für eine Seite $E = \operatorname{conv}\{y_1, y_2\}$ und $f \in P_1(E)$ gilt

$$\int_E |f|^2 \, ds = \begin{pmatrix} f(y_1) & f(y_2) \end{pmatrix} A \begin{pmatrix} f(y_1) \\ f(y_2) \end{pmatrix}$$

mit

$$A_{jk} = \int_E \lambda_j \lambda_k \, ds$$

für barycentrische Koordinaten λ_j zu y_j. Es gilt $A_{11} = A_{22} = |E|/3$ und $A_{12} = A_{21} = |E|/6$ (kann durch Berechnung auf Referenzseite und Transformation gezeigt werden). Nach dem Gerschgorinschen Kreissatz gilt also, dass die Eigenwerte von A größer oder gleich $(\frac{1}{3} - \frac{1}{6})|E| = |E|/6$ sind. Also

$$|f(y_1)|^2 + |f(y_2)|^2 \leq 6|E|^{-1} \, \|f\|^2_{L^2(E)}.$$

Mit $y_1 = z_j$ und $E_k = T_k \cap T_{k+1}$ und y_2 mit $E_k = \operatorname{conv}\{y_1, y_2\}$ und $f = (v_h|_{T_{k+1}} - v_h|_{T_k})|_{E_k} = [v_h]_{E_k}$ folgt

$$|(v_h|_{T_{k+1}} - v_h|_{T_k})(z_j)|^2 \leq |f(y_1)|^2 + |f(y_2)|^2 \leq 6|E_k|^{-1} \, \|[v_h]_{E_k}\|^2_{L^2(E_k)}.$$

Insgesamt also

$$\begin{aligned}
\|\nabla(v_h|_T - J_1 v_h)\|^2_{L^2(T)} &\lesssim \sum_{j=1}^{3} |(v_h|_T - J_1 v_h)(z_j)|^2 \lesssim \sum_{j=1}^{3} |\mathcal{T}(z_j)|^{-2} \sum_{k=0}^{J_{z_j}-1} \left| (v_h|_{T_k} - v_h|_{T_{k+1}})(z_j) \right|^2 \\
&\lesssim \sum_{j=1}^{3} \underbrace{|\mathcal{T}(z_j)|^{-2}}_{\leq 1} \sum_{k=0}^{J_{z_j}-1} |E_k|^{-1} \, \|[v_h]_{E_k}\|^2_{L^2(E_k)} \\
&\leq \sum_{z \in \mathcal{N} \cap T} \sum_{E \in \mathcal{E}(z) \setminus \mathcal{E}(\partial\Omega)} |E|^{-1} \, \|[v_h]_E\|^2_{L^2(E)}.
\end{aligned}$$

Für die L^2-Abschätzung kann man genauso vorgehen und mit

$$\int_T \lambda_j \lambda_k \, dx \leq |T| \quad \text{statt} \quad \int_T \nabla\lambda_j \cdot \nabla\lambda_k \, dx \leq 1$$

folgt die Behauptung. □

Lemma 5.31. *Es sei $v \in H^1(\Omega)$. Dann gilt für alle $E \in \mathcal{E}(\Omega)$ mit $E = T_+ \cap T_-$ und $\omega_E = T_+ \cup T_-$, dass*

$$|E|^{-1} \|[\Pi_0 v]_E\|^2_{L^2(E)} \lesssim \|\nabla v\|^2_{L^2(\omega_E)},$$

wobei $\Pi_0 : L^2(\Omega) \to P_0(\mathcal{T})$ die L^2-Projektion auf stückweise Konstante bezeichne.

Beweis. Da $v|_E \in L^2(E)$ und v stetig auf E im Sinne der Spur, gilt

$$\begin{aligned}|E|^{-1}\|[\Pi_0 v]_E\|_{L^2(E)}^2 &= |E|^{-1}\|[\Pi_0 v - v]_E\|_{L^2(E)}^2 \\ &\leq 2|E|^{-1}\|(\Pi_0 v - v)|_{T_+}\|_{L^2(E)}^2 + 2|E|^{-1}\|(\Pi_0 v - v)|_{T_-}\|_{L^2(E)}^2.\end{aligned}$$

Mit der Spurungleichung folgt

$$\|(\Pi_0 v - v)|_{T_+}\|_{L^2(E)}^2 \lesssim h_T^{-1}\|\Pi_0 v - v\|_{L^2(T_+)}^2 + h_T\|\underbrace{\nabla(\Pi_0 v - v)|_{T_+}}_{=\nabla v}\|_{L^2(T_+)}^2.$$

Mit einer Poincaré-Ungleichung folgt wegen $\int_{T_+}(v - \Pi_0 v)\,dx = 0$

$$\|\Pi_0 v - v\|_{L^2(T_+)}^2 \lesssim h_T^2\|\nabla(\Pi_0 v - v)|_{T_+}\|_{L^2(T_+)}^2 = h_T^2\|\nabla v\|_{L^2(T_+)}^2.$$

Also insgesamt

$$|E|^{-1}\|[\Pi_0 v]_E\|_{L^2(E)}^2 \lesssim \underbrace{h_T|E|^{-1}}_{\approx 1}\|\nabla v\|_{L^2(\omega_E)}.$$ □

Korollar 5.32 (Approximations- und Stabilitätseigenschaften der Quasi-Interpolation)**.** *Der Operator $I_h : H_0^1(\Omega) \to S_0^1(\mathcal{T})$, der durch $I_h = J_1 \circ \Pi_0$ gegeben ist, erfüllt*

$$\|h_T^{-1}(v - I_h v)\|_{L^2(T)} + \|\nabla I_h v\|_{L^2(T)} \lesssim \|\nabla v\|_{L^2(\Omega_T)}, \tag{5.5}$$

wobei $\Omega_T := \bigcup\{K \in \mathcal{T} \mid K \cap T \neq \emptyset\}$. Wegen der endlichen Überlappung der Ω_T folgt

$$\|h_{\mathcal{T}}^{-1}(v - I_h v)\| + \|\nabla I_h v\| \lesssim \|\nabla v\|, \tag{5.6}$$

wobei $h_{\mathcal{T}} \in P_0(\mathcal{T})$ die stückweise konstante Gitterweite ist, also $h_{\mathcal{T}}|_T = h_T$ für alle $T \in \mathcal{T}$.

Beweis. Die erste Abschätzung folgt aus Proposition 5.30 und Lemma 5.31. Die zweite folgt, da jedes $K \in \mathcal{T}$ nur für endlich viele $T \in \mathcal{T}$ in Ω_T vorkommt. Die Anzahl ist beschränkt in Abhängigkeit vom minimalen Winkel in $\mathcal{T}$, aber unabhängig von der Gitterweite. □

Bemerkung. Operatoren wie I_h werden auch Quasi-Interpolation genannt, da sie auf ganz $H_0^1(\Omega)$ definiert sind. Es gibt viele solcher Operatoren, zum Beispiel die Clément-Quasi-Interpolation $I_{\mathrm{Cl}}v \in S_0^1(\mathcal{T})$, die gegeben ist durch

$$(I_{\mathrm{Cl}}v)(z) = \fint_{\omega_z} v\,dx \qquad \text{für alle } z \in \mathcal{N}(\Omega)$$

und $\omega_z := \bigcup\{T \in \mathcal{T} \mid z \in T\}$. ♢

Beweis von Lemma 5.29. Wir zeigen die Existenz eines Fortin-Interpolationsoperators (siehe Satz 5.14). Es sei $v \in H_0^1(\Omega;\mathbb{R}^2)$ gegeben. Definiere $v_h \in V_h(\mathcal{T}) = (S_0^1(\mathcal{T}) + \mathcal{B}(\mathcal{T}))^2$ durch $v_h := v_c + v_b$, wobei $v_c \in S_0^1(\mathcal{T};\mathbb{R}^2)$ und $v_b \in \mathcal{B}(\mathcal{T})^2$ definiert sind durch $v_c := I_h v$ und für alle $T \in \mathcal{T}$ sei $v_b|_T = \alpha_T\lambda_a\lambda_b\lambda_c$ mit

$$\alpha_T = \frac{\int_T (v - v_c)\,dx}{\int_T \lambda_a\lambda_b\lambda_c\,dx} = \underbrace{(2|T|/5!)^{-1}}_{=60/|T|}\int_T (v - v_c)\,dx.$$

