Die Finite Elemente Methode für partielle Differentialgleichungen

Thomas Richter
thomas.richter@ovgu.de

Magdeburg

Inhaltsverzeichnis

Literaturverzeichnis

[1] H.W. Alt. *Lineare Funktionalanalysis. Eine anwendungsorientierte Einführung*. Springer Verlag, Berlin, 5. auflage edition, 2008.

[2] D. Braess. *Finite Elemente*. Springer, 1997.

[3] V. Girault and P.-A. Raviart. *Finite Elements for the Navier Stokes Equations*. Springer, 1986.

[4] T.J.R. Hughes, J.A. Cottrell, and Y. Bazilevs. Isogeometric analysis: Cad, finite elements, nurbs, exact geometry and mesh refinement. *Computer Methods in Applied Mechanics and Engineering*, 194(39–41):4135 – 4195, 2005.

[5] M. Lenoir. Optimal isoparametric finite elements and error estimates for domains involving curved boundaries. *SIAM Journal on Numerical Analysis*, 23(3):562–580, 1986.

[6] R. Rannacher. *Special Topics in Numerics I (FEM for Nonlinear Problems)*. Universität Heidelberg, http://numerik.uni-hd.de/~lehre/notes/, 2016. Vorlesungsskriptum.

[7] B. Schweizer. *Partielle Differentialgleichungen. Eine Anwendungsorientierte Einführung*. Springer, 2013.

[8] J. Wloka. *Partielle Differentialgleichungen*. Teubner, Stuttgart, 1982.

1 Einleitung

Partielle Differentialgleichungen treten zur Beschreibung einer Vielzahl physikalischer Prozesse auf. Üblicherweise suchen wir in einem Gebiet $\Omega \subset \mathbb{R}^d$ mit $d > 1$ eine Funktion $u : \Omega \to \mathbb{R}^c$ mit $c \in \mathbb{N}$, welche einer Differentialvorschrift genügt:

$$F(x, u, \nabla u, \nabla^2 u, \ldots, \nabla^n u) = 0.$$

Von einer partiellen Differentialgleichung spricht man, wenn partielle Ableitungen in mehrere Richtungen auftreten, d.h. nur im Fall $d > 1$ (und nur dann, wenn auch wirklich verschiedene Richtungen auftauchen). Im Fall $c = 1$ sprechen wir von einer *skalaren Differentialgleichung*, im Fall $c > 1$ von einem *System von Differentialgleichungen*. Die höchste Stufe n der auftretenden Ableitungen heißt die Ordnung der Differentialgleichung. Wir beschränken uns hier zunächst auf skalare partielle Differentialgleichungen bis zur Ordnung $n = 2$. Zumeist behandeln wir *lineare* partielle Differentialgleichungen. Diese schreiben wir in der Form:

$$Lu = f, \quad Lu := -\sum_{i,j=1}^{d} a_{ij}\partial_i\partial_j u(x) + \sum_{i=1}^{d} a_i\partial_i u + au,$$

mit einem *Differentialoperator* L und mit (reellen) Koeffizienten a_{ij}, a_i, a. Zur vollständigen Beschreibung von partiellen Differentialgleichungen gehören wie bei gewöhnlichen Differentialgleichungen üblicherweise Randwerte bzw. Anfangswerte.

1.1 Beispiele von partiellen Differentialgleichungen

Wir erwähnen hier kurz einige Beispiele für partielle Differentialgleichungen, welche Prozesse aus der Natur beschreiben. Zunächst sei $\Omega \subset \mathbb{R}^d$ für $d \geqslant 1$ ein Gebiet. Auf diesem Gebiet beschreiben wir die Ausbreitung von Wellen durch die *Wellengleichung*. Wir suchen auf dem Orts-Zeit Gebiet $[0, T] \times \mathbb{R}^d$ die Lösung $u(x, t)$ so dass

$$\partial_t^2 u(x, t) = \Delta u(x, t),$$

wobei durch Δ der *Laplace-Operator* als Summe der (örtlich) zweiten partiellen Ableitungen bestimmt ist:

$$\Delta := \sum_{i=1}^{d} \partial_i^2.$$

Die Wellengleichung hat eine ausgezeichnete Variable t, die physikalische Zeit. Auch für eindimensionale Gebiete Ω mit $d = 1$ liegt eine partielle Differentialgleichung mit den partiellen Ableitungen ∂_{tt} und ∂_{xx} vor. Für den Fall $d = 1$ ist eine mögliche Lösung der Wellengleichung gegeben durch

$$u(x,t) = \sin(x)\sin(t).$$

Wir haben hier noch keine Rand- oder Anfangswerte angegeben. Die Wellengleichung ist der Prototyp einer *hyperbolischen partiellen Differentialgleichung*.

Ein weiteres Beispiel für eine einfache partielle Differentialgleichung ist die Wärmeleitungsgleichung: In einem Gebiet $\Omega \subset \mathbb{R}^d$ mit $d \geqslant 1$ beschreiben wir die Ausbreitung von Wärme $u(x,t)$ im Laufe der Zeit. Zu Beginn $t = 0$ sei eine Wärmeverteilung vorgeschrieben $u(x,0) = u^0(x)$, der Rand des Gebiets sei "isoliert", d.h. es existiert kein Wärmefluss über den Rand. Mathematisch formuliert bedeutet dies, dass sich die Wärmeverteilung in Richtung Rand (also in Normalrichtung n) nicht ändert, $\partial_n u(x,t) = 0$ für $x \in \partial\Omega$. Die zeitliche Verteilung der Wärme ist beschrieben durch die Gleichung:

$$\partial_t u(x,t) - \lambda\Delta u(x,t) = f(x,t) \quad x \in \Omega,\ t \geqslant 0, \quad u(x,0) = u^0(x), \quad \partial_n u = 0 \text{ auf } \partial\Omega,$$

wobei durch $f(x,t)$ ein externer Wärmezufluss (oder eine Wärmesenke) gegeben ist und λ eine physikalische Konstante, welche die Wärmeleitfähigkeit des Mediums beschreibt. Die Wärmeleitungsgleichung hat wieder eine ausgezeichnete Zeitrichtung und ist unabhängig von der Dimension des Gebietes d eine partielle Differentialgleichung. Sie ist der Prototyp einer *parabolischen partiellen Differentialgleichung*. Sie ist ein *Anfangs-Randwertproblem*.

Oft ist man nicht am zeitlichen Verlauf der Lösung interessiert, sondern nur am *stationären Zustand* welcher sich (gegebenenfalls) für $t \to \infty$ einstellt. Ein stationärer Zustand ist ein Gleichgewichtszustand, an dem sich die Lösung nicht mehr ändert (im Verlauf der Zeit), für den also $\partial_t u(x,t) = 0$ gilt. Die *stationäre Wärmeleitungsgleichung* ist bestimmt durch die Vorschrift:

$$-\Delta u(x) = f(x), \quad x \in \Omega, \quad \partial_n u(x) = 0 \quad x \in \partial\Omega.$$

Ein stationärer Zustand kann sich natürlich nur dann einstellen, wenn der Quellterm f nicht explizit von der Zeit abhängt. Die Gleichung $-\Delta u = f$ wird die *Poisson-Gleichung* genannt und wird uns in dieser Vorlesung hauptsächlich beschäftigen. Zusammen mit der Vorgabe der Normalableitung auf dem Rand spricht man vom *Neumann-Problem* der Poisson-Gleichung. Für $\Omega \subset \mathbb{R}$, also $d = 1$ ist diese Gleichung eine gewöhnliche Differentialgleichung, genauer ein Spezialfall des Sturm-Liouville Randwertproblems. Für $d \geqslant 2$ ist sie der Prototyp einer *elliptischen partiellen Differentialgleichung*.

Bei den bisherigen Beispielen handelt es sich um lineare partielle Differentialgleichungen zweiter Ordnung. Die Ordnung ist definiert als die höchste Stufe der auftretenden partiellen Ableitung. Ein Beispiel für eine partielle Differentialgleichung erster Ordnung ist die *Transportgleichung*. Dafür sei durch $\beta : \Omega \to \mathbb{R}^d$ ein *Transportfeld* gegeben (etwa der Wind). Der Transport einer Größe (etwa einer Dichteverteilung) $\rho(x,t)$ in einem Gebiet $\Omega \subset \mathbb{R}^d$ wird dann durch die Gleichung

$$\partial_t \rho(x,t) + \operatorname{div}\big(\beta(x,t)\rho(x,y)\big) = 0,$$

beschrieben. Dabei bezeichnen wir mit $\operatorname{div}\phi$ die *Divergenz* einer Funktion $\phi : \mathbb{R}^d \to \mathbb{R}^d$:

$$\operatorname{div}\phi = \sum_{i=1}^{d} \partial_i \phi_i.$$

Die Transportgleichung ist so wie die Wellengleichung eine *hyperbolische* Gleichung. Mit der Produktregel können wir die Transportgleichung schreiben als

$$\partial_t \rho(x,t) + (\beta \cdot \nabla)\rho(x,y) + (\operatorname{div}\ \beta(x,y))\rho(x,y) = 0,$$

bzw. im Fall eines *divergenzfreien* Transportfelds $\operatorname{div}\ \beta(x,y) \equiv 0$ kurz

$$\partial_t \rho(x,t) + (\beta \cdot \nabla)\rho(x,y) = 0.$$

Mit der Schreibweise

$$\beta \cdot \nabla = \sum_{i=1}^{d} \beta_i \partial_i =: \nabla_\beta \text{ oder auch } \partial_\beta$$

bezeichnen wir die *Richtungsableitung* in Richtung β. Der Punkt meint also einfach das Skalarprodukt zwischen dem Vektor $\beta(x,y)$ und dem Gradientenoperator ∇.

Falls mehr als eine Lösungsvariable involviert ist, so spricht man von einem *System von partiellen Differentialgleichungen*. Ein bekannter Vertreter eines (nichtlinearen) Systems von partiellen Differentialgleichungen sind die *Navier-Stokes Gleichungen*. Gesucht werden in einem Gebiet $\Omega \subset \mathbb{R}^2$ (hier zweidimensional) und $I := [0,T]$ Druck $p(x,t) : \Omega \to \mathbb{R}$ und Geschwindigkeitsfeld $\boldsymbol{v}(x,t) : \Omega \to \mathbb{R}^2$ einer inkompressiblen Flüssigkeit (oder eines Gases) so dass

$$\begin{aligned} \partial_t v^1 - \nu\Delta v^1 + v \cdot \nabla v^1 + \partial_x p &= f^1 \\ \partial_t v^2 - \nu\Delta v^2 + v \cdot \nabla v^2 + \partial_y p &= f^2 \\ \partial_x v^1 + \partial_y v^2 &= 0, \end{aligned}$$

wobei $\boldsymbol{v} = (v^1, v^2)$ die beiden komponenten der Geschwindigkeit sind. Unter Ausnutzung der Vektornotationen schreiben wir kompakter

$$\partial_t \boldsymbol{v} - \nu\boldsymbol{\Delta v} + (\boldsymbol{v} \cdot \nabla)\boldsymbol{v} + \nabla p = \mathbf{f}, \quad \operatorname{div}\boldsymbol{v} = 0,$$

wobei $\mathbf{f} = (f^1, f^2)$ ist und der Vektorlaplace $\boldsymbol{\Delta}$ definiert ist als

$$\boldsymbol{\Delta v} = \begin{pmatrix} \Delta v^1 \\ \Delta v^2 \end{pmatrix}.$$

1.1.1 Typeinteilung von linearen partiellen Differentialgleichungen zweiter Ordnung

Wir werden ausschließlich lineare partielle Differentialgleichungen zweiter Ordnung betrachten. Diese schreiben wir mit Hilfe eines Differentialoperators

$$L : C^2(\Omega) \to C(\Omega),$$

d.h. einer Abbildung vom Funktionenraum der zweimal stetig differenzierbaren Funktionen in den Funktionenraum der stetigen Funktionen in der Form:

$$Lu = f, \quad L := \sum_{i,j=1}^{d} a_{ij}\partial_i\partial_j + \sum_{j=1}^{d} a_j\partial_j + a.$$

Die Koeffizienten sind reell und können vom Ort abhängen, d.h. $a_{ij} = a_{ij}(x,y)$, $a_i = a_i(x,y)$ und $a = a(x,y)$. Da die partiellen Ableitungen in $C^2(\Omega)$ vertauschbar sind $\partial_i\partial_j u = \partial_j\partial_i u$ können die Koeffizienten $a_{ij} = a_{ji}$ als symmetrisch angenommen werden.

Bevor wir zu der angesprochenen Typeinteilung in *elliptische, parabolische* und *hyperbolische* Differentialgleichungen kommen stellen wir einen Zusammenhang zu den gewöhnlichen Differentialgleichungen vom Typ

$$u'(t) = f(t, u(t)), \quad u(t_0) = u^0,$$

her. Hier kann bei Kenntnis der rechten Seite f und des Anfangswert die Lösung aus der Taylor-Entwicklung rekonstruiert werden:

$$u(t) = \sum_{k=0}^{\infty} \frac{(t-t_0)^k}{k!} u^{(k)}(t_0) = u^0 + (t-t_0)\sum_{k=0}^{\infty} \frac{(t-t_0)^k}{(k+1)!} f^{(k)}(t_0, u(t_0)).$$

Ist die Differentialgleichung linear, d.h. im Fall

$$u'(t) = Au(t), \quad u(t_0) = u^0,$$

mit $A \in \mathbb{R}^{d\times d}$, so gilt

$$u(t) = u^0 + \sum_{k=0}^{\infty} \frac{(t-t_0)^k}{(k+1)!} A^k u^0.$$

Wir wollen nun versuchen, auch die Lösung von partiellen Differentialgleichungen aus den Anfangsdaten zu rekonstruieren. Wir betrachten den Fall $\Omega \in \mathbb{R}^2$. Dann sei $\Gamma \in \Omega$ eine Kurve durch Ω mit beliebig glatter Parametrisierung

$$\Gamma = \{(\tau_x(\gamma), \tau_y(\gamma)),\ \gamma \in [a,b],\ \tau_x, \tau_y \in C^\infty(a,b)\}.$$

Gesucht ist die Lösung $u \in C^\infty(\Omega)$ zur skalaren, linearen Differentialgleichung zweiter Ordnung

$$a_{11}\partial_{xx}u + 2a_{12}\partial_{xy}u + a_{22}\partial_{yy}u + a_1\partial_x u + a_2\partial_y u + au = f. \tag{1.1}$$

Angenommen, entlang Γ sei u und auch $\partial_n u$ bekannt. Dabei ist

$$\partial_n = \mathbf{n}\cdot\nabla$$

die Richtungsableitung in Richtung der (orientierten) Normalvektors $\mathbf{n} : \Gamma \to \mathbb{R}^2$ an Γ. Das wir zwei Werte vorgeben entspricht der Tatsache, dass wir es mit einer Differentialgleichung zweiter Ordnung zu tun haben. Da wir u entlang der glatten Kurve Γ kennen, kennen wir

auch die Tangentialableitung ∂_τ mit dem Tangentialvektor $\boldsymbol{\tau} : \Gamma \to \mathbb{R}^2$. Sind $\partial_n u$ und $\partial_\tau u$ bekannt, so kennen wir natürlich den gesamten Gradienten ∇u. Wir werden nun versuchen mit dem Wissen um u und ∇u aus der Gleichung (1.1) alle höheren Ableitungen $\nabla^k u$ zu rekonstruieren. Gelingt dies, so können wir die Lösung auf Ω als Taylor-Reihe darstellen.

Wir führen nun einige Bezeichnungen ein:

$$p := \partial_x u, \quad q := \partial_y u, \quad r := \partial_{xx} u, \quad s := \partial_{xy} u, \quad t := \partial_{yy} u$$

Differentiation von p und q in Richtung des Kurvenparameters $\gamma \in [a, b]$ ergibt

$$\begin{aligned} \partial_\gamma p &= \partial_x p \tau'_x + \partial_y p \tau'_y = r\tau'_x + s\tau'_y \\ \partial_\gamma q &= \partial_x q \tau'_x + \partial_y q \tau'_y = s\tau'_x + t\tau'_y, \end{aligned}$$

wobei τ'_x und τ'_y bekannt sind. Wir erinnern an die Ausgangslage: u sowie $p = \partial_x u$ und $q = \partial_y u$ sind bekannt. Gesucht sind die zweiten Ableitungen r, s, t. Mit (1.1) können wir ein lineares Gleichungssystem zur Bestimmung dieser Unbekannten (in jedem Punkt der Kurve) aufstellen

$$\begin{pmatrix} a_{11} & 2a_{12} & a_{22} \\ \tau'_x & \tau'_y & 0 \\ 0 & \tau'_x & \tau'_y \end{pmatrix} \begin{pmatrix} r \\ s \\ t \end{pmatrix} = \begin{pmatrix} f - a_1 p - a_2 q - au \\ \partial_\gamma p \\ \partial_\gamma q \end{pmatrix}, \tag{1.2}$$

wobei

$$\partial_\gamma p = (\boldsymbol{\tau} \cdot \nabla)\partial_x u, \quad \partial_\gamma q = (\boldsymbol{\tau} \cdot \nabla)\partial_y u,$$

also die höheren Ableitungen entlang der Kurve bekannt sind. Das System ist lösbar, wenn die Matrix invertiertbar ist, wenn also für die Determinante dieser Matrix - wir nennen sie $\mathbf{L}$ - gilt $\det(\mathbf{L}) \neq 0$. Wir Berechnen die Determinante durch Entwickeln nach der ersten Zeile

$$\det(\mathbf{L}) = a_{11}(\tau'_y)^2 - 2a_{12}\tau'_x\tau'_y + a_{22}(\tau'_x)^2.$$

Die Lösbarkeit hängt also nur von den Koeffizienten a_{11}, a_{12} und a_{22} ab. Diese bestimmen den *Hauptteil der Differentialgleichung*, die Terme vor der zweiten Ableitung. Weiter geht die Form der Kurve Γ ein.

Fall 1: $\det(L) \neq 0$ **in allen Punkten der Kurve** Γ**.** In diesem Fall ist das lineare Gleichungssystem in jedem Punkt der Kurve lösbar und wir erhalten als Lösung die zweiten Ableitungen $\nabla^2 u$ entlang der Kurve. Wir können nun das lineare Gleichungssystem (1.2) nach x und y differenzieren und erhalten so ein Gleichungssystem für die Unbekannten $\partial_x r, \partial_x r, \partial_x t$ sowie eines für $\partial_y r, \partial_y r, \partial_y t$. Die Koeffizientenmatrix ist die gleiche, da die alle Ableitungen der a_{ij}, bzw. die höheren Ableitungen der Kurve τ''_x und τ''_y mit den bekannten Größen r, s, t auf der rechten Seite des Gleichungssystems auftauchen. Auf diese Weise können sukzessive alle höheren Ableitungen der Lösung u bestimmt werden.

Ist $x_0, y_0 \in \Gamma$ ein Punkt auf der Kurve, so gilt

$$u(x, y) = \sum_{i,j=0}^{\infty} \frac{(x - x_0)^i (y - y_0)^j}{i!j!} \partial_x^i \partial_y^j u(x_0, y_0),$$

und die Lösung u ist auf einer Umgebung der Kurve bekannt. Man nennt diese Aufgabe die *Cauchysche Anfangswertaufgabe* entlang der Anfangskurve Γ. Bei partiellen Differentialgleichungen zweiter Ordnung ist es eher ungewöhnlich, dass entlang einer Kurve Γ mehrere Anfangswerte, also der Wert und die Ableitungen, vorgegeben werden.

Fall 2: $\det(L) = 0$ in einem Punkt $(x_0, y_0) \in \Gamma$ auf der Kurve. Die Gleichung

$$a_{11}(\tau'_y)^2 - 2a_{12}\tau'_x\tau'_y + a_{22}(\tau'_x)^2 = 0 \tag{1.3}$$

ist in (x_0, y_0) also lösbar. Wir gehen davon aus, dass $a_{11} \neq 0$ und $\tau'_x \neq 0$.[1] Dann ist:

$$\left(\frac{\tau'_y}{\tau'_x}\right)^2 - \frac{2a_{12}}{a_{11}}\frac{\tau'_y}{\tau'_x} + \frac{a_{22}}{a_{11}} = 0.$$

Der Quotient $\delta := \tau'_y/\tau'_x = dy/dx$ bestimmt die Richtung einer Kurve, welche wir in $(x_0, y_0) \in \Gamma$ durch einen Graph $y = y(x)$ oder $x = x(y)$ beschreiben können. Entlang dieses Graphen ist die quadratische Gleichung (1.3) lösbar. Dies bedeutet gerade, dass entlang des Graphen das lineare Gleichungssystem (1.2) nicht lösbar ist. Wir nennen eine solche Kurve eine *Charakteristik*. Für die Steigungen dieser Kurve gilt:

$$\delta_{1/2} = \frac{a_{12}}{a_{11}} \pm \frac{1}{a_{11}}\sqrt{a_{12}^2 - a_{11}a_{22}}.$$

Je nach Vorzeichen von $a_{12}^2 - a_{11}a_{22}$ existieren nun keine, eine oder zwei verschiedene Charakteristiken der Differentialgleichung.

Wir unterscheiden

Elliptischer Fall $a_{12}^2 - a_{11}a_{22} < 0$: Es existiert keine reelle Nullstelle und somit existiert keine Charakteristik durch den Punkt (x_0, y_0). Dieser Punkt ist zwar Nullstelle von (1.3), jedoch eine isolierte Nullstelle. In benachbarten Punkten ist die Taylor-Entwicklung durchf"uhrbar und die Lösung u kann bestimmt werden.

Parabolischer Fall $a_{12}^2 - a_{11}a_{22} = 0$: Es existiert somit genau eine Richtung $\delta = a_{12}/a_{11}$ in welcher die Gleichung (1.3) lösbar ist, das lineare Gleichungssystem (1.2) also nicht lösbar ist. Die Differentialgleichung lässt sich also bei Kenntnis von u und ∇u entlang einer Kurve Γ nur in gewisse Richtungen entwickeln, nicht jedoch über die Charakteristik hinweg.

Hyperbolischer Fall $a_{12}^2 - a_{11}a_{22} > 0$: Es existieren zwei Charakteristiken und die Lösung lässt sich nur in den entsprechenden Abschnitten entwickeln, in denen die Kurve Γ nicht mit den Charakteristiken zusammenfällt.

[1] Wenn dies doch der Fall ist, so ist die gehen wir davon aus, dass $a_{22} \neq 0$ und $\tau'_y \neq 0$. Können beide Bedingungen nicht erfüllt werden so kann nicht $\tau'_x = 0$ und $\tau'_y = 0$ gelten. (Sonst wäre die Kurve ein Punkt). Es folgt dann z.B. $a_{11} = 0$, $a_{22} \neq 0$ und $\tau'_y = 0$ aber $\tau'_x \neq 0$ und somit $\det(\mathbf{L}) = a_{22}(\tau'_x)^2 \neq 0$ im Widerspruch zu $\det(\mathbf{L}) = 0$.

Die Herkunft der Namen *elliptisch, parabolisch* und *hyperbolisch* ist einfach erklärt und kommt von der Betrachtung von quadratischen Gleichungen des Typs

$$Q(x,y) := a_{11}x^2 + 2a_{12}xy + a_{22}y^2 + a_1x + a_2y + a,$$

deren Null-Niveau $Q(x,y) = 0$ je nach Koeffizienten gerade eine Ellipse, Parabel oder Hyperbel beschreibt:

$$a_{12}^2 - a_{11}a_{22} \begin{cases} < 0 & \text{Ellipse} \\ = 0 & \text{Parabel} \\ > 0 & \text{Hyperbel} \end{cases}.$$

Der Typ des Schnittes hängt nur vom Hauptteil, also von a_{11}, a_{22} und a_{12} ab.

1.1.2 Normalformen von linearen partiellen Differentialgleichungen zweiter Ordnung

Die Typeinteilung partieller Differentialgleichungen hängt nur vom *Hauptteil* L_0 ab:

$$L_0 = a_{11}\partial_x^2 + 2a_{12}\partial_{xy} + a_{22}\partial_y^2.$$

Wir schreiben diesen als

$$L_0 = \begin{pmatrix} \partial_x \\ \partial_y \end{pmatrix}^T \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix} \begin{pmatrix} \partial_x \\ \partial_y \end{pmatrix},$$

mit einer symmetrischen Matrix A. Diese Matrix hat zwei Eigenwerte $\lambda_1, \lambda_2 \in \mathbb{R}$ und zwei orthonormalen Eigenvektoren $\omega_1, \omega_2 \in \mathbb{R}^2$. Hiermit gilt die Normalform

$$\begin{pmatrix} \mu_1^1 & \mu_2^1 \\ \mu_1^2 & \mu_2^2 \end{pmatrix}^T \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix} \begin{pmatrix} \mu_1^1 & \mu_2^1 \\ \mu_1^2 & \mu_2^2 \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix},$$

bzw. die Normalform des Differentialoperators

$$L_0 = \lambda_1\partial_{\mu_1}^2 + \lambda_2\partial_{\mu_2}^2.$$

Die Eigenwerte sind die Nullstellen des charakteristischen Polynoms

$$\begin{aligned} \det(A - \lambda I) &= \lambda^2 - (a_{11} + a_{22})\lambda + a_{11}a_{22} - a_{12}^2 \\ &= (\lambda - \lambda_1)(\lambda - \lambda_2) \\ &= \lambda^2 - (\lambda_1 + \lambda_2)\lambda + \lambda_1\lambda_2. \end{aligned} \tag{1.4}$$

Wir können nun die verschiedenen Fälle unterscheiden:

Elliptischer Fall $a_{11}a_{22} - a_{12} = \lambda_1\lambda_2 > 0$**:** Beide Eigenwerte sind von Null verschieden und o.E. beide positiv. Die Normalform einer *elliptischen partiellen Differentialgleichung* ist die *Laplace-Gleichung*

$$\partial_{\mu_1}^2 + \partial_{\mu_2}^2 u = f,$$

bzw.

$$\Delta u = f.$$

Parabolischer Fall $a_{12}^2 - a_{11}a_{22} = \lambda_1\lambda_2 = 0$**:** Genau ein Eigenwert ist null, ein Eigenwert ist ungleich null. (Ansonsten könnte nicht $a_{11} \neq 0$ und $\tau_x' \neq 0$ sein). Die Normalform der *parabolischen partiellen Differentialgleichung* ist

$$L_0 = \partial_{\mu_1}^2 u + \psi(u, \nabla u) = f,$$

Die prototypische parabolische Gleichung ist die Wärmeleitungsgleichung

$$\partial_t u - \Delta u = f,$$

mit einer ausgezeichneten Richtung t.[2] Dies kennzeichet gerade die Existenz einer Charakteristik.

Hyperbolischer Fall $a_{12}^2 - a_{11}a_{22} = \lambda_1\lambda_2 < 0$**:** Beide Eigenwerte sind ungleich Null und haben unterschiedliches Vorzeichen. Die Normalform der *hyperbolischen partiellen Differentialgleichung* ist

$$L_0 = \partial_{\mu_1}^2 u - \psi(u, \nabla u) = f,$$

und der prototypische Fall ist die *Wellengleichung*

$$\partial_{tt} u - \Delta u = f.$$

1.1.3 Propagation of information

The three different types of second order differential equation show a kind of information propagation. For elliptic equations, we can reconstruct all higher order derivatives from $u(x_0, y_0)$ and $\nabla u(x_0, y_0)$. This however also means, that change of $u(x_0, y_0)$ or $\nabla u(x_0, y_0)$ will influence the Taylor expansion of u around (x_0, y_0). Therefore, change of u or ∇u in a single point (x_0, y_0) has possible impact on u on the whole domain Ω. In elliptic equations, the solution $u(x, y)$ for $(x, y) \in \Omega$ will depend on the solution $u(x', y')$ in any other point $(x', y') \in \Omega$. Propagation of information has infinite velocity in all directions. We depict this sitation in Figure 1.1 (right).

Next, we consider the parabolic problem

$$\partial_t u(t, x) - \partial_{xx} u(t, x) = f.$$

Comparing the derivation in Section 1.1.1, $a_{11} = -1$, $a_{22} = 0$ and $a_{12} = 0$ reveals the direction

$$\delta = \frac{\tau_y'}{\tau_x'} = 0,$$

and the linear system (1.2) is *not* regular along all curves with $\tau_y' = 0$ which are the horizontal lines. We show the situation in the middle plot of Figure 1.1. In x-direction, information is spread like in the elliptic case. In t-direction we can write

$$u(t, x) = \int_{t_0}^{t} \big\{ f(s, x) + \partial_{xx} u(s, x) \big\} ds.$$

[2] Wir hätten genauso gut $\partial_x u - \partial_{yy} u = f$ schreiben können. In Anwendungen, z.B. bei der Wärmeleitung nimmt jedoch t die Rolle der Zeit ein und $x, y, \ldots$ die Rolle des Orts.

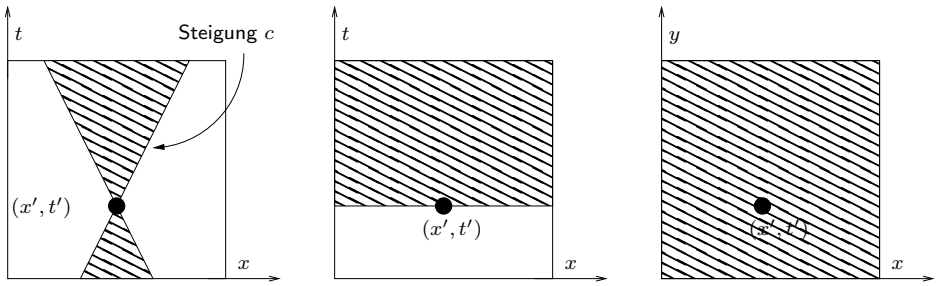


Abbildung 1.1: Typeeinteilung und Ausbreitung von Informationen bei partiellen Differentialgleichungen. Von links nach rechts: hyperbolisch, parabolisch und elliptisch.

The solution in (t, x) depends on all (s, y) for $t_0 \leqslant s \leqslant t$ and all t. This corresponds to our understanding of the *idealized heat equation*: assume a piece of metal is our spatial domain. At time t_0 it has a given temperature. The temperature at times $t > t_0$ will depend on this initial temperature, while times $t < t_0$ do not depend on the "future" t_0. In the spatial domain,

information spreads out with infinite speed. Heating of the plate at time t_0 in a point x may influence the temperature in all points y for any time $t > t_0$. This of course is idealized, as temperature will have a finite spatial propagation speed in reality. (The mathematical heat equation only gives an approximation of reality).

Finally, we consider the wave equation.

$$\partial_{tt} u(t, x) - \partial_{xx} u(t, x) = 0.$$

Hyperbolic equations allow for a second prototypical form. Change of variables

$$t \mapsto r - s, \quad x \mapsto r + s \quad \Leftrightarrow \quad r \mapsto \frac{x + t}{2}, \quad s \mapsto \frac{x - t}{2}$$

with $v(r, s) := u(t, x) = u(r - s, r + s)$ gives

$$\partial_{tt} u(t, x) = \partial_{tt} v(r, s) = \frac{1}{2}\partial_t(\partial_r v - \partial_s v) = \frac{1}{4}\left(\partial_{rr} v - 2\partial_{rs} v + \partial_{ss} v\right)$$
$$\partial_{xx} u(t, x) = \partial_{xx} v(r, s) = \frac{1}{2}\partial_x(\partial_r v + \partial_s v) = \frac{1}{4}\left(\partial_{rr} v + 2\partial_{rs} v + \partial_{ss} v\right)$$

The wave equation is then equivalent to

$$\partial_{rs} v(r, s) = 0.$$

This shows, that a solution must take the form

$$v(r, s) = v_1(r) + v_2(s),$$

and hence

$$u(t, x) = u_1(x + t) + u_2(x - t),$$

where u_1 and u_2 are set accordingly. Assume $\Gamma = \{(0, s),\ s \in \mathbb{R}\}$ and $u(0, x) = u^0(x)$ and $\partial_t u(0, x) = u^1(x)$ are known. Then,

$$\begin{aligned} u^0(x) &= u_1(x) + u_2(x) \\ u^1(x) &= u_1(x) - u_2(x) \end{aligned} \quad \Rightarrow \quad \begin{aligned} u_1(x) &= \frac{1}{2}(u^0(x) + u^1(x)) \\ u_2(x) &= \frac{1}{2}(u^0(x) - u^1(x)) \end{aligned}$$

and we can construct the solution as

$$u(t, x) = \frac{1}{2}\big(u^0(x + t) + u^1(x + t) + u^0(x - t) - u^1(x - t)\big).$$

Hereby we can deduce, that the speed of information propagation is finite and can never exceed the characteristics $t = \pm x$, see Figure 1.1 (left).

2 Theoretische Grundlagen

Die theoretische Analyse von partiellen Differentialgleichungen ist zumeist nicht mit den elementaren Methoden möglich, welche bei gewöhnlichen Differentialgleichungen ausreichend waren. Die Frage nach der Existenz einer Lösung muss bei partiellen Differentialgleichungen neu formuliert werden: statt dem bloßen "existiert eine Lösung?" tritt die Frage "in welchem Sinne und in welchem Funktionenraum kann die Existenz einer Lösung gezeigt werden?" in den Vordergrund.

2.1 The Laplace Equation and the Sobolev space $H^1(\Omega)$

We are studying the Laplace equation on a domain $\Omega \subset \mathbb{R}^d$ with $d = 2, 3$

$$u \in C^2(\Omega) \cap C(\bar{\Omega}): \quad -\Delta u = f \text{ in } \Omega,\ u = 0 \text{ on } \partial\Omega, \tag{2.1}$$

where $f \in C(\Omega)$. We call such a solution the *classical solution* to the Laplace problem, also the *strong solution*.

Notation

Let $f, g \in L^2(\Omega)$, we denote by (all the same)

$$(f, g)_{L^2(\Omega)} = (f, g)_\Omega = (f, g) = \int_\Omega f(x) \cdot g(x)\, dx$$

the L^2-scalar product on Ω and by

$$\|f\|_{L^2(\Omega)} = \|f\|_\Omega = \|f\| = \left(\int_\Omega f(x)^2\, dx \right)^{\frac{1}{2}}$$

the L^2-norm on Ω.

We show, that a solution $u \in C^2(\Omega) \cap C(\bar{\Omega})$ to (2.1) is also given as solution to a minimization problem

Lemma 2.1 (Variational formulation and minimization Problem)**.** *Let $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ be a bounded domain. Let $u \in C^2(\Omega) \cap C(\bar{\Omega})$ be solution to the Laplace equation (2.1). Then, u is also solution to the* variational formulation

$$u \in V_0: \quad (\nabla u, \nabla v) = (f, v) \quad \forall v \in V_0, \tag{2.2}$$

which is equivalent to the solution of the minimization problem

$$u \in V_0 : \quad E(u) \leqslant E(v) := \frac{1}{2}\|\nabla v\|^2 - (f, v), \tag{2.3}$$

where

$$V_0 := \{\phi \in C^1(\Omega) \cap C(\bar{\Omega}),\ \phi = 0 \text{ on } \partial\Omega\}.$$

Every solution $u \in V_0$ *to the* minimization problem *and* variational formulation, *that also satisfies* $u \in C^2(\Omega) \cap C(\bar{\Omega})$ *is a* classical solution *to the Laplace problem* (2.1).

Proof. (*i*) We show (2.1) $\Rightarrow$ (2.2). Let $v \in V_0$ be arbitrary. Then, every solution u to (2.1) is solution to

$$-\Delta u \cdot v = f \cdot v \quad \Rightarrow \quad \int_\Omega -\Delta u \cdot v \,\mathrm{d}x = \int_\Omega f \cdot v \,\mathrm{d}x.$$

As $v \in C^1(\Omega) \cap C(\bar{\Omega})$ we can apply integration by parts (Greens formula) to get

$$\int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x - \int_{\partial\Omega} \partial_n u \underbrace{\cdot v}_{=0} \,\mathrm{d}s = \int_\Omega f v \,\mathrm{d}x.$$

This is exactly the *variational formulation*.

(*ii*) We show (2.2) $\Leftrightarrow$ (2.3). For $u, \phi \in V_0$ (note, that $u + \phi \in V_0$) it holds

$$\begin{aligned} E(u+\phi) &= \frac{1}{2}\|\nabla(u+\phi)\|^2 - (f, u+\phi) \\ &= \frac{1}{2}\|\nabla u\|^2 + (\nabla u, \nabla\phi) + \frac{1}{2}\|\nabla\phi\|^2 - (f,u) - (f,\phi) \\ &= E(u) + (\nabla u, \nabla\phi) - (f,\phi) + \frac{1}{2}\|\nabla\phi\|^2. \end{aligned} \tag{2.4}$$

First assume, that u solves the *variational formulation*. Then

$$E(u+\phi) = E(u) + \underbrace{(\nabla u, \nabla\phi) - (f,\phi)}_{=0} + \frac{1}{2}\underbrace{\|\nabla\phi\|^2}_{\geqslant 0},$$

and it holds

$$E(u) \leqslant E(v) \quad \forall v := u + \phi \in V_0,$$

such that u is also a minimizer. Now, let u be solution to the minization problem. By (2.4) we have

$$(f,\phi) - (\nabla u, \nabla\phi) = \underbrace{E(u) - E(u+\phi)}_{\leqslant 0} + \frac{1}{2}\|\nabla\phi\|^2 \leqslant \frac{1}{2}\|\nabla\phi\|^2.$$

We choose $\phi := s\phi_0$ with $\|\nabla\phi_0\| = 1$ to get

$$s\Big((f,\phi_0) - (\nabla u, \nabla\phi_0)\Big) \leqslant \frac{s^2}{2}.$$

As $\phi \in V_0$ is arbitrary and s has an arbitrary sign, it follows, that

$$-\frac{|s|}{2} \leqslant (f, \phi_0) - (\nabla u, \nabla \phi_0) \leqslant \frac{|s|}{2}.$$

Choosing $s \to 0$ shows, that $u \in V_0$ is also solution to the variational problem.

(iii) We show $(2.2) \Rightarrow (2.1)$, if $u \in V_0 \cap C^2(\Omega)$ has sufficient regularity. Then, all steps in (i) are equivalent transformations until we get

$$\int_\Omega -\Delta u \cdot v \, dx = \int_\Omega f \cdot v \, dx \quad \forall v \in V_0.$$

As $C_0^\infty(\Omega) \subset V_0$, the fundamental *Theorem of Variations*, see [**?**], gives the point-wise relation

$$-\Delta u = f \text{ in } \Omega.$$

The boundary condition $u = 0$ was already enforced in V_0. Therefore, given sufficient regularity, a variational solution (such as the solution of the minimization problem) is also a classical solution. □

This lemma reveals a new definition of the Laplace problem. It defined a *weaker* solution, as it is well-defined for $u \in C^1(\Omega) \cap C(\bar{\Omega})$ and no second derivatives are required. However, there may be (and there are) variational solutions, that are not classical solutions.

We can easily show uniqueness of a variational solution

Lemma 2.2 (Uniqueness of variational solutions)**.** *Let $u_1, u_2 \in V_0$ be solutions to the variational formulation (2.2) for one $f \in C(\Omega)$. Then, $u_1 = u_2$.*

Proof. Let $w := u_1 - u_2 \in V_0$. It holds for all $v \in V_0$

$$(\nabla w, \nabla v) = (\nabla u_1, \nabla v) - (\nabla u_2, \nabla v) = (f, v) - (f, v) = 0.$$

We choose $v = w$ to get with Poincaré's inequality

$$\|\nabla w\|^2 \geqslant c_p^{-2} \|w\|^2 = 0 \quad \Rightarrow \quad w = 0.$$

Therefore, $u_1 = u_2$. □

Likewise, we can show, that a solution to the variational formulation will continuously depend on the right hand side.

Lemma 2.3 (Continuity of the variational solution)**.** *Let $f \in C \cap L^2(\Omega)$ and let $u \in V_0$ be solution to the variational formulation (2.2). It holds*

$$\|\nabla u\| \leqslant c_p \|f\|,$$

where $c_p > 0$ is the constant from Poincaré's inequality, see Lemma 2.5.

Proof. It holds for all $v \in V_0$

$$(\nabla u, \nabla v) = (f, v),$$

and therefore, choosing $v = u$ and using Poincarés inequality

$$\|\nabla u\|^2 \leqslant \|f\| \, \|u\| \leqslant c_p \|f\| \, \|\nabla u\|.$$

Hence

$$\|\nabla u\| \leqslant c_p \|f\|.$$

□

The minimization formulation allows for a physical interpretation. By

$$E(v) = \frac{1}{2}\|\nabla v\|^2 - (f, v)$$

we can describe (in a simplified way) the energy of a membrane, that is fixed $v = 0$ at the boundary of the domain $\partial\Omega$ and that undergoes a deflection (given by v) due to a force f that is acting on the domain. Physical principles say, that the system will take the state of minimal energy, which is given by (2.3). The model is accurate for (theoretical) membranes without a mass and without a width.

We show

Theorem 2.4 (Solution to the minimization problem)**.** *Let $\Omega \subset \mathbb{R}^d$ for $d \geqslant 1$ be a domain and $f \in L^2(\Omega)$. Then, there exists a solution*

$$u \in H_0^1(\Omega),$$

that solves the minimization problem

$$E(u) \leqslant E(v) \quad \forall v \in H_0^1(\Omega).$$

By $H_0^1(\Omega)$ we denote the completion of $C_0^\infty(\Omega)$ with respect to $\|\nabla \cdot \|_{L^2(\Omega)}$.

Proof. The proof takes several steps. We start with the formulation of the minimization problem (2.3) using the Hölder space V_0.
(i) We show, that $E(\cdot)$ is bounded

$$E(v) = \frac{1}{2}\|\nabla v\|^2 - (f, v) \geqslant \frac{1}{2}\|\nabla v\|^2 - \|f\| \, \|v\|.$$

We note, that $v = 0$ on $\partial\Omega$ and use the *Poincaré inequality*, see Lemma 2.5, to get

$$E(v) \geqslant \frac{1}{2}\|\nabla v\|^2 - c_p \|f\| \, \|\nabla v\|.$$

With *Youngs inequality* $ab \leqslant \frac{\varepsilon}{2}a^2 + \frac{1}{2\varepsilon}b^2$ for all $a, b \in \mathbb{R}$ and $\varepsilon > 0$ we get (choosing $\varepsilon = 1$)

$$E(v) \geqslant \frac{1}{2}\|\nabla v\|^2 - \frac{c_p^2}{2}\|f\|^2 - \frac{1}{2}\|\nabla v\|^2 = -\frac{c_p^2}{2}\|f\|^2 > -\infty.$$

(*ii*) As

$$\inf_{v \in V_0} E(v) > -\infty,$$

we can choose a *minimal sequence* $v_n \in V_0$ of $E(\cdot)$

$$E(v_n) \to \int_{v \in V} E(v) =: d > -\infty.$$

We show, that v_n is Cauchy in $\|\nabla \cdot\|$. Using the *parallelogram law* on $v, w \in V_0$ we have

$$\|\nabla(v-w)\|^2 + \|\nabla(v+w)\|^2 = 2\|\nabla v\|^2 + 2\|\nabla w\|^2,$$

and hence

$$\begin{aligned}
\|\nabla(v_n - v_m)\|^2 &= 2\|\nabla v_n\|^2 + 2\|\nabla v_m\|^2 - \|\nabla(v_n + v_m)\|^2 \\
&= 2\|\nabla v_n\|^2 + 2\|\nabla v_m\|^2 - 4\|\frac{1}{2}\nabla(v_n + v_m)\|^2 \\
&= 4E(v_n) + 4E(v_m) - 8E\big(\frac{1}{2}(v_n + v_m)\big) \\
&\qquad + 4(f, v_n) + 4(f, v_m) - 8(f, \frac{1}{2}(v_n + v_m))
\end{aligned}$$

It holds $E(v_n) \to d$ for $n \to \infty$. For the mixed term, it holds

$$E(\frac{1}{2}(v_n + v_m)) \geqslant d.$$

For the difference $\|\nabla(v_n - v_m)$ it therefore holds

$$\limsup_{n,m \to \infty} \|v_n - v_m\|^2 \leqslant 4d + 4d - 8d = 0.$$

We must consider $\limsup$ as the mixed term is not necessarily converging.

(*iii*) The last crucial step is to show convergence of this Cauchy sequence. The space $V_0 = C^1(\Omega) \cap C(\bar{\Omega})$ is not complete with respect to the gradient L^2-norm $\|\nabla \cdot\|_{L^2(\Omega)}$. We give a counter-example. On can show, that

$$v_n := \log\left(\log\left(\frac{1}{|x| + \frac{1}{n}}\right) + 1\right) \in V_0$$

on $\Omega = \{x \in \mathbb{R}^2 : \ |x| \leqslant 1\}$. Further, V_n is Cauchy. The limit function

$$v := \log\left(\log\left(\frac{1}{|x|}\right) + 1\right)$$

still satisfies $\|\nabla v\| < \infty$. However, v is not even bounded in the origin, therefore $v \notin V_0$. We define the solution space

$H_0^1(\Omega) := \{\phi$ is the equivalence class of Cauchy sequences in V_0 with respect to $\|\nabla \cdot\|_{L^2\Omega)}\}$.

The space $H_0^1(\Omega)$ is the completion of V_0 (and also of $C_0^\infty(\Omega)$) with respect to $\|\nabla \cdot\|$. □

Similar to $H_0^1(\Omega)$ we define $H^1(\Omega)$ as completion of $C^\infty(\Omega)$ with respect to $\|\nabla \cdot\|$. We will call these function spaces *Sobolev spaces*. The characterizaiton of functions $v \in H^1(\Omega)$ or $v \in H_0^1(\Omega)$ is difficult and often, we will simply use the fact, that for $v \in H^1(\Omega)$ there exists a Cauchy sequence $v_n \in C^\infty(\Omega)$ with $\|\nabla(v_n - v)\| \to 0$. We have already seen, that H^1 functions must not even be continuous.

As an example for a typical argumentation, we give a proof of the very important *Poincaré inequality*

Lemma 2.5 (Poincaré inequality)**.** *Let Ω be a domain. There exists a constant $c_p > 0$, such that for all $v \in H_0^1(\Omega)$ it holds*

$$\|v\|_{L^2(\Omega)} \leqslant c_p \|\nabla v\|_{L^2(\Omega)}.$$

The constant $c_p = c_P(\operatorname{diam}(\Omega)$ depends on the size of the domain.

Proof. We show the inequality for $\Omega \subset \mathbb{R}^2$. Let $Q \in \mathbb{R}^2$ be a square with length $L > 0$ such that $\Omega \subset Q$. Without loss of generality, let $Q = (0, L)^2$.

(*i*) Let $v \in C_0^\infty(\Omega)$. By $\hat{v}$ we denote the trivial extension of v from Ω to Q by zero. It holds $v \in C_0^\infty(Q)$. Let $(x, y) \in \Omega$. It holds

$$v(x, y) = \underbrace{v(0, y)}_{=0} + \int_0^x \partial_x v(s, y)\, ds$$

and by taking the square and using Hölder's inequality

$$v(x, y)^2 \leqslant \int_0^x |\partial_x v(s, y)|^2\, ds \int_0^x 1\, ds.$$

We estimate

$$v(x, y)^2 \leqslant L \int_0^L |\nabla v|^2\, ds$$

and integrate over Q to get

$$\|v\|_{L^2(Q)}^2 = \|v\|_{L^2(\Omega)}^2 \leqslant L^2 \|\nabla v\|_{L^2(\Omega)}^2.$$

We choose $c_p = L$.

(*ii*) Now, let $v \in H_0^1(\Omega)$ and $v_n \in C_0^\infty(\Omega)$ be such, that $\|\nabla(v_n - v)\| \to 0$ and $\|v - v_n\| \to 0$ for $n \to \infty$. It holds

$$\|v\| \leqslant \|v - v_n\| + \|v_n\| \leqslant \|v - v_n\| + c_p \|\nabla v_n\| \leqslant \|v - v_n\| + c_p \|\nabla(v - v_n)\| + c_p \|\nabla v_n\|.$$

Going to the limit $n \to \infty$ gives Poincaré's inequality for $v \in H_0^1(\Omega)$. □

This kind of argumentation is typical. We first show an estimate in Hölder spaces and then use continuity to go to $H^1(\Omega)$.

Poincaré's inequality says, that the *semi-norm* $\|\nabla \cdot \|$ is a *norm* on the space $H_0^1(\Omega)$, as

$$0 = \|\nabla v\| \geqslant c_p^{-1}\|v\| \quad \Rightarrow \quad v = 0.$$

Notation

Let $\Omega \subset \mathbb{R}^d$ be a domain. Let $v \in H^1(\Omega)$. By

$$\|\nabla v\|_{L^2(\Omega)} = \left(\int_\Omega |\nabla v|^2 \, dx\right)^{\frac{1}{2}}$$

we define the H^1-seminorm, which - due to Poincaré's inequality - is a norm on $H_0^1(\Omega)$. By

$$\|v\|_{H^1(\Omega)} = \left(\|v\|^2_{L^2(\Omega)} + \|\nabla v\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}}.$$

we denote the H^1-norm. In $H_0^1(\Omega)$ semi-norm and norm are equivalent. In $H_0^1(\Omega)$ we denote by $\|\nabla v\|_{L^2(\Omega)}$ the energy norm as this is the main part of the energy functional $E(v)$ describing the "energy of a membrane".

We can now define the Laplace problem in variational formulation for all right hand sides $f \in L^2(\Omega)$ using test-functions $\phi \in H_0^1(\Omega)$.

Definition 2.6 (Variational formulation, weak solution)**.** *Let $f \in L^2(\Omega)$ with "smooth boundary" $\partial\Omega$. We call $u \in H_0^1(\Omega)$ the* weak solution *to the* variational formulation

$$(\nabla u, \nabla \phi) = (f, \phi) \quad \forall \phi \in H_0^1(\Omega)$$

Another basic inequality is the *trace inequality*.

Lemma 2.7 (Trace inequality)**.** *Let $\Omega \subset \mathbb{R}^d$ be a domain. There exists a constant $c_t = c_t(\Omega)$, such that for all $v \in C^1(\bar{\Omega}) \cap C(\bar{\Omega})$ it holds*

$$\|\gamma(v)\|_{L^2(\partial\Omega)} \leqslant c_t \|v\|_{H^1(\Omega)},$$

where by $\gamma : H^1(\Omega) \to L^2(\partial\Omega)$ we denote the trace operator.

Proof. (i) First, let $\Omega = (0, L)^2$ be a square. Let $(x, y) \in \Omega$. It holds

$$v(x, 0) = v(x, y) - \int_0^x \partial_x v(s, y) \, ds,$$

and therefore (as $(a-b)^2 \leqslant 2a^2 + 2b^2$)

$$\begin{aligned} v(x,0)^2 &\leqslant 2v(x,y)^2 + 2\int_0^x |\partial_x v(s,y)|^2 \,\mathrm{d}s \int_0^x 1 \,\mathrm{d}x \\ &\leqslant 2v(x,y)^2 + 2x\int_0^x |\partial_x v(s,y)|^2 \,\mathrm{d}s. \end{aligned}$$

We integrate over x and y to get

$$L\|v\|^2_{\Gamma_{\text{left}}} \leqslant 2\|v\|^2_\Omega + L\|\nabla v\|^2_\Omega \quad \Rightarrow \quad \|v\|^2_{\Gamma_{\text{left}}} \leqslant c_t^2 \|v\|^2_{H^1(\Omega)},$$

where by Γ_{left} we denote the left segment of the boundary. We can use the same argument for the other three parts of the boundary to get $\|v\|_\Gamma \leqslant c_t \|v\|_{H^1(\Omega)}$.

(*ii*) For domains with polygonal boundary, we can use this argument for every segment of the boundary. On general curved domains Ω we can locally introduce mappings to map Ω onto a domain with straight boundary. Here, we use the argument as above. Definition of these mappings as well as glueing the different parts together is technical and we refer to the literature [8, 3, 7]. □

We have derived the trace inequality for functions in $C^1(\Omega) \cap C(\bar{\Omega})$. Let $v \in H^1(\Omega)$ and $v_n \in C^1(\Omega) \cap C(\bar{\Omega})$ a Cauchy sequence with $\|\nabla(v_n - v)\| \to 0$. Then, it holds

$$\|v_n - v_m\|_{\partial\Omega} \leqslant c_t \|\nabla(v_n - v_m)\|_{H^1(\Omega)} \to 0 \quad (n \to \infty).$$

Hence, v_n is Cauchy on $\partial\Omega$ with respect to the norm $L^2(\partial\Omega)$. As $L^2(\partial\Omega)$ is a Banach space, this sequence converges in $L^2(\partial\Omega)$, such that we can introduce the *trace of* $H^1(\Omega)$*-functions* in the space $L^2(\Omega)$.

Every H^1-function has a L^2-trace on the boundary, e.g.

$$\|v\|_{\partial\Omega} \leqslant c_t \|v\|_{H^1(\Omega)} \quad \forall v \in H^1(\Omega).$$

This estimate is not sharp, i.e., there are $L^2(\partial\Omega)$-functions, that are not trace of a $H^1(\Omega)$ function. We can however define the exact space of functions, that are trace of a H^1 function.

Definition 2.8 (Trace)**.** *Let $\Omega \in \mathbb{R}^d$ be a domain. We define the space of traces as*

$$H^{1/2}(\partial\Omega) = \{\phi \in L^2(\partial\Omega) \,|\, \phi \textit{ is trace of } \phi' \in H^1(\Omega) \textit{ on } \partial\Omega\}.$$

By

$$\|v\|_{H^{1/2}(\partial\Omega)} := \inf\{\|\phi\|_{H^1(\Omega)} \,|\, \forall \phi \in H^1(\Omega) \textit{ where } v \textit{ is trace of } \phi \textit{ on } \partial\Omega\}$$

we denote the $H^{1/2}$-norm.

There is a constant $c = c(\Omega)$, such that

$$\|v\|_{L^2(\partial\Omega)} \leqslant c\|v\|_{H^{1/2}(\partial\Omega)} \quad \forall v \in H^{1/2}(\partial\Omega).$$

However, there exists no such constant $c = c(\Omega) < \infty$ that

$$\lightning \quad \|v\|_{H^{1/2}(\partial\Omega)} \leqslant c\|v\|_{L^2(\partial\Omega)} \quad \forall v \in L^2(\partial\Omega) \quad \lightning.$$

A further important theorem - that however is less simple to proof - is the inverse of the trace inequality

Lemma 2.9 (Inverse trace lemma)**.** *Let $\Omega \subset \mathbb{R}^d$ be a domain. There exists a constant $c_{it} = c(\Omega)$, such that for all $v \in H^{1/2}(\partial\Omega)$ there exists a $\tilde{v} \in H^1(\Omega)$ with*

$$\|\tilde{v}\|_{H^1(\Omega)} \leqslant c_{it}\|v\|_{H^{1/2}(\partial\Omega)}.$$

For a proof, see [8].

Hereby, we can generalize the Laplace problem to non-homogenous Dirichlet conditions

Lemma 2.10 (Dirichlet problem)**.** *Let $\Omega \subset \mathbb{R}^d$ be a domain, $f \in L^2(\Omega)$ and $g \in H^{1/2}(\partial\Omega)$ be given right hand side. Then, the variational formulation of the Laplace problem*

$$u \in H^1(\Omega): \quad (\nabla u, \nabla\phi) = (f, \phi) \quad \forall \phi \in H^1_0(\Omega), \quad u = g \text{ on } \partial\Omega$$

has a unique solution.

Proof. Be Lemma 2.9, there exists a function $\tilde{g} \in H^1(\Omega)$ with $\tilde{g}|_{\partial\Omega} = g$. Let

$$u_0 := u - \tilde{g} \in H^1_0(\Omega).$$

Any solution to the variational formulation is given by

$$(\nabla u_0, \nabla\phi) = (f, \phi) + (\nabla\tilde{g}, \nabla\phi) \quad \forall \phi \in H^1_0(\Omega).$$

Now, existence and uniqueness can be shown similar to the previous argumentation. □

We have seen, that H^1-functions are not necessarily continuous or even differentiable. The *embedding*

$$\lightning \quad H^1(\Omega) \hookrightarrow C(\Omega) \quad \lightning$$

is not true in the general case. It holds however, that

Lemma 2.11 (Rellich Theorem)**.** *The embedding $H^1(\Omega) \hookrightarrow L^2(\Omega)$ is compact. Every sequence $v_n \in H^1(\Omega)$ that is bound in $H^1(\Omega)$ has a converging subsequence $v_{n'}$ in $L^2(\Omega)$. The limit $v_{n'} \to v \in L^2(\Omega)$ is a member of $v \in H^1(\Omega)$.*

See Wloka [8].

For one-dimensional domains $I = (a, b) \subset \mathbb{R}$ we can easily show, that $H_0^1(I)$-functions must be bounded, as

$$v(x) = \int_0^x v'(s)\,ds \quad \Rightarrow \quad |v(x)|^2 \leqslant \int_0^x 1\,ds \int_0^x |v'(s)|^2\,ds \leqslant |b-a|\|\nabla v\|_I^2.$$

Another approach to the Sobolev spaces $H^1(\Omega)$ and $H_0^1(\Omega)$ is given by the concept of *weak derivatives*.

Notation

Let $\Omega \subset \mathbb{R}^d$ be a domain. For $1 \leqslant p < \infty$ we define the L^p-norm as

$$\|f\|_{L^p(\Omega)} := \left(\int_\Omega |f|^p\right)^{\frac{1}{p}}.$$

For $p = \infty$ we define the L^∞-norm as

$$\|f\|_{L^\infty(\Omega)} := \operatorname{ess\,sup}_{x\in\Omega} |f(x)|.$$

For $1 \leqslant p \leqslant \infty$ we define the spaces

$$L^p(\Omega) := \{\phi : \Omega \to \mathbb{R} : \ \|\phi\|_{L^p(\Omega)} < \infty\}.$$

Definition 2.12 (Weak derivative)**.** *Let* $u \in L^1(\Omega)$*. A function* $w \in L^1(\Omega)$ *is called* first weak derivative *of* u*, if it holds*

$$\int_\Omega w \cdot \phi\,dx = -\int_\Omega u \cdot \partial_x \phi\,dx \quad \forall \phi \in C_0^\infty(\Omega).$$

The generalization is obvious. For a multiindex $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ we denote by $w \in L^1(\Omega)$ the weak derivative of degree $|\alpha| = \alpha_1 + \cdots + \alpha_d$, if

$$\int_\Omega w \cdot \phi\,dx = (-1)^{|\alpha|} \int_\Omega u \cdot D^\alpha \phi\,dx \quad \forall \phi \in C_0^\infty(\Omega),$$

for

$$D^\alpha = \partial_1^{\alpha_1} \cdots \partial_d^{\alpha_d}.$$

If v is differentiable in the common sense, weak and classical derivatives are the same. With this concept, we can define

Definition 2.13 (Sobolev spaces)**.** *For* $k \in \mathbb{N}$ *and* $1 \leqslant p < \infty$ *we define the* $W^{k,p}(\Omega)$ *norm as*

$$\|f\|_{W^{k,p}(\Omega)} := \|f\|_{k,p} = \left(\sum_{l=0}^{k} \sum_{|\alpha|=l} \|D^\alpha f\|_{L^p(\Omega)}^p\right)^{\frac{1}{p}}$$

For $p = \infty$ we define the $W^{k,\infty}$ norm as

$$\|f\|_{W^{k,\infty}(\Omega)} := \|f\|_{k,\infty} = \max_{l \leqslant k} \max_{|\alpha|=l} \|D^\alpha f\|_{L^\infty(\Omega)}.$$

Further, we define the corresponding semi norms as

$$|f|_{W^{k,p}(\Omega)} := |f|_{k,p} = \left(\sum_{|\alpha|=k} \|D^\alpha f\|^p_{L^p(\Omega)} \right)^{\frac{1}{p}}$$

and

$$|f|_{W^{k,\infty}(\Omega)} := |f|_{k,\infty} = \max_{|\alpha|=k} \|D^\alpha f\|_{L^\infty(\Omega)}.$$

For $1 \leqslant p \leqslant \infty$ we define the Sobolev space

$$W^{k,p}(\Omega) := \{\phi \in L^p(\Omega) : \ \|\phi\|_{W^{k,p}(\Omega)} < \infty\}.$$

An important result of linear functional analysis shows, that

$$H^k(\Omega) = W^{k,2}(\Omega),$$

where $H^k(\Omega)$ is the completion of $C^\infty(\Omega)$ with respect to the Sovolev norm $\|\cdot\|_{W^{k,2}(\Omega)}$. The space $H^1(\Omega)$ is therefore the space $W^{1,2}(\Omega)$, which is the space of L^2-functions with first weak derivative in L^2.

For a further characterization of the different Sovolev spaces we cite powerfull embedding theorems.

Lemma 2.14 (Embedding Theorems)**.** *Let $\Omega \subset \mathbb{R}^d$ and let $W^{k,p}(\Omega)$ be a Sobolev space with $k \geqslant 0$ and $1 \leqslant p < \infty$.*

1. *Let $W^{k',p'}(\Omega)$ be a second Sobolev space with $k' \geqslant 0$ and $1 \leqslant p' < \infty$. If it holds*

$$k - \frac{d}{p} \geqslant k' - \frac{d}{p'} \text{ and } k \geqslant k',$$

the embedding

$$W^{k,p}(\Omega) \hookrightarrow W^{k',p'}(\Omega)$$

is continuous. If it holds

$$k - \frac{d}{p} > k' - \frac{d}{p'} \text{ and } k > k',$$

the embedding

$$W^{k,p}(\Omega) \hookrightarrow W^{k',p'}(\Omega)$$

is compact.

2. *Let* $C^{k',\alpha}(\bar{\Omega})$ *for* $0 \leqslant \alpha \leqslant 1$ *be a Hölder space. If it holds*

$$k - \frac{d}{p} = k' + \alpha \text{ and } 0 < \alpha < 1,$$

then, the embedding

$$H^{k,p}(\Omega) \hookrightarrow C^{k,\alpha}(\bar{\Omega})$$

is continuous. If it holds

$$k - \frac{d}{p} > k' + \alpha,$$

this embedding

$$H^{k,p}(\Omega) \hookrightarrow C^{k,\alpha}(\bar{\Omega})$$

is compact.

The continuity of an embedding $X \hookrightarrow Y$ means, that for every $u \in X$ there exists a function in Y - we again call it $u \in Y$ - that coincides with $u \in X$ almost everywhere and such that

$$\|u\|_Y \leqslant C(\Omega, X, Y)\|u\|_X.$$

The constant can depend on k, p, d, k', p', α and the domain Ω. By the Sobolev embeddings we can conclude, that a function $u \in H^k(\Omega)$ is continuous, e.g. $u \in C(\Omega)$, if

$$k - \frac{d}{2} \geqslant 0 \quad \Leftrightarrow \quad k \geqslant \frac{d}{2}.$$

Therefore, H^1-functions are continuous in one dimensios, whereas H^2 is required in two or three dimensions. Compactness of an embedding $X \hookrightarrow Y$ means, that a bounded sequence $x_n \in X$ has a converging subsequence $x_{n'} \in Y$.

2.2 Elliptic Problems

In diesem Abschnitt vertiefen wir die theoretische Analyse von elliptischen Differentialgleichungen. Auf einem Lipschitzgebiet $\Omega \subset \mathbb{R}^d$ mit $d \geqslant 2$ betrachten wir allgemein Probleme vom Typ

$$Lu = f, \quad Lu := -\sum_{i,j=1}^{d} \partial_i(a_{ij}(x)\partial_j u(x)) + \sum_{j=1}^{d} b_j(x)\partial_j u(x) + c(x)u(x), \tag{2.5}$$

mit reellen Koeffizientenfunktionen a_{ij}, b_j und c. Es gilt wegen der Vertauschbarkeit der zweiten Ableitungen ohne Einschränkung $a_{ij} = a_{ji}$. Mit der Matrix $A = (a_{ij})_{i,j}$ sowie dem Vektor $b = (b_j)_j$ schreiben wir kurz:

$$Lu = -\nabla \cdot (A\nabla u) + b \cdot \nabla u + cu.$$

Bei elliptischen Problemen sind die Eigenwerte von A ungleich Null und haben das gleiche Vorzeichen. Wir nehmen ohne Einschränkung an, dass die Koeffizientenmatrix A positiv definit ist.

Elliptische partielle Differentialgleichungen sind Randwertprobleme. Wir unterscheiden drei verschiedene Arten von Randwerten:

1. Dirichlet-Problem Auf dem Rand $\partial\Omega$ geben wir den Funktionswert der Lösung vor:

$$\text{suche } u: \quad Lu = f \text{ in } \Omega, \quad u = g \text{ auf } \partial\Omega.$$

Es sei nun $\bar{g} \in H^1(\Omega)$ eine Fortsetzung von g, d.h., g ist die Spur. Dann können wir das Dirichlet-Problem stets in ein Problem mit homogenen Randdaten transformieren:

$$\text{suche } u = \bar{u} + \bar{g}: \quad L\bar{u} = f - L\bar{g}, \quad \bar{u} = 0 \text{ auf } \partial\Omega.$$

Die strenge Voraussetzung an die Regularität der Dirichlet-Daten g ist also, dass diese Funktion als Spur einer $H^1(\Omega)$-Funktion gegeben ist. Wir bezeichnen den Raum aller Spuren von $H^1(\Omega)$ Funktionen auf $\partial\Omega$ als $H^{\frac{1}{2}}(\partial\Omega)$. Es gilt $H^{\frac{1}{2}}(\partial\Omega) \subset L^2(\partial\Omega)$.

2. Neumann-Problem Auf dem Rand von Ω geben wir die Ableitung von u in *Normal-Richtung* vor:

$$\text{suche } u: \quad Lu = f \text{ in } \Omega, \quad \partial_n u = g \text{ auf } \partial\Omega.$$

Bei Neumann-Problemen ist der Funktionswert von u auf dem Rand nicht fixiert. Die Lösung muss also im vollen Raum $H^1(\Omega)$ gesucht werden. Betrachten wir z.B. das homogene Neumann-Problem

$$-\Delta u = f \text{ in } \Omega, \quad \partial_n u = 0 \text{ auf } \partial\Omega,$$

so hat dies zur Konsequenz, dass die Lösung nicht mehr eindeutig sein muss. Denn ist $u \in H^1(\Omega)$ eine Lösung, so ist auch durch $u + c$ für jedes $c \in \mathbb{R}$ eine Lösung gegeben. Um Eindeutigkeit zu erreichen muss der Lösungsraum künstlich eingeschränkt werden, etwa durch die Wahl:

$$H^1(\Omega)/\mathbb{R} := \{v \in H^1(\Omega), \int_\Omega v \, dx = 0\}.$$

We can easily show, that Poincaré's inequality also holds on this space.

Lemma 2.15 (Modified Poincaré's inequality)**.** *There exists a constant* $c_p = c(\Omega) > 0$ *such that*

$$\|v\|_\Omega \leqslant c_p \left(\left| \int_\Omega v \, dx \right| + \|\nabla v\|_\Omega \right) \quad \forall v \in H^1(\Omega).$$

Proof. Using Rellich's Theorem we can show (by contradiction), that the inequality holds for

$$\|v\|_\Omega \leqslant c\|\nabla v\|_\Omega \quad \forall v \in H^1(\Omega) \subset \mathbb{R}.$$

Then, the general case is obtained by

$$\bar{v} := \frac{1}{|\Omega|}\int_\Omega v\,dx \quad \Rightarrow \quad \|v\|_\Omega \leqslant \|\bar{v}\|_\Omega + \|v-\bar{v}\|_\Omega \leqslant \sqrt{|\Omega|}\left|\int_\Omega v\,dx\right| + c_p\|\nabla v\|_\Omega.$$

□

By this modification, the Neumann problem can be written as minimization problem with the functional

$$u \in H^1(\Omega)\setminus\mathbb{R}: \quad E(u) \leqslant E(v) := \frac{1}{2}\|\nabla v\|^2 - (f,v) \quad \forall v \in H^1(\Omega)\setminus\mathbb{R}.$$

3. Robin-Problem Auf dem Rand $\partial\Omega$ geben wir gemischte Randdaten vor:

$$\text{suche } u: \quad Lu = f \text{ in } \Omega, \quad \partial_n u + \alpha u = g \text{ auf } \partial\Omega,$$

mit einem $\alpha \neq 0$.

Wir werden zunächst ausschließlich das (homogene) Dirichlet-Problem betrachten. Hierzu leiten wir eine variationelle Formulierung der allgemeinen elliptischen Differentialgleichung (2.5) her:

Lemma 2.16 (Variationsproblem)**.** *Die Koeffizientenmatrix $A \in [L^\infty(\Omega)]^{d\times d}$ sei positiv definit mit $\langle Ax, x\rangle \geqslant \gamma|x|^2$ mit $\gamma > 0$, der Vektor $b \in [W^{1,\infty}(\Omega)]^d$ sei divergenzfrei (also $\nabla\cdot b = \sum_j \partial_i b_i = 0$) und es sei $c \in L^\infty(\Omega)$ mit $c \geqslant 0$. Jede Lösung $u \in C^2(\Omega)\cap C(\bar{\Omega})$ von (2.5) ist Lösung der Variationsgleichung:*

$$a(u,\phi) = (f,\phi) \quad \forall \phi \in H_0^1(\Omega).$$

Die Bilinearform

$$a(u,\phi) = (A\nabla u, \nabla\phi) + (b\cdot\nabla u, \phi) + (cu, \phi).$$

ist stetig

$$a(u,v) \leqslant M\|\nabla u\|\,\|\nabla v\| \quad \forall u,v \in H_0^1(\Omega)$$

mit einer Konstante

$$M = \max\{d\|A\|_\infty, c_p\sqrt{d}\|b\|_\infty, \|c\|_{L^\infty(\Omega)}c_p^2\},$$

sowie positiv definit (elliptisch)

$$a(u,u) \geqslant \gamma\|\nabla u\|^2 \quad \forall u \in H_0^1(\Omega).$$

Proof: (*i*) Wir multiplizieren (2.5) für festes $t \in \bar{I}$ mit einer beliebigen Funktion $\phi \in C_0^\infty(\Omega)$ und integrieren über das Gebiet:

$$\begin{aligned}(f,\phi) &= -(\nabla\cdot(A\nabla u),\phi) + (b\cdot\nabla u,\phi) + (cu,\phi)\\ &= (A\nabla u,\nabla\phi) + \underbrace{\int_{\partial\Omega} n\cdot(A\nabla u)\phi\,ds}_{=0} + (b\cdot\nabla u,\phi) + (cu,\phi).\end{aligned}$$

Wegen der Dichtheit von $C_0^\infty(\Omega)$ in $H_0^1(\Omega)$ kann der Testraum erweitert werden.

