Numerik Partieller Differentialgleichungen I

Version vom 23. März 2015

Vorlesungsskriptum
von
Prof. Dr. Vincent Heuveline
Fakultät für Mathematik und Informatik
Universität Heidelberg

Wintersemester 2014/2015

Inhaltsverzeichnis

1. Einführung

Was sind partielle Differentialgleichungen? Eine partielle Differentialgleichung ist eine Gleichung, welche Ableitungen einer gesuchten Funktion $u : \Omega \to \mathbb{R}^d$ enthält, wobei Ω eine offene Teilmenge des $\mathbb{R}^d$ ist. Partielle Differentialgleichungen spielen eine wesentliche Rolle bei der Modellierung von physikalischen, chemischen oder biologischen Phänomenen in mehreren Raumdimensionen bzw. in Raum