Dann gilt für alle $q_h \in S^1(\mathcal{T})/\mathbb{R} \subseteq H^1(\Omega)$

$$\begin{aligned}\int_\Omega q_h \operatorname{div} v_h \, dx &= -\int_\Omega v_h \cdot \nabla q_h \, dx \overset{\nabla q_h \in P_0(\mathcal{T};\mathbb{R}^2)}{=} -\sum_{T\in\mathcal{T}} \left(\int_T v_c \, dx + \int_T v_b \, dx \right) \cdot \nabla q_h \\ &= \sum_{T\in\mathcal{T}} \left(-\int_T v_c \, dx - \int_T (v - v_c) \, dx \right) \cdot \nabla q_h \\ &= -\int_\Omega v \cdot \nabla q_h \, dx = \int_\Omega q_h \operatorname{div} v \, dx.\end{aligned}$$

Außerdem gilt

$$\|\nabla v_h\|^2 \leq 2\big(\|\nabla v_c\|^2 + \|\nabla v_b\|^2\big)$$

und mit Korollar 5.32

$$\|\nabla v_c\| \lesssim \|\nabla v\|.$$

Außerdem folgt mit Korollar 5.32

$$\int_T |v - v_c| \, dx \leq |T|^{1/2} \|v - J_1(\Pi_0 v)\|_{L^2(T)} \lesssim |T|^{1/2} h_T \|\nabla v\|_{L^2(\Omega_T)},$$

also

$$|\alpha_T| \leq \frac{60}{|T|} \int_T |v - v_c| \, dx \lesssim \|\nabla v\|_{L^2(\Omega_T)}.$$

Also gilt wegen $\nabla(\lambda_a\lambda_b\lambda_c) = \lambda_a\lambda_b\nabla\lambda_c + \lambda_a\lambda_c\nabla\lambda_b + \lambda_b\lambda_c\nabla\lambda_a$ mit $|\lambda_a| \leq 1$ und $|\nabla\lambda_a| \lesssim h_T^{-1}$, dass

$$\|\nabla v_b\|_{L^2(T)} = |\alpha_T| \underbrace{\|\nabla(\lambda_a\lambda_b\lambda_c)\|_{L^2(T)}}_{\lesssim 1} \lesssim \|\nabla v\|_{L^2(\Omega_T)}.$$

Aus der endlichen Überlappung von Ω_T folgt schließlich

$$\|\nabla v_b\| \lesssim \|\nabla v\|,$$

und damit

$$\|\nabla v_h\| \lesssim \|\nabla v\|.$$

Mit Satz 5.14 folgt, dass die diskrete inf-sup-Bedingung erfüllt ist. □

Korollar 5.33. *Es existiert eine eindeutige Lösung $(u_h, p_h) \in V_h(\mathcal{T}) \times Q_h(\mathcal{T})$ zur Mini-FEM und es gilt*

$$\|\nabla(u - u_h)\| + \|p - p_h\| \lesssim \inf_{v_h \in V_h(\mathcal{T})} \|\nabla(u - v_h)\| + \inf_{q_h \in Q_h(\mathcal{T})} \|p - q_h\|.$$

Ist $u \in H^2(\Omega)$ und $p \in H^1(\Omega)$, dann konvergiert die Mini-FEM also mit Rate h.

Die folgende Proposition berechnet die zur Implementierung nötigen Größen.

Proposition 5.34. *Es sei $T \in \mathcal{T}$ und $b_T = \lambda_a\lambda_b\lambda_c$ bezeichne die zugehörige Volumen-Bubble. Außerdem bezeichne φ_z die Standard P_1 Basisfunktion zu einem Knoten $z \in \mathcal{N}$. Dann gilt*

$$\begin{aligned}
\int_T \nabla\varphi_z \cdot \nabla b_T \, dx &= 0,\\
\int_T \nabla b_T \cdot \nabla b_T \, dx &= \frac{|T|}{180}(|\nabla\lambda_a|^2 + |\nabla\lambda_b|^2 + |\nabla\lambda_c|^2),\\
\int_T \varphi_z \operatorname{div}(\varphi_y e_j) \, dx &= \frac{|T|}{3} \operatorname{div}(\varphi_y e_j),\\
\int_T \varphi_z \operatorname{div}(b_T e_j) \, dx &= -\frac{|T|}{60} e_j \cdot \nabla\varphi_z.
\end{aligned}$$

Beweis. Die erste Gleichung folgt aus einer partiellen Integration und aus $\nabla\varphi_z \in P_0(\mathcal{T};\mathbb{R}^2)$ und $b_T|_{\partial T} = 0$. Die dritte Formel folgt aus $\int_T \varphi_z \, dx = 1/3$. Die vierte folgt aus einer partiellen Integration und der Integrationsformel aus Proposition 4.53.
Für die zweite Formel berechnen wir $\nabla b_T = \lambda_a\lambda_b\nabla\lambda_c + \lambda_a\lambda_c\nabla\lambda_b + \lambda_c\lambda_b\nabla\lambda_a$. Da die Integrationsformel aus Proposition 4.53 nicht von der Wahl von a, b, c abhängt, ergibt ausmultiplizieren dann

$$\begin{aligned}
\int_T \nabla b_T \cdot \nabla b_T \, dx = \big(|\nabla\lambda_a|^2 + |\nabla\lambda_b|^2 + |\nabla\lambda_c|^2\big) \int_T \lambda_a^2\lambda_b^2 \, dx\\
+ 2\big(\nabla\lambda_a \cdot \nabla\lambda_b + \nabla\lambda_a \cdot \nabla\lambda_c + \nabla\lambda_c \cdot \nabla\lambda_b\big) \int_T \lambda_a\lambda_b\lambda_c^2 \, dx.
\end{aligned}$$

Da $\nabla\lambda_a + \nabla\lambda_b + \nabla\lambda_c = 0$, folgt

$$\begin{aligned}
&\nabla\lambda_a \cdot \nabla\lambda_b + \nabla\lambda_a \cdot \nabla\lambda_c + \nabla\lambda_c \cdot \nabla\lambda_b\\
&\quad = -|\nabla\lambda_a|^2 - \nabla\lambda_a \cdot \nabla\lambda_c - |\nabla\lambda_c|^2 - \nabla\lambda_b \cdot \nabla\lambda_c - |\nabla\lambda_b|^2 - \nabla\lambda_a \cdot \nabla\lambda_b.
\end{aligned}$$

Hieraus ergibt sich

$$2\big(\nabla\lambda_a \cdot \nabla\lambda_b + \nabla\lambda_a \cdot \nabla\lambda_c + \nabla\lambda_c \cdot \nabla\lambda_b\big) = -\big(|\nabla\lambda_a|^2 + |\nabla\lambda_b|^2 + |\nabla\lambda_c|^2\big).$$

Mit der Integrationsformel aus Proposition 4.53 folgt

$$\int_T \lambda_a^2\lambda_b^2 \, dx = \frac{|T|}{90} \qquad \text{und} \qquad \int_T \lambda_a\lambda_b\lambda_c^2 \, dx = \frac{|T|}{180}.$$

Daraus folgt die Behauptung. □

6 A-Posteriori-Analysis

In diesem Kapitel werden wir A-Posteriori-Fehlerabschätzungen beweisen. Bisher haben wir A-Priori-Fehlerabschätzungen bewiesen. Dies sind Fehlerabschätzungen, die Informationen der exakten Lösung benutzen, wie zum Beispiel $u \in H^2(\Omega)$, aber *keine* Informationen der diskreten Lösung. Für die A-Posteriori-Abschätzungen in diesem Kapitel werden wir Fehlerschätzer definieren, die von der diskreten Lösung abhängen. Wir zeigen, dass die Fehlerschätzer effizient und zuverlässig sind, also bis auf Konstanten äquivalent zum Fehler.