(*ii*) Weiter gilt für alle $u, v \in H_0^1(\Omega)$ wegen der Beschränktheit der Koeffizienten und mit Poincaré:

$$\begin{aligned}|a(u,v)| &\leqslant \sum_{i,j=1}^{d} |(a_{ij}\partial_j u, \partial_i v)| + \sum_{j=1}^{d} |(b_j\partial_j u, v)| + |(cu,v)|\\ &\leqslant \|A\|_\infty \sum_{ij} \|\partial_j u\|_{L^2(\Omega)}\,\|\partial_i v\|_{L^2(\Omega)}\\ &\qquad + \|b\|_\infty \sum_{j} \|\partial_j u\|_{L^2(\Omega)}\,\|v\|_{L^2(\Omega)} + |c|_\infty \|u\|_{L^2(\Omega)}\,\|v\|_{L^2(\Omega)}\\ &\leqslant \|A\|_\infty d \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} + \|b\|_\infty \sqrt{d}\|\nabla u\|_{L^2(\Omega)} c_p \|\nabla v\|_{L^2(\Omega)}\\ &\qquad + |c|_\infty c_p^2 \|\nabla u\|_{L^2(\Omega)}\,\|\nabla v\|_{L^2(\Omega)}\\ &\leqslant M\|\nabla u\|_{L^2(\Omega)}\,\|\nabla v\|_{L^2(\Omega)}, \quad M := \max\{\|A\|_\infty d, c_p\|b\|_\infty\sqrt{d}, |c|_\infty c_p^2\},\end{aligned}$$

wobei wir die folgende Ungleichung verwendet haben:

$$\sum_{i=1}^{d} |a_i| \leqslant \sqrt{d}\left(\sum_{i=1}^{d} |a_i|^2\right)^{\frac{1}{2}}.$$

(*iii*) Zum Nachweis der *Elliptizität* untersuchen wir die verschiedenen Terme der Bilinearform getrennt. Zunächst existiert wegen der positiven Definitheit von A ein γ mit:

$$(A\nabla u,\nabla u) = \int_\Omega \langle A(x)\nabla u(x), \nabla u(x)\rangle dx \geqslant \gamma \int_\Omega |\nabla u|^2\,dx = \gamma\|\nabla u\|^2_{L^2(\Omega)}. \tag{2.6}$$

Wegen der Divergenzfreiheit von b gilt:

$$\begin{aligned}(b\cdot\nabla u, u) &= \sum_{i=1}^{d}(b\partial_i u, u) = \sum_{i=1}^{d}(b_i u, \partial_i u) = -\sum_{i=1}^{d}(\partial_i(b_i u), u) + \sum_{i=1}^{d}\underbrace{\int_{\partial\Omega} n_i(b_i u)u\,ds}_{=0}\\ &= -\sum_{i=1}^{d}(\partial_i b_i u, u) - \sum_{j=1}^{d}(b_i\partial_i u, u) = \underbrace{-((\nabla\cdot b)u,u)}_{=0} - (b\cdot\nabla u, u),\end{aligned}$$

und also $(b\cdot\nabla u, u) = 0$. Schließlich gilt mit $c \geqslant 0$

$$(cu,u) \geqslant \min_\Omega c\cdot\|u\| \geqslant 0$$

und zusammen mit (2.6) ergibt sich die Behauptung. □

Die stetige, positiv definite Bilinearform unterscheidet sich von einem Skalarprodukt durch die fehlende Symmetrie.

Definition 2.17 (Schwache Lösung)**.** *Eine Funktion* $u \in H_0^1(\Omega)$ *heißt verallgemeinerte* (*schwache*) *Lösung von* (2.5)*, falls*

$$a(u, \phi) = (f, \phi) \quad \forall \phi \in H_0^1(\Omega). \tag{2.7}$$

Wie bereits argumentiert gilt:

Lemma 2.18. *Jede schwache Lösung* $u \in H_0^1(\Omega)$ *von* (2.7) *für welche* $u \in C^2(\Omega) \cap C(\bar{\Omega})$ *gilt ist auch klassische Lösung von* (2.5)*.*

Und weiter gilt:

Lemma 2.19 (Minimierung des Energiefunktionals)**.** *Im Fall* $b = 0$ *ist die Lösung der variationellen Formulierung* (2.7) *äquivalent zur Minimierung des Energiefunktionals. Suche* $u \in H_0^1(\Omega)$*:*

$$E(u) \leqslant E(v) \quad \forall v \in V, \quad E(v) := \frac{1}{2}a(v, v) - (f, v). \tag{2.8}$$

Proof: (*0*) Die stetige und elliptische Bilinearform $a(\cdot, \cdot)$ ist symmetrisch und somit ein Skalarprodukt. Dies folgt aus $b = 0$ sowie der Symmetrie von A.

(*i*) Wir gehen zunächst davon aus, dass u eine variationelle Lösung von (2.7) ist. Dann folgt für alle $v \in H_0^1(\Omega)$:

$$\begin{aligned} E(u) - E(v) &= \frac{1}{2}a(u, u) - (f, u) - \frac{1}{2}a(v, v) + (f, v) \\ &= \frac{1}{2}a(u, u) - a(u, u) - \frac{1}{2}a(v, v) + a(u, v) \\ &= -\frac{1}{2}\Big\{a(u, u) - 2a(u, v) + a(v, v)\Big\} \\ &= -\frac{1}{2}a(u - v, u - v) \leqslant -\frac{\gamma}{2}\|\nabla(u - v)\| \leqslant 0. \end{aligned}$$

Also, $E(u) \leqslant E(v)$.

(*ii*) Nun sei u Minimum des Energiefunktionals. Dann muss gelten:

$$\frac{d}{ds}E(u + sv)\Big|_{s=0} = 0 \quad \forall v \in H_0^1(\Omega).$$

Also, wegen der Symmetrie von $a(\cdot, \cdot)$:

$$\frac{d}{ds}\Big\{\frac{1}{2}a(u + sv, u + sv) - (f, u + sv)\Big\}\Big|_{s=0} = a(u, v) - (f, v) = 0 \quad \forall v \in H_0^1(\Omega).$$

□

Existence of solutions

For considering the general (non-symmetric) case, we need some additional background from linear functional analysis.

Definition 2.20 (Banach space, Hilbert space)**.** *Let V be a linear space. Let $\|\cdot\| : V \to \mathbb{R}$ be a norm on V. If V is complete with $\|\cdot\|$ it is called a* Banach space*. By $(\cdot,\cdot) : V \times V \to \mathbb{R}$ we introduce a* scalar product *on V. If V is complete with respect to the induced norm $\|\cdot\|_V = (\cdot,\cdot)_{V\times V}^{1/2}$ it is called* Hilbert space*.*

Examples for Hilbert spaces are the space $L^2(\Omega)$ with the scalar product and norm

$$(f,g)_{L^2(\Omega)} := \int_\Omega fg\,dx, \quad \|f\|_{L^2(\Omega)} := \int_\Omega |f|^2\,dx,$$

or the space $H_0^1(\Omega)$ with scalar product and norm

$$(f,g)_{H_0^1(\Omega)} := (\nabla f, \nabla g)_{L^2(\Omega)} = \int_\Omega \nabla f \cdot \nabla g\,dx, \quad \|f\|_{H_0^1(\Omega)} := \|\nabla f\|_{L^2(\Omega)} := \int_\Omega |\nabla f|^2\,dx.$$

Further, $H^1(\Omega)$ (without zero trace) is a Hilbert space with

$$(f,g)_{H^1(\Omega)} := (f,g)_{L^2(\Omega)} + (\nabla f, \nabla g)_{L^2(\Omega)}, \quad \|f\|_{H^1(\Omega)} := \left(\|f\|_{L^2(\Omega)}^2 + \|\nabla f\|_{L^2(\Omega)}^2\right)^{\frac{1}{2}}.$$

Finally, $H^k(\Omega) = W^{k,2}(\Omega)$ is a Hilbert space for every k together with the scalar product

$$(f,g)_{H^k(\Omega)} = \sum_{l=0}^{k} (\nabla^l f, \nabla^l g)_{L^2(\Omega)}.$$

Definition 2.21 (Dual space)**.** *Let V be a vector space. By V^* we denote the* dual space *of V. The space V^* consists of all* linear functionals

$$l : V \to \mathbb{R}$$

and V^* carries the induces linear structure:

$$l(v+w) = l(v) + l(w) \quad \forall v, w \in V,$$

and

$$l(\alpha v) = \alpha l(v) \quad \forall \alpha \in \mathbb{R},\ \forall v \in V.$$

If V is a normed space with norm $\|\cdot\|_V$, the dual norm *$\|\cdot\|_{V^*}$ on V^* is induced by*

$$\|l\|_{V^*} := \sup_{v\in V,\ \|v\|_V \neq 0} \frac{|l(v)|}{\|v\|_V} = \sup_{v\in V,\ \|v\|_V \leqslant 1} |l(v)|.$$

Example 2.22. *Let* $V = L^2(\Omega)$. *A linear functional* $l \in V^*$ *is given by*

$$l(v) := \int_\Omega v \, dx.$$

Linearity of $l(\cdot)$ *comes from the linearity of the integral. It holds*

$$\|l\|_{L^2(\Omega)^*} = \sup_{\|v\| \leqslant 1} \int_\Omega v \, dx \leqslant \underbrace{\|v\|}_{\leqslant 1} \|1\|_\Omega \leqslant \sqrt{|\Omega|}.$$

On the other hand,

$$\|l\|_{L^2(\Omega)^*} = \sup_{\|v\| \leqslant 1} \int_\Omega v \, dx \geqslant \int_\Omega \frac{1}{\sqrt{|\Omega|}} \, dx = \sqrt{|\Omega|}.$$

Therefore, $\|l\|_{L^2(\Omega)^2} = \sqrt{|\Omega|}$.

Next, let $V = H_0^1(\Omega)$ *and* $\omega \in H_0^1(\Omega)$ *be arbitrary. Then,*

$$l_\omega(v) := \int_\Omega \nabla v \cdot \nabla \omega \, dx$$

is a linear functional. Once more, linearity follows by using the linearity of the integral and the scalar product. Here, it holds

$$\|l\|_{H_0^1(\Omega)^*} = \sup_{\|\nabla v\| \leqslant 1} (\nabla v, \nabla \omega) \leqslant \|\nabla \omega\|.$$

Then,

$$\|l\|_{H_0^1(\Omega)^*} = \sup_{\|\nabla v\| \leqslant 1} (\nabla v, \nabla \omega) \geqslant \left(\frac{\nabla \omega}{\|\nabla \omega\|}, \nabla \omega \right) = \|\nabla \omega\|.$$

Hence, $\|l\|_{L^2(\Omega)^2} = \|\nabla \omega\|$.

By Hölder's inequality

$$\|fg\|_{L^1(\Omega)} \leqslant \|f\|_p \|g\|_q, \quad p, q \in [1, \infty], \quad \frac{1}{p} + \frac{1}{q} = 1,$$

we can identify the dual space of $L^p(\Omega)$ for $1 < p < \infty$ with $L^p(\Omega)^* \cong L^q(\Omega)$ where $q = (1 + 1/p)^{-1}$ is the dual exponent. For every $g \in L^q(\Omega)$, the mapping

$$f \mapsto \int_\Omega fg \, dx$$

is s linear functional.

One of the very import theorems in linear functional analysis is

Lemma 2.23 (Riesz representation theorem). *Let V be a Hilbert space with scalar product $(\cdot,\cdot)_{V\times V}$. For every linear functional $l \in V^*$ there exists a unique element $w \in V$, such that*

$$l(v) = (v,w)_{V\times V} \quad \forall v \in V, \quad \|l\|_{V^*} = \|w\|_V.$$

Vice versa, every $w \in V$ defines a linear functional $l \in V^*$ by

$$l(v) := (v,w)_{V\times V}.$$

The proof is found in the literature [1].

A consequence of Riesz representation theorem is the unique existence of the Laplace solution

Lemma 2.24 (Existence of Laplace). *Let $f \in L^2(\Omega)$. The variational solution $u \in H_0^1(\Omega)$ of the Laplace equation*

$$(\nabla v, \nabla\phi) = (f,\phi) \quad \forall\phi \in V$$

is unique.

Proof. By

$$l(\phi) := (f,\phi) \quad \forall\phi \in H_0^1(\Omega)$$

a linear functional in $H_0^1(\Omega)^*$ is given, as it is bounded

$$l(\phi) \leqslant (f,\phi) \leqslant \|f\|\,\|\phi\| \leqslant c_p\|f\|\,\|\nabla\phi\|,$$

and obviously linear. The variational formulation is the H_0^1-scalar product. The problem is therefore equivalent to the formulation

$$(u,\phi)_{H_0^1(\Omega)} = l(\phi).$$

According to Riesz, such a $u \in H_0^1(\Omega)$ exists uniquely. Furthermore it holds

$$\|u\|_{H_0^1(\Omega)} = \|\nabla u\| = \|l\|_{H^1(\Omega)^*} \leqslant c_p\|f\|.$$

□

Auf das allgemeine elliptische Problem kann der Riesz'sche Darstellungssatz nicht angewendet werden. Das Problem ist nicht durch ein Skalarprodukt gegeben. Wir benötigen den allgemeineren:

Lemma 2.25 (Lax-Milgram). *Sei V ein Hilbertraum, $l \in V^*$ ein stetiges lineares Funktional und $a : V \times V \to \mathbb{R}$ eine stetige und elliptische Bilinearform:*

$$a(u,v) \leqslant M\|u\|_V\,\|v\|_V \quad \forall u,v \in V, \quad a(u,u) \geqslant \gamma\|u\|_V^2 \quad \forall u \in V.$$

Dann existiert eine eindeutige Lösung $u \in V$ *der Variationsgleichung:*

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

welche stetig von den Daten abhängt:

$$\|u\|_V \leqslant \frac{1}{\gamma}\|l\|_{V^*}.$$

Proof: (i) Für jedes feste $u \in V$ ist wegen der Stetigkeit der Bilinearform ein stetiges lineares Funktional $A_u \in V^*$ definiert:

$$A_u(v) := a(u, v) \leqslant M_u\|v\|_V \quad \forall v \in V.$$

Da V ein Hilbertraum ist gilt der Darstellungssatz von Riesz und es existiert ein Element $Au \in V$, so dass gilt:

$$(Au, v)_V = A_u(v) = a(u, v) \quad \forall v \in V.$$

Es ist also durch den Riesz'schen Darstellungssatz ein Operator $A : V \to V$ definiert, welcher jedem Element $u \in V$ ein Element $Au \in V$ zuordnet, so dass sich die Bilinearform durch das Skalarprodukt ausdrücken lässt. Es gilt:

$$\|Au\|_V = \|A_u\|_{V^*} = \sup_{v\in V} \frac{a(u, v)}{\|v\|_V} \leqslant \sup_{v\in V} \frac{M\|u\|_V\,\|v\|_V}{\|v\|_V} \leqslant M\|u\|_V. \tag{2.9}$$

(ii) Für ein festes $r \in \mathbb{R}_+$ definieren wir einen weiteren Operator $T_r : V \to V$ durch

$$(T_r u, v)_V := (u, v)_V + r\big(l(v) - (Au, v)_V\big) \quad \forall v \in V.$$

Es gilt für zwei Elemente $u, w \in V$:

$$(T_r u - T_r w, v)_V = (u - w - rA(u - w), v)_V.$$

Wir wählen die Testfunktion $v := T_r u - T_r w$

$$\begin{aligned} \|T_r u - T_r w\|_V^2 &= (u - w - rA(u - w), u - w - rA(u - w))_V \\ &= \|u - w\|_V^2 - 2r(A(u - w), u - w)_V + r^2(A(u - w), A(u - w))_V. \end{aligned} \tag{2.10}$$

Es gilt wegen der Elliptizität

$$(A(u - w), u - w)_V = a(u - w, u - w) \geqslant \gamma\|u - w\|_V^2,$$

und mit (2.9)

$$(A(u - w), A(u - w))_V = \|A(u - w)\|_V^2 \leqslant M^2\|u - w\|_V^2.$$

Zusammen folgt aus (2.10)

$$\|T_r u - T_r w\|_V^2 \leqslant |1 - 2r\gamma + M^2r^2|\|u - w\|_V^2.$$

Und für

$$|1 - 2r\gamma + M^2 r^2| < 1 \quad \Leftrightarrow \quad r \in \left(0, \frac{2\gamma}{M^2}\right),$$

ist T_r eine Kontraktion und mit dem Banach'schen Fixpunktsatz existiert ein eindeutig bestimmtes $u \in V$ mit $T_r u = u$. Und dann:

$$T_r u = u \quad \Rightarrow \quad (Au, v) = l(v) \quad \forall v \in V.$$

(*iii*) Es existiert also eine Lösung. Angenommen es würden zwei Lösungen $u_1, u_2 \in V$ mit $w := u_1 - u_2$ existieren. Dann gilt:

$$a(w, v) = 0 \quad \forall v \in V.$$

Und wegen der Elliptizität:

$$0 = a(w, w) \geqslant \gamma \|w\|_V^2 \quad \Rightarrow \quad w = 0.$$

(*iv*) Abschließend gilt für die Lösung die *a priori* Schranke:

$$\gamma \|u\|_V^2 \leqslant a(u, u) = l(u) \leqslant \|l\|_{V^*} \|u\|_V.$$

□

Dieser Satz kann nun unmittelbar auf das allgemeine elliptische Problem angewendet werden:

Korollar 2.26. *Sei* L *ein elliptischer Operator mit den Eigenschaften aus Satz 2.16. Weiter sei* $f \in L^2(\Omega)$. *Dann hat die Gleichung:*

$$Lu = f \text{ in } \Omega, \quad u = 0 \text{ auf } \partial\Omega,$$

eine eindeutig bestimmte schwache Lösung $u \in H_0^1(\Omega)$ *und es gilt:*

$$\|\nabla u\|_{L^2(\Omega)} \leqslant \frac{c_p}{\gamma} \|f\|_{L^2(\Omega)},$$

mit der Poincaré-Konstante c_p.

Proof: Durch $l(v) = (f, v)$ ist ein lineares, stetiges Funktional bestimmt:

$$l(v) = (f, v) \leqslant \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \leqslant c_p \|f\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} \quad \forall v \in H_0^1(\Omega).$$

Mit dem Satz von Lax-Milgram 2.25 folgt die Aussage. □

This result is not optimal in terms of the right hand side. Every $f \in L^2(\Omega)$ defines a functional $l \in H_0^1(\Omega)^*$ by

$$l(\phi) = (f, \phi)_\Omega \quad \Rightarrow \quad |l(\phi)| \leqslant c_p \|f\| \, \|\nabla \phi\|.$$

It is however not necessary to choose $f \in L^2(\Omega)$. Lax-Milgram requires $f \in H_0^1(\Omega)^*$ to guarantee a solution $u \in H_0^1(\Omega)$. We can identify $L^2(\Omega)$ as a subspace of the dual space $H_0^1(\Omega)^*$. We define

Definition 2.27 (Dual space of $H_0^1(\Omega)$)**.** *By*

$$\|l\|_{H^{-1}(\Omega)} := \sup_{\phi \in H_0^1(\Omega)} \frac{l(\phi)}{\|\nabla \phi\|}$$

we define the H^{-1}*-dual norm. By*

$$H^{-1}(\Omega) := \{l : H_0^1(\Omega) \to \mathbb{R}, \quad \|l\|_{H^{-1}(\Omega)} < \infty\}$$

we denote the dual space of $H_0^1(\Omega)$*.*

The space $H^{-1}(\Omega)$ is the dual space of $H_0^1(\Omega)$ with respect to the scalar product $(\nabla \cdot, \nabla \cdot)_\Omega$ that only by Poincaré's inequality is a homogenous.

Further,

Definition 2.28 (Duality product)**.** *Let* V *be a vector space with dual space* V^**. We define the duality product*

$$\langle l, v \rangle_{V^* \times V} := l(v) \quad \forall v \in V, \ \forall l \in V^*.$$

Using this more general concept, the Laplace problem has a weak solution $u \in H_0^1(\Omega)$ for all $f \in H^{-1}(\Omega)$ and it holds

$$(\nabla u, \nabla \phi) = \langle f, \phi \rangle \quad \forall \phi \in H_0^1(\Omega) \quad \Rightarrow \quad \|\nabla u\| = \|f\|_{H^{-1}(\Omega)}.$$

One examples for a H^{-1}-functional $l \in H^{-1}(\Omega)$, that cannot be represented as an L^2-function via

$$l(\phi) = (f, \phi)_\Omega \quad \forall \phi \in H^1(\Omega),$$

is the evaluation of a line integral, i.e.

$$l(\phi) = \int_{\Gamma_{in}} \phi(x)\, dx,$$

where $\Gamma_{in} \subset \Omega$ is a line segment within the domain. By the trace inequality we know, that such a trace exists, i.e. we know the estimate

$$|l(\phi)| = \Big| \int_{\Gamma_{in}} \phi(x)\, dx \Big| \leqslant \|\phi\|_{\Gamma_{in}} \sqrt{|\Gamma_{in}|} \leqslant \sqrt{|\Gamma_{in}|} c_{\Gamma_{in}} \|\phi\|_{H^1(\Omega)}.$$

Hence, $l \in H^{-1}(\Omega)$. There exists however no such $f \in L^2(\Omega)$ that represents this functional. Considering the problem

$$(\nabla u, \nabla \phi) = \int_{\Gamma_{in}} \phi(x)\, dx \quad \forall \phi \in H_0^1(\Omega),$$

we must use this generalized concept of the right hand side.

Regularität der Lösung Von entscheidender Bedeutung für das weitere Vorgehen ist die Regularität der Lösung $u \in H_0^1(\Omega)$. Falls wir z.B. $u \in H^2(\Omega)$ zeigen können, so folgt aus dem Einbettungssatz 2.14 $H^2(\Omega) \hookrightarrow C(\Omega)$ die Stetigkeit der Lösung. Um zu garantieren, dass u eine klassische Lösung ist, also $u \in C^2(\Omega)$ benötigen wir höhere Regularität. Der Einbettungssatz 2.14 fordert in d räumlichen Dimensionen für die Einbettung $H^m(\Omega) \hookrightarrow C^2(\Omega)$:

$$m - \frac{d}{2} > 2 \quad \Leftrightarrow \quad m > 2 + \frac{d}{2},$$

also benötigen wir $m = 4$ und $u \in H^4(\Omega)$.

Zunächst gilt einfach:

Lemma 2.29 (Schwacher Laplace)**.** *Für die verallgemeinerte Lösung $u \in H_0^1(\Omega)$ der Laplace-Gleichung ist Δu im schwachen Sinne definiert und liegt in $L^2(\Omega)$.*

Proof: Sei $u \in H_0^1(\Omega)$ die schwache Lösung von $(\nabla u, \nabla v) = (f, v)$ für alle v. Dann gilt:

$$(f, v) = (\nabla u, \nabla v) = -(u, \Delta v) \quad \forall v \in C_0^\infty(\Omega).$$

Also ist $f = -\Delta u$ im Sinne der schwachen Ableitung mit

$$f = -\Delta u \in L^2(\Omega).$$

□

Es zeigt sich nun, dass sich die Regularität der Daten auf die Regularität der Lösung überträgt. Im Gegensatz zu gewöhnlichen Differentialgleichungen spielt jedoch die Regularität des Gebiets eine entscheidende Bedeutung. Ohne Beweis:

Lemma 2.30 (Regularität der Lösung)**.** *Sei Ω polygonal und konvex, oder besitze einen Rand der Klasse C^2 (lokal parametrisierbar durch eine zweimal stetig differenzierbare Funktion). Im Falle $f \in L^2(\Omega)$ gilt die* a priori *Abschätzung:*

$$\|u\|_{H^2(\Omega)} \leqslant c_s \|f\|_{L^2(\Omega)},$$

mit einer von f unabhängigen Stabilitätskonstante c_s.

Dieser Satz sichert uns die Stetigkeit der Lösung $u \in H_0^1(\Omega) \cap C(\bar{\Omega})$. Die Konvexität des Gebiets ist entscheidend. Ausgeschlossen sind *einspringende Ecken*, also Kanten mit Innenwinkel ω größer $180°$. Wir betrachten auf dem "Tortenstück" aus Abbildung 2.1 die Poisson-Gleichung:

$$-\Delta u = 0, \quad u(r, \theta) = 0 \quad \theta \in [0, \frac{\pi}{\omega}], \quad u(1, \theta) = \sin(\frac{\pi}{\omega}).$$

In Polarkoordinaten ist die Lösung gegeben durch:

$$u(r, \theta) = r^{\frac{\pi}{\omega}} \sin\left(\frac{\pi}{\omega}\theta\right).$$

Abbildung 2.1: Gebiet mit *einspringender Ecke*.

Für $\omega \in (\pi, 2\pi)$ gilt $u \in H^1(\Omega)$, aber $u \notin H^2(\Omega)$. Der Fall $\omega = 2\pi$ wird *Schlitz-Gebiet* genannt.

Allgemein gilt:

Lemma 2.31 (Höhere Regularität der Lösung)**.** *Für ein* $m \geqslant 0$ *sei* $f \in H^{m-2}(\Omega)$*. Weiter sei* Ω *ein Gebiet der Klasse* C^{m+2}*. Dann gilt für die Lösung des Poisson-Problems*

$$\|u\|_{H^m(\Omega)} \leqslant c\|f\|_{H^{m-2}(\Omega)},$$

mit einer von f *unabhängigen Konstante* c*.*

Abschließend beweisen wir als wichtige und grundlegende Aussage für elliptische Gleichungen

Lemma 2.32 (Elliptisches Maximumprinzip)**.** *Für den elliptischen Operator*

$$Lu := -\Delta u + au,$$

mit $a \geqslant 0$ *gilt auf dem Gebiet* $\Omega \subset \mathbb{R}^d$ *das* Maximumprinzip*. D.h., eine Funktion* $u \in C^2(\Omega) \cap C(\bar{\Omega})$ *mit der Eigenschaft* $Lu \leqslant 0$ *hat kein positives Maximum im Innern. Es gilt also*

$$u(x) \leqslant 0 \text{ } \textit{für alle } x \in \Omega \textit{ oder } \max_{x\in\Omega} u(x) = \max_{x\in\partial\Omega} u(x)$$

Proof. Wir beweisen den einfachen Fall $a > 0$. Sei also $Lu \leqslant 0$ und $z \in \Omega$ ein inneres Maximum mit $u(z) > 0$. Dann gilt notwendig

$$\nabla u(z) = 0, \quad \partial_{xx}u(z) \leqslant 0, \quad \partial_{zz}u(z) \leqslant 0.$$

Also folgt:

$$-\Delta u(z) \geqslant 0$$

aus $au(z) > 0$ folgt der Widerspruch. □

In einer Dimension besagt das Maximumprinzip gerade die Konvexität einer Funktion mit $\partial_{xx}u(x) > 0$.

2.3 Parabolische Probleme

Es sei nun L ein elliptischer Differentialoperator auf dem Gebiet $\Omega \subset \mathbb{R}^d$. Dann betrachten wir auf dem Zeit-Orts-Gebiet $I \times \Omega$ mit $I = [0, T]$ die parabolische Anfangs-Randwertaufgabe

$$\partial_t u(x,t) + Lu(x,t) = f(x,t) \quad \forall (x,t) \in \Omega \times I,$$

mit der Anfangsbedingung

$$u(x,0) = u^0(x),$$

und den Randbedingungen von Dirichlet, Neumann oder Robin-Typ.

Als Prototyp einer parabolischen Differentialgleichung betrachten wir die Wärmeleitungsgleichung auf einem Gebiet $\Omega \subset \mathbb{R}^d$ und einem Zeitintervall $I = [0, T]$ mit Dirichlet-Randwerten:

$$\partial_t u(x,t) - \Delta u(x,t) = f(x,t) \quad \text{in } I \times \Omega, \quad u(x,0) = u^0, \quad u(x,t) = 0 \text{ auf } I \times \partial\Omega. \quad (2.11)$$

Zur Beschreibung einer Lösung dieser Gleichung werden wir wieder eine schwache Formulierung herleiten. Hierzu ist es wesentlich, die notwendige Regularität der auftretenden Funktionen zu untersuchen. Neben der örtlichen Komponente hat die Lösung $u : I \times \Omega \to \mathbb{R}$ nun auch eine zeitliche. Im Ort werden wir wie bei den elliptischen Differentialgleichungen den schwachen Regularitätsbegriff der Sobolew-Räume verwenden. Wir definieren:

Definition 2.33. *Sei $V := W^{n,p}(\Omega)$ ein Sobolew-Raum im Ort. Sei $I := (0, T)$ ein Intervall. Der Raum $C(\bar{I}; V)$ besteht aus allen stetigen Funktionen $u : \bar{I} \to V$ mit Werten in V mit der Norm*

$$\|u\|_{C(\bar{I};V)} := \max_{t \in \bar{I}} \|u(t)\|_V.$$

Wir interpretieren also jede solche Funktion $u : \bar{I} \times \Omega \to \mathbb{R}$ auch als Funktion $u : \bar{I} \to V$ mittels

$$u(t,x) =: [u(t)](x),$$

mit $u(t) \in V$.

Diese Funktionen sind stetig in der Zeit und haben einen Wertevorrat im Sobolew-Raum V. Wir müssen diesen Regularitätsbegriff in der Zeit weiter abschwächen und definieren weiter:

Definition 2.34 (Zeitliche Sobolew-Räume)**.** *Sei $V := W^{n,p}(\Omega)$ ein Sobolew-Raum im Ort. Der Raum $W^{m,q}(I; V)$ besteht aus allen Funktionen $u : I \times \Omega \to \mathbb{R}$ deren m-te Zeitableitung $\partial_t^m u$ im schwachen Sinne existiert und welche beschränkt sind bezüglich der Norm*

$$\|u\|_{W^{m,q}(I;V)} = \left(\sum_{j=0}^{m} \|\partial_t^j u(t)\|_V^q \right)^{\frac{1}{q}},$$

beziehungsweise im Fall $q = \infty$

$$\|u\|_{W^{m,\infty}(I;V)} = \operatorname{ess\,sup} \left(\sum_{j=0}^{m} \|\partial_t^j u(t)\|_V \right)$$

Für diese Räume gelten wieder die Einbettungssätze. Man beachte, dass die Dimension des "zeitlichen Gebiets" stets 1 ist:

Lemma 2.35. *Sei* $u \in W^{1,p}(I;V)$. *Dann gilt:*

$$\begin{aligned} &(i) \quad u \in C(\bar{I}, V), \\ &(ii) \quad u(t) = u(0) + \int_0^t \partial_t u(s)\, ds \quad s \in I, \\ &(iii) \quad \|u\|_{C(\bar{I};V)} \leqslant c\|u\|_{W^{1,p}(I,V)}. \end{aligned}$$

Im Folgenden leiten wir eine schwache Formulierung der parabolischen Gleichung (2.11) her. Für die rechte Seite f und den Startwert fordern wir die Regularität

$$f \in L^2(I; L^2(\Omega)), \quad u^0 \in L^2(\Omega),$$

d.h., die rechte Seite f ist bezüglich Ort und Zeit in L^2. Wir multiplizieren (2.11) für festes $t \in \bar{I}$ mit einer Funktion $\phi \in C_0^\infty(\Omega)$ und integrieren über das räumliche Gebiet Ω. Mit partieller Integration erhalten wir

$$\int_\Omega \partial_t u(x,t)\phi(x)\, dx + \int_\Omega \nabla u(x,t) \cdot \nabla\phi(x) dx = \int_\Omega f(x,t)\phi(x)\, dx, \qquad \text{für fast alle } t \in I.$$

Wir dürfen diese Gleichheit nur für *fast alle* $t \in I$ ansetzen, da die rechte Seite $f \in L^2(I; L^2(\Omega))$ aufgefasst als Funktion $f : I \to L^2(\Omega)$ ja auch nur für fast alle $t \in \bar{I}$ definiert ist.

Die gesuchte Lösung $u : I \to V$ kann zeitlich gesehen also nur in $L^2(I;V)$ liegen. Die optimale örtliche Regularität ergibt sich wie bei den elliptischen Gleichungen für $V = H_0^1(\Omega)$, so dass die Ableitungen schwach definiert sind und alle Integrale existieren. Mit $u : I \to V$ und $f : I \to L^2(\Omega)$ schreiben wir kürzer:

$$(\partial_t u(t), \phi)_\Omega + (\nabla u(t), \nabla\phi)_\Omega = (f(t), \phi)_\Omega \quad \forall \phi \in H_0^1(\Omega) \quad \text{und für fast alle } t \in \bar{I}. \tag{2.12}$$

Für die Lösung $u : I \to V$ gilt also:

$$u \in L^2(I; H_0^1(\Omega)).$$

Es bleibt, die Regularität der Zeitableitung $\partial_t u$ zu diskutieren, da u zeitlich zunächst nur in L^2 liegt (und von Ableitungen im strengen Sinne nicht die Rede sein kann). Es gilt für jedes $\phi \in H_0^1(\Omega)$:

$$\begin{aligned} |(\partial_t u(t), \phi)| &= \big|(f(t), \phi) - (\nabla u(t), \nabla\phi)\big| \\ &\leqslant \|f(t)\|\,\|\phi\| + \|\nabla u(t)\|\,\|\nabla\phi\| \\ &\leqslant (c_P\|f(t)\| + \|\nabla u(t)\|)\|\nabla\phi\|, \end{aligned}$$

mit der Poincaré-Konstante c_P. D.h., die Zeitableitung $\partial_t \tilde{u}(t)$ definiert für fast jedes $t \in I$ ein stetiges lineares Funktional auf dem $H_0^1(\Omega)$, oder anders ausgedrückt: $\partial_t \tilde{u}(t) \in H^{-1}(\Omega)$ für fast jedes $t \in I$, denn:

$$\|\partial_t u(t)\|_{H^{-1}} = \sup_{\phi \in H_0^1(\Omega)} \frac{(\partial_t u(t), \phi)}{\|\nabla \phi\|} \leqslant c_P \|f(t)\| + \|\nabla u(t)\|.$$

Wir definieren

Definition 2.36 (Schwache Formulierung der Wärmeleitungsgleichung)**.** *Sei $\Omega \subset \mathbb{R}^d$ ein Gebiet und $I = (0, T)$ das zeitliche Intervall. Wir suchen die Lösung*

$$u \in L^2(I; H_0^1(\Omega)), \quad \partial_t u \in L^2(I; H^{-1}(\Omega)), \tag{2.13}$$

so dass

$$(\partial_t u(t), \phi) + (\nabla u(t), \nabla \phi) = (f(t), \phi) \quad \forall \phi \in H_0^1(\Omega), \tag{2.14}$$

für fast alle $t \in I$ und

$$u(0) = u^0, \quad u\big|_{I \times \partial\Omega} = 0.$$

Es gilt der folgende Regularitätssatz für den Lösungsraum

Lemma 2.37. *Angenommen $u \in L^2(I; H_0^1(\Omega))$ und $\partial_t u \in L^2(I; H^{-1}(\Omega))$. Dann gilt*

$$u \in C(\bar{I}; L^2(\Omega)),$$

und

$$\|u\|_{C(\bar{I}; L^2(\Omega))} \leqslant c \left(\|u\|_{L^2(I; H_0^1(\Omega))} + \|\partial_t u\|_{L^2(I; H^{-1}(\Omega))} \right),$$

mit einer Konstante $c > 0$ unabhängig von u. Weiter ist die Funktion $t \mapsto \|u(t)\|^2_{L^2(\Omega)}$ absolut stetig und es gilt

$$\frac{d}{dt} \|u(t)\|^2 = 2(\partial_t u(t), u(t))_\Omega \quad \textit{für fast alle } t \in I.$$

Dieser Satz besagt also, dass jede schwache Lösung von (2.14) mit der Regularität (2.13) eine bezüglich der Zeit stetige Funktion mit Werten in $L^2(\Omega)$ ist. Dies erlaubt insbesondere die Vorgabe eines Anfangswertes $u^0 \in L^2(\Omega)$.

Existenz einer Lösung Der Nachweis der Existenz einer Lösung der Wärmeleitungsgleichung ist aufwändiger als bei den elliptischen Gleichungen. Wir skizzieren hier nur das Vorgehen.

Zunächst sei durch $\{\omega_k\}_{k \geqslant 0}$ ein Orthogonalsystem von Eigenvektoren des Laplace-Operators $L := -\Delta$ gegeben. Die ω_k seien normiert bzgl $L^2(\Omega)$. Für ein festes m definieren wir den endlich dimensionalen Raum

$$V_m := \operatorname{span}\{\omega_1, \ldots, \omega_m\},$$

und suchen die Galerkin-Approximation $u_m \in L^2(I; V_m)$ mit der Darstellung

$$u_m(x,t) = \sum_{k=1}^{m} d_k(t)\omega_k(x), \tag{2.15}$$

als Lösung von

$$(\partial_t u_m, \omega_l) + (\nabla u_m, \nabla \omega_l) = (f, \omega_l), \quad l = 1, \dots, m \text{ und für fast alle } t \in I. \tag{2.16}$$

Mit dem Ansatz 2.15 gilt:

$$\sum_{k=1}^{m} \Big\{ \partial_t d_k(t)(\omega_k, \omega_l) + d_k(t)(\nabla\omega_k, \nabla\omega_l) \Big\} = (f, \omega_k), \quad k = 1, \dots, m.$$

Die ω_k sind eine L^2-Orthonormalbasis, das System ist also äquivalent zu einem System aus linearen gewöhnlichen Differentialgleichungen für $d(t) : I \to \mathbb{R}^m$:

$$\partial_t d(t) + Ad(t) = \bar{f}(t), \tag{2.17}$$

mit der Matrix $A = (a_{kl})_{k,l=1}^m$ und $a_{kl} := (\nabla\omega_k, \nabla\omega_l)$ und $\bar{f} : I \to \mathbb{R}^m$ mit $\bar{f} = (f(t), \omega_k)_{k=1}^m$.

Nun gilt:

Lemma 2.38. *Für jedes $m = 1, 2, \dots$ besitzt das System (2.17) eine eindeutige Lösung $u_m \in L^2(I; V_m)$ mit*

$$\max_{t\in\bar{I}} \|u_m(t)\| + \|u_m\|_{L^2(I;H_0^1(\Omega))} + \|\partial_t u_m\|_{L^2(I;H^{-1}(\Omega))} \leqslant \|f\|_{L^2(I;L^2(\Omega))} + \|u^0\|_{L^2(\Omega)}.$$

Jetzt beweisen wir

Lemma 2.39 (Existenz einer Lösung der Wärmeleitungsgleichung)**.** *Die Wärmeleitungsgleichung hat für jedes $f \in L^2(I; L^2(\Omega))$ und $u^0 \in L^2(\Omega)$ eine schwache Lösung mit*

$$\max_{t\in\bar{I}} \|u(t)\| + \|u\|_{L^2(I;H_0^1(\Omega))} + \|\partial_t u\|_{L^2(I;H^{-1}(\Omega))} \leqslant \|f\|_{L^2(I;L^2(\Omega))} + \|u^0\|_{L^2(\Omega)}.$$

Proof: (*i*) Es existiert also nach Satz 2.38 eine beschränkte Folge $u_m \in L^2(I; H_0^1(\Omega))$. D.h., es existiert eine Teilfolge $(u_{m_l})_{l\geqslant 1}$, welche *schwach* in $L^2(I; H_0^1(\Omega))$ gegen eine Funktion $u \in L^2(I; H_0^1(\Omega))$ konvergiert, also

$$(u_{m_l}, v) \to (u, v) \quad \forall v \in L^2(I; H_0^1(\Omega)) \quad (l \to \infty).$$

(*ii*) Wir wählen nun ein festes $N \in \mathbb{N}$ und eine Funktion $v \in C^1(I; H_0^1(\Omega))$ gemäß

$$v(x,t) = \sum_{k=1}^{N} d_k(t)\omega_k(x), \tag{2.18}$$

mit stetig differenzierbaren $d_k \in C^1(I)$. Für jedes $m > N$ folgt dann mit der schwachen Konvergenz:

$$(\partial_t u_m, v) + (\nabla u_m, \nabla v) \xrightarrow{m\to\infty} (\partial_t u, v) + (\nabla u, \nabla v) \quad \forall v \text{ gemäß (2.18)}.$$

Da die Funktionen gemäß (2.18) für $N \to \infty$ dicht in $L^2(I; H_0^1(\Omega))$ liegen gilt:

$$(\partial_t u, v) + (\nabla u, \nabla v) = (f, v) \quad \forall v \in L^2(\Omega),$$

und der schwache Grenzwert u ist Lösung der Differentialgleichung.

(*iii*) Es bleibt zu zeigen, dass u den Startwert erfüllt, dass also gilt $u(\cdot, 0) = u^0$. Für $v \in C^1(I; H_0^1(\Omega))$ mit $v(T) = 0$ folgt wegen $u \in C(\bar{I}; H_0^1(\Omega))$ aus der schwachen Formulierung (2.14):

$$\begin{aligned}\int_I (f, v)\,dt &= \int_I (\partial_t u, v)\,dt + \int_I (\nabla u, \nabla v)\,dt \\ &= \underbrace{(u(T), v(T))}_{=0} - \underbrace{(u(0), v(0))}_{(u^0, v(0))} - \int_I (u, \partial_t v)\,dt + \int_I (\nabla u, \nabla v)\,dt \\ &= -(u^0, v(0)) - \int_I (u, \partial_t v)\,dt + \int_I (\nabla u, \nabla v)\,dt\end{aligned}$$

Ebenso gilt für jedes $u_m \in L^2(I; V_m)$:

$$\int_I (f, v)\,dt = -(u_m(0), v(0)) - \int_I (u_m, \partial_t v)\,dt + \int_I (\nabla u_m, \nabla v)\,dt.$$

Aus der schwachen Konvergenz $u_m \rightharpoonup u$ folgt dann

$$(u_m(0), v(0)) \to (u(0), v(0)) = (u^0, v(0)) \quad \forall v(0) \in H_0^1(\Omega).$$

□

Eindeutigkeit der Lösung Angenommen $u_1, u_2 \in L^2(I; H_0^1(\Omega))$ mit $\partial_t u_1, \partial_t u_2 \in L^2(I; H^{-1}(\Omega))$ seien zwei schwache Lösungen von (2.14). Dann gilt für $w := u_1 - u_2$ mit $w(x, 0) = 0$

$$(\partial_t w(t), \phi) + (\nabla w(t), \nabla \phi) = 0 \text{ für fast alle } t \in I.$$

Für $\phi = w(t)$ erhalten wir hieraus mit Satz 2.37

$$0 = (\partial_t w(t), w(t)) + \|\nabla w(t)\|^2 = \frac{1}{2}\partial_t \|w(t)\|^2 + \|\nabla w(t)\|^2 \quad \text{für fast alle } t \in I.$$

Wir integrieren über t von 0 bis $s \in I$:

$$0 = \int_0^s \left\{ \frac{\partial}{\partial t}\|w(t)\|^2 + \|\nabla w(t)\|^2 \right\} dt = \|w(s)\|^2 - \underbrace{\|w(0)\|^2}_{=0} + \int_0^s \|\nabla w(t)\|^2\,dt.$$

Also

$$\|w(s)\|^2 = -\int_0^s \|\nabla w(t)\|^2\,dt \leqslant 0 \quad \Rightarrow \quad \|w(s)\| = 0 \quad \forall s > 0,$$

und die Lösung ist eindeutig.

Eigenschaften der Lösung Wie bei den elliptischen Differentialgleichung kann eine analytische Lösung nur in Spezialfällen angegeben werden. Wir versuchen aber formal durch die Methode der "Variablenseparation" eine Lösung zu erstellen. Wir betrachten das homogene Problem

$$\partial_t u - \Delta u = 0, \quad u(x,0) = u^0(x), \quad u = 0 \text{ auf } I \times \partial\Omega. \tag{2.19}$$

Mit dem Ansatz $u(x,t) = v(x)\psi(t)$ erhalten wir:

$$\partial_t u = \Delta u \quad \Rightarrow \quad \partial_t \psi v = \psi \Delta v \quad \Rightarrow \quad \frac{\partial_t \psi(t)}{\psi(t)} = \frac{\Delta v(x)}{v(x)} =: -\lambda.$$

Orts- und Zeitvariablen können unabhängig voneinander variert werden, die Gleichheit gilt für alle $x \in \Omega$ und $t \in I$. D.h., die beiden Faktoren $v(x)$ sowie $\psi(t)$ müssen Lösungen der Eigenwertprobleme

$$-\Delta v(x) = \lambda v(x) \; x \in \Omega, \quad -\psi'(t) = \lambda\psi(t) \; t \geqslant 0$$

mit homogenen Dirichlet-Bedingungen $v = 0$ auf $\partial\Omega$ im Ort und $\psi(0) = 1$ in der Zeit. Für das zeitliche Problem ergibt sich für jedes $\lambda > 0$ die Lösung:

$$\psi_\lambda(t) = e^{-\lambda t}, \quad t \geqslant 0.$$

Im Ort erhalten wir die Eigenwerte des Laplace-Operators, also mit

$$-\Delta\omega_j(x) = \lambda_j\omega_j(x) \quad j = 1, 2, \ldots,$$

ein L^2-Orthonormalsystem aus Eigenwerten $\omega_j \in L^2(\Omega)$ und Eigenwerte $0 < \lambda_{\min} = \lambda_1 \leqslant \lambda_2 \leqslant \cdots$. Jede Lösung von (2.19) lässt sich schreiben als

$$u(x,t) = \sum_{i=1}^{\infty} \mu_i e^{-\lambda_i t}\omega_i(x),$$

wobei die Entwicklungskoeffizienten μ_i gerade die Entwicklungskoeffizienten des Startwerts sind:

$$u^0(x) = u(x,0) = \sum_{i=1}^{\infty} \mu_i\omega_i.$$

Für eine parabolische Differentialgleichung mit rechter Seite $f = 0$ hängt die Lösung naturgemäß lediglich von den Anfangswerten ab. Die einzelnen Lösungskomponenten fallen exponentiell in der Zeit ab. Je größer der Eigenwert λ_i, umso schneller fällt die entsprechende Lösungskomponenten ab. Im Spezialfall $\Omega = (0,1) \subset \mathbb{R}$ gilt

$$-\partial_{xx}\omega_j(x) = \lambda_j\omega_j(x),$$

mit

$$\omega_j(x) = \frac{\sin(j\pi x)}{\left(\int_I \sin^2(j\pi x)\,dx\right)^{\frac{1}{2}}}, \quad \lambda_j = j^2\pi^2.$$

Die Lösung setzt sich also aus Sinus-Schwingungen zusammen. Je größer der Eigenwert, umso größer die Frequenz der Schwingung. Hieraus kann bei genauer Analyse gefolgert werden, dass die Lösung u von (2.19) zu jedem Zeitpunkt $t > 0$ beliebig regulär ist, auch wenn nur $u^0 \in L^2(\Omega)$ gilt. Dies liegt daran, dass gerade die hochfrequenten, also irregulären Komponenten im Startwert besonders schnell geglättet werden, da $\exp(-j^2\pi^2 t)$ für $j \to \infty$ schnell klein wird.

Wir beweisen für die Lösung der klassischen Formulierung

Lemma 2.40 (Parabolisches Maximumprinzip)**.** *Für jede Lösung der Wärmeleitungs-Ungleichung*

$$\partial_t u - \Delta u \leqslant 0 \text{ in } \Omega$$

gilt das Maximumprinzip. *Sie nimmt im Innern des halboffenen Zylinders* $Q := \Omega \times (0, T]$ *kein striktes Maximum an.*

Proof: Es sei $u \in C^1(I; C^2(\Omega) \cap C(\bar{\Omega}))$ eine klassische Lösung der Wärmeleitungsgleichung dann ist die Funktion $v_\varepsilon := u - \varepsilon t$ für jedes ε stetig und nimmt in einem Punkt $(x_0, t_0) \in \bar{Q}$ ihr Maximum an. Wir nehmen nun an, dass $(x_0, t_0) \in Q$, also im Innern des Zylinders liegt. Dann ist $\partial_{xx} v_\varepsilon(x_0, t_0) \leqslant 0$ und also

$$\partial_t v_\varepsilon(x_0, t_0) \leqslant \partial_t v_\varepsilon(x_0, t_0) - \partial_{xx} v_\varepsilon(x_0, t_0) = \partial_t u(x_0, t_0) - \varepsilon - \partial_{xx} u(x_0, t_0) \leqslant -\varepsilon.$$

Da die Funktion v_ε auf $\bar{Q}$ stetig ist, gilt dann auch in einer Umbebung (x_0, t) für $t \in (t_0 - h, t_0]$, dass

$$\partial_t v_\varepsilon(x_0, t) \leqslant -\frac{1}{2}\varepsilon.$$

Dies führt auf einen Widerspruch zu $v_\varepsilon(x_0, t_0) < v_\varepsilon(x_0, t')$ und die Funktion v_ε nimmt ihr Maximum notwendig bei $t = 0$ an. Da v_ε stetig ist und ε beliebig gilt diese Aussage auch im Grenzfall $\varepsilon \to 0$. □

Aus dem parabolischen Maximumprinzip lässt sich wie im elliptischen Fall die Eindeutigkeit der klassischen Lösung herleiten. Weiter kann gezeigt werden, dass die Lösung der homogenen Gleichung

$$\partial_t u - \Delta u = 0, \quad u(x, 0) = u^0(x), \quad u = 0 \text{ auf } I \times \partial\Omega,$$

mit positiven Anfangsdaten $u^0 \geqslant 0$ für alle Zeiten nicht negativ bleibt:

$$0 \leqslant u(x, t) \leqslant \max_{z \in \Omega} u^0(z).$$

Im folgenden befassen wir uns mit dem allgemeinen Fall der inhomogenen Wärmeleitungsgleichung. Es gilt

Lemma 2.41 (Beschränktheit der Lösung)**.** *Für die Lösung der inhomogenen Wärmeleitungsgleichung (2.11) gilt die a priori Abschätzung*

$$\|u(t)\| \leqslant e^{-\lambda t}\|u^0\| + \lambda^{-1} \sup_{t \geqslant 0} \|f(t)\|,$$

mit dem kleinsten Eigenwert λ des Laplace-Operators $L = -\Delta$ auf Ω mit homogenen Dirichlet-Randdaten.

Proof: (*i*) Wir stellen die Lösung $u = v + w$ als Summe der Lösungen zweier Hilfsprobleme dar:

$$\begin{aligned} \partial_t v - \Delta v = 0 \text{ in } I \times \Omega, \quad v\big|_{t=0} = u^0, \quad v\big|_{\partial\Omega} = 0 \\ \partial_t w - \Delta w = f \text{ in } I \times \Omega, \quad v\big|_{t=0} = 0, \quad v\big|_{\partial\Omega} = 0. \end{aligned}$$

Diese Aufteilung (Superpositionsprinzip) gilt offenbar wegen der Linearität der Wärmeleitungsgleichung und sie hilft, die beiden Dateneinflüsse durch Anfangswert und rechter Seite zu trennen $\|u(t)\| \leqslant \|v(t)\| + \|w(t)\|$.

(*ii*) Wir starten mit einer Abschätzung für $v(t)$. Wir überführen die erste Gleichung gemäß (2.14) in die variationelle Formulierung und setzen als Testfunktion $\phi = v$. Es gilt:

$$\frac{1}{2} d_t \|v(t)\|^2 + \|\nabla v(t)\|^2 = 0.$$

Weiter multiplizieren wir mit $e^{2\lambda t}$ und erhalten

$$\frac{1}{2} d_t (e^{2\lambda t}\|v(t)\|^2) - \lambda e^{2\lambda t}\|v(t)\|^2 + e^{2\lambda t}\|\nabla v(t)\|^2 = 0.$$

λ ist der kleinste Eigenwert des Laplace-Operators, es gilt also $\lambda\|v\|^2 \leqslant \|\nabla v\|^2$ und es folgt:

$$\frac{1}{2} d_t (e^{2\lambda t}\|v(t)\|^2) \leqslant 0.$$

Integration über t liefert

$$e^{2\lambda t}\|v(t)\|^2 \leqslant \|v(0)\|^2 = \|u^0\|^2 \quad \Rightarrow \quad \|v(t)\| \leqslant e^{-\lambda t}\|u^0\|.$$

(*iii*) Wir betrachten nun $\|w(t)\|$. Wir transformieren entsprechend durch Multiplikation mit $w(t)$ und Integration im Ort. Mit der Young'schen Ungleichung $|ab| \leqslant \frac{1}{2}\lambda^{-1}a^2 + \frac{1}{2}\lambda b^2$ gilt:

$$\frac{1}{2} d_t \|w(t)\|^2 + \|\nabla w(t)\|^2 = (f, w) \leqslant \frac{1}{2}\lambda^{-1}\|f\|^2 + \frac{1}{2}\lambda\|w\|^2$$

und mit $\lambda\|w(t)\|^2 \leqslant \|\nabla w(t)\|^2$ folgt

$$d_t \|w(t)\|^2 + \|\nabla w(t)\|^2 \leqslant \lambda^{-1}\|f\|^2.$$

Wir multiplizieren mit $e^{\lambda t}$

$$d_t(e^{\lambda t}\|w(t)\|^2) - \lambda e^{\lambda t}\|w(t)\|^2 + e^{\lambda t}\|\nabla w(t)\|^2 \leqslant e^{\lambda t}\lambda^{-1}\|f\|^2,$$

und erhalten also wie in Teil (ii)

$$d_t(e^{\lambda t}\|w(t)\|^2) \leqslant \lambda^{-1}e^{\lambda t}\|f\|^2.$$

Integration über t liefert

$$e^{\lambda t}\|w(t)\|^2 - \underbrace{\|w(0)\|^2}_{=0} \leqslant \lambda^{-1}\int_0^t e^{\lambda s}\|f(s)\|^2\,ds,$$

bzw.

$$\begin{aligned}\|w(t)\|^2 &\leqslant \lambda^{-1}e^{-\lambda t}\int_0^t e^{\lambda s}\|f(s)\|^2\,ds\\ &\leqslant \lambda^{-1}\sup_{t\geqslant 0}\|f(t)\|^2 e^{-\lambda t}\int_0^t e^{\lambda s}\,ds.\end{aligned}$$

Es gilt

$$e^{-\lambda t}\int_0^t e^{\lambda s}\,ds \leqslant \lambda^{-1},$$

und somit folgt das Ergebnis:

$$\|w(t)\| \leqslant \lambda^{-1}\sup_{t\geqslant 0}\|f(t)\|.$$

□

Wie bereits im eindimensionalen diskutiert fällt der Einfluss des Startwerts exponentiell ab. Von der rechten Seite hängt die Lösung stetig ab. Die Wärmeleitungsgleichung hat eine sehr wichtige Glättungseigenschaft. Auf für irreguläre Anfangsdaten ist die Lösung für $t > 0$ stets regulär (wenn dies die rechte Seite erlaubt):

Lemma 2.42 (Regularität der Lösung)**.** *Für die Lösung der homogenen Wärmeleitungsgleichung mit $f = 0$ gilt*

$$\begin{aligned}\|\partial_t u(t)\| + \|\Delta u(t)\| &\leqslant 2\|\Delta u^0\|\\ \|\partial_t u(t)\| + \|\Delta u(t)\| &\leqslant 2t^{-1}\|u^0\|,\end{aligned}$$

vorrausgesetzt der Anfangswert erfüllt jeweils die notwendige Regularität.

Proof: (i) Für den Beweis nutzen wir wieder die Darstellung der Lösung in den Eigenfunktionen ω_j des Laplace-Operators

$$u(x,t) = \sum_{j\geqslant 0} u_j^0\omega_j(x)e^{-\lambda_j t},$$

wobei u_j^0 gerade die Entwicklungskoeffizienten des Anfangswerts sind. Es gilt:

$$\partial_t u(x,t) = \Delta u(x,t) = -\sum_{j\geqslant 0} u_j^0 \omega_j(x)\lambda_j e^{-\lambda_j t},$$

(*ii*) Wir wollen jetzt die Norm der Lösung bestimmen. Die ω_j sind L^2-orthonormal und mit der *Parseval'schen Identität* gilt:

$$\|\partial_t u(t)\|^2 = \|\Delta u(t)\|^2 = \sum_{j\geqslant 0} (u_j^0)^2 \lambda_j^2 e^{-2\lambda_j t}. \tag{2.20}$$

Hieraus folgern wir unmittelbar die erste Ungleichung:

$$\|\partial_t u(t)\|^2 = \|\Delta u(t)\|^2 \leqslant \sum_{j\geqslant 0} (u_j^0)^2 \lambda_j^2 = \|\Delta u^0\|^2.$$

(*iii*) Weiter mit (2.20) und der Ungleichung $xe^{-x} \leqslant 1$ für $x \geqslant 0$:

$$\|\partial_t u(t)\|^2 = \|\Delta u(t)\|^2 = t^{-2} \sum_{j\geqslant 0} (u_j^0)^2 \underbrace{(t\lambda_j)^2 e^{-2\lambda_j t}}_{\leqslant 1} \leqslant t^{-2}\|u^0\|^2.$$

□

Dieser Satz bedeutet also, dass auch für Anfangswerte u^0 welche lediglich in $L^2(\Omega)$ liegen die Lösung für jeden Zeitpunkt $t > 0$ glatt ist, dass also z.B. $u(t) \in H^2(\Omega)$ liegt. Diese Argumentation kann beliebig fortgeführt werden und wir erhalten die Ungleichung

$$\|\partial_t^p u(t)\| + \|\nabla^{2p} u(t)\| \leqslant c(p) t^{-p} \|u^0\|, \quad t > 0,\ p \in \mathbb{N}.$$

Auf

3 Die Finite Elemente Methode für elliptische Probleme

Wir betrachten elliptische partielle Differentialgleichungen auf einem Lipschitz-Gebiet $\Omega \subset \mathbb{R}^d$ mit $d \geqslant 2$:

$$-\operatorname{div}(A\nabla u) + b \cdot \nabla u + cu = f.$$

Es sei $A \in [L^\infty(\Omega)]^{d\times d}$ eine positiv definite Koeffizientenmatrix, $b \in [L^\infty(\Omega)]^d$ ein divergenzfreier *Transportvektor* (also $\nabla \cdot u = 0$), sowie $c \in L^\infty(\Omega)$ mit $c \geqslant 0$.

Laut Satz 2.16 ist jede Lösung der Differentialgleichung auch Lösung des Variationsproblems

$$u \in V: \ a(u,\phi) = (f,\phi) \quad \forall \phi \in V, \tag{3.1}$$

mit

$$a(u,\phi) = \big(A\nabla u, \nabla\phi\big)_\Omega + \big((b\cdot\nabla)u, \phi\big) + \big(cu,\phi\big).$$

Die genaue Wahl des Test- und Ansatzraums V hängt von den Randwerten des Problems ab. Bei Verwendung von Dirichlet-Werten ist $V = H^1_0(\Omega)$, bei reinen Neumann-Werten ist $V = H^1(\Omega)/\mathbb{R}$.

Die Lösung dieses unendlich dimensionalen Problems ist mit analytischen Mitteln nur in Spezialfällen möglich. Wir werden daher Verfahren zur Approximation von Lösungen untersuchen. Die sogenannten *Finite Differenzen-Verfahren* zielen auf eine Approximation der Ableitungen im Differentialoperator: Die Lösung wird nicht in ganz Ω berechnet, sondern nur in *diskreten Punkten* $x_i \in \Omega$. Zwischen diesen Punkten werden die Ableitungen durch Differenzenapproximationen dargestellt. In dieser Vorlesung betrachten wir ausschließlich *Variationsmethoden*. Wir diskretisieren die variationelle Formulierung durch die Wahl von endlich dimensionalen Test- und Ansatzräumen $V_h \subset V$.

3.1 Allgemeine Galerkin-Verfahren

Es sei $V_h \subset V$ ein endlich dimensionaler Teilraum mit

$$\dim V_h = N_h \in \mathbb{N}.$$

Der *Diskretisierungsparameter* h beschreibt die *Feinheit* der Diskretisierung. Je kleiner der Parameter h, umso größer die Räume:

$$N_h \to \infty \quad (h \to 0).$$

Definition 3.1 (Galerkin-Approximation)**.** *Sei* $a(\cdot,\cdot)$ *eine stetige, elliptische Bilinearform. Die Galerkin-Approximation von* (3.1) *im Teilraum* $V_h \subset V$ *ist gegeben durch:*

$$u_h \in V_h: \quad a(u_h, \phi_h) = (f, \phi_h) \quad \forall \phi_h \in V_h. \tag{3.2}$$

Zur Diskretisierung werden Test- und Ansatzräume auf endlich dimensionale Teilräume $V_h \subset V$ eingeschränkt. Dieser endlich dimensionale Teilraum V_h erbt die Eigenschaften von V, er ist wieder ein Hilbertraum. In der Analyse von Galerkin-Approximationen werden wir viele Elemente der theoretischen Betrachtungen wieder anwenden können.

3.1.1 Lösbarkeit und Galerkin-Orthogonalität

Lemma 3.2 (Lösbarkeit der Galerkin-Approximation)**.** *Sei* $a : V \times V \to \mathbb{R}$ *eine stetige* (*mit Konstante* $M > 0$) *und elliptische* (*mit Konstante* $\gamma > 0$) *Bilinearform und sei* $f \in V^*$ *ein stetiges lineares Funktional. Dann gibt es eine eindeutige Lösung* $u_h \in V_h$ *von* (3.2) *und es gilt:*

$$\|u_h\|_V \leqslant \frac{1}{\gamma}\|f\|_{V^*}.$$

Proof: Der endlich dimensionale Teilraum $V_h \subset V$ erbt die lineare Struktur von V sowie Skalarprodukt und Norm. Als endlich dimensionaler Teilraum ist V_h abgeschlossen, also ein vollständig und hier ein Hilbertraum.

Der Satz von Lax-Milgram kann angewendet werden und liefert Existenzaussage und Eindeutigkeit. □

Aufgrund der Teilraumbeziehung $V_h \subset V$ können bei der Analyse von diskreten Galerkin-Approximationen die funktionalanalytischen Eigenschaften des Hilbertraums V unmittelbar übertragen werden.

Lemma 3.3. *Unter den Voraussetzungen von Satz 3.2 ist Problem* (3.2) *äquivalent zur Lösung eines linearen Gleichungssystems*

$$\mathbf{A}_h \mathbf{u}_h = \mathbf{b}_h,$$

mit einem Koeffizientenvektor $\mathbf{u}_h \in \mathbb{R}^{N_h}$*, einer rechten Seite* $\mathbf{b}_h \in \mathbb{R}^{N_h}$ *und mit einer positiv definiten Matrix* $\mathbf{A}_h \in \mathbb{R}^{N_h \times N_h}$*. Falls* $b = 0$ (*kein Transport-Term*) *so ist die Matrix symmetrisc*

$$\mathbf{A}_h = \mathbf{A}_h^T.$$

Proof: (*i*) Wir wählen eine Basis $\{\phi_1, \ldots, \phi_{N_h}\} \subset V_h$ von V_h. Jede Funktion $v_h \in V_h$ kann in dieser Basis entwickelt werden

$$v_h(x) = \sum_{j=1}^{N_j} \mathbf{v}_j \phi_j(x).$$

Der Vektor $\mathbf{b}_h$ auf der rechten Seite berechnet sich als

$$(\mathbf{b}_h)_i := \mathbf{b}_i = (f, \phi_i), \quad i = 1, \dots, N_h.$$

Für die Lösung u_h mit Koeffizientenvektor $\mathbf{u}_h$ gilt wegen der Linearität der Bilinearform:

$$\begin{aligned} & a(u_h, \phi_i) = (f, \phi_i) \quad \forall i = 1, \dots, N_h \\ \Leftrightarrow \quad & \sum_{j=1}^{N_h} \underbrace{a(\phi_j, \phi_i)}_{=:\mathbf{A}_{ij}} \mathbf{u}_j = \mathbf{b}_i \quad \forall i = 1, \dots, N_h \\ \Leftrightarrow \quad & \mathbf{A}_h \mathbf{u}_h = \mathbf{b}_h, \quad \text{mit } \mathbf{A}_h := (\mathbf{A}_{ij})_{i,j=1}^{N_h}, \quad \mathbf{A}_{ij} := a(\phi_j, \phi_i). \end{aligned}$$

(*ii*) Für beliebigen Koeffizientenvektor $\mathbf{v}_h \in \mathbb{R}^{N_h}$ gilt:

$$\langle \mathbf{A}_h \mathbf{v}_h, \mathbf{v}_h \rangle = a(v_h, v_h) \geqslant \gamma \|v_h\|_V^2 > 0,$$

wobei γ die Elliptizitätskonstante der Bilinearform ist.

(*iii*) Schließlich gilt im Fall $b = 0$

$$\begin{aligned} \mathbf{A}_{ij} = a(\phi_j, \phi_i) &= (A\nabla\phi_j, \nabla\phi_i) + c(\phi_j, \phi_i) \\ &= (\nabla\phi_j, A\nabla\phi_i) + c(\phi_i, \phi_a) = a(\phi_i, \phi_j) = \mathbf{A}_{ji}. \end{aligned}$$

Hier haben wir genutzt, dass die Koffizientenmatrix $A : \Omega \to \mathbb{R}^{d \times d}$ symmetrisch ist. □

Die Matrix $\mathbf{A}_h$ wird *Steifigkeitsmatrix* genannt. Diese Matrix "erbt" die Eigenschaften der Bilinearform: ist diese symmetrisch, so ist auch die Steifigkeitsmatrix symmetrisch. Aus der Beschränktheit (Stetigkeit) der Bilinearform folgt die gleichmäßige Beschränktheit der Matrix. Als positiv definite Matrix ist $\mathbf{A}_h$ stets regulär.

Die Wahl der Basis war für den weiteren Beweis unwesentlich. Die Matrix $\mathbf{A}$ ändert sich natürlich je nach Basis, die Eigenschaften Regularität, positive Definiteheit und gegebenenfalls Symmetrie bleiben jedoch stets bestehen.

Das allgemeine Galerkin-Verfahren besteht nun aus den folgenden Schritten:

1. Wähle einen diskreten Teilraum $V_h \subset V$.
2. Wähle eine Basis $\{\phi_i,\ i = 1, \dots, N_h\}$ von V_h.
3. Löse das lineare Gleichungssystem $\mathbf{A}_h \mathbf{u}_h = \mathbf{b}_h$.

Um ein konkretes Galerkin-Verfahren zu erhalten muss also zunächst ein Teilraum V_h gewählt werden. Dieser Teilraum soll möglichst groß sein und nahe bei V liegen, nur so ist eine gute Approximation zu erwarten. Andererseits soll das lineare Gleichungssystem möglichst effizient zu lösen sein, d.h., die Matrix $\mathbf{A}_h$ soll gut konditioniert und nicht zu groß sein.

Schlüssel zur Analyse von Galerkin-Approximationen ist die:

Lemma 3.4 (Galerkin-Orthogonalität)**.** *Der Fehler* $u - u_h \in V$ *der Galerkin-Approximation* $u_h \in V_h \subset V$ *"steht orthogonal" auf dem Testraum* V_h*:*

$$a(u - u_h, \phi_h) = 0 \quad \forall \phi_h \in V_h.$$

Proof: Diese wichtige Eigenschaft folgt durch Subtraktion von (3.1) und (3.2). □

Remark 3.5. *Falls die Bilinearform* $a(\cdot,\cdot)$ *symmetrisch ist, so stellt sie ein Skalarprodukt auf* V *und* V_h *dar. In diesem Fall fällt die Galerkin-Orthogonalität mit dem üblichen Orthogonalitätsbegriff zusammen.*

Mit der Galerkin-Orthogonalität erhalten wir schnell den folgenden Approximationssatz:

Lemma 3.6 (Lemma von Cea)**.** *Es seien die Voraussetzungen von Satz 3.2 erfüllt. Es sei* $u \in V$ *die Lösung von* (3.1) *und* $u_h \in V_h$ *die Lösung von* (3.2)*. Es gilt die Fehlerabschätzung*

$$\|u - u_h\|_V \leqslant \frac{M}{\gamma} \inf_{v_h \in V_h} \|u - v_h\|_V.$$

Proof: Aus der Elliptizität folgt zusammen mit der Galerkin-Orthogonalität und der Stetigkeit für ein beliebiges $v_h \in V_h \subset V$:

$$\begin{aligned}
\gamma \|u - u_h\|_V^2 &\leqslant a(u - u_h, u - u_h) = a(u - u_h, u - u_h + v_h - v_h) \\
&= a(u - u_h, u - v_h) + a(u - u_h, \underbrace{v_h - u_h}_{\in V_h}) \\
&= a(u - u_h, u - v_h) \\
&\leqslant M \|u - u_h\|_V \, \|u - v_h\|_V.
\end{aligned}$$

□

Das Lemma von Cea besagt, dass die Galerkin-Lösung bis auf die Konstante $\frac{M}{\gamma}$ die beste mögliche Approximation von u im Funktionenraum V_h ist. Die Approximationsgüte eines Galerkin-Ansatzes hängt also einzig von der Approximierbarkeit der Lösung $u \in H_0^1(\Omega)$ mit Funktionen $\phi_h \in V_h$ ab. Für die Herleitung von Fehlerabschätzungen muss eine spezielle Approximation $\phi_h \in V_h$ gewählt werden, für welche Fehlerabschätzungen existieren. Hier werden sich lokale Interpolationen anbieten. Interpolationen in diskrete Galerkin-Räume werden wir detailliert untersuchen.

Remark 3.7 (Bestapproximation)**.** *Im Fall, dass* $a(\cdot,\cdot)$ *symmetrisch ist, ist durch*

$$\|v\|_a := \sqrt{a(v,v)} \quad \forall v \in H_0^1(\Omega),$$

eine Norm definiert. In dieser Norm gilt dann die Bestapproximationseigenschaft

$$\|u - u_h\|_a = \inf_{\phi_h \in V_h} \|u - \phi_h\|_a.$$

3.1.2 Einige Begriffe

Das Prinzip der Galerkin-Approximationen lässt sich noch erweitern und wir führen abschließend einige weitere Begriffe ein.

Definition 3.8 (Konforme und konsistente Galerkin-Approximation)**.** *Die Approximation*

$$u_h \in V_h : \quad a(u_h, \phi_h) = (f, \phi_h) \quad \forall \phi_h \in V_h,$$

der Variationsgleichung

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

mit einem endlich dimensionalen Raum V_h *heißt* V-konform, *falls* $V_h \subset V$, *ansonsten, im Fall* $V_h \not\subset V$ *heißt sie* nicht V-konform.

Die Galerkin-Approximation heißt konsistent, *falls die exakte Lösung die diskrete Gleichung erfüllt, d.h.*

$$a(u, \phi_h) = (f, \phi_h) \quad \forall \phi_h \in V_h.$$

Der Begriff der *Konsistenz* scheint zunächst überflüssig. Jeder *konforme* Ansatz ist bisher auch *konsistent*. Wir werden jedoch Verfahren kennenlernen, welche die Galerkin-Eigenschaft nicht im strengen Sinne erfüllen und bei denen das diskrete Problem mit einer modifizierten diskreten Bilinearform definiert ist, d.h.

$$u_h \in V_h : \quad a_h(u_h, \phi_h) = (f, \phi_h) \quad \forall \phi_h \in V_h,$$

mit $a(\cdot,\cdot) \neq a_h(\cdot,\cdot)$. Bei Betrachtung dieser Verallgemeinerung fallen die beiden Begriffe *konform* und *konsistent* nicht mehr zusammen.

Oft ist es notwendig, nicht-konforme Approximationen zu untersuchen. Die fehlende Konformität macht die Analyse weit aufwändiger, da ohne die Teilraumbeziehung auch die Galerkin-Orthogonalität verloren geht. Dann gilt das Lemma von Cea nicht mehr.

Ein anderer Spezialfall ist die Wahl von unterschiedlichen Ansatz- und Test-Räumen:

Definition 3.9 (Petrov-Galerkin-Verfahren)**.** *Es seien* $V_h \subset V$ *und* $W_h \subset V$ *zwei endlich dimensionale Teilräume. Die Approximation*

$$u_h \in V_h : \quad a(u_h, \phi_h) = (f, \phi_h) \quad \forall \phi_h \in W_h$$

heißt Petrov-Galerkin-Approximation.

Petrov-Galerkin-Approximationen können gegebenenfalls wieder nicht-konform sein, wenn $V_h \not\subset V$ oder $W_h \not\subset V$ (oder beides). Für Petrov-Galerkin-Verfahren, welche konform im Testraum $W_h \subset V$ sind, gilt zwar die Galerkin-Orthogonalität bzgl. des Testraums $W_h \subset V$.

Die diskrete Lösung $u_h \in V_h \neq W_h$ liegt jedoch nicht in diesem Raum. Im Beweis zum Lemma von Cea 3.6 erhalten wir eine Abschätzung durch den Ansatz:

$$\gamma\|u-u_h\|_V \leqslant a(u-u_h,u-u_h) = a(u-u_h,u-v_h) + a(u-u_h,v_h-\underbrace{u_h}_{\notin W_h}) \quad \forall v_h \in W_h.$$

Wegen der fehlenden Galerkin-Orthogonalität bleibt ein *Konformitätsterm* erhalten. Die Herleitung einer Abschätzung für diesen Term kann sehr aufwändig sein.

3.1.3 Wahl der Ansatzräume

Wie bereits argumentiert sollte der Ansatzraum V_h so gewählt werden, dass er über gute Approximationseigenschaften von Funktionen $u \in V$ verfügt. Daneben muss das lineare Gleichungssystem einfach zu lösen sein.