6.1 A-Posteriori-Analysis für die P_1-FEM für das PMP

Wir betrachten in diesem Kapitel das Poisson-Problem mit der Diskretisierung aus Definition 4.48. Wir nehmen der Einfachheit halber $\Gamma_D = \partial\Omega$ an.

Definition 6.1 (Residuum)**.** Das Residuum $\mathrm{Res} \in H^{-1}(\Omega) := (H^1_0(\Omega))'$ ist definiert durch

$$\mathrm{Res}(v) := a(u - u_h, v) = \int_\Omega f\, v\, dx - a(u_h, v) \qquad \text{für alle } v \in H^1_0(\Omega),$$

wobei $a(u,v) := \int_\Omega \nabla u \cdot \nabla v\, dx$. ♢

Bemerkung. Das Residuum ist der Riesz-Darsteller des Fehlers im Dualraum, wenn $H^1_0(\Omega)$ mit Skalarprodukt a betrachtet wird. ♢

Proposition 6.2. *Es gilt*

$$\|\mathrm{Res}\|_{H^{-1}(\Omega)} = \|\nabla(u - u_h)\|.$$

Bemerkung. Da die Riesz-Abbildung eine Isometrie ist, folgt die Proposition. Der folgende Beweis ist direkt. ♢

Beweis von Proposition 6.2. Es sei $v \in H^1_0(\Omega)$. Dann gilt

$$\mathrm{Res}(v) = a(u - u_h, v) \leq \|\nabla(u - u_h)\|\, \|\nabla v\|,$$

also

$$\|\mathrm{Res}\|_{H^{-1}(\Omega)} = \sup_{v \in H^1_0(\Omega)} \frac{\mathrm{Res}(v)}{\|\nabla v\|} \leq \|\nabla(u - u_h)\|.$$

Andererseits folgt mit $v = u - u_h$, dass $\mathrm{Res}(v) = \|\nabla(u - u_h)\|^2$, also $\|\mathrm{Res}\|_{H^{-1}(\Omega)} = \|\nabla(u - u_h)\|$. □

Für ein festes $v \in H^1_0(\Omega)$ lässt sich $\mathrm{Res}(v) = \int_\Omega f\, v\, dx - a(u_h, v)$ berechnen, ohne dass die Kenntnis von u nötig wäre. Trotzdem lässt sich $\|\mathrm{Res}\|_{H^{-1}(\Omega)}$ nicht berechnen.

Definition 6.3 (residualer Fehlerschätzer für konforme P_1-FEM)**.** Für ein gegebenes $u_h \in S^1_0(\mathcal{T})$ definiere für alle $T \in \mathcal{T}$

$$\eta(T) := \sqrt{\|h_T f\|^2_{L^2(T)} + h_T \sum_{E \in \mathcal{E}(T) \cap \mathcal{E}(\Omega)} \|[\nabla u_h \cdot \nu_E]_E\|^2_{L^2(E)}}$$

und

$$\eta := \sqrt{\sum_{T \in \mathcal{T}} \eta^2(T)}.$$ ♢

Wir wollen zeigen, dass η äquivalent zum Fehler ist. Dazu betrachten wir den Quasi-Interpolationsoperator I_h aus Korollar 5.32.

Satz 6.4 (Zuverlässigkeit des Fehlerschätzers)**.** *Es gilt*

$$\|\nabla(u-u_h)\| \lesssim \eta.$$

Beweis. Es sei $v \in H^1_0(\Omega)$. Dann gilt für ein beliebiges $v_h \in S^1_0(\mathcal{T})$ wegen der Galerkin-Orthogonalität

$$\operatorname{Res}(v) = \int_\Omega f\, v\, dx - a(u_h, v) = \underbrace{\int_\Omega f\,(v-v_h)\, dx}_{=:(1)} \underbrace{- \int_\Omega \nabla u_h \cdot \nabla(v-v_h)\, dx}_{=:(2)}.$$

Es bezeichne $I_h : H^1_0(\Omega) \to S^1_0(\mathcal{T})$ den Quasi-Interpolationsoperator aus Korollar 5.32. Mit $v_h := I_h v$ folgt für den ersten Term mit (5.6)

$$(1) = \int_\Omega h_\mathcal{T} f h_\mathcal{T}^{-1}(v-v_h)\, dx \le \|h_\mathcal{T} f\|\, \underbrace{\|h_\mathcal{T}^{-1}(v-v_h)\|}_{\lesssim \|\nabla v\|}.$$

Eine stückweise partielle Integration zeigt

$$\begin{aligned}(2) = -\sum_{T\in\mathcal{T}} \int_{\partial T} (v-v_h)\nabla u_h \cdot \nu_T\, ds &= \sum_{E\in\mathcal{E}(\Omega)} \int_E -(v-v_h)\,[\nabla u_h \cdot \nu_E]_E\, ds \\ &\le \sum_{E\in\mathcal{E}(\Omega)} |E|^{-1/2}\|v-v_h\|_{L^2(E)} |E|^{1/2} \|[\nabla u_h \cdot \nu_E]_E\|_{L^2(E)}.\end{aligned}$$

Mit der Spurungleichung und $|E| \approx h_T$, wenn $E \in \mathcal{E}(T)$ gilt für $T_E \in \mathcal{T}$ mit $E \in \mathcal{E}(T_E)$

$$|E|^{-1/2}\|v-v_h\|_{L^2(E)} \lesssim h_{T_E}^{-1}\|v-v_h\|_{L^2(T_E)} + \|\nabla(v-v_h)\|_{L^2(T_E)}.$$

Mit einer Cauchy-Ungleichung im $\mathbb{R}^{|\mathcal{E}(\Omega)|}$ folgt

$$\begin{aligned}&\sum_{E\in\mathcal{E}(\Omega)} |E|^{-1/2}\|v-v_h\|_{L^2(E)} |E|^{1/2} \|[\nabla u_h \cdot \nu_E]_E\|_{L^2(E)} \\ &\qquad \lesssim \left(\sum_{E\in\mathcal{E}(\Omega)} |E| \|[\nabla u_h \cdot \nu_E]_E\|^2_{L^2(E)} \right)^{1/2} \\ &\qquad\qquad \times \left(\sum_{E\in\mathcal{E}(\Omega)} \left(h_{T_E}^{-2}\|v-v_h\|^2_{L^2(T_E)} + \|\nabla(v-v_h)\|^2_{L^2(T_E)} \right) \right)^{1/2}\end{aligned}$$

Wegen der endlichen Überlappung gilt für den zweiten Term

$$\begin{aligned}\left(\sum_{E\in\mathcal{E}(\Omega)} \left(h_{T_E}^{-2}\|v-v_h\|^2_{L^2(T_E)} + \|\nabla(v-v_h)\|^2_{L^2(T_E)} \right) \right)^{1/2} &\lesssim \|h_\mathcal{T}^{-1}(v-v_h)\| + \|\nabla(v-v_h)\| \\ &\overset{(5.6)}{\lesssim} \|\nabla v\|.\end{aligned}$$

Daraus folgt

$$|\operatorname{Res}(v)| \lesssim \eta\, \|\nabla v\|$$

und daraus die Behauptung. □

Satz 6.5 (lokale Effizienz des Fehlerschätzers)**.** *Für alle $T \in \mathcal{T}$ gilt*

$$\eta(T) \lesssim \|\nabla(u-u_h)\|_{L^2(\omega_T)} + \operatorname{osc}(f, \mathcal{T}(\omega_T)),$$

wobei $\omega_T := \operatorname{int}\left(\bigcup\{K \in \mathcal{T} \mid K \cap T \in \mathcal{E}\}\right)$ *und* $\operatorname{osc}(f, \mathcal{T}(\omega_T)) := \|h_{\mathcal{T}}(f - \Pi_0 f)\|_{L^2(\omega_T)}$.