Globale Polynomräume Als diskreten Funktionenraum V_h wählen wir den Raum aller Polynome bis zu einem bestimmten Grad. So können wir z.B. in d räumlichen Dimensionen den Raum

$$V_h^{(r)} := \operatorname{span}\{x_1^{\alpha_1}x_2^{\alpha_2}\cdots x_d^{\alpha_d} \mid 0 \leqslant \alpha_1,\ldots,\alpha_d \leqslant r\}$$

wählen. Dieser Raum hat für großes r gute Approximationseigenschaften und wir kennen Interpolationsabschätzungen um den Fehler angeben zu können. Es gilt

$$\dim\left(V_h^{(r)}\right) = (r+1)^d.$$

Wir betrachten den einfachen Fall $d = 1$ und das Laplace-Problem

$$a(u,\phi) = (\partial_x u, \partial_x \phi)$$

auf dem Gebiet $\Omega = (0,1)$. Mit der eindimensionalen Monombasis

$$V_h = \operatorname{span}\{x,\ldots,x^r\},$$

also $\phi_i = x^i$ für $i = 1,\ldots,r$ gilt dann

$$A_{ij} = a(\phi_j,\phi_i) = \int_0^1 (ij)x^{i+j-2}\,dx = \frac{ij}{i+j-1}x^{i+j-1}\Big|_0^1 = \frac{ij}{i+j-1}.$$

Das konstante Monom $x^0 = 1$ berücksichtigen wir nicht im Ansatzraum, da dieses Im Kern des Laplace-Operators liegt. Durch Vorgabe von Randwerten ist der konstante Wert eindeutig festgelegt.

Nun gilt

$$\mathbf{A}_h = \begin{pmatrix} 1 & 1 & 1 & 1 & \cdots \\ 1 & \frac{4}{3} & \frac{3}{2} & \frac{8}{5} & \\ 1 & \frac{3}{2} & \frac{9}{5} & 2 & \\ 1 & \frac{8}{5} & \frac{16}{7} & \frac{5}{2} & \\ \vdots & & & & \ddots \end{pmatrix} \approx \begin{pmatrix} 1 & 1 & 1 & 1 & \cdots \\ 1 & 1.33 & 1.5 & 1.6 & \\ 1 & 1.5 & 1.8 & 2 & \\ 1 & 1.6 & 2.29 & 2.5 & \\ \vdots & & & & \ddots \end{pmatrix}$$

Diese Matrix ist sehr ähnlich zur Hilbertmatrix mit

$$\mathbf{H}_{ij} = \frac{1}{i+j-1}.$$

Es gilt

$$\mathbf{A}_h = \mathbf{D}^{-1}\mathbf{H}\mathbf{D}^{-1},$$

mit der Diagonalmatrix

$$\mathbf{D} = \begin{pmatrix} 1 & 0 & \cdots \\ 0 & 2 & \\ \vdots & & \ddots \end{pmatrix}$$

voll besetzt. Wie bei der Hilbertmatrix gilt $\mathrm{cond}_2(\mathbf{A}_h) = O(e^{N_h})$. Z.B. ist

$$\begin{aligned} r = 1 \quad & \mathrm{cond}_2(A_h) \approx 1, \\ r = 2 \quad & \mathrm{cond}_2(A_h) \approx 14, \\ r = 3 \quad & \mathrm{cond}_2(A_h) \approx 280, \\ r = 4 \quad & \mathrm{cond}_2(A_h) \approx 7 \cdot 10^3, \\ r = 5 \quad & \mathrm{cond}_2(A_h) \approx 2 \cdot 10^5, \\ r = 6 \quad & \mathrm{cond}_2(A_h) \approx 6 \cdot 10^6, \\ r = 7 \quad & \mathrm{cond}_2(A_h) \approx 2 \cdot 10^8. \end{aligned}$$

Ein lineares Gleichungssystem mit dieser Matrix kann selbst für moderate $N_h \approx 100$ nicht numerisch gelöst werden. Dieser Galerkin-Ansatz scheidet also aus.

Wir nehmen zunächst an, dass die Bilinearform $a(\cdot,\cdot)$ symmetrisch, also ein Skalarprodukt ist. Angenommen, die Basis $\{\phi_i,\ i = 1,\dots,N_h\}$ sei a-orthogonal:

$$a(\phi_i, \phi_j) = 0 \quad \forall i \neq j.$$

In diesem Fall wäre die Matrix $\mathbf{A}_h$ eine Diagonalmatrix und das Lösen des linearen Gleichungssystems wäre trivial. Zur Orthogonalisierung der Basis könnte etwa das Gram-Schmidt-Verfahren verwendet werden. In der Praxis treten bei dieser Orthogonalisierung jedoch zu große Rundungsfehler auf. Die resultierende Basis wäre nicht mehr exakt orthogonal. Auch dieser Ansatz kann numerisch nicht realisiert werden.

Trigonometrische Ansatzräume Es sei $\Omega = (0,1) \times (0,1) \subset \mathbb{R}^2$. Als Ansatzraum wählen wir

$$V_h := \{ \sum_{i,j=0}^{m} c_{ij} \sin(i\pi x) \sin(j\pi y),\ c_{ij} \in \mathbb{R}\}, \quad h := \frac{1}{m}, \quad N_h := (m+1)^2.$$

Als Basisfunktionen werden die Fourier-Koeffizienten

$$\phi_{ij} = \sin(i\pi x) \sin(j\pi y)$$

gewählt. Angewendet auf die Poisson-Gleichung ist die resultierende Matrix sehr gut konditioniert mit $\mathrm{cond}_2(\boldsymbol{A}_h) = O(N)$ und das Gleichungssystem kann sehr effizient (schnelle Fourier-Transformation) gelöst werden. Dieser Ansatz wird *Spektral-Verfahren* genannt. Er ist höchst leistungsfähig, falls das Gebiet Ω und der Differentialoperator einfach gehalten sind, jedoch nicht auf allgemeine Fälle übertragbar.

"Fast orthogonale" Ansätze (Finite-Elemente) Dieser Ansatz stellt einen Kompromiss zwischen Approximationsgenauigkeit und Effizienz dar. Als Basisfunktionen werden stückweise polynomiale Funktionen gewählt, die jeweils nur einen möglichst kleinen Träger besitzen. Insbesondere soll für verschiedene Basisfunktionen gelten:

$$\mathrm{supp}(\phi_i) \cap \mathrm{supp}(\phi_j) \neq \emptyset \text{ nur für sehr wenige } i \neq j.$$

Durch diese Konstruktion ist die Steifigkeitsmatrix $\boldsymbol{A}_h$ automatisch *dünn besetzt* mit nur sehr wenigen Einträgen in jeder Matrix-Zeile. Diesen Ansatz werden wir im Detail untersuchen.

3.2 Finite Elemente Methode

Bei der Finite Elemente Methode werden Basisfunktionen des Raums V_h mit einem kleinen, lokalen Träger gewählt. Diese Basisfunktionen sind als stückweise Polynome definiert. Grundlage ist zunächst eine *Triangulierung* des Gebiets Ω in ein *Finite-Elemente Gitter*.

3.2.1 Triangulierung und lineare Finite Elemente

Es sei $\Omega \subset \mathbb{R}^d$ mit $d \geqslant 2$ ein offenes Gebiet mit zunächst stückweise polygonalem Rand. Alle Innenwinkel seien positiv. Wir definieren:

Definition 3.10 (Finite Elemente Gitter)**.** *Eine Zerlegung von Ω in offene Teilgebiete* T, Zellen *oder* Elemente *genannt, heißt* Triangulierung *oder* Gitter $\Omega_h = \{T_i;\ i = 1, \ldots, N_h^T\}$ *von Ω. Die Elemente sind üblicherweise Dreiecke, Tetraeder, konvexe Vierecke oder konvexe Hexaeder. Wir definieren die Zellgröße $h_T := diam(T)$ und maximale Gitterweite $h := \max_{T \in \Omega_h} h_T$. Die Eckpunkte der Elemente heißen* Knoten *und wir fassen das Gitter bei Bedarf auch als Knoten-Gitter $\Omega_h := \{x_i,\ i = 1, \ldots, N_h\}$ auf. Das Gitter heißt*

Strukturregulär *Je zwei unterschiedliche Zellen $T_1, T_2 \in \Omega_h$, $T_1 \neq T_2$ überlappen sich entweder nicht $T_1 \cap T_2 = \emptyset$, oder in einem gemeinsamen Eckpunkt, oder in einer ganzen Seite.*

Formregulär (*Im Fall von Dreiecken*) *Alle Dreiecke der Triangulierung Ω_h sind von ähnlicher Gestalt. Für den Inkreisradius ρ_T und Zelldurchmesser h_T gilt gleichmäßig:*

$$\max_{T \in \Omega_h} \frac{h_T}{\rho_T} \leqslant c_f \quad (h \to 0).$$

Größenregulär *Die Elemente $T \in \Omega_h$ sind von ähnlicher Größe. Es gilt gleichmäßig:*

$$\max_{T \in \Omega_h} h_T \leqslant c_g \min_{T \in \Omega_h} h_T \quad (h \to 0).$$

Es müssen nicht unbedingt alle drei Regularitätseigenschaften gelten um von einem sinnvollen Finite Elemente Gitter zu sprechen. Um zum Beispiel *lokal verfeinerte Gitter* zu ermöglichen wird es notwendig sein, die Größenregularität fallen zu lassen. Auf mögliche Abschwächungen der Regularitätsbedingungen kommen wir bei Bedarf zurück.

In einem Finite Elemente Gitter können verschiedene Elemente, also z.B. Kombinationen aus Hexaedern, Tetraedern und Prismen betrachtet werden. Wir werden jedoch nur reine Dreiecks-, Vierecks-, Tetraeder- oder Hexaedergitter untersuchen.

Eine Finite Elemente Basis wird nun unter Zuhilfenahme des Gitters definiert. Das einfachste Beispiel, die stückweise linearen Finite sind gegeben als:

$$V_h^{(1)} := \{v_h \in C(\bar{\Omega}) : \ v_h\big|_T \in P_1(T), \ T \in \mathcal{T}_h, \ v_{h|\partial\Omega} = 0\}.$$

Für diesen Vektorraum gilt $V_h^{(1)} \subset H_0^1(\Omega)$. Weiter ist jedes $v_h \in V_h^{(1)}$ durch die Vorgabe der Funktionswerte an den Eckpunkten $x_i \in \Omega_h$ der Triangulierung eindeutig beschrieben.

Durch die Vorschrift:

$$\phi_h^{(i)}(x_j) = \delta_{ij} \quad \forall i, j = 1, \ldots, N_h,$$

wird jedem inneren Knoten eine Basisfunktion zugeordnet. Diese Basis heißt die *Knoten-Basis* des Raums $V_h^{(1)}$ der linearen Finiten Elemente. Siehe Abbildung 3.1. Die Basis erstreckt sich nur über die inneren Knoten des Gebiets $x_i \notin \partial\Omega$. Auf diese Weise ist sichergestellt, dass $V_h^{(1)} \subset H_0^1(\Omega)$, dass also Null-Randwerte von allen diskreten Funktionen erfüllt sind.

Jede Funktion $v_h \in V_h^{(1)}$ besitzt die eindeutige Basisdarstellung:

$$v_h(x) = \sum_{i=1}^{N_h} \mathbf{v}_i \phi_h^{(i)}(x),$$

mit einem Koeffizientenvektor $\mathbf{v} \in \mathbb{R}^{N_h}$.

Die Summe erstreckt sich nur über die inneren Knoten der Triangulierung. In den Randknoten ist der Wert der diskreten Vektoren aufgrund der Dirichlet-Randwerte stets Null. Wir

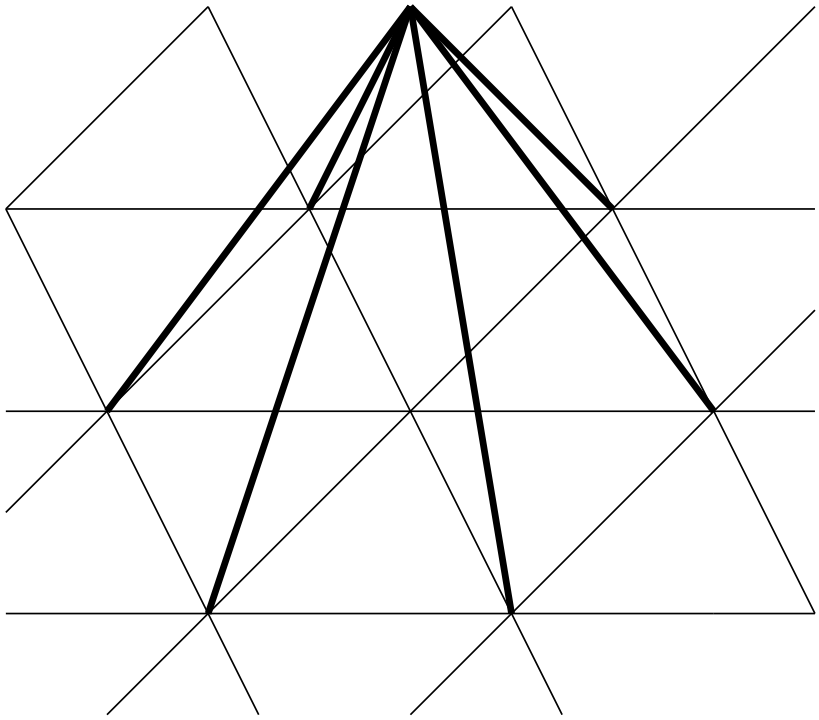


Abbildung 3.1: Lineare Knotenbasisfunktion.

nennen die Koeffizienten in den inneren Knoten die *Freiheitsgrade* des diskreten Problems. Die Anzahl der Freiheitsgrade entspricht der Dimension des linearen Gleichungssystems.

Dieser Ansatz erfüllt eine wesentliche Forderung an ein Galerkin-Verfahren. Für die Steifigkeitsmatrix $\boldsymbol{A}_h \in \mathbb{R}^{N_h \times N_h}$ gilt:

$$\boldsymbol{A}_{ij} = a(\phi_h^{(j)}, \phi_h^{(i)}) \neq 0 \quad \text{nur wenn ein } T \in \Omega_h \text{ existiert mit } x_i \in \bar{T} \text{ und } x_j \in \bar{T}.$$

Die Steifigkeitsmatrix ist also dünn besetzt.

Example 3.11 (Lineare Finite Elemente auf gleichmäßigem Dreiecksgitter)**.** *Es sei* $\Omega = (0,1)^2$ *und* Ω_h *ein gleichmäßiges Dreiecksgitter: Das Gebiet wird in Quadrate der größe* $h \times h$ *mit* $h = 1/m$ *geteilt. Jedes dieser Quadrate wird in zwei Dreiecke durch eine diagonale Linie von links unten nach recht oben geteilt.*

Das Gitter Ω_h *hat mit*

$$N_D = \frac{2}{m^2}$$

Dreiecke und

$$N = \frac{1}{(m+1)^2}$$

Punkte.

Die Finite Elemente Basis wird lokal erstellt. Wir betrachten hierzu exemplarisch das Beispiel

$$T_0 = \{(x, y) : \ 0 < x + y < 1 \text{ und } 0 \leqslant x, y \leqslant 1\}.$$

Und in den drei Knoten

$$X_1 = (0, 0), \quad X_2 = (h, 0), \quad X_3 = (h, h)$$

definieren wir die drei linearen Knotenbasisfunktionen

$$\hat{\phi}_1(x, y) = 1 - \frac{x}{h}, \quad \hat{\phi}_2(x, y) = \frac{x - y}{h}, \quad \hat{\phi}_3(x, y) = \frac{y}{h}.$$

Man überprüfe die Eigenschaft $\hat{\phi}_i(X_j) = \delta_{ij}$.

In allen anderen Dreiecken ergeben sich die lokalen Basisfunktionen durch Translation der $\hat{\phi}_i$ *und gegebenenfalls durch eine Spiegelung oder Drehung. Die* N *globalen Basisfunktionen setzen sich stückweise aus den* $\hat{\phi}_i$ *auf den jeweiligen Dreiecken zusammen.*

Wir betrachten nun die Matrix der Laplace-Diskretisierung, d.h.

$$a(u, \phi) = (\nabla u, \nabla \phi)_\Omega,$$

also

$$\mathbf{A}_{ij} = \int_0^1 \int_0^1 \nabla \phi_j(x, y) \cdot \nabla \phi_i(x, y) \, dx \, dy.$$

Wir schreiben dieses Integral als Summe über alle Dreiecke

$$\mathbf{A}_{ij} = \sum_{T \in \Omega_h} \underbrace{\int_T \nabla \phi_j(x, y) \cdot \nabla \phi_i(x, y) \, dxy}_{a_{ij}^T}.$$

Das Aufstellen der Systemmatrix *wird* Assemblieren *genannt. Im ersten Schritt werden stets lokale Integrale auf den einzelnen Elementen der Triangulierung berechnet (oder approximiert), in einem zweiten Schritt werden die einzelnen Einträge zur globalem Matrix zusammengefasst.*

Wir kümmern uns nun um die lokalen Beiträge a_{ij}^T. *Für die drei lokalen Basisfunktionen* $\hat{\phi}_i$ *gilt auf* T_0

$$a_{ij}^T = \int_0^h \int_0^x \nabla \hat{\phi}_j(x, y) \cdot \nabla \hat{\phi}_i(x, y) \, dy \, dx, \quad i, j = 1, 2, 3.$$

Nun gilt

$$\nabla \hat{\phi}_1 = \frac{1}{h} \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad \nabla \hat{\phi}_2 = \frac{1}{h} \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \nabla \hat{\phi}_3 = \frac{1}{h} \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Und da wir nur über konstante Funktionen integrieren folgt mit $|T| = h^2/2$

$$\begin{aligned} a_{11}^T &= \frac{1}{2} & a_{12}^T &= -\frac{1}{2} & a_{13}^T &= 0 \\ a_{21}^T &= -\frac{1}{2} & a_{22}^T &= 1 & a_{23}^T &= -\frac{1}{2} \\ a_{31}^T &= 0 & a_{32}^T &= -\frac{1}{2} & a_{33}^T &= \frac{1}{2} \end{aligned}$$

Auf diesem einfach gestalteten Gitter können wir die globalen Einträge der Matrix leicht zusammenfassen. Hierzu wählen wir eine lexikographische Sortierung der Knoten, d.h.

$$x_i = x_{kl}, \quad i = (1+m)k + l.$$

Dann folgt

$$\begin{aligned} \mathbf{A}_{ii} &= \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + 1 + 1 = 4, \\ \mathbf{A}_{i,i\pm1} &= -\frac{1}{2} - \frac{1}{2} = -1 \\ \mathbf{A}_{i,i\pm m} &= -\frac{1}{2} - \frac{1}{2} = -1 \end{aligned}$$

und ansonsten $\mathbf{A}_{ij} = 0$. Die Matrix hat damit die Blockform

$$\mathbf{A} = \begin{pmatrix} \mathbf{B} & -I & 0 & \cdots & 0 \\ -I & \mathbf{B} & -I & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -I & \mathbf{B} & -I \\ 0 & \cdots & 0 & -I & \mathbf{B} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 4 & -1 & 0 & \cdots & 0 \\ -1 & 4 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 4 & -1 \\ 0 & \cdots & 0 & -1 & 4 \end{pmatrix}.$$

Die rechte Seite ergibt sich durch entsprechende Integration

$$\mathbf{b}_i = \int_\Omega f(x,y)\phi_i(x,y)\,dx\,dy = \sum_{T\in\Omega_h} \underbrace{\int_T f(x,y)\phi_i(x,y)\,dx\,dy}_{=:\mathbf{b}_i^T}.$$

Für die rechte Seite $f \equiv 1$ erhalten wir exemplarisch

$$\begin{aligned} \mathbf{b}_1^T &= \int_0^h \int_0^x 1 - h^{-1}x\,dy\,dx = \frac{1}{6}h^2 \\ \mathbf{b}_2^T &= \int_0^h \int_0^x h^{-1}(x-y)\,dy\,dx = \frac{1}{6}h^2 \\ \mathbf{b}_3^T &= \int_0^h \int_0^x h^{-1}x\,dy\,dx = \frac{1}{6}h^2 \end{aligned}$$

Der globale Vektor hat damit im Innern (d.h. für Punkte $x_i \in \Omega_h$ die nicht auf dem Rand liegen) die Einträge

$$\mathbf{b}_i = h^2.$$

Zusammengefasst ergibt sich bei der Diskretisierung der Laplace-Gleichung mit linearen Finiten Elementen auf einem gleichmäßigen Dreiecksgitter ein Problem, welches äquivalent (gleich bis auf Skalierung mit h^2) zur 5-Punkte Stern Diskretisierung mit finiten Differenzen ist.

Für das allgemeine Galerkin-Verfahren haben wir bereits das Lemma von Cea 3.6 bewiesen. Wir betrachten das Poisson-Problem. Für die Finite-Elemente Lösung $u_h \in V_h^{(1)}$ gilt die Bestapproximation:

$$\|\nabla(u - u_h)\| = \min_{\phi\in V_h^{(1)}} \|\nabla(u - \phi_h)\|.$$

Konvergenz für $h \to 0$ liegt vor, wenn es ein $\phi_h \in V_h^{(1)}$ gibt, so dass $\|\nabla(u - \phi_h)\| \to 0$ gilt. Als Wahl für eine Approximation verwenden wir die *Interpolation* $I_h u \in V_h^{(1)}$:

$$I_h : V \to V_h, \quad I_h u(x) = \sum_{i=1}^{N_h} u(x_i)\phi_h^{(i)}(x),$$

für welche wir die folgenden *Interpolationsabschätzungen* kennenlernen werden:

$$\|u - I_h u\|_T \leqslant c_i h_T^2 \|\nabla^2 u\|_T, \quad \|\nabla(u - I_h u)\|_T \leqslant c_i h_T \|\nabla^2 u\|_T. \tag{3.3}$$

Hieraus folgt dann unmittelbar mit der Bestapproximationseigenschaft die *a priori* Fehlerabschätzung für lineare Finite Elemente

$$\|\nabla(u - u_h)\|_\Omega \leqslant \|\nabla(u - I_h u)\|_\Omega \leqslant c_i h \|\nabla^2 u\|_\Omega.$$

3.2.2 Allgemeine Finite Elemente Räume

Wir betrachten ein Dreiecks- oder Viereckgitter Ω_h des Gebiets Ω. Finite Elemente Räume werden üblicherweise als stückweise polynomial eingeführt. Hierzu betrachten wir auf einem Dreieck $T \in \Omega_h$ oder Viereck $K \in \Omega_h$ die Polynomräume

$$P^r := \text{span}\{x^{\alpha_1} y^{\alpha_2},\ 0 \leqslant \alpha_1 + \alpha_2 \leqslant r\}, \quad Q^r := \text{span}\{x^{\alpha_1} y^{\alpha_2},\ 0 \leqslant \alpha_1, \alpha_2 \leqslant r\},$$

der stückweise linearen, quadratischen u.s.w., bzw. bilinearen, biquadratischen u.s.w. Funktionen. Der Raum P^r ist einfach für Tetraeder, der Raum Q^r auf Hexaeder zu erweitern. Der Raum der linearen Finite Elemente auf einem Dreiecksgitter, bzw. der bilinearen Finiten Elemente auf einem Viereckgitter ist definiert als:

$$V_h^{(1)} := \{\phi_h \in C(\bar{\Omega}),\ \phi_h\big|_T \in P^1, T \in \Omega_h\}, \quad V_h^{(1)} := \{\phi_h \in C(\bar{\Omega}),\ \phi_h\big|_K \in Q^1,\ K \in \Omega_h\}.$$

Allgemeine Finite Elemente Räume werden über die Angabe von *Knotenwerten* definiert. Dabei wird der Wert der diskreten Funktion ϕ_h oder Ableitungen der diskreten Funktion $\nabla^s \phi_h$ in Eckpunkten, inneren Punkten oder Punkten auf Kanten der Elemente fixiert. Weiter ist es möglich (und manchmal auch zweckdienlich) die diskrete Funktion über gewisse Mittelwerte, etwa über eine Kante zu bestimmen. Die *Knotenwerte* sind Punkte auf Ecken, Kanten oder im Innern. In diesen Knotenwerten werden sogenannte *Knotenfunktionale* $P(T)$, also Punktwerte, Ableitungswerte oder Mittelwerte vorgeschrieben.

Ein Finite Elemente Ansatzraum wird durch die Vorgabe eines Polynomraums $P(T) \subset P^r$ oder $P(K) \subset Q^r$ sowie durch einen Satz von Knotenwerten und Knotenfunktionalen beschrieben, welche unter Hinzunahme von globalen Eigenschaften (z.B. Stetigkeit der Knotenfunktionale über Elemente hinweg) eine eindeutige Zuordnung erlauben.

Definition 3.12 (Unisolvenz)**.** *Ein Polynomraum* $P(T)$ *und ein zugehöriger Satz von linearen Knotenfunktionalen* $K(T)$ *heißt* unisolvent*, wenn jedes* $p \in P(T)$ *eindeutig durch die Vorgabe von* $\chi \in K(T)$ *bestimmt ist.*

Jedes mögliche Polynom $p \in P(T)$ muss sich also eindeutig durch die Vorgabe der Knotenfunktionale beschreiben lassen. Eine notwendige Bedingung ist natürlich, dass die Dimension des lokalen Polynomraums $P(T)$ mit der Anzahl der Knotenfunktionale übereinstimmt. Eine hinreichende Bedingung ist aufgrund der Linearität der Knotenfunktionale, dass für ein $p \in P(T)$ aus $\chi(p) = 0$ für alle Knotenfunktionale notwendig $p = 0$ folgt.

Falls wir zum Beispiel den Raum der linearen Finite Elemente auf Drei- oder Vierecken betrachten müssen wir auf jedem Element das Polynom eindeutig beschreiben. Ein Dreieck hat 3 Eckpunkte, somit muss folgerichtig $\dim(P^1) = 3$ gelten, in einem Viereck mit 4 Eckpunkten benötigen wir den Raum der bilinearen Funktionen mit $\dim(Q^1) = 4$.

Ist ein Satz von unisolventen Knotenfunktionalen gegeben, kann die *lokale Interpolation* einer Funktion $v \in H^m(\Omega)$ in diesen Polynomraum definiert werden:

Definition 3.13 (Finite Elemente Interpolation)**.** *Für jede Zelle $T \in \Omega_h$ sei ein Polynomraum $P(T)$ und ein Satz von unisolventen Knotenfunktionalen $K(T)$ mit Dimension R gegeben:*

$$\chi_r : H^m(T) \to \mathbb{R} \quad (r = 1, \dots, R).$$

Durch die Vorgabe:

$$\chi_r(I_h v) = \chi_r(v), \quad r = 1, \dots, R$$

ist dann eindeutig die Finite Elemente Interpolation $I_h v \in P(T)$ *beschrieben.*

Remark 3.14. *Bei der genauen Definition von Interpolationsoperatoren wird die Regularität der Räume eine wesentliche Rolle spielen. Angenommen $u \in H^1(\Omega)$, dann ist u nicht unbedingt stetig. Die Zuordnung eines Punktwertes:*

$$\chi(u) = u(a), \quad a \in \Omega,$$

ist kein lineares Funktional in $H^1(\Omega) \to \mathbb{R}$ und kann somit nicht zur Definition einer Interpolation genutzt werden. Oft liefert die Existenz- und Regulartiätstheorie aber gerade $u \in H^1(\Omega)$ und keine höhere Regularität. Für diese Zwecke werden wir einen speziellen Interpolationsoperator betrachten, die Clement-Interpolation*, welche auch für Funktionen $u \in H^1(\Omega)$ definiert ist.*

Der übliche Finite Elemente Ansatz ist der:

Definition 3.15 (Lagrange-Ansatz)**.** *Im Fall, dass alle Knotenfunktionale auf der Vorgabe von Punktwerten beruhen spricht man von einem* Lagrange-Ansatz*.*

Und dem gegenüber:

Definition 3.16 (Hermite-Ansatz)**.** *Im Fall, dass als Knotenfunktionale auch die Vorgabe von Ableitungswerten auftauchen, spricht man von einem* Hermite-Ansatz*.*

Beispiele von Finite Elemente Räumen

Wir betrachten einige Beispiele von Finite Elemente Räumen. Besonderes Interesse haben wir an H^1 konformen Finite Elemente Räumen, da hier die Galerkin-Orthogonalität sowie das Lemma von Cea gilt. Wir haben bereits mehrfach argumentiert, dass H^1-Funktionen nicht in jedem Punkt definiert sein müssen, insbesondere müssen diese Funktionen also auch nicht stetig sein. Es zeigt sich aber, dass in zwei und drei Dimensionen, eine H^1-Funktion nicht entlang eines ganzen Linienstückes, also z.B. einer Elementkante unstetig sein darf. Unter H^1-Konformität stellen wir uns vereinfacht Stetigkeit vor.

Dreieckselemente

a) stückweise konstante Funktionen $P(T) = \mathbf{span}\{1\}$ Der Ansatz niedrigster Ordnung verwendet den lokalen Polynomraum $P^0(T) = \text{span}\{1\}$, also die konstanten Funktionen. Ein mögliches Knotenfunktional in jedem Element ist die Vorgabe des Mittelwerts:

$$p \in P: \quad (p,1)_T = (v,1)_T,$$

Der Mittelwert stellt ein lineares Funktional dar, auch ist dieser Ansatz unisolvent.

Ein globaler, nicht-konformer Finite-Elemente Raum ist definiert durch die Vorgabe:

$$V_h^{0,nc} := \{\phi_h \in L^2(\Omega),\ \phi_h|_T \in \text{span}\{1\}\}.$$

Dieser Raum spielt bei sogenannten *dual-gemischten* Formulierungen eine Rolle. Zur direkten Approximation von elliptischen partiellen Differentialgleichungen eignet er sich nicht, da alle Ableitungen verschwinden, also $\nabla\phi_h = 0$ für alle $\phi_h \in V_h^{0,nc}$.

Ein H^1-konformer Finite Elemente Raum lässt sich mit den stückweise konstanten Funktionen nicht erstellen. Durch die Forderung nach globaler Stetigkeit bleiben nur die global konstanten Funktionen übrig.

b) stückweise lineare Polynome $P(T) = \mathbf{span}\{1, x, y\}$ Dieser Raum hat die lokale Dimension 3 und als Knotenfunktionale geben wir den Wert in den Eckpunkten vor. Dieser Ansatz ist unisolvent. Denn eingeschränkt auf eine Kante (also ein Linienstück) ist jedes $p \in P$ eine lineare Funktion. Ist diese lineare Funktion Null in zwei Punkten, so ist sie komplett Null, also ist auch die Richtungsableitung in Richtung der Kante Null. Da zwei Kanten sich jeweils schneiden gilt in den Eckpunkten $\nabla p = 0$. Also p (als lineare Funktion) konstant und somit $p = 0$.

Ein globaler H_0^1-konformer Ansatz ist gegeben durch:

$$V_h^{(1)} := \{\phi_h : \bar{\Omega} \to \mathbb{R} : \phi_h|_T \in \text{span}\{1, x, y\},\ \phi_h \text{ stetig in Eckpunkten } a_i \in T,$$
$$\phi_h = 0 \text{ in Eckpunkten } a_i \in \partial\Omega\}.$$

Die globale Stetigkeit $\phi_h \in C(\bar{\Omega})$ erhält man, da ϕ_h auf jeder Kante durch die Vorgabe beider Eckpunkte eindeutig bestimmt ist. Aus der Stetigkeit in den Eckpunkten folgt die Stetigkeit entlang der ganzen Kante. Stetige, stückweise polynomiale Funktionen sind stets H^1-konform (Übungsaufgabe).

Einen zunächst ungewöhnlichen, in der Anwendung (Navier-Stokes-Gleichungen) jedoch verbreiteten nicht-konformen Ansatz erhält man durch die Vorgabe der Knotenwerte in den drei Kantenmitten:

$$V_h^{(1),nc} := \{\phi_h : \bar{\Omega} \to \mathbb{R} : \phi_h\big|_T \in \text{span}\{1, x, y\},\ \phi_h \text{ stetig in Kantenmitten},\\ \phi_h = 0 \text{ in Kantenmitten auf } \partial\Omega\}.$$

c) stückweise quadratische Polynome $P(T) = \textbf{span}\{1, x, y, x^2, y^2, xy\}$ Die lokale Dimension des Raumes ist 6. Der einfachste Raum verwendet als Knotenwertvorgabe die drei Eckpunkte sowie drei Kantenmitten der Elemente:

$$V_h^{(2)} := \{\phi_h : \bar{\Omega} \to \mathbb{R} : \phi_h\big|_T \in \text{span}\{1, x, y\},\ \phi_h \text{ stetig in Eckpunkten und Kantenmitten},\\ \phi_h = 0 \text{ in Eckpunkten und Kantenmitten auf } \partial\Omega\}.$$

Dieser Raum ist global stetig. Denn entlang jeder Kante ist $\phi_h \in V_h^{(2)}$ durch Vorgabe von drei Funktionswerten eindeutig bestimmt. Diese Funktionswerte liegen alle auf einer Kante und sind stetig.

Ähnlich kann bei der Untersuchung der Unisolvenz argumentiert werden. Angenommen $\chi_r(p) = 0$ für $r = 1, \ldots, 6$. D.h., das Polynom ist entlang jeder der drei Kanten konstant Null. Hieraus folgt, dass der Gradient des Polynoms in den Eckpunkten und somit entlang der ganzen Kante Null ist. p ist also konstant und somit ganz Null.

Statt die Knotenwerte in den Kantenmitten vorzugeben, kann jeweils der Mittelwert über eine Kante vorgeschrieben werden. Dieser Ansatz ist auch unisolvent:

$$p \in P: \quad (p, 1)_e = (v, 1)_e \quad e \in \partial T.$$

Ein nicht konformes quadratisches Element ist das *Morley Platten-Element*. Neben der Vorgabe der Funktionswerte in den Eckpunkten werden die Normalableitungen in den Kantenmitten vorgegeben:

$$V_h^{Morley} := \{\phi_h : \bar{\Omega} \to \mathbb{R} : \phi_h\big|_T \in \text{span}\{1, x, y, x^2, xy, y^2\},\ \phi_h \text{ stetig in Eckpunkten},\\ \partial_n \phi_h \text{ stetig in Kantenmitten}\}.$$

Dieser Ansatz ist nicht konform, denn nur die Normalableitung, nicht aber die Funktionen selbst müssen stetig sein in den Kantenmitten. Es kann jedoch gezeigt werden, dass dieser Ansatz unisolvent ist.

d) stückweise quintisches *Argyris*-Plattenelement Als letztes Beispiel betrachten wir das *Argyris-Element*. Lokal wird der Raum der quintischen Polynome verwendet:

$$P(T) = \text{span}\{x^\alpha y^\beta,\ 0 \leqslant \alpha + \beta \leqslant 5\},$$

mit $\dim(P(T)) = 21$. Als Knotenfunktionale werden in den Eckpunkten die Funktionswerte, der Gradient, sowie die zweiten Ableitungen vorgegeben. Dies sind $3 \cdot (1+2+3) = 18$ (nur 3 zweite Ableitungen wegen $\partial_x \partial_y \phi_h = \partial_y \partial_x \phi_h$) Werte. Die fehlenden drei Werte werden durch Vorgabe der Normalableitung in den Kantenmitten vorgegeben. Dieser Ansatz ist unisolvent und darüber hinaus H^2-konform, d.h., es gilt $V_h^{Argyris} \subset H^2(\Omega)$.

Viereckselemente

Die meisten Ansätze lassen sich auf Vierecke bzw. Hexaeder verallgemeinern. Es ist üblicherweise notwendig mehr lokale Freiheitsgrade hinzuzunehmen. Daher sind als Grundlage Polynomräume der Art $Q^1 = \text{span}\{1, x, y, xy\}$ mit den gemischten Termen xy üblich.

stückweise bi-Polynome $P(T) = Q^r$ Das Viereck K wird durch ein uniformes $(r+1) \times (r+1)$-Gitter überdeckt, wobei die Eckpunkte enthalten sind. In jedem dieser $(r+1)^2$ Punkte wird der Funktionswert fixiert. Dieser Ansatz ist unisolvent und $H^1(\Omega)$-konform, da auf jeder Kante $r+1$ Freiheitsgrade zur Verfügung stehen, um die Funktion hier eindeutig vorzugeben.

Ein nicht konformer bilinearer Ansatz könnte durch die Vorgabe der Funktionswerte in den Kantenmitten erzeugt werden. Dieser Ansatz ist jedoch nicht unisolvent, denn die Funktion

$$\phi_h = xy \in Q^1,$$

ist ungleich Null, eingebettet in das Viereck $K = (-1,1) \times (-1,1)$ haben jedoch alle Knotenfunktionale den Wert 0. Einen nicht konformen, jedoch unisolventen Ansatz erhält man durch die Wahl der lokalen Räume

$$Q^{1,rot} := \text{span}\{1, x, y, x^2 - y^2\},$$

welcher durch *Rotation* des Q^1 um $\pi/4$ erzeugt wird. Dieser Ansatz ist bei Vorgabe der Funktionswerte in den Kantenmitten unisolvent.

Hexaederelemente

Das Lagrange-Viereckselement lässt sich unmittelbar in beliebige Dimensionen übertragen, da dieser Ansatzraum in einem Tensorprodukt-Sinne aufgebaut ist. Allgemein hat der Raum

$$P(K) = Q^r := \text{span}\{\prod_{j=1}^{d} x_j^{\alpha_j},\ 0 \leqslant \alpha_1, \ldots, \alpha_d \leqslant r\},$$

die Dimension $\dim(Q^r) = (r+1)^d$. Über die Vorgabe von Knotenfunktionalen in einem uniformen inneren Punktgitter ist ein unisolventes Element definiert.

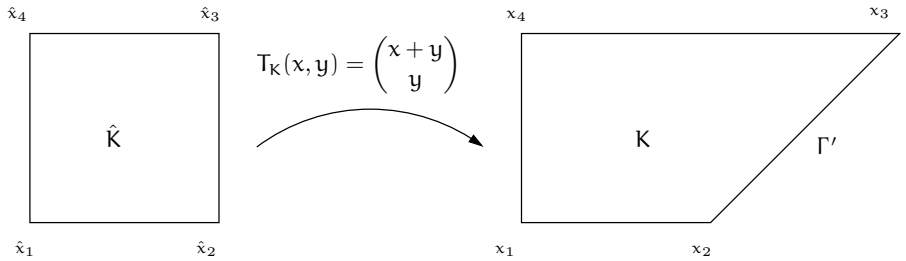


Abbildung 3.2: Allgemeines Viereck, welches nicht durch affin-lineare Transformation erzeugt werden kann.

3.2.3 Parametrische Finite Elemente

Die Definition von Finite Elemente Räumen über die Angabe von lokalen Knotenfunktionalen ist oft schwierig. Z.B. liefert der lokale Polynomraum $P(K) := \text{span}\{1, x, y, xy\}$ mit der Vorgabe der Funktionswerte in den Eckpunkten einen unisolventen Ansatz auf dem Viereck $K = (0,1) \times (0,1)$, jedoch nicht auf dem gedrehten Viereck $K := \{(x,y) \in \mathbb{R}^2,\ |x| + |y| < 1\}$. Denn hier ist die Funktion $p = xy$ in jedem Eckpunkt Null. D.h., alle Knotenfunktionale $\chi_i(p) = 0$ verschwinden, für das Polynom selbst gilt allerdings $p \neq 0$. In einem allgemeinen Gitter ist es jedoch nicht möglich diese Fälle auszuschließen.

Das Problem ist noch größer, wenn allgemeine Vierecke betrachtet werden, also auch solche, die keine Rechtecke sind, wie in Abbildung 3.2. Auf dem Linienstück Γ' mit Parametrisierung $y = x = t$ gilt für eine bilineare Funktion $\phi_h \in Q^1 := \text{span}\{1, x, y, xy\}$:

$$\phi_h(x,y)\Big|_{\Gamma'} = \text{span}\{1, t, t^2\},$$

diese Funktion ist entlang dieser Linie quadratisch und nicht linear. Durch die Vorgabe der zwei Eckpunkte ist diese Funktion nicht eindeutig festgelegt. Alle vier Basisfunktionen beeinflussen den Wert entlang dieser Kante. Der globale Ansatz ist nicht stetig, wenn nur die Stetigkeit in den Knoten vorgeschrieben wird.

Ein sinnvoller bilinear Ansatz muss aus Basisfunktionen bestehen, die entlang jeder Kante linear sind. Dieses erreichen wir durch Transformation:

1. In einem ersten Schritt wird der Ansatz auf einem *Referenzelement*, etwa dem Einheitsviereck $\hat{T} := (0,1) \times (0,1)$ definiert. Dies ist einfach, da nur dieses eine Element betrachtet werden muss.

2. Auf diesem Referenzelement wird eine Referenzbasis $\{\hat{\phi}^1, \ldots, \hat{\phi}^r\}$ definiert. Die Dimension dieser Basis entspricht der Anzahl der Knotenfunktionale (z.B. 4 bei bilinearen Finiten Elementen auf dem Viereck).

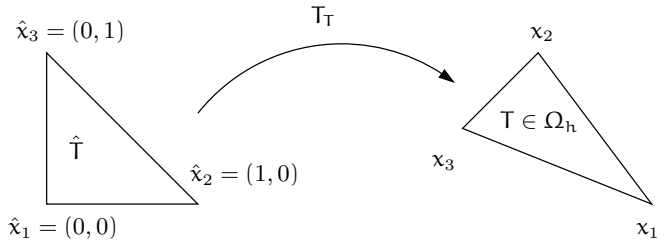


Abbildung 3.3: Referenzdreieck und Transformation auf beliebiges $T \in \Omega_h$

3. Um einen Finite Elemente Ansatz auf einer Zelle $T \in \Omega_h$ des Gitters zu definieren wird eine Abbildung $T_T : \hat{T} \to T$ definiert und mit Hilfe dieser Abbildung werden Basisfunktionen transformiert:

$$x \in T: \quad \phi_h^{(i)}(x)\Big|_T = \hat{\phi}^{(i)}(T_T^{-1}(x)). \tag{3.4}$$

Einen solchen Finite Elemente Ansatz nennt man *parametrisch*. Falls die Transformation T_T im gleichen Polynomraum ist, wie die Basis auf $\hat{T}$, also $\hat{\phi}^i \in P(T)$ und $T_T \in P(T)$, so spricht man von einem *iso-parametrischen Ansatz*.

Example 3.17 (Isoparametrische lineare Dreieckselemente)**.** *Es sei* $\hat{T} = \{(x,y) : \ 0 \leqslant x + y \leqslant 1\}$ *das Referenzdreieck wie in Abbildung 3.3. In den drei Knoten definieren wir die linearen Basisfunktionen*

$$\hat{\phi}^{(1)}(\hat{x},\hat{y}) = 1 - \hat{x} - \hat{y}, \quad \hat{\phi}^{(2)}(\hat{x},\hat{y}) = \hat{x}, \quad \hat{\phi}^{(3)}(\hat{x},\hat{y}) = \hat{y}.$$

Jetzt sei $T \in \Omega_h$ *ein beliebiges Dreieck mit Eckpunkten* $x_1, x_2, x_3 \in \Omega$*. Die Transformation* $T_T : \hat{T} \to T$ *soll isoparametrisch sein, also auch aus dem Raum der linearen Funktionen kommen. Damit ist sie affin linear und setzt sich mit Hilfe der Basisfunktionen zusammen:*

$$T_T(\hat{x}) = x_1\hat{\phi}^{(1)}(\hat{x}) + x_2\hat{\phi}^{(2)}(\hat{x}) + x_3\hat{\phi}^{(3)}(\hat{x}).$$

dies sieht man einfach wegen:

$$T_T(\hat{x}_i) = \sum_{j=1}^{3} x_j \underbrace{\hat{\phi}^{(j)}(\hat{x}_i)}_{=\delta_{ij}} = x_i.$$

Wir können diese affin lineare Transformation auch schreiben als:

$$T_T(\hat{x}) = B_T\hat{x} + b_T,$$

mit einer Matrix $B_T \in \mathbb{R}^{2\times 2}$ *und einem Vektor* $b_T \in \mathbb{R}^2$*. Diese Matrix ist regulär, wenn die drei Punkte nicht auf einer Geraden liegen und es gilt* $\det(B_T) > 0$*, falls die drei Eckpunkte* x_i *in gleicher*

Reihenfolge (also hier gegen Uhrzeigersinn) wie die Referenzpunkte $\hat{x}_i$ nummeriert sind. Auf dem Dreieck T gelten nun die Basisfunktionen:

$$\phi^{(i)}(x) = \hat{\phi}^{(i)}(T_T^{-1}(x)) = \hat{\phi}^{(i)}(B_T^{-1}x - B_T^{-1}b_T),$$

mit der Eigenschaft:

$$\phi^{(i)}(x_j) = \hat{\phi}^{(i)}(T_T^{-1}(x_j)) = \hat{\phi}^{(i)}(\hat{x}_j) = \delta_{ij}.$$

Im Fall von linearen Dreieckselementen fallen die isoparametrischen Ansätzen mit der einfachen Dreiecksbasis auf jedem Element zusammen. Dies liegt an der einfachen Struktur am Ansatzraum: Die Transformation T_T sowie ihre Inverse T_T^{-1} sind affin linear. Auch die Aneinanderkettung $\phi^{(i)} \circ T_T^{-1}$ ist wieder eine lineare Funktion.

Example 3.18. *Isoparametrische bilineare Viereckselemente Wir betrachten wieder den einfachen, bi-linearen Viereckraum. Auf dem Referenzelement (links in Abbildung 3.2) wird der Raum der Bilinearen Funktionen $P(\hat{K}) = Q^1$ durch die Knotenbasis aufgespannt:*

$$\hat{\phi}^{(1)} = 1 - x - y + xy, \quad \hat{\phi}^{(2)} = x - xy, \quad \hat{\phi}^{(3)} = xy, \quad \hat{\phi}^{(4)} = y - xy.$$

Für diese Basisfunktionen gilt $\hat{\phi}^{(i)}(\hat{x}_j) = \delta_{ij}$. Entlang jeder Kante des Referenzelements sind die Basisfunktionen linear. Es sei nun $K \in \Omega_h$ ein beliebiges Viereck mit Eckpunkten $x_1, \dots, x_4 \in$. Die Transformation T_K soll isoparametrisch, also $T_K \in P(\hat{K})$. Wir können die Transformation wieder einfach mit Hilfe der Knotenbasis erklären:

$$T_K(\hat{x}) = \sum_{i=1}^{4} x_i \hat{\phi}^{(i)}(\hat{x}),$$

denn es gilt wieder:

$$T_K(\hat{x}_j) = \sum_{i=1}^{4} x_i \hat{\phi}^{(i)}(\hat{x}_j) = x_j.$$

Wir betrachten nun den speziellen Fall aus Abbildung 3.2 und die vier Eckpunkte in $K \in \Omega_h$ seien gegeben durch:

$$x_1 = (0,0),\ x_2 = (1,0),\ x_3 = (2,1),\ x_4 = (0,1).$$

Dann gilt für die Transformation:

$$T_K(\hat{x}) = (0,0)^T \hat{\phi}^1(\hat{x}) + (1,0)^T \hat{\phi}^2(\hat{x}) + (2,1)^T \hat{\phi}^3(\hat{x}) + (0,1)^T \hat{\phi}^4(\hat{x}) = \begin{pmatrix} \hat{x} + \hat{x}\hat{y} \\ \hat{y} \end{pmatrix}$$

Diese Transformation ist bilinear (in jeder Komponente) und liegt also im Polynomraum $P(\hat{K})$. Für ihre Inverse gilt allerdings:

$$T_K^{-1}(x) = \begin{pmatrix} \frac{x}{1+y} \\ y \end{pmatrix},$$

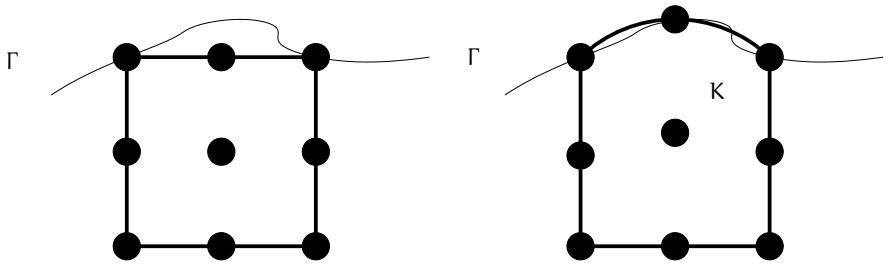


Abbildung 3.4: Iso-parametrische Finite Elemente höherer Ordnung zur besseren Rand-Approximation.

sie ist eine rationale Funktion und liegt nicht mehr im Ansatzraum. Mit (3.4) können wir die Basisfunktionen auf der "Rechenzelle" $K \in \Omega_h$ *aufstellen:*

$$\phi^{(1)}(x) = 1 - y + \frac{xy - x}{1+y}, \quad \phi^{(2)}(x) = \frac{x(1-y)}{1+y}, \quad \phi^{(3)}(x) = \frac{xy}{1+y}, \quad \phi^{(4)}(x) = \frac{y(1+y-x)}{1+y}.$$

Man mache sich klar, dass $\phi^{(i)}(x_j) = \delta_{ij}$ *gilt. Es handelt sich nicht mehr um einen Polynomansatz! Entlang jeder Kante, z.B.* $\Gamma := \overline{x_2 x_3} = \{(1+s, s),\ s \in (0,1)\}$ *sind die Basisfunktionen linear und hängen insbesondere nur noch von den beiden begrenzenden Eckpunkten ab:*

$$\phi^{(1)}|_\Gamma(s) = 0, \quad \phi^{(2)}|_\Gamma(s) = 1 - s, \quad \phi^{(3)}|_\Gamma(s) = s, \quad \phi^{(4)}|_\Gamma(s) = 0.$$

Die Verwendung von isoparametrischen Finiten Elementen ist der übliche Zugang für allgemeine Finite Elemente Ansätze. Die Konstruktion erscheint zunächst kompliziert, da die verwendeten Basisfunktionen keine Polynome mehr sind. Bei Betrachtungen zur Realisierung der Finite Elemente Methode werden wir jedoch feststellen, dass es zu keinem Zeitpunkt notwendig sein wird, die Funktionen $\phi^{(i)}$ auf der Zelle $K \in \Omega_h$ aufzustellen. Auswerten sowie Integration erfolgt stets durch Transformation auf das Referenzelement. Und hier liegen die klassischen Polynomansätze vor. Im folgenden Abschnitt zur Interpolation mit Finiten Elementen werden wir diese Argumentation bereits kennenlernen.

Isoparametrische Ansätze haben eine besondere Bedeutung bei Verwendung höherer Ansatzgrade. Angenommen wir betrachten biquadratische Ansätze. Dann stehen auch zur Definition der Transformation von $\hat{K}$ auf K biquadratische Funktionen zur Verfügung, für jedes Viereck also 9 Freiheitsgrade. Auf diese Weise können zusätzliche Freiheitsgrade auf dem Rand des Elementes verwendet werden. Wie in Abbildung 3.4 erlaubt dieses Vorgehen eine bessere Approximation von Gebieten mit "krummen Rändern". Das eigentliche Element $K \in \Omega_h$ ist somit gar kein Viereck mehr!

3.3 Interpolation mit Finiten Elemente

Wesentlich für das weitere Vorgehen ist die Herleitung von Abschätzungen für den Interpolationsfehler in Finite Elemente Räumen. Es sei also V_h ein Finite Elemente Raum. Wir

betrachten ausschließlich parametrische Finite Elemente. Dazu sei $\hat{T}$ das Referenzelement, also ein festes Dreieck oder Viereck sowie $T \in \Omega_h$ eine beliebiges Element des Gitters. Die Transformation sei $T_T : \hat{T} \to T$. Zunächst sei diese Transformation affin linear. D.h., für jedes Element $T \in \Omega_h$ existiert eine Matrix $B_T \in \mathbb{R}^{d\times d}$ sowie ein Vektor $b_T \in \mathbb{R}^d$ so dass gilt:

$$T_T(\hat{x}) = B_T x + b_T x.$$

Die Matrix B_T sei ferner invertierbar.

Lemma 3.19 (Referenz-Interpolation)**.** *Auf dem Referenzelement $\hat{T}$ sei durch $P(\hat{T})$ ein Polynomraum mit $\dim(P(\hat{T})) = R$ sowie ein Satz von linearen Knotenfunktionalen $K(\hat{T}) = \{\chi_1, \ldots, \chi_R\}$ mit den folgenden Eigenschaften gegeben:*

1. *Der Ansatz sei unisolvent*

$$\chi_i(p) = 0 \quad i = 1, \ldots, R \quad \Rightarrow \quad p = 0.$$

2. *Für ein $m \geqslant 1$ gilt $P^{m-1} \subset P(\hat{T})$, d.h., der Polynomraum enthält alle Polynome bis zum Grad $m-1$.*

3. *Die Knotenfunktionale sind stetig auf $H^m(\Omega)$:*

$$i = 1, \ldots, R: \quad \chi_i(v) \leqslant c\|v\|_{H^m(\Omega)} \quad \forall v \in H^m(\Omega).$$

Unter den Bedingungen 1. und 3. ist die Interpolationsaufgabe für jedes $v \in H^m(\hat{T})$ eindeutig lösbar, d.h., es existiert ein eindeutig bestimmtes $I_{\hat{T}} v \in P(\hat{T})$ mit

$$\chi_i(I_{\hat{T}} v) = \chi_i(v) \quad \forall i = 1, \ldots, R.$$

Die Interpolation hat die Darstellung

$$I_{\hat{T}} v = \sum_{i=1}^{R} \chi_i(v)\hat{\phi}^{(i)}, \tag{3.5}$$

mit der verallgemeinerten Lagrange-Basis $\{\hat{\phi}^{(i)}, i = 1, \ldots, R\}$, welche eindeutig durch die Bedingung $\chi_i(\hat{\phi}^{(j)}) = \delta_{ij}$ gegeben ist.

Proof: Wegen der Stetigkeit der Knotenfunktionale auf $H^m(\Omega)$ ist die Aufgabe wohlgestellt. Die eindeutige Lösbarkeit folgt unmittelbar aus der Unisolvenz. □

Bedingung 2., also die Reichhaltigkeit des Polynomraums werden wir zur Herleitung von Fehlerabschätzungen für die Interpolation benötigen.

3.3.1 Das Bramble-Hilbert-Lemma

Zunächst benötigen wir einige Hilfsätze:

Hilfsatz 3.20 (Nullraum von Ableitungsoperatoren)**.** *Jede Funktion* $v \in H^m(\hat{T})$ *mit der Eigenschaft:*

$$D^\alpha v = 0 \quad \forall |\alpha| = m,$$

ist fast überall ein Polynom $v \in P^{m-1}(\hat{T})$.

Proof: Wegen $D^\alpha v = 0$ für $|\alpha| = m$ folgt für beliebigen Multiindex β also $D^\beta D^\alpha = 0$, d.h., es gilt:

$$v \in \cap_{k=1}^{\infty} H^k(\hat{T}).$$

Und mit dem Einbettungssatz in Räume stetiger Funktionen **??** folgt $v \in C^\infty(\hat{T})$ und insbesondere $v \in C^m(\hat{T})$. Die Aussage folgt nun etwa durch m-maliges bilden der Stammfunktion von $D^\alpha v = 0$. □

Hilfsatz 3.21 (Polynomprojektion)**.** *Für jede Funktion* $v \in H^m(\hat{T})$ *existiert eine eindeutig bestimmte Projektion* $q \in P^{m-1}(\hat{T})$ *mit den Eigenschaften:*

$$\int_{\hat{T}} D^\alpha (v - q)\, dx = 0 \quad 0 \leqslant |\alpha| \leqslant m - 1.$$

Proof: Wir konstruieren $q \in P^{m-1}(\hat{T})$ gemäß dem Ansatz:

$$q(x) = \sum_{|\beta| \leqslant m-1} \xi_\beta x^\beta,$$

mit den Koeffizienten $\xi_\beta \in \mathbb{R}$. Es muss gelten:

$$\sum_{|\beta| \leqslant m-1} \xi_\beta \int_{\hat{T}} D^\alpha x^\beta \, dx = \int_{\hat{T}} D^\alpha v \, dx \quad \forall |\alpha| \leqslant m - 1.$$

Dieses Problem ist ein quadratisches lineares Gleichungssystem mit Matrix

$$A = (A_{\alpha\beta})_{|\alpha|,|\beta| \leqslant m-1}, \quad A_{\alpha\beta} := \int_{\hat{T}} D^\alpha x^\beta \, dx.$$

Diese Matrix ist regulär. Ansonsten würde es Koeffizienten $\xi := (\xi^\beta)_{|\beta| \leqslant m-1}$ geben mit $\xi \neq 0$ aber $A\xi = 0$. Das zugehörige Polynom $q(x) = \sum \xi^\beta x^\beta$ vom Grad $m - 1$ hätte die Eigenschaft $\int_{\hat{T}} D^\alpha q \, dx = 0$ für alle $|\alpha| \leqslant m-1$. Hieraus folgt im Widerspruch zur Annahme $q = 0$. □

Hilfsatz 3.22 (Verallgemeinerte Poincare-Ungleichung). *Für jede Funktion* $v \in H^m(\hat{T})$ *mit der Eigenschaft:*

$$\int_{\hat{T}} D^\alpha v \, dx = 0 \quad \forall |\alpha| \leqslant m-1, \tag{3.6}$$

gilt

$$\|v\|_{H^m(\hat{T})} \leqslant c_0 |v|_{H^m(\hat{T})},$$

mit einer Konstante c_0*, welche nicht von* v *abhängt.*

Proof: Angenommen, diese Ungleichung würde nicht gelten. D.h., zu jeder Konstante $c_v > 0$ existiert ein $v \in H^m(\hat{T})$ mit Eigenschaft (3.6), aber mit $\|v\|_{H^m} > c_v |v|_{H^m(\hat{T})}$. Dann existiert also auch eine Folge von Funktionen $v_n \in H^m(\hat{T})$ mit:

$$1 = \|v_n\|_{H^m(\hat{T})} \geqslant n |v_n|_{H^m(\hat{T})}, \quad n \in \mathbb{N}. \tag{3.7}$$

Die Eigenschaft $\|v_n\|_{H^m(\hat{T})} = 1$ folgt durch einfache Normierung. Die Einbettung von $H^m(\hat{T}) \hookrightarrow H^{m-1}(\hat{T})$ ist kompakt, siehe Satz 2.14. Also hat die beschränkte Folge v_n eine in $H^{m-1}(\hat{T})$ konvergente Teilfolge (welche wir wieder mit v_n bezeichnen):

$$\|v_n - v\|_{H^{m-1}(\hat{T})} \to 0. \tag{3.8}$$

Mit der Annahme (3.7) folgt:

$$|v_n|_{H^m(\hat{T})} \leqslant \frac{1}{n} \to 0 \quad (n \to \infty). \tag{3.9}$$

Zusammen mit (3.8) folgt, dass v_n in $H^m(\hat{T})$ Cauchy-Folge ist, denn:

$$\begin{aligned} \|v_k - v_l\|^2_{H^m(\Omega)} &= \|v_k - v_l\|^2_{H^{m-1}(\Omega)} + |v_k - v_l|^2_{H^m(\Omega)} \\ &\leqslant \|v_k - v\|^2_{H^{m-1}(\Omega)} + \|v_l - v\|^2_{H^{m-1}(\Omega)} + |v_k|^2_{H^m(\Omega)} + |v_l|^2_{H^m(\Omega)} \to 0. \end{aligned}$$

Wegen der Vollständigkeit existiert ein $v_n \to \tilde{v} \in H^m(\Omega)$ und wegen (3.8) gilt $v = \tilde{v}$ aus (3.8). Aus Hilfsatz 3.20 folgt jetzt $v \in P^{m-1}(\hat{T})$. Weiter gilt mit (3.8)

$$\int_{\hat{T}} D^\alpha v \, dx = \lim_{k\to\infty} \int_{\hat{T}} D^\alpha v_k \, dx \quad |\alpha| \leqslant m-1.$$

Hieraus folgt $v = 0$ im Widerspruch zur Annahme (3.7). □

Mit diesen Hilfsätzen kann nun der folgende Satz bewiesen werden:

Lemma 3.23 (Bramble-Hilbert-Lemma). *Es sei* $F(\cdot) : H^m(\hat{T}) \to \mathbb{R}$ *ein Funktional mit den folgenden Eigenschaften:*

1. *Beschränktheit*

$$|F(v)| \leqslant c_1 \|v\|_{H^m(\hat{T})} \quad \forall v \in H^m(\hat{T})$$

2. *Sublinearität*
$$|F(u+v)| \leqslant c_2(|F(u)| + |F(v)|) \quad \forall u, v \in H^m(\hat{T})$$

3. *Verschwindet auf* $P^{m-1}(\hat{T})$
$$F(q) = 0 \quad \forall q \in P^{m-1}(\hat{T})$$

Dann gilt mit der Konstante c_0 *aus der verallgemeinerten Poincaré-Ungleichung, Hilfsatz 3.22*

$$|F(v)| \leqslant c_0 c_1 c_2 |v|_{H^m(\hat{T})}.$$

Proof: Für ein $v \in H^m(\hat{T})$ gilt mit beliebigem Polynom $q \in P^{m-1}(\hat{T})$

$$|F(v)| = |F(v-q+q)| \leqslant c_2\big(|F(v-q)| + |F(q)|\big) \leqslant c_1 c_2 \|v-q\|_{H^m(\hat{T})}.$$

Das Polynom $q \in P^{m-1}(\hat{T})$ wird nun gemäß Hilfsatz 3.21 gewählt, dann folgt mit der verallgemeinerten Poincaré-Ungleichung

$$|F(v)| \leqslant c_0 c_1 c_2 |v-q|_{H^m(\hat{T})} = c_0 c_1 c_2 |v|_{H^m(\hat{T})}.$$

□

Eine einfache Anwendung des Bramble-Hilbert-Lemmas ist die Herleitung einer Interpolationsabschätzung auf dem Referenzelement $\hat{T}$, denn der Interpolationsoperator I_h erfüllt gerade die Anforderungen an das Funktional aus dem Bramble-Hilbert-Lemma:

Lemma 3.24 (Allgemeiner Interpolationssatz)**.** *Es sei* $I_{\hat{T}}$ *ein Interpolationsoperator mit den Eigenschaften von Satz 3.19. Dann gilt für jede Funktion* $v \in H^m(\hat{T})$ *die Abschätzung:*

$$|v - I_{\hat{T}} v| \leqslant c|v|_{H^m(\hat{T})},$$

bzgl. einer beliebigen, auf $H^m(\hat{T})$ *stetigen Halbnorm* $|\cdot|$.

Proof: Ohne Einschränkung sei $|v| \leqslant \|v\|_{H^m(\hat{T})}$ für alle $v \in H^m(\hat{T})$.

Wir betrachten das Funktional:
$$F(v) := |v - I_{\hat{T}} v|.$$

Dieses Funktional erfüllt die Eigenschaften des Bramble-Hilbert-Lemmas. Denn mit Darstellung (3.5) gilt

$$|F(v)| \leqslant |v| + |I_{\hat{T}} v| \leqslant |v| + \sum_{i=1}^{R} |\chi_i(v)|\,|\hat{\phi}^{(i)}| \leqslant (1 + Rc \max_{i=1,\dots,R} |\hat{\phi}^{(i)}|)\|v\|_{H^m(\hat{T})},$$

falls alle Knotenfunktionale χ_i auf H^m beschränkt sind. Dies muss von Fall zu Fall untersucht werden. Etwa die Vorgabe von Ableitungswerten in Eckpunkten benötigt sehr hohe

Regularität, Beschränktheit gilt also nur in Räumen $H^m(\hat{T})$ mit großem m. Für die Interpolation gilt $I_{\hat{T}} q = q$ für alle Polynome $q \in P^{m-1}(\hat{T})$. Also gilt $F(q) = 0$ für alle $q \in P^{m-1}(\hat{T})$. Somit ergibt das Bramble-Hilbert-Lemma die gewünschte Abschätzung. □

Dieses abstrakte Resultat ist sehr allgemein und kann nun auf verschiedene Normen konkretisiert werden. Die Halbnorm $|\cdot|$ muss lediglich stetig auf $H^m(\hat{T})$ sein. Wir betrachten einige Beispiele:

1. L^2-Fehler. Es gilt:
$$|v - I_{\hat{T}} v| := \|v - I_{\hat{T}} v\|_{L^2(\hat{T})} \leqslant \|v - I_{\hat{T}} v\|_{H^m(\hat{T})} \quad m \in \mathbb{N}.$$

2. $H^k(\hat{T})$-Halbnorm. Wie oben gilt:
$$|v - I_{\hat{T}} v| := |v - I_{\hat{T}} v|_{H^k(\hat{T})} \leqslant \|v - I_{\hat{T}} v\|_{H^m(\hat{T})} \quad 0 \leqslant k \leqslant m.$$

3. Maximum-Fehler. In zwei und drei räumlichen Dimensionen gilt wegen der Einbettung von Sobolew-Räumen in die Räume stetiger Funktionen:
$$|v - I_{\hat{T}} v| := \max_{x \in \hat{T}} |v - I_{\hat{T}} v| \leqslant c\|v - I_{\hat{T}} v\|_{H^2(\hat{T})}.$$

4. Auf jeder Kante $\Gamma \subset \hat{T}$ gilt mit dem Spursatz:
$$|v - I_{\hat{T}} v| := \|v - I_{\hat{T}} v\|_{\Gamma} \leqslant c\|v - I_{\hat{T}} v\|_{H^1(\hat{T})}.$$

5. Der Fehler kann anstelle von Normen auch in Mittelwerten gemessen werden, etwa:
$$|v - I_{\hat{T}} v| := \left| \int_{\hat{T}} (v - I_{\hat{T}} v)\, dx \right| \leqslant c\|v - I_{\hat{T}} v\|_{L^1(\hat{T})}.$$

Auf dem festen Referenzelement $\hat{T}$ gilt also eine Interpolationsabschätzung in sehr beliebigen Halbnormen. Die eigentliche Aufgabe ist es nun, eine Interpolationsabschätzung auf den Elementen $T \in \Omega_h$ der Triangulierung herzuleiten. Jedes Element wird durch eine Transformation $T_T : \hat{T} \to T$ erzeugt. Die Interpolation $I_T v$ auf dem Element $T \in \Omega_h$ ist durch einen parametrischen Ansatz gebildet, es gilt nicht $I_T v \in P(\hat{T})$, sondern $I_{\hat{T}} \hat{v} \in P(\hat{T})$ mit:
$$I_{\hat{T}} \hat{v}(\hat{x}) = I_T v(x), \quad x = T_T(\hat{x}) = B_T \hat{x} + b_T.$$

Hilfsatz 3.25 (Eigenschaften der Transformation)**.** *Es sei* $T_T : \hat{T} \to T$ *eine affin lineare Transformation* $T_T(\hat{x}) = B_T \hat{x} + b_T$ *mit einer regulären Matrix* B_T *mit* $\det(B_T) > 0$ *sowie einem Vektor* b_T*. Für eine Funktion* $\hat{f} \in W^{1,\infty}(\hat{T})$ *und* $f \in W^{1,\infty}(T)$ *mit* $\hat{f}(\hat{x}) = f(x)$ *gilt:*

$$\int_T f(x)\, dx = |\det B_T| \int_{\hat{T}} \hat{f}(\hat{x})\, d\hat{x}, \quad \int_{\hat{T}} \hat{f}(\hat{x})\, d\hat{x} = |\det B_T^{-1}| \int_T f(x)\, dx, \tag{3.10}$$

$$\partial_i f(x) = \sum_{j=1}^{d} b_{ji}^{(-1)} \hat{\partial}_j \hat{f}(\hat{x}), \quad \hat{\partial}_i f(\hat{x}) = \sum_{j=1}^{d} b_{ji} \partial_j f(x), \tag{3.11}$$

$$|f|_{H^k(T)} \leqslant c|\det B_T|^{\frac{1}{2}} \|B_T^{-1}\|^k |\hat{f}|_{H^k(\hat{T})}, \quad |\hat{f}|_{H^k(\hat{T})} \leqslant c|\det B_T^{-1}|^{\frac{1}{2}} \|B_T\|^k |f|_{H^k(T)}, \tag{3.12}$$

mit $B = (b_{ij})_{ij=1}^{d}$ *und* $B_T^{-1} = (b_{ij}^{(-1)})_{ij=1}^{d}$ *und einer Konstante* c*, welche nur von der Dimension des Gebiets* d *abhängt.*

Proof: (i) Zunächst gilt mit dem Transformationssatz unmittelbar:

$$\int_T f(x)\,dx = \int_{T_T(\hat{T})} f(x)\,dx = \int_{\hat{T}} |\det(\nabla T_T)| \hat{f}(\hat{x})\,d\hat{x} = |\det B_T| \int_{\hat{T}} \hat{f}(\hat{x})\,d\hat{x}. \tag{3.13}$$

Entsprechend wird die Rückrichtung bewiesen.

(ii) Für die Ableitung gilt:

$$\partial_i f(x) = \partial_i \hat{f}(\hat{x}) = \sum_{j=1}^{d} \hat{\partial}_j \hat{f}(\hat{x}) \frac{\partial \hat{x}_j}{\partial_i}.$$

Es ist $\hat{x} = T_K^{-1}(x) = B_T^{-1}x - B_T^{-1}b_T$ also gilt weiter:

$$\partial_i f(x) = \sum_{j=1}^{d} \hat{\partial}_j \hat{f}(\hat{x}) b_{ji}^{(-1)}. \tag{3.14}$$

Auch hier folgt die Rückrichtung entsprechend.

(iii) Aus 3.14 leiten wir eine Abschätzung her:

$$|\partial_i f| \leqslant \|B_T^{-1}\|_\infty \max_{j=1,\dots,d} |\hat{\partial}_j \hat{f}|,$$

wobei $\|\cdot\|_\infty$ eine für eine beliebige Matrix-Norm stehen kann. Nun sei α ein beliebiger Multiindex. Mehrfache Anwendung liefert:

$$|D^\alpha f| \leqslant \|B_T^{-1}\|_\infty^{|\alpha|} \max_{|\beta|=|\alpha|} |\hat{D}^\beta \hat{f}|.$$

Quadrieren und integrieren liefert zusammen mit 3.13:

$$|f|^2_{H^k(T)} = \int_T \sum_{|\alpha|=k} |D^\alpha f|^2\,dx \leqslant |\det B_T|\, \|B_T^{-1}\|_\infty^{2k} \int_T \sum_{|\alpha|=k} \max_{|\beta|=|\alpha|} |\hat{D}^\beta \hat{f}|^2\,d\hat{x}.$$

Wurzelziehen liefert das Ergebnis. Die Rückrichtung transformiert sich entsprechend. □.

Hilfsatz 3.26 (Eigenschaften der Transformation)**.** *Es sei* $T_T : \hat{T} \to T$ *mit* $T_T(\hat{x}) = B_T\hat{x} + b_T$ *eine affin lineare Transformation. Es seien* $\hat{\rho}$ *und* $\hat{h}$ *Inkreisradius und Durchmesser von* $\hat{T}$*, sowie* ρ_T *und* h_T *Inkreisradius und Durchmesser von* T*. Dann gilt:*

$$\|B_T\| \leqslant c\frac{h_T}{\hat{\rho}}, \quad \|B_T^{-1}\| \leqslant c\frac{\hat{h}}{\rho_T}.$$

Proof: Übungsaufgabe. □

Mit diesen Vorbereitungen kann der folgende, für uns wichtige Interpolationssatz bewiesen werden:

Lemma 3.27 (Spezielle Interpolation)**.** *Es sei* $T \in \Omega_h$ *ein Dreieck mit Inkreisradius* ρ_T *und Durchmesser* h_T*. Für jedes* $v \in H^m(T)$ *und die zugehörige Interpolation* $I_T v \in P(T)$ *in den Raum der parametrischen Finite Elemente von Ordnung* $m-1$ *gilt:*

$$|v - I_T v|_{H^k(T)} \leqslant c_I \frac{h_T^m}{\rho_T^k} |v|_{H^m(T)} \quad 0 \leqslant k \leqslant m.$$

Proof: (i) Auf dem Dreieck T ist die Interpolation konstruiert gemäß:

$$I_{\hat{T}} \hat{v}(\hat{x}) = I_T v(T_T(\hat{x})),$$

wobei $\hat{v}(\hat{x}) = v(T_T(\hat{x})) = v(x)$. Auf dem Referenzdreieck gilt mit Satz 3.24

$$|\hat{v} - I_{\hat{T}} \hat{v}|_{H^k(\hat{T})} \leqslant c |v|_{H^m(\hat{T})},$$

denn $|\cdot|_{H^k(\hat{T})}$ ist auf $H^m(\hat{T})$ mit $k \leqslant m$ eine stetige Halbnorm.