Beweis. Es sei $T = \operatorname{conv}\{a,b,c\} \in \mathcal{T}$. Definiere die Bubble-Funktion $\flat_T := 27\lambda_a\lambda_b\lambda_c$ und $\varphi_T := \Pi_0 f\, \flat_T$. Dann gilt

$$|T|\,|\Pi_0 f|^2 \approx \int_T \Pi_0 f\, \varphi_T\, dx = \int_T f\, \varphi_T\, dx + \int_T (\Pi_0 f - f)\varphi_T\, dx. \tag{6.1}$$

Für den zweiten Term gilt

$$\int_T (\Pi_0 f - f)\varphi_T\, dx \le \|h_T(f-\Pi_0 f)\|_{L^2(T)}\, \underbrace{\|h_T^{-1}\varphi_T\|_{L^2(T)}}_{\lesssim |\Pi_0 f|}.$$

Da ∇u_h stückweise konstant ist und φ_T auf dem Rand von T verschwindet, folgt, dass $\int_T \nabla u_h \cdot \nabla \varphi_T\, dx = 0$. Für den ersten Term von (6.1) folgt damit

$$\int_T f\, \varphi_T\, dx = \int_T \nabla u \cdot \nabla \varphi_T\, dx = \int_T \nabla(u-u_h)\cdot\nabla\varphi_T\, dx \le \|\nabla(u-u_h)\|_{L^2(T)}\, \underbrace{\|\nabla\varphi_T\|_{L^2(T)}}_{\lesssim |\Pi_0 f|}.$$

Damit folgt

$$\|h_T f\|_{L^2(T)} \le \|h_T(f-\Pi_0 f)\|_{L^2(T)} + \underbrace{h_T|T|^{1/2}|\Pi_0 f|}_{\lesssim \|\nabla(u-u_h)\|_{L^2(T)} + \|h_T(f-\Pi_0 f)\|_{L^2(T)}}.$$

Für die Abschätzung des Sprung-Terms sei $E \in \mathcal{E}(T)$ und $\flat_E := 6\lambda_a\lambda_b$ für $E = \operatorname{conv}\{a,b\}$ und $\varphi_E := [\nabla u_h \cdot \nu_E]_E \flat_E$. Dann gilt

$$\begin{aligned}
|E|\|[\nabla u_h\cdot\nu_E]_E\|^2 &\approx \int_E [\nabla u_h\cdot\nu_E]_E \varphi_E\, ds \overset{\text{partielle Integration}}{=} \int_{\omega_E} \nabla u_h \cdot \nabla\varphi_E\, dx \\
&\overset{u \text{ Lösung}}{=} \int_{\omega_E} \nabla(u_h - u)\cdot\nabla\varphi_E\, dx + \int_\Omega f\,\varphi_E\, dx \\
&\le \|\nabla(u-u_h)\|_{L^2(\omega_E)} \underbrace{\|\nabla\varphi_E\|_{L^2(\omega_E)}}_{\approx |[\nabla u_h\cdot\nu_E]_E|} + \|h_{\mathcal{T}} f\|_{L^2(\omega_E)} \underbrace{\|h_{\mathcal{T}}^{-1}\varphi_E\|_{L^2(\omega_E)}}_{\approx |[\nabla u_h\cdot\nu_E]_E|}.
\end{aligned}$$

Mit der Abschätzung für $\|h_T f\|_{L^2(\omega_T)}$ folgt die Behauptung. □

Bemerkung. Für den Beweis der Effizienz muss u_h nicht das diskrete Problem lösen, für die Zuverlässigkeit wurde aber benutzt, dass u_h das diskrete Problem löst. ♢

Bemerkung. Die Effizienz gilt lokal. Die Zuverlässigkeit hingegen gilt nur global. ♢

6.2 A-Posteriori-Analysis für die Mini-FEM

In diesem Abschnitt werden wir einen Fehlerschätzer für die Mini-FEM aus Definition 5.28 für das Stokes-Problem herleiten und seine Effizienz und Zuverlässigkeit zeigen. Wir werden wieder ähnlich wie in Abschnitt 6.1 vorgehen und zuerst ein Residuum definieren, dessen Norm äquivalent zum Fehler ist.
Wir haben in Abschnitt 5.3 gezeigt, dass die Abbildung $L : H_0^1(\Omega;\mathbb{R}^d) \times (L^2(\Omega)/\mathbb{R}) \to (H_0^1(\Omega;\mathbb{R}^d))' \times ((L^2(\Omega)/\mathbb{R}))'$, die zu dem Stokes-Problem gehört ein Isomorphismus ist. Es

sei $(u,p) \in H^1_0(\Omega) \times L^2(\Omega)/\mathbb{R}$ die exakte Lösung zur Stokes-Gleichung aus Definition 5.25 und $(u_h, p_h) \in V_h(\mathcal{T}) \times Q_h(\mathcal{T})$ die Mini-FEM Approximation aus Definition 5.28. Also gilt

$$\begin{aligned}\|L(u-u_h, p-p_h)\|_{(H^1_0(\Omega;\mathbb{R}^d))' \times (L^2(\Omega)/\mathbb{R})'} &\lesssim \|(u-u_h, p-p_h)\|_{H^1_0(\Omega;\mathbb{R}^d)\times(L^2(\Omega)/\mathbb{R})} \\ &\lesssim \|L(u-u_h, p-p_h)\|_{(H^1_0(\Omega;\mathbb{R}^d))' \times (L^2(\Omega)/\mathbb{R})'}.\end{aligned}$$

Dies ist der Ausgangspunkt für unsere Analysis. Definiere $\mathrm{Res}_1 \in (H^1_0(\Omega;\mathbb{R}^d))'$ und $\mathrm{Res}_2 \in (L^2(\Omega)/\mathbb{R})'$ durch

$$\begin{aligned}\mathrm{Res}_1(v) &:= \int_\Omega f \cdot v \, dx - \int_\Omega \nabla u_h \cdot \nabla v \, dx - \int_\Omega p_h \operatorname{div} v \, dx \qquad \text{für alle } v \in H^1_0(\Omega;\mathbb{R}^d), \\ \mathrm{Res}_2(q) &:= \int_\Omega q \operatorname{div} u_h \, dx \qquad \text{für alle } q \in L^2(\Omega)/\mathbb{R}.\end{aligned}$$

Proposition 6.6. *Es gilt*

$$\|\nabla(u-u_h)\| + \|p-p_h\| \approx \|\mathrm{Res}_1\|_{(H^1_0(\Omega;\mathbb{R}^d))'} + \|\mathrm{Res}_2\|_{(L^2(\Omega)/\mathbb{R})'}.$$

Beweis. Mit der Argumentation von oben müssen wir nur $L(u-u_h, p-p_h) = (\mathrm{Res}_1, \mathrm{Res}_2)$ zeigen. Nach Definition ist dies aber erfüllt, wenn

$$\begin{aligned}\int_\Omega \nabla(u-u_h)\cdot\nabla v \, dx + \int_\Omega (p-p_h) \operatorname{div} v \, dx &= \mathrm{Res}_1(v) \qquad \text{für alle } v \in H^1_0(\Omega;\mathbb{R}^d), \\ \int_\Omega q \operatorname{div}(u-u_h) \, dx &= \mathrm{Res}_2(q) \qquad \text{für alle } q \in L^2(\Omega)/\mathbb{R}.\end{aligned}$$