(ii) Wir transformieren den Interpolationsfehler mit Hilfe von (3.12) auf das Referenzdreieck:

$$|v - I_T v|_{H^k(T)} \leqslant |\det B_T|^{\frac{1}{2}} \, \|B_T^{-1}\|^k \, |\hat{v} - I_{\hat{T}} \hat{v}|_{H^k(\hat{T})}. \tag{3.15}$$

(iii) Auf der Referenzzelle wenden wir nun den allgemeinen Interpolationssatz 3.24 an und transformieren entsprechend mit 3.12 zurück:

$$|\hat{v} - I_{\hat{T}} \hat{v}|_{H^k(\hat{T})} \leqslant c_I |\hat{v}|_{H^m(\hat{T})} \leqslant c_I |\det B_T^{-1}|^{\frac{1}{2}} \, \|B_T\|_\infty^m \, |v|_{H^m(T)}.$$

Zusammen mit (3.15) erhalten wir:

$$|v - I_T v|_{H^k(T)} \leqslant c_I |\det B_T|^{\frac{1}{2}} \, |\det B_T^{-1}|^{\frac{1}{2}} \, \|B_T^{-1}\|_\infty^k \, \|B_T\|_\infty^m \, |v|_{H^m(T)}.$$

Mit Hilfsatz 3.26 und $|\det B_T^{-1}| = |\det B_T|^{-1}$ folgt die Abschätzung. □

Die Interpolation ist eine lokale Eigenschaft auf jedem Element $T \in \Omega_h$. Dennoch lässt sich für eine größen- und formreguläre Triangulierung eine globale Interpolationsabschätzung herleiten:

Lemma 3.28 (Globale Interpolationsabschätzung)**.** *Es sei* Ω_h *eine form-, größen- und strukturreguläre Triangulierung mit* $h := \max_{T \in \Omega_h} h_T$*. Dann gilt für jedes* $v \in H^m(\Omega)$ *und die zugehörige Interpolation* $I_h v$ *in den Raum der parametrischen Finiten Elemente von Ordnung* $m-1$ *die Abschätzung:*

$$|v - I_h v|_{H^k(\Omega)} \leqslant c_I h^{m-k} |v|_{H^m(\Omega)} \quad 0 \leqslant k \leqslant m.$$

Proof: Übungsaufgabe. □

Dieser Satz liefert uns die wichtigen Interpolationsabschätzungen für lineare Finite Elemente:

$$\|\nabla(u - I_h u)\| \leqslant c_I h \|\nabla^2 u\|, \quad \|u - I_h u\| \leqslant c_I h^2 \|\nabla^2 u\| \quad \forall u \in H^2(\Omega).$$

Zusammen mit dem Lemma von Cea, Satz 3.6 und der Regularitätsabschätzung der Lösung, Satz 2.30 folgt unmittelbar eine Fehlerabschätzung für den Fehler der Galerkin-Approximation der Poisson-Gleichung in der Energie-Norm:

$$\|\nabla(u - u_h)\| \leqslant \frac{M}{\gamma} \|\nabla(u - I_h u)\| \leqslant c_I h \frac{M}{\gamma} \|\nabla^2 u\| \leqslant c_I c_s h \frac{M}{\gamma} \|f\|.$$

Abschließend beweisen wir noch eine für die Analyse von Finite Elemente Verfahren wichtige Ungleichung:

Lemma 3.29 (Inverse Beziehung)**.** *Es sei $v_h \in V_h$ eine parametrische Finite Elemente Funktion vom Grad $m - 1$. Auf jeder Zelle des Gitters gilt die Beziehung:*

$$|v_h|_{H^k(T)} \leqslant c \frac{h_T^s}{\rho_T^k} |v|_{H^s(T)} \quad 0 \leqslant s \leqslant k \leqslant m - 1.$$

Proof: Es sei $\hat{q} \in P(\hat{T})$ ein Polynom vom maximalen Grad $m-1$ auf der Referenzzelle. Dieser Polynomraum ist endlich-dimensional, d.h., alle Normen sind auf diesem Raum äquivalent. D.h., die Ungleichung gilt auf dem Referenzelement

$$|\hat{q}|_{H^k(\hat{T})} \leqslant \hat{c} |\hat{q}|_{H^s(\hat{T})} \quad s \leqslant k \leqslant m - 1.$$

Die Aussage folgt nun durch Transformation auf $T \to \hat{T} \to T$. □

3.3.2 Die Clement-Interpolation

Der natürliche Raum zur Analyse von elliptischen Differentialgleichungen ist der Sobolew-Raum $H_0^1(\Omega)$. Der Makel der Knoteninterpolation ist die Verwendung von Punktwerten zur Definition des Interpolationsoperators mittels $I_h v(a_i) = v(a_i)$. Diese Knotenfunktionale sind auf $H_0^1(\Omega)$ nicht stetig definiert. Oft benötigen wir jedoch Interpolationen von Funktionen mit dieser minimalen Regularität. Die Clement-Interpolation ist ein Interpolationsoperator welcher anstelle von Funktionsauswertungen lokale Mittelwerte verwendet. Diese Mittelwerte sind als Funktionale auf dem H^1 beschränkt. Wir definieren:

Definition 3.30 (Patch)**.** *Sei Ω_h ein strukturreguläres Gitter. Für jeden Knoten $a \in \Omega_h$, jedes Element $T \in \Omega_h$ und jede Kante $E \in \Omega_h$ definieren wir den* Knotenpatch $P_a \in \Omega_h$*, die* Zellpatche

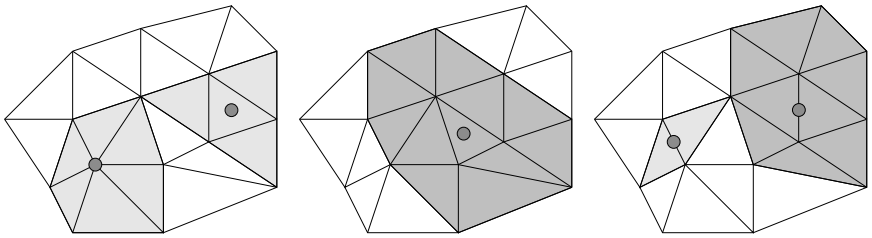


Abbildung 3.5: Definition der Patche. Links: Knotenpatch und "kleine Zellpatch" P_T. Mitte: "großer Zellpatch" $\tilde{P}_T$. Rechts: "kleiner Kantenpatch" P_E (hell) und "großer Kantenpatch" $\tilde{P}_E$ (dunkel).

$P_T \in \Omega_h$ *und* $\tilde{P}_T \in \Omega_h$ *sowie die* Kantenpatche $P_E \in \Omega_h$ *und* $\tilde{P}_E \in \Omega_h$

$$
\begin{aligned}
P_a &:= \bigcup_{T \in \Omega_h,\, x_i \in \bar{T}} T \\
P_T &:= \bigcup_{T' \in \Omega_h,\, \exists E \in \Omega_h,\, E = \bar{T} \cap \bar{T}'} T', \\
\tilde{P}_T &:= \bigcup_{P_a \in \Omega_h,\, a \in \bar{T}} P_a,
\end{aligned}
$$

$$
\begin{aligned}
P_E &:= \bigcup_{T' \in \Omega_h,\, E \subset T'} T', \\
\tilde{P}_E &:= \bigcup_{P_a \in \Omega_h,\, a \in \bar{E}} P_a
\end{aligned}
$$

In Abbildung (3.5) zeigen wir ein Beispiel solcher Patche. Zur Definition des Clement-Interpolationsoperators $C_h : V \to V_h$ werden nun anstelle von Punktwerten $v(a)$ Mittelwerte über die Knotenpatche P_a verwendet. Diese Funktionale sind auf dem $V = H_0^1(\Omega)$ beschränkt.

Lemma 3.31 (Clement-Interpolation)**.** *Es sei Ω_h ein form- größen- und strukturreguläres Gitter. Dann gibt es einen stetigen linearen Operator $C_h : V \to V_h$ in den Raum der linearen Finiten Elemente mit den folgenden Eigenschaften:*

$$
\|v - C_h v\|_{L^2(T)} \leqslant c h_T \|\nabla v\|_{L^2(\tilde{P}_T)}, \quad \|v - C_h v\|_{L^2(E)} \leqslant c h_E^{\frac{1}{2}} \|\nabla v\|_{L^2(\tilde{P}_E)} \quad \forall v \in V = H_0^1(\Omega).
$$

Proof: Wir definieren zunächst die Knotenfunktionale der Clement-Interpolation. Für jeden Knoten $x_i \in \Omega_h$ sei:

$$
\chi_i : L^2(P_{x_i}) \to \mathbb{R}, \quad \chi_i(v) := \begin{cases} \frac{1}{|P_{x_i}|} \int_{P_{x_i}} v(x)\,dx & x_i \notin \partial\Omega \\ 0 & x_i \in \partial\Omega \end{cases}
$$

Diese Knotenfunktionale sind linear, auf $L^2(\Omega)$ und also auch auf $V = H_0^1(\Omega)$ beschränkt. Nun sei $\hat{P}_x$ ein Referenzpatch. Hier gilt:

$$|\hat{\chi}_i(\hat{v})| = \frac{1}{|\hat{P}_x|}\int_{\hat{P}_x} \hat{v}\,d\hat{x} \leqslant c\|\hat{v}\|_{\hat{P}_x}.$$

Weiter sei $T_{x_i} : \hat{P}_x \to P_{x_i}$ mit $\det(\nabla T_x) = |P_{x_i}| = O(h^2)$ und $\|\nabla T_{x_i}\|_\infty = O(h_i)$. Dann gilt:

$$\|v - \chi_i(v)\|^2_{L^2(P_{x_i})} \leqslant h^2\|\hat{v} - \hat{\chi}_i(\hat{v})\|^2_{L^2(\hat{P}_x)}.$$

Weiter mit dem Bramble-Hilbert-Lemma und Rücktransformation:

$$\|v - \chi_i(v)\|^2_{L^2(P_{x_i})} \leqslant h^2 c_{bhl}\|\hat{\nabla}\hat{v}\|^2_{\hat{P}_x} \leqslant c_{bhl}h_i^2\|\nabla v\|^2_{L^2(P_{x_i})}.$$

(*ii*) Wir definieren die Interpolierende als:

$$C_h v(x) := \sum_{x_i \in \Omega_h} \chi_i(v)\phi_h^{(i)}(x) \quad \forall v \in V.$$

Für die Knotenbasis gilt

$$\sum_{x_i \in \bar{T}} \phi_h^{(i)}(x)\Big|_T \equiv 1,$$

da die Interpolation auch in den Randknoten definiert ist. Weiter folgt mit $\|\phi_h^{(i)}\|_{L^\infty(T)} \leqslant 1$:

$$\begin{aligned}
\|v - C_h v\|_T &= \left\| v\left(\sum_{x_i\in\bar{T}} \phi_h^{(i)}\right) - \sum_{x_i\in\bar{T}} \chi_i(v)\phi_h^{(i)} \right\|_T \\
&\leqslant \sum_{x_i\in\bar{T}} \|(v - \chi_i(v))\phi_h^{(i)}\|_T \leqslant \sum_{x_i\in\bar{T}} \|v - \chi_i(v)\|_{P_{x_i}} \leqslant \sum_{x_i\in\bar{T}} c_{bhl}h_i\|\nabla v\|_{P_{x_i}} \\
&\leqslant c_{bh}\sqrt{c(T)}\tilde{h}_T\|\nabla v\|_{\tilde{P}_T},
\end{aligned}$$

mit $\tilde{h}_T = \operatorname{diam}(\tilde{P}_T)$ und $c(T)$ der Anzahl der Zellen in einem Patch. Aus der Formregularität des Gitters folgt $c(T) \leqslant c_T$ gleichmäßig in $h \to 0$. Die Abschätzung des Interpolationsfehlers auf der Kante folgt entsprechend durch geeignete Transformation auf die Referenzkante. □

Die Clement-Interpolation wird H^1-stabil genannt. Es gilt:

Lemma 3.32 (H^1-Stabilität der Clement-Interpolation)**.** *Auf einem form-, größen und strukturregulärem Gitter Ω_h folgt für die Clement-Interpolation:*

$$\|\nabla C_h v\|_{L^2(T)} \leqslant c\|\nabla v\|_{\tilde{P}_T}.$$

Proof: Die Knotenfunktionale sind H^1-stabil. Deswegen kann auf gleichem Wege die H^1-Fehlerabschätzung hergeleitet werden:

$$\|\nabla(v - C_h v)\|_{L^2(T)} \leqslant c_I\|\nabla v\|_{\tilde{P}_T}.$$

Dann gilt:

$$\|\nabla C_h v\|_T \leqslant \|\nabla(v - C_h v)\|_T + \|\nabla v\|_T.$$

Hieraus folgt die Behauptung. □

3.4 A priori error analysis

We consider a general elliptic problem with homogenous Dirichlet boundary data

$$u \in \mathcal{V} := H_0^1(\Omega): \quad a(u,\phi) = (f,\phi) \quad \forall \phi \in \mathcal{V},$$

where $a(\cdot,\cdot)$ is an elliptic and continuous bilinear form.

The natural norm for elliptic problems is the gradient norm $\|\nabla \cdot\|_\Omega$, the norm of $H_0^1(\Omega)$. Using this norm we can use the Lamma of Cea, or in the case of the Laplace equation $a(u,\phi) = (\nabla u, \nabla \phi)$ the best approximation property. This will directly give us a the finite element error estimate in the energy norm.

Lemma 3.33 (A priori estimate in the energy norm). *Let $m \geqslant 1$. Let Ω be a domain with convex polygonal boundary for $m = 1$ or smooth boundary with C^{m+1}-parametrization, $f \in L^2(\Omega) \cap H^{m-1}(\Omega)$ and Ω_h a form-, shape- and size-regular domain. Let $u_h \in V_h^{(m)}$ be the finite element solution with parametric finite elements of degree m. Given $u \in H^{m+1}(\Omega) \cap H_0^1(\Omega)$ it holds*

$$\|\nabla(u - u_h)\| \leqslant \frac{c_I c_s}{\gamma} h^m \|f\|_{H^{m-1}(\Omega)}.$$

Proof: By Cea's lemma it holds

$$\|\nabla(u - u_h)\| \leqslant \frac{1}{\gamma} \|\nabla(u - \phi_h)\| \quad \forall \phi_h \in V_h^{(m)}.$$

We choose the interpolation $\phi_h := I_h^{(m)} u$. As $u \in H^{m+1}(\Omega)$ Lemma 3.28 gives

$$\|\nabla(u - u_h)\| \leqslant \frac{c_I}{\gamma} h^m \|\nabla^{m+1} u\|.$$

Using the stability estimate from Lemma 2.31 we get the estimate. □

3.4.1 A duality argument - the Aubin-Nitsche Trick

For the energy estimate we need the relation between norm $\|\nabla \cdot\|$ and the bilinear form given its ellipticity. This relation is typical for finite element discretizations of elliptic problems. Considering finite differences, the typical norm would be the maximum norm. Finite element a priori error estimates in other norms will require additional work and a way to replace the ellipticity estimate. The key idea will be the *Aubin-Nitsche-Trick*.

Lemma 3.34 (Adjoint problem). *Let $u \in \mathcal{V}$ be the variational solution of*

$$a(u,\phi) = l(\phi) \quad \forall \phi \in \mathcal{V},$$

where $l \in \mathcal{V}^*$. *For a linear functional* $J \in \mathcal{V}^*$ *we define the* adjoint problem

$$z \in \mathcal{V} \quad a(\phi, z) = J(\phi) \quad \forall \phi \in \mathcal{V}. \tag{3.16}$$

The adjoint solution $z \in \mathcal{V}$ *is uniquely determined and it holds*

$$\|z\|_{\mathcal{V}} \leqslant \frac{M}{\gamma} \|J\|_{\mathcal{V}^*},$$

where $M > 0$ *is the constant of ellipticity.*

Proof: We define the adjoint bilinear form

$$a^*(z, \phi) := a(\phi, z) \quad \forall z, \phi \in \mathcal{V}.$$

This bilinear form is elliptic and bounded. Lax-Milgram gives us a unique solution $z \in \mathcal{V}$ and also the stability estimate. □

The idea of the introducing the adjoint problem is to obtain new error measures. Assume, that $J : \mathcal{V} \to \mathbb{R}$ is the quantity, that we want to measure. Then, using the corresponding adjoint solution $z \in \mathcal{V}$ we get the error identity

$$J(u - u_h) = a^*(z, u - u_h) = a(u - u_h, z).$$

Galerkin-Orthogonalität and continuity of $a(\cdot, \cdot)$ then gives

$$J(u - u_h) = a(u - u_h, z - I_h z) \leqslant M \|\nabla(u - u_h)\| \, \|\nabla(z - I_h z)\| \quad \forall \phi_h \in V_h.$$

Using the adjoint problem, we can estimate the error in general error functionals J by the energy error $\|\nabla(u - u_h)\|$ and the approximation error of the adjoint solution $\|\nabla(z - I_h z)\|$. It remains to find an estimate for this adjoint interpolation error. We will discuss different adjoint problems.

Example 3.35 (Adjoint problem)**.** *We will consider different adjoint problems. We can express the* L^2*-error with help of an linear functional* (*even if the* L^2*-norm is no linear functional*)

$$J(\phi) := (u - u_h, \phi) \, \|u - u_h\|^{-1}.$$

This functional is linear in ϕ *and bounded in* $L^2(\Omega)$ (*and hence in* $H^1(\Omega)$)

$$J(\phi) \leqslant \|\phi\|_\Omega.$$

It holds

$$J(u - u_h) = \|u - u_h\|.$$

By Riesz, there exists a $j \in L^2(\Omega) \cong L^2(\Omega)^*$ *given by*

$$(j, \phi) = J(\phi) \quad \forall \phi \in L^2(\Omega), \quad \|j\|_{L^2(\Omega)} = \|J\|_{L^2(\Omega)^*} = 1.$$

As $j \in L^2(\Omega)$ *the adjoint solution on convex polygonal domains will satisfy* $z \in H^2(\Omega) \cap H_0^1(\Omega)$ *with*

$$\|z\|_{H^2(\Omega)} \leqslant c_s \|j\| = c_s.$$

Another example for a linear functional wouled be the average of the x-derivative

$$J(\phi) = \int_\Omega \partial_x \phi \, dx$$

$$\Rightarrow \quad |J(\phi)| \leqslant \int_\Omega |\partial_x \phi| \, dx \leqslant \int_\Omega |\nabla \phi| \, dx \leqslant c(\Omega) \|\nabla \phi\|_{L^2(\Omega)} \quad \forall \phi \in H_0^1(\Omega)$$

It holds $J \in H^{-1}(\Omega)$ *but* $J \notin L^2(\Omega)$*. Hence, we can only expect* $u \in H_0^1(\Omega)$*.*

The adjoint bilinear form is defined as $a^*(z, \phi) = a(\phi, z)$*. For symmetric problems - like Poisson - it holds*

$$a^*(z, \phi) := a(\phi, z) = (\nabla \phi, \nabla z) = (\nabla z, \nabla \phi) = a(z, \phi).$$

This problem is selfadjoint *with* $-\Delta z = j$*.*

If the elliptic problem includes a transport term, the adjoint problem is changed. We consider the general diffusion-transport problem

$$-\Delta u + \partial_x u = f \quad \Rightarrow \quad a(u, \phi) = (\nabla u, \nabla \phi) + (\partial_x u, \phi).$$

The adjoint variational formulation is given with integration by parts

$$a^*(z, \phi) := a(\phi, z) = (\nabla \phi, \nabla z) + (\partial_x \phi, z) = (\nabla z, \nabla \phi) - (\partial_x z, \phi) + \underbrace{\int_{\partial\Omega} n_x \cdot (\phi z) \, ds}_{=0}.$$

Here $a^*(z, \phi) \neq a(z, \phi)$ *and the direction of transport is reversed*

$$-\Delta z - \partial_x z = j.$$

The original idea of Aubin-Nitsche was the derivation of an a priori L^2-estimate.

Lemma 3.36 (A priori error estimate in the L^2-norm)**.** *Let the assumtions of Lemma 3.33 hold true. Let* $u \in V_h^{(m)}$ *be the finite element solution of polynomial degree* m*. Given* $u \in H^{m+1}(\Omega)$ *it holds*

$$\|u - u_h\| \leqslant \frac{M c_i^2 c_s^2}{\gamma} h^{m+1} \|f\|_{H^{m-1}(\Omega)}.$$

Proof: We introduce the adjoint problem

$$z \in H_0^1(\Omega): \quad a(\phi, z) = J(\phi), \quad J(\phi) := (e_h, \phi) \|e_h\|^{-1}.$$

It holds $e_h \in H_0^1(\Omega)$ and the adjoint right hand side J is a bounded linear functional in $L^2(\Omega) \to \mathbb{R}$ as

$$|J(\phi)| = |(\phi, e_h)| \, \|e_h\|^{-1} \leqslant \|\phi\| \quad \forall \phi \in L^2(\Omega).$$

By Riesz, there exists a $j \in L^2(\Omega)$ with

$$(j, \phi) = J(\phi) \quad \phi \in L^2(\Omega) \quad \Rightarrow \quad \|j\| = 1.$$

Using Lemma 2.30 and assuming sufficient regularity of the domai it holds $z \in H^2(\Omega)$ and

$$\|z\|_{H^2(\Omega)} \leqslant c_s \|j\|_{L^2(\Omega)} = c_s.$$

Galerkin-Orthogonality gives

$$\|e_h\| = J(e_h) = a(e_h, z) = a(e_h, z - I_h z) \leqslant M \|\nabla e_h\| \, \|\nabla(z - I_h z)\|. \tag{3.17}$$

The first part is the energy norm estimate from Lemma 3.33. For the second part we use the interpolation estimate and the stability of the adjoint problem to get

$$\|e_h\| \leqslant M \frac{c_i c_s}{\gamma} h^m \|f\|_{H^{m-1}(\Omega)} \, c_i h \|\nabla^2 z\| \leqslant \frac{M c_i^2 c_s^2}{\gamma} h^{m+1} \|f\|_{H^{m-1}(\Omega)}. \tag{3.18}$$

□

The adjoint problem is used as analytical tool only. We do not have to really compute a solution but can estiamte with the stability estimate. Later on we will discuss examples, where we cannot remove the adjoint solution but where we will need a numerical approximation of it.

Remark 3.37 (Optimality of the L^2-estimate)**.** *The L^2 estimates is better by one power in h. One could get the idea that more is to gain, if we use higher order polynomials and if the adjoint solution would allow for higher regularity. Assume quadratic polynomials are used and assume, that $z \in H^3(\Omega)$ holds. Then interpolation gives*

$$\|\nabla(z - I_h z)\| \leqslant c_I h^2 \|\nabla^3 z\|_\Omega \leqslant c_I c_s \|j\|_{H^1(\Omega)},$$

where

$$j := \frac{u - u_h}{\|u - u_h\|},$$

such that

$$\|\nabla(z - I_h z) \leqslant c_I c_s h^2 \frac{\|\nabla(u - u_h)\|}{\|u - u_h\|}.$$

The overall L^2-estimate gets

$$\|u - u_h\| \leqslant c_I c_s h^2 \frac{\|\nabla(u - u_h)\|^2}{\|u - u_h\|} \quad \Leftrightarrow \quad \|u - u_h\| \leqslant \sqrt{c_I c_s} h \|\nabla u - u_h\|$$

and once again get the result of an L^2-error which is one order better than the energy error.

For the L^2 error we get one additional order of convergence. What is the maximum possible convergence order if we go to even weaker error functionals?

Example 3.38 (Average error)**.** *We consider the Poisson problem with a given right hand side* f*. We assume that* $f \in C^\infty$ *is regular and we also assume, that the domain is sufficiently regular, such that* $u \in C_0^\infty(\Omega)$*.*

$$u \in H_0^1(\Omega) \quad (\nabla u, \nabla \phi) = (f, \phi) \quad \forall \phi \in H_0^1(\Omega).$$

Given a finite element approximation of degree m *we get the energy error estimate*

$$\|\nabla(u - u_h)\| \leqslant c_I c_s h^m \|f\|_{H^{m-1}(\Omega)}.$$

As error functional we consider the average of the solution on the entire domain

$$J(\phi) = \int_\Omega \phi \, dx.$$

The adjoint solution z *is given as*

$$z \in H_0^1(\Omega): \quad (\nabla z, \nabla \phi) = (\phi, 1) \quad \forall \phi \in H_0^1(\Omega).$$

The adjoint right hand side is the constant $j \equiv 1$*-function satisfying* $j \in C^\infty(\Omega)$*. Hence*

$$\|z\|_{H^{m+1}(\Omega)} \leqslant c_s \|1\|_{H^{m-1}(\Omega)} = c_s c(\Omega).$$

By inserting the interpolation and using the energy norm estimate we obtain

$$|J(e_h)| = |(\nabla e_h, \nabla z)| \leqslant \|\nabla e_h\| \, \|\nabla(z - I_h z)\| \leqslant c_I c_s h^m \|f\|_{H^{m-1}(\Omega)} \, c_I h^m \|\nabla^{m+1} z\|$$

such that the error estimate for the average gets

$$|J(e_h)| \leqslant c h^{2m} \|f\|_{H^{m-1}(\Omega)}.$$

We can double the convergence order.

One can argue, that h^{2m} is the optimal order if we consider linear functionals $J \in H^{-1}(\Omega)$. All these functionals are bound in the H^{-1}-norm. Let $u \in C^2(\Omega) \cap C(\bar{\Omega})$ be the classical solution with $f = -\Delta u$. Then, it holds

$$(u - u_h, f) = -(u - u_h, \Delta u) = \big(\nabla(u - u_h), \nabla u\big) = \big(\nabla(u - u_h), \nabla(u - u_h)\big).$$

Next, with the definition of the H^{-1}-norm

$$\|u - u_h\|_{-1} = \sup_{\phi \in H_0^1(\Omega)} \frac{\big(u - u_h, \phi\big)_\Omega}{\|\nabla \phi\|_\Omega} \geqslant \frac{\big(u - u_h, f\big)}{\|\nabla f\|} = \frac{\|\nabla(u - u_h)\|^2}{\|\nabla f\|}$$

which showsm that no error estimate (in a linear functional) can be better than twice the energy error.

L^∞ estimates L^∞ estimates take a special role. They are important for applications, as a bound in the L^∞ norm will prevent the error to be large in single points. Assuming the simulation of a technical problem, e.g. a mechanical load simulation of a bridge, where the L^2 norm of the error is small, the error in single points can still be very large. If such a point is the joint between different structural parts, this could result in a failure of the overall structure.

The L^∞ norm however is not natural in the world of finite elements and for variational formulations. Finite element application shows, that the convergence order in the L^∞ norm is usually the same as in the L^2-norm, $\mathcal{O}(h^2)$ in the case of linear finite elements. A proper analysis will show, that $\mathcal{O}(h^2)$ convergence is not fully reached, instead we optimal analysis reveals $\mathcal{O}(h^2 \log(h))$.

L^∞ estimates are called *pointwise error estimates* and proofs will aim at showing convergence in single points $a \in \Omega$.

Lemma 3.39 (Suboptimal L^∞-estimate)**.** *Assume $u \in H^2(\Omega) \cap C(\Omega)$. Then the pointwise error for the linear finite element approximation can be estimated (suboptimal) as*

$$\max_{x \in \bar{\Omega}} |e_h| \leqslant ch \|\nabla^2 u\|.$$

Proof: Let $a \in \bar{\Omega}$. Then there exists an element $T \in \Omega_h$ with $a \in \bar{T}$. For every discrete function $v_h \in V_h$ it holds

$$\max_T |v_h| \leqslant ch^{-1} \|v_h\|_T. \tag{3.19}$$

This can be shown by transformation to the reference element. Use of norm equivalence and transformation back to T gives

$$\max_T |v_h| = \max_{\hat{T}} |\hat{v}_h| \leqslant c \|\hat{v}_h\|_{\hat{T}} = c|T|^{-\frac{1}{2}} \|v_h\|_T$$

using $|T| = O(h^2)$ (in two dimensions) (3.19). Inserting the interpolation $I_h u$ and use of the L^∞-estimate $\max_T |u - I_h u| \leqslant ch \|\nabla^2 u\|$ gives

$$\begin{aligned} \max_T |e_h| &\leqslant \max_T |u - I_h u| + \max_T |I_h e_h| \\ &\leqslant ch \|\nabla^2 u\|_T + ch^{-1} \|I_h e_h\|. \end{aligned}$$

By the continuity of the nodal interpolation in $H^2(\Omega) \cap C(\bar{\Omega})$ and the L^2 error estimate we can show the estimate. □

In most application we will observe quadratic convergence for the linear finite element approximation in the L^∞ norm. In the general case it holds

Lemma 3.40 (Optimal L^∞-error estimate)**.** *Let $u \in H_0^1(\Omega) \cap C^2(\bar{\Omega})$. It holds*

$$\max_\Omega |e_h| \leqslant ch^2 \left\{ |\ln(h)| + 1 \right\} \max_\Omega |\nabla^2 u|.$$

Proof: The complete proof is given in [6]. We use a duality argument to estimate the error in $a \in \bar{\Omega}$. The proper right hand side

$$j(x) := \delta_a(x) = \begin{cases} \operatorname{sign}(e_h(a)) & x = a \\ 0 & x \neq 0, \end{cases}$$

is a signed Dirac function. The corresponding error functional $J(\phi)$ is not sufficiently regular for defining an H^1 adjoint solution. Instead we introduce a *regularized Dirac* on an element $a \in T_* \in \Omega_h$:

$$\delta_a^h(x) = \begin{cases} \frac{\operatorname{sign}(e_h)}{|T_*|} & x \in T_* \\ 0 & x \notin T_*. \end{cases}$$

This gives

$$(\nabla e_h, \nabla g^h) = \frac{1}{T_*} \int_{T_*} |e_h| \, dx.$$

For $|T^*| \to 0$ this expression is an approximation of the pointwise error. The technical difficulty of the optimal error estimate is to estimate the adjoint solution g^h. A function g with the property $(\nabla g, \nabla e_h) = e_h(a)$ is called *Green's function* at $a \in \Omega$. The adjoint solution g^h is a *regularized Green's function*. For details we refer to [6]. □

The logarithmic part $\ln(h) \to \infty$ in the estimate is no weakness of the proof. On specially constructed meshes it is possible to numerically observe this term. Using finite elements of degree $m \geqslant 2$, the optimal estimate holds without this logarithmic factor

$$\sup_\Omega |e_h| \leqslant c h^m \sup_\Omega |\nabla^m u|.$$

And if we consider the maximum norm of the gradient error, we also do not get this logarithm

$$\sup_\Omega |\nabla e_h| \leqslant c h \sup_\Omega |\nabla^2 u|.$$

3.4.2 Finite Elements on Curved domains

The standard finite element analysis is heavily depending on the conformity of the Galerkin approach $V_h \subset \mathcal{V}$ which is essential for getting Galerkin-Orthogonality. If the domain Ω is curved and cannot be matched by the finite element mesh $\Omega_h \neq \Omega$, the finite element space will not be conforming. In this section, we shortly discuss the approximation of the Laplace problem

$$u \in H_0^1(\Omega): \quad (\nabla u, \nabla \phi)_\Omega = (f, \phi)_\Omega \quad \forall \phi \in H_0^1(\Omega), \tag{3.20}$$

on a domain $\Omega \subset \mathbb{R}^d$ that is curved and smooth, i.e., the boundary $\partial\Omega$ locally allows for a C^{r+1}-parametrization, with $r \in \mathbb{N}_+$. Finite elements on curved domains must deal with two difficulties.

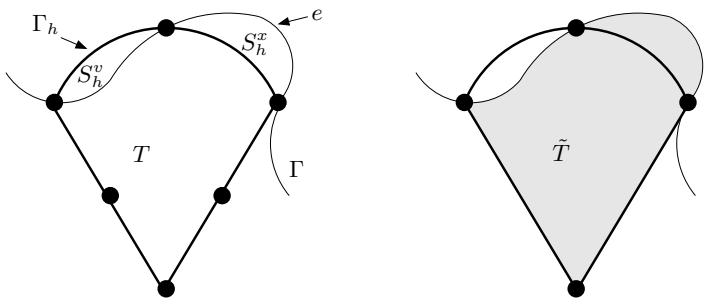


Abbildung 3.6: Left: geometric remainders for curved boundary approximation. Definition of the mesh snippets $S_h^v = \Omega_h \setminus \Omega$ and $S_h^x = \Omega \setminus \Omega_h$. Right: Definition of the curved extended element $\tilde{T}$ fitting the domain Ω. Exemplarily for quadratic iso-parametric elements.

1. A polygonal mesh will never exactly match the domain Ω. Hence, the discrete equation

$$u_h \in V_h : \quad (\nabla u_h, \nabla \phi_h)_{\Omega_h} = (\tilde{f}, \phi_h)_{\Omega_h} \quad \forall \phi_h \in V_h,$$

   is given on a different domain. The right hand side f must not even be defined on all of Ω_h, which is the case, if the domain Ω has concave boundary parts, where Ω_h might reach out. For this reason, we denoted a modified (extended) right hand side by $\tilde{f}$. For details, we refer to Remark 3.45.

2. The boundary conditions cannot be exactly satisfied. We consider homogenous Dirichlet conditions only. While $u \in H_0^1(\Omega)$ is zero on all of $\partial\Omega$, $u_h \in V_h$ is zero in the boundary nodes on $\partial\Omega$ but otherwise, it is zero on $\partial\Omega_h \neq \partial\Omega$.

Finite element analysis on curved domains is discussed in literature [2]. General proofs for isoparametric finite elements on curved domains, including optimal order a priori error bounds for the energy error are given in [5].

To cope with the two problems mentioned above, we will start by stating some definitions and lemma. Parts of the boundary can be convex or concave. We define the remainders by

$$S_h^x = \Omega \setminus \Omega_h, \quad S_h^v = \Omega_h \setminus \Omega, \quad S_h = S_h^x \cup S_h^v. \tag{3.21}$$

For a parametric triangulation Ω_h of Ω it holds

Lemma 3.41 (Isoparametric triangulation of Curved domains)**.** *Let $\Omega \subset \mathbb{R}^d$ be a domain with smooth boundary allowing for a C^{r+1}-parametrization with $r \geqslant 1$. Let Ω_h be an isoparametric mesh of Ω with polynomial degree r. For the area of the mesh snippets S_h^x, S_h^v, S_h it holds*

$$|S_h^x| = |S_h^v| = |S_h| = O(h^r).$$

Proof. This follows by simple geometrical arguments. Let $T \in \Omega_h$ be an element at the boundary and S be that part of S_h which is conntected to the element T, see Figure 3.6. Further, let $e \in \partial T$ be the (curved) edge at the boundary Γ_h, which is a $d-1$-dimensional manifold in $\mathbb{R}^d$ with area $|e| = O(h^{d-1})$. Assume, that $\psi : e \to \mathbb{R}$ is the parametrization of $\partial\Omega$ over e (see again Figure 3.6). $\psi(s)$ has $r+1$ zero's along the edge in 2d. Hence,

$$\max_{[0,h]} |\psi| \leqslant c h^{r+1} \max_{[0,h]} |\psi^{r+1}|.$$

Therefore, as $|e| = O(h^{d-1})$, it holds

$$|S| = O(h^{r+d}) \quad \Rightarrow \quad |S_h| = O(h^{r+1}),$$

as $\mathcal{O}(1/|e|) = \mathcal{O}(h^{-d+1})$. □

The previous lemma shows that standard finite elements will always suffer from a geometrical error. By the use of isogeometric analysis [4] this error could be completely avoided for domains, that can be described by splines.

Another technical difficulty is given by the mismatch of Ω and Ω_h. Functions $u \in H_0^1(\Omega)$ and $u_h \in V_h$ are defined on different domains, such that the meaning of the expression $u - u_h$ must be discussed. The following lemma will show a way to give $u_h \in V_h$ a meaning both on Ω_h and on Ω.

Lemma 3.42 (Boundary extension of discrete functions)**.** *Under the assumptions of Lemma 3.41, let $h \leqslant h_0 \in \mathbb{R}$ and $T \in \Omega_h$ be an element at the boundary $\partial\Omega$ with boundary edge $e \in \partial T$. By $\tilde{T}$ we denote the curved triangle fitting the domain's boundary, see Figure 3.6. For $u_h \in V_h$ we define by $\tilde{u}_h|_T$ the polynomial extension of $u_h|_T$ to $\tilde{T}$. It holds*

$$c_1 \|u_h\|_{H^s(T)} \leqslant \|\tilde{u}_h\|_{H^s(\tilde{T})} \leqslant c_2 \|u_h\|_{H^s(T)}, \ s = 0, 1, 2,$$

with two constants $c_1, c_2 > 0$ that do not depend on T or h.

Proof. This follows by considering equivalence of (discrete) norms and the negligible size of the remainders.

$$|T| = |\tilde{T}| = O(h^d), \quad |(T \setminus \tilde{T}) \cup (\tilde{T} \setminus T)| = O(h^{r+d}).$$

□

In the following, we will always use the notation u_h even on $\tilde{T}$.

While $u_h \in V_h$ is well-defined on Ω_h (including S_h^v) and can be extended to Ω includig S_h^x, functions $u \in H_0^1(\Omega)$ are only well-defined on Ω including S_h^x. An extension to the concave part S_h^v might fail due to limited regularity. For the analysis, we need one further - trace inequality-like - estimate:

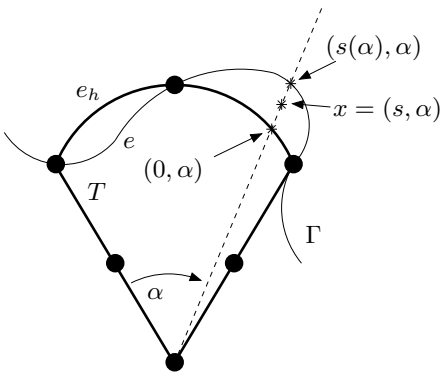


Abbildung 3.7: Local coordinate system on curved elements. Sketch for the proofs of Lemma 3.43 and 3.44. The boundary Γ with segment $e \subset \Gamma$ is given as parametrization of Γ_h with segments e_h, i.e. $s_T : e_h \to e$.

Lemma 3.43 (Geometric boundary error)**.** *Let $u \in H_0^1(\Omega)$. There exists a constant $c > 0$, such that for the convex remainder S_h^x, it holds*

$$\|u\|_{S_h^x} \leqslant c h^{\frac{r+1}{2}} \|u\|_{H^1(\Omega)}.$$

Further, let $u_h \in V_h$. It holds

$$\|u_h\|_{H^s(S_h)} \leqslant c h^{\frac{r}{2}} \|u_h\|_{H^s(\Omega)}, \ s = 0, 1.$$

Proof. For the proof, we refer to Figure 3.7. Let $T \in \Omega_h$ be an element on the boundary, $e_h \in \partial T$ be the edge of the element, $\tilde{T}$ the extended element and $e \in \tilde{T}$ be the edge at the boundary $\partial\Omega$. By S we denote the remainder between T and $\tilde{T}$.

(*i*) Let $x \in S$ be given as $x = (s, \alpha)$, where α is the angle and s the radial coordinate, see Figure 3.7. The local coordinate system is such, that $(0, \alpha) \in e_h \subset \Gamma_h$ is a point on the boundary of the (curved) triangle and $(s(\alpha), \alpha)$ is the corresponding point on the domain's boundary part $e \subset \Gamma$. It holds $|s(\alpha)| = O(h^{r+1})$, compare Lemma 3.41. Let $u \in C^1(\bar{S})$. It holds

$$u(s, \alpha) = u(0, \alpha) + \int_0^s \partial_r u(t, \alpha) \, dt,$$

and hence

$$|u(s, \alpha)|^2 \leqslant c \left(|u(0, \alpha)|^2 + |s| \int_0^s |\partial_r u(t, \alpha)|^2 \, dt \right).$$

Integration over S (in s and α) and noting that $|s| \leqslant |s(\alpha)| \leqslant c h^{r+1}$ gives

$$\|u\|_S^2 \leqslant c h^{r+1} \|u\|_e^2 + h^{2r+1} \|\nabla u\|_S^2. \tag{3.22}$$

(*ii*) To proof the first estimate, we continue with (3.22) by summing over all boundary elements, using trace inequality and Poincaré and extending S_h^x to Ω

$$\|u\|_{S_h} \leqslant c h^{\frac{r+1}{2}} \|\nabla u\|_\Omega.$$

(*iii*) For the second inequality, we apply the local trace inequality and extend from S to T

$$\|u_h\|_S^2 \leqslant ch^{r+1}\left(h^{-1}\|u_h\|_T^2 + h\|\nabla u_h\|_T^2\right) + h^{2r+1}\|\nabla u_h\|_T^2.$$

Using the inverse inequality, we get

$$\|u_h\|_S^2 \leqslant ch^r\|u_h\|_T^2,$$

such that the result follows by summing over all boundary snippets. This argumentation is also valid for ∇u_h. □

Discrete functions $\phi_h \in V_h$ are not zero on $\partial\Omega$ but zero on $\partial\Omega_h$.

Lemma 3.44 (Curved boundary error)**.** *Let $\phi_h \in V_h$ be arbitrary. It holds*

$$\|\phi_h\|_{\partial\Omega} \leqslant ch^{r+\frac{1}{2}}\|\nabla\phi_h\|_\Omega$$

Proof. We again refer to Figure 3.7. Let $T \in \Omega_h$ and $(s(\alpha), \alpha) \in e$ be a point on the boundary of $\partial\Omega$. By $(0, \alpha) \in e_h \subset \partial T$ we denote the corresponding point on the boundary of the triangle. It holds for $\phi_h \in V_h$

$$\phi_h(s(\alpha), \alpha) = \phi_h(0, \alpha) + \int_0^{s(\alpha)} \partial_r\phi_h(t, \alpha)\,dt,$$

and hence by squaring and integrating over α and by noting that $|s(\alpha)| = O(h^{r+1})$ we get

$$\|\phi_h\|_e^2 \leqslant \|\phi_h\|_{e_h}^2 + ch^{r+1}\|\nabla\phi_h\|_S^2. \tag{3.23}$$

With Lemma 3.43 and using $\phi_h = 0$ on e_h, gives

$$\|\phi_h\|_e^2 \leqslant ch^{2r+1}\|\nabla\phi_h\|_\Omega^2,$$

such that the result follows by summing over all boundary parts. □

Remark 3.45 (Extension of the right hand side at concave domain boundaries)**.** *As discussed in the beginning of this section, problems might already arise with the definition of the right hand side $f : \Omega \to \mathbb{R}$, which is not necessarily well-defined on the discrete domain Ω_h. This issue is easily handled by defining a projection or interpolation $f_h \in V_h$ to be used as discrete right hand side:*

$$(f_h, \phi_h)_\Omega = (f, \phi_h)_\Omega \quad \forall\phi_h \in V_h.$$

An additional error of type

$$(f - f_h, \phi)_\Omega \leqslant c\|f - f_h\|_{H^{-1}(\Omega)}\|\nabla\phi\|_\Omega,$$

will arise. By exploiting the weak norm and orthogonality of $f - f_h$ *such estimates can be given with optimal order and without requiring additional regularity of* $f \in H^{r-1}(\Omega)$*:*

$$\begin{aligned}\|f - f_h\|_{H^{-1}(\Omega)} &= \sup_{\phi \in H_0^1(\Omega)} \frac{(f - f_h, \phi)_\Omega}{\|\nabla \phi\|} \\ &= \sup_{\phi \in H_0^1(\Omega)} \frac{(f - f_h, \phi - \bar{\phi})_\Omega}{\|\nabla \phi\|} \\ \leqslant c h^r \|\nabla^{r-1} f\|_\Omega.\end{aligned}$$

To shorten the proof of the following lemma we will not give details on this issue and just consider f *as a well-defined right hand side function.*

With these preparations, we can show the following essential theorem, that gives the a priori error estimate for the Laplace equation on smooth and curved domains:

Theorem 3.46 (A priori error on curved domains)**.** *Let* $r \in \mathbb{N}_+$*. Let* Ω *be a domain with boundary that allows for parametrization of degree* $r+1$*. Let* $f \in H^{r-1}(\Omega) \cap L^2(\Omega)$*. Let* $u_h \in V_h$ *be the isoparametric finite element solution of degree* r*. It holds*

$$\|u - u_h\|_{H^1(\Omega)} \leqslant c h^r \|f\|_{H^{r-1}(\Omega)}$$

and

$$\|u - u_h\| \leqslant c h^{r+1} \|f\|_{H^{r-1}(\Omega)}.$$

Proof. (*i*) We start with the H^1 error estimate and derive a modified Galerkin orthogonality. For $\phi_h \in V_h$ it holds (where we use the extension $\tilde{\phi}_h \cong \phi_h$ defined by Lemma 3.42 without further notice)

$$(f, \phi_h)_\Omega = (-\Delta u, \phi_h)_\Omega = (\nabla u, \nabla \phi_h)_\Omega - \langle \partial_n u, \phi_h \rangle_{\partial\Omega}.$$

The discrete problem is defined on Ω_h with $\Omega_h = (\Omega \cup S_h^\nu) \setminus S_h^x$. It holds

$$(f, \phi_h)_\Omega + (f, \phi_h)_{S_h^\nu} - (f, \phi_h)_{S_h^x} = (\nabla u_h, \nabla \phi_h)_\Omega + (\nabla u_h, \nabla \phi_h)_{S_h^\nu} - (\nabla u_h, \nabla \phi_h)_{S_h^x}.$$

Then, for the finite element error $e_h = u - u_h$, we get the following disturbed Galerkin orthogonality:

$$\begin{aligned}(\nabla e_h, \nabla \phi_h)_\Omega = &-(f, \phi_h)_{S_h^\nu} + (f, \phi_h)_{S_h^x} \\ &+ \langle \partial_n u, \phi_h \rangle_{\partial\Omega} + (\nabla u_h, \nabla \phi_h)_{S_h^\nu} - (\nabla u_h, \nabla \phi_h)_{S_h^x}.\end{aligned} \tag{3.24}$$

(*ii*) Now, we can estimate the energy error by picking $\phi_h = I_h u - u_h$:

$$\begin{aligned}\|\nabla e_h\|_\Omega^2 \leqslant &\|\nabla e_h\|_\Omega \|\nabla (u - I_h u)\|_\Omega + \|f\|_S \|I_h u - u_h\|_S \\ &+ \|\nabla u_h\|_S \|\nabla (I_h u - u_h)\|_S + \|\partial_n u\|_{\partial\Omega} \|I_h u - u_h\|_{\partial\Omega},\end{aligned} \tag{3.25}$$

where we enlarged S_h^x and S_h^y to S. The single terms can be estimates with help of Lemma 3.43 and 3.44 and the standard interpolation estimate. Exemplarily we discuss the boundary term. With Lemma 3.44

$$\begin{aligned}\|\partial_n u\|_{\partial\Omega}\|I_h u - u_h\|_{\partial\Omega} &\leqslant c\|u\|_{H^2(\Omega)} c h^{\frac{r+1}{2}} \|\nabla(I_h u - u_h)\|_\Omega \\ &\quad c\|u\|_{H^2(\Omega)} h^{\frac{r+1}{2}} \left(\|\nabla(u - I_h u)\|_\Omega + \|\nabla(u - u_h)\|_\Omega\right).\end{aligned}$$

The remaining terms can be handled in a similar fashion, such that combination with Young's inequality gives the final estimate.

(*iii*) For estimating the L^2-error, we define the adjoint problem:

$$-\Delta z = \frac{e_h}{\|e_h\|} \text{ on } \Omega \text{ with } z = 0 \text{ on } \partial\Omega,$$

such that

$$\|z\|_{H^2(\Omega)} \leqslant c_s.$$

Multiplication with e_h and integration over Ω yields

$$\|e_h\|_\Omega = (e_h, -\Delta z)_\Omega = (\nabla e_h, \nabla z)_\Omega + \langle u_h, \partial_n z\rangle_{\partial\Omega},$$

as $u = 0$ on $\partial\Omega$. Using (3.24) with $\phi_h = I_h z$, it follows

$$\begin{aligned}\|e_h\| \leqslant{}& \|\nabla e_h\|_\Omega \|\nabla(z - I_h z)\|_\Omega + \|u_h\|_{\partial\Omega}\|\partial_n z\|_{\partial\Omega} + \|\partial_n u\|_{\partial\Omega}\|I_h z\|_{\partial\Omega} \\ &+ \|f\|_S \|I_h z\|_S + \|\nabla u_h\|_S \|\nabla I_h z\|_S.\end{aligned} \tag{3.26}$$

The first term can be estimated with help of the energy estimate and the interpolation estimates, followed by the stability of the adjoint solution $\|z\|_{H^2(\Omega)} \leqslant c_s$. For the second term, we first use (3.23) and get by introducing $\pm u$

$$\|u_h\|_{\partial\Omega} \leqslant c h^{\frac{r+1}{2}} \|\nabla u_h\|_s \leqslant c h^{\frac{r+1}{2}} \left(\|\nabla e_h\|_S + \|\nabla u\|_S\right) \leqslant c h^{\frac{r+1}{2}} \|\nabla e_h\| + c h^{r+1} \|u\|_{H^2(\Omega)}.$$

This procedure will also be used for the third term. The right hand side part in the fourth term of (3.26) is estimated with Lemma 3.43

$$\|f\|_S \leqslant c h^{\frac{r+1}{2}} \|f\|_{H^1(\Omega)}.$$

For the interpolation part $\|I_h z\|$ we first use the intermediate result (3.22) from the proof of Lemma 3.43 and introduce $\pm z$ on the boundary to get with interpolation estimates

$$\begin{aligned}\|I_h z\|_S &\leqslant c h^{\frac{r+1}{2}} \|z - I_h z\|_{\partial\Omega} + c h^{\frac{r+1}{2}} \underbrace{\|z\|_{\partial\Omega}}_{=0} + c h^{r+\frac{1}{2}} \|\nabla I_h z\|_S \\ &\leqslant c h^{2+\frac{r}{2}} \|z\|_{H^2(\Omega)} + c h^{r+\frac{1}{2}} \|z\|_{H^1(\Omega)}.\end{aligned}$$

Overall, the fourth term in (3.26) is estimated as

$$\|f\|_S \|I_h z\|_S \leqslant c h^{r+\frac{3}{2}} \|f\|_{H^1(\Omega)}.$$

This trick is also used in the final term of (3.26). As $(r+1)/2 \leqslant r+1/2$

$$\begin{aligned}\|\nabla I_h z\|_S &\leqslant ch^{\frac{r+1}{2}}\|\nabla I_h z\|_{\partial\Omega} + ch^{r+\frac{1}{2}}\|\nabla^2 I_h z\|_{S_h} \\ &\leqslant ch^{\frac{r+1}{2}}\Big(\|\nabla(z-I_h z)\|_{\partial\Omega} + \|\nabla^2(z-I_h z)\|_{\Omega} + \|z\|_{H^2(\Omega)}\Big)\end{aligned}$$

The same estimate is applied to ∇u_h

$$\|\nabla u_h\|_S \leqslant ch^{\frac{r+1}{2}}\Big(\|\nabla(u-u_h)\|_{\partial\Omega} + \|\nabla^2(u-u_h)\|_{\Omega} + \|u\|_{H^2(\Omega)}\Big).$$

Together with the stability estimates of the interpolation and higher order estimates of the discrete solution (that can be shown by introducing $\pm I_h u$ and applying the inverse estimate to the discrete parts) we get

$$\|\nabla u_h\|_S \|\nabla I_h z\|_S \leqslant ch^{r+1}\|f\|_{L^2(\Omega)}.$$

□

3.5 Praktische Aspekte der Finite Elemente Methode

Die wesentlichen Schritte beim Lösen einer partiellen Differentialgleichung mit der Finite Elemente Methode sind:

1. Wahl eines Finite Elemente Gitters Ω_h
2. Wahl einer Finite Elemente Basis $\{\phi_h^{(i)},\ i = 1,\dots,N\}$
3. Aufstellen der rechten Seite
$$\mathbf{b}_h = (\mathbf{b}_i)_{i=1}^N, \quad \mathbf{b}_i = (f, \phi_h^{(i)})_\Omega$$
4. Aufstellen der Systemmatrix
$$\boldsymbol{A}_h = (\boldsymbol{A}_{ij})_{i,j=1}^N, \quad \boldsymbol{A}_{ij} = a(\phi_h^{(j)}, \phi_h^{(i)})$$
5. Lösen des linearen Gleichungssystems
$$\boldsymbol{A}_h \mathbf{u}_h = \mathbf{b}_h.$$

Schritte 3. und 4. bestehen im Wesentlichen aus Integration von Testfunktionen, der rechten Seite und eventuell von weiteren Daten. Zum automatischen Aufbauen von rechter Seite und Matrix muss diese Integration durch numerische Approximation erfolgen. Hierdurch werden zusätzliche numerische Fehler in das diskrete Problem eingeführt. Im Folgenden werden wir den Einfluss der numerischen Integration auf das Gesamtverfahren genauer untersuchen.

3.5.1 Numerischer Aufbau der Gleichungen

Das Finite Elemente Verfahren ist durch die lokale Vorgabe einer Basis gegeben. Für jede Basisfunktion $\phi_h^{(i)}$ gilt, dass $\mathrm{supp}(\phi_h^{(i)}) \cap T \neq \emptyset$ nur für sehr wenige Elemente des Gitters gilt. Auf jedem Element ist jede Basisfunktion ein Polynom und daher sehr effizient zu integrieren.

Der Aufbau von rechter Seite und Systemmatrix bei der Finite Elemente Methode wird *Assemblierung* genannt und nutzt diese lokale Definition durch geeignetes Sortieren der Terme. Es gilt:

$$\mathbf{b}_i = \int_\Omega f\phi_h^{(i)}\,dx = \sum_{T\in\Omega_h}\int_T f\phi_h^{(i)}\,dx = \sum_{T\cap\mathrm{supp}(\phi_h^{(i)})\neq\emptyset}\int_T f\phi_h^{(i)}\,dx$$

Zur Integration der Vektorkomponente $\mathbf{b}_i$ muss also über all diejenigen Elemente $T \in \Omega_h$ integriert werden, welche den Träger von $\phi_h^{(i)}$ schneiden. Bei der praktischen Realisierung von Finite Elemente Methoden ist es meist (aufgrund der verwendeten Datenstrukturen) schwierig zu einer Basisfunktion alle Elemente zu finden, in denen die Basisfunktion definiert ist. Da hingegen die Basisfunktionen lokal durch Vorgabe von Knotenfunktionalen auf jedem Element definiert sind, besteht für jedes Element $T \in \Omega_h$ unmittelbar Zugriff auf alle Basisfunktionen $\phi_h^{(i)}$ welche auf T definiert sind.

Die Integration von rechter Seite und Matrix ist deshalb lokal aufgebaut: auf jeder Zelle $T \in \Omega_h$ werden die Integrale über alle auf T definierten Basisfunktionen berechnet und lokal gespeichert:

$$\mathbf{b}_{T,j} = (f, \phi_h^{(T,j)})_T \quad j = 1,\dots,N_T, \quad \mathbf{A}_{T,ij} = a(\phi_h^{(T,j)}, \phi_h^{(T,i)})_T \quad i,j = 1,\dots,N_T,$$

wobei N_T die lokale Anzahl Basisfunktionen ist, also die Dimension des zugrundeliegenden Polynomraums $P(T)$. Bei linearen Finiten Elementen auf Dreiecken ist $N_T = 3$. In einem ersten Schritt werden also sogenannte *Element-Steifigkeits-Matrizen* und *Element-Lastvektoren* gebildet. Im Anschluss wird die globale Systemmatrix $\mathbf{A}_h$ und der Lastvektor $\mathbf{b}_h$ aus den lokalen Beiträgen zusammengesetzt:

$$\mathbf{A}_{ij} = \sum_{T\in\Omega_h}\sum_{k,l=1}^{N_T}\mathbf{A}_{T,i(k)j(l)}, \quad \mathbf{b}_i = \sum_{T\in\Omega_h}\sum_{k=1}^{N_T}\mathbf{b}_{T,i(k)}.$$

Dieses Vorgehen hat den Vorteil, dass die Integration stets nur auf einzelnen Elementen des Gitters erfolgt. Hier sind die Basisfunktionen Polynome (oder im Fall von parametrischen Finiten Elementen rationale Funktionen) und also auf jedem Element beliebig regulär. Dies ist wichtig für die Verwendung von numerischen Quadraturregeln.

Bei parametrischen Finite Elemente Ansätzen erfolgt die Integration der lokalen Element-Matrizen und Element-Lastvektoren durch Transformation auf ein Referenzelement $\hat{T}$. Mit der Transformation $T_T : \hat{T} \to T$ gilt für die rechte Seite:

$$\mathbf{b}_{T,j} = \int_T f(x)\phi_h^{(T,j)}(x)\,dx = \int_{\hat{T}} \det(T_T(\hat{x}))\hat{f}(\hat{x})\hat{\phi}_h^{(T,j)}(\hat{x})\,d\hat{x},$$

und für die Matrix (im Fall der Poisson-Gleichung):

$$\mathbf{A}_{T,ij} = \int_T \nabla\phi_h^{(j)}(x)\nabla\phi_h^{(i)}(x)\,dx = \int_{\hat{T}} \det(T_T(\hat{x}))B_T^{-T}\hat{\nabla}\hat{\phi}_h^{(j)} \cdot B_T^{-T}\hat{\nabla}\hat{\phi}_h^{(i)}\,d\hat{x}.$$

Diese Integrale auf der Referenzzelle $\hat{T}$ werden mit Hilfe einer numerischen Quadraturformel berechnet. Wir betrachten im Folgenden also das Problem, eine gegebene Funktion $\hat{v}$ auf einer Zelle $\hat{T}$ zu integrieren, bzw. das Integral mit hinreichender Genauigkeit zu approximieren. Hierzu werden *interpolatorische Quadraturformeln* verwendet, welche die Funktion $\hat{v}$ mit einem Polynom $\hat{p} \in P_{\hat{T}}$ mit $\dim(P_{\hat{T}}) = S$ interpolieren und das Integral über $\hat{v}$ durch das Integral über $\hat{p}$ approximieren

$$\int_{\hat{T}} \hat{v}\,d\hat{x} = \int_{\hat{T}} \hat{p}\,d\hat{x} + \mathcal{R}_{\hat{T}},$$

mit einem Restglied $\mathcal{R}_{\hat{T}}$. Ist durch $\hat{L}_k(\hat{x}),\ k = 1,\dots,S$ eine Basis des Polynomraums $P_{\hat{T}}$ gegeben, so schreibt sich die Quadraturformel einfach als

$$Q_{\hat{T}}(\hat{v}) = \sum_{k=1}^{S} \hat{\omega}_k\hat{v}(\hat{x}_k), \quad \hat{\omega}_k = \int_{\hat{T}} \hat{L}_k(\hat{x})\,d\hat{x},$$

mit Gewichten $\hat{\omega}_k$ und Stützstellen $\hat{x}_k \in \hat{T}$.

Definition 3.47 (Quadratur). *Eine interpolatorische Quadraturformel Q_h auf der Referenzzelle $\hat{T}$ heißt von Ordnung r, wenn sie Polynome bis zum Grad $r-1$ exakt integriert. Sie heißt* zulässig *für den Polynomansatz $P(\hat{T})$, falls die Stützstellenmenge reichhaltig genug ist, so dass:*

$$q \in P(\hat{T}): \quad \nabla q(x_k) = 0\ (k = 1,\dots,S) \quad \Rightarrow \quad q \equiv \textit{konstant.}$$

Eine Quadraturformel $Q_{\hat{T}}$ auf der Referenzzelle kann nun zur Integration einer beliebigen Funktion v auf T verwendet werden. Mit der Schreibweise $\hat{v}(\hat{x}) = v(x)$ und $x = T_T\hat{x}$ gilt:

$$Q_T(v) := \sum_{k=1}^{S} \omega_k\, v(x_k) = \sum_{k=1}^{S} \underbrace{\det(B_T(x_k))\hat{\omega}_k}_{=:\omega_k}\, \hat{v}(\hat{x}_k). \tag{3.27}$$

Für den Fehler einer Quadraturformel einer gegebenen Ordnung r auf einer Zelle $T \in \Omega_h$ gilt

Lemma 3.48 (Quadraturfehler). *Für eine interpolatorische Quadraturformel Q_T der Ordnung $r \geqslant d$ auf einer Zelle $T \in \Omega_h$ angewendet auf eine Funktion $v \in W^{r,1}(T)$ gilt:*

$$\left|\int_T v\,dx - Q_T(v)\right| \leqslant c_T h_T^r \int_T |\nabla^r v|\,dx,$$

mit einer Konstante c_T welche von T abhängt.

Proof: Mit $T_T : \hat{T} \to T$ und $x = T_T\hat{x}$ verwenden wir wieder die Schreibweise $\hat{v}(\hat{x}) = v(x)$. Auf der Referenzzelle definieren wir ein Fehlerfunktional:

$$F(\hat{v}) := \left| \int_{\hat{T}} \hat{v}(\hat{x})\, d\hat{x} - Q_{\hat{T}}(\hat{v}) \right|.$$

Dieses Funktional ist wohldefiniert, wenn die Auswertung von Punktwerten $\hat{v}(\hat{x}_k)$ erlaubt ist. Der Einbettungssatz fordert hierfür:

$$r - \frac{d}{1} \geqslant 0 \quad \Leftrightarrow \quad r \geqslant d.$$

Es gilt somit:

$$|F(\hat{v})| \leqslant c\|v\|_{W^{r,1}(\hat{T})}.$$

Weiter ist F sublinear $|F(\hat{v} + \hat{w})| \leqslant c(|F(\hat{v})| + |F(\hat{w})|)$ und es verschwindet auf dem Polynomraum P^{r-1}, denn r ist gerade die Ordnung der Quadraturregel. Wir können das Bramble-Hilbert Lemma anwenden und erhalten:

$$|F(\hat{v})| \leqslant c\|\hat{\nabla}^r\hat{v}\|_{L^1(\hat{T})}.$$

Wir haben hier eine Variante des Bramble-Hilbert-Lemmas für die L^1-Norm verwendet.

Dieses Fehlerfunktional wird auf die Zelle T transformiert. Hierzu nehmen wir an, dass die Transformation affin linear ist. Es gilt (vergleiche den Beweis zur speziellen Interpolationsabschätzung, Satz 3.27) :

$$\left| \int_T v(x)\, dx - Q_T(v) \right| = |\det B_T|\, |F(\hat{v})|,$$

sowie

$$\|\hat{\nabla}^r\hat{v}\|_{L^1(\hat{T})} \leqslant c|\det B_T|^{-1} h_T^r \|\nabla^r v\|_{L^1(T)}.$$

Kombination beider Transformationen ergibt die Behauptung. □

Die Konstruktion von Quadraturformeln auf allgemeinen Elementen, also Dreiecken, Vierecken, Tetraeder, usw. ist weitaus komplizierter als im eindimensionalen Fall. Der Fall von Vierecks-, sowie Hexaedergittern ist einfach, da hier ein *Tensorprodukt-Ansatz* möglich ist. Es sei $\hat{T} = (0,1)^d$ das Referenzelement und $\hat{I} = (0,1)$ das Einheitsintervall. Auf $\hat{I}$ sei eine interpolatorische Quadraturregel $Q_{\hat{I}}$ einer gegebenen Ordnung r gegeben:

$$Q_{\hat{I}}(\hat{v}) = \sum_{k=1}^{S} \hat{\omega}_k \hat{v}(\hat{x}_k).$$

Auf Intervallen stehen zum Beispiel die sehr effizienten Gauß'schen Quadraturregeln zur Verfügung. Diese Regeln haben bei Verwendung von S Stützstellen die Ordnung $r = 2S$. Für die Einheitszelle gilt $\hat{T} = \hat{I}^d$ und wir definieren die Quadraturregel

$$Q_{\hat{T}}(\hat{v}) = \sum_{k_1=1}^{S} \cdots \sum_{k_d=1}^{S} \hat{\omega}_{k_1} \cdots \hat{\omega}_{k_d} \hat{v}(\hat{x}^1_{k_1}, \ldots, \hat{x}^d_{k_d}),$$

mit S^d Punkten und Gewichten. Ist die ursprüngliche Quadraturformel $Q_{\hat{I}}$ von Ordnung r, so ist die Tensorprodukt-Formel $Q_{\hat{T}}$ auch von Ordnung r und integriert alle Polynome aus dem Raum Q^{r-1} exakt

$$Q^{r-1} := \{x_1^{\alpha_1} \cdots x_d^{\alpha_d},\ 0 \leqslant \alpha_1, \ldots, \alpha_d \leqslant r-1\}.$$

Auf allgemeinen Elementen (also etwa Dreiecken) ist diese Konstruktion nicht möglich. Einfache Quadraturformeln auf Dreiecken sind z.B. die Mittelpunktsregel

$$Q_T(v) = |T| v(s), \quad s \text{ Schwerpunkt von } T,$$

von Ordnung 1 oder die Trapezregel:

$$Q_T(v) = |T| \sum_{k=1}^{3} \frac{1}{3} v(x_i), \quad x_i \text{ Eckpunkt von} T,$$

von Ordnung 2. In Figure 3.8 we show different quadrature rules on the reference triangle. They can be mapped via T_T to every triangle $T \in \Omega_h$.

By numerical quadrature, system matrix $\mathbf{A}_h$ and right hand side $\mathbf{b}_h$ are not exact. Instead of

$$\mathbf{A}_h \mathbf{x}_h = \mathbf{b}_h$$

we solve the disturbed system

$$\tilde{\mathbf{A}}_h \tilde{\mathbf{x}}_h = \tilde{\mathbf{b}}_h$$

that gives rise to the disturbed finite element solution

$$\tilde{u}_h(x) = \sum_{i=1}^{N} \tilde{\mathbf{x}}_i \phi_i(x).$$

We must include the additional quadrature error into the analysis. Prior however we musst assure, that the disturbed system matrix $\tilde{\mathbf{A}}_h$ is still regular, such that we get a unique solution $\tilde{\mathbf{u}}_h$ at all. It holds

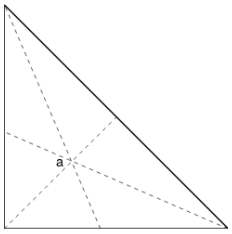
Lemma 3.49 (Finite Elements with numerical quadrature)**.** *Let $Q_{\hat{T}}$ be a permissable quadrature role of order $r \geqslant d$. The disturbed Finite Elemente solution $\tilde{u}_h \in V_h$ of the Laplace problem is uniquely determined and for the finite element approximation of degree $m-1$ it holds*

$$\|u - \tilde{u}_h\| \leqslant c h^{\min\{m, r+3-m\}} \|u\|_{H^m(\Omega)}, \quad \|\nabla(u - \tilde{u}_h)\| \leqslant c h^{\min\{m-1, r+2-m\}} \|u\|_{H^m(\Omega)}.$$

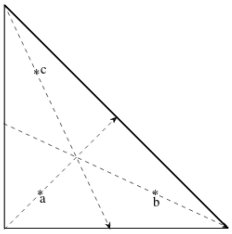
Remark 3.50. *To avoid loss of convergence order we need a quadrature rule of degree*

$$r + 3 - m \geqslant m \quad \Leftrightarrow \quad r \geqslant 2m - 3.$$

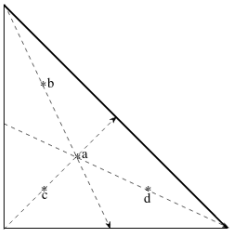
For linear finite elements, $r = 1$ is sufficient such that constant polynomials are exactly integrated. For quadratic finite element we need $r \geqslant 3$ such that quadratic polynomials are exactly integrated. If $\phi_h|_T \in P^{m-1}(T)$, the product $\nabla\phi_h \cdot \nabla\phi_h$ is a polynomial of degree $2(m-2)$. By this rule, the matrix (without a coefficient) will always be integrated without a quadrature error.



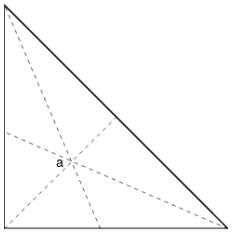
(a) Linear
$a = \left(\frac{1}{3}, \frac{1}{3}\right),\ w = 1$



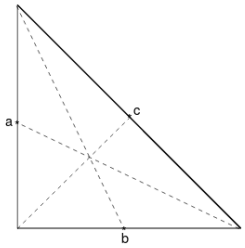
(b) Quadratic
$a = \left(\frac{1}{6}, \frac{1}{6}\right),\ w = \frac{1}{3}$
$b = \left(\frac{2}{3}, \frac{1}{6}\right),\ w = \frac{1}{3}$
$c = \left(\frac{1}{6}, \frac{2}{3}\right),\ w = \frac{1}{3}$



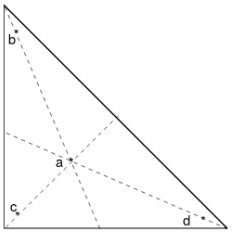
(c) Cubic
$a = \left(\frac{1}{3}, \frac{1}{3}\right),\ w = -\frac{27}{48}$
$b = \left(\frac{1}{5}, \frac{3}{5}\right),\ w = \frac{25}{48}$
$c = \left(\frac{1}{5}, \frac{1}{5}\right),\ w = \frac{25}{48}$
$d = \left(\frac{3}{5}, \frac{1}{5}\right),\ w = \frac{25}{48}$



(a) Linear
$a = \left(\frac{1}{3}, \frac{1}{3}\right),\ w = 1$



(b) Quadratic
$a = \left(0, \frac{1}{2}\right),\ w = \frac{1}{3}$
$b = \left(\frac{1}{2}, 0\right),\ w = \frac{1}{3}$
$c = \left(\frac{1}{2}, \frac{1}{2}\right),\ w = \frac{1}{3}$



(c) Cubic
$a = \left(\frac{1}{3}, \frac{1}{3}\right),\ w = -\frac{27}{48}$
$b = \left(\frac{2}{15}, \frac{11}{15}\right),\ w = \frac{25}{48}$
$c = \left(\frac{2}{15}, \frac{2}{15}\right),\ w = \frac{25}{48}$
$d = \left(\frac{11}{15}, \frac{2}{15}\right),\ w = \frac{25}{48}$

Abbildung 3.8: Numerical quadrature rules on triangles.

3.5.2 Eigenschaften der Systemmatrix

In die Lösung der linearen Gleichungssysteme

$$\mathbf{A}_h \mathbf{u}_h = \mathbf{b}_h,$$

fließt oft der größte Teil des numerischen Aufwands. Die Steifigkeitsmatrix $\mathbf{A}_h \in \mathbb{R}^{N \times N}$ ist sehr groß, oft gilt $N \gg 1\,000\,000$, dafür aber nach Konstruktion der Finite Elemente Basis dünn besetzt mit je nach Ansatzgrad und Dimension zwischen 5 und über 100 Einträgen pro Zeile. Üblicherweise, insbesondere bei Verwendung von allgemeinen Gittern, hat die Matrix keine Bandstruktur sondern die Struktur ändert sich von Zeile zu Zeile.

Durch numerische Quadratur treten zwangsläufig Rundungsfehler auf. Beim Lösen des linearen Gleichungssystems werden diese Fehler verstärkt. Es gilt die folgende Abschätzung.

Lemma 3.51 (Störungssatz)**.** *Let $\delta\mathbf{A}_h$ and $\delta\mathbf{b}_h$ be distortions of system matrix and right hand side satisfying*

$$\mu := \operatorname{cond}_2(\mathbf{A}_h)\frac{\|\delta\mathbf{A}_h\|}{\|\mathbf{A}_h\|} < 1, \quad \|\delta\mathbf{A}_h\| < \|\mathbf{A}_h^{-1}\|^{-1}.$$

Then, for the distorted solution $\tilde{\mathbf{u}}_h = \mathbf{u}_h + \delta\mathbf{u}_h$ it holds

$$\frac{\|\delta\mathbf{u}_h\|}{\|\mathbf{u}_h\|} \leqslant \frac{\operatorname{cond}_2(\mathbf{A}_h)}{1-\mu}\left\{\frac{\|\delta\mathbf{A}_h\|}{\|\mathbf{A}_h\|} + \frac{\|\delta\mathbf{b}_h\|}{\|\mathbf{b}_h\|}\right\}.$$

Proof. (*i*) We start with the simple case and consider an error in the right hand side only $\tilde{\mathbf{b}}_h = \mathbf{b}_h + \delta\mathbf{b}_h$. It holds

$$\mathbf{A}_h\mathbf{u}_h = \mathbf{b}_h, \quad \mathbf{A}_h\hat{\mathbf{u}}_h = \tilde{\mathbf{b}}_h \quad \Rightarrow \quad \mathbf{u}_h - \hat{\mathbf{u}}_h = \mathbf{A}_h^{-1}\delta\mathbf{b}_h$$

and hence with $\mathbf{A}\mathbf{u}_h = \mathbf{b}_h$

$$\frac{\|\mathbf{u}_h - \hat{\mathbf{u}}_h\|}{\|\mathbf{A}_h\|\,\|\mathbf{u}_h\|} \leqslant \frac{\|\mathbf{u}_h - \hat{\mathbf{u}}_h\|}{\|\mathbf{A}_h\mathbf{u}_h\|} = \frac{\|\mathbf{u}_h - \hat{\mathbf{u}}_h\|}{\|\mathbf{b}_h\|} \leqslant \|\mathbf{A}_h^{-1}\|\frac{\|\delta\mathbf{b}_h\|}{\|\mathbf{b}_h\|}.$$

Multiplication with $\|\mathbf{A}\|$ gives the first result

$$\frac{\|\mathbf{u}_h - \hat{\mathbf{u}}_h\|}{\|\mathbf{u}_h\|} \leqslant \operatorname{cond}(\mathbf{A}_h)\frac{\|\delta\mathbf{b}_h\|}{\|\mathbf{b}_h\|} \tag{3.28}$$

if we consider an error in the right hand side only.

(*ii*) Now, we consider an error in the matrix $\tilde{\mathbf{A}}_h = \mathbf{A}_h + \delta\mathbf{A}_h$. It holds

$$\mathbf{A}_h\mathbf{u}_h = \mathbf{b}_h, \quad \tilde{\mathbf{A}}_h\bar{\mathbf{u}}_h = \mathbf{b}_h \quad \Rightarrow \quad \mathbf{u}_h - \bar{\mathbf{u}}_h = \tilde{\mathbf{A}}_h^{-1}\delta\mathbf{A}_h\mathbf{u}_h.$$

Using the assumption $\|\mathbf{A}_h^{-1}\delta\mathbf{A}_h\| \leqslant \|\mathbf{A}_h^{-1}\|\,\|\delta\mathbf{A}_h\| < 1$ gives

$$\|(\mathbf{A}_h + \delta\mathbf{A}_h)^{-1}\delta\mathbf{A}_h\| = \|(\mathrm{I} + \mathbf{A}_h^{-1}\delta\mathbf{A})^{-1}\mathbf{A}_h^{-1}\delta\mathbf{A}_h\| < \frac{\|\mathbf{A}_h^{-1}\|\,\|\delta\mathbf{A}_h\|}{1 - \|\mathbf{A}_h^{-1}\delta\mathbf{A}_h\|}.$$

Hence

$$\|\mathbf{u}_h - \bar{\mathbf{u}}_h\| \leqslant \frac{\|\mathbf{A}_h^{-1}\|\,\|\delta\mathbf{A}_h\|}{1 - \|\mathbf{A}_h^{-1}\delta\mathbf{A}_h\|}\|\mathbf{u}_h\|.$$

We define by $\|\mathbf{u}_h\|$ and extend the fraction by $\|\mathbf{A}_h\|/\|\mathbf{A}_h\|$ (twice) to get

$$\frac{\|\mathbf{u}_h - \bar{\mathbf{u}}_h\|}{\|\mathbf{u}_h\|} \leqslant \frac{\operatorname{cond}(\mathbf{A}_h)}{1 - \operatorname{cond}(\mathbf{A}_h)\frac{\|\delta\mathbf{A}_h\|}{\|\mathbf{A}_h\|}}\frac{\|\delta\mathbf{A}_h\|}{\|\mathbf{A}_h\|}$$

(*iii*) The final result follows by combining both estimates. □

Lemma 3.52. *Let $\|\cdot\|$ be a matrix norm, induced by a corresponding vector nom. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be such that $\|\mathbf{A}\| < 1$. Then, $I + \mathbf{A}$ is a regular matrix and it holds*

$$\|(I+\mathbf{A})^{-1}\| \leqslant \frac{1}{1-\|\mathbf{A}\|}.$$

Proof. As $\|\cdot\|$ is an induced norm

$$\|(I+\mathbf{A})a\| \geqslant \|x\| - \|\mathbf{A}x\| \geqslant (1-\|\mathbf{A}\|)\|x\|.$$

Therefore $1-\|\mathbf{A}\| > 0$ shows, that $I+\mathbf{A}$ is injective and regular. Finally

$$\begin{aligned} 1 = \|I\| &= \|(I+\mathbf{A})(I+\mathbf{A})^{-1}\| \| + \|(I+\mathbf{A})^{-1} + \mathbf{A}(I+\mathbf{A}) \quad -1\| \\ &\geqslant \|(I+\mathbf{A})^{-1}\| - \|\mathbf{A}\|\,\|(I+\mathbf{A})^{-1}\| \\ &= \|(I+\mathbf{A})^{-1}\|(1-\|\mathbf{A}\|) > 0. \end{aligned}$$

□

Die bestimmende Größe für die Fehlerfortpflanzung ist also die Konditionszahl der Matrix. Wir beweisen:

Lemma 3.53 (Konditionierung der Steifigkeitsmatrix)**.** *Auf einer Folge von regulären Gittern Ω_h gelten für die Konditionszahlen der Steifigkeitsmatrix $\mathbf{A}_h$ (der Poisson-Gleichung) sowie für die Massenmatrix $\mathbf{M}_h$:*

$$\mathrm{cond}_2(\mathbf{A}_h) = O(h^{-2}), \quad \mathrm{cond}_2(\mathbf{M}_h) = O(1).$$

Proof: (i) Beide Matrizen sind positiv definit. Die Spektralkondition ist also gegeben durch:

$$\mathrm{cond}_2(\mathbf{A}_h) = \frac{\lambda_{\max}(\mathbf{A}_h)}{\lambda_{\min}(\mathbf{A}_h)}, \quad \mathrm{cond}_2(\mathbf{M}_h) = \frac{\lambda_{\max}(\mathbf{M}_h)}{\lambda_{\min}(\mathbf{M}_h)}.$$

Für die Eigenwerte einer positiv definiten Matrix $\mathbf{A}$ gilt

$$\lambda_{\min}(\mathbf{A}) = \min_{\mathbf{v}\in\mathbb{R}^N} \frac{\langle \mathbf{A}\mathbf{v},\mathbf{v}\rangle}{|\mathbf{v}|^2} \leqslant \max_{\mathbf{v}\in\mathbb{R}^N} \frac{\langle \mathbf{A}\mathbf{v},\mathbf{v}\rangle}{|\mathbf{v}|^2} = \lambda_{\max}(\mathbf{A}).$$

(ii) Wir bestimmen zunächst die Konditionszahl der Massenmatrix. Mit den Element- Massenmatrizen $\mathbf{M}_T$ und der elementweisen Einschränkung $\mathbf{v}_T = \mathbf{v}_{|T}$ gilt für einen Vektor $v_h \in V_h$ mit Koeffizienten $\mathbf{v}$:

$$\begin{aligned} \langle \mathbf{M}_h\mathbf{v},\mathbf{v}\rangle &= \sum_{T\in\Omega_h} \frac{\langle \mathbf{M}_T\mathbf{v}_T,\mathbf{v}_T\rangle}{|\mathbf{v}_T|^2}|\mathbf{v}_T|^2 \\ &\geqslant \min_{T\in\Omega_h,\,\mathbf{v}\in\mathbb{R}^N} \frac{\langle \mathbf{M}_T\mathbf{v}_T,\mathbf{v}_T\rangle}{|\mathbf{v}_T|^2} \sum_{T\in\Omega_h} |\mathbf{v}_T|^2 \geqslant \min_{T\in\Omega_h}\{\lambda_{\min}(\mathbf{M}_T)\}|\mathbf{v}|^2, \end{aligned}$$

Entsprechend gilt für den maximalen Eigenwert:

$$\langle \boldsymbol{M}_h \mathbf{v}, \mathbf{v}\rangle \leqslant \max_{T\in\Omega_h, \mathbf{v}\in\mathbb{R}^N} \frac{\langle M_T \mathbf{v}_T, \mathbf{v}_T\rangle}{|\mathbf{v}_T|^2} \sum_{T\in\Omega_h} |\mathbf{v}_T|^2 \leqslant \min_{T\in\Omega_h} \{\lambda_{max}(\boldsymbol{M}_T)\} d_{max} |\mathbf{v}|^2,$$

wobei d_{max} die maximale Zahl Zellen ist, die sich in einem Knoten treffen. (Diese Konstante ist auf formregulären Gittern gleichmäßig in $h > 0$ beschränkt).

Für die Einträge der Massenmatrix gilt bei Transformation auf das Referenzelement:

$$\mathbf{m}_{ij} = |\det B_T| \hat{\mathbf{m}}_{ij},$$

und es gilt also mit $|\det B_T| = O(h_T^d)$ für die Eigenwerte von $\boldsymbol{M}_T$:

$$\lambda_{max}(\boldsymbol{M}_T) = |\det B_T| \lambda_{max}(M_{\hat{T}}) \leqslant c h_T^d, \quad \lambda_{min}(\boldsymbol{M}_T) = |\det B_T| \lambda_{min}(M_{\hat{T}}) \geqslant c h_T^d.$$

Die Matrix $\boldsymbol{M}_{\hat{T}}$ ist fest, die Eigenwerte können durch Konstanten abgeschätzt werden. Es folgt:

$$\lambda_{min}(\boldsymbol{M}_h) \geqslant c h^d, \quad \lambda_{max}(\boldsymbol{M}_h) \leqslant c h^d \quad \Rightarrow \quad \mathrm{cond}_2(\boldsymbol{M}_h) = O(1).$$

(*iii*) Die Eigenwerte der Steifigkeitsmatrix $\boldsymbol{A}_h$ wollen wir auf die Eigenwerte der Massenmatrix $\boldsymbol{M}_h$ zurückführen. Es gilt:

$$\lambda_{min}(\boldsymbol{A}_h) \geqslant \min_{\mathbf{v}\in\mathbb{R}^N} \frac{\langle \boldsymbol{A}_h \mathbf{v}, \mathbf{v}\rangle}{\langle \boldsymbol{M}_h \mathbf{v}, \mathbf{v}\rangle} \min_{\mathbf{v}\in\mathbb{R}^N} \frac{\langle \boldsymbol{M}_h \mathbf{v}, \mathbf{v}\rangle}{|\mathbf{v}|^2} = \min_{v_h\in V_h} \frac{a(v_h, v_h)}{\|v_h\|^2} \lambda_{min}(\boldsymbol{M}_h),$$
$$\lambda_{max}(\boldsymbol{A}_h) \leqslant \max_{\mathbf{v}\in\mathbb{R}^N} \frac{\langle \boldsymbol{A}_h \mathbf{v}, \mathbf{v}\rangle}{\langle \boldsymbol{M}_h \mathbf{v}, \mathbf{v}\rangle} \max_{\mathbf{v}\in\mathbb{R}^N} \frac{\langle \boldsymbol{M}_h \mathbf{v}, \mathbf{v}\rangle}{|\mathbf{v}|^2} = \max_{v_h\in V_h} \frac{a(v_h, v_h)}{\|v_h\|^2} \lambda_{max}(\boldsymbol{M}_h).$$

Weiter gilt wegen $V_h \subset V := H_0^1(\Omega)$:

$$\min_{v_h\in V_h} \frac{a(v_h, v_h)}{\|v_h\|^2} \geqslant \inf_{v\in H_0^1(\Omega)} \frac{a(v, v)}{\|v\|^2} =: \lambda_{min}(\Delta),$$

mit dem kleinsten Eigenwert des Laplace-Operators auf Ω. Mit der inversen Beziehung, Satz 3.29 gilt ferner:

$$a(v_h, v_h) \leqslant \sum_{T\in\Omega_h} \|\nabla v_h\|_T^2 \leqslant c \sum_{T\in\Omega_h} h_T^{-2} \|v_h\|_T^2 \leqslant c \max_{T\in\Omega_h} h_T^{-2} \|v_h\|^2,$$

Insgesamt gilt also:

$$\lambda_{min}(\Delta)\lambda_{min}(\boldsymbol{M}_h) \leqslant \lambda_{min}(\boldsymbol{A}_h) \leqslant \lambda_{max}(\boldsymbol{A}_h) \leqslant c \max_{T\in\Omega_h} h_T^{-2} \lambda_{max}(\boldsymbol{M}_h).$$

Mit $\lambda_{min}(\Delta) = c_0 > 0$ und den Eigenwerten der Massenmatrix folgt die Behauptung. □

It is important to note, that the h-dependency of the condition number

$$\mathrm{cond}_2(\boldsymbol{A}_h) = \mathcal{O}(h^{-2})$$

comes from the degree of the differential operator. $-\Delta$ is a second order differential operator. The mass matrix corresponds to the identity operator id and here, the condition behaves like $\mathcal{O}(1)$. Finite element approximations of the operator Δ^2, where

$$a(u, \phi) = (-\Delta u, -\Delta \phi)$$

have a system matrix with a condition number that scales as $\mathcal{O}(h^{-4})$. The condition number does not depend on the polynomial degree of V_h and it also does not depend on the dimension d of $\Omega \subset \mathbb{R}^d$.

3.6 A posteriori Fehlerschätzung und adaptive Finite Elemente

In diesem Abschnitt befassen wir uns mit der *a posteriori Fehlerschätzung*. Hier geht es zum einen um die Frage, eine berechenbare Schranke $\eta_h \in \mathbb{R}$ für den Fehler der Finite Elemente Approximation angeben zu können:

$$|J(u - u_h)| \leqslant \eta_h(\Omega_h, u_h, f),$$

also eine Größe η_h, welche bei Kenntnis des Gitters, der Lösung und der Problemdaten berechenbar ist. J soll hier ein beliebiges Fehlerfunktional sein. Im Gegensatz zu *a priori* Fehlerabschätzungen muss auf unbekannte Konstanten soweit wie möglich verzichtet werden, d.h., zur Berechnung des Schätzers η_h dürfen nur das Gitter Ω_h, die Daten f sowie die berechnete diskrete Lösung u_h eingehen, nicht aber etwa die Lösung $u \in H_0^1(\Omega)$ oder Konstanten welche nicht berechnet werden können (wie die Konstante des Spurlemmas, der Interpolation, oder die Poincaré Konstante).

Der zweite Aspekt in diesem Kapitel ist die Berechnung von *Fehlerindikatoren* $\{\eta_T\}_{T\in\Omega_h}$. Das sind verteilte Größen, welche den lokalen Fehleranteil angeben. Lokal kann hier bedeuten, dass etwa η_T den Fehlerbeitrag der Gitterzelle $T \in \Omega_h$ angibt, oder η_i den Fehlerbeitrag des Einzugbereichs einer Finite Elemente Basisfunktion. Solche lokalen Fehlerindikatoren sind Grundlage von *adaptiven Verfahren* nach dem folgenden Muster:

1. Berechne diskrete Lösung $u_h \in V_h$
2. Schätze den Fehler η_h. Falls $\eta_h < TOL$ einer vorgegebenen Fehlertoleranz, Abbruch.
3. Erstelle lokale Fehlerindikatoren $\{\eta_T\}_{T\in\Omega_h}$ und *verfeinere das Gitter* $\Omega_h \xrightarrow{\{\eta_T\}} \Omega_h'$. Weiter bei 1 mit V_h' auf Ω_h'.

Wir betrachten in diesem Abschnitt exemplarisch die Poisson-Gleichung:

$$u \in V := H_0^1(\Omega) \qquad (\nabla u, \nabla\phi) = (f, \phi) \qquad \forall \phi \in V, \tag{3.29}$$

$$u_h \in V_h \subset V \qquad (\nabla u_h, \nabla\phi_h) = (f, \phi_h) \quad \forall \phi_h \in V_h. \tag{3.30}$$

3.6.1 Residuenbasierte Fehlerschätzer

Bei der Berechnung von a posteriori Fehlerschätzern darf die unbekannte Lösung $u \in V$ nicht eingehen. Ein zentraler Begriff ist das *Residuum*:

Definition 3.54 (Residuum)**.** *Das Residuum der Gleichung* (3.29) *an der Stelle* $u_h \in V_h$ *ist ein Funktional* $R_h : V \to \mathbb{R}$*:*

$$R_h(u_h)(\phi) = (f, \phi) - (\nabla u_h, \nabla\phi) \quad \forall \phi \in V.$$

Das Residuum steht im engen Zusammenhang zum Fehler $e_h := u - u_h$ der Finite Elemente Approximation:

Lemma 3.55 (Residuum). *Das Residuum $R_h(\cdot)$ ist ein stetiges lineares Funktional auf V. Ist $u_h \in V_h$ Lösung von (3.30) und $u \in V$ Lösung von (3.29) so gilt:*

$$R_h(u_h)(\phi_h) = 0 \quad \forall \phi_h \in V_h \tag{3.31}$$

$$\|\nabla(u - u_h)\|_{L^2(\Omega)} = \|R_h(u_h)\|_{-1}, \tag{3.32}$$

mit der Dualnorm

$$\|R_h(u_h)\|_{-1} := \sup_{\phi \in V} \frac{R_h(u_h)(\phi)}{\|\nabla \phi\|}.$$

Proof: (*i*) Wir zeigen zunächst, dass das Residuum ein stetiges lineares Funktional ist. Es gilt:

$$|R_h(u_h)(\phi)| = |(f, \phi) - (\nabla u_h, \nabla \phi)| \leqslant \|f\| \, \|\phi\| + \|\nabla u_h\| \, \|\nabla \phi\| \leqslant (c_p \|f\| + \|\nabla u_h\|) \|\nabla \phi\|,$$

mit der Poincare-Konstante c_p. Für $u_h \in V_h$ fest gilt also $R_h(u_h) \in V^*$.

(*ii*) Für die diskrete Lösung $u_h \in V_h$ gilt:

$$R_h(u_h)(\phi_h) = (f, \phi_h) - (\nabla u_h, \nabla \phi_h) = 0 \quad \forall \phi_h \in V_h.$$

Gleichung (3.31) folgt also unmittelbar aus der Definition des Residuums und der diskreten Lösung u_h.

(*iii*) Für die Lösungen $u \in V$ und $u_h \in V_h \subset V$ gilt:

$$R_h(u_h)(\phi) = (f, \phi) - (\nabla u_h, \nabla \phi) = (\nabla u - \nabla u_h, \phi) = (\nabla e_h, \nabla \phi). \tag{3.33}$$

Das heißt, für die Dualnorm des Residuums folgt:

$$\|R_h(u_h)\|_{-1} = \sup_{\phi \in V} \frac{(\nabla e_h, \nabla \phi)}{\|\nabla \phi\|} \leqslant \sup_{\phi \in V} \frac{\|\nabla e_h\| \, \|\nabla \phi\|}{\|\nabla \phi\|} = \|\nabla e_h\|.$$

Umgekehrt gilt mit (3.33):

$$\|\nabla e_h\|^2 = R_h(u_h)(e_h) = \frac{R_h(u_h)(e_h)}{\|\nabla e_h\|} \|\nabla e_h\| \leqslant \|\nabla e_h\| \, \|R_h(u_h)\|_{-1}.$$

Die letzten beiden Ungleichungen ergeben (3.32). □

Die Dualnorm des Residuums ist also eng mit der Energienorm des Fehlers verwandt. Ist $u_h \in V_h$ bekannt, so kann für jede gegebene Größe $\phi \in V$ auch das Residuum berechnet werden. Es ist im Allgemeinen jedoch nicht möglich die Dualnorm $\|R_h(u_h)\|_{-1}$ zu berechnen. Wir benötigen Abschätzungen für diese Dualnorm. Zunächst definieren wir als Hilfsgrößen:

Definition 3.56 (Kantensprung)**.** *Sei* $u_h \in V_h$ *und* $E \in \Omega_h$ *die Kante einer Zelle* $T \in \Omega_h$*. Wir definieren den Kantensprung über die Normalableitung:*

$$[n_E \cdot \nabla u_h] := \begin{cases} n_{T_1} \cdot \nabla u_h\big|_{T_1} + n_{T_2} \cdot \nabla u_h\big|_{T_2} & E \subset \bar{T}_1 \cap \bar{T}_2, \quad T_1 \neq T_2 \\ 0 & E \subset \partial\Omega, \end{cases}$$

wobei n_{T_i} *die bzgl.* T_i *nach außen gerichteten Normalvektoren sind. Da* $n_{T_1} = -n_{T_2}$ *gilt folgt:*

$$|[n_E \cdot \nabla u_h]| = |n_E \cdot (\nabla u_h\big|_{T_1} - \nabla u_h\big|_{T_2})|,$$

bei beliebiger Wahl von $n_E = n_{T_i}$*.*

Lemma 3.57 (Residuenbasierter a posteriori Fehlerschätzer für den Energiefehler)**.** *Es sei* $u \in V$ *Lösung von (3.29) sowie* $u_h \in V_h$ *die lineare Finite Elemente Approximation gemäß (3.30). Dann gilt für den Fehler* $e_h := u - u_h$

$$\|\nabla e_h\| \leqslant c\eta_h, \quad \eta_h := \left(\sum_{T \in \Omega_h} (\rho_T^2 + \sum_{E \in \partial T} \rho_E^2) \right)^{\frac{1}{2}},$$

mit den Zellresiduen ρ_T *und den Kantenresiduen* ρ_E*:*

$$\rho_T := h_T \|f\|_{L^2(T)}, \quad \rho_E := \frac{1}{2} h_E^{\frac{1}{2}} \|[n_E \cdot \nabla u_h]\|_{L^2(E)}.$$

Proof: Es gilt mit Satz 3.55 für die diskrete Lösung u_h:

$$\|\nabla e_h\|_{L^2(\Omega)} = \|R_h(u_h)\|_{-1}.$$

Nun seien $\phi \in V$ sowie $\phi_h \in V_h$ beliebig. Dann gilt mit (3.31)

$$\begin{aligned} R_h(u_h)(\phi) &= R_h(u_h)(\phi - \phi_h) = (f, \phi - \phi_h) - (\nabla u_h, \nabla(\phi - \phi_h)) \\ &= \sum_{T \in \Omega_h} \left(\int_T f(\phi - \phi_h)\, dx - \int_T \nabla u_h \cdot \nabla(\phi - \phi_h)\, dx \right) \\ &= \sum_{T \in \Omega_h} \left(\int_T (f + \Delta u_h)(\phi - \phi_h)\, dx - \int_{\partial T} (n_T \cdot \nabla u_h)(\phi - \phi_h)\, ds \right) \\ &= \sum_{T \in \Omega_h} \left(\int_T f\,(\phi - \phi_h)\, dx \sum_{E \in \partial T} \frac{1}{2} \int_{\partial T} [n_E \cdot \nabla u_h](\phi - \phi_h)\, ds \right). \end{aligned}$$

Es gilt $\Delta u_h\big|_T = 0$ wegen der Linearität der Lösung u_h.

An dieser Stelle schätzen wir mit Cauchy-Schwarz weiter ab:

$$|R_h(u_h)(\phi)| \leqslant \sum_{T \in \Omega_h} \left(\|f\|_{L^2(T)}\, \|\phi - \phi_h\|_{L^2(T)} + \sum_{E \in \partial T} \frac{1}{2} \|[n_E \cdot \nabla u_h]\|_{L^2(E)} \|\phi - \phi_h\|_{L^2(E)} \right)$$

Wir wollen aus den Termen $\phi - \phi_h$ positive Potenzen in der Gitterweite h gewinnen. Die Funktion ϕ nimmt Werte aus $V := H_0^1(\Omega)$ an, verfügt im Allgemeinen jedoch nicht über höhere Regularität. Wir dürfen daher nicht mit dem Ansatz $\phi_h := I_h\phi$, also der Wahl von ϕ_h als der Knoteninterpolation von ϕ weiter rechnen. Denn diese Knoteninterpolation ist im $H_0^1(\Omega)$ nicht definiert. Stattdessen wählen wir mit $\phi_h := C_h\phi$ die H^1-stabile Clement-Interpolation aus Satz 3.31 und erhalten:

$$\begin{aligned}
|R_h(u_h)(\phi)| &\leqslant \sum_{T\in\Omega_h}\left(c_I h_T\|f\|_T\,\|\nabla\phi\|_{\tilde{P}_T} + \sum_{E\in\partial T}\frac{1}{2}c_I h_T^{\frac{1}{2}}\|[n_E\cdot\nabla u_h]\|\|\nabla\phi\|_{\tilde{P}_E}\right)\\
&\leqslant c_I\left(\sum_{T\in\Omega_h}\left(\rho_T^2+\sum_{E\in\partial T}\rho_E^2\right)\right)^{\frac{1}{2}}\left(\sum_{T\in\Omega_h}\|\nabla\phi\|_{\tilde{P}_T}^2+\sum_{E\in\Omega_h}\|\nabla\phi\|_{\tilde{P}_E}^2\right)^{\frac{1}{2}}\\
&\leqslant c_T c_I\left(\sum_{T\in\Omega_h}\left(\rho_T^2+\sum_{E\in\partial T}\rho_E^2\right)\right)^{\frac{1}{2}}\|\nabla\phi\|_\Omega.
\end{aligned}$$

Die Konstante c_T beschreibt den Überlappungsgrad der Patche $\tilde{P}_T$ sowie $\tilde{P}_E$. Aus der Formregularität des Gitters folgt, dass c_T unabhängig von h eine kleine Konstante ist. Aus Satz 3.55 und der Definition der Dualnorm folgt die Behauptung. □

Um den vorgestellten Energiefehlerschätzer auswerten zu können muss die rechte Seite f vorliegen. Darüber hinaus müssen die Kantensprünge berechnet werden. Die Zellresiduen $\rho_T := \|f+\Delta u_h\|_T$ (mit $\Delta u_h = 0$ auf jedem T) messen das Residuum der *klassischen Formulierung* der Poisson-Gleichung $-\Delta u = f$. Die Kantensprünge messen die *Glattheit* der diskreten Lösung. Für $u \in C^1(\Omega)$, also für stetige Differenzierbarkeit über die Elementkanten hinaus gilt $\rho_E(u) = 0$.

Als Unbekannte gehen in den Fehlerschätzer die Konstante der Clement-Interpolation c_I sowie die Konstante c_T ein. Die Konstante c_T kann für ein gegebenes Gitter berechnet werden. Sie misst lediglich den Überlappungsgrad der Patche P_E. Die Konstante der Clement-Interpolation kann nur in Spezialfällen bestimmt werden. Üblicherweise muss eine Schätzung $c_I \approx 0.1 - 1$ vorgenommen werden.

Der Fehlerschätzer eignet sich nun für eine Schätzung des Energiefehlers, es gilt:

$$\|\nabla e_h\| \leqslant c_I\eta_h(\Omega_h, u_h, f).$$

Weiter können wir einfach zellweise Fehlerindikatoren definieren

$$\eta_T := (\rho_T^2 + \sum_{E\in\partial T}\rho_E^2)^{\frac{1}{2}} = \left(h_T^2\|f_T\|_T^2 + \frac{1}{2}\sum_{E\in\partial T}h_E\|[n_E\cdot\nabla u_h]\|_E^2\right)^{\frac{1}{2}}, \tag{3.34}$$

und diese zur Verfeinerung des Gitters verwenden. Algorithmen zur Verfeinerung des Gitters werden in einem folgenden Abschnitt vorgestellt. Idee ist, solche Elemente T in kleinere Elemente aufzuteilen, welche einen großen Fehlerbeitrag η_T haben.

Wir haben bisher eine Abschätzung $\|\nabla e_h\| \leqslant c\eta_h$, also eine obere Schranke für den Fehler bewiesen. Diese Abschätzung ist wichtig, um die Genauigkeit der Lösung zu garantieren. Soll der Fehlerschätzer jedoch zur Verfeinerung des Gitters verwendet werden, so muss er in gewissem Sinne "scharf" sein. D.h., wir benötigen ferner eine umgekehrte Abschätzung der Art:

$$c_1\eta_h \leqslant \|\nabla e_h\| \leqslant c_2\eta_h.$$

Ein Fehlerschätzer mit dieser Eigenschaft wird *effizient* genannt. Wenn diese Abschätzung nicht gilt, so ist es möglich, dass die lokale Gitterverfeinerung ineffizient verfeinert, dass also Bereiche verfeinert werden, welche keinen wesentlichen Fehlerbeitrag haben.

Wir benötigen als Hilfsatz eine spezielle Spurabschätzung:

Hilfsatz 3.58. *Auf jedem Element $T \in \Omega_h$ gilt*

$$\|\partial_n v\|_{L^2(\partial T)} \leqslant c\left(h^{\frac{1}{2}}\|\Delta v\|_{L^2(T)} + h^{-\frac{1}{2}}\|\nabla v\|_{L^2(\Omega)}\right) \quad \forall v \in H^2(T).$$

Proof: (i) Zunächst sei $\hat{T}$ ein Referenzelement. Hier gilt mit der Spurabschätzung:

$$\|\partial_n \hat{v}\|_{\partial\hat{T}} \leqslant c\|\hat{v}\|_{H^2(\hat{T})}.$$

Mit der elliptischen Regularität folgt dann

$$\|\partial_n \hat{v}\|_{\partial\hat{T}} \leqslant c\|\hat{v}\|_{H^2(\hat{T})} \leqslant cc_s\left(\|\hat{\Delta}\hat{v}\|_{\hat{T}} + \|\hat{v}\|_{\hat{T}}\right)$$

(ii) Es sei $\bar{v} \in \mathbb{R}$ der Mittelwert von $\hat{v}$ auf $\hat{T}$. Dann gilt mit der Poincaré Ungleichung:

$$\begin{aligned}\|\partial_n \hat{v}\|_{\partial\hat{T}} &= \|\partial_n(\hat{v}-\bar{v})\|_{\partial\hat{T}} \\ &\leqslant c\|\hat{v}-\bar{v}\|_{H^2(\hat{T})} \leqslant c\left(\|\hat{\Delta}(\hat{v}-\bar{v})\|_{\hat{T}} + \|\hat{v}-\bar{v}\|_{\hat{T}}\right) \leqslant c\left(\|\hat{\Delta}\hat{v}\|_{\hat{T}} + c_p\|\hat{\nabla}\hat{v}\|_{\hat{T}}\right).\end{aligned}$$

(iii) Jetzt sei $T_T(\hat{x}) = B_T\hat{x}+b_T$ die affin lineare Referenztransformation. Es gilt mit $\det(\nabla T_T) = O(h^2)$, $\det(\nabla T_T\big|_{\partial T}) = O(h)$ sowie $\|B_T\|_\infty = O(h)$:

$$\begin{aligned}\|\partial_n v\|^2_{L^2(\partial T)} &= ch\,h^{-2}\|\hat{\partial}_n\hat{v}\|^2_{L^2(\partial\hat{T})} \leqslant ch^{-1}(\|\hat{\Delta}\hat{v}\|^2_{\hat{T}} + \|\hat{\nabla}\hat{v}\|^2_{\hat{T}}) \\ &= ch^{-1}(h^{-2}\,h^4\|\Delta v\|^2_T + h^{-2}\,h^2\|\nabla v\|_T) \\ &= c(h\|\Delta v\|^2_T + h^{-1}\|v\|^2_T).\end{aligned}$$

□

Jetzt beweisen wir:

Lemma 3.59 (Effizienz des Energiefehlerschätzers)**.** *Seien $u \in V$ und $u_h \in V_h$ Lösungen der Poisson-Gleichung. Auf einer Folge von formregulären Triangulierungen Ω_h ist der Energiefehlerschätzer asymptotisch exakt:*

$$\eta_h \leqslant c\|\nabla e_h\| + ch\|f\|.$$

Proof: Es ist:

$$\eta_h^2 = \sum_{T\in\Omega_h} \left(h_T^2\|f\|_T^2 + h_T\|[\partial_n u_h]\|_{\partial T}^2\right).$$

Für die Lösung $u \in H^2(\Omega)$ gilt auf jeder Kante $[\partial_n u]_E = 0$. Also mit Hilfsatz 3.58

$$\|[\partial_n u_h]\|_{\partial T}^2 = \|[\partial_n e_h]\|_{\partial T}^2 \leqslant 2\|[\partial_n e_h]\|_{\partial T}^2 \leqslant c(h\|\Delta e_h\|_T^2 + h^{-1}\|\nabla e_h\|_T^2)$$

Es ist $\|\Delta e_h\|_T = \|f + \Delta u_h\|_T = \|f\|_T$. Also:

$$\eta_h^2 \leqslant c\sum_{T\in\Omega_h} \left(h_T^2\|f\|_T^2 + h_T^2\|f\|_T^2 + \|\nabla e_h\|_T^2\right) = c\|\nabla e_h\|^2 + ch^2\|f\|_{L^2(\Omega)}.$$

□

Weiter kann bewiesen werden, dass bei der Verwendung von linearen Finiten Elementen die Sprungterme in den Fehlerindikatoren überwiegen, dass also gilt:

Lemma 3.60 (Dominanz der Sprünge)**.** *Für den Fehler der linearen Finite Elemente Approximation gilt für rechte Seiten $f \in H^1(\Omega)$*

$$\|\nabla e_h\| \leqslant c\left(\sum_{E\in\Omega_h} h_E\|[n_E\cdot\nabla u_h]\|_{L^2(E)}^2\right)^{\frac{1}{2}} + \left(\sum_{x_i\in\Omega_h} h_i^4\|\nabla f\|_{L^2(P_i)}^2\right)^{\frac{1}{2}}.$$

Der zweite Term konvergiert mit zweiter Ordnung in Bezug auf die Zellweite h, ist also asymptotisch zu vernachlässigen. Erstaunlicherweise dreht sich die Dominanz der lokalen Fehlerbeiträge stets um. Bei quadratischen Finiten Elementen überwiegen die Zellbeiträge.

Es stellt sich im folgenden wieder die Frage nach der Schätzung des Fehlers in anderen Fehlerfunktionalen, etwa in der L^2-Norm. Dies erfordert wieder den Aubin-Nitsche-Trick:

Lemma 3.61 (A posteriori Fehlerschätzer in der L^2-Norm)**.** *Für den Fehler der linearen Finite Elemente Approximation gilt auf formregulären Gittern die a posteriori Abschätzung:*

$$\|u - u_h\|_{L^2(\Omega)} \leqslant c\left(\sum_{T\in\Omega_h}\left(h_T^4\|f\|_{L^2(T)}^2 + \frac{1}{2}\sum_{E\in\bar{T}} h_E^3\|[n_E\cdot\nabla u_h]\|_{L^2(E)}^2\right)\right)^{\frac{1}{2}}.$$

Proof: (*i*) Wir betrachten das duale Problem

$$z \in V: \quad (\nabla\phi, \nabla z) = (e_h, \phi)\|e_h\|^{-1}, \quad -\Delta z = e_h\,\|e_h\|^{-1},$$

mit einer dualen Lösung $z \in H^2(\Omega)$ und mit $\|z\|_{H^2(\Omega)} \leqslant c_s\|\Delta z\| = c_s$.

(*ii*) Es gilt:

$$R_h(u_h)(\phi) = (\nabla e_h, \nabla\phi) \quad \forall\phi \in V.$$

Also für $z = \phi$

$$R_h(u_h)(z) = (\nabla e_h, \nabla z) = \|e_h\|_{L^2(\Omega)}.$$

Weiter gilt mit der Galerkin-Orthogonalität und Einschub der Knoteninterpolation:

$$\|e_h\| = (\nabla e_h, \nabla(z - I_h z)) = (f, z - I_h z) - (\nabla u_h, \nabla(z - I_h z)).$$

Durch partielle Integration auf jeder Zelle $T \in \Omega_h$ entstehen wieder Kantenterme

$$\|e_h\| = \sum_{T \in \Omega_h} \left(\int_T f\,(z - I_h z)\,dx - \int_{\partial T} n_E \cdot \nabla u_h \cdot (z - I_h z)\,ds \right),$$

welche wir zu Sprüngen zusammenfassen können:

$$\|e_h\| = \sum_{T \in \Omega_h} \left(\int_T (f + \Delta u_h)\,(z - I_h z)\,dx - \sum_{E \in \partial T} \frac{1}{2} \int_{\partial T} [n_E \cdot \nabla u_h] \cdot (z - I_h z)\,ds \right)$$

Da $z \in H^2(\Omega)$ gelten die Interpolationsabschätzungen auf jeder Zelle und auf jeder Kante und zusammen mit der Stabilität der dualen Lösung ergibt sich

$$\|z - I_h z\|_T \leqslant c_I h_T^2 \|\nabla^2 z\| \leqslant c_I c_s h_T^2, \quad \|z - I_h z\|_E \leqslant c_I h_E^{\frac{3}{2}} \|\nabla^2 z\|_{P_E} \leqslant c_I c_s h_T^{\frac{3}{2}}.$$

Weiter mit Cauchy-Schwarz:

$$\begin{aligned}
\|e_h\| &\leqslant c_I \sum_{T \in \Omega_h} \left(h_T^2 \|f\|_T \|\nabla^2 z\|_T + \sum_{E \in \partial T} \frac{1}{2} h_E^{\frac{3}{2}} \|[n_E \cdot \nabla u_h]\| \|\nabla^2 z\|_{P_E} \right) \\
&\leqslant c_I c_s c_T \left(\sum_{T \in \Omega_h} \left(h_T^4 \|f\|_T^2 + \sum_{E \in \partial T} \frac{1}{2} h_E^3 \|[n_E \cdot \nabla u_h]\|_E^2 \right) \right)^{\frac{1}{2}} \|\nabla^2 z\|_{L^2(\Omega)}.
\end{aligned}$$

Die Aussage folgt unter Verwendung der Stabilitätsabschätzung für die duale Lösung. □

3.6.2 Der dual gewichtete Fehlerschätzer

Bei technischen Simulationen stellt sich oft die Frage nach einer guten Approximation von speziellen Funktionalwerten. Das kann in der Strukturmechanik etwa die Spannung in einem Punkt sein, in der Strömungsmechanik die Kraft, die auf ein umströmtes Objekt wirkt. Für solche Funktionale ist die Schätzung des Fehlers in globalen Normen nur von geringem Interesse. Die Aubin-Nitsche Trick erlaubt, den Fehler in beliebigen linearen Funktionalen mit Hilfe einer dualen Lösung darzustellen. Mit der Lösung $z \in V$ des dualen Problems

$$(\nabla \phi, \nabla z) = J(\phi),$$

zu einem gegebenen Fehlerfunktional gilt wie im Beweis zu Satz 3.61

$$J(e_h) = R_h(u_h)(z). \tag{3.35}$$

Zur Herleitung von a priori Fehlerschätzern und auch bei der a posteriori-Schätzung des L^2-Fehlers haben wir die duale Lösung z stets mit Hilfe einer Stabilitätsabschätzung gegen die entsprechende Rechte Seite (welche a priori bekannt ist) abgeschätzt. Für allgemeine Funktionale ist dieser Zugang oft nicht möglich.

A posteriori Fehlerschätzer nutzen zur Schätzung des Fehlers die numerische Approximationen $u_h \in V_h$ der Lösung u. Für allgemeine Fehlerfunktionale wollen wir nun auch eine diskrete duale Lösung $z_h \in V_h$ verwenden. Das Duale Problem ist also nicht mehr nur ein mathematisches Hilfskonstrukt, es wird numerische approximiert und die Lösung z_h geht in die Fehlerschätzung ein:

Lemma 3.62 (Dual gewichteter Fehlerschätzer)**.** *Sei $J \in V^*$ ein beschränktes lineares Fehlerfunktional. Für die Finite Elemente Approximation der Poisson-Gleichung gilt die Fehleridentität*

$$J(u-u_h) = \sum_{T\in\Omega_h} \Big\{ \int_T (f+\Delta u_h)\,(z-I_h z)dx - \sum_{E\in\partial T} \frac{1}{2}\int_E [n_E\cdot\nabla u_h]\cdot(z-I_h z)\,ds \Big\}, \quad (3.36)$$

mit der Lösung $z \in V$ des dualen Problems

$$(\nabla\phi,\nabla z) = J(\phi) \quad \forall \phi \in V,$$

sowie die Fehlerabschätzung

$$|J(u-u_h)| \leqslant \sum_{T\in\Omega_h} \eta_T, \quad \eta_T := \rho_T\omega_T + \rho_{\partial T}\omega_{\partial T} \quad (3.37)$$

mit den Zell- *und* Kantenresiduen ρ_T *bzw.* $\rho_{\partial T}$ *sowie den* Zell- *und* Kantengewichten ω_T *und* $\omega_{\partial T}$*:*

$$\rho_T := \|f+\Delta u_h\|_T, \quad \rho_{\partial T} := \frac{1}{2}h_T^{-\frac{1}{2}}\|[n\cdot\nabla u_h]\|_{\partial T}, \quad \omega_T := \|z-I_h z\|_T, \quad \omega_{\partial T} := h_T^{\frac{1}{2}}\|z-I_h z\|_{\partial T}.$$

Proof: Die Fehleridentität folgt mit der Galerkin-Orthogonalität sofort aus (3.35)

$$J(e_h) = (\nabla e_h, \nabla(z-I_h z)) = \sum_{T\in\Omega_h} \Big\{ \int_T (f+\Delta u_h)\,(z-I_h z)\,dx - \int_{\partial T} n_E\cdot\nabla u_h\cdot(z-I_h z)\,ds \Big\},$$

und Übergang zu den Sprungtermen. Abschätzen mit Cauchy-Schwarz liefert

$$|J(u-u_h)| \leqslant \sum_{T\in\Omega_h} \big(\|f+\Delta u_h\|_T\,\|z-I_h z\|_T + \frac{1}{2}\|[n\cdot\nabla u_h]\|_{\partial T}\,\|z-I_h z\|_{\partial T}\big)$$

die Fehlerabschätzung mit den lokalen Fehlerindikatoren. □

Diese Fehleridentität ist noch kein a posteriori Fehlerschätzer in dem Sinne, dass er ohne unbekannte Größen auswertbar ist. Die duale Lösung $z \in V$ ist im Allgemeinen nicht verfügbar. Um den Fehlerschätzer auswerten zu können muss der Interpolationsfehler $z - I_h z$ approximiert werden.

Numerische Approximation Zunächst wäre es naheliegend, die duale Lösung $z_h \in V_h$ durch einen Finite Elemente Ansatz zu diskretisieren

$$(\nabla\phi_h, \nabla z_h) = J(\phi_h) \quad \forall \phi_h \in V_h,$$

und als Approximation $z \approx z_h$ aufzufassen. Wegen $z_h \in V_h$ und Satz 3.55 gilt jedoch

$$R_h(u_h)(z_h) = 0,$$

und auf diese Weise lässt sich keine Fehlerapproximation erstellen. Alternativ kann das duale Problem in einem Raum höherer Ordnung $V_h^{(*)}$ berechnet werden

$$(\nabla\phi_h^*, \nabla z_h^*) = J(\phi_h^*) \quad \forall \phi_h \in V_h^*,$$

welcher echt größer ist als der diskrete Raum V_h. Dann kann $z \approx z_h^*$ approximiert werden und der Fehlerschätzer ist auswertbar. Es gilt:

Lemma 3.63 (Fehlerapproximation höherer Ordnung)**.** *Sei $u_h \in V_h^{(1)}$ die lineare Finite Elemente Approximation der Poisson-Gleichung. Sei $J \in V^*$ ein Fehlerfunktional und durch $z_h^* \in V_h^{(2)}$ die quadratische Finite Elemente Approximation der dualen Lösung. Im Fall $z \in H^3(\Omega)$ ist der a posteriori Fehlerschätzer*

$$\eta_h^* := \sum_{T\in\Omega_h} \left\{ \int_T f(z_h^* - I_h z_h^*)\,dx - \sum_{E\in\partial T} \frac{1}{2} \int_E [n_E \cdot \nabla u_h] \cdot (z_h^* - I_h z_h^*)\,ds \right\}.$$

auf einer Folge von größenregulären Gittern asymptotisch effizient:

$$\frac{|\eta_h^*|}{|J(e_h)|} = 1 + O(h).$$

Proof: (i) Wir wollen zeigen, dass der Fehlerschätzer schneller gegen den Fehler konvergiert $|J(e_h) - \eta_h^*| \to 0$ als der Fehler $J(e_h) \to 0$ selbst. Bei der linearen Finite Elemente Approximation ist die optimale Konvergenzordnung eines linearen beschränkten Funktionals $O(h^2)$.

(ii) Es gilt mit der Fehlerabschätzung aus Satz 3.62:

$$|J(e_h) - \eta_h^*| \leqslant \sum_{T\in\Omega_h} \left\{ \|f\|_T \, \|z - z_h^*\|_T + \sum_{E\in\partial T} \frac{1}{2} \|[n_E \cdot \nabla u_h]\|_E \, \|z - z_h^*\|_E \right\}$$

Auf jeder Kante E nutzen wir das lokale Spur-Lemma:

$$\|v\|_E \leqslant c\big(h^{-\frac{1}{2}} \|v\|_{P_E} + h^{\frac{1}{2}} \|\nabla v\|_{P_E}\big).$$

Dann gilt:

$$\begin{aligned}|J(e_h) - \eta_h^*| \leqslant \sum_{T\in\Omega_h} \{&\|f\|_T \, \|z - z_h\|_T + \\ &ch^{-\frac{1}{2}} \sum_{E\in\partial T} \frac{1}{2}\|[n_E \cdot \nabla u_h]\|_E \left(\|z - z_h^*\|_{P_E} + h\|\nabla(z - z_h^*)\|_{P_E}\right)\Bigg\} \\ \leqslant \|f\|_{L^2(\Omega)} &\, \|z - z_h^*\|_{L^2(\Omega)} \\ &+ c_T c \left(\sum_{E\in\partial T} h^{-1}\|[n_E \cdot \nabla u_h]\|_E^2\right)^{\frac{1}{2}} \left(\|z - z_h^*\|_{L^2(\Omega)} + h\|\nabla(z - z_h^*)\|_{L^2(\Omega)}\right).\end{aligned}$$

Für die quadratische Finite Elemente Approximation der dualen Lösung gelten die a priori Abschätzungen:

$$\|z - z_h^*\| \leqslant ch^3\|\nabla^3 z\|, \quad \|\nabla(z - z_h^*)\| \leqslant ch^2\|\nabla^3 z\|.$$

Mit diesen folgt:

$$|J(e_h) - \eta_h^*| \leqslant c_I h^3 \left\{\|f\|_{L^2(\Omega)} + c_T c h^{-\frac{1}{2}} \left(\sum_{E\in\partial T} \|[n_E \cdot \nabla u_h]\|_E^2\right)^{\frac{1}{2}}\right\} \|\nabla^3 z\|_{L^2(\Omega)}.$$

(*iii*) Es bleibt, die Beschränktheit der Sprünge nachzuweisen. Hierzu betrachten wir eine Kante $E \in \Omega_h$ zwischen $T_1, T_2 \in \Omega_h$. Mit der Schreibweise $\partial_n := n \cdot \nabla$ gilt:

$$[n_E \cdot \nabla u_h] = \partial_{n_E} u_h\big|_{T_1} - \partial_{n_E} u_h\big|_{T_2} = h\frac{\partial_{n_E} u_h\big|_{T_1} - \partial_{n_E} u_h\big|_{T_2}}{h} \approx\approx h\frac{\partial_{n_E} u\big|_{T_1} - \partial_{n_E} u\big|_{T_2}}{h} \approx h\partial_{n_E}^2 u\big|_E.$$

Der Kantensprung über die Normalableitung verhält sich also wie die zweiten Ableitungen der Lösungen. Diese Abschätzung setzt voraus, dass die Ableitungen der diskreten Lösung u_h lokal eine gute Approximation der Ableitung von u sind. Auf regulären Gittern lässt sich eine solche Abschätzung beweisen (Super-Approximation). Zusammen gilt:

$$|J(e_h) - \eta_h^*| \leqslant c_I h^3 \left\{\|f\|_{L^2(\Omega)} + c_T c h^{\frac{1}{2}} \|u\|_{H^2(\Omega)}\right\} \|\nabla^3 z\|.$$

Der Abstand zwischen Fehler und Schätzwert konvergiert mindestens eine Ordnung besser als der Fehler selbst. □

Durch die numerische Approximation des dualen Problems mit einer erhöhten Genauigkeit kann der Fehler asymptotisch effizient geschätzt werden. Dieses Vorgehen ist jedoch im Allgemeinen nicht zu rechtfertigen, bedeutet es doch, dass zum Schätzen des Fehlers ein höherer Aufwand betrieben werden muss als zur Lösung des eigentlichen Problem selbst.

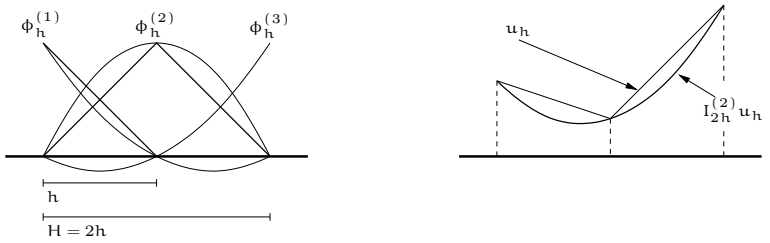


Abbildung 3.9: Lineare und quadratische Basisfunktionen, sowie diskrete Interpolation in den Raum höherer Ordnung.

Numerische Interpolation Eine weitere Idee zur Konstruktion von auswertbaren Fehlerschätzern ist eine nachträgliche Rekonstruktion einer dualen Lösung höherer Ordnung. Zunächst wird $z_h \in V_h$ aus dem gleichen Finite Elemente Raum der Lösung $u_h \in V_h$ berechnet. In einem zweiten Schritt wird die Lösung z_h in einen Raum von höherer Ordnung interpoliert:

$$I_h^* : V_h \to V_h^{(*)}.$$

Hier bietet sich zum Beispiel der Raum $V_{2h}^{(2)}$ an, der Raum der Finiten Elemente vom doppelten Grad auf dem doppelt so groben Gitter. Dieser Raum teil sich die gleichen Knoten wie der Raum V_h und es gilt für die Knotenbasis-Funktionen

$$\phi_h^{(i)}(x_j) = \phi_{2h}^{(2),i}(x_j) = \delta_{ij}.$$

Somit hat die Interpolierende die einfache Darstellung:

$$I_h^* u_h = \sum_{i=1}^{N} \phi_{2h}^{(2),i} \mathbf{u}_i.$$

In Abbildung 3.9 zeigen wir auf einem (eindimensionalen) Gitter einige stückweise lineare Testfunktionen $\phi_h^{(i)}$, sowie in den gleichen Gitterknoten die stückweise quadratischen Testfunktionen auf dem Gitter mit Gitterweite $H = 2h$. Rechts in der Abbildung wird die Interpolation einer diskreten Funktion in diesen Raum höherer Ordnung gezeigt.

Der Fehlerschätzer ist auswertbar als

$$\eta_h^* := \sum_{T \in \Omega_h} \left\{ \int_T f(I_h^* z_h - z_h)\,dx - \sum_{E \in \partial T} \frac{1}{2} \int_E [n_E \cdot \nabla u_h] \cdot (I_h^* z_h - z_h)\,ds \right\}. \tag{3.38}$$

Die Effizienz dieses Fehlerschätzers hängt nun an der Frage, in wie weit $I_h^* z_h$ eine bessere Approximation zu z ist als z_h selbst. Wir benötigen eine Abschätzung der Art:

$$\|z - I_h^* z\| \leqslant ch\|z - z_h\|.$$

Eine Rechtfertigung für eine nachträgliche Verbesserung der Lösung kann wieder durch das Konzept der *Superapproximation* geschehen. Bei gewisser Gitterregularität kann gezeigt

werden, dass die Finite-Elemente Lösung in den Gitterpunkten mit höherer Ordnung konvergiert. Bei linearen Finiten Elementen gilt zum Beispiel falls $f \in C^1(\Omega)$ auf gleichmäßigen Tensorproduktgittern eine Abschätzung der Art

$$|u(x_i) - u_h(x_i)| = o(h^2) \quad \text{für Gitterpunkte } x_i \in \Omega_h.$$

Diese höhere Ordnung kann dann genutzt werden, um über eine Interpolation global bessere Genauigkeit zu erreichen. In der praktischen Anwendung ist der Fehlerschätzer (3.38) höchst erfolgreich. Die Berechnung des dualen Problems ist "billig" und zum Auswerten muss lediglich die diskrete Lösung z_h mit anderen Basisfunktionen dargestellt werden.

Abschätzung der Interpolationsfehler Falls nicht die Fehleridentität (3.36), sondern nur die Abschätzung in Form (3.37) numerisch ausgewertet werden muss, so gilt es, die Residuen

$$\rho_T = \|f + \Delta u_h\|_T, \quad \rho_{\partial t} = \frac{1}{2} h^{-\frac{1}{2}} \|[n \cdot \nabla u_h]\|_{\partial T},$$

und die Gewichte

$$\omega_T = \|z - I_h z\|_T, \quad \omega_{\partial T} = h^{\frac{1}{2}} \|z - I_h z\|_T$$

zu berechnen. Die Residuen können mit der vorhandenen diskreten Lösung $u_h \in V_h$ unmittelbar ausgewertet werden. Bei den Gewichten wird zunächst die Interpolationsabschätzung im Raum $V_h^{(m-1)}$ vom Grad $m-1$ genutzt:

$$\|z - I_h z\|_T + h^{\frac{1}{2}} \|z - I_h z\|_{\partial T} \leqslant c_I h^m \|\nabla^m z\|_{P_T}.$$

Die m-ten Ableitungen von z können durch Differenzenquotienten approximiert werden:

$$\omega_T + \omega_{\partial_T} \leqslant c_I h^m \|\nabla^m z\|_T \approx c_I h^m |T|^{\frac{1}{2}} |\nabla_h^m z_h|_{T,\infty}.$$

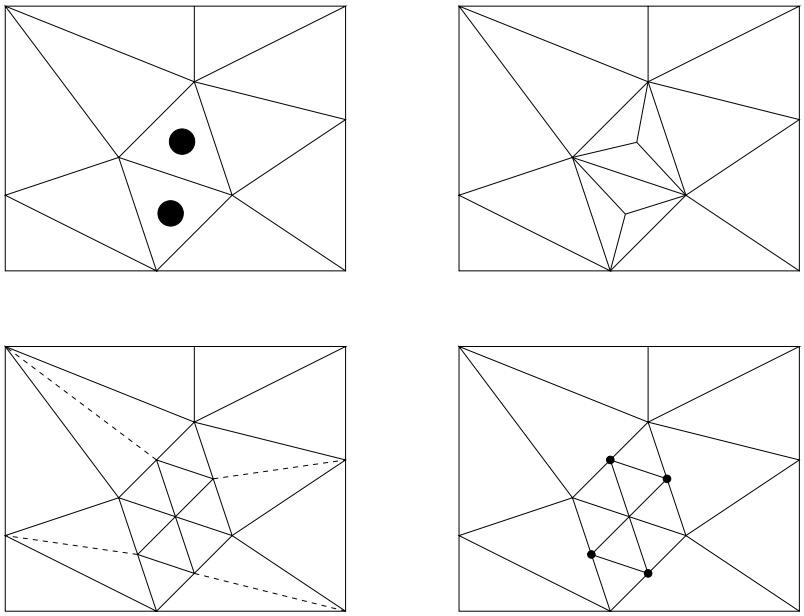


Abbildung 3.10: Drei verschiedene Methoden zur Verfeinerung des Gitters. Rechts oben: Verfeinerung mit neuen Inneren Knoten. Links unten: Verfeinerung mit Hilfe von *Anschlusselementen*. Rechts unten: Verfeinerung mit Hilfe von *hängenden Knoten*. Hier entstehen vier hängende Knoten.

3.6.3 Adaptive Gitterverfeinerung

Bei einem adaptiven Finite Elemente Verfahren werden die Fehlerindikatoren genutzt, um mit Informationen über die lokale Verteilung der Fehler die Triangulierung anzupassen und um dort die Diskretisierungsgenauigkeit zu erhöhen, wo der Fehler entsteht. Hierzu gibt es zwei alternative Optionen. Beim *Remeshing* wird mit Hilfe der Fehlerindikatoren ein komplett neues Gitter erzeugt. Hierzu wird zunächst eine *Dichtefunktion* $H(x)$ erzeugt, welche angibt, welche Gitterweite in welchem Bereich des Gebiets realisiert werden soll. Im Anschluss wird ein neues Gitter mit einem *Gittergenerator* erzeugt.

Wir betrachten hier ausschließlich die *Gitterverfeinerung*. Bei dieser Methode wird mit Hilfe der Fehlerindikatoren $\{\eta_T\}_{T\in\Omega_h}$ das Gitter Ω_h zu einem neuen Gitter Ω_h' verändert. Dabei können entweder Elemente von Ω_h zusammengefasst werden, wenn ihr Indikator-Beitrag zu dem Gesamtfehler zu vernachlässigen ist, oder aber Gitter-Elemente $T \in \Omega_h$ werden verfeinert, also in kleinere Elemente aufgeteilt.

In Abbildung 3.10 zeigen wir einige Methoden zur Verfeinerung von Dreiecks-Gittern. Bei der Gitterverfeinerung ist darauf zu achten, dass die Sequenz von Gittern Ω_h für $h \to 0$ immer noch gewissen Regularitätsbedingungen genügt:

Größenregularität Die Größenregularität, also die Forderung

$$\max_{T\in\Omega_h} h_T \leqslant c \min_{T\in\Omega_h} h_T,$$

kann sicher nicht mehr beibehalten werden. Denn das Ziel der lokalen Gitterverfeinerung ist es gerade, lokal angepasste Diskretisierungen und somit Elemente zu Verwenden. Die üblichen a priori Abschätzungen für den Fehler und die Interpolation lauten z.B.:

$$\|\nabla(u - u_h)\| \leqslant c h_{max}\|f\|,$$

mit der maximalen Gitterweite h_{max}. Abschätzungen dieser Art verlieren bei adaptiven Finiten Elementen ihre Aussagekraft, da der Gesamtfehler nicht von der maximalen Gitterweite abhängt, sondern von der optimalen Verteilung.

Formregularität Die Formregularität ist wesentlich für die Herleitung von lokalen Interpolationsabschätzungen und darf nicht durch lokale Verfeinerung verletzt werden. Bei der Aufteilung eines Dreiecks in kleinere Dreiecke muss darauf geachtet werden, dass die Innenwinkel weiterhin nicht gegen 0 oder 180 Grad gehen.

Strukturregularität In Abbildung 3.10 ist oben rechts eine Methode der Verfeinerung dargestellt, welche die Struktur-Regularität erhält. Bei dieser Verfeinerung degenerieren allerdings die Dreiecke und verletzen die Formregularität. Bei den beiden Methoden in der unteren Zeile wird die Formregularität gewahrt, die vier neuen Dreiecke haben die gleichen Winkel wie das große. Durch das Einführen von zusätzlichen Knoten auf den Eckpunkten wird jedoch die Strukturregularität verletzt.

Unten links wird die Verwendung von sogenannten *Anschlusselementen* gezeigt. Die angrenzenden Elemente werden verfeinert, diese Verfeinerung wird jedoch zurückgenommen, falls diese Elemente selbst verfeinert werden sollen. In Abbildung 3.11 werden zwei Verfeinerungsschritte bei der Verwendung von Anschlusselementen gezeigt. Im zweiten Schritt werden Anschlusselemente aufgelöst, bzw. modifiziert.

Bei der Methode unten rechts bleibt die Strukturregularität verletzt. Die neuen Knoten auf den Ecken sind sogenannte *hängende Knoten*. Diese Knoten sind nicht echte Freiheitsgrade der Triangulierung sondern werden durch den Mittelwert der beiden angrenzenden Eckknoten ersetzt.

Ziel der Gittersteuerung ist es mit Hilfe von auswertbaren Fehlerindikatoren das Gitter, also die Diskretisierung, zu verfeinern, so dass der Gesamtfehler möglichst effizient reduziert wird. Wir wählen als Beispiel die Darstellung (3.37)

$$\eta_h := \sum_{T\in\Omega_h} \eta_T, \quad \eta_T := \rho_T \omega_T,$$

welche bereits in lokalisierter Form vorliegt. Verfahren zur Gittersteuerung müssen nun solche Elemente $T \in \Omega_h$ des Gitters wählen, welche einen großen Beitrag zum Gesamtfehler haben.

Folgende Leitideen liegen jeder Gitteradaption zu Grunde:

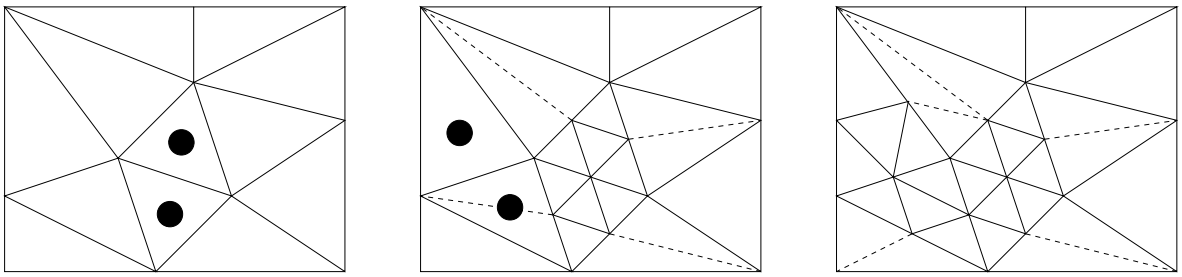


Abbildung 3.11: Verfeinerung bei Verwendung von Anschlusselementen.

1. Wenn ein Element $T \in \Omega_h$ verfeinert wird, so werden auch alle Elemente mit $T' \in \Omega_h$ mit einem größeren Indikatorwert $\eta_{T'} > \eta_T$ verfeinert.
2. Es wird versucht, ein Gitter mit *ausbalancierten Indikatoren* zu erreichen:

$$\eta_T \approx \eta_{T'} \quad \forall T, T', \in \Omega_h.$$

3. Wenn die Fehlerindikatoren balanciert sind, so wird global, also das ganze Gitter verfeinert.

Im Folgenden stellen wir einige Methoden zur Gitterverfeinerung vor. Dazu seien η_i für $i = 1, \ldots, N$ die Fehlerindikatoren absteigend sortiert, also mit $\eta_i \geqslant \eta_{i+1}$:

Fixed Number Es werden die $p\%$ der Elemente mit höchsten Fehlerindikatoren verfeinert.

Fixed Fraction Es werden diejenigen Elemente mit höchsten Fehlerindikatoren verfeinert, die zusammen $p\%$ des Gesamtfehlers ausmachen.

Balancierung Es werden alle Elemente Verfeinert, deren Fehlerindikator über dem Mittelwert liegt.

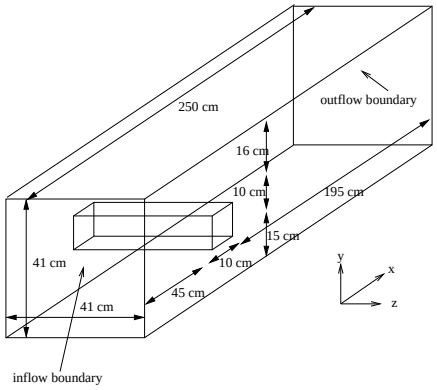


Abbildung 3.12: Umströmung eines Hindernis.

Ein numerisches Beispiel In einem Kanal $\Omega \subset \mathbb{R}^3$ soll die Strömung um ein Objekt mit Rand Γ_o berechnet werden. Die Konfiguration ist in Abbildung 3.12 dargestellt. Ziel der Berechnung ist es, den Strömungswiderstand des Objektes zu Berechnen. Dieser ist durch ein lineares stetiges Funktional gegeben:

$$J(\nu, p) = \int_{\Gamma_o} \nu \vec{n} \cdot \nabla \vec{v} \cdot \vec{e}_1 - n p \cdot \vec{e}_1 \, ds,$$

wobei $\vec{v}$ die Geschwindigkeit der Strömung und p der Druck ist. Auf dem Rand Γ_o des Hindernis ist $\vec{n}$ der Normalvektor und $\vec{e}_1 = (1, 0, 0)$ ist der Einheitsvektor in Strömungsrichtung. Der exakte Wert des Fehlerfunktionals, also $J(u)$ ist durch Vergleichsrechnungen bekannt $J(u) \approx 7.767$.

Zunächst wird diese numerische Simulation mit stückweise quadratischen Finiten Elementen unter Verwendung von gleichmäßigen Gittern durchgeführt. Die Ergebnisse sind in Tabelle 3.1 zusammengefasst. Um eine Fehlertoleranz von 1% einzuhalten muss bereits ein Problem mit über $1\,000\,000$ Freiheitsgraden gelöst werden. D.h., wir müssen lineare Gleichungssysteme der Dimension $A \in \mathbb{R}^{1\,000\,000 \times 1\,000\,000}$ lösen!

Im Anschluss wird die gleiche Berechnung mit Hilfe von lokal verfeinerten Gittern wiederholt. Als Fehlerschätzer kommt der dual gewichtete Fehlerschätzer zum Einsatz. Das duale Problem wird im gleichen Ansatzraum approximiert wie das eigentliche Problem, also $u_h \in V_h$ und $z_h \in V_h$. Zur Approximation der Fehleridentität wird die oben beschriebene Interpolation von höherer Ordnung verwendet. In Tabelle 3.2 fassen wir die Ergebnisse zusammen.

Bei der Verwendung von adaptiven Finiten Elementen auf lokal verfeinerten Gittern wird ein relativer Fehler von 1% bereits mit $85\,000$ Freiheitsgraden erreicht. Auf global verfeinerten Gittern ist mit $1\,300\,000$ die 15 fache Anzahl von Freiheitsgraden notwendig. Wird eine Fehlertoleranz von unter 0.1% angestrebt, so ist die Ersparnis sogar ein Faktor 150. In Abbil-

Gitterzellen	Freiheitsgrade	Funktionalwert	Fehler	relativer Fehler
78	3 696	13.3149	5.5479	71.4%
624	24 544	8.0450	0.2780	3.58%
4 992	177 600	7.9759	0.2089	2.69%
39 936	1 348 480	7.7878	0.0208	0.27%
319 488	10 787 840	7.7579	0.0091	0.12%
2 255 904	86 302 720	7.7612	0.0058	0.07%

Tabelle 3.1: Umströmung eines Hindernis und Widerstandsberechnung mit quadratischen Finiten Elementen unter Verwendung von uniform verfeinerten Gittern.

Gitterzellen	Freiheitsgrade	Funktionalwert	Fehler	relativer Fehler
78	3 696	13.3149	5.5479	71.4%
624	24 544	8.0450	0.2780	3.58%
2 427	84 945	7.7942	0.0272	0.35%
7 120	242 080	7.7881	0.0211	0.27%
16 808	571 472	7.7595	0.0075	0.10%
54 880	1 811 040	7.7620	0.0050	0.06%

Tabelle 3.2: Widerstandsberechnung mit quadratischen Finiten Elementen. Lokal verfeinerte Gitter mit dem dual gewichteten Fehlerschätzer.

dung 3.13 zeigen wir einige Gitter und Ausschnitte von Gittern aus den Berechnungen mit lokal verfeinerten Elementen.

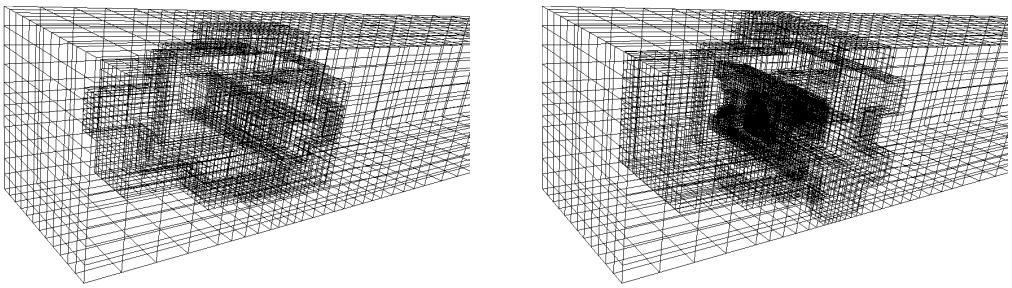


Abbildung 3.13: Gitterausschnitte aus der dreidimensionalen Umströmung eines Hindernis.

4 Solution of the linear systems

Das Bestimmen der Finite Elemente Approximation $u_h \in V_h$ einer linearen Randwertaufgabe erfordert das Lösen eines linearen Gleichungssystems

$$\boldsymbol{A}_h \mathbf{u}_h = \mathbf{b}_h.$$

In diesem Kapitel befassen wir uns mit iterativen Approximationsverfahren für große lineare Gleichungssysteme, welche von der Finite Elemente Diskretisierung einer partiellen Differentialgleichung stammen. Wir beschränken uns dabei wieder im Wesentlichen auf das Dirichlet-Problem des Laplace-Operators:

$$-\Delta u = f \quad \text{in } \Omega \subset \mathbb{R}^2, \quad u = 0 \text{ auf } \partial\Omega.$$

Bei Bedarf gehen wir auf andere, z.B. nicht-symmetrische Gleichungen, oder auch die (einfache) Erweiterung auf dreidimensionale Probleme ein.

Die Matrix $A \in \mathbb{R}^{N \times N}$ erbt die Eigenschaften des Differentialoperators, d.h., falls der Operator symmetrisch positiv definit ist, so gilt dieses auch für die Matrix. Üblicherweise ist $N \gg 1000$, oft $N \gg 1\,000\,000$. Dann scheiden direkte Verfahren, wie die LR- oder die Cholesky-Zerlegung als Lösungsmethoden aus.

4.1 Eigenschaften der linearen Gleichungssysteme

Das lineare Gleichungssystem

$$\boldsymbol{A}\mathbf{u} = \mathbf{b}$$

stamme von der Finite Elemente Diskretisierung der Poisson-Gleichung. Wir wissen, dass die Matrix $\boldsymbol{A}$ symmetrisch und positiv definit, also insbesondere auch regulär ist. Für die Kondition der Systemmatrix $\boldsymbol{A}$ gilt

$$\operatorname{cond}_2(\boldsymbol{A}) = O(h^{-2}).$$

Hierbei ist das quadratische Anwachsen der Kondition durch die zweite Ordnung des betrachteten Laplace-Operators bedingt und gilt für beliebige Finite Elemente Ansatzräume. Aufgrund des Konstruktionsprinzip von Finite Elemente Ansätzen sind die Matrizen dünn besetzt, denn z.B. für Lagrange-Ansätze gilt:

$$\operatorname{supp}(\phi_h^{(i)}) \cap \operatorname{supp}(\phi_h^{(j)}) \neq \emptyset \quad \Leftrightarrow \quad \exists T \in \Omega_h,\ x_i, x_j \in \bar{T}.$$

Die genaue Anzahl von Matrix-Einträgen pro Matrixzeile hängt jedoch stark vom jeweiligen Finite Elemente Ansatz und auch von der zugrundeliegenden Triangulierung ab.

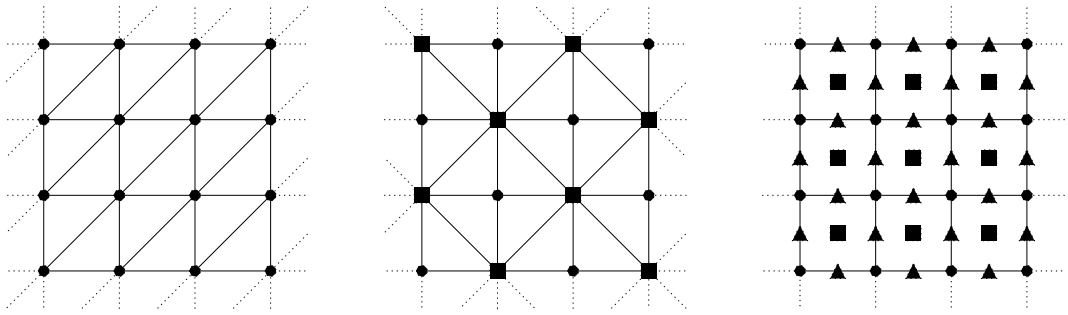


Abbildung 4.1: Ausschnitt von zwei uniformen Dreiecksgittern und jeweilige Knotenfunktionale der linearen Finiten Elemente Diskretisierung, sowie uniformes Viereckgitter mit den Knotenfunktionalen der biquadratischen Finiten Elemente.