Da (u,p) die exakte Lösung zum Stokes-Problem ist, ist dies erfüllt. □

Definition 6.7 (residualer Fehlerschätzer für Mini-FEM)**.** Für ein gegebenes $u_h \in V_h(\mathcal{T})$ definiere für alle $T \in \mathcal{T}$ den Fehlerschätzer für die Mini-FEM

$$\begin{aligned}\eta(T) := \Bigg(&\|h_T(f + \Delta u_h + \nabla p_h)\|^2_{L^2(T)} + \|\operatorname{div} u_h\|^2_{L^2(T)} \\ &+ h_T \sum_{E \in \mathcal{E}(T)\cap\mathcal{E}(\Omega)} \|[(\nabla u_h + p_h I_{2\times 2})\cdot \nu_E]_E\|^2_{L^2(E)} \Bigg)^{1/2}\end{aligned}$$

und

$$\eta := \sqrt{\sum_{T\in\mathcal{T}} \eta^2(T)}. \qquad \diamondsuit$$

Satz 6.8 (Zuverlässigkeit des Fehlerschätzers)**.** *Es gilt*

$$\|\nabla(u-u_h)\| + \|p-p_h\| \lesssim \eta.$$

Beweis. Es sei $v \in H^1_0(\Omega;\mathbb{R}^2)$ beliebig und es sei $v_h = I_h v$. Dann gilt wegen der Galerkin-Orthogonalität mit einer stückweisen partiellen Integration

$$\begin{aligned}\mathrm{Res}_1(v) = \mathrm{Res}_1(v-v_h) &= \int_\Omega f\cdot(v-v_h)\,dx - \int_\Omega \nabla u_h \cdot \nabla(v-v_h)\,dx - \int_\Omega p_h \operatorname{div}(v-v_h)\,dx \\ &= \sum_{T\in\mathcal{T}} \int_T (f + \Delta u_h + \nabla p_h)\cdot(v-v_h)\,dx - \sum_{E\in\mathcal{E}(\Omega)} \int_E [(\nabla u_h + p_h I_{2\times 2})\,\nu_E]_E \cdot (v-v_h)\,ds.\end{aligned}$$

Wie im Beweis von Satz 6.4 schätzen wir den ersten Term wieder mit den Approximationseigenschaften von I_h ab

$$\begin{aligned}\int_T (f+\Delta u_h+\nabla p_h)\cdot(v-v_h)\,dx &\le \|h_{\mathcal{T}}(f+\Delta u_h+\nabla p_h)\|_{L^2(T)}\|h_{\mathcal{T}}^{-1}(v-v_h)\|_{L^2(T)}\\ &\lesssim \|h_{\mathcal{T}}(f+\Delta u_h+\nabla p_h)\|_{L^2(T)}\,\|\nabla v\|_{L^2(\Omega_T)}.\end{aligned}$$

Für den zweiten Term benutzen wir auch wieder die Approximations- und Stabilitätseigenschaften und eine Spurungleichung und erhalten

$$\begin{aligned}-\int_E [(\nabla u_h + p_h I_{2\times 2})\,\nu_E]_E\cdot(v-v_h)\,ds&\\ &\le |E|^{-1/2}\|v-v_h\|_{L^2(E)}\,|E|^{1/2}\|[(\nabla u_h+p_hI_{2\times2})\,\nu_E]_E\|_{L^2(E)}\\ &\lesssim \left(h_{T_E}^{-1}\|v-v_h\|_{L^2(T_E)}+\|\nabla(v-v_h)\|_{L^2(T_E)}\right)h_{T_E}^{1/2}\|[(\nabla u_h+p_hI_{2\times2})\,\nu_E]_E\|_{L^2(E)}\\ &\lesssim \|\nabla v\|_{L^2(\Omega_{T_E})}\,h_{T_E}^{1/2}\|[(\nabla u_h+p_hI_{2\times2})\,\nu_E]_E\|_{L^2(E)}.\end{aligned}$$

Mit Cauchy-Ungleichungen folgt wie im Beweis von Satz 6.4

$$\|\mathrm{Res}_1\|_{(H_0^1(\Omega;\mathbb{R}^2))'}\lesssim\eta.$$

Für die Abschätzung des zweiten Residuums benutzen wir eine Cauchy-Ungleichung und erhalten

$$\mathrm{Res}_2(q)=\int_\Omega q\,\mathrm{div}\,u_h\,dx\le\|q\|\,\|\mathrm{div}\,u_h\|.$$

Also folgt

$$\|\mathrm{Res}_2\|_{(L^2(\Omega)/\mathbb{R})'}=\sup_{q\in(L^2(\Omega)/\mathbb{R})'}\frac{\mathrm{Res}_2(q)}{\|q\|}\le\|\mathrm{div}\,u_h\|\lesssim\eta$$

und damit die Behauptung. □

Satz 6.9 (lokale Effizienz des Fehlerschätzers)**.** *Für alle $T\in\mathcal{T}$ gilt*

$$\eta(T)\lesssim\|\nabla(u-u_h)\|_{L^2(\omega_T)}+\|p-p_h\|_{L^2(\omega_T)}+\mathrm{osc}(f,\mathcal{T}(\omega_T)).$$

Beweis. Wir werden wie im Beweis von Satz 6.5 die sogenannte *bubble function technique* anwenden. Dazu definieren wir $\varphi_T:=\flat_T(\Pi_0 f+\Delta u_h+\nabla p_h)|_T\in H_0^1(T)$ mit der Bubble-Funktion $\flat_T$ aus dem Beweis von Satz 6.5. Dann gilt mit $g:=(\Pi_0 f+\Delta u_h+\nabla p_h)|_T$

$$\begin{aligned}\|\varphi_T\|_{L^2(T)}&\le\|\flat_T\|_{L^\infty(T)}\|g\|_{L^2(T)}=\|g\|_{L^2(T)},\\ \|\nabla\varphi_T\|_{L^2(T)}&\le\|\flat_T\nabla g\|_{L^2(T)}+\|g\nabla\flat_T\|_{L^2(T)}\\ &\le\|\flat_T\|_{L^\infty(T)}\|\nabla g\|_{L^2(T)}+\|\nabla\flat_T\|_{L^\infty(T)}\|g\|_{L^2(T)}\\ &\lesssim\|\nabla g\|_{L^2(T)}+h_T^{-1}\|g\|_{L^2(T)}\overset{\text{inverse Ungl.}}{\lesssim}h_T^{-1}\|g\|_{L^2(T)}.\end{aligned}\tag{6.2}$$

Es gilt $(\Pi_0 f+\Delta u_h+\nabla p_h)|_T\in P_1(T;\mathbb{R}^d)$. Da $P_1(T;\mathbb{R}^d)$ ein endlich-dimensionaler Raum ist und $\flat_T>0$ auf $\mathrm{int}(T)$ sind die Normen $\|\bullet\|_{L^2(T)}$ und $\|\bullet\,\flat_T^{1/2}\|_{L^2(T)}$ äquivalent. Ein Skalierungsargument zeigt, dass die Konstante nicht von der Gitterweite abhängt. Damit gilt

$$\begin{aligned}\|(\Pi_0 f+\Delta u_h+\nabla p_h)|_T\|_{L^2(T)}^2&\lesssim\left\|\flat_T^{1/2}(\Pi_0 f+\Delta u_h+\nabla p_h)|_T\right\|_{L^2(T)}^2\\ &=\int_T(\Pi_0 f+\Delta u_h+\nabla p_h)|_T\cdot\varphi_T\,dx\\ &=\int_T(f+\Delta u_h+\nabla p_h)|_T\cdot\varphi_T\,dx+\int_T(\Pi_0 f-f)\cdot\varphi_T\,dx.\end{aligned}$$

Wie im Beweis von Satz 6.5 schätzen wir den zweiten Term mit Hilfe von (6.2) ab gegen

$$\begin{aligned}\int_T (\Pi_0 f - f)\cdot\varphi_T\,dx &\lesssim \|h_{\mathcal{T}}(f-\Pi_0 f)\|_{L^2(T)}\|h_T^{-1}\varphi_T\|_{L^2(T)}\\ &\le h_T^{-1}\|h_{\mathcal{T}}(f-\Pi_0 f)\|_{L^2(T)}\,\|(\Pi_0 f+\Delta u_h+\nabla p_h)|_T\|_{L^2(T)}\,.\end{aligned}$$

Da $\varphi_T\in H_0^1(T)$ und (u,p) die Lösung des Stokes-Problems ist, folgt für den ersten Term mit einer partiellen Integration

$$\begin{aligned}\int_T (f+\Delta u_h+\nabla p_h)|_T\cdot\varphi_T\,dx &= \int_T \nabla(u-u_h)\cdot\nabla\varphi_T\,dx+\int_T (p-p_h)\operatorname{div}\varphi_T\,dx\\ &\le \big(\|\nabla(u-u_h)\|_{L^2(T)}+\|p-p_h\|_{L^2(T)}\big)\,\|\nabla\varphi_T\|_{L^2(T)}\\ &\overset{(6.2)}{\lesssim} h_T^{-1}\big(\|\nabla(u-u_h)\|_{L^2(T)}+\|p-p_h\|_{L^2(T)}\big)\,\|(\Pi_0 f+\Delta u_h+\nabla p_h)|_T\|_{L^2(T)}\,.\end{aligned}$$

Indem wir durch $h_T^{-1}\,\|(\Pi_0 f+\Delta u_h+\nabla p_h)|_T\|_{L^2(T)}$ teilen, folgt die Abschätzung für den Term $h_T\,\|(\Pi_0 f+\Delta u_h+\nabla p_h)|_T\|_{L^2(T)}$.
Da $\operatorname{div} u=0$, folgt direkt

$$\|\operatorname{div} u_h\|_{L^2(T)}=\|\operatorname{div}(u-u_h)\|_{L^2(T)}\le\|\nabla(u-u_h)\|_{L^2(T)}.$$

Die Abschätzung für den Sprungterm folgt mit den Argumenten aus Satz 6.5 und ist eine Übungsaufgabe. □

6.3 A-Posteriori-Analysis für die Raviart-Thomas FEM

In diesem Kapitel werden wir einen Fehlerschätzer für die Raviart-Thomas FEM definieren und seine Effizienz und Zuverlässigkeit zeigen.
In Abschnitt 5.2 haben wir gezeigt, dass die Abbildung $L: H(\operatorname{div},\Omega)\times L^2(\Omega)\to (H(\operatorname{div},\Omega)\times L^2(\Omega))'$, die zum gemischten Problem gehört, ein Isomorphismus ist. Wie in den vorangegangenen Abschnitten ist deswegen der Fehler äquivalent zum Residuum. Allerdings lässt sich das Residuum

$$\int_\Omega p_{\mathrm{RT}}\cdot q\,dx+\int_\Omega u_h\operatorname{div} q\,dx$$

nicht mit den üblichen Methoden abschätzen. Wir werden deswegen hier einen anderen Weg gehen und eine A-Posteriori-Abschätzung für $p-p_{\mathrm{RT}}$ in $L^2(\Omega)$ herleiten.

Satz 6.10. *Es sei $p=\nabla u\in H(\operatorname{div},\Omega)$ der Gradient der exakten Lösung zu (4.4) und $(p_{\mathrm{RT}},u_h)\in \mathrm{RT}_0(\mathcal{T})\times P_0(\mathcal{T})$ die Lösung zur RT-FEM aus Definition 5.18. Dann gilt*

$$\|p-p_{\mathrm{RT}}\|^2=\|f-\Pi_0 f\|^2_{H^{-1}(\Omega)}+\inf_{\varphi\in H_0^1(\Omega)}\|p_{\mathrm{RT}}-\nabla\varphi\|^2.$$

Beweis. Ist $(q_j)_{j\in\mathbb{N}}=(\nabla v_j)_{j\in\mathbb{N}}$ eine Cauchy-Folge in $L^2(\Omega)$, dann ist wegen der Poincaré-Ungleichung $(v_j)_{j\in\mathbb{N}}$ auch eine Cauchy-Folge in $H_0^1(\Omega)$. Deswegen existiert ein Grenzwert $v\in H_0^1(\Omega)$ und ∇v ist auch der Grenzwert von der Folge $(q_j)_{j\in\mathbb{N}}$ in $L^2(\Omega)$. Also ist der Raum $\nabla H_0^1(\Omega)$ als Unterraum in $L^2(\Omega)$ abgeschlossen. Aus der Funktionalanalysis folgt, dass wir p_{RT} auf diesen Unterraum projizieren können und dass es ein $\alpha\in H_0^1(\Omega)$ gibt mit

$$\|p_{\mathrm{RT}}-\nabla\alpha\|=\inf_{\varphi\in H_0^1(\Omega)}\|p_{\mathrm{RT}}-\nabla\varphi\|.$$

Außerdem erfüllt dieses α

$$\int_\Omega (p_{\mathrm{RT}} - \nabla\alpha)\cdot\nabla v\,dx = 0 \qquad \text{für alle } v \in H^1_0(\Omega).$$

Daher gilt

$$\begin{aligned}\|p - p_{\mathrm{RT}}\|^2 &= \int_\Omega (p - p_{\mathrm{RT}})\cdot(p - \nabla\alpha)\,dx + \int_\Omega (\nabla u - p_{\mathrm{RT}})\cdot(\nabla\alpha - p_{\mathrm{RT}})\,dx\\ &= \|p - \nabla\alpha\|^2 + \|p_{\mathrm{RT}} - \nabla\alpha\|^2\end{aligned}$$

Wegen der Orthogonalität $(p_{\mathrm{RT}} - \nabla\alpha)\perp_{L^2(\Omega)}\nabla H^1_0(\Omega)$ gilt für den ersten Term in der ersten Zeile der obigen Gleichung auch

$$\begin{aligned}\|p - \nabla\alpha\|^2 &= \int_\Omega (p - \nabla\alpha)\cdot\nabla(u - \alpha)\,dx = \int_\Omega (p - p_{\mathrm{RT}})\cdot\nabla(u - \alpha)\,dx\\ &= -\int_\Omega \operatorname{div}(p - p_{\mathrm{RT}})\,(u - \alpha)\,dx = \int_\Omega (f - \Pi_0 f)\,(u - \alpha)\,dx\\ &\le \|f - \Pi_0 f\|_{H^{-1}(\Omega)}\|\nabla(u - \alpha)\|.\end{aligned}$$

Andererseits gilt für alle $v \in H^1_0(\Omega)$

$$\begin{aligned}\int_\Omega (f - \Pi_0 f)v\,dx &= -\int_\Omega v\operatorname{div}(p - p_{\mathrm{RT}})\,dx = \int_\Omega (p - p_{\mathrm{RT}})\cdot\nabla v\,dx\\ &= \int_\Omega (p - \nabla\alpha)\cdot\nabla v\,dx \le \|p - \nabla\alpha\|\,\|\nabla v\|.\end{aligned}$$

Also gilt

$$\|p - \nabla\alpha\| = \|f - \Pi_0 f\|_{H^{-1}(\Omega)}.$$

Damit folgt die Behauptung.

□

Den ersten Term können wir direkt gegen Oszillationen von f abschätzen.