Example 4.1 (Matrix-Struktur)**.** *Zunächst betrachten wir die Diskretisierung der Poisson-Gleichung mit linearen Finiten Elementen auf einem gleichmäßigen Dreiecksgitter wie in Abbildung 4.1 links. Auf jeder Zelle* T *liegen drei Knotenpunkte* x_i*, das heißt pro Zelle überschneiden sich drei Basisfunktionen. In jedem Knoten* x_i *kommen auf diesem Gitter sechs Eckpunkte zusammen. Jede Matrix-Zeile (die* j*-te Zeile beinhaltet jeweils die Kopplungen der Testfunktion* $\phi_h^{(j)}$ *zu allen anderen Testfunktionen) hat neben dem Diagonaleintrag noch sechs Nebendiagonaleinträge.*

In der mittleren Abbildung ist ein weiteres, auch sehr regelmäßiges Dreiecksgitter gezeigt. Hier gibt es Knoten x_i*, welche Eckpunkte (die rund markierten) von vier Dreiecken sind, sowie Knoten (die eckigen), in welchen 8 Dreiecke zusammenkommen. Neben der Nebendiagonale gibt es also noch vier oder noch 8 weitere Matrixeinträge pro Zeile.*

In der Abbildung rechts wird ein uniformes Gitter mit den Knotenpunkten x_i *der biquadratischen Finiten Elemente gezeigt. Hier muss die Analyse der Matrixeinträge je nach Typ der Knotenfunktionale erfolgen, ob die Punkte im Innern (Viereck), in den Eckpunkten (Kreis) oder auf den Kanten (Dreieck) liegen. In jedem Elemente* T *überschneiden sich die Träger von* 9 *Basisfunktionen. Die Matrixzeile, welche zu einem Knoten im Innern eines Elementes gehört hat demnach inklusive Diagonalelement* 9 *Einträge, da für die zugehörige Basisfunktion gilt* $\operatorname{supp}(\phi_h^{(i)}) = T$*. Basisfunktionen zu Punkten auf den Kanten haben einen Träger auf genau zwei Vierecken und erzeugen somit 15 Matrixeinträge. Für Knotenpunkte auf einer Ecke entstehen 25 Matrixeinträge.*

Die betrachteten Beispiele zeigen, dass auch bei einfachen Ansätzen und gleichmäßigen Gittern keine einheitliche Matrixstruktur erwartet werden kann. Vielmehr muss davon ausgegangen sein, dass die Matrix generell unstrukturiert ist und sich von Zeile zu Zeile ändern kann. Dies ist insbesondere der Fall, wenn allgemeine Gitter zugelassen werden, in denen sich pro Eckknoten unterschiedlich viele Elemente treffen.

Das "Aussehen" der Matrix $\boldsymbol{A}$ hängt zusätzlich noch von der gewählten Nummerierung der Freiheitsgrade im Gitter ab. Angenommen das Gitter habe $N = M^2$ Knoten, mit M Knoten in jeder Raumrichtung. Werden z.B. die Freiheitsgrade im Gitter links von Abbildung 4.1

lexikographisch, also von unten links, nach oben rechts nummeriert ergibt sich eine Block-Bandmatrix mit der Struktur:

$$\mathbf{A} = \left.\begin{pmatrix} B & -I & & & & \\ -I & B & -I & & & \\ & -I & B & -I & & \\ & & \ddots & \ddots & \ddots & \\ & & & -I & B & -I \\ & & & & -I & B \end{pmatrix}\right\} M, \quad B = \left.\begin{pmatrix} 4 & -1 & & & & \\ -1 & 4 & -1 & & & \\ & -1 & 4 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 4 & -1 \\ & & & & -1 & 4 \end{pmatrix}\right\} M,$$

Man beachte, dass jede Matrixzeile nur 5 Einträge ungleich Null hat, obwohl nach obiger Diskussion 7 Einträge entstehen sollten. Die beiden Einträge zur jeweils Diagonalen Kopplung verschwinden jedoch auf diesem gleichmäßigen Gitter aus Symmetriegründen und sind im Allgemeinen vorhanden. Auf allgemeinen Gittern, oder bei anderer Nummerierung der Freiheitsgrade im Gitter ist üblicherweise keine Bandstruktur gegeben. Zum effizienten Speichern dieser dünn besetzten Matrizen sind daher spezielle Speicherstrukturen notwendig.

4.2 Krylow-Raum-Methoden

In diesem Abschnitt diskutieren wir sogenannte *Krylow-Raum-Methoden*. Hierzu zählt insbesondere das *Verfahren der konjugierten Gradienten*, *CG-Verfahren* genannt. Für die weitere Diskussion sei also durch

$$\mathbf{A}\mathbf{u} = \mathbf{b}, \tag{4.1}$$

das lineare Gleichungssystem mit einer positiv definiten symmetrischen (und dünn besetzten) Matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ gegeben. Weiter sei durch $\langle\cdot,\cdot\rangle$ das euklidische Skalarprodukt im $\mathbb{R}^N$ und durch $|\cdot|$ die entsprechende euklidische Vektornorm gegeben. Es gilt:

Lemma 4.2 (Minimierungsaufgabe)**.** *Die Lösung des linearen Gleichungssystems* (*4.1*) *ist unter den gegebenen Voraussetzungen an die Matrix* $\mathbf{A}$ *äquivalent zur Lösung einer quadratischen Minimierungsaufgabe:*

$$\mathbf{A}\mathbf{u} = \mathbf{b} \quad \Leftrightarrow \quad Q(\mathbf{u}) = \min_{\mathbf{x}\in\mathbb{R}^N} Q(\mathbf{x}), \quad Q(\mathbf{x}) := \frac{1}{2}\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle - \langle \mathbf{b}, \mathbf{x}\rangle \tag{4.2}$$

Proof: Es gilt für die notwendige Bedingung an das Minimum:

$$\nabla Q(\mathbf{x}) = \frac{1}{2}(\mathbf{A} + \mathbf{A}^T)\mathbf{x} - \mathbf{b} = \mathbf{A}\mathbf{x} - \mathbf{b}.$$

Für die zweiten Ableitungen der quadratischen Form $Q(\cdot)$ gilt:

$$\nabla^2 Q(\mathbf{x}) = \mathbf{A},$$

und aus der positiven Definitheit der Matrix $\mathbf{A}$ folgt, dass ein eindeutig bestimmtes Minimum von (4.2) existiert, welches Lösung von (4.1) ist. □

Der Gradient der Form $Q(\cdot)$ in einem Punkt $\mathbf{x}$ ist gerade das (negative) Residuum der linearen Gleichung:

$$Q(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b} =: -r(\mathbf{x}).$$

Für das weitere Vorgehen verwenden wir:

Lemma 4.3 (A-Norm, A-Skalarprodukt)**.** *Es sei $\mathbf{A} \in \mathbb{R}^{N\times N}$ eine symmetrisch positiv definite Matrix. Dann sind durch:*

$$\langle \mathbf{x}, \mathbf{y}\rangle_A := \langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle, \quad |\mathbf{x}|_A := \langle \mathbf{x}, \mathbf{x}\rangle_A^{\frac{1}{2}},$$

ein Skalarprodukt, bzw. eine Norm definiert. Man nennt diese Norm die diskrete Energie-Norm*.*

Proof: Übungsaufgabe. □

Sei $\mathbf{x}$ die Lösung von $\mathbf{A}\mathbf{x} = \mathbf{b}$. Dann gilt für das Minimum $Q(\mathbf{x})$ auch:

$$\begin{aligned}
2Q(\mathbf{x}) &= \langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle - 2\langle \mathbf{b}, \mathbf{x}\rangle \\
&= \langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle - \langle \mathbf{b}, \mathbf{x}\rangle - \langle \mathbf{A}\mathbf{x}, \mathbf{A}^{-1}\mathbf{b}\rangle + \langle \mathbf{b}, \mathbf{A}^{-1}\mathbf{b}\rangle - \langle \mathbf{b}, \mathbf{A}^{-1}\mathbf{b}\rangle \\
&= \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{x} - \mathbf{A}^{-1}\mathbf{b}\rangle - \langle \mathbf{b}, \mathbf{A}^{-1}\mathbf{b}\rangle \\
&= \langle \mathbf{A}^{-1}(\mathbf{A}\mathbf{x} - \mathbf{b}), \mathbf{A}\mathbf{x} - \mathbf{b}\rangle - \langle \mathbf{b}, \mathbf{A}^{-1}\mathbf{b}\rangle = |\mathbf{A}\mathbf{x} - \mathbf{b}|_{A^{-1}} - |\mathbf{b}|_{A^{-1}} \\
&= \langle \mathbf{x} - \mathbf{A}^{-1}\mathbf{b}, \mathbf{A}(\mathbf{x} - \mathbf{A}^{-1}\mathbf{b})\rangle - \langle \mathbf{A}\mathbf{A}^{-1}\mathbf{b}, \mathbf{A}^{-1}\mathbf{b}\rangle = |\mathbf{x} - \mathbf{u}|_A - |\mathbf{u}|_A.
\end{aligned}$$

Wir bekommen für die Lösung des linearen Gleichungssystems zwei weitere Charakterisierungen: einerseits ist die Minimierung von $Q(\cdot)$ äquivalent zur Minimierung der Defektnorm $|\mathbf{A}\mathbf{x} - \mathbf{b}|_{A^{-1}}$ andererseits zur Minimierung der Energienorm $|\mathbf{x} - \mathbf{u}|_A$. Krylow-Raum Methoden basieren nun auf einer schrittweisen Approximation des Minimierungsproblems in einer der gegebenen Formulierungen.

4.2.1 Abstiegsverfahren

Wir beschreiben zunächst das einfache *Abstiegsverfahren*. Ausgehend von einem *Startwert* $\mathbf{x}^{(0)} \in \mathbb{R}^N$ werden Abstiegsrichtungen $\mathbf{d}^{(i)}$ gewählt. Dann wird eine Folge von Iterierten durch die Vorschrift

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \alpha_t \mathbf{d}^{(t)}, \quad t \geqslant 0,$$

bestimmt. Dabei ist der Parameter $\alpha_t \in \mathbb{R}$ die Lösung des (eindimensionalen) Minimierungsproblems:

$$Q(\mathbf{x}^{(t+1)}) = \min_{\alpha\in\mathbb{R}} Q(\mathbf{x}^{(t)} + \alpha\mathbf{d}^{(t)}).$$

Die Optimalitätsbedingung an α ergibt:

$$\frac{d}{d\alpha} Q(\mathbf{x}^{(t)} + \alpha\mathbf{d}^{(t)}) = \nabla Q(\mathbf{x}^{(t)} + \alpha\mathbf{d}^{(t)})\mathbf{d}^{(t)} = \langle \mathbf{A}\mathbf{x}^{(t)} - \mathbf{b}, \mathbf{d}^{(t)}\rangle + \alpha\langle \mathbf{A}\mathbf{d}^{(t)}, \mathbf{d}^{(t)}\rangle = 0,$$

Mit dem Residuum $\mathbf{r}^{(t)} := \mathbf{b} - \mathbf{A}\mathbf{x}^{(t)}$ ergibt dies:

$$\alpha_t = \frac{\langle \mathbf{r}^{(t)}, \mathbf{d}^{(t)} \rangle}{\langle \mathbf{A}\mathbf{d}^{(t)}, \mathbf{d}^{(t)} \rangle}$$

Wir fassen zusammen

Algorithmus 4.4 (Abstiegsverfahren). *Sei $\mathbf{x}^{(0)} \in \mathbb{R}^N$ ein gegebener Startvektor, $\mathbf{d}^{(t)}$ Abstiegsrichtungen. Iteriere für $t \geqslant 0$:*

1. $\mathbf{r}^{(t)} := \mathbf{b} - \mathbf{A}\mathbf{x}^{(t)}$
2. $\alpha_t = \frac{\langle \mathbf{r}^{(t)}, \mathbf{d}^{(t)} \rangle}{\langle \mathbf{A}\mathbf{d}^{(t)}, \mathbf{d}^{(t)} \rangle}$
3. $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \alpha_t \mathbf{d}^{(t)}$

Ein konkretes Abstiegsverfahren erreicht man nun durch Wahl von entsprechenden Abstiegsrichtungen $\mathbf{d}^{(t)}$. Die einfachste Möglichkeit ist die Wahl $\mathbf{d}^{(t)} := e^{(t)}$, wobei $e^{(t)}$ der t-te kartesische Einheitsvektor ist. Es kann gezeigt werden, dass ein kompletter Durchlauf des Abstiegsverfahren mit diesen Abstiegsrichtungen dem Gauß-Seidel-Verfahren entspricht.

Der Gradient $-\nabla Q(x^{(t)}) = \mathbf{r}^{(t)}$ gibt die Richtung des stärksten Abfalls an und ist die Basis des sogenannten *Gradientenverfahrens*.

Lemma 4.5 (Direction of steepest descent). *Let $x \in \mathbb{R}^n$. The direction of the steepest descent in $Q(\cdot)$ is given by the negative gradient*

$$d = -\nabla Q(x).$$

Proof. We are looking for the direction $d \in \mathbb{R}^n$, such that

$$N(d) := \frac{d}{ds} \frac{Q(x + sd)}{\|d\|}\Big|_{s=0}$$

gets minimal. We assume, that $\|d\| = 1$. It holds

$$N(d)\Big|_{\|d\|=1} = \langle \mathbf{A}x - b, d \rangle,$$

which is minimized for

$$d = \frac{1}{\|b - \mathbf{A}x\|}(b - \mathbf{A}x).$$

□

Algorithmus 4.6 (Gradientenverfahren). *Sei $\mathbf{x}^{(0)} \in \mathbb{R}^N$ ein gegebener Startvektor. Iteriere für $t \geqslant 0$:*

1. $\mathbf{r}^{(t)} := \mathbf{b} - \mathbf{A}\mathbf{x}^{(t)}$

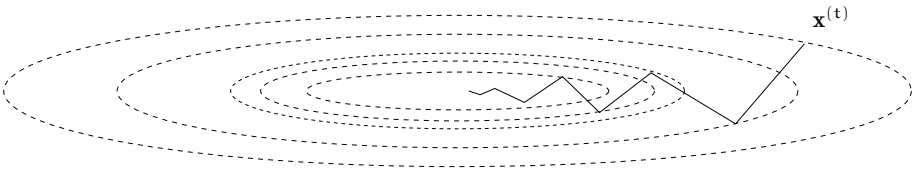


Abbildung 4.2: Niveaulinien des Gradientenverfahrens.

2. $\alpha_t = \frac{|\mathbf{r}^{(t)}|^2}{|\mathbf{r}^{(t)}|_A^2}$
3. $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \alpha_t \mathbf{r}^{(t)}$

Dieses sehr einfache Verfahren konvergiert im Allgemeinen jedoch nur äußerst langsam, etwa wie die einfache Richardson-Iteration. Es gilt eine Orthogonalitätsbeziehung zwischen aufeinander folgenden Abstiegsrichtungen:

Lemma 4.7 (Orthogonalität im Gradientenverfahren)**.** *Je zwei aufeinander folgende Abstiegsrichtungen im Gradientenverfahren stehen orthogonal aufeinander:*

$$\langle \mathbf{r}^{(t)}, \mathbf{r}^{(t+1)} \rangle = 0 \quad t \geqslant 0.$$

Proof: Sei $x^{(t)}$ gegeben. Es gilt:

$$\mathbf{r}^{(t)} := \mathbf{b} - \mathbf{A}\mathbf{x}^{(t)}, \quad \mathbf{r}^{(t+1)} := \mathbf{b} - \mathbf{A}\mathbf{x}^{(t+1)},$$

sowie

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \frac{|\mathbf{r}^{(t)}|^2}{|\mathbf{r}^{(t)}|_A^2}\mathbf{r}^{(t)} \quad \Rightarrow \quad \mathbf{r}^{(t+1)} = \mathbf{r}^{(t)} - \frac{|\mathbf{r}^{(t)}|^2}{|\mathbf{r}^{(t)}|_A^2}\mathbf{A}\mathbf{r}^{(t)}.$$

Dann gilt:

$$\langle \mathbf{r}^{(t)}, \mathbf{r}^{(t+1)} \rangle = |\mathbf{r}^{(t)}|^2 - \frac{|\mathbf{r}^{(t)}|^2}{|\mathbf{r}^{(t)}|_A^2} \langle \mathbf{A}\mathbf{r}^{(t)}, \mathbf{r}^{(t)} \rangle = 0.$$

□

Im Allgemeinen müssen die Abstiegsrichtungen jedoch nicht orthogonal aufeinander stehen, d.h. $\langle \mathbf{r}^{(t)}, \mathbf{r}^{(t+2)} \rangle = 0$ gilt im Allgemeinen nicht (und zwar nicht einmal annähernd). Dies führt in der Anwendung zu einem stark oszillierenden Konvergenzverhalten, insbesondere, wenn die Matrix $\mathbf{A}$ stark unterschiedliche Eigenwerte hat.

Example 4.8 (Konvergenz der Gradientenverfahrens)**.** *Wir betrachten das lineare Gleichungssystem*

$$\mathbf{A}\mathbf{x} = \mathbf{b},$$

mit der Matrix und rechten Seite

$$\mathbf{A} := \begin{pmatrix} 16 & 0 \\ 0 & 2 \end{pmatrix}, \quad \mathbf{b} := \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

mit den Eigenwerten 16 *und* 2. *Die Lösung dieses Gleichungssystems ist äquivalent zur Minimierung des quadratischen Funktionals*

$$Q(\mathbf{x}) = 8\mathbf{x}_1^2 + \mathbf{x}_2^2 - \mathbf{x}_1 - \mathbf{x}_2.$$

Die Niveaulinien $Q(\mathbf{x}) = c$ *dieser quadratischen Form sind stark elliptisch, siehe Abbildung* (4.2). *Angenommen,* $\mathbf{x}^{(t)}$ *sein eine gegebene Approximation. Das Residuum* $r(\mathbf{x}^{(t)})$ *als neue Abstiegsrichtung steht orthogonal auf der Durch* $\mathbf{x}^{(t)}$ *laufenden Niveaulinie. Jeweils zwei Richtungen stehen orthogonal aufeinander,* $\mathbf{r}^{(t)}$ *und* $\mathbf{r}^{(t+2)}$ *sind hingegen fast parallel. Dies führt zu sehr langsamen, ineffizienten Konvergenzverhalten.*

4.2.2 Das Verfahren der konjugierten Gradienten (CG-Verfahren)

Das CG-Verfahren versucht nun die Struktur der Matrix und der quadratischen Form besser auszunutzen: die Abstiegsrichtungen $\mathbf{d}^{(t)}$ sollen so gewählt werden, dass sie alle aufeinander orthogonal stehen.

Anstelle der Lösung eines eindimensionalen Optimierungsproblems in jedem Schritt des Abstiegsverfahren suchen wir nun die neue Lösung in der Form

$$\mathbf{x}^{(t)} \in \mathbf{x}^{(0)} + K_t, \quad Q(\mathbf{x}^{(t)}) = \min_{\mathbf{x} \in \mathbf{x}^{(0)} + K_t} Q(\mathbf{x}), \quad Q(\mathbf{x}) := \frac{1}{2}\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle - \langle \mathbf{b}, \mathbf{x}\rangle, \tag{4.3}$$

mit dem Raum, welcher durch die Abstiegsrichtungen aufgespannt wird:

$$K_t := \operatorname{span}\{\mathbf{d}^{(0)}, \ldots, \mathbf{d}^{(t-1)}\}.$$

Ein solches Verfahren terminiert nach N Schritten zwangsläufig und liefert die gesuchte Lösung. D.h., wir können das Verfahren als ein *direktes Verfahren* interpretieren. Im Allgemeinen werden wir jedoch nur mit einer kleinen Zahl von Schritten eine Näherungslösung erstellen.

Es gilt:

Lemma 4.9 (Galerkin-Gleichung)**.** *Die Lösung* $\mathbf{x}^{(t)} \in \mathbf{x}^{(0)} + K_t$ *der Minimierungsaufgabe* (4.3) *ist eindeutig durch die Galerkin-Gleichung beschrieben:*

$$\langle \mathbf{b} - \mathbf{A}\mathbf{x}^{(t)}, \mathbf{y}\rangle = 0 \quad \forall \mathbf{y} \in K_t. \tag{4.4}$$

Proof: Übung. □

Die Lösung $\mathbf{x}^{(t)}$ stellen wir nun als Entwicklung in den Abstiegsrichtungen dar:

$$\mathbf{x}^{(t)} := \mathbf{x}^{(0)} + \sum_{i=0}^{(t-1)} \alpha_i \mathbf{d}^{(i)}.$$

Wir setzen diese Entwicklung in die Galerkin-Gleichung (4.4) ein und erhalten als bestimmendes Gleichungssystem

$$\sum_{i=0}^{t-1} \alpha_i \langle \mathbf{A}\mathbf{d}^{(i)}, \mathbf{d}^{(j)} \rangle = \langle \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}, \mathbf{d}^{(j)} \rangle.$$

Wenn es uns nun gelingt, eine $\mathbf{A}$-orthogonale Basis der Abstiegsrichtungen $\{\mathbf{d}^{(0)}, \ldots, \mathbf{d}^{(t-1)}\}$ zu erzeugen, so kann dieses lineares Gleichungssystem trivial gelöst werden.

Für die Ansatzräume K_t wählen wir sogenannte

Definition 4.10 (Krylow-Räume)**.** *Sei $\mathbf{A} \in \mathbb{R}^{N\times N}$ eine Matrix, sowie $\mathbf{r}^{(0)} \in \mathbb{R}^N$ das Residuum zum Startwert $\mathbf{r}^{(0)} := \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$. Wir definieren den* Krylow-Raum

$$K_t(\mathbf{r}^{(0)}; \mathbf{A}) := \mathit{span}\{\mathbf{r}^{(0)}, \mathbf{A}\mathbf{r}^{(0)}, \ldots, \mathbf{A}^{t-1}\mathbf{r}^{(0)}\},$$

und schreiben kurz $K_t := K_t(\mathbf{r}^{(0)}, \mathbf{A})$.

Remark 4.11. *Wir betrachten den Fall $\mathbf{x}^{(0)} = 0$ und somit $\mathbf{r}^{(0)} = \mathbf{b}$. Jeder Vektor $\mathbf{y} \in K_t$ hat die Form*

$$\mathbf{y} = p(\mathbf{A})\mathbf{b},$$

mit einem Polynom $p(\cdot)$ vom Grad $t-1$. Das bedeutet, die Minimierung von $Q(\cdot)$ im Krylow-Raum kann als eine polynomiale Approximation der Lösung in der Form

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \approx p(\mathbf{A})\mathbf{b} = \mathbf{x}^{(t)}$$

verstanden werden.

Es gilt nun für die Galerkin-Lösung im Krylow-Raum K_t der Satz:

Lemma 4.12 (Residuen der Krylow-Raum Minimierung)**.** *Für die Residuuen der Krylow-Raum Minimierung gilt:*

$$\begin{aligned} &(i) \quad \mathbf{r}^{(t)} \in K_{t+1}, \\ &(ii) \quad \mathbf{r}^{(t)} \perp K_t, \\ &(iii) \quad K_{t+1} \subset K_t \quad \Rightarrow \quad \mathbf{x}^{(t)} = \mathbf{x}. \end{aligned}$$

Proof: (*i*) Zunächst gilt:

$$\begin{aligned} \mathbf{r}^{(t)} &= \mathbf{b} - \mathbf{A}\mathbf{x}^{(t)} = \mathbf{r}^{(0)} - \mathbf{r}^{(0)} + \mathbf{b} - \mathbf{A}\mathbf{x}^{(t)} \\ &= \mathbf{r}^{(0)} - \mathbf{b} + \mathbf{A}\mathbf{x}^{(0)} + \mathbf{b} - \mathbf{A}\mathbf{x}^{(t)} \\ &= \mathbf{r}^{(0)} + \mathbf{A}(\mathbf{x}^{(0)} - \mathbf{x}^{(t)}) \in K_t + \mathbf{A}K_t \subset K_{t+1}(\mathbf{r}^{(0)}; \mathbf{A}). \end{aligned}$$

(*ii*) Aus der Galerkin-Gleichung (4.4) folgt unmittelbar:

$$\mathbf{r}^{(t)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(t)} \perp K_t.$$

Wegen $\mathbf{r}^{(i)} \in K_t$ für alle $i < t$ gilt:

$$\mathbf{r}^{(i)} \perp \mathbf{r}^{(j)} \quad \forall i < j < t$$

(*iii*) Angenommen $K_{t+1} \subset K_t$. Dann ist auch $\mathbf{r}^{(t)} \in K_{t+1} \subset K_t$, aber nach (*ii*) gilt $\mathbf{r}^{(t)} \perp K_t$. Also ist $|\mathbf{r}^{(t)}| = 0$, somit $\mathbf{r}^{(t)} = 0$ und durch $\mathbf{A}\mathbf{x}^{(t)} = \mathbf{b}$ ist die Lösung bestimmt. □

Sobald das Krylow-Raum Verfahren abbricht, weil die Krylow-Räume nicht mehr größer werden, ist die Lösung gefunden. Für diese Konstruktion ist es wesentlich, dass der Krylow-Raum um das erste Residuum herum aufgespannt wird.

Zum Durchführen des Krylow-Raum Verfahrens muss nun eine $\mathbf{A}$-orthogonale Basis der Räume K_t erstellt werden. Diese Basis besteht dann aus den Abstiegsrichtungen

$$K_t = \operatorname{span}\{\mathbf{d}^{(0)}, \dots, \mathbf{d}^{(t-1)}\}.$$

Prinzipiell wäre es möglich diese Basis mit dem *Gram-Schmidt* oder anderen Orthogonalisierungsverfahren wie der *Householder-Transformation* oder den *Givens-Rotations* zu erstellen. Diese Verfahren haben allerdings den Nachteil, dass der Aufwand Orthogonalisierung eines Vektors $\mathbf{d}^{(t)}$ mit größer werdendem t steigt. Darüber hinaus ist insbesondere das Gram-Schmidt-Verfahren numerische instabil und sehr anfällig für Akkumulation von Rundungsfehlern. Stattdessen werden wir für die symmetrische Matrix $\mathbf{A}$ eine zweistufige Rekursionsformel zur Orthogonalisierung entsprechend dem Vorgehen bei den orthogonalen *Legendre-Polynomen* herleiten.

Wir gehen nun also davon aus, dass durch $\{\mathbf{d}^{(0)}, \dots, \mathbf{d}^{(t-1)}\}$ eine $\mathbf{A}$-orthogonale Basis des K_t besteht. Weiter sei durch $\mathbf{x}^{(t)} \in \mathbf{x}^{(0)} + K_t$ die Galerkin-Lösung und durch $\mathbf{r}^{(t)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(t)} \in K_{t+1}$ das entsprechende Residuum gegeben. Wir suchen nun $\mathbf{d}^{(t)} \in K_{t+1}$ als Orthogonalisierung von $\mathbf{r}^{(t)}$ bzgl. der $\{\mathbf{d}^{(0)}, \dots, \mathbf{d}^{(t-1)}\}$ gemäß dem Ansatz:

$$\mathbf{d}^{(t)} = \mathbf{r}^{(t)} + \sum_{j=0}^{t-1} \beta_j^{t-1} \mathbf{d}^{(j)} \in K_{t+1}. \tag{4.5}$$

Ohne Einschränkung sei $\mathbf{r}^{(t)} \notin K_t$, denn ansonsten folgt laut Satz 4.12 $\mathbf{r}^{(t)} = 0$ und die Lösung liegt vor. Wir multiplizieren Gleichung (4.5) mit einem $\mathbf{A}\mathbf{d}^{(i)}$ für ein $i = 0, \dots, t-1$ also ein $\mathbf{d}^{(i)} \in K_t$.

$$0 = \langle \mathbf{d}^{(t)}, \mathbf{A}\mathbf{d}^{(i)} \rangle = \langle \mathbf{r}^{(t)}, \mathbf{A}\mathbf{d}^{(i)} \rangle + \sum_{j=0}^{t-1} \beta_j^{t-1} \langle \mathbf{d}^{(j)}, \mathbf{A}\mathbf{d}^{(i)} \rangle, \quad i = 0, \dots, t-1.$$

Es gilt $\langle \mathbf{d}^{(j)}, \mathbf{A}\mathbf{d}^{(i)} \rangle = 0$ für alle $i \neq j$, also bleibt nur ein Term für $j = i$ in der Summe erhalten:

$$\langle \mathbf{r}^{(t)} - \beta_i^{t-1} \mathbf{d}^{(i)}, \mathbf{A}\mathbf{d}^{(i)} \rangle = 0, \quad i = 0, \dots, t-1.$$

Laut Satz 4.12 gilt $\mathbf{r}^{(t)} \perp K_t$ und wegen $\mathbf{A}\mathbf{d}^{(i)} \in K_t$ für $i < t-1$ folgt $\langle \mathbf{r}^{(t)}, \mathbf{A}\mathbf{d}^{(i)} \rangle = 0$ und

$$i < t-1: \quad \beta_i^{t-1} \langle \mathbf{d}^{(i)}, \mathbf{A}\mathbf{d}^{(i)} \rangle = 0 \quad \Rightarrow \quad \beta_i^{(t-1)} = 0.$$

Es bleibt der Fall $i = t - 1$:

$$i = t-1: \quad \langle \mathbf{r}^{(t)}, \mathbf{A}\mathbf{d}^{(t-1)} \rangle + \beta_{t-1}^{t-1} \langle \mathbf{d}^{(t-1)}, \mathbf{A}\mathbf{d}^{(t-1)} \rangle = 0 \quad \Rightarrow \quad \beta_{j-1}^{j-1} = -\frac{\langle \mathbf{r}^{(t)}, \mathbf{A}\mathbf{d}^{(t-1)} \rangle}{\langle \mathbf{d}^{(t-1)}, \mathbf{A}\mathbf{d}^{(t-1)} \rangle}.$$

Das heißt, mit dem Ansatz (4.5) erhalten wir den neuen $\mathbf{A}$-orthogonalen Basis-Vektor als:

$$\beta_{t-1} := \beta_{t-1}^{t-1} = \frac{\langle \mathbf{r}^{(t)}, \mathbf{A}\mathbf{d}^{(t-1)} \rangle}{\langle \mathbf{d}^{(t-1)}, \mathbf{A}\mathbf{d}^{(t-1)} \rangle}, \quad \mathbf{d}^{(t)} = \mathbf{r}^{(t)} - \beta_{t-1}\mathbf{d}^{(t-1)}. \tag{4.6}$$

$\mathbf{d}^{(t)}$ ist die neue Abstiegsrichtung und die nächste Approximation $\mathbf{x}^{(t+1)}$ wird gemäß dem Abstiegsverfahren bestimmt als:

$$\alpha_t = \frac{\langle \mathbf{r}^{(t)}, \mathbf{d}^{(t)} \rangle}{\langle \mathbf{d}^{(t)}, \mathbf{A}\mathbf{d}^{(t)} \rangle}, \quad \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \alpha_t \mathbf{d}^{(t)}. \tag{4.7}$$

Mit dieser Notation können wir auch für das neue Residuum $\mathbf{r}^{(t+1)}$ eine einfache Rekursionsformel herleiten. Es gilt:

$$\mathbf{r}^{(t+1)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(t+1)} = \mathbf{r}^{(t)} - \alpha_t \mathbf{A}\mathbf{d}^{(t)}. \tag{4.8}$$

Diese drei Rekursionsformeln (4.6-4.8) bilden das klassische CG-Verfahren. Ausgehend von einem Startwert $\mathbf{x}^{(0)}$ werden rekursiv Residuen, die $\mathbf{A}$-orthogonale Basis und die neue Lösung berechnet. Die Formeln zur Berechnung der Faktoren β_{t-1} und α_t können effizienter geschrieben werden. Denn wegen der Orthogonalität $\mathbf{r}^{(t)} \perp K_t$ gilt für den Nenner von α_t und β_t

$$\alpha_t \langle \mathbf{d}^{(t)}, \mathbf{A}\mathbf{d}^{(t)} \rangle = \langle \mathbf{d}^{(t)}, \underbrace{\alpha_t \mathbf{A}\mathbf{d}^{(t)} + \mathbf{r}^{(t+1)}}_{\mathbf{r}^{(t)}} \rangle = \langle \underbrace{\mathbf{d}^{(t)} - \beta_{t-1}\mathbf{d}^{(t-1)}}_{\mathbf{r}^{(t)}, \mathbf{r}^{(t)}} \rangle = |\mathbf{r}^{(t)}|^2,$$

sowie für den Zähler von β_t:

$$-\alpha_t \langle \mathbf{r}^{(t+1)}, \mathbf{A}\mathbf{d}^{(t)} \rangle = \langle \mathbf{r}^{(t+1)}, \underbrace{\mathbf{r}^{(t)} - \alpha_t \mathbf{A}\mathbf{d}^{(t)}}_{\mathbf{r}^{(t+1)}} \rangle = |\mathbf{r}^{(t+1)}|^2.$$

Solange $\mathbf{r}^{(t)} \neq 0$ gelten die vereinfachten Formeln:

$$\alpha_t = \frac{|\mathbf{r}^{(t)}|^2}{\langle \mathbf{d}^{(t)}, \mathbf{A}\mathbf{d}^{(t)} \rangle}, \quad \beta_{t-1} := \frac{|\mathbf{r}^{(t)}|^2}{|\mathbf{r}^{(t-1)}|^2}.$$

Wir fassen zusammen

Algorithmus 4.13 (CG-Verfahren). *Sei* $\mathbf{x}^{(0)} \in \mathbb{R}^N$ *ein gegebener Startwert sowie* $\mathbf{d}^{(0)} = \mathbf{r}^{(0)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$. *Iteriere für* $t \geqslant 0$

$$\begin{aligned}
(i) \quad & \alpha_t = \frac{|\mathbf{r}^{(t)}|^2}{\langle \mathbf{A}\mathbf{d}^{(t)}, \mathbf{d}^{(t)} \rangle} \\
(ii) \quad & \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \alpha_t \mathbf{d}^{(t)} \\
(iii) \quad & \mathbf{r}^{(t+1)} = \mathbf{r}^{(t)} - \alpha_t \mathbf{A}\mathbf{d}^{(t)} \\
(iv) \quad & \beta_t = \frac{|\mathbf{r}^{(t+1)}|^2}{|\mathbf{r}^{(t)}|^2} \\
(v) \quad & \mathbf{d}^{(t+1)} = \mathbf{r}^{(t+1)} + \beta_t \mathbf{d}^{(t)}.
\end{aligned}$$

Das CG-Verfahren bricht ab, wenn $|\mathbf{r}^{(t)}| = 0$. In diesem Fall ist die gesuchte Lösung erreicht. Üblicherweise wird das CG-Verfahren jedoch nur zur Approximation der Lösung mit $t \ll N$ verwendet. In jedem Schritt des CG-Verfahrens muss eine Matrix-Vektor Multiplikation, sowie 2 Skalarprodukte und 3 Additionen von Matrizen durchgeführt werden. Angenommen, die dünn besetzte Matrix hat maximal M Einträge pro Zeile (siehe Diskussion in Abschnitt 4.1), dann ist der Gesamtaufwand pro Schritt $O((5+M)N)$ arithmetische Operationen. Falls das CG-Verfahren als direkter Löser eingesetzt wird, wäre die Gesamtkomplexität $O(N^2M)$.

Es gilt:

Lemma 4.14 (CG-Verfahren). *Das CG-Verfahren bricht für jeden Startwert* $\mathbf{x}^{(0)} \in \mathbb{R}^N$ *nach maximal* $N-1$ *Schritten ab und liefert* $\mathbf{x}^{(t)} = \mathbf{x}$. *Für* $0 \leqslant t \leqslant N-1$ *gilt die Fehlerabschätzung:*

$$|\mathbf{x} - \mathbf{x}^{(t)}|_A \leqslant 2 \left(\frac{1 - \frac{1}{\sqrt{\kappa}}}{1 + \frac{1}{\sqrt{\kappa}}} \right)^t |\mathbf{x} - \mathbf{x}^{(0)}|_A,$$

mit der Spektralkondition $\kappa := \mathrm{cond}_2(\mathbf{A})$. *Zur Reduktion des Fehlers um den Faktor* $\varepsilon > 0$ *sind maximal:*

$$t(\varepsilon) \leqslant \frac{\sqrt{\kappa}}{2} \ln\left(\frac{2}{\varepsilon}\right) + 1$$

Schritte notwendig.

Proof: Siehe [Rannacher]. □

Remark 4.15 (CG-Verfahren als iteratives Lösungsverfahren). *Angenommen, die Finite Elemente Diskretisierung der Poisson-Gleichung soll mit dem CG-Verfahren gelöst werden. Hierzu soll eine feste Genauigkeit* ε *erreicht werden. Die Konvergenzgeschwindigkeit und die benötigte Anzahl an Schritten wird durch die Spektralkondition der Systemmatrix* $\mathbf{A}_h$ *bestimmt. Es gilt:*

$$\kappa(\mathbf{A}_h) = O(h^{-2}),$$

und also:

$$t(\varepsilon) = O(h^{-1}).$$

In einem d*-dimensionalen Gebiet mit uniformen Triangulierung gilt die Relation* $N = O(h^{-d})$*. Also sind*

$$t(\varepsilon) = O(N^{\frac{1}{d}})$$

Schritte erforderlich. Der Gesamtaufwand zum Lösen des Gleichungssystems ist somit

$$N_{CG} = O((5+M)N^{1+\frac{1}{d}})$$

arithmetische Operationen. In zweidimensionalen Gebieten ergibt sich $O(N^{\frac{3}{2}})$*. Für große* N *ist das CG-Verfahren weit effizienter als direkte Methoden, oder als z.B. die Jacobi oder die Gauß-Seidel-Iteration.*

4.2.3 Krylow-Raum Verfahren für nicht-symmetrische Gleichungssysteme

Satz 4.2 ist Grundlage des CG-Verfahrens, da er die Äquivalenz zwischen linearem Gleichungssystem und Minimierung der quadratischen Form $Q(\cdot)$ herstellt. Dieser Satz gilt nur für symmetrische Matrizen. Allgemeiner erhalten wir:

Lemma 4.16 (Minimierung des Residuums)**.** *Es sei* $\mathbf{A} \in \mathbb{R}^{N\times N}$ *eine reguläre Matrix. Dann ist das lineare Gleichungssystem* $\mathbf{Ax} = \mathbf{b}$ *äquivalent zur Minimierung des Residuums*

$$\|\mathbf{b} - \mathbf{Ax}\| = \min_{\mathbf{y}\in\mathbb{R}^N} \|\mathbf{b} - \mathbf{Ay}\|.$$

Proof: Sei $\mathbf{x}$ das Minimum des Residuums in der l_2-Norm. Dann gilt:

$$0 = \frac{d}{ds}\|\mathbf{b} - \mathbf{A}(\mathbf{x} + s\mathbf{y})\|^2\Big|_{s=0} \quad \forall \mathbf{y} \in \mathbb{R}^N,$$

und also

$$0 = \frac{d}{ds}\|\mathbf{b} - \mathbf{Ax} - s\mathbf{Ay}\|^2\Big|_{s=0} = 2\langle \mathbf{b} - \mathbf{Ax}, \mathbf{Ay}\rangle \quad \forall \mathbf{y} \in \mathbb{R}^N.$$

Da $\mathbf{A}$ regulär ist folgt $\langle \mathbf{b} - \mathbf{Ax}, \mathbf{y}\rangle = 0$ für alle $\mathbf{y} \in \mathbb{R}^N$ und schließlich $\mathbf{Ax} = \mathbf{b}$.

Umgekehrt sei nun $\mathbf{x}$ die Lösung des linearen Gleichungssystems. Dann gilt für $\mathbf{y}$ beliebig:

$$\begin{aligned}\|\mathbf{b} - \mathbf{Ay}\|^2 - \|\mathbf{b} - \mathbf{Ax}\|^2 &= \langle \mathbf{Ay}, \mathbf{Ay}\rangle - \langle \mathbf{Ax}, \mathbf{Ax}\rangle - 2\langle \mathbf{b}, \mathbf{Ay}\rangle + 2\langle \mathbf{b}, \mathbf{Ax}\rangle \\ &= \langle \mathbf{Ay}, \mathbf{Ay}\rangle + \langle \mathbf{Ax}, \mathbf{b}\rangle - 2\langle \mathbf{Ax}, \mathbf{Ay}\rangle = \|\mathbf{Ax} - \mathbf{Ay}\|^2 > 0.\end{aligned}$$

□

Anstelle der quadratischen Form soll jetzt direkt das Residuum der Gleichung minimiert werden. Mit dem Krylow-Raum

$$K_t := K_t(\mathbf{r}^{(0)}; \mathbf{A}) = \text{span}\{\mathbf{r}^{(0)}, \mathbf{Ar}^{(0)}, \ldots, \mathbf{A}^{t-1}\mathbf{r}^{(0)}\},$$

wird ein $\mathbf{x}^{(t)} \in \mathbf{x}^{(0)} + K_t$ gesucht, so dass gilt:

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}^{(t)}\|^2 = \min_{\mathbf{y} \in K_t} \|\mathbf{b} - \mathbf{A}\mathbf{y}\|^2. \tag{4.9}$$

Für dieses Minimum gilt wieder eine Orthogonalitätsbeziehung:

Lemma 4.17 (Galerkin-Gleichung)**.** *Die Lösung $\mathbf{x}^{(t)} \in \mathbf{x}^{(0)} + K_t$ der Minimierungsaufgabe (4.9) ist eindeutig durch die Galerkin-Gleichung beschrieben:*

$$\langle \mathbf{b} - \mathbf{A}\mathbf{x}^{(t)}, A\mathbf{y} \rangle = 0 \quad \forall \mathbf{y} \in K_t. \tag{4.10}$$

Proof: Übung. □

Zum Ansatz sei zunächst eine orthogonale Basis $\{\mathbf{d}^{(0)}, \ldots, \mathbf{d}^{(t-1)}\}$ des K_t gegeben. Durch

$$\mathbf{Q}_t := [\mathbf{d}^{(0)}, \ldots, \mathbf{d}^{(t-1)}] \in \mathbb{R}^{N \times t},$$

ist eine Matrix mit $\mathbf{Q}^T\mathbf{Q} = I$ gegeben. Jede Lösung $\mathbf{x}^{(t)} \in \mathbf{x}^{(0)} + K_t$ kann dann in der Form

$$\mathbf{x}^{(t)} = \mathbf{x}^{(0)} + \mathbf{Q}_t\mathbf{y}^{(t)}, \quad \mathbf{y}^{(t)} \in \mathbb{R}^t$$

geschrieben werden. Wir setzen diesen Ansatz in die Galerkin-Gleichung (4.10) ein und erhalten:

$$\langle \mathbf{r}^{(0)} - \mathbf{A}\mathbf{Q}_t\mathbf{y}^{(t)}, \mathbf{A}\mathbf{Q}\mathbf{z}^{(i)} \rangle = 0 \quad \forall \mathbf{z}^{(i)} \in \mathbb{R}^t.$$

Um die gesuchte Approximation $\mathbf{x}^{(t)} = \mathbf{Q}_t\mathbf{y}^{(t)}$ zu finden, muss also nur ein lineares Gleichungssystem mit t Unbekannten und Gleichungen gelöst werden.

Diese Idee ist Grundlage des *Generalized Minimal Residual Verfahrens* (GMRES). In einem ersten Schritt wird eine Orthonormalbasis des K_t erstellt. Anschließend wird das lineare Gleichungssystem im $\mathbb{R}^t$ gelöst um schließlich mittels $\mathbf{x}^{(t)} = \mathbf{Q}_t\mathbf{y}^{(t)}$ die Approximation zu erhalten.

Für allgemeine Matrizen $\mathbf{A}$ sind Konvergenzaussagen schwer zu treffen. Wegen der Herleitung über die Minimierung des Residuums konvergiert dieses aber monoton, wie bei dem CG-Verfahren. Im Fall fehlerfreier Arithmetik ist das GMRES-Verfahren ein direktes Lösungsverfahren. In der Anwendung werden jedoch stets nur wenige Schritte durchgeführt. Kern des Verfahrens ist die Orthogonalisierung des Krylow-Raums K_t. Je größer der Raum K_t umso aufwändiger ist auch die Orthogonalisierung. Aus diesem Grund wird üblicherweise nur eine feste Zahl m von Schritten des GMRES-Verfahrens ausgeführt und anschließend mit einer besseren Startlösung $\mathbf{x}^{(0)'} = \mathbf{x}^{(m)}$ neu gestartet. (GMRES-Verfahren mit *Restart*).

4.2.4 Vorkonditionierung

Satz 4.14 sagt, dass die Konvergenzgeschwindigkeit des CG-Verfahrens im Wesentlichen von der Spektralkondition der Matrix $\mathbf{A}$ abhängt. Die Idee der *Vorkonditionierung* ist es, das Gleichungssystem durch Multiplikation einer Matrix $\mathbf{P}^{-1}$ derart zu ändern, so dass für die Matrix $\mathbf{P}^{-1}\mathbf{A}$ gilt:

$$\operatorname{cond}_2(\mathbf{P}^{-1}\mathbf{A}) \ll \operatorname{cond}_s(\mathbf{A}).$$

Dann wird anstelle von $\mathbf{Ax} = \mathbf{b}$ das System

$$\mathbf{P}^{-1}\mathbf{Ax} = \mathbf{Pb},$$

gelöst, welches wegen der weitaus besseren Spektralkondition von $\mathbf{P}^{-1}\mathbf{A}$ schneller konvergiert. Wir gehen zunächst davon aus, dass die Matrix $\mathbf{A}$ symmetrisch positiv definit ist und mit dem CG-Verfahren gelöst werden soll. Das vorkonditionierte System muss dann auch symmetrisch positiv definit sein. Der symmetrisch positiv definite Vorkonditionierer $\mathbf{P}$ liege also in der Form:

$$\mathbf{P} = \mathbf{K}\mathbf{K}^{\mathsf{T}},$$

vor. Dann lösen wir:

$$\mathbf{K}^{-\mathsf{T}}\mathbf{K}^{-1}\mathbf{A}(\mathbf{K}^{-\mathsf{T}}\mathbf{K}^{\mathsf{T}})\mathbf{x} = \mathbf{K}^{-\mathsf{T}}\mathbf{K}^{-1}\mathbf{b} \quad \Rightarrow \quad \underbrace{\mathbf{K}^{-1}\mathbf{A}\mathbf{K}^{-\mathsf{T}}}_{=:\tilde{\mathbf{A}}}\underbrace{\mathbf{K}^{\mathsf{T}}\mathbf{x}}_{=\tilde{\mathbf{x}}} = \underbrace{\mathbf{K}^{-1}\mathbf{b}}_{=\tilde{\mathbf{b}}}.$$

Für die Matrix $\tilde{\mathbf{A}}$ gilt:

$$\mathbf{K}^{-\mathsf{T}}\tilde{\mathbf{A}}\mathbf{K}^{\mathsf{T}} = (\mathbf{K}^{-\mathsf{T}}\mathbf{K}^{-1})\mathbf{A}(\mathbf{K}^{-\mathsf{T}}\mathbf{K}^{\mathsf{T}}) = \mathbf{P}^{-1}\mathbf{A}.$$

Die Matrix $\tilde{\mathbf{A}}$ ist also ähnlich zur Matrix $\mathbf{P}^{-1}\mathbf{A}$. Ähnliche Matrizen haben die gleichen Eigenwerte. Im Fall $\mathbf{P} = \mathbf{A}$ ist die Matrix $\tilde{\mathbf{A}}$ also ähnlich zur Einheitsmatrix mit der Kondition $\operatorname{cond}_2(\mathrm{I}) = 1$. In diesem Fall wäre das Konvergenzverhalten des CG-Verfahrens optimal.

Im Folgenden formulieren wir das sogenannte *Vorkonditionierte CG-Verfahren* oder *Preconditioned Conjugate Gradient* (PCG):

Algorithmus 4.18 (PCG-Verfahren)**.** *Sei* $\mathbf{P} = \mathbf{K}\mathbf{K}^{\mathsf{T}}$ *ein Vorkonditionierer sowie* $\mathbf{x}^{(0)} \in \mathbb{R}^N$ *ein Startwert. Setze* $\mathbf{d}^{(0)} = \mathbf{r}^{(0)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$ *sowie* $\mathbf{z}^{(0)} = \mathbf{P}^{-1}\mathbf{d}^{(0)}$*. Iteriere*

$$\begin{aligned}
&(\mathrm{i}) && \alpha_t = \frac{\langle \mathbf{r}^{(t)}, \mathbf{z}^{(t)} \rangle}{\langle \mathbf{A}\mathbf{d}^{(t)}, \mathbf{d}^{(t)} \rangle} \\
&(\mathrm{ii}) && \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \alpha_t \mathbf{d}^{(t)} \\
&(\mathrm{iii}) && \mathbf{r}^{(t+1)} = \mathbf{r}^{(t)} - \alpha_t \mathbf{A}\mathbf{d}^{(t)} \\
&(\mathrm{iv}) && \mathbf{z}^{(t+1)} = \mathbf{P}^{-1}\mathbf{r}^{(t+1)} \\
&(\mathrm{v}) && \beta_t = \frac{\langle \mathbf{r}^{(t+1)}, \mathbf{z}^{(t+1)} \rangle}{\langle \mathbf{r}^{(t)}, \mathbf{z}^{(t)} \rangle} \\
&(\mathrm{vi}) && \mathbf{d}^{(t+1)} = \mathbf{r}^{(t+1)} + \beta_t \mathbf{d}^{(t)}.
\end{aligned}$$

Im Vergleich zum einfachen CG-Verfahren fällt als zusätzlicher Aufwand insbesondere das Lösen des Vorkonditionierersystems $\mathbf{P}\mathbf{z}^{(t+1)} = \mathbf{r}^{(t+1)}$ an. Hierzu kann die Zerlegung $\mathbf{P} = \mathbf{K}\mathbf{K}^T$ genutzt werden. Der Vorkonditionierer sollte also einerseit möglichst einfach zu invertieren sein, auf der anderen Seite sollte $\mathbf{P} \approx \mathbf{A}$, insbesondere sollten das Eigenwerte von $\mathbf{P}^{-1}\mathbf{A}$ möglichst nahe beieinander liegen.

Jacobi-Vorkonditionierung Für $\mathbf{P}$ eignet sich z.B. das Jacobi-Verfahren mit

$$\mathbf{P}_J = \operatorname{diag}(\mathbf{A}) =: \mathbf{D}, \quad \mathbf{P}_J = \mathbf{D}^{\frac{1}{2}}\mathbf{D}^{\frac{1}{2}}.$$

Dieser Vorkonditionierer ist sehr "billig" anzuwenden und sorgt dafür, dass die Einträge der Matrix $\mathbf{A}$ skaliert werden. Insbesondere gilt $\tilde{a}_{ii} = 1$ für $i = 1, \ldots, N$. Dies kann zu einer Reduktion der Kondition führen. Mit Hilfe der Gerschgorin-Kreise kann eine einfache Abschätzung für die Eigenwerte gefunden werden. Bei $\tilde{A}$ liegt der Mittelpunkt stets bei 1.

SOR-Vorkonditionierung Das Gauß-Seidel Verfahren kann nicht in der Form $\mathbf{P} = \mathbf{K}\mathbf{K}^T$ faktorisiert werden, es ist nicht symmetrisch. Die Matrix $\mathbf{A}$ sei additiv zerlegt in $\mathbf{A} = \mathbb{L} + \mathbb{D} + \mathbb{R} = \mathbb{L} + \mathbb{D} + \mathbb{L}^T$, da $\mathbf{A}$ symmetrisch. Das SOR-Verfahren hat die Iterationsmatrix:

$$P_{SOR} = (\mathbf{D} + \omega\mathbf{L})\mathbb{D}^{-1}(\mathbf{D} + \omega\mathbf{L}^T) = \underbrace{(\mathbf{D}^{\frac{1}{2}} + \omega\mathbf{L}\mathbf{D}^{-\frac{1}{2}})}_{K}\underbrace{(\mathbf{D}^{\frac{1}{2}} + \omega\mathbf{D}^{-\frac{1}{2}}\mathbf{L}^T)}_{K^T}$$

Bei optimaler Wahl der Relaxationsparameters ω (dies ist im Allgemeinen jedoch nicht möglich) wird eine wesentliche Reduktion der Kondition erreicht:

$$\operatorname{cond}_2(\tilde{\mathbf{A}}) = \sqrt{\operatorname{cond}_2(\mathbf{A})}.$$

Unvollständige Cholesky-Vorkonditionierung Abschließend stellen wir noch ein schwer zu analysierendes, jedoch höchst erfolgreiches Verfahren vor. Durch

$$\tilde{\mathbf{A}} \approx \tilde{\mathbf{C}}\tilde{\mathbf{C}^T},$$

sei die *unvollständige Cholesky-Zerlegung* von $\mathbf{A}$ gegeben. Die Cholesky-Zerlegung als Spezialfall der LR-Zerlegung ist üblicherweise voll besetzt. Mit $\tilde{\mathbf{C}}\tilde{\mathbf{C}^T}$ bezeichnen wir die Approximative Zerlegung der Matrix $\mathbf{A}$. Elemente c_{ij} von $\tilde{\mathbf{C}}$ werden künstlich auf Null gesetzt, wenn für den Matrixeintrag $a_{ij} = 0$ gilt.

Durch gute Vorkonditionierung, etwa mit dem SOR-Verfahren oder der Cholesky-Zerlegung kann die Konditionierung des vorkonditionierten Systems (bei der Poisson-Gleichung) auf $\operatorname{cond}(\tilde{\mathbf{A}}) = O(h^{-1})$ verbessert werden. Hierdurch wird die Konvergenzrate des CG-Verfahrens von $1 - O(h)$ auf $1 - O(\sqrt{h})$ verbessert. Dies führt zu einem Gesamtaufwand von $O(N^{\frac{5}{4}})$.

4.3 Mehrgitterverfahren

Alle bisher betrachteten iterativen Lösungsverfahren hängen von der Kondition der Matrix und damit bei der Behandlung der FE Diskretisierung der Poisson-Gleichung vom Gitter.

Im Folgenden betrachten wir die eindimensionale Diskretisierung mit linearen Finiten Elementen auf einem uniformen Gitter mit N Elementen. Als prototypisches Iterationsverfahren werden wir die gedämpfte Richardson-Iteration genauer untersuchen. Es ist:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \theta(\mathbf{b} - \mathbf{A}\mathbf{x}^{(t)}) = \underbrace{(I - \theta\mathbf{A})}_{=:\mathbf{B}_\theta}\mathbf{x}^{(t)} + \theta\mathbf{b},$$

mit der Iterationsmatrix $\mathbf{B}_\theta$.

Die positiv definite Matrix $\mathbf{A}$ hat in *Stencil-Notation* die Form:

$$\mathbf{A} = \begin{bmatrix} -1 & 2 & -1 \end{bmatrix}.$$

Diese Matrix hat die Eigenvektoren $\omega_k \in \mathbb{R}^{N+1}$ mit Eigenwerten λ_k:

$$\begin{aligned} \lambda_k &= 2\left(1 - \cos\left(\frac{k}{N}\pi\right)\right) \\ (\omega_k)_i &= \sin\left(\frac{ki}{N}\pi\right), \quad i = 0, \ldots, N. \end{aligned} \tag{4.11}$$

Die Konvergenzordnung der Iterativen Verfahren ist im Allgemeinen sehr langsam, sie hängt vom Spektralradius der Iterationsmatrix $\mathbf{B} := I - \theta\mathbf{A}$ ab. Es gilt wieder in Stencil-Notation

$$\mathbf{B}_\theta = \begin{bmatrix} \theta & (1 - 2\theta) & \theta \end{bmatrix}$$

Wir wählen nun $\theta = 4^{-1} \approx \lambda_{max}(\mathbf{A}_h)^{-1}$. Dann ist

$$\mathbf{B} = \frac{1}{4}\begin{bmatrix} 1 & 2 & 1 \end{bmatrix}$$

Diese Matrix $\mathbf{B} := \mathbf{B}_{\frac{1}{4}}$ hat wieder die gleichen Eigenvektoren (4.11) diesmal mit den Eigenwerten:

$$\lambda_k^B = \frac{1}{2}\left(1 + \cos\left(\frac{k}{N}\pi\right)\right)$$

Die Eigenwerte liegen alle im Intervall $\lambda_k^B \in (0, 1)$, der größte Eigenwert von $\mathbf{B}$ verhält sich wie $\lambda_{max}(\mathbf{B}) = 1 - O(h^2)$.

Wir untersuchen die Konvergenz des Richardson-Verfahrens für die Poisson-Gleichung im Detail. Es sei also $\mathbf{x}^{(t)}$ die letzte Approximation mit Fehler $\mathbf{e}^{(t)} := \mathbf{x} - \mathbf{x}^{(t)}$. Für den Fehler in der nächsten Iteration gilt:

$$\mathbf{e}^{(t+1)} = \mathbf{B}\mathbf{e}^{(t)}.$$

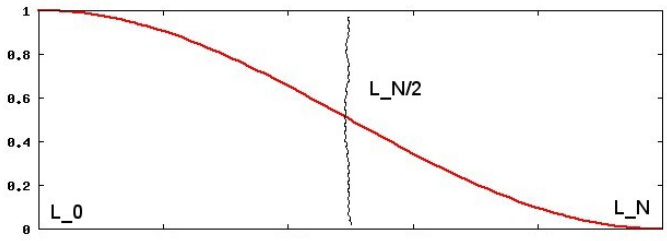


Abbildung 4.3: Fehlerreduktion des gedämpften Richardson-Verfahrens für die eindimensionale Poisson-Gleichung in Abhängigkeit der Fehlerfrequenz.

Wir schreiben den Vektor $\mathbf{e}^{(t)}$ in einer Entwicklung in Eigenwerten

$$\mathbf{e}^{(t)} = \sum_{k=1}^{N-1} \mathbf{e}_k^{(t)} \omega_k.$$

Dabei sind die $\mathbf{e}_k^{(t)} \in \mathbb{R}$ die Entwicklungskoeffizienten in der Eigenvektor-Darstellung und nicht die kartesischen Koeffizienten. Wir nennen (wegen der Sinus-Form der Eigenvektoren) die einzelnen Komponenten ω_k die *Frequenzen* des Fehlers. Wir können nun die Wirkung des Richardson-Verfahrens für jede einzelne Frequenz ω_k verfolgen. Es gilt

$$\mathbf{e}_k^{(t+1)} = \lambda_k^B \mathbf{e}_k^{(t)},$$

das heißt, die k-te Komponente des Fehlers wird genau um den Eigenwert λ^B gedämpft. In Abbildung 4.3 zeigen wir den Verlauf der Eigenwerte λ_k^B. Fehlerkomponenten, welche zu niedrigen Frequenzen gehören werden nur sehr langsam reduziert, während alle hochfrequenten Anteile schnell reduziert werden. Wir definieren:

Definition 4.19 (Fehlerfrequenzen)**.** *Komponenten welche zu Eigenwerten λ_k für $k > \frac{N}{2}$ gehören heißen* hochfrequente Anteile*, alle anderen Komponenten heißen* niederfrequent*. Hochfrequente Anteile sind gerade die, welche auf gröberen Gittern nicht dargestellt werden können.*

Das Richardson-Verfahren ist zwar ein schlechter Löser für die Poisson-Gleichung, es *glättet* hochfrequente Fehleranteile allerding sehr schnell aus. In Abbildung 4.4 zeigen wir den Fehler der Richardson-Iteration über einige Schritte. Der Gesamtfehler wird nach zehn Schritten nicht wesentlich kleiner, allerdings ist der Fehler bereits nach zwei Schritten stark geglättet. Das Richardson-Verfahren scheint also in der Lage zu sein, lokale Fehleranteile sehr schnell zu *glätten*.

Das Mehrgitterverfahren beruht nun auf der Idee, dass die Frage, ob ein Fehleranteil hochfrequent ist oder nicht vom zugrundeliegenden Gitter abhängt. Auf einem Gitter mit $N = 10$ Elementen ist die Frequenz

$$(\omega_4)_i = \sin\left(\frac{4i}{10}\pi\right), \quad i = 0, \ldots, 10,$$

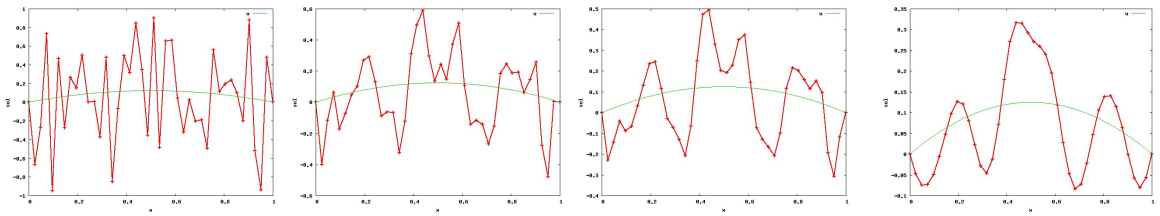


Abbildung 4.4: Fehlerverlauf (rot) und approximierte Lösung (grün) der gedämpften Richardson-Iteration. Von links nach rechts: Startfehler, nach einem Schritt, zwei Schritten und nach 9 Schritten.

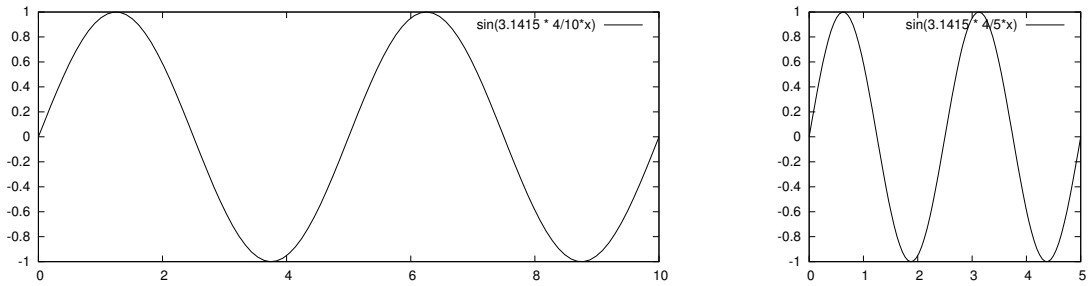


Abbildung 4.5: Die Fehlerfrequenz λ_4 auf einem Gitter mit 10 Elementen und auf eine Gitter mit 5 Elementen.

niederfrequent, auf einem gröberen Gittern mit $N' = 5$ Elementen ist die gleiche Schwingung jedoch hochfrequent:

$$(\omega_4)'_i = \sin\left(\frac{4i}{5}\pi\right), \quad i = 0, \ldots, 5.$$

In Abbildung 4.5 ist diese Frequenz auf beiden Gittern aufgetragen. Wir fassen zusammen:

- Einfache Iterationsverfahren wie die Richardson-Iteration sind schlechte Löser aber gute Glätter für hochfrequente Fehleranteiler.
- Niederfrequente Anteile auf einem Gitter Ω_h sind hochfrequente Anteile auf einem gröberen Gitter Ω_{2h}.
- Und ganz trivial: je gröber das Gitter, umso geringer der Aufwand.

Beim Mehrgitter-Verfahren sollen diese Leitsätze kombiniert werden. Auf dem feinsten Gitter Ω_h wird das Gleichungssystem nicht gelöst, es werden zunächst nur hochfrequente Fehleranteile ausgeglättet und es bleiben nur niederfrequente Fehler $\mathbf{e}_h \to \mathbf{e}_h^{nf}$. Diese werden auf ein gröberes Gitter Ω_H transferiert $\mathbf{e}_h^{nf} \to \mathbf{e}_H^{hf}$, wo sie wieder hochfrequent sind. Auf diesem Grobgitter ist das verbleibende Problem sehr viel kleiner und kann einfacher gelöst werden.

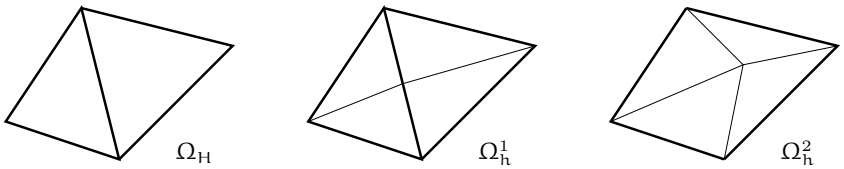


Abbildung 4.6: Grobgitter Ω_H und zwei feinere Gitter. Das Gitter Ω_h^1 ist durch Verfeinerung entstanden und es gilt $\Omega_H \Subset \Omega_h^1$. Das Gitter Ω_h^2 stammt nicht von Ω_H ab.

4.3.1 Hierarchische Finite Elemente Ansätze

Wir müssen zur Notation zunächst einige Begriffe einführen. Durch

$$\Omega_H = \Omega_0, \Omega_1, \ldots, \Omega_L = \Omega_h,$$

sei eine Familie von Gittern des Gebiets Ω gegeben. Wir definieren:

Definition 4.20 (Geschachtelte Gitter)**.** *Seien Ω_H, Ω_h zwei Triangulierungen des Gebiets Ω. Die Gitter heißen geschachtelt $\Omega_H \Subset \Omega_h$, falls für jeden Knoten $x_i \in \Omega_H$ gilt $x_i \in \Omega_h$ und jedes Element $K \in \Omega_h$ durch Verfeinerung eines Elementes $K' \in \Omega_H$ entstanden ist.*

In Abbildung 4.6 zeigen einfache Beispiele von Gittern die geschachtelt sind und Gittern, die nicht geschachtelt sind. Auf jedem Gitter Ω_l sei durch V_l ein Finite Elemente Ansatzraum definiert und wir definieren die lokalen Probleme:

$$u_l \in V_l : \quad a(u_k, \phi_l) = (f, \phi_l) \quad \forall \phi_l \in V_l$$

Als kompakte Schreibweise definieren wir einen Operator $\mathcal{A}_l : V_l \to V_l$ mittels

$$(\mathcal{A}_l u_l, v_l) = a(u_l, v_l) \quad \forall u_l, v_l \in V_l.$$

Mit der Vektordarstellung

$$u_l = \sum_{i=1}^{N_l} \mathbf{u}_l^i \phi_h^{(i)},$$

und der Steifigkeitsmatrix $\mathbf{A}_l$ ist dies Problem äquivalent zu dem linearen Gleichungssystem

$$\mathbf{A}_l \mathbf{u}_l = \mathbf{b}_l, \quad (\mathbf{A}_l)_{ij} = a(\phi_h^{(j)}, \phi_h^{(i)}), \quad (\mathbf{b}_l)_i = (f, \phi_h^{(i)}).$$

Es gilt:

Lemma 4.21 (Geschachtelte Finite Elemente Räume)**.** *Es seien $\Omega_H \Subset \Omega_h$ geschachtelte Gitter und V_H sowie V_h isoparametrische Finite Elemente Räume mit dem gleichen Finite Elemente Ansatz $\{P(\hat{T}), \chi(\hat{T})\}$ auf den Gittern Ω_H bzw. Ω_h. Es gilt:*

$$V_H \subset V_h.$$

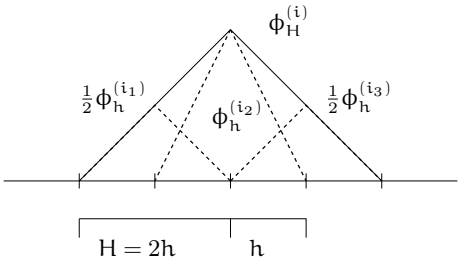


Abbildung 4.7: Darstellung einer groben Basisfunktion $\phi_H^{(i)}$ (durchgezogene Linie) durch drei Basisfunktionen des feinen Gitters.

Proof: Übung. □

Definition 4.22 (Gittertransfer)**.** *Es seien $V_{l-1} \subset V_l$ zwei geschachtelte Finite-Elemente Räume. Für eine Funktion $v_h \in V_h$ definieren wir den* Restriktionsoperator $\mathcal{R}_{l-1} : V_l \to V_{l-1}$ *durch*

$$(\mathcal{R}_{l-1} v_h, \phi_H) = (v_h, \phi_H) \quad \forall \phi_H \in V_H,$$

sowie für eine Funktion $v_H \in V_H$ den Prolongationsoperator $\mathcal{P}_l : V_{l-1} \to V_l$ *durch*

$$\mathcal{P}_l v_H = v_H.$$

Remark 4.23 (Eigenschaften der Restriktion)**.** *Seien $\Omega_H \Subset \Omega_h$ zwei geschachtelte Gitter. Die Restriktion $\mathcal{R}_H : V_h \to V_H$ ist gerade die L^2-Projektion von V_h in den gröberen Raum V_H. Die L^2-Projektion ist eine globale Operation, um sie zu berechnen muss ein lineares Gleichungssystem mit der Massenmatrix $\mathbf{M}_H$ gelöst werden. Obwohl ein lineares Gleichungssystem mit der Massenmatrix relativ einfach zu lösen ist (da $cond_2(\mathbf{M}_H) = O(1)$) sollte dies aus Effizienzgründen vermieden werden.*

In der Anwendung des Mehrgitterverfahrens werden wir allerdings gar nicht die Finite Elemente Funktion $\mathcal{R}_H v_h \in V_H$ benötigen, sondern nur den Vektor $\mathbf{x}_H \in \mathbb{R}^{N_H}$ mit

$$(\mathbf{x}_H)_i := (\mathcal{R}_H v_h, \phi_H^{(i)}) = (v_h, \phi_H^{(i)}), \quad i = 1, \ldots, N_H. \tag{4.12}$$

Zu der beliebigen Funktion $v_h \in V_h$ definieren wir auf dem feinen Gitter Ω_h den Vektor $\mathbf{x}_h \in \mathbb{R}^{N_h}$ als

$$(\mathbf{x}_h)_i = (v_h, \phi_h^{(i)}).$$

Die Räume $V_H \subset V_h$ sind geschachtelt, d.h., jede Knotenbasisfunktion $\phi_H^{(i)} \in V_H$ ist auch im feinen Raum $\phi_H^{(i)} \in V_h$ und hat dort die Basisdarstellung

$$\phi_H^{(i)} = \sum_{j=1}^{N_h} \mu_{ij} \phi_h^{(j)}, \tag{4.13}$$

mit einer Koeffizientenmatrix $\mathbf{R}_H \in \mathbb{R}^{N_H \times N_h}$ *mit* $(\mathbf{R}_H)_{ij} = \mu_{ij}$ *welche sehr dünn besetzt ist, also mit* $\mu_{ij} \neq 0$ *nur für sehr wenige (insbesondere unabhängig von* h *und* H*) Einträge. In Abbildung 4.7 zeigen wir die Kombination einer Basisfunktion* $\phi_H^{(i)}$ *durch drei Funktionen des feinen Gitters.*

Die Einträge des gesuchten Vektors $\mathbf{x}_H$ *aus (4.12) sind dann bestimmt durch die Gleichungen:*

$$(\mathbf{x}_H)_i = (v_h, \phi_H^{(i)}) = \sum_{j=1}^{N_h} \mu_{ij} \underbrace{(v_h, \phi_h^{(j)})}_{=:(\mathbf{x}_h)_j} = (\mathbf{R}_H \mathbf{x}_h)_i.$$

Das heißt, wenn nicht die Funktion $\mathcal{R}_H v_h$ *von Interesse ist, sondern nur die Funktion* r_h *getestet mit den Basisfunktionen* $\phi_h^{(i)}$*, also ein Vektor* $\mathbf{x}_h = ((r_h, \phi_h^{(i)}))_{i=1}^{N_h}$*, dann besteht der einfache Zusammenhang*

$$\mathbf{x}_H = \mathcal{R}_H \mathbf{x}_h.$$

Ebenso betrachten wir nun

Remark 4.24 (Eigenschaften der Prolongation)**.** *Die Prolongation* $\mathcal{P}_h : V_H \to V_h$ *ist die Identität auf* V_H*. Mit der Darstellung (4.13) erhalten wir für jeden Vektor* $v_H \in V_H$ *unmittelbar*

$$\begin{aligned} v_H(x) &= \sum_{i=1}^{N_H} (\mathbf{v}_H)_i \phi_H^{(i)}(x) = \sum_{i=1}^{N_H} \sum_{j=1}^{N_h} \mu_{ij} (\mathbf{v}_H)_i \phi_h^{(j)}(x) \\ &= \sum_{j=1}^{N_h} \underbrace{\left(\sum_{i=1}^{N_H} \mu_{ij} (\mathbf{v}_H)_i \right)}_{=:(\mathbf{v}_h)_j} \phi_h^{(j)}(x) \end{aligned}$$

Für den Vektor $\mathbf{v}_h \in \mathbb{R}^{N_h}$ *gilt also*

$$\mathbf{v}_h = \mathbf{R}_H^T \mathbf{v}_H.$$

Die Wirkung der Prolongation auf einen Knotenvektor ist also gerade durch die transponierte Matrix der Restriktion eines integrierten Vektors beschrieben. Man beachte, dass die Hintereinanderausführung von Restriktion und Prolongation nicht die Identität ergibt!

Der enge Zusammenhang zwischen Formulierungen in den Funktionenräumen V_h und V_H sowie zwischen den Vektorräumen $\mathbb{R}^{N_H}$ sowie $\mathbb{R}^{N_h}$ ist typisch für die Analyse des Mehrgitterverfahrens.

4.3.2 Das Zweigitter-Verfahren

Als Vorstufe zum Mehrgitterverfahren beschreiben wir zunächst die *Zweigitteriteration*. Diese formulieren wir im Finite-Elemente Kontext:

Algorithmus 4.25 (Zweigitteriteration in Finite-Elemente Räumen). *Seien $V_H \subset V_h$ zwei geschachtelte FE-Räume sowie $u_h^{(0)} \in V_h$ eine Approximation der Lösung $u_h \in V_h$. Weiter sei $\omega \in (0,1]$ ein Dämpfungsparameter und durch $\mathcal{G}_h : V_h \to V_h$ eine Glättungsiteration gegeben. Iteriere:*

$$u_h^{(t+1)} = \mathcal{Z}\mathcal{G}(u_h^{(t)}, f_h),$$

mit der Zweigitter-Iteration:

(i)	*Vorglätten:*	$\bar{u}_h = \mathcal{G}_h^{\nu_1}(u_h^{(t)}) := \mathcal{G}_h^{\nu_1}(u_h^{(t)}, a(\cdot,\cdot), f)$
(ii)	*Grobgitterproblem:*	$w_H \in V_H : \quad a(\bar{u}_h + w_H, \phi_H) = (f, \phi_H) \quad \forall \phi_H \in V_H$
(iii)	*Nachglätten:*	$u_h^{(t+1)} = \mathcal{G}_h^{\nu_2}(\bar{u}_h + \omega w_H)$

Das Verfahren besteht aus zwei Schritten: zunächst werden im feinen Finite Elemente Raum V_h die hochfrequenten Fehleranteile geglättet. Im Anschluss wird durch die Grobgitter-Korrektur das Problem im groben Raum $V_H \subset V_h$ approximiert. Schließlich kann ein weiteres Mal geglättet werden. Die Approximation wird in zwei Stufen aufgeteilt: die hohen Fehlerfrequenzen werden mit einigen wenigen (üblicherweise $\nu \sim 3$) Schritten eines einfachen Iterationsverfahrens reduziert, alle niedrig frequenten Fehleranteile werden auf dem Grobgitter behandelt.

Remark 4.26 (Nachglätten). *Die Zweigitteriteration (und später auch die Mehrgitteriteration) konvergieren auch ohne Nachglättung nur mit Vorglättung. Auch für die Beweise ist die Nachglättung nicht wesentlich. Wir setzen daher im Folgenden ohne Einschränkung $\nu_2 = 0$.*

Im Folgenden beschreiben wir die einzelnen Schritte der Zweigitter-Iteration genauer und leiten darüber hinaus eine Vektor-Schreibweise des Zweigitter-Verfahrens her.

Die Vorglättung Mit $\bar{u}_h = \mathcal{G}_h^{\nu}(u_h^{(0)})$ ist die ν-fache Ausführung der Glättungsoperation bezeichnet. Allgemein sei $u_h^{(t)} = \mathcal{G}_h(u_h^{(t-1)}) := \mathcal{G}_h u_h^{(t-1)} + g_h$ affin linear und eine Fixpunkt-Iteration mit $\mathcal{G}_h(u_h) = u_h$. Für den Fehler $\bar{e}_h := u_h - \bar{u}_h = u_h - u_h^{(\nu)}$ gilt dann

$$\bar{e}_h = u_h - u_h^{(\nu)} = \mathcal{G}_h(u_h) - \mathcal{G}_h(u_h^{(\nu-1)}) = \mathcal{G}_h e_h^{(\nu-1)} = \mathcal{G}_h^{\nu} e_h^{(0)} = \mathcal{G}_h^{\nu}(u_h - u_h^{(0)}). \quad (4.14)$$

Wir betrachten als Beispiel die gedämpfte Richardson-Iteration. In Finite Elemente Schreibweise suchen wir $u_h^{(t)} \in V_h$, so dass

$$(u_h^{(t)}, \phi_h) = (\mathcal{G}_h(u_h^{(t-1)}), \phi) := (u_h^{(t-1)}, \phi_h) + \theta\big((f_h, \phi_h) - a(u_h^{(t-1)}, \phi_h)\big) \quad \forall \phi_h \in V_h,$$

also

$$u_h^{(t+1)} = (\mathcal{I}_h - \theta \mathcal{A}_h) u_h^{(t)} + \theta f_h.$$

Formuliert mit Vektoren gilt die Vorschrift:

$$\mathbf{M}_h \mathbf{u}_h^{(t)} = \mathbf{M}_h \mathbf{u}_h^{(t-1)} + \theta(\mathbf{b}_h - \mathbf{A}_h \mathbf{u}_h^{(t)}),$$

oder mit einer Verfahrensmatrix:

$$\mathbf{u}_h^{(t)} = \mathbf{G}_h(\mathbf{u}_h^{(t-1)}) := \mathbf{G}_h\mathbf{u}_h^{(t-1)} + \mathbf{g}_H, \quad \mathbf{G}_h := I - \theta\mathbf{M}_h^{-1}\mathbf{A}_h, \quad \mathbf{g}_h := \theta\mathbf{M}_h^{-1}\mathbf{b}_h.$$

Dann erhalten wir die Fehlerfortpflanzungen in Finite Elemente und Vektorschreibweise:

$$\bar{\mathbf{e}}_h = \mathbf{u}_h - \bar{\mathbf{u}}_h = \mathbf{G}_h^\nu \mathbf{e}_h^{(0)}, \quad \bar{e}_h = u_h - \bar{u}_h = \mathcal{G}_h^\nu e_h^{(0)}. \tag{4.15}$$

Die Grobgitterkorrektur Wir suchen die Lösung $w_H \in V_H$ von

$$a(w_H, \phi_H) = (f, \phi_H) - a(\bar{u}_h, \phi_H) \quad \forall \phi_H \in V_H. \tag{4.16}$$

Die rechte Seite dieses Problems ist das Residuum zur vorgeglätteten Approximation $\bar{u}_h \in V_h$ bezüglich der groben Basisfunktionen $\phi_H \in V_H$. Wir betrachten das Residuum $\bar{r}_h \in V_h$ als

$$(\bar{r}_h, \phi_h) := (f, \phi_h) - a(\bar{u}_h, \phi_h) \quad \forall \phi_h \in V_h.$$

Dann ist die rechte Seite $\bar{r}_H \in V_H$ von (4.16) gegeben als

$$\bar{r}_H = \mathcal{R}_H \bar{r}_h. \tag{4.17}$$

Mit dem Operator $\mathcal{A}_H$ schreiben wir (4.16) in der Form

$$w_H = \mathcal{A}_H^{-1}\bar{r}_H, \tag{4.18}$$

In Vektor-Schreibweise definieren wir zunächst

$$\mathbf{r}_h := \mathbf{b}_h - \mathbf{A}_h\bar{\mathbf{u}}_h,$$

für welchen gilt:

$$(\mathbf{r}_h)_i = (r_h, \phi_h^{(i)}).$$

Also, mit Bemerkung 4.23 ist

$$\mathbf{r}_H = \mathbf{R}_H\mathbf{r}_h, \quad (\mathbf{r}_H)_i = (r_H, \phi_h^{(i)}),$$

und die Grobgitterlösung $\mathbf{w}_H \in \mathbb{R}^{N_H}$ berechnet sich als

$$\mathbf{w}_H = \mathbf{A}_H^{-1}\bar{\mathbf{r}}_H, \tag{4.19}$$

oder eben

$$\mathbf{w}_H = \mathbf{A}_H^{-1}\mathbf{R}_H(\mathbf{b}_h - \mathbf{A}_h\bar{\mathbf{u}}_h). \tag{4.20}$$

Update Schließlich ergibt sich die neue Iteration durch:

$$u_h^{(1)} = \bar{u}_h + \omega \mathcal{P}_h w_H, \tag{4.21}$$

bzw. in Vektorschreibweise

$$\mathbf{u}_h^{(1)} = \bar{\mathbf{u}}_h + \omega \mathbf{R}_H^T \mathbf{W}_H. \tag{4.22}$$

Wir fassen zusammen:

Algorithmus 4.27 (Zweigitteriteration)**.** *Seien $\Omega_H \Subset \Omega_h$ zwei geschachtelte Triangulierungen von Ω und $V_H \subset V_h$ zwei geschachtelte Finite Elemente Räume mit $\dim(V_H) = N_H$ und $\dim(V_h) = N_h$. Sei $\mathbf{u}_h^{(0)} \in \mathbb{R}^{N_h}$ ein Startvektor. $\nu > 0$ sei die Anzahl an Glättungsschritten, $\omega \in (0, 1]$ ein Dämpfungsparameter. Iteriere*

$$u_h^{(t+1)} = \mathcal{ZG}(u_h^{(t)}, f_h),$$

mit der Zweigitter-Iteration $\mathcal{ZG}(u_h^{(t)}, f_h)$:

(i)	*Vorglätten:*	$\bar{\mathbf{u}}_h = \mathbf{G}_h^{\nu_1}(\mathbf{u}_h^{(t)})$	$\bar{u}_h = \mathcal{G}_h^{\nu_1}(u_h^{(t)})$
(ii)	*Residuum:*	$\mathbf{r}_h = \mathbf{b}_h - \mathbf{A}_h \bar{\mathbf{u}}_h$	$r_h = f_h - \mathcal{A}_h \bar{u}_h$
(iii)	*Restriktion:*	$\mathbf{r}_H = \mathbf{R}_H \mathbf{r}_h$	$r_H = \mathcal{R}_H r_H$
(iv)	*Grobgitterproblem:*	$\mathbf{w}_H = \mathbf{A}_H^{-1} \mathbf{r}_H$	$w_H = \mathcal{A}_H^{-1} r_H$
(v)	*Prolongation:*	$\mathbf{u}_h^{(t+1)} = \bar{\mathbf{u}}_h + \omega \mathbf{R}_H^T \mathbf{w}_H$	$u_h^{(t+1)} = \bar{u}_h + \omega \mathcal{P}_h w_H$

Fehlerfortpflanzung Zur Analyse der Zweigitter-Konvergenz müssen wir nun eine Fehlerfortpflanzung herleiten, also einen Zusammenhang zwischen $e_h^{(t+1)}$ nach einem Schritt und $e_h^{(t)}$ vor dem Schritt. Wir entwickeln diese Fehlerdarstellung wieder simultan in Funktionenschreibweise und für Vektoren. Für den neuen Fehler gilt:

$$e_h^{(t+1)} = u_h - u_h^{(t+1)}, \quad \text{bzw.} \quad \mathbf{e}_h^{(t+1)} = \mathbf{u}_h - \mathbf{u}_h^{(t+1)}.$$

Mit (4.21) bzw. (4.22) erhalten wir

$$e_h^{(t+1)} = \bar{e}_h - \mathcal{P}_h w_H, \quad \text{bzw.} \quad \mathbf{e}_h^{(t+1)} = \bar{\mathbf{e}}_h - \mathbf{R}_H^T \mathbf{w}_H$$

Weiter, mit (4.18) bzw. (4.19)

$$e_h^{(t+1)} = \bar{e}_h - \mathcal{P}_h \mathcal{A}_H^{-1} \bar{r}_H, \quad \text{bzw.} \quad \mathbf{e}_h^{(t+1)} = \bar{\mathbf{e}}_h - \mathbf{R}_H^T \mathbf{A}_H^{-1} \bar{\mathbf{r}}_H.$$

Die Rechte Seite des Grobgitterproblems ist durch (4.17) bzw. durch (4.20) in Vektorschreibweise gegeben:

$$e_h^{(t+1)} = \bar{e}_h - \mathcal{P}_h \mathcal{A}_H^{-1} \mathcal{R}_H (f - \mathcal{A}_h \bar{u}_h), \quad \text{bzw.} \quad \mathbf{e}_h^{(t+1)} = \bar{\mathbf{e}}_h - \mathbf{R}_H^T \mathbf{A}_H^{-1} \mathbf{R}_H (\mathbf{b}_h - \mathbf{A}_h \bar{\mathbf{u}}_h).$$

Wir nutzen nun für die exakte Lösung $\mathcal{A}_h u_h = f$ bzw. $\mathbf{A}_h \mathbf{u}_h = \mathbf{b}_h$ und erhalten einen Bezug zum Fehler nach der Vorglättung

$$e_h^{(t+1)} = \bar{e}_h - \mathcal{P}_h \mathcal{A}_H^{-1} \mathcal{R}_H \mathcal{A}_h (u_h - \bar{u}_h) \quad \text{bzw.} \quad \mathbf{e}_h^{(t+1)} = \bar{\mathbf{e}}_h - \mathbf{R}_H^T \mathbf{A}_H^{-1} \mathbf{R}_H \mathbf{A}_h (\mathbf{u}_h - \bar{\mathbf{u}}_h),$$

also:

$$e_h^{(t+1)} = (\mathcal{I}_h - \mathcal{P}_h \mathcal{A}_H^{-1} \mathcal{R}_H \mathcal{A}_h) \bar{e}_h, \quad \text{bzw.} \quad \mathbf{e}_h^{(t+1)} = [I_h - \mathbf{R}_H^T \mathbf{A}_H^{-1} \mathbf{R}_H \mathbf{A}_h] \bar{\mathbf{e}}_h.$$

Für diesen gilt mit Darstellung (4.14) und (4.15)

$$e_h^{(t+1)} = (\mathcal{I}_h - \mathcal{P}_h \mathcal{A}_H^{-1} \mathcal{R}_H \mathcal{A}_h) \mathcal{G}_h^{\nu_1} e_h^{(t)} \quad \text{bzw.} \quad \mathbf{e}_h^{(t+1)} = [I_h - \mathbf{R}_H^T \mathbf{A}_H^{-1} \mathbf{R}_H \mathbf{A}_h] \mathbf{G}_h^{\nu_1} \mathbf{e}_h^{(t)}.$$

Zusammengefasst erhalten wir den *Zweigitter-Operator* $\mathcal{B}_{ZG}(\nu)$ und die *Zweigitter-Matrix* $\mathbf{B}_{ZG}(\nu)$:

$$\mathcal{B}_{ZG}(\nu) = (\mathcal{I}_h - \mathcal{P}_h \mathcal{A}_H^{-1} \mathcal{R}_H \mathcal{A}_h) \mathcal{G}_h^{\nu}, \quad \mathbf{B}_{ZG}(\nu) = [I_h - \mathbf{R}_H^T \mathbf{A}_H^{-1} \mathbf{R}_H \mathbf{A}_h] \mathbf{G}_h^{\nu}$$

Die Zweigitter-Iteration spaltet sich in zwei Bestandteile: die Glättung und die Grobgitter-Korrektur. Bei der mathematischen Konvergenzanalyse des Verfahrens werden diese beiden Anteile getrennt. Wir beweisen getrennt:

Lemma 4.28 (Glättungseigenschaft)**.** *Die gedämpfte Richardson-Iteration mit* $\theta = \lambda_{max}(A_h)^{-1}$ *erfüllt die Glättungseigenschaft*

$$\|\mathcal{A}_H \mathcal{G}_h^{\nu} v_h\|_{L^2(\Omega)} \leqslant \frac{c_G}{\nu h^2} \|v_h\|_{L^2(\Omega)} \quad \forall v_h \in V_h.$$

sowie

Lemma 4.29 (Approximationseigenschaft)**.** *Sei* $\Omega_H \Subset \Omega_h$ *mit* $H \leqslant cH$ *und* $c > 0$ *unabhängig von* h *und* H *zwei geschachtelte Gitter. Für die Grobgitterkorrektur gilt die Approximationseigenschaft*

$$\|(\mathcal{A}_h^{-1} - \mathcal{P}_h \mathcal{A}_H^{-1} \mathcal{R}_H) v_h\|_{L^2(\Omega)} \leqslant c_A h^2 \|v_h\|_{L^2(\Omega)} \quad \forall v_h \in V_h.$$

Mit diesen beiden Sätzen folgt unmittelbar das Konvergenzresultat für die Zweigitter-Iteration:

Lemma 4.30 (Konvergenz der Zweigitter-Iteration)**.** *Es sei* $\mathcal{G}$ *ein Glättungsoperator mit der Eigenschaft von Satz 4.28. Weiter sei* $H \leqslant ch$*. Dann gilt für hinreichend viele Glättungsschritte* $\nu_1 > \nu$ *mit* ν *unabhängig von* h

$$\|\mathcal{B}_{ZG}(\nu)\| \leqslant \rho_{ZG}(\nu) = \frac{c}{\nu} < 1.$$

Proof: Mit Satz 4.28 und Satz 4.29 gilt:

$$\|\mathcal{B}_{ZG}(\nu)\nu_h\| \leqslant \frac{c_A c_G}{\nu}\|\nu_h\| \quad \forall \nu_h \in V_h.$$

Mit $\nu > c_A c_G$ folgt die Aussage des Satzes. □

Wir beweisen zunächst die Glättungseigenschaft:

Beweis von Satz 4.28: Der Operator $\mathcal{A}_h : V_h \to V_h$ mit $(\mathcal{A}_h u_h, \nu_h) = (u_h, \mathcal{A}_h \nu_h)$ ist selbstadjungiert, und hat positive reelle Eigenwerte $0 < \lambda_1 \leqslant \cdots \leqslant \lambda_{N_h}$ und verfügt über ein zugehöriges System aus L^2-orthonormalen Eigenvektoren $\omega_h^{(1)}, \ldots, \omega_h^{(N_h)}$. Jede Funktion $\nu_h \in V_h$ schreiben wir nun in der Form

$$\nu_h = \sum_{i=1}^{N_h} \gamma_i \omega_h^{(i)}, \quad \gamma_i = (\nu_h, \omega_h^{(i)}), \quad \|\nu_h\|^2_{L^2(\Omega)} = \sum_{i=1}^{N_h} \gamma_i^2. \tag{4.23}$$

Mit $\theta := \lambda_{N_h}^{-1}$ ist durch die Richardson-Iteration

$$\mathcal{G}_h := \mathcal{I}_h - \frac{1}{\lambda_{N_h}}\mathcal{A}_h,$$

ein Operator $\mathcal{G}_h : V_h \to V_h$ definiert. Für ein beliebiges $\nu_h \in V_h$ gilt in Schreibweise (4.23)

$$\mathcal{A}_h \mathcal{G}_h^\nu \nu_h = \sum_{i=1}^{N_h} \gamma_i \lambda_i \left(1 - \frac{\lambda_i}{\lambda_{N_h}}\right)^\nu \omega_h^{(i)}.$$

In der Norm:

$$\begin{aligned}
\|\mathcal{A}_h \mathcal{G}_h^\nu \nu_h\|^2 &= \sum_{i=1}^{N_h} \gamma_i^2 \lambda_i^2 \left(1 - \frac{\lambda_i}{\lambda_{N_h}}\right)^{2\nu} \\
&\leqslant \lambda_{N_h}^2 \max_{1 \leqslant i \leqslant N_h} \left\{ \left(\frac{\lambda_i}{\lambda_{N_h}}\right)^2 \left(1 - \frac{\lambda_i}{\lambda_{N_h}}\right)^{2\nu} \right\} \sum_{i=1}^{N_h} \gamma_i^2 \\
&= \lambda_{N_h}^2 \max_{1 \leqslant i \leqslant N_h} \left\{ \left(\frac{\lambda_i}{\lambda_{N_h}}\right)^2 \left(1 - \frac{\lambda_i}{\lambda_{N_h}}\right)^{2\nu} \right\} \|\nu_h\|^2.
\end{aligned}$$

Es gilt $0 < \lambda_1/\lambda_{N_h} \leqslant 1$ und mit der Ungleichung

$$\max_{0 \leqslant x \leqslant 1} \{x(1-x)^\nu\} \leqslant (1+\nu)^{-1}, \quad \nu \geqslant 1$$

folgt

$$\|\mathcal{A}_h \mathcal{G}_h^\nu \nu_h\|^2 \leqslant \lambda_{N_h}^2 (1+\nu)^{-2} \|\nu_h\|^2. \tag{4.24}$$

Für den größten Eigenwert des Operators $\mathcal{A}_h$ gilt der Zusammenhang

$$\lambda_{N_h} \|\omega_h^{(N_h)}\|^2 = (\mathcal{A}_h \omega_h^{(N_h)}, \omega_h^{(N_h)}) = \langle \mathbf{A}_h \boldsymbol{\omega}_h^{(N_h)}, \boldsymbol{\omega}_h^{(N_h)} \rangle \leqslant \lambda_{\max}(\mathbf{A}_h) |\boldsymbol{\omega}_h^{(N_h)}|^2,$$

mit dem Koeffizientenvektor $\boldsymbol{\omega}_h^{(N_h)}$ von $\omega_h^{(N_h)}$ in der Knotenbasisdarstellung. Jetzt gilt

$$|\boldsymbol{\omega}_h^{(N_h)}|^2 = \langle \mathbf{M}_h^{-1}\mathbf{M}_h\boldsymbol{\omega}_h^{(N_h)}, \boldsymbol{\omega}_h^{(N_h)}\rangle \leqslant \lambda_{\min}(\mathbf{M}_h)^{-1}(\omega_h^{(N_h)}, \omega_h^{(N_h)}).$$

Mit $\lambda_{\min}(\boldsymbol{M}_h) = O(h^2)$ sowie $\lambda_{\max}(\boldsymbol{A}_h) = O(1)$ folgt $\lambda_{N_h} \leqslant ch^{-2}$ und aus (4.24) schließlich die Behauptung. □

Es bleibt, die Approximationseigenschaft zu zeigen:

Beweis von Satz 4.29: Es sei $f_h \in V_h$ beliebig. Dann ist $v_h = \mathcal{A}_h^{-1} f_h$ definiert durch

$$a(v_h, \phi_h) = (f_h, \phi_h) \quad \forall \phi_h \in V_h. \tag{4.25}$$

Weiter sei $f_H \in V_H$ gegeben durch $f_H = \mathcal{R}_H v_h$, also

$$(f_H, \phi_H) = (f_h, \phi_H) \quad \forall \phi_H \in V_H.$$

Also ist $v_H = \mathcal{A}_H^{-1} f_H = \mathcal{A}_H^{-1}\mathcal{R}_H f_h$ definiert durch

$$a(v_H, \phi_H) = (f_h, \phi_H) \quad \forall \phi_H \in V_H. \tag{4.26}$$

Schließlich definieren wir eine Funktion $v \in H_0^1(\Omega) \cap H^2(\Omega)$ als Lösung der Randwertaufgabe

$$a(v, \phi) = (f_h, \phi) \quad \forall \phi \in H_0^1(\Omega). \tag{4.27}$$

Für diese Lösung gilt die a priori Abschätzung

$$\|\nabla^2 v\| \leqslant c\|f_h\|. \tag{4.28}$$

Die Lösungen $v_h \in V_h$ von (4.25) und $v_H \in V_H$ (4.26) sind gerade die Galerkin-Approximationen der Lösung $v \in H_0^1(\Omega) \cap H^2(\Omega)$ von (4.27). Es gilt demnach die L^2-Fehlerabschätzung:

$$\|v - v_h\| \leqslant ch^2\|\nabla^2 v\|, \quad \|v - v_H\| \leqslant cH^2\|\nabla^2 v\|.$$

Mit der Annahme $H \leqslant ch$ und der a priori Abschätzung (4.28) folgt

$$\|v_h - v_H\| \leqslant \|v - v_h\| + \|v - v_H\| \leqslant ch^2\|\nabla^2 v\| \leqslant ch^2\|f_h\|.$$

Also folgt für den Grobgitter-Operator

$$\|\mathcal{A}_h^{-1} f_h - \mathcal{P}_h\mathcal{A}_H^{-1}\mathcal{R}_H f_h\| \leqslant ch^2\|f_h\| \quad \forall f_h \in V_h.$$

Dies vervollständigt den Beweis. □

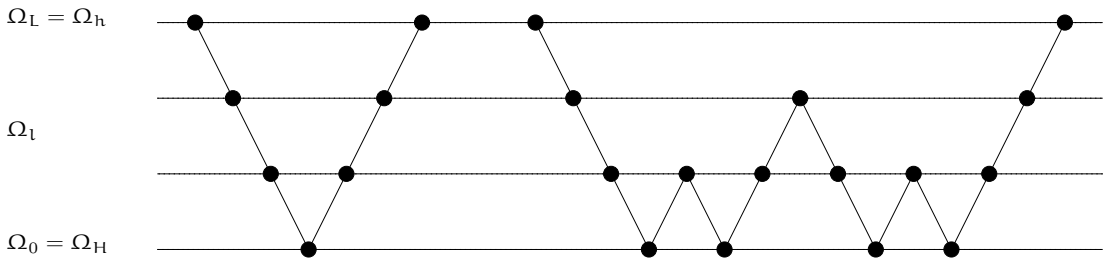


Abbildung 4.8: Schematische Darstellung des *V*- und *W*-Zyklus.

4.3.3 Mehrgitter-Verfahren

Beim Zweigitter-Verfahren erfolgt die Grobgitter-Korrektur durch Lösen eines Systems $\mathbf{A}_H u_H = f_H$. Dieses Problem ist zwar kleiner als das ursprüngliche, kann jedoch immer noch zu groß zum effizienten Lösen sein.

Beim Mehrgitter-Verfahren zur Lösung von $\mathbf{A}_L u_L = f_L$ im Raum $V_h = V_L$ wird zur Lösung des Grobgitter-Problems $A_{L-1} w_{L-1} = r_{L-1}$ wieder das Zweigitter-Verfahren verwendet. Dies geschieht auf rekursive Art bis wir beim Problem auf dem gröbsten Gitter $\mathbf{A}_0 u_0 = r_0$ ankommen. Dieses nun sehr kleine Gleichungssystem kann mit einem direkten Verfahren gelöst werden.

Algorithmus 4.31 (Mehrgitter-Verfahren)**.** *Durch* $\Omega_H = \Omega_0 \Subset \Omega_1 \Subset \ldots \Subset \Omega_L = \Omega_h$ *sei eine Familie von geschachtelten Triangulierungen mit* $h_{l-1} \leqslant c h_l$ *gegeben. Sei* $u_L^{(0)} := u_h^{(0)}$ *ein Startwert. Durch* $\nu_1, \nu_2 \geqslant 0$ *sei die Anzahl der Vor- bzw. Nachglättungsschritte gegeben. Weiter sei* $R \geqslant 1$ *und* $\omega_l \in (0, 1]$ *ein Dämpfungsparameter. Iteriere für* $t \geqslant 0$

$$u_L^{(t+1)} = \mathcal{MG}(L, u_L^{(t)}, f_L),$$

mit der Mehrgitter-Iteration $\mathcal{MG}(l, u_l^{(t)}, f_l)$*:*

$l = 0$		*Direkter Löser:*	$u_0^{(t+1)} = \mathcal{A}_0^{-1} f_0$
$l > 0$	(i)	*Vorglätten:*	$\bar{u}_l = \mathcal{G}_l^{\nu_1}(u_l^{(t)})$
	(ii)	*Residuum:*	$r_l = f_l - \mathcal{A}_l \bar{u}_l$
	(iii)	*Restriktion:*	$r_{l-1} = \mathcal{R}_{l-1} r_l$
	(iv)	*Grobgitterproblem:*	$w_{l-1}^{(0)} = 0$
		$1 \leqslant r \leqslant R$:	$w_{l-1}^{(r)} = \mathcal{MG}(l-1, w_{l-1}^{(r-1)}, r_{l-1})$
	(v)	*Prolongation:*	$w_l = \mathcal{P}_l w_{l-1}^{(R)}$
	(vi)	*Update:*	$\bar{\bar{u}}_l = \bar{u}_l + \omega_l w_l$
	(vii)	*Nachglätten:*	$u_l^{(t+1)} = \mathcal{G}_l^{\nu+2}(\bar{\bar{u}}_l)$

Da die Grobgitter-Korrektur nun lediglich eine Approximation ist, führen wir $R \geqslant 1$ Korrekturschritte aus. R wird üblicherweise sehr klein als $R = 1$ oder als $R = 2$ gewählt. Im Fall $R = 1$ spricht man vom *V*-Zyklus, im Fall $R = 2$ vom *W*-Zyklus der Mehrgitter-Iteration, siehe Abbildung 4.8 für eine Darstellung des Iterationsschemas. Neben dem *V* und *W*-Zyklus existieren weitere Varianten wie der F-Zyklus, beim dem in einer Richtung $R = 1$ und in der anderen Richtung $R = 2$ gewählt wird. Weiter kann es zweckdienlich sein, den Mehrgitterprozess nicht auf dem feinsten Gitter Ω_h sondern auf dem gröbsten Gitter Ω_H zu starten. Iterativ werden mit dem Mehrgitter-Verfahren auf den Gittern $\Omega_0, \Omega_1, \ldots$ Lösungen mit hinreichender Genauigkeit erstellt. Dieses *geschachtelte Mehrgitter-Verfahren* erreicht die optimale Komplexität $O(N_L)$.

Wenn der *V*-Zyklus konvergiert, so ist er sehr effizient. Bei Problemen mit Unsymmetrien, nicht-glatten Koeffizienten, Singularitäten ist der *V*-Zyklus hingegen oft instabil. Hier bietet sich dann der robuste *W*-Zyklus an.

Wir beweisen nun:

Lemma 4.32 (Konvergenz des Mehrgitter-Verfahrens)**.** *Der Zweigitter-Zyklus sei auf jedem Paar der Familie* $V_0 \subset V_1 \subset \cdots \subset V_L$ *konvergent mit* $\rho_{ZG}(\nu) \to 0$ *für* $\nu \to 0$ *gleichmäßig bezüglich* l*. Dann konvergiert für* $\nu \geqslant \nu_0$ *groß genug der Mehrgitteralgorithmus im W-Zyklus mit einer von* L (*und somit* h) *unabhängigen Konvergenzrate* $\rho_{MG} < 1$ *bezüglich der* L^2*-Norm:*

$$\|u_L - \mathcal{MG}(L, u_L^{(t)}, f_L)\| \leqslant \rho_{MG} \|u_L - u_L^{(t)}\|.$$

Proof: (*i*) Wir setzen $\nu_2 = 0$. Wir führen den Beweis per Induktion nach L. Zunächst sei $\nu \geqslant \nu_0$ so gewählt, dass für die Zweigitter-Konvergenz gilt $\rho_{ZG} \leqslant \frac{1}{8}$. Wir wollen zeigen, dass dann gilt $\rho_{MG} \leqslant \frac{1}{4}$. Im Fall $L \leqslant 2$ liegt gerade das Zweigitter-Verfahren vor und die Aussage ist richtig.

(*ii*) Sei nun $L > 2$. Für das Gitterlevel $L - 1$ gilt nach Voraussetzung $\rho_{MG} \leqslant \frac{1}{4}$. Sei $u_L^{(t)}$ die letzte Iteration.

Wir führen nun als Hilfsgröße die Lösung der Zweigitter-Iteration (mit exakter Lösung des Grobgitter-Problems) ein:

$$\bar{u}_L^{(t+1)} = \mathcal{ZG}(L, u_L^{(t)}, f_L).$$

Dann gilt für die Differenz zwischen dieser Lösung und zweimaliger Grobgitter-Korrektur mit dem Mehrgitter-Verfahren ($0 = \omega_{l-1}^{(0)} \xrightarrow{\mathcal{MG}(l-1)} \omega_{l-1}^{(1)} \xrightarrow{\mathcal{MG}(l-1)} \omega_{l-1}^{(2)}$)

$$u_L^{(t+1)} - \bar{u}_L^{(t+1)} = \mathcal{P}_L(w_{l-1}^{(2)} - \bar{w}_{l-1}),$$

wobei $w_{l-1}^{(2)}$ die mit Mehrgitter approximierte Grobgitterlösung ist und $\bar{w}_{l-1}$ die exakte Lösung des Grobgitterproblems. Es gilt:

$$\|w_{l-1}^{(2)} - \bar{w}_{l-1}\| \leqslant \rho_{MG}^2 \|w_{l-1}^{(0)} - \bar{w}_{l-1}\| = \rho_{MG}^2 \|\bar{w}_{l-1}\|, \tag{4.29}$$

da der Startwert in Schritt (*iv*) von Algorithmus 4.31 gerade $w_{l-1}^{(0)} = 0$ ist. Mit

$$\begin{aligned}\bar{w}_{l-1} &= \mathcal{A}_{l-1}^{-1}\mathcal{R}_{l-1}(f_l - \mathcal{A}_l\mathcal{G}_l^{\nu}(u_l^{(t)})) \\ &= \mathcal{A}_{l-1}^{-1}\mathcal{R}_{l-1}\mathcal{A}_l(u_l - \mathcal{G}_l^{\nu}(u_l^{(t)})) \\ &= \mathcal{A}_{l-1}^{-1}\mathcal{R}_{l-1}\mathcal{A}_l\mathcal{G}_l^{\nu}(u_l - u_l^{(t)}),\end{aligned}$$

da $\mathcal{G}_l(u_l) = u_l$ eine Fixpunktiteration ist. Also gilt mit (4.29)

$$\|w_{l-1}^{(2)} - \bar{w}_{l-1}\| \leqslant \rho_{MG}^2 \|\mathcal{A}_{l-1}^{-1}\mathcal{R}_{l-1}\mathcal{A}_l\mathcal{G}_l^{\nu}\| \, \|u_l - u_l^{(t)}\|.$$

Wir schätzen nun die Norm des Grobgitter-Operators mit Hilfe des Zweigitter-Operators ab:

$$\mathcal{A}_{l-1}^{-1}\mathcal{R}_{l-1}\mathcal{A}_l\mathcal{G}_l^{\nu} = \mathcal{G}_l^{\nu} - (\mathcal{A}_l^{-1} - \mathcal{P}_l\mathcal{A}_{l-1}^{-1}\mathcal{R}_{l-1})\mathcal{A}_l\mathcal{G}_l^{\nu} = \mathcal{G}_l^{\nu} - \mathcal{Z}\mathcal{G}_l$$

Hieraus folgt:

$$\|\mathcal{A}_{l-1}^{-1}\mathcal{R}_{l-1}\mathcal{A}_l\mathcal{G}_l^{\nu}\| \leqslant \|\mathcal{G}_l^{\nu}\| + \|\mathcal{Z}\mathcal{G}\| \leqslant 1 + \rho_{ZG} \leqslant 2.$$

Insgesamt erhalten wir:

$$\|u_l - u_l^{(t+1)}\| \leqslant (\rho_{ZG} + 2\rho_{MG}^2)\|u_l - u_l^{(t)}\|.$$

Mit $\rho_{ZG} \leqslant \frac{1}{8}$ und $\rho_{MG} \leqslant \frac{1}{4}$ aus der Induktionsannahme folgt die Aussage des Satzes. □

Im Abschluss betrachten wir nun den numerischen Aufwand in jedem Mehrgitterschritt. Hierzu benötigen wir einige Hilfsgrößen, welche auf jedem Gitter Ω_l den Aufwand der einzelnen Mehrgitterkomponenten bezüglich der Anzahl der Freiheitsgrade messen. Es sei

$$C_0(N_0) := OP(\mathcal{A}_0^{-1})/N_0,$$

der Aufwand zur Grobgitterlösung pro Grobgitter-Freiheitsgrad. Man beachte, dass C_0 von N_0 anhängt, da die Grobgitterlösung im Allgemeinen nicht mit linearer Laufzeit erfolgen kann. Dennoch, da N_0 fest ist, kann $C_0(N_0)$ in diesem Zusammenhang als konstant angesehen werden. Weiter sei $C_s := OP(\mathcal{S}_l)/N_l$ der Aufwand pro Glättungs-Iteration pro Knoten auf dem Gitterlevel Ω_l sowie $C_T := OP(\mathcal{R}_l)/N_l$ der Aufwand zur Restriktion und auch Prolongation pro Freiheitsgrad und $C_r := OP(r_l)/N_l$ der numerische Aufwand zur Berechnung des Residuums. Dann gilt:

Lemma 4.33 (Komplexität des Mehrgitterverfahrens)**.** *Es sei* $\kappa := \max(N_{l-1}/N_l) < 1$, $R = 1$ *oder* $R = 2$ *die Anzahl der Grobgitteriterationen und* ν_1, ν_2 *die Anzahl der Vor- bzw. Nachglättungsschritte. Dann gilt im Fall* $q := \kappa R < 1$ *für den numerischen Aufwand der Mehrgitteriteration*

$$OP(\mathcal{MG}(L)) = \frac{1}{1-q}\left((\nu_1 + \nu_2)C_s + C_r + 2C_T\right)N_L + C_0(N_0)q^L N_L$$

Der Gesamtaufwand zur Reduktion des L^2*-Fehlers bezüglich der* L^2*-Norm auf die Diskretisierungsgenauigkeit* $O(h^2)$ *beträgt daher* $O(N_L \ln(N_L))$.

Proof: Wir betrachten zunächst eine Gitterebene l mit N_l Freiheitsgraden und zählen die auf dieser Ebene notwendigen arithmetischen Operation mit Ausnahme der Grobgitterkorrektur. Mit der Vorbereitung gilt hier:

$$OP(\mathcal{G}_l) = \tilde{C}N_l, \quad \tilde{C} := (\nu_1 + \nu_2)C_s + C_r + 2C_T \tag{4.30}$$

D.h., auf der feinsten Gitterebene gilt rekursiv:

$$\begin{aligned} OP(\mathcal{MG}_L) &= \tilde{C}N_L + R \cdot OP(\mathcal{MG}_{L-1}) \\ &= \tilde{C}N_L + R\tilde{C}N_{L-1} + R^2 \cdot OP(\mathcal{MG}_{L-2}) \\ &= \tilde{C}N_L + R\tilde{C}N_{L-1} + R^2CN_{L-2} + \cdots + R^{L-1}\tilde{C}N_1 + R^L C_0(N_0)N_0, \end{aligned}$$

wobei $C_0(N_0)N_0$ der Aufwand zum Lösen des Grobgitterproblems ist. Mit $N_{l-1} \leqslant \kappa N_l$ und $q := \kappa R < 1$ folgt:

$$\begin{aligned} OP(\mathcal{MG}_L) &\leqslant \tilde{C} \sum_{l=0}^{L-1} (\kappa R)^l N_L + C_0(N_0)(\kappa R)^L N_L \\ &= \tilde{C}\frac{1-q^L}{1-q}N_L + C_0(N_0)q^L N_L \\ &\leqslant \left(\frac{\tilde{C}}{1-q} + C_0(N_0)q^L\right) N_L. \end{aligned}$$

Mit (4.30) folgt die Komplexitätsabschätzung. Zur Abschätzung der Gesamtkomplexität machen wir den Ansatz:

$$\rho_{MG}^t = h_L^2 \sim N_L^{-\frac{2}{d}} \quad \rightarrow \quad t \sim -\frac{\ln(N_L)}{\ln(\rho_{MG})}.$$

□

Die Mehrgitter-Iteration erreicht also fast die optimale Komplexität $O(N_L)$ zum Lösen des Gleichungssystems mit hinreichender Genauigkeit. Der Logarithmische Term ist hierbei in der Anwendung nicht wesentlich. Für den Beweis der optimalen Komplexitätsabschätzung ist die Bedingung $\kappa R < 1$ wesentlich. D.h., im Fall des *W*-Zyklus mit $R = 2$ bedeutet dies $\kappa := \max N_{l-1}/N_l < \frac{1}{2}$. Bei der Verwendung von lokal verfeinerten Gittern ist diese Bedingung oft nicht erfüllt. Um dennoch optimale Mehrgitter-Komplexität zu erhalten muss der Algorithmus modifiziert werden.

5 Die Finite Elemente Methode für parabolische Probleme

Wir betrachten im Folgenden die Wärmeleitungsgleichung auf dem Zeit-Ortsgebiet $I \times \Omega$ mit einem Intervall $I = (0, T)$ und $\Omega \subset \mathbb{R}^d$. Wir suchen die Lösung der Wärmeleitungsgleichung mit homogenen Dirichlet-Randwerten

$$\partial_t u(x,t) - \Delta u(x,t) = f(x,t) \text{ in } I \times \Omega, \quad u = u^0 \text{ für } t = 0, \quad u = 0 \text{ auf } I \times \partial\Omega, \tag{5.1}$$

mit einer rechten Seite $f \in L^2(I; L^2(\Omega))$. Der Fall von nicht-homogenen Randdaten $u = g$ auf $I \times \partial\Omega$ kann wie bei elliptischen Differentialgleichungen in ein Problem mit modifizierter rechter Seite aber homogenen Randdaten übertragen werden. Zu (5.1) korrespondiert die variationelle Formulierung:

$$(\partial_t u(t), \phi) + (\nabla u(t), \nabla \phi) = (f(t), \phi), \quad (u(0), \phi) = (u^0, \phi) \quad \forall \phi \in H_0^1(\Omega) \tag{5.2}$$

mit der Lösung $u \in W$ in

$$u \in W, \quad W := \{v \in L^2(I; H_0^1(\Omega)),\ \partial_t v \in L^2(I; L^2(\Omega))\}, \tag{5.3}$$

Zur Diskretisierung von parabolischen Gleichungen existieren einige verschiedene Ansätze:

Orts-Zeit-Diskretisierung Die Diskretisierung erfolgt gleichzeitig in Ort und Zeit. Bei einer Finite Differenzendiskretisierung wird hierbei das Gebiet $I \times \Omega$ in ein Orts-Zeit-Gitter $I_h \times \Omega_h$ zerlegt und wir suchen in den diskreten Punkten $(t_i, x_j)_{i,j}$ eine Differenzenapproximation zu (5.1).

Alternativ kann die Lösung auch mit einem globalen Galerkin-Ansatz approximiert werden. Hierzu wird ein diskreter Teilraum $W_{kh} \subset W$ gewählt und die Lösung $u_{kh} \in W_{kh}$ gesucht, so dass sie die schwache Formulierung erfüllt. Diesen Ansatz werden wir später näher untersuchen. Er hat seine Stärken in der einfachen mathematischen Untersuchung.

Linienmethode Bei der Linienmethode wird die Wärmeleitungsgleichung zunächst im Ort diskretisiert. Dies geschieht zum Beispiel mit der Finite Elemente Methode. D.h., die variationelle Formulierung (5.2) wird im Ort mit einem Finite Elemente Ansatz $V_h \subset H_0^1(\Omega)$ diskretisiert und wir erhalten eine Ortsdiskrete Funktion

$$u_h : I \to V_h.$$

Wir suchen also die örtliche diskrete, aber in der Zeit kontinuierliche Lösung $u_h(t)$ von

$$(\partial_t u_h(t), \phi_h)_{\Omega_h} + (\nabla u_h(t), \nabla \phi_h)_{\Omega_h} = (f(t), \phi_h) \quad \forall \phi_h \in V_h.$$

Mit der üblichen Schreibe für die Massen- sowie Steifigkeitsmatrix und der Koeffizientendarstellung der Lösung

$$u_h(x,t) = \sum_{i=1}^{N_h} \mathbf{u}_i(t)\phi_h(x),$$

wird die Wärmeleitungsgleichung überführt in ein System von linearen Anfangswertaufgaben:

$$\mathbf{M}_h \mathbf{u}_h'(t) + \mathbf{A}_h \mathbf{u}_h(t) = \mathbf{b}_h(t), \quad \mathbf{u}_h(0) = \mathbf{u}_h^0, \tag{5.4}$$

mit rechter Seite $\mathbf{b}_h$ und Startwert $\mathbf{u}_h^0$ als

$$\mathbf{b}_h(t) = (\mathbf{b}_i(t))_{i=1}^{N_h}, \quad \mathbf{b}_i(t) = (f(t), \phi_h^{(i)})_{\Omega_h},$$

und

$$\mathbf{u}_h^0 = (\mathbf{u}_i^0)_{i=1}^{N_h}, \quad \mathbf{u}_i^0 = (u^0, \phi_h^{(i)})_{\Omega_h}.$$

Das System (5.4) kann nun mit einem üblichen Differenzenverfahren zur Approximation von Anfangswertaufgaben in der Zeit approximiert werden. Hierbei sollten A-stabile Verfahren genutzt werden, da durch die Kondition der Steifigkeitsmatrix $\operatorname{cond}_2(\mathbf{A}_h) = O(h^{-2})$ die ODE sehr steif ist.

Dieser Verfahrensansatz ist einfach zu analysieren, da die Theorie der gewöhnlichen Differentialgleichungen unmittelbar angewendet werden kann. Die Linienmethode ist jedoch wenig flexibel. Insbesondere baut die Linienmethode auf einer örtlichen Diskretisierung der Gleichung auf, welche für alle Zeitschritte fest ist. Es können also nicht unterschiedliche Gitter zu unterschiedlichen Zeitpunkten gewählt werden. Bei instationären Prozessen ist dies allerdings oft wesentlich.

Rothe-Methode Die Rothe-Methode geht umgekehrt vor. Zunächst diskretisieren wir die Wärmeleitungsgleichung in der Zeit durch Einführen eines Zeitgitters

$$t_0 < t_1 < \cdots < t_M,$$

mit der zeitdiskreten Funktion $u_k = (u_k^m)_{m=0}^M$, $u_k^m \in H_0^1(\Omega)$. Mit dem impliziten Euler-Verfahren erhalten wir

$$u_k^0 = u^0, \quad u_k^m - k_m \Delta u_k^m = u_k^{m-1} + k_m \bar{f}^m, \quad m = 1, \ldots, M.$$

Hier ist $\bar{f}^m$ eine Approximation zu $f(t_m)$, welche meistens zur besseren Genauigkeit im zeitlichen Mittel ausgewertet wird, etwa

$$f^m = k_m^{-1} \int_{t_{m-1}}^{t_m} f(t)\,dt.$$

Wir erhalten also eine Abfolge von *quasi-stationären* partiellen Differentialgleichungen, welche nun im Ort diskretisiert werden. Hier betrachten wir im Ort die Finite Elemente Methode, indem zu jedem Zeitschritt t_m eine Triangulierung Ω_h^m von Ω und ein entsprechender Finite Elemente Raum $V_h^m \subset H_0^1(\Omega)$ erstellt wird. Dann ist in jedem Zeitschritt das folgende Galerkin-Problem zu lösen:

$$u_{kh}^m \in V_h^m : \quad (u_{kh}^m, \phi_h) + k_m(\nabla u_{kh}^m, \nabla \phi_h) = (u_{kh}^{m-1}, \phi_h) + k_m(\bar{f}^m, \phi_h), \quad \forall \phi_h \in V_h^m, \tag{5.5}$$

sowie für den Startwert

$$u_{kh}^0 \in V_h^0 : \quad (u_{kh}^0, \phi_h) = (u^0, \phi_h) \quad \forall \phi_h \in V_h^0.$$

Eine detaillierte mathematische Analyse der verschiedenen Diskretisierungsmethoden für die Wärmeleitungsgleichung ist sehr komplex. Jeder der drei Ansätze hat verschiedene Vorteile bei der Untersuchung. Wir werden deshalb die grundsätzlichen Fragen nach *Existenz*, *Konvergenz* und *Stabilität* mit zum Teil verschiedenen Ansätzen untersuchen. Es lässt sich aber Zeigen, dass die drei Ansätze eng miteinander verwandt sind. Oft handelt es sich - bis auf einen numerischen Integrationsfehler - um äquivalente Formulierungen.

5.1 Die Rothe-Methode für parabolische Differentialgleichungen

Wir betrachten die Rothe-Methode zum Lösen der Wärmeleitungsgleichung. Suche u mit

$$\partial_t u - \Delta u = f, \qquad u\big|_{t=0} = u^0, \quad u\big|_{\partial\Omega} = 0. \tag{5.6}$$

Zur Diskretisierung zerlegen wir zunächst das Zeitintervall $I = [0, T]$ in diskrete Zeitpunkt

$$0 = t_0 < t_1 < \cdots < t_M = T, \quad k_m := t_m - t_{m-1},$$

Auf diesen Zeitpunkten definieren wir durch $u_k = (u_k^m)_{m=0}^M$ eine zeitdiskrete Funktion.Gleichung (5.6) wird nun durch ein beliebiges Einschritt-Verfahren diskretisiert. Prototypisch betrachten wir:

Explizites Euler-Verfahren

$$u_k^0 = u^0, \quad u_k^m = u_k^{m-1} + k_m \Delta u_k^{m-1} + k_m \bar{f}^m$$

Implizites Euler-Verfahren

$$u_k^0 = u^0, \quad u_k^m - k_m \Delta u_k^m = u^0 + k_m \bar{f}^m$$

Crank-Nicolson-Verfahren

$$u_k^0 = u^0, \quad u_k^m - k_m \frac{1}{2} \Delta u_k^m = u^0 + k_m \frac{1}{2} \Delta u_k^{m-1} + k_m \bar{f}^m.$$

Diese drei Verfahren lassen sich mit einem Parameter $\theta \in [0,1]$ einheitlich mit der Vorschrift

$$u_k^0 = u^0, \quad u_k^m - k_m \theta \Delta u_k^m = u^0 + k_m(1-\theta)\Delta u_k^{m-1} + k_m f^m \tag{5.7}$$

formulieren. Diese Einschrittmethode wird das allgemeine *Theta-Verfahren* genannt. Für $\theta = 0$ ist das Verfahren explizit, d.h. im Kontext von parabolischen Gleichungen, dass der elliptische Operator $L = -\Delta$ nur an Stelle des alten Zeitschrittes u_k^{m-1} ausgewertet werden muss. Die rechte Seite $\bar{f}^m$ wird jeweils in einem geeigneten Mittel ausgewertet.

Zur Ortsdiskretisierung leiten wir gemäß (5.2) eine schwache Formulierung des Theta-Verfahrens her. Für $m = 0, \dots, M$, suche $u_k^m \in H_0^1(\Omega)$ so dass

$$\begin{aligned} (u_k^0, \phi) &= (u^0, \phi) \quad \forall \phi \in H_0^1(\Omega) \\ (u_k^m, \phi) + k_m\theta(\nabla u_k^m, \nabla\phi) &= (u_k^{m-1}, \phi) - k_m(1-\theta)(\nabla u_k^{m-1}, \nabla\phi) + k_m(\bar{f}^m\phi) \quad \forall \phi \in H_0^1(\Omega). \end{aligned} \tag{5.8}$$

Wie bei gewöhnlichen Differentialgleichungen handelt es sich um ein Zeitschritt-Verfahren. Die Lösung u_k^m hängt einzig von den Problemdaten und vom vorherigen Zeitschritt u_k^{m-1} ab. In jedem Schritt ist also eine *quasi-stationäre* partielle Differentialgleichung mit dem Differentialoperator

$$L_{k,\varepsilon} := -k\theta\Delta + \mathrm{id},$$

zu lösen, welcher für alle $k\theta > 0$ elliptisch ist.

Wir diskretisieren, indem wir das Gebiet Ω zu jedem Zeitpunkt t_m triangulieren Ω_h^m. Auf jeder dieser - variierenden - Triangulierungen erstellen wir einen Finite Elemente Raum $V_h^m \subset H_0^1(\Omega)$. Auch wenn es theoretische möglich wäre, in jedem Zeitschritt einen anderen Finite Elemente Ansatz zu wählen, nehmen wir an, dass alle Räume alle vom gleichen parametrischen Typ sind, also dass z.B. V_h^m ein stückweise linearer Finite Elemente Raum für alle m ist.

Das diskrete Theta Verfahren sucht nun die Lösung $u_{kh} = (u_{kh}^m)_{m=0}^M$ von:

$$\begin{aligned} (u_{kh}^0, \phi_0) &= (u^0, \phi_0) \quad \forall \phi_0 \in V_h^0 \\ (u_{kh}^m, \phi_m) + k_m\theta(\nabla u_{kh}^m, \nabla\phi_m) &= (u_{kh}^{m-1}, \phi_m) - k_m(1-\theta)(\nabla u_{kh}^{m-1}, \nabla\phi_m) \quad \forall \phi_m \in V_h^m. \end{aligned} \tag{5.9}$$

In jedem Zeitschritt stellen wir die Lösung u_{kh}^m in der üblichen Basisdarstellung dar:

$$u_{hk}^m(x) = \sum_{i=1}^{N_m} \mathbf{u}_i^m \phi_m^{(i)}(x), \quad \mathbf{u}_h^m = (\mathbf{u}_i^m)_{i=1}^{N_m}.$$

Jetzt können wir mit der L^2-Projektion $\mathcal{P}_h^m : L^2(\Omega) \to V_h^m$ Gleichung (5.9) schreiben als:

$$\mathbf{u}_h^0 = \mathcal{P}_h^0 u^0, \quad (\mathbf{M}_h^m + \theta k_m \mathbf{A}_h^m)\,\mathbf{u}_h^m = \left(\mathbf{M}_h^{m-1,m} - (1-\theta)k_m \mathbf{A}_h^{m-1,m}\right)\mathbf{u}_h^{m-1} + k_m \mathbf{M}_h^m \mathcal{P}_h^m \bar{f}^m.$$

Dabei sind $\mathbf{A}_h^m \in \mathbb{R}^{N_m \times N_m}$ und $\mathbf{M}_h^m \in \mathbb{R}^{N_m \times N_m}$ Steifigkeits- und Massenmatrix im Raum V_h^m:

$$\mathbf{A}_h^m = (\mathbf{A}_{ij}^m)_{i,j=1}^{N_m}, \quad \mathbf{A}_{ij}^m = (\nabla\phi_m^{(j)}, \nabla\phi_m^{(i)}), \quad \mathbf{M}_h^m = (\mathbf{M}_{ij}^m)_{i,j=1}^{N_m}, \quad \mathbf{M}_{ij}^m = (\phi_m^{(j)}, \phi_m^{(i)}).$$

Die Matrizen $\mathbf{M}_h^{m-1,m} \in \mathbb{R}^{N_m \times N_{m-1}}$ sowie $\mathbf{A}_h^{m-1,m} \in \mathbb{R}^{N_m \times N_{m-1}}$ sind entsprechende Massen- und Steifigkeits-Transfermatrizen. Hier kommen die Testfunktionen aus dem neuen Raum V_h^m und die Ansatzfunktionen aus dem vorherigen Raum V_h^{m-1}:

$$\begin{aligned}\mathbf{A}_h^{m-1,m} &= (\mathbf{A}_{ij}^{m-1,m})_{i,j=1}^{N_m,N_{m-1}}, \quad \mathbf{A}_{ij}^{m-1,m} = (\nabla\phi_{m-1}^{(j)}, \nabla\phi_m^{(i)}) \\ \mathbf{M}_h^{m-1,m} &= (\mathbf{M}_{ij}^{m-1,m})_{ij=1}^{N_m,N_{m-1}}, \quad \mathbf{M}_{ij}^{m-1,m} = (\phi_{m-1}^{(j)}, \phi_m^{(i)}).\end{aligned}$$

Das Verfahren vereinfacht sich, wenn $V_h^m = V_h$ für alle Zeitschritte gleich gewählt wird. Dann gilt in jedem Schritt:

$$(\mathbf{M}_h + \theta k_m \mathbf{A}_h)\mathbf{u}_h^m = (\mathbf{M}_h - (1-\theta)k_m\mathbf{A}_h)\mathbf{u}_h^{m-1} + k_m\mathbf{M}_h\mathcal{P}_h\bar{f}^m.$$

mit der Steifigkeits und Massenmatrix im Raum V_h sowie der L^2-Projektion $\mathcal{P}_h : L^2(\Omega) \to V_h$.

Der Vorteil der Rothe-Methode ist die große Flexibilität. In jedem Zeitschritt kann ein speziell angepasstes Ortsgitter verwendet werden.

Die Konvergenzanalyse ist sehr aufwändig. Für das implizite Euler-Verfahren gilt

Lemma 5.1 (Implizites Euler-Verfahren)**.** *Für das implizite Euler-Verfahren mit Verbindung einer linearen Finite Elemente Diskretisierung im Ort gelten die Fehlerabschätzungen:*

1. *Für allgemeine, variierende Ortsgitter*

$$\max_{1\leqslant m\leqslant M}\|e_{kh}^m\| \leqslant cT^{\frac{1}{2}} \max_{0\leqslant m\leqslant M}\{k_m^{-\frac{1}{2}}h_m^2\|\nabla^2 u^m\|\} + c\left(\sum_{m=1}^{M} k_m^2 \int_{t_{m-1}}^{t_m} \|\nabla\partial_t u\|^2\,dt\right)^{\frac{1}{2}}$$

2. *Für feste Ortsgitter mit $V_h^m = V_h$:*

$$\max_{1\leqslant m\leqslant M}\|e_{kh}^m\| \leqslant cT^{\frac{1}{2}} \max_{0\leqslant m\leqslant M}\{h^2\|\nabla^2 u^m\|\} + c\left(\sum_{m=1}^{M} k_m^2 \int_{t_{m-1}}^{t_m} \|\nabla\partial_t u\|^2\,dt\right)^{\frac{1}{2}},$$

jeweils mit dem Fehler $e_{k,h} = u - u_{k,h}$.

Für den zweiten Fall mit festen Ortsgittern gilt falls $k = k_m$ uniform gewählt wird

$$\|e_{kh}^m\| = O(h^2 + k),$$

wir erhalten also die bekannte lineare Ordnung des impliziten Euler-Verfahrens in der Zeit und die quadratische Ordnung h^2 für den L^2-Fehler der linearen Finite Elemente Approximation im Ort.

Auf wechselnden Gittern ist der Faktor $k_m^{-\frac{1}{2}} h_m^2$ nicht optimal. Fordert man die Schrittweitenbedingung

$$k \approx h^{\frac{4}{3}},$$

so erhält man einen balancierten Fehler

$$\|e_{kh}^m\| = O(k + h^2 k^{-\frac{1}{2}}) = O(h^{\frac{4}{3}} + h^2 h^{-\frac{2}{3}}) = O(h^{\frac{4}{3}}).$$

Im Fall $V_h^{m-1} \subset V_h^m$ für alle m, wenn also die Gitter von Zeitschritt zu Zeitschritt nur feiner werden, aber nie gröber, so kann wieder das optimale Resultat gezeigt werden. Für einen allgemeinen Beweis dieses Satzes verweisen wir auf [Rannacher].

Wir werden einen später alternativen Beweis nachtragen. Dazu machen wir einen Umweg und zeigen die Verwandtschaft des impliziten Euler-Verfahrens zu einer Galerkin-Diskretisierung in Ort und Zeit.

5.1.1 Praktische Aspekte der Rothe-Methode

In jedem Zeitschritt der Rothe-Methode muss ein lineares Gleichungssystem gelöst werden:

$$\mathbf{u}_h^m \in \mathbb{R}^{N_m}: \quad (\mathbf{M}_h^m + \theta k_m \mathbf{A}_h^m)\, \mathbf{u}_h^m = \left(\mathbf{M}_h^{m-1,m} - (1-\theta) k_m \mathbf{A}_h^{m-1,m}\right) \mathbf{u}_h^{m-1} + k_m \mathbf{b}_h^m$$

Hierzu sind die folgenden Schritte notwendig

1. Aufbauen der rechten Seite

$$\mathbf{b}_h^m = (\mathbf{b}_i^m)_{i=1}^{N_m}, \quad \mathbf{b}_i^m = (\bar{f}^m, \phi_m^i).$$

2. Einfluss des alten Zeitschritts

$$(\mathbf{M}_h^{m-1,m} \mathbf{u}_h^{m-1})_{i=1}^{N_m} = (\mathbf{u}_h^{m-1}, \phi_m^{(i)}).$$

3. Falls $\theta < 1$

$$(\mathbf{A}_h^{m-1,m} \mathbf{u}_h^{m-1})_{i=1}^{N_m} = (\nabla \mathbf{u}_h^{m-1}, \nabla \phi_m^{(i)}).$$

4. Aufbau der Matrix

$$\mathbf{M}_h^m + \theta k_m \mathbf{A}_h^m.$$

5. Lösen des linearen Gleichungssystems

Im Folgenden beschreiben wir die notwendigen Schritte. Der Aufbau der rechten Seite unterscheidet sich nicht vom Vorgehen bei elliptischen partiellen Differentialgleichungen. Auch die Systemmatrix wird aus Steifigkeitsmatrix $\mathbf{A}_h$ und Massenmatrix $\mathbf{M}_h$ zusammengesetzt. Dies entspricht der Tatsache, dass jeder Zeitschritt gerade die Diskretisierung eines elliptischen Differentialoperators $L = -k_m\theta\Delta + \mathrm{id}$ ist. Dies bedeutet auch, dass zum Lösen des Systems in Schritt 5. die bekannten Verfahren wie CG oder auch das Mehrgitterverfahren genutzt werden können.

Der Gittertransfer Falls $V_h^m \neq V_h^{m-1}$ gehen zur Berechnung von Schritten 2. und 3. beide Finite Elemente Räume ein. Die Funktion u_{kh}^{m-1} stammt aus V_h^{m-1} und hat die Darstellung

$$u_{kh}^{m-1} = \sum_{i=1}^{N_{m-1}} \mathbf{u}_i^{m-1}\phi_{m-1}^{(i)} \quad \phi_{m-1}^{(i)} \in V_h^{m-1}$$

und z.B. in Schritt 2. müssen die Produkte

$$\sum_{i=1}^{N_{m-1}} \mathbf{u}_i^{m-1}(\phi_{m-1}^{(i)}, \phi_m^{(j)}), \quad j = 1, \ldots, N_m$$

berechnet werden. Dies führt zu praktischen Problemen, da die Knotenbasisfunktionen auf den jeweiligen Gittern definiert sind. Im einfachen Fall, dass $V_h^{m-1} \subset V_h^m$, dass das Gitter Ω_h^m also eine Verfeinerung von Ω_h^{m-1} ist, kann jede Basisfunktion $\phi_{m-1}^{(i)} \in V_h^{m-1}$ durch Basisfunktionen aus V_h^m dargestellt werden. Siehe hierzu die entsprechende Diskussion beim Gittertransfer des Mehrgitter-Verfahrens. In diesem Fall kann die alte Lösung $\mathbf{u}_h^{m-1}$ zunächst auf dem alten Gitter integriert werden

$$\mathbf{X}_i^{m-1} = (\mathbf{u}_i^{m-1}, \phi_{m-1}^{(i)}) \quad i = 1, \ldots, N_{m-1},$$

anschließend wird dieser Vektor auf das feine Gitter prolongiert

$$\mathbf{X}_h^m = \mathbf{R}_{m-1}^T \mathbf{X}_h^{m-1}.$$

Ein ähnliches Vorgehen lässt sich herleiten, wenn $V_h^m \subset V_h^{m-1}$, wenn die Gitter also stets gröber werden. Falls die Gitter jedoch von Schritt zu Schritt verfeinert und vergröbert werden, oder falls in jedem Schritt überhaupt ein gänzlich neues Gitter verwendet wird, muss im Allgemeinen eine L^2-Projektion von u_h^{m-1} in V_h^m berechnet werden.

Die Matrix Die Steifigkeitsmatrix $\mathbf{A}_h^m$ ist gerade die Matrix des Poisson-Problems. In Abschnitt 3.5.2 haben wir bereits die Eigenwerte dieser Matrix sowie der Massenmatrix untersucht. Zusammen ergibt sich für die Diskretisierung der Wärmeleitungsgleichung (im Fall zweidimensionaler Gebiete):

$$\lambda_{\max}(\mathbf{M}_h + \theta k\mathbf{A}_h) = O\left(h^2 + \theta k\right), \quad \lambda_{\min}(\mathbf{M}_h + \theta k\mathbf{A}_h) = O\left(h^2 + \theta k h^2\right),$$

und also für die Kondition

$$\operatorname{cond}_2(\mathbf{M}_h + \theta k \mathbf{A}_h) = O\left(\frac{h^2 + \theta k}{1 + \theta k} h^{-2}\right).$$

Im Fall von großen Zeitschritten $k = O(1)$ bezogen auf die Ortsgitterweite h gilt also die bekannte Kondition $\operatorname{cond}_2 = O(h^{-2})$. Wird die Zeitschrittweite allerdings an die Ortsgitterweite gekoppelt, so verbessert sich die Gesamtkondition. Die Fehlerabschätzung für das implizite Euler-Verfahren in Satz 5.1 legt die Wahl $k \approx h^2$ nahe. Für das Crank-Nicolson Verfahren erhalten wir optimal $\|e_{kh}^m\| = O(k^2 + h^2)$ und mit $k \approx h$ einen äquilibrierten Orts-Zeitfehler. Zusammengefasst gilt:

$$\begin{aligned}
\operatorname{cond}_2(\mathbf{M}_h + \theta k \mathbf{A}_h) &= O\left(\frac{h^2 + \theta k}{1 + \theta k} h^{-2}\right) = O(h^{-2}) && (k = O(1)) \\
\operatorname{cond}_2(\mathbf{M}_h + \theta k \mathbf{A}_h) &= O\left(\frac{h^2 + \theta h}{1 + \theta h} h^{-2}\right) = O(h^{-1}) && (k = O(h)) \\
\operatorname{cond}_2(\mathbf{M}_h + \theta k \mathbf{A}_h) &= O\left(\frac{h^2 + \theta h^2}{1 + \theta h^2} h^{-2}\right) = O(1) && (k = O(h^2))
\end{aligned}$$

Je kleiner die Zeitschrittweite (in Bezug auf die Ortsgitterweite h) umso einfacher ist also das Lösen der algebraischen Gleichungen. Diesen Umstand kann man mit einem Vergleich zu Finite Differenzen-Approximationen der Wärmeleitungsgleichung erklären. Hier ist in jedem Schritt die Matrix

$$I + \theta k \mathbf{A}_h$$

zu invertieren. An die Stelle der Massenmatrix tritt die Einheitsmatrix, welche trivial zu invertieren ist. Für $k \to 0$ ist die Gesamtmatrix nur eine kleine Störung der Einheitsmatrix.

Bei der Finite Elemente Methode ist selbst bei rein expliziten Verfahren wie dem expliziten Euler-Verfahren ist in jedem Schritt ein Gleichungssystem zu lösen

$$\mathbf{M}_h \mathbf{u}_h^m = (\mathbf{M}_h - k\mathbf{A}_h)\mathbf{u}_h^{m-1}. \tag{5.10}$$

Als Matrix ist die Massenmatrix mit Kondition $\operatorname{cond}_2(\mathbf{M}_h) = O(1)$. Aus Stabilitätsgründen (welche wir weiter unten diskutieren) müssen explizite Verfahren einer Schrittweitenbedingung $k \leqslant ch^2$ genügen. Unter dieser Schrittweitenbedingung haben jedoch auch die impliziten Verfahren eine Matrix mit Kondition der Größenordnung $O(1)$, d.h., der Vorteil einer expliziten Methode fällt hier geringer aus.

Werden lineare Finite Elemente zur Diskretisierung im Ort verwendet und sind alle Innenwinkel der Dreiecke stumpf also $\alpha_i \leqslant \pi/2$, so ist die Steifigkeitsmatrix $\mathbf{A}_h$ eine *M-Matrix*, es gilt:

$$\mathbf{A}_{ii} > 0, \quad \mathbf{A}_{ij} \leqslant 0 \ (i \neq j), \quad (\mathbf{A}_h^{-1})_{ij} \geqslant 0.$$

Aus dieser Eigenschaft folgt das *diskrete Maximumsprinzip*, insbesondere die *inverse Monotonie*:

$$\mathbf{A}_h x \leqslant 0 \quad x \leqslant 0 \quad \text{(elementweise)}.$$

Das diskrete Maximumprinzip garantiert, dass die diskrete Lösung wichtige Eigenschaften der kontinuierlichen Lösung u widerspiegelt. Darüber hinaus garantiert die Eigenschaft M-Matrix zu sein die Konvergenz grundlegender iterativer Lösungsmethoden. Auf einem uniformen Dreiecksgitter hat die Massenmatrix die Stencil-Notation:

$$\mathbf{M}_h = \frac{h^2}{12}\begin{bmatrix} 1 & 1 & 0 \\ 1 & 6 & 1 \\ 0 & 1 & 1 \end{bmatrix},$$

und ist mit positiven Nebendiagonaleinträgen keine M-Matrix. Für die zusammengesetzte Matrix $\mathbf{M}_h + k\theta\mathbf{A}_h$ gilt diese Eigenschaft auch nicht zwingend für jede Kombination aus Schrittweite k und Ortsgitterweite h.

Um diese Eigenschaft immer garantieren zu können greift man oft zu einem Trick beim Aufstellen der Massenmatrix. Die Einträge werden nicht exakt, sondern mit der Trapez-Regel integriert:

$$I_T(v) = \frac{|T|}{3}\sum_{i=1}^{3} v(x_i),$$

mit den drei Eckpunkten des Dreiecks T. So ist die Massenmatrix stets diagonal und z.B. auf dem gleichmäßigen Dreiecksgitter gilt:

$$\mathbf{M}_h^{\text{lump}} = h^2\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Dieser Prozess wird *Massen-Lumping* genannt. Auf diese Weise kann auch der Aufwand zum Lösen des Systems (5.10) bei expliziten Verfahren wesentlich reduziert werden. Die Trapezregel hat bei der Integration einen Fehler der Ordnung $O(h^2)$. Dieser Integrationsfehler schlägt sich in gleicher Größenordnung auf den Gesamtfehler durch.

5.1.2 Stabilitätsanalyse

Zur Untersuchung der numerischen Stabilität von Zeitdiskretisierungen nehmen wir vereinfacht an, dass die Ortsgitter in jedem Zeitschritt gleich gewählt werden, also $\Omega_h = \Omega_h^m$ und auch $V_h = V_h^m$ für $m = 0, \dots, M$. In diesem Fall sind Rothe- und Linienmethode äquivalent, die voll diskrete Lösung v_{kh} kann als Diskretisierung eines Systems von gewöhnlichen Differentialgleichungen in der Zeit verstanden werden,

$$u_h : \bar{I} \to V_h : \quad u_h(0) = u_h^0, \quad \mathbf{M}_h u_h'(t) + \mathbf{A}_h u_h(t) = \bar{f}_h(t) \quad t \in I,$$

wobei u_h^0 die L^2-Projektion von $u^0 \in L^2(\Omega)$ in den Ansatzraum V_h ist. Nach Multiplikation mit der inversen Massenmatrix gilt

$$u_h(0) = u_h^0, \quad u_h'(t) + \mathbf{M}_h^{-1}\mathbf{A}_h v_h(t) = \mathbf{M}_h^{-1}\bar{f}_h(t) \quad t \in I. \tag{5.11}$$

Zur Untersuchung der numerischen Stabilität einer Anfangswertaufgabe haben wir das skalare Modellproblem

$$u(0) = 1, \quad u'(t) + \lambda u(t) = 0 \quad t \geqslant 0, \tag{5.12}$$

mit der Lösung $u(t) = \exp(-\lambda t)$ betrachtet. Für $\lambda \in \mathbb{C}$ mit $\mathrm{Re}(\lambda) \geqslant 0$ ist $u(t)$ für alle t beschränkt. Wir definieren

Definition 5.2 (Absolut stabil)**.** *Ein Differenzenverfahren heißt* absolut stabil *für ein* $\lambda k \neq 0$*, wenn sie angewendet auf das Modellproblem* (5.12) *für alle* $\mathrm{Re}(\lambda) \geqslant 0$ *beschränkte Lösungen* $\sup_{m \geqslant 0} \|u^m\|$ *erzeugt.*

Das Verfahren heißt also stabil, wenn es die Beschränktheit der Lösung auch numerisch wiedergibt.