Proposition 6.11. *Es gilt*

$$\|f - \Pi_0 f\|_{H^{-1}(\Omega)} \lesssim \operatorname{osc}(f, \mathcal{T}).$$

Beweis. Mit einer Poincaré-Ungleichung gilt

$$\begin{aligned}\|f - \Pi_0 f\|_{H^{-1}(\Omega)} &= \sup_{v\in H^1_0(\Omega)\setminus\{0\}} \frac{\int_\Omega (f - \Pi_0 f)v\,dx}{\|\nabla v\|} = \sup_{v\in H^1_0(\Omega)\setminus\{0\}} \frac{\int_\Omega (f - \Pi_0 f)(v - \Pi_0 v)\,dx}{\|\nabla v\|}\\ &\le \sup_{v\in H^1_0(\Omega)\setminus\{0\}} \frac{\|h_{\mathcal{T}}(f - \Pi_0 f)\|\,\|h_{\mathcal{T}}^{-1}(v - \Pi_0 v)\|}{\|\nabla v\|} \lesssim \|h_{\mathcal{T}}(f - \Pi_0 f)\|.\end{aligned}$$

□

Für die Abschätzung des zweiten Terms benötigen wir die Helmholtz-Zerlegung.

Satz 6.12 (Helmholtz-Zerlegung in $2d$)**.** *Es gilt*

$$L^2(\Omega;\mathbb{R}^2) = \nabla H^1_0(\Omega) \oplus \operatorname{Curl} H^1(\Omega)/\mathbb{R}, \tag{6.3}$$

wobei

$$\operatorname{Curl} v := \begin{pmatrix} -\partial_2 v \\ \partial_1 v \end{pmatrix}$$

und die Summe in (6.3) *ist orthogonal in* L^2.

Beweisskizze. Für glatte Funktionen gilt

$$\operatorname{div}\operatorname{Curl} v = -\partial_1\partial_2 v + \partial_2\partial_1 v = 0.$$

Die L^2-Orthogonalität folgt dann mit einem Dichtheitsargument: Ist $v \in H^1_0(\Omega)$ und $\beta \in H^1(\Omega)/\mathbb{R}$, dann gibt es eine Folge $(\beta_j)_{j\in\mathbb{N}}$ in $C^\infty(\Omega)$ mit $\beta_j \to \beta$ in $H^1(\Omega)$. Damit gilt

$$\begin{aligned}\int_\Omega \nabla v \cdot \operatorname{Curl}\beta\,dx &= \int_\Omega \nabla v \cdot \operatorname{Curl}(\beta-\beta_j)\,dx + \int_\Omega \nabla v \cdot \operatorname{Curl}\beta_j\,dx \\ &= \int_\Omega \nabla v \cdot \operatorname{Curl}(\beta-\beta_j)\,dx - \int_\Omega v \underbrace{\operatorname{div}\operatorname{Curl}\beta_j}_{=0}\,dx + \int_{\partial\Omega} \underbrace{v}_{=0} \operatorname{Curl}\beta_j \cdot \nu\,ds \\ &\le \|\nabla v\|\,\|\nabla(\beta-\beta_j)\| \to 0.\end{aligned}$$

Daraus folgt die L^2-Orthogonalität.
Nach Definition von $H^1(\Omega)$ bzw. $H^1_0(\Omega)$ gilt

$$\nabla H^1_0(\Omega) \oplus \operatorname{Curl} H^1(\Omega)/\mathbb{R} \subseteq L^2(\Omega;\mathbb{R}^2).$$

Für die Inklusion in die andere Richtung sei $p \in L^2(\Omega;\mathbb{R}^2)$. Es sei $\alpha \in H^1(\Omega)/\mathbb{R}$ definiert durch

$$\int_\Omega \operatorname{Curl}\alpha \cdot \operatorname{Curl}\beta\,dx = \int_\Omega p \cdot \operatorname{Curl}\beta\,dx \qquad \text{für alle } \beta \in H^1(\Omega)/\mathbb{R}.$$

Eine eindeutige Lösung α zu diesem Problem existiert, da die linke Seite eine stetige und koerzitive Bilinearform auf $H^1(\Omega)/\mathbb{R}$ definiert. Definiere den Differentialoperator rot durch

$$\operatorname{rot} q = -\partial_2 q_1 + \partial_1 q_2.$$

Dann gilt für $p - \operatorname{Curl}\alpha$, dass $\operatorname{rot}(p-\operatorname{Curl}\alpha) = 0$ im schwachen Sinne gilt. Für glatte Funktionen wird in den Grundvorlesungen gezeigt, dass rotationsfreie Funktionen Gradientenfelder sind. Für L^2-Funktionen, deren schwache Rotation verschwindet, kann dies mithilfe einer Fourier-Transformation gezeigt werden. Wir verweisen für den Beweis auf [GR86, Satz 3.2]. Es folgt jedenfalls, dass es ein $u \in H^1_0(\Omega)$ gibt mit

$$\nabla u = p - \operatorname{Curl}\alpha. \qquad \square$$

Mit der Helmholtz-Zerlegung können wir den zweiten Term aus Satz 6.10 abschätzen. Wir definieren

$$\mu(T) := h_T^{1/2} \sqrt{\sum_{E\in\mathcal{E}(T)} \|[p_{\mathrm{RT}}]_E \cdot \tau_E\|^2_{L^2(E)}},$$
$$\mu := \sqrt{\sum_{T\in\mathcal{T}} \mu^2(T)}.$$

Satz 6.13. *Es gilt*

$$\min_{v\in H^1_0(\Omega)} \|p_{\mathrm{RT}} - \nabla v\| \lesssim \mu$$

und

$$|E|^{1/2} \|[p_{\mathrm{RT}}]_E \cdot \tau_E\|_{L^2(E)} \lesssim \|p_{\mathrm{RT}} - \nabla\alpha\|_{L^2(\omega_E)},$$

wobei $\alpha \in H^1_0(\Omega)$ *den Minimierer in* $\min_{v\in H^1_0(\Omega)}\|p_{\mathrm{RT}} - \nabla v\|$ *bezeichne.*

Beweis. Wie wir auch im Beweis von Satz 6.10 gesehen haben, gilt für das minimierende $\alpha \in H^1_0(\Omega)$

$$\int_\Omega (p_{\mathrm{RT}} - \nabla\alpha) \cdot \nabla v \, dx = 0 \qquad \text{für alle } v \in H^1_0(\Omega),$$

d.h., dass $p_{\mathrm{RT}} - \nabla\alpha$ orthogonal (in L^2) auf $\nabla H^1_0(\Omega)$ ist. Mit der Helmholtz-Zerlegung aus Satz 6.12 folgt, dass es ein $\beta \in H^1(\Omega)/\mathbb{R}$ gibt mit

$$p_{\mathrm{RT}} - \nabla\alpha = \operatorname{Curl}\beta.$$

Es sei $\widetilde{J_1} : P_1(\mathcal{T}) \to S^1(\mathcal{T})$ der Mittelungsoperator, der wie J_1 definiert ist, aber an den Knoten am Rand nicht (notwendigerweise) Null ist, sondern dort auch durch Mittelung definiert ist. Man kann zeigen, dass für diesen Operator die gleichen Eigenschaften gelten wie für J_1. Definiere $\beta_h := \widetilde{J_1}(\Pi_0(\beta)) \in S^1(\mathcal{T})$. Dann gilt $\operatorname{Curl}\beta_h \in P_0(\mathcal{T};\mathbb{R}^2)$ und da β_h stetig ist

$$[\operatorname{Curl}\beta_h \cdot \nu_E]_E = [\nabla\beta_h \cdot \tau_E]_E = [\partial\beta_h/\partial\tau_E]_E = 0,$$

wobei $\tau_E = \begin{pmatrix} -\nu_{E,2} \\ \nu_{E,1} \end{pmatrix}$ die Tangente an E bezeichne. Also folgt $\operatorname{Curl}\beta_h \in \mathrm{RT}_0(\mathcal{T})$. Außerdem folgt aus der L^2-Orthogonalität von $\operatorname{Curl} H^1(\Omega)$ und $\nabla H^1_0(\Omega)$, dass $\operatorname{div}\operatorname{Curl}\beta_h$ existiert und $\operatorname{div}\operatorname{Curl}\beta_h = 0$ in L^2 gilt. Deswegen dürfen wir $\operatorname{Curl}\beta_h$ als Testfunktion in Definition 5.20 verwenden und erhalten

$$\int_\Omega p_{\mathrm{RT}} \cdot \operatorname{Curl}\beta_h \, dx = 0.$$

Dann gilt mit der Orthogonalität von $\operatorname{Curl}(H^1(\Omega)/\mathbb{R})$ zu $\nabla H^1_0(\Omega)$ und einer stückweisen partiellen Integration

$$\begin{aligned}
\|\operatorname{Curl}\beta\|^2 &= \int_\Omega \operatorname{Curl}\beta \cdot (p_{\mathrm{RT}} - \nabla\alpha)\, dx = \int_\Omega \operatorname{Curl}\beta \cdot p_{\mathrm{RT}}\, dx = \int_\Omega \operatorname{Curl}(\beta - \beta_h) \cdot p_{\mathrm{RT}}\, dx \\
&= -\sum_{T\in\mathcal{T}} \int_T (\beta - \beta_h)\,(-\partial_2 p_{\mathrm{RT},1} + \partial_1 p_{\mathrm{RT},2})\, dx \\
&\qquad\qquad + \sum_{E\in\mathcal{E}} \int_E (\beta - \beta_h)\,[-p_{\mathrm{RT},1}\,\nu_{E,2} + p_{\mathrm{RT},2}\,\nu_{E,1}]_E\, ds \\
&= -\sum_{T\in\mathcal{T}} \int_T (\beta - \beta_h)\,\operatorname{rot}(p_{\mathrm{RT}}|_T)\, dx + \sum_{E\in\mathcal{E}} \int_E (\beta - \beta_h)\,[p_{\mathrm{RT}} \cdot \tau_E]_E\, ds.
\end{aligned}$$

Da $p_{\mathrm{RT}}|_T(x) = \begin{pmatrix} a_T \\ b_T \end{pmatrix} + c_T x$ für Konstanten $a_T, b_T, c_T \in \mathbb{R}$, gilt $\operatorname{rot}(p_{\mathrm{RT}}|_T) = 0$. Wir haben also

$$\|\operatorname{Curl}\beta\|^2 = \sum_{E\in\mathcal{E}} \int_E (\beta - \beta_h)\,[p_{\mathrm{RT}} \cdot \tau_E]_E\, ds$$

gezeigt.
Wir verfahren nun wieder wie im Beweis der Zuverlässigkeit in den Sätzen 6.4 und 6.8. Mit den Approximations- und Stabilitätseigenschaften und einer Spurungleichung folgt also

$$\begin{aligned}
\int_E (\beta - \beta_h)\,[p_{\mathrm{RT}} \cdot \tau_E]_E\, ds &\le |E|^{-1/2}\|\beta - \beta_h\|_{L^2(E)}\; |E|^{1/2}\|[p_{\mathrm{RT}} \cdot \tau_E]_E\|_{L^2(E)} \\
&\lesssim \left(h_{T_E}^{-1}\|\beta - \beta_h\|_{L^2(T_E)} + \|\nabla(\beta - \beta_h)\|_{L^2(T_E)}\right) h_{T_E}^{1/2}\|[p_{\mathrm{RT}} \cdot \tau_E]_E\|_{L^2(E)} \\
&\lesssim \|\nabla\beta\|_{L^2(\Omega_{T_E})} h_{T_E}^{1/2}\|[p_{\mathrm{RT}} \cdot \tau_E]_E\|_{L^2(E)}.
\end{aligned}$$

Da $|\nabla\beta| = |\operatorname{Curl}\beta|$, folgt also wie in den Sätzen 6.4 und 6.8

$$\|\operatorname{Curl}\beta\| \lesssim \mu.$$

Für die Effizienz benutzen wir wieder die *bubble function technique* und wir gehen wieder ähnlich vor wie im Beweis von den Sätzen 6.5 und 6.9. Da $[p_{\mathrm{RT}}\cdot\tau_E]_E$ nun nicht konstant ist auf E, müssen wir den Sprung auf ω_E fortsetzen. Dies kann z.B. dadurch geschehen, dass der Sprung konstant in die Normalenrichtung fortgesetzt wird. Definiere $\varphi_E := \flat_E [p_{\mathrm{RT}}\cdot\tau_E]_E$, wobei der Sprung hier als die Fortsetzung angesehen wird. Dann gilt mit einer stückweisen partiellen Integration und der Orthogonalität von $p = \nabla u$ zu $\operatorname{Curl} H^1(\Omega)$

$$\begin{aligned}
\|[p_{\mathrm{RT}}\cdot\tau_E]_E\|_{L^2(E)}^2 &\lesssim \|[p_{\mathrm{RT}}\cdot\tau_E]_E \flat_E^{1/2}\|_{L^2(E)}^2 = \int_E [p_{\mathrm{RT}}\cdot\tau_E]_E\,\varphi_E\,ds \\
&= \sum_{T\in\mathcal{T}, T\subseteq\overline{\omega}_E} \int_T \varphi_E \underbrace{\operatorname{rot} p_{\mathrm{RT}}}_{=0}\,dx + \int_{\omega_E} p_{\mathrm{RT}}\cdot\operatorname{Curl}\varphi_E\,dx \\
&= \int_{\omega_E} (p_{\mathrm{RT}} - \nabla\alpha)\cdot\operatorname{Curl}\varphi_E\,dx \\
&\lesssim \|\nabla\alpha - p_{\mathrm{RT}}\|_{L^2(\omega_E)}\,\|\operatorname{Curl}\varphi_E\|_{L^2(\omega_E)}.
\end{aligned}$$

Mit

$$\|\operatorname{Curl}\varphi_E\|_{L^2(\omega_E)} \lesssim |E|^{-1}\,\|[p_{\mathrm{RT}}\cdot\tau_E]_E\|_{L^2(\omega_E)} \lesssim |E|^{-1/2}\,\|[p_{\mathrm{RT}}\cdot\tau_E]_E\|_{L^2(E)}$$

folgt die Behauptung. □

Korollar 6.14. *Es gilt*

$$\|p - p_{\mathrm{RT}}\| + \operatorname{osc}(f,\mathcal{T}) \approx \mu + \operatorname{osc}(f,\mathcal{T}).$$

Beweis. Dies folgt aus den Sätzen 6.10 und 6.13 und Proposition 6.11. □

Literatur

[Alt16] Hans Wilhelm Alt. *Linear functional analysis*. Universitext. Springer-Verlag London, Ltd., London, 2016. An application-oriented introduction, Translated from the German edition by Robert Nürnberg.

[Bar16] Sören Bartels. *Numerical approximation of partial differential equations*, volume 64 of *Texts in Applied Mathematics*. Springer, [Cham], 2016.

[Bra13] Dietrich Braess. *Finite Elemente*. Springer, 5. edition, 2013.

[BS08] Susanne C. Brenner and L. Ridgway Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008.

[Eva98] Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1998.

[GR86] Vivette Girault and Pierre-Arnaud Raviart. *Finite element methods for Navier-Stokes equations*, volume 5 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1986. Theory and algorithms.

[Gri11] Pierre Grisvard. *Elliptic problems in nonsmooth domains*, volume 69 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011. Reprint of the 1985 original [MR0775683], With a foreword by Susanne C. Brenner.

[Gud10] T. Gudi. A new error analysis for discontinuous finite element methods for linear elliptic problems. *Math. Comp.*, 79(272):2169–2189, 2010.