Bei einem System von gewöhnlichen Differentialgleichungen wie (5.11) muss zu einer gewählten Schrittweite k der Faktor $-\lambda_i k$ für alle Eigenwerte λ_i von $\mathbf{M}_h^{-1}\mathbf{A}_h$ im Stabilitätsgebiet gelten. Diese Herleitung beruht auf der Annahme, dass das System diagonalisierbar ist und in Komponenten $\omega_i'(t) + \lambda_i \omega_i(t) = f_i(t)$ zerfällt. Jede dieser Komponenten muss stabil gelöst werden.

Um die Stabilität beziehungsweise Instabilität eines Systems besser charakterisieren zu können haben wir den Begriff der *Steifheit* eingeführt. Dieser ist durch den Quotienten

$$\frac{\max_{\mathrm{Re}(\lambda)\geqslant 0} |\lambda|}{\min_{\mathrm{Re}(\lambda)\geqslant 0} |\lambda|}, \quad \lambda \text{ Eigenwerte von } \mathbf{M}_h^{-1}\mathbf{A}_h,$$

zwischen Betrag des größten sowie kleinsten Eigenwerts mit positivem Realteil gegeben. Für die Matrix $\mathbf{M}_h^{-1}\mathbf{A}_h$ aus (5.11) zeigen wir:

Lemma 5.3. *Die Matrix* $\mathbf{M}_h^{-1}\mathbf{A}_h$ *hat positive Eigenwerte und es gilt*

$$\lambda_{max}(\mathbf{M}_h^{-1}\mathbf{A}_h) = O(h^{-2}), \quad \lambda_{min}(\mathbf{M}_h^{-1}\mathbf{A}_h) = O(1),$$

und also

$$cond_2(\mathbf{M}_h^{-1}\mathbf{A}_h) = O(h^{-2}).$$

(*i*) Sei $\lambda \in \mathbb{R}$ ein Eigenwert und $\mathbf{w}_h$ der zugehörige Eigenvektor von $\mathbf{M}_h^{-1}\mathbf{A}_h$. Dann gilt

$$\mathbf{M}_h^{-1}\mathbf{A}_h\mathbf{w}_h = \lambda_h\mathbf{w}_h \quad \Leftrightarrow \quad (\mathbf{M}_h - \lambda\mathbf{A}_h)\omega = 0 \quad \Rightarrow \quad \lambda = \frac{\langle \mathbf{A}_h\mathbf{w}_h, \mathbf{w}_h\rangle}{\langle \mathbf{M}_h\mathbf{w}_h, \mathbf{w}_h\rangle}$$

Hieraus folgt zunächst (da $\mathbf{A}_h$ und $\mathbf{M}_h$ symmetrisch positiv definit) dass alle Eigenwerte positiv sind.

(*ii*) Für größten und kleinsten Eigenwert gilt nun

$$\lambda_{\min}(\mathbf{M}_h^{-1}\mathbf{A}_h) = \min_{x\in\mathbb{R}^N} \frac{\langle \mathbf{A}_h x_h, x_h\rangle}{\langle \mathbf{M}_h x_h, x_h\rangle}, \quad \lambda_{\max}(\mathbf{M}_h^{-1}\mathbf{A}_h) = \max_{x\in\mathbb{R}^N} \frac{\langle \mathbf{A}_h x_h, x_h\rangle}{\langle \mathbf{M}_h x_h, x_h\rangle}. \tag{5.13}$$

Denn Berechnen des stationären Punktes liefert für alle $\mathbf{y}_h \in \mathbb{R}^N$:

$$\begin{aligned} 0 &= \frac{d}{ds}\left(\frac{\langle \mathbf{A}_h(\mathbf{x}_h + s\mathbf{y}_h), \mathbf{x}_h + s\mathbf{y}_h\rangle}{\langle \mathbf{M}_h(\mathbf{x}_h + s + \mathbf{y}_h), (\mathbf{x}_h + s + \mathbf{y}_h)\rangle}\right)\Big|_{s=0} \\ &= \frac{2\langle \mathbf{A}_h\mathbf{x}_h, \mathbf{y}_h\rangle\langle \mathbf{M}_h\mathbf{x}_h, \mathbf{x}_h\rangle - 2\langle \mathbf{M}_h\mathbf{x}_h, \mathbf{y}_h\rangle\langle \mathbf{A}_h\mathbf{x}_h, \mathbf{x}_h\rangle}{\langle \mathbf{M}_h\mathbf{x}_h, \mathbf{x}_h\rangle^2} \end{aligned}$$

also mit (5.13) z.B. für $\lambda_{\min}$ nach Multiplikation mit $\langle \mathbf{M}_h\mathbf{x}_h, \mathbf{x}_h\rangle > 0$

$$\langle \mathbf{A}_h\mathbf{x}_h - \lambda_{\min}\mathbf{M}_h\mathbf{x}_h, \mathbf{y}_h\rangle = 0 \quad \forall \mathbf{y}_h \in \mathbb{R}^N.$$

Hieraus folgt:

$$\mathbf{M}_h^{-1}\mathbf{A}_h\mathbf{x}_h = \lambda_{\min}\mathbf{x}_h.$$

(*iii*) Wir schätzen nun ab:

$$\lambda_{\min}(\mathbf{M}_h^{-1}\mathbf{A}_h) = \min_{x\in\mathbb{R}^N} \frac{\langle \mathbf{A}_h\mathbf{x}_h, \mathbf{x}_h\rangle}{\langle \mathbf{M}_h\mathbf{x}_h, \mathbf{x}_h\rangle} = \min_{v_h\in V_h} \frac{(\nabla v_h, \nabla v_h)}{\|v_h\|^2} \geqslant \inf_{v\in H_0^1(\Omega)} \frac{(\nabla v, \nabla v)}{\|v\|^2} = \lambda_{\min}(-\Delta)$$

$$\lambda_{\max}(\mathbf{M}_h^{-1}\mathbf{A}_h) = \max_{x\in\mathbb{R}^N} \frac{\langle \mathbf{A}_h\mathbf{x}_h, \mathbf{x}_h\rangle}{\langle \mathbf{M}_h\mathbf{x}_h, \mathbf{x}_h\rangle} \leqslant \max_{x\in\mathbb{R}^N} \frac{\langle \mathbf{A}_h\mathbf{x}_h, \mathbf{x}_h\rangle}{\|\mathbf{x}_h\|^2} \max_{x\in\mathbb{R}^N} \frac{\|\mathbf{x}_h\|^2}{\langle \mathbf{M}_h\mathbf{x}_h, \mathbf{x}_h\rangle} = \lambda_{\max}(\mathbf{A}_h)\lambda_{\max}(\mathbf{M}_h^{-1}).$$

Also erhalten wir

$$\lambda_{\min}(\mathbf{M}_h^{-1}\mathbf{A}_h) = O(1), \quad \lambda_{\max}(\mathbf{M}_h^{-1}\mathbf{A}_h) = O(h^{-2}).$$

□

Das System von gewöhnlichen Differentialgleichungen (5.11) ist also mit $h \to 0$ beliebig steif. Wir benötigen zur Diskretisierung Zeitschrittverfahren, welche über eine möglichst hohe Stabilität verfügen. Zur Untersuchung der Stabilität eines Verfahrens führen wir den sogenannten *Verstärkungsfaktor* $R(z)$ der Einschrittmethode für ein $z = \lambda k \in \mathbb{C}$ ein, welcher, angewendet auf das skalare Modellproblem (5.12) die diskrete Lösung liefert:

$$\mathbf{u}^{m+1} = R(-\lambda k)\mathbf{u}^m \quad m \geqslant 1.$$

Remark 5.4 (Verstärkungsfaktor)**.** *Die Wahl des Vorzeichens* $-\lambda k$ *geschieht hier nur aus Konventionsgründen. Bei der Stabilitätsuntersuchung der gewöhnlichen Differentialgleichungen wird üblicherweise die Gleichung* $u'(t) = \lambda u(t)$ *mit* $Re(\lambda) \leqslant 0$ *anstelle von* $u'(t) + \lambda u(t) = 0$ *mit* $Re(\lambda) \geqslant 0$ *betrachtet. Durch die Definition* $\mathbf{u}^{m+1} = R(-\lambda k)\mathbf{u}^m$ *stimmen die Verstärkungsfaktoren wieder überein.*

Für die gebräuchlichen und bereits angesprochenen Verfahren gilt:

$$\begin{aligned} \text{Explizites Euler: } R(z) &= 1 + z \\ \text{Implizites Euler: } R(z) &= (1 - z)^{-1} \\ \text{Crank-Nicolson: } R(z) &= \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z} \\ \text{Theta-Verfahren: } R(z) &= \frac{1 + (1-\theta)z}{1 - \theta z} \end{aligned}$$

Gilt nun für ein $z = -\lambda k$ dass $|R(z)| \leqslant 1$, so ist die Einschrittmethode absolut stabil. Wir definieren:

Definition 5.5 (Stabilitätsgebiet)**.** *Wir nennen die Teilmenge der komplexen Zahlenebene*

$$SG = \{z = -\lambda k \in \mathbb{C},\ |R(z)| \leqslant 1\} \subset \mathbb{C}$$

das Stabilitätsgebiet *einer Einschrittmethode.*

Angewendet auf die Wärmeleitungsgleichung muss das Stabilitätsgebiet also einen möglichst großen Anteil der reellen Achse enthalten, da

$$-k\lambda_{\min}(\mathbf{M}_h^{-1}\mathbf{A}_h) = O(-k), \quad -k\lambda_{\max}(\mathbf{M}_h^{-1}\mathbf{A}_h) = O(-kh^{-2}).$$

Für das explizite Euler-Verfahren gilt:

$$SG_{EE} \cap \mathbb{R} = [-2, 0],$$

und damit muss:

$$-2 \leqslant -kh^{-2} \quad \Leftrightarrow \quad k \leqslant 2h^2.$$

Das explizite Euler-Verfahren ist somit wenig geeignet. Durch Rundungsfehlereinfluss können die Einträge der Systemmatrix fehlerbehaftet sein. Auch kleine Fehler in der Matrix können durch ein Abweichen von der Symmetrie dazu führen, dass Eigenwerte komplexwertig sind. Um auch in diesem Fall Stabilität zu erhalten reicht es nicht, dass nur die negative reelle Achse im Stabilitätsgebiet enthalten ist, wir benötigen auch einen imaginären Anteil. Wir definieren weiter

Definition 5.6 (Stabilitätsbegriffe)**.** *Im Fall* $\{z \in \mathbb{C},\ Re(z) \leqslant 0\} \subset SG$ *heißt die Methode* A-stabil*.*

Die Methode heißt streng A-Stabil *falls darüber hinaus mit einem* $c > 0$ *gilt:*

$$|R(z)| \leqslant 1 - ck \quad Re(z) \to -\infty.$$

Gilt unabhängig von der Schrittweite

$$|R(z)| \leqslant \kappa < 1 \quad Re(z) \to -\infty,$$

so heißt die Methode stark A-stabil*.*

Aus der A-Stabilität einer Methode folgt, dass die Iteration für beliebige Zeitschrittweiten k stabil bleibt

$$\sup_{m \geqslant 1} |\mathbf{u}_h^m| < \infty,$$

falls ein homogenes System mit rechter Seite $f = 0$ betrachtet wird. Die A-Stabilität reicht also gerade um zu garantieren, dass der Einfluss des Startwerts im Laufe der Zeit beschränkt

bleibt. Liegt strenge A-Stabilität vor, so bleibt die Lösung auch bei inhomogener rechter Seite beschränkt:

$$\sup_{m\geqslant 1} |\mathbf{u}_h^m| < c \sup_{m\geqslant 0} |\bar{f}^m|$$

Im Fall von starker A-Stabilität werden alle Fehleranteile exponentiell gedämpft. Stark A-stabile Verfahren sind besonders robust gegenüber Störungen (z.B. Rundungsfehler). Das implizite Euler-Verfahren ist stark A-stabil, es gilt:

$$|1-z|^{-1} = \sqrt{(1-\mathrm{Re}(z))^2 + \mathrm{Im}(z)^2|^{-1}} \leqslant \sqrt{\frac{1}{2}} \quad \mathrm{Re}(z) \leqslant -1.$$

Dieser Dämpfungsfaktor ist allerdings sehr gering und das implizite Euler-Verfahren neigt zur *Überdämpfung*. In der Anwendung ist es mit diesem Verfahren sehr schwer möglich natürliche Schwingungsvorgänge (die gewünscht sind) wieder zu geben. Wir definieren

Definition 5.7 (Dissipation)**.** *Ein Verfahren mit Verstärkungsfaktor* $R(z)$ *heißt* wenig dissipativ*, falls*

$$R(\pm i) \approx 1.$$

Remark 5.8 (Dissipation)**.** *Wir betrachten das Modellproblem*

$$u'(t) + iu(t) = 0, \quad u(0) = 1,$$

mit der Lösung

$$u(t) = \exp(-it) = \cos(t) + i\sin(t).$$

Für diese Lösung gilt $|u(t)| = 1$ *für alle* $t \geqslant 0$*. Die Lösung ist also eine Kosinus-, bzw. Sinus-Schwingung. Das implizite Euler-Verfahren liefert jedoch:*

$$\mathbf{u}^m = R(-ik)\mathbf{u}^{m-1} = (1+ik)^{-1}\mathbf{u}^{m-1}.$$

Mit

$$|(1+ik)^{-1}| = \sqrt{\frac{1}{1+k^2}}$$

folgt die Abschätzung:

$$|\mathbf{u}^m| \leqslant \left|\frac{1}{1+k^2}\right|^{\frac{m}{2}} \to 0 \quad (m\to\infty).$$

Das implizite Euler-Verfahren kann also für dieses Problem nicht sinnvoll eingesetzt werden. Betrachten wir die Diskretisierung mit dem Crank-Nicolson-Verfahren, so gilt

$$\mathbf{u}^m = \frac{1+\frac{ik}{2}}{1-\frac{ik}{2}}\mathbf{u}^{m-1},$$

mit

$$\left|\frac{1+\frac{ik}{2}}{1-\frac{ik}{2}}\right| = 1.$$

Das Crank-Nicolson-Verfahren erhält die Energie *des Systems also optimal.*

Für das Theta-Verfahren beweisen wir nun

Lemma 5.9 (Stabilität des Theta-Verfahrens)**.** *Für das Theta-Verfahren mit* $\theta \in [0,1]$ *gilt:*

1. *Das Theta-Verfahren ist genau dann A-stabil, falls* $\theta \geqslant \frac{1}{2}$.
2. *Für jedes* $\theta \geqslant \frac{1}{2} + ck$ *ist das Theta-Verfahren streng A-stabil.*
3. *Für jedes* $\theta > \frac{1}{2}$ *unabhängig von der Schrittweite* k *ist das Theta-Verfahren stark A-stabil.*
4. *Für die Dissipativität des Theta-Verfahrens gilt:*

$$|R(i)| \approx 1, \quad \textit{falls } \theta \approx \frac{1}{2}$$

5. *Für* $\theta \in [0, \frac{1}{2})$ *ist das Stabilitätsgebiet durch eine Ellipse in der linken komplexen Halbebene gegeben. Für das Stabilitätsintervall gilt:*

$$SG \cap \mathbb{R} = [-2(1-2\theta)^{-1}, 0] \quad 0 \leqslant \theta < \frac{1}{2}.$$

Proof: *1*) Es gilt:

$$|R(z)| = \left|\frac{1+(1-\theta)z}{1-\theta z}\right| \leqslant 1 \quad \Leftrightarrow \quad |1+(1-\theta)z| \leqslant |1-\theta z|$$

also genau dann, wenn gilt:

$$1+(1-\theta)^2|z|^2+2(1-\theta)Re(z) \leqslant 1+\theta^2|z|^2-2Re(z)\theta \quad \Leftrightarrow \quad (1-2\theta)|z|^2 \leqslant -2Re(z). \tag{5.14}$$

Wir betrachten $z \in \mathbb{C}$ mit $Re(z) \leqslant 0$ also $-Re(z) \geqslant 0$. Die Ungleichung ist dann genau für alle $\theta \geqslant \frac{1}{2}$ erfüllt (mit $Im(z)$ beliebig).

2) Wir betrachten den Grenzwert

$$\lim_{Re(\lambda)\to\infty} |R(-\lambda k)| = \lim_{Re(\lambda)\to\infty} \left|\frac{1+(1-\theta)\lambda k}{1-\theta\lambda k}\right| = \left|\frac{(1-\theta)k}{\theta k}\right| = \frac{1-\theta}{\theta} \tag{5.15}$$

Es gilt in erster Ordnung

$$\frac{1-\theta}{\theta} = 1 - 4\left(\theta - \frac{1}{2}\right) + O\left(\left|\theta - \frac{1}{2}\right|^2\right),$$

und strenge A-Stabilität folgt, wenn $\theta - \frac{1}{2} \geqslant ck$.

3) Aus (5.15) folgt sofort

$$\lim_{Re(\lambda)\to\infty} |R(-\lambda k)| = \frac{1-\theta}{\theta} \leqslant \kappa < 1 \quad \Leftrightarrow \quad \theta \geqslant \frac{1}{1+\kappa}.$$

4) Es gilt:

$$|R(\pm i)|^2 = \frac{\theta^2 - 2\theta + 2}{1 + \theta^2} = 1 - \frac{8}{5}\left(\theta - \frac{1}{2}\right) + O\left(\left|\theta - \frac{1}{2}\right|^2\right).$$

5) Nach (5.14) gilt absolute Stabilität für alle $z = x + iy$ mit

$$(1 - 2\theta)(x^2 + y^2) + 2x \leqslant 0.$$

Diese Gleichung beschreibt für $1 - 2\theta > 0$ also $\theta < \frac{1}{2}$ eine Ellipse. Auf der reellen Achse gilt (beachte $x \leqslant 0$)

$$(1 - 2\theta)x^2 + 2x \leqslant 0 \quad \Leftrightarrow \quad (1 - 2\theta)x \geqslant -2 \quad \Leftrightarrow \quad -\frac{2}{1 - 2\theta} \leqslant x \leqslant 0.$$

□

Remark 5.10 (Crank-Nicolson-Verfahren)**.** *Das Crank-Nicolson Verfahren scheint auf den ersten Blick ein idealer Kandidat zur Diskretisierung der Wärmeleitungsgleichung zu sein. Es ist A-stabil, von zweiter Ordnung genau und erhält die Energie optimal. Aufgrund fehlender stärkerer Regularität werden Störungen in den Anfangsdaten nur sehr langsam ausgedämpft. Für die homogene Wärmeleitungsgleichung gilt laut Satz 2.41*

$$\|u(t)\| \leqslant e^{-\lambda t}\|u^0\|,$$

für alle Startwerte $u^0 \in L^2(\Omega)$ mit dem kleinsten Eigenwert λ des Laplace-Operators. Hochfrequente Anteile im Anfangswert werden vom Crank-Nicolson Verfahren nur unzureichend ausgedämpft. Auch Rundungsfehler, welche in jedem Zeitschritt entstehen können, stören den Lösungsverlauf.

Um bessere Stabilität zu erreichen verwendet man oft:

Lemma 5.11 (Das implizit geshiftete Crank-Nicolson-Verfahren)**.** *Für $\theta = \frac{1}{2} + ck$ mit einer Konstante $c > 0$ hat das Theta-Verfahren quadratische Konsistenzordnung und ist streng A-stabil.*

Proof: Die strenge A-Stabilität wurde bereits in Satz 5.9 gezeigt. Wir schätzen für eine Anfangswertaufgabe $u'(t) = f(t, u(t))$ den Abschneidefehler ab:

$$\begin{aligned}\tau_k &= \frac{u(t_m) - u(t_{m-1})}{k} + \left(\frac{1}{2} + ck\right) f(t_m, u(t_m)) + \left(\frac{1}{2} - ck\right) f(t_{m-1}, u(t_{m-1})) \\ &= u'(t_{m-\frac{1}{2}}) + O(k^2) + f(t_{m-\frac{1}{2}}, u(t_{m-\frac{1}{2}})) + O(k^2) + ck(f(t_m, u(t_m)) - f(t_{m-1}, u(t_{m-1}))) \\ &= O(k^2) + O(k^2)|\nabla f|.\end{aligned}$$

□

Dieses Verfahren verfügt in der Regel über genügend Stabilität um hochfrequente Anteile in der Startlösung hinreichend schnell zu dämpfen. Darüber hinaus ist es von zweiter Ordnung (also balanciert mit einer Ortsdiskretisierung mit linearen Finiten Elementen) und jeder Schritt des Verfahrens ist sehr einfach durchzuführen, der Aufwand entspricht dem Crank-Nicolson Verfahren.

5.1.3 Verfahren höherer Ordnung

Padé-Approximationen Die Verstärkungsfaktoren für die Einschrittmethoden sind alles rationale Funktionen. Bei Betrachtung des Modellproblems ergibt sich der Zusammenhang:

$$u(t_m) = \exp(-\lambda t_m), \quad u_k^m = R(-\lambda k)^m, \quad R(z) \approx \exp(z).$$

Die Verstärkungsfaktoren sind somit rationale Approximationen der Exponentialfunktion. Dieser Zusammenhang kann nun als Konstruktionsprinzip für Einschrittmethoden verwendet werden. Wir suchen zwei Polynome $p \in P^r$ und $q \in P^s$, so dass

$$\mathrm{Re}(z) \leqslant 0: \quad \frac{p(z)}{q(z)} - e^z \to 0, \quad e^z q(z) - p(z) \to 0 \quad (|z| \to 0).$$

Die allgemeine Theorie der sogenannten *Padé-Approximationen* besagt, dass es zu jedem Polynomgrad $r \geqslant 0$ und $s \geqslant 0$ eine eindeutig bestimmte beste Approximation gibt, so dass

$$|e^z q(z) - p(z)| = O(|z|^{r+s+1}), \quad \mathrm{Re}(z) \leqslant 0.$$

Wir fassen diese Approximationen sortiert nach Polynomgrad von Zähler und Nenner im *Padé-Schemata* zusammen:

$$\begin{vmatrix} 1 & \frac{1+z}{1} & \frac{1+z+\frac{1}{2}z^2}{1} & \frac{1+z+\frac{1}{2}z^2+\frac{1}{6}z^3}{1} & \cdots \\ \frac{1}{1-z} & \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z} & \frac{1+\frac{2}{3}z+\frac{1}{6}z}{1-\frac{1}{3}z} & \frac{1+\frac{3}{4}z+\frac{1}{4}z^2+\frac{1}{24}z^3}{1-\frac{1}{4}z} & \\ \frac{1}{1-z+\frac{1}{2}z} & \frac{1+\frac{1}{3}z}{1-\frac{2}{3}z+\frac{1}{6}z^2} & \frac{1+\frac{1}{2}z+\frac{1}{12}z^2}{1-\frac{1}{2}z+\frac{1}{12}z^2} & \ddots & \\ \vdots & & \ddots & \frac{1+\frac{1}{2}z+\frac{1}{10}z^2+\frac{1}{120}z^3}{1-\frac{1}{2}z+\frac{1}{10}z^2-\frac{1}{120}z^3} & \ddots \\ \vdots & & & \ddots & \ddots \end{vmatrix}$$

Die bisherigen Verfahren sind (bis auf das allgemeine Theta-Verfahren) lassen sich in das Padé-Schema einordnen. Sie sind also alle Ordnungsoptimal. Das Konstruktionsprinzip lässt sich umgekehrt auch einfach nutzen, um die Konsistenzordnung eines Verfahrens zu prüfen. Für das Theta-Verfahren etwa gilt:

$$R(z) = \frac{1 + (1-\theta)z}{1 - \theta z} = 1 + z + tz^2 + O(|z|^3).$$

Wir vergleichen mit der Reihenentwicklung der Exponentialfunktion

$$\exp(z) = 1 + z + \frac{1}{2}z^2 + O(|z|^3).$$

Für $\theta = \frac{1}{2}$ erhalten wir eine Approximation der Exponentialfunktion dritter Ordnung. Und für $\theta = \frac{1}{2} + ck$ folgt mit $z = -\lambda k$ ebenso:

$$R(z) - \exp(z) = ckO(k^2) + O(k^3) = O(k^3).$$

Jetzt betrachten wir die Anfangswertaufgabe

$$u_h'(t) + \mathbf{A}_h u_h(t) = f_h(t), \quad u_h(0) = u_h^0 \in \mathbb{R}^N,$$

mit einer symmetrisch positiven Matrix $\mathbf{A}_h \in \mathbb{R}^{N\times N}$ und die zugehörigen Komponenten

$$\omega_i'(t) + \lambda_i \omega_i(t) = f_i(t), \quad \omega_i(0) = \omega_i^0, \quad i = 1, \ldots, N,$$

mit den Eigenwerten $\lambda_i > 0$ von $\mathbf{A}_h$. Jede dieser Gleichungen wird nun mit einem Padé-Schema approximiert:

$$\omega_i^m = \frac{p(-\lambda_i k)}{q(-\lambda_i k)}\omega_i^{m-1} \quad \Leftrightarrow \quad q(-\lambda_i k)\omega_i^m = p(-\lambda_i k)\omega_i^{m-1}, \quad m \geqslant 1.$$

Mit der Diagonalmatrix $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$ gilt kurz:

$$q(-kD)\omega_k^m = p(-kD)\omega_k^{m-1}. \tag{5.16}$$

Die symmetrisch positiv definite Matrix $\mathbf{A}_h$ ist mit einer unitären Matrix $\mathbf{Q}_h$ diagonalisierbar. Es gilt für jedes Polynom:

$$p(A) = p(Q^T D Q) = Q^T p(D) Q.$$

Eingesetzt in (5.16) kann das Padé-Schema unmittelbar für das lineare System $u_k' + \mathbf{A}_h u_k = 0$ formuliert werden:

$$q(-k\mathbf{A}_h)\mathbf{u}_k^m = p(-k\mathbf{A}_h)\mathbf{u}_k^{m-1}.$$

In jedem Schritt der Padé-Approximation muss also ein Gleichungssystem mit Matrix $q(-k\mathbf{A}_h)$ mit N_h Unbekannten gelöst werden.

Es gilt (ohne Beweis):

Lemma 5.12 (Stabilität der Padé-Approximationen)**.** *Alle diagonalen Padé-Approximationen* $(r = s)$ *sind A-stabil. Alle subdiagonalen Padé-Approximationen* $(s = r + 1)$ *sind stark A-stabil.*

Zur Approximation von parabolischen Gleichungen kommen also nur diagonale oder subdiagonale Padé Verfahren in Frage. Verfahren oberhalb der Diagonale haben zu geringe Stabilitätseigenschaften, Verfahren mit $\mathtt{Grad}(q) \gg \mathtt{Grad}(p)$ sind zwar stabil, erfordern jedoch zur Lösung in jedem Schritt das Lösen eines komplexen Gleichungssystems. Für das subdiagonale $s = 2$ und $r = 1$ Verfahren gilt in jedem Schritt:

$$[\mathbf{I}_h - \frac{2}{3}k\mathbf{A}_h + \frac{1}{6}k^2\mathbf{A}_h^2]\,\mathbf{u}_k^m = \mathbf{u}_k^{m-1} + \frac{1}{3}k\mathbf{A}_h\mathbf{u}_k^{m-1} \quad m \geqslant 1.$$

Die Matrix auf der linken Seite des Gleichungssystems ist sehr viel dichter besetzt als die Matrix $\mathbf{A}_h$ selbst. Darüber hinaus gilt für die Kondition einer Matrix:

$$\operatorname{cond}_2(\mathbf{A}_h^2) = \operatorname{cond}_2(\mathbf{A}_h)^2,$$

d.h. im Fall der Steifigkeitsmatrix erhalten wir mit $\mathrm{cond}_2(\mathbf{A}_h^2) = O(h^{-4})$ eine äußerst schlecht konditionierte Matrix.

Die Padé-Verfahren wären einfach durchzuführen, wenn der rationale Anteil $q(z)$ in Linearfaktoren zerfallen würde:

$$q(z) = (1 - \theta_1 z)(1 - \theta_2 z) \cdots (1 - \theta_s z). \tag{5.17}$$

Angenommen alle $\theta_i \in \mathbb{R}$ wären reell, dann könnte ein Schritt des Padé-Verfahrens durch s einfache Schritte eines Theta-Verfahrens ersetzt werden:

$$\begin{aligned} [I + k\theta_1 \mathbf{A}_h]\mathbf{u}_1^m &= p(-k\mathbf{A}_h)\mathbf{u}_k^{m-1} \\ [I + k\theta_2 \mathbf{A}_h]\mathbf{u}_2^m &= \mathbf{u}_1^m \\ &\vdots \\ [I + k\theta_s \mathbf{A}_h]\mathbf{u}^m &= \mathbf{u}_{s-1}^m. \end{aligned} \tag{5.18}$$

Es gilt allerdings der Satz:

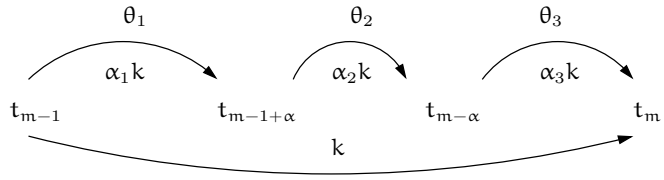
Lemma 5.13. *Der rationale Anteil eine diagonalen bzw. subdiagonalen Padé-Approximation hat höchstens eine reelle Nullstelle.*

D.h., die Nullstellen von $q(z) = 0$ sind (bis auf eine) alle komplex. Prinzipiell könnte das System (5.18) auch mit komplexer Arithmetik, also mit $k\theta_i \notin \mathbb{R}$ gelöst werden. Der numerische Aufwand hingegen ist deutlich größer.

Remark 5.14. *Die Bedeutung der Padé-Approximationen liegt zunächst in einer einfachen einheitlichen Theorie bzgl. Konsistenzordnung und Stabilität. Das Problem bei der Durchführung ist die zu "teure" Invertierung der Matrix $q(-k\mathbf{A}_h)$. Es stellt sich allerdings heraus, dass die Padé-Approximationen angewendet auf das lineare System $u_h'(t)+\mathbf{A}_h u_h(t) = f_h(t)$ speziellen impliziten Runge-Kutta Verfahren entsprechen.*

Eine weitere Analogie werden wir bei der Diskussion von Galerkin-Verfahren zur Zeitdiskretisierung feststellen: die diagonalen Padé-Verfahren entsprechen (bis auf numerischen Quadraturfehlern) einer Galerkin-Diskretisierung in der Zeit mit stetigen Ansatzfunktionen. Die subdiagonalen Padé-Verfahren lassen sich durch eine Galerkin-Diskretisierung mit unstetigen Ansatzfunktionen herleiten.

Das Teilschritt-Theta-Verfahren Auch wenn die Padé-Verfahren prinzipiell viele gewünschte Eigenschaften wie hohe Ordnung und gute Stabilität vereinen, sind sie wenig zur praktischen Verwendung geeignet, da in jedem Schritt Gleichungssysteme von großer Dimension gelöst werden müssen. Wir verfolgen nun zum Abschluss den umgekehrten Ansatz und konstruieren ein Gesamtverfahren, dessen Verstärkungsfaktor aus Linearfaktoren zusammengesetzt ist. Hierzu führen wir drei Teilschritte mit dem Theta-Verfahren hintereinander aus:



Dabei wählen wir für den ersten Teilschritt die Schrittweite $\alpha_1 k$ und für daneben $\theta = \theta_1$. Schritt zwei wird entsprechend mit $\alpha_2 k$ und θ_2, Schritt drei mit $\alpha_3 k$ und θ_3 ausgeführt. Dabei muss für die Parameter gelten:

$$0 \leqslant \theta_1, \theta_2, \theta_3, \alpha_1, \alpha_2 \alpha_3 \leqslant 1, \quad \alpha_1 + \alpha_2 + \alpha_3 = 1.$$

Das Gesamtverfahren hat den Verstärkungsfaktor:

$$R(z) = \left(\frac{1 + (1-\theta_1)\alpha_1 z}{1 - \theta_1 \alpha_1 z}\right)\left(\frac{1 + (1-\theta_2)\alpha_2 z}{1 - \theta_2 \alpha_2 z}\right)\left(\frac{1 + (1-\theta_3)\alpha_3 z}{1 - \theta_3 \alpha_3 z}\right)$$

Bei symmetrische Verfahren wird üblicherweise eine höhere Ordnung erreicht, da sich Terme in der Restgliedentwicklung gegenseitig aufheben. Wir reduzieren daher die freien Parameter machen einen einfachen Ansatz mit nur 2 Parametern:

$$\theta_1 = \theta_3 = \theta, \quad \alpha_1 = \alpha_3 = 1, \quad \theta_2 = 1 - \theta, \quad \alpha_2 = 1 - 2\alpha.$$

Für den Parameter α muss nun $\alpha \in (0, \frac{1}{2})$ gelten, ansonsten wäre $\alpha_2 < 0$. Theta ist weiterhin beliebig in $\theta \in [0, 1]$ zu wählen. Nun gilt:

$$R(z) = \underbrace{\left(\frac{1 + (1-\theta)\alpha z}{1 - \theta\alpha z}\right)}_{R_1(z)}\underbrace{\left(\frac{1 + \theta(1-2\alpha) z}{1 - (1-\theta)(1-2\alpha) z}\right)}_{R_2(z)}\underbrace{\left(\frac{1 + (1-\theta)\alpha z}{1 - \theta\alpha z}\right)}_{R_1(z)}$$

Wir wollen diese beiden Parameter nun so bestimmen, dass das Gesamtverfahren im Zeitpunkt t_m (also nach den drei Teilschritten) eine möglichst hohe Ordnung hat, stark A-stabil ist und von geringer Dissipativität.

Für den Grenzwert $Re(z) \to -\infty$ gilt:

$$|R(z)| \xrightarrow[Re(z)\to-\infty]{} \frac{\theta(1-\theta)^2\alpha^2(1-2\alpha)}{\theta^2\alpha^2(1-\theta)(1-2\alpha)} = \frac{1-\theta}{\theta}$$

Der Grenzwert ist kleiner 1 für alle $\theta > \frac{1}{2}$. Dann liegt - unabhängig von α - starke A-Stabilität vor. Zur Bestimmung der Fehlerordnung entwickeln wir $R(z)$ um $z = 0$ und vergleichen mit der Entwicklung der Exponentialfunktion $\exp(z)$

$$R(z) - \exp(z) = \left(\alpha^2 - 2\alpha + \frac{1}{2}\right)(1 - 2\theta)z^2 + O(|z|^3),$$

und erhalten für

$$\alpha = 1 - \sqrt{\frac{1}{2}} \approx 0.293\ldots$$

Konvergenz zweiter Ordnung. Der Parameter θ darf hierbei im Intervall $\theta \in (\frac{1}{2}, 1]$ frei gewählt werden. In jedem der drei Teilschritte ist im Fall $\theta \neq 1$ ein implizites Verfahren zu lösen. Mit der speziellen Wahl:

$$\chi := \theta\alpha = (1-\theta)(1-2\alpha) \quad \Leftrightarrow \quad \theta = \frac{1-2\alpha}{1-\alpha},$$

wird erreicht, dass in jedem Teilschritt der rationale Anteil im Verstärkungsfaktor identisch ist und somit die gleiche Matrix zu invertieren ist:

$$R(z) = \left(\frac{1+(1-\theta)\alpha z}{1-\chi z}\right)\left(\frac{1+\theta(1-2\alpha)z}{1-\chi z}\right)\left(\frac{1+(1-\theta)\alpha z}{1-\chi z}\right)$$

Dieses Verfahren wird das *Teilschritt-Theta-Verfahren* genannt. Schließlich bestimmen wir noch die Dissipativität. Es gilt:

$$|R(\pm i)| \approx 0.9997\ldots$$

Das Teilschritt-Theta-Verfahren stellt ein nahezu optimales Verfahren dar: es ist von zweiter Ordnung in der Zeit, somit gut balanciert mit linearen Finiten Elementen. Es ist stark A-stabil, hat aber im Gegensatz zum impliziten Euler-Verfahren eine äußerst geringe Dissipativität, so dass es in der Lage ist, die Energie des Systems sehr gut zu erhalten. Insbesondere bei der Diskretisierung von Strömungsproblemen, wo sowohl Glättungseigenschaften als auch Erhaltungseigenschaften eine wesentliche Rolle spielen ist das Teilschritt-Theta-Verfahren von großer Bedeutung.

5.2 Zeitdiskretisierung mit Galerkin-Verfahren

Abschließend betrachten wir wie angekündigt Galerkin-Verfahren zur Diskretisierung der parabolischen Gleichung in Zeit und Ort. Wir suchen wieder $u : I \times \Omega \to \mathbb{R}$ als Lösung von

$$\partial_t u - \Delta u = f \quad \text{in } I \times \Omega, \quad u(0) = u^0. \tag{5.19}$$

Wir wissen aus Satz 2.38, dass es für jede rechte Seite $f \in L^2(I; L^2(\Omega))$ und für jeden Startwert $u^0 \in L^2(\Omega)$ eine schwache Lösung der Wärmeleitungsgleichung (5.19) im Raum:

$$u \in W(I), \quad W(I) := \{v \in L^2(I; H_0^1(\Omega)),\ \partial_t v \in L^2(I; H^{-1}(\Omega))\}.$$

Darüber hinaus gilt dann wegen Satz 2.37 für diese Lösung auch die Regularität

$$u \in C(\bar{I}; L^2(\Omega)),$$

das heißt, die Lösung ist (bis zum Rand von I) stetig in der Zeit.

Zur Orts-Zeit Diskretisierung mit Galerkin-Verfahren leiten wir zunächst eine variationelle Formulierung her, welche sowohl in Ort als auch Zeit variationell ist. Dazu multiplizieren (5.19) mit einer Testfunktion $\phi \in L^2(I; H^1(\Omega))$ und integrieren über Ort und Zeit. Es gilt:

Lemma 5.15 (Variationelle Formulierung 1)**.** *Die schwache Lösung* $u \in W(I)$ *der Wärmeleitungsgleichung (5.19) ist charakterisiert durch das Variationsproblem:*

$$\begin{aligned} \int_I \Big\{ (\partial_t u, \phi) + (\nabla u, \nabla \phi) \Big\}\, dt = \int_I (f, \phi)\, dt \qquad & \forall \phi \in L^2(I; H^1(\Omega)) \quad t > 0 \\ (u(0), \phi) = (u^0, \phi) \qquad & \forall \psi \in H_0^1(\Omega) \qquad\quad t = 0. \end{aligned} \tag{5.20}$$

Proof: Die Hinrichtung folgt unmittelbar durch Multiplikation mit einer Testfunktion $\phi \in L^2(I; H_0^1(\Omega))$ und Integration über I und Ω. Die Rückrichtung folgt durch partielle Integration:

$$\int_I (\partial_t u - \Delta u - f, \phi)\, dt = 0 \quad \forall \phi \in L^2(I; H^1(\Omega)).$$

Zu $t_0 \in I$ wählen wir eine Dirac-Folge $\delta_{t_0,\varepsilon} \in C^\infty(I)$ mit der Eigenschaft:

$$\int_I \nu(t)\delta_{t_0,\varepsilon}(t)\, dt \xrightarrow[\varepsilon\to 0]{} \nu(t_0) \quad \forall \nu \in C(I).$$

Für die Testfunktion $\phi_{t_0,\varepsilon}(x,t) := \delta_{t_0,\varepsilon}(t)\psi(x)$ mit $\psi \in H_0^1(\Omega)$ gilt weiter:

$$\int_I (\partial_t u - \Delta u - f, \phi_{t_0,\varepsilon})\, dt \xrightarrow[\varepsilon\to 0]{} (\partial_t u(t_0) - \Delta u(t_0) - f(t_0), \psi) = 0 \quad \forall \psi \in C_0^\infty(\Omega) \quad t_0 \in I.$$

Abschließend wird entsprechend im Ort vorgegangen. □

Zur Diskretisierung (zunächst in der Zeit) suchen wir nun endlich dimensionale Teilräume $X_k \subset W(I)$ und $V_k \subset L^2(I; H_0^1(\Omega))$. Dabei unterscheiden wir zwischen zwei alternativen Verfahren:

dG(r)-Verfahren Das *discontinuous Galerkin-Verfahren* verwendet zur Diskretisierung in der Zeit unstetige Ansatzräume W_k mit lokalem Polynomgrad r. Aufgrund von Satz 2.37 ist $u \in C(\bar{I}; L^2(\Omega))$ und somit handelt es sich wegen $W_k \not\subset W(I)$ um einen *nichtkonformen* Galerkin-Ansatz.

cG(r)-Verfahren Beim *continuous Galerkin-Verfahren* wird zur Diskretisierung ein stetiger, stückweise polynomialer Ansatzraum $W_k \subset W(I)$ gewählt. Hierbei handelt es sich um eine konforme Methode.

In beiden Fällen darf der V_k Testraum als unstetig gewählt werden, da der Raum $L^2(I; H^1(\Omega))$ keine Stetigkeitsanforderungen in der Zeit stellt.

5.2.1 Das dG(r)-Verfahren zur Zeitdiskretisierung der Wärmeleitungsgleichung

In diesem Abschnitt untersuchen wir die Zeitdiskretisierung der Wärmeleitungsgleichung mit dem dG(r)-Verfahren. Dazu definieren wir zunächst halboffene Teilintervalle $I_m := (t_{m-1}, t_m]$ als

$$0 = t_0 < t_1 < \cdots < t_M = T, \quad k_m := t_m - t_{m-1}, \quad I_m := (t_{m-1}, t_m].$$

Auf dieser Zerlegung führen wir einen nicht konformen (bezogen auf $W(I)$), stückweise definierten Raum ein:

$$W_I := \{v : I \times \Omega \to \mathbb{R},\ v\big|_{I_m} \in W(I_m)\}.$$

Dieser Raum ist nicht in $W(I)$ enthalten, da Unstetigkeiten an den Gitterpunkten t_m zugelassen sind. Jede Funktion $v \in W_I$ ist jedoch in jedem Intervall I_m stetig bis zu beiden Intervallenden definiert, da ja $v\big|_{I_m} \in W(I_m)$ auch $v_{I_m} \in C(\bar{I}_m; L^2(\Omega))$ impliziert. Für eine Funktion $v \in W_I$ definieren wir nun in einem Gitterpunkt:

$$v_m^+ := v(t_m)^+ := \lim_{s \downarrow 0} v(t_m + s), \quad v_m^- := \lim_{s \downarrow 0} v(t_m - s), \quad [v]^m := v_m^+ - v_m^-.$$

Der Sprung $[v]^m$ misst die Unstetigkeit der Funktion $v \in W_I$.

Wir definieren ein zweites Variationsproblem:

Lemma 5.16 (Variationelle Formulierung 2)**.** *Das Variationsproblem (5.20) ist äquivalent zum Variationsproblem: suche $u \in W_I$, so dass*

$$\int_I \Big\{(\partial_t u, \phi) + (\nabla u, \nabla \phi)\Big\}\, dt + \sum_{m=1}^{M} ([u]^{m-1}, \phi_{m-1}^+) = \int_I (f, \phi)\, dt \qquad \forall \phi \in L^2(I; H^1(\Omega)), \tag{5.21}$$

mit der Notation

$$u_0^- := u^0.$$

Proof: (*i*) Sei zunächst $u \in W(I)$ eine Lösung von (5.20). Es gilt $W(I) \subset W_I$. Weiter gilt für $u \in W(I)$ auch $u \in C(\bar{I}; L^2(\Omega))$, also in jedem Gitterpunkt t_m

$$u(t_m)^+ = u(t_m)^- \quad \text{im } L^2\text{-Sinne},$$

also

$$([u]^m, \psi) = 0 \quad \forall \psi \in L^2(\Omega).$$

Hieraus folgt, dass $u \in W(I) \subset W_I$ auch Lösung von (5.21) ist.

(*ii*) Nun sei $u \in W_I$. Wir müssen zeigen, dass $u \in W(I)$, dass die Funktion u also in den Gitterpunkten stetig sein muss. Dann löst u automatisch auch Gleichung (5.20). Hierzu wählen wir spezielle Testfunktionen $\delta_{m,\varepsilon} \in I_m = (t_{m-1}, m]$ mit der Eigenschaft:

$$\operatorname{supp}(\delta_{m,\varepsilon}) \subset I_m, \quad \delta_{m,\varepsilon}(t_m) = 1, \quad \delta_{m,\varepsilon}(t) \xrightarrow[\varepsilon \to 0]{} 0 \quad \text{für } t \neq t_m.$$

Mit $\phi_{m,\varepsilon}(x,t) := \delta_{m,\varepsilon}(t)\psi(x)$ und $\psi \in H_0^1(\Omega)$ folgt dann

$$\begin{aligned} 0 &= \int_I (f, \phi)\, dt - \int_I \Big\{(\partial_t u, \phi_{m,\varepsilon}) + (\nabla u, \nabla \phi_{m,\varepsilon})\Big\}\, dt - \sum_{m=1}^{M} ([u]^{m-1}, (\phi_{m,\varepsilon})_m^+) \\ &= \int_{I_m} (f, \phi)\, dt - \Big\{(\partial_t u, \phi_{m,\varepsilon}) + (\nabla u, \nabla \phi_{m,\varepsilon})\Big\}\, dt - ([u]^m, \psi) \xrightarrow[\varepsilon \to 0]{} ([u]^{m-1}, \psi) \quad m = 1, \dots, M \end{aligned}$$

die Stetigkeit von $u \in W_I$ in den Zeitpunkten t_m, also dass $u \in W(I)$. □

Die Grundlage des $dG(r)$-Verfahrens ist jetzt eine Galerkin-Diskretisierung von (5.21). Hierzu wählen als konforme Teilräume $W_k^{(r)}$ von W_I unstetige, stückweise polynomielle Räume vom Grad r:

$$W_k^{(r)} := \{v \in W_I,\ v\big|_{I_m} \in P^r(I_m; H_0^1(\Omega))\},$$

wobei $P^r(I_m; H_0^1(\Omega))$ der Raum der Polynome vom Grad r mit Werten in $H_0^1(\Omega)$ ist.

Wir formulieren:

Lemma 5.17 ($dG(r)$-Verfahren)**.** *Die Lösung des* $dG(r)$*-Verfahrens* $u_k \in W_k^{(r)}$ *der variationellen Formulierung*

$$\sum_{m=1}^{M} \int_{I_m} \Big\{(\partial_t u_k, \phi_k) + (\nabla u_k, \nabla \phi_k)\Big\}\, dt + \sum_{m=1}^{M} ([u_k]^{m-1}, \phi_{k,m-1}^+) = \int_I (f, \phi_k)\, dt \quad \forall \phi_k \in W_k^{(r)} \tag{5.22}$$

ist eindeutig bestimmt und es gilt die a priori Abschätzung

$$\|u_{k,M}^-\|^2 + \int_I \|\nabla u_k\|^2\, dt \leqslant \|u^0\|^2 + c_P^2 \int_I \|f\|^2\, dt.$$

Proof: (*i*) *Eindeutigkeit:* Angenommen es existieren zwei Lösungen $u_k^1, u_k^2 \in W_k^{(r)}$ von (5.22). Dann gilt für $w_k := u_k^1 - u_k^2$:

$$\int_I \Big\{(\partial_t w_k, \phi_k) + (\nabla w_k, \nabla \phi_k)\Big\}\, dt + \sum_{m=1}^{M} ([w_k]^{m-1}, \phi_{k,m-1}^+) = 0 \quad \forall \phi_k \in W_k^{(r)}.$$

Wir wählen nun auf dem ersten Intervall

$$\phi_k\big|_{I_1} = w_k, \quad \phi_k = 0 \text{ sonst}$$

und erhalten mit $w_{k,0}^- = {u_{k,0}^1}^- - {u_{k,0}^2}^- = u^0 - u^0 = 0$:

$$\int_{I_1} \Big\{(\partial_t w_k, w_k) + (\nabla w_k, \nabla w_k)\Big\}\, dt + (w_{k,0}^+, w_{k,0}^+) = 0,$$

was äquivalent ist zu

$$\int_{I_1} \Big\{\frac{1}{2}\partial_t \|w_k\|^2 + \|\nabla w_k\|^2\Big\}\, dt + \|w_{k,0}^+\|^2 = 0,$$

bzw mit dem Hauptsatz der Differential- und Integralrechnung zu

$$\|w_{k,1}^-\|^2 + \|w_{k,0}^+\|^2 + \int_{I_1} \|\nabla w_k\|^2\, dt = 0.$$

Hieraus folgt $\|\nabla w_k\| = 0$ auf I_1 und mit der Poincaré-Ungleichung auch $\|w_k\| = 0$ auf I_1. Diese Argumentation kann nun auf $I_2, I_3, \ldots, I_m$ fortgesetzt werden, da für den Startwert in jedem Intervall gilt $w^-_{k,m-1} = 0$.

(*ii*) *Existenz:* Der $H^1_0(\Omega)$ kann in eine L^2-orthogonale Summe aus Eigenräumen des Laplace-Operators aufgespalten werden:

$$H^1_0(\Omega) = H_1 \oplus H_2 \oplus \cdots .$$

Dabei gilt $\dim(H_m) < \infty$, d.h., alle Eigenräume sind endlich dimensional. Dies folgt aus dem Spektralsatz für kompakte normale Operatoren (angewendet auf den inversen Laplace-Operator). Wir schreiben die Lösung $u_k \in W_I$ auf einem Intervall I_m in der Form:

$$u_k(x,t)\Big|_{I_m} = \sum_{l \geqslant 1} \mu^l_m(t)\omega^l(x),$$

wobei $\omega^l \in H_l$ und $\mu^l_m(t)$ ein Polynom im P^r ist. Die Räume H_m sind L^2-orthogonal und invariant gegenüber dem Laplace-Operator. Das Problem (5.22) zerfällt somit in endlich dimensionale Probleme in den jeweiligen Teilräumen. Auf dem ersten Intervall gilt:

$$m \geqslant 0: \quad (\omega_l, \phi)\int_I (\partial_t \mu^l_0(t) + \lambda_l \mu^l_0(t))\,dt + (\omega_l - u^0, \phi)\mu^l_0(0) = \int_I (f,\phi)\,dt \quad \forall \phi \in H_m \tag{5.23}$$

Jedes dieser Problem ist endlich-dimensional, da $\dim(P^r) < \infty$ und da $\dim(H_m) < \infty$. Die Eindeutigkeit einer Lösung impliziert somit die Existenz.

(*iii*) *A priori Abschätzung:* Wir setzen in (5.22) $\phi_k = u_k$ und erhalten:

$$\int_I \Big\{\frac{1}{2}\partial_t \|u_k(t)\|^2 + \|\nabla u_k(t)\|^2\Big\}\,dt + \sum_{m=1}^{M} ([u_k]^{m-1}, u^+_{k,m-1}) = \int_I (f, u_k)\,dt. \tag{5.24}$$

Wir verarbeiten nun die Terme einzeln. Für die Zeitableitung erhalten wir mit dem Hauptsatz der Differential- und Integralrechnung auf jedem Intervall I_m:

$$\frac{1}{2}\int_I \partial_t \|u_k(t)\|^2\,dt = \frac{1}{2}\sum_{m=1}^{M}\int_{I_m} \partial_t \|u_k(t)\|^2\,dt = \frac{1}{2}\sum_{m=1}^{M}\Big\{\|u^-_{k,m}\|^2 - \|u^+_{k,m-1}\|^2\Big\} \tag{5.25}$$

Den Gradienten-Term $\|\nabla u_k(t)\|$ in (5.24) lassen wir unberührt, die Sprünge schreiben wir zunächst als:

$$([u_k]^{m-1}, u^+_{k,m-1}) = (u^+_{k,m-1} - u^-_{k,m-1}, u^+_{k,m-1}) = \|u^+_{k,m-1}\|^2 - (u^-_{k,m-1}, u^+_{k,m-1})$$

Dann schätzen wir mit Young'scher Ungleichung nach unten ab und erhalten:

$$([u_k]^{m-1}, u^+_{k,m-1}) \geqslant \|u^+_{k,m-1}\|^2 - \frac{1}{2}\|u^-_{k,m-1}\|^2 - \frac{1}{2}\|u^+_{k,m-1}\|^2 = \frac{1}{2}\|u^+_{k,m-1}\|^2 - \frac{1}{2}\|u^-_{k,m-1}\|^2 \tag{5.26}$$

Zuletzt gilt für die rechte Seite von (5.24) mit Cauchy-Schwarz, Poincaré und Young'scher Ungleichung:

$$\int_I (f, u_k)\,dt \leqslant \int_I \|f\|\,\|u_k\|\,dt \leqslant \int_I c_P\|f\|\,\|\nabla u_k\|\,dt \leqslant \int_I \frac{c_P^2}{2}\Big\{\|f\|^2 + \frac{1}{2}\|\nabla u_k\|^2\Big\}\,dt \quad (5.27)$$

Wir fassen nun (5.24) mit (5.25), (5.26) und (5.27) zusammen und erhalten die Abschätzung:

$$\sum_{m=1}^{M}\Big\{\|u_{k,m}^-\|^2 \underbrace{-\|u_{k,m-1}^+\|^2 + \|u_{k,m-1}^+\|^2}_{=0} - \|u_{k,m-1}^-\|^2\Big\} + \int_{I_m}\|\nabla u_k\|^2\,dt \leqslant c_P^2\int_I \|f\|^2\,dt.$$

Zusammenfassen der Summe und $u_{k,0}^- = u^0$ liefert dann die gewünschte Abschätzung. □

Durchführung des dG(r)-Verfahrens

Die Lösung $u \in W_k^{(r)}$ ist unstetig über die Intervallgrenzen hinaus. Wir definieren auf jedem Intervall

$$u_k^m := u_k\big|_{I_m} \in P^r(I_m; H_0^1(\Omega)),$$

und weil die Testfunktionen auch unstetig sind koppelt u_k^m zu u_k^{m-1} nur durch den Sprungterm

$$([u_k]^m, \phi_{k,m-1}^+) = (u_{k,m-1}^{m,+} - u_{k,m-1}^{m-1,-}, \phi_{k,m-1}^+)$$

Wenn die Lösung u_k^{m-1} auf I_{m-1} bekannt ist, so ist insbesondere auch $u_{k,m-1}^{m-1,-}$ bekannt und u_k^m lässt sich berechnen als Lösung des lokalen Variationsproblems

$$\begin{aligned} u_k^m \in P^r(I_m; H_0^1(\Omega)): \quad & \int_{I_m}\Big\{(\partial_t u_k^m, \phi_k^m) + (\nabla u_k^m, \nabla\phi_k^m)\Big\}\,dt + (u_{k,m-1}^{m,+}, \phi_{k,m-1}^{m,+}) \\ & = \int_{I_m}(f, \phi_k^m)\,dt + (u_{k,m-1}^{m-1,-}, \phi_{k,m-1}^{m,+}) \quad \forall \phi_k^m \in P^r(I_m; H_0^1(\Omega)). \end{aligned}$$

Bei dem dG(r)-Verfahren handelt es sich also um Zeitschritt-Verfahren. Im Gegensatz zu den einfachen Zeitschritt-Verfahren wie dem Theta-Verfahren müssen in jedem Zeitschritt $r+1$ Unbekannte bestimmt werden, da u_k^m ein Polynom im Raum $P^r(I_m; H_0^1(\Omega))$ ist.

Example 5.18 (Das dG(0)-Verfahren)**.** *Wir betrachten den einfachsten Fall des* dG(0)*-Verfahrens. Es sei also:*

$$X_k^{(0)} := \{v \in W_I,\ v\big|_{I_m} \in H_0^1(\Omega)\},$$

d.h., eine Funktion $v \in X_k^{(0)}$ ist zeitlich konstant auf jedem Intervall (siehe Abbildung (5.1) und wir schreiben:

$$v_k^m := v_k\big|_{I_m}$$

Da die Intervall $I_m := (t_{m-1}, t_m]$ links offen sind gilt

$$v_{k,m}^+ = v_k^{m+1}, \quad v_{k,m}^- = v_k^m, \quad [v_k]^m = v_k^{m+1} - v_k^m.$$

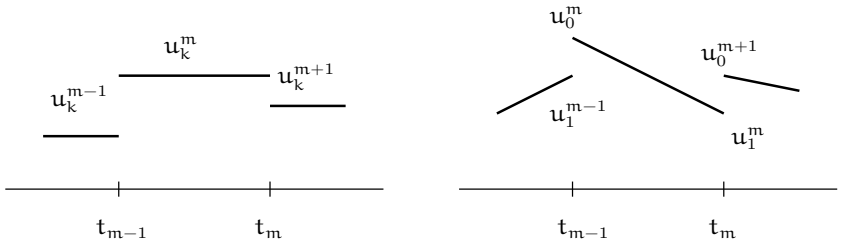


Abbildung 5.1: Schematische Darstellung der Ansatzräume im $dG(0)$- (links) bzw. $dG(1)$-Verfahren (rechts).

Weiter gilt $\partial_t v_k^m = 0$. Damit vereinfacht sich die Variationsgleichung auf jedem Intervall I_m zu:

$$u_k^m \in H_0^1(\Omega): \quad k_m(\nabla u_k^m, \nabla \phi_k^m) + (u_k^m, \phi_k^m) = \int_{I_m} (f, \phi_k^m)\, dt + (u_k^{m-1}, \phi_k^m) \quad \forall \phi_k^m \in H_0^1(\Omega).$$

Wenn wir das Zeitintegral über die rechte Seite f mit Hilfe der Box-Regel auswerten, so erhalten wir gerade das implizite Euler-Verfahren:

$$(u_k^m, \phi_k^m) + k_m(\nabla u_k^m, \nabla \phi_k^m) = (u_k^{m-1}, \phi_k^m) + k_m(f(t_m), \phi_k^m) \quad \forall \phi_k^m \in H_0^1(\Omega).$$

Example 5.19 (Das $dG(1)$-Verfahren)**.** *Auf jedem Intervall I_m ist $u_k \in W_k^{(1)}$ stückweise linear. Wir schreiben:*

$$u_k(x,t)\big|_{I_m} = \xi_0^m(t) u_0^m(x) + \xi_1^m(t) u_1^m(x),$$

mit zwei Funktionen $u_0^m, u_1^m \in H_0^1(\Omega)$ und Basisfunktionen des P^1 welche gegeben sind durch:

$$\xi_0^m(t) = \frac{t_m - t}{t_m - t_{m-1}}, \quad \xi_1^m(t) = \frac{t - t_{m-1}}{t_m - t_{m-1}}.$$

Dann gilt:

$$u_{k,m-1}^+ = u_0^m, \quad u_{k,m-1}^- = u_1^{m-1}, \quad \partial_t u_k\big|_{I_m} = \frac{u_1^m - u_0^m}{k_m}.$$

Das lokale Gleichungssystem zur Bestimmung der Lösung u_k^m bei bekannter Größe $u_{k,m-1}^- = u_0^{m-1}$ ist dann mit der Testfunktion $\phi_k^m = \xi_0^m \phi_0^m + \xi_1^m \phi_1^m$ und mit $\phi_0^m, \phi_1^m \in H_0^1(\Omega)$:

$$\begin{aligned}
\frac{1}{2}(u_1^m + u_0^m, \phi_0^m) + \frac{1}{3} k_m(\nabla u_0^m, \nabla \phi_0^m) + \frac{1}{6} k_m(\nabla u_1^m, \nabla \phi_0^m) &= (u_1^{m-1}, \phi_0^m) \\
\frac{1}{2}(u_1^m - u_0^m, \phi_1^m) + \frac{1}{6} k_m(\nabla u_0^m, \nabla \phi_1^m) + \frac{1}{3} k_m(\nabla u_1^m, \nabla \phi_1^m) &= 0.
\end{aligned}$$

Hier haben wir der Einfachheit halber $f = 0$ angenommen. Wir haben also ein gekoppeltes Gleichungssystem mit zwei Unbekannten u_0^m und u_1^m. Wir führen für den schwachen Laplace-Operator wieder eine Operatorschreibweise ein:

$$(\mathcal{A}u, v) = (\nabla u, \nabla v) \quad \forall u, v \in H_0^1(\Omega),$$

und können das Gleichungssystem kompakt schreiben:

$$\begin{pmatrix} \mathcal{I} + \frac{2}{3}k_m\mathcal{A} & \mathcal{I} + \frac{1}{3}k_m\mathcal{A} \\ -\mathcal{I} + \frac{1}{3}k_m\mathcal{A} & \mathcal{I} + \frac{2}{3}k_m\mathcal{A} \end{pmatrix} \begin{pmatrix} u_0^m \\ u_1^m \end{pmatrix} = \begin{pmatrix} 2u_1^{m-1} \\ 0 \end{pmatrix},$$

wobei wir die rechte Seite kurz zusammenfassen. Dieses (unendlich dimensionale) Gleichungssystem lösen wir mit der Gauß-Elimination formal nach u_1^m. Wir multiplizieren dazu die erste Zeile von links mit $(-\mathcal{I}+\frac{1}{3}k_m\mathcal{A})$ und die zweite von rechts mit $\mathcal{I}+\frac{2}{3}k_m\mathcal{A}$ und ziehen dann die erste von der zweiten ab. So erhalten wir:

$$(\mathcal{I} + \frac{2}{3}k_m\mathcal{A})(\mathcal{I} + \frac{2}{3}k_m\mathcal{A})u_1^m - (-\mathcal{I} + \frac{1}{3}k_m\mathcal{A})(\mathcal{I} + \frac{1}{3}k_m\mathcal{A})u_1^m = 2(\mathcal{I} - \frac{1}{3}k_m\mathcal{A})u_1^{m-1}$$

Ausmultiplizieren liefert:

$$(\mathcal{I} + \frac{2}{3}k_m\mathcal{A} + \frac{1}{6}k_m^2\mathcal{A}^2)u_1^m = u_1^{m-1} - \frac{1}{3}k_m\mathcal{A}u_1^{m-1}.$$

Zur Bestimmung von u_1^m ist also eine Gleichung zu lösen, welche als Polynom in $\mathcal{A}$ gegeben ist. Dieses Polynom $1 + \frac{2}{3}z + \frac{1}{6}z^2$ für $z = -k_m\mathcal{A}$ ist gerade der rationale Anteil der subdiagonalen $\{2,1\}$-Padé-Approximation. Im expliziten Anteil auf der rechten Seite tritt für $z = -k_m\mathcal{A}$ mit $1-\frac{1}{3}z$ gerade der Zähler der Padé-Approximation auf.

Remark 5.20 ($dG(r)$-Verfahren)**.** *Das $dG(0)$-Verfahren zur Diskretisierung der Wärmeleitungsgleichung entspricht bis auf numerischen Quadraturfehler gerade dem impliziten Euler-Verfahren. Die allgemeinen $dG(r)$-Verfahren zur Diskretisierung der Wärmeleitungsgleichung entsprechen bis auf Quadratur jeweils einem Runge-Kutta-Verfahren und können auch durch eine subdiagonale Padé-Approximationen $\{r+1, r\}$ ausgedrückt werden. Hieraus können wir folgern, dass jedes $dG(r)$-Verfahren (bei entsprechender Integration) in den Gitterpunkten die Ordnung $2r+1$ hat und stark A-Stabil ist.*

A priori Fehleranalyse

In diesem Abschnitt wollen die eine a priori Fehlerabschätzung für das $dG(r)$ Verfahren herleiten. Dazu soll zunächst der reine Zeitfehler, also die Größe $e_k := u - u_k$, der Fehler zwischen Lösung $u \in W_I$ von (5.21) und $u_k \in W_k^{(r)}$ der Lösung von (5.22) abgeschätzt werden. Aufgrund der Konstruktion $W_k^{(r)} \subset W_I$ können wir von einer konformen Methode sprechen und es gilt:

Lemma 5.21 (Galerkin-Orthogonalität)**.** *Für den Fehler $e_k := u - u_k$ der Lösungen $u \in W_I$ der Wärmeleitungsgleichung (5.21) und $u_k \in W_k^{(r)}$ von (5.22) gilt die* Galerkin-Orthogonalität

$$B_I(e_k, \phi_k) = 0 \quad \forall \phi_k \in W_k^{(r)},$$

in der Orts-Zeit Bilinearform

$$B_I(u, \phi) := \sum_{m=1}^{M} \int_{I_m} \Big\{ (\partial_t u, \phi)_\Omega + (\nabla u, \nabla \phi)_\Omega \Big\} dt + \sum_{m=1}^{M} ([u]^{m-1}, \phi_{m-1}^+)_\Omega.$$

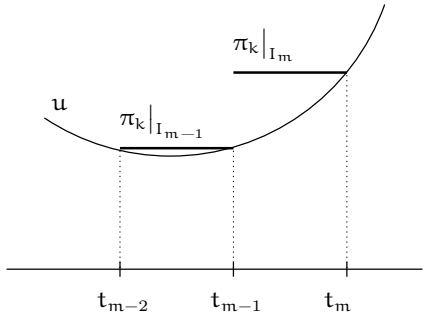


Abbildung 5.2: Projektion π_k in den Raum der stückweise konstanten Funktionen.

Proof: $u \in W_I$ ist Lösung von:

$$B_I(u, \phi) = \int_I (f, \phi)\, dt \quad \forall \phi \in W_I,$$

und $u_k \in W_k^{(r)}$ ist Lösung von

$$B_I(u_k, \phi_k) = \int_I (f, \phi_k)\, dt \quad \forall \phi_k \in W_k^{(r)}.$$

Der Beweis folgt unmittelbar aus der Beziehung $W_k^{(r)} \subset W_I$ sowie der Linearität von $B_I(\cdot, \cdot)$.

□

Wir werden nun den folgenden Satz beweisen:

Lemma 5.22 (A priori Fehlerabschätzung für das dG(r)-Verfahren)**.** *Für die Lösung* $u \in W_I$ *von (5.21) gelte* $\partial_t^{r+1} u \in L^2(I; L^2(\Omega))$*. Dann gilt für den Fehler der* dG(r)*-Lösung die a priori Abschätzung*

$$\int_I \|e_k\|\, dt \leqslant c k_{max}^{r+1} \int_I \|\partial_t^{r+1} u\|\, dt.$$

Proof:

Wir werden Beweis nur für den Spezialfall $r = 0$ zeigen. Der allgemeine Teil ist technisch aufwändiger.

(i) Wir definieren zunächst einen Projektionsoperator

$$\pi_k : W(I) \to W_k^{(0)},$$

durch die folgende Vorschrift: auf jedem Intervall I_m gelte:

$$\pi_k u\big|_{I_m} \in P^0(I_m; H_0^1(\Omega)), \quad (\pi_k u)_m^- = u_m^-. \tag{5.28}$$

In Abbildung (5.2) ist diese Projektion dargestellt. Am rechten Intervallende stimmen Projektion und Funktion überein.

Wir spalten den Fehler nun auf in einen Projektionsfehler $\eta_k := u - \pi_k u$ und einen diskreten Fehleranteil $\xi_k := \pi_k u - u_k$:

$$e_k = u - u_k = \underbrace{u - \pi_k u}_{=\eta_k} + \underbrace{\pi_k u - u_k}_{=\xi_k}.$$

(*ii*) Wir werden zeigen, dass der Gesamtfehler e_k durch den Projektionsfehler η_k beschränkt ist:

$$\int_I \|e_k\| \, dt \leqslant c \int_I \|\eta_k\| \, dt. \tag{5.29}$$

Hierzu definieren wir ein duales Problem:

$$z_k \in W_k^{(0)}: \quad B_I(\phi_k, z_k) = \int_I (\phi, e_k) \, dt \quad \forall \phi_k \in W_k^{(0)}. \tag{5.30}$$

Wir teilen den Fehler e_k nun auf und nutzen die Definition des dualen Problems für $\phi_k = \xi_k \in W_k^{(0)}$:

$$\int_I \|e_k\|^2 \, dt = \int_I (\eta_k, e_k) \, dt + \int_I (\xi_k, e_k) \, dt = \int_I (\eta_k, e_k) \, dt + B_I(\xi_k, z_k). \tag{5.31}$$

Es gilt:

$$0 = B_I(e_k, \phi_k) = B_I(\xi_k, \phi_k) + B_I(\eta_k, \phi_k) \quad \forall \phi_k \in W_k^{(0)}.$$

Also, weiter bei (5.31)

$$\int_I \|e_k\|^2 \, dt = \int_I (\eta_k, e_k) \, dt - B_I(\eta_k, z_k) \tag{5.32}$$

Nun gilt für die *duale Bilinearform* mit partieller Integration:

$$\begin{aligned} B_I(\eta_k, z_k) &= \sum_{m=1}^{M} \int_{I_m} \Big\{ (\partial_t \eta_k, z_k)_\Omega + (\nabla \eta_k, \nabla z_k)_\Omega \Big\} dt + \sum_{m=1}^{M} (\eta_{k,m-1}^+ - \eta_{k,m-1}^-, z_{k,m-1}^+)_\Omega \\ &= \sum_{m=1}^{M} \int_{I_m} \Big\{ -(\eta_k, \underbrace{\partial_t z_k}_{=0})_\Omega + (\nabla \eta_k, \nabla z_k) \Big\} \, dt \\ &\quad \sum_{m=1}^{M} \Big\{ (\eta_{k,m}^-, z_{k,m}^-) \underbrace{-(\eta_{k,m-1}^+, z_{k,m-1}^+) + (\eta_{k,m-1}^+, z_{k,m-1}^+)}_{=0} - (\eta_{k,m-1}^-, z_{k,m-1}^+) \Big\} \\ &= \int_I (\nabla \eta_k, \nabla z_k) \, dt = - \int_I (\eta_k, \Delta z_k) \, dt, \end{aligned} \tag{5.33}$$

da die Terme mit $\eta_{k,m}^-$ und $\eta_{k,m-1}^-$ nach der Definition der Projektion (5.28) verschwinden. Weiter in (5.32) ergibt sich mit Young'scher Ungleichung:

$$\begin{aligned} \int_I \|e_k\|^2 \, dt &= \int_I (\eta_k, e_k) \, dt + \int_I (\eta_k, \Delta z_k) \, dt \\ &\leqslant \int_I \Big\{ \|\eta_k\| \, \|e_k\| + \|\eta_k\| \, \|\Delta z_k\| \Big\} \, dt \\ &\leqslant \int_I \Big\{ (1+\varepsilon) \|\eta_k\|^2 + \frac{1}{4} \|e_k\|^2 + \frac{1}{4\varepsilon} \|\Delta z_k\|^2 \Big\} \, dt. \end{aligned}$$

Wir nutzen die Stabilitätsabschätzung des $dG(r)$-Verfahren aus Satz 5.17 angewendet auf das duale Problem mit rechter Seite e_k:

$$\int_I \|\Delta z_k\|^2 \, dt \leqslant c \int_I \|e_k\|^2 \, dt.$$

Wir bekommen nun mit $\varepsilon = c$:

$$\int_I \|e_k\|^2 \, dt \leqslant \int_I \left\{ (1+c)\|\eta_k\|^2 + \frac{1}{2}\|e_k\|^2 \right\} dt \;\Leftrightarrow \int_I \|e_k\|^2 \, dt \leqslant 2(1+c) \int_I \|\eta_k\|^2 \, dt.$$

(*iii*) Schließlich benötigen wir eine Abschätzung für den Projektionsfehler $\eta_k = u - \pi_k u$. Diese folgt durch Transformation von jedem Teilintervall I_m auf ein Referenzintervall $\hat{I} = (0,1)$ Anwendung von Poincaré ($\text{da } \eta^-_{k,m} = 0$) und Rücktransformation:

$$\int_{I_m} \|\eta_k\|^2 \, dt = k_m \int_{\hat{I}} \|\widehat{\eta_k}\|^2 \, dt \leqslant k_m c_P^2 \int_{\hat{I}} \|\hat{\partial}_t \widehat{\eta_k}\|^2 \, dt = k_m^2 c_P^2 \int_{\hat{I}} \|\partial_t \eta_k\|^2 \, dt = k_m^2 c_P^2 \int_{\hat{I}} \|\partial_t u_k\|^2 \, dt.$$

□

Ortsdiskretisierung mit Finiten Elementen

A posteriori Fehlerschätzung

5.2.2 Das $cG(r)$-Verfahren

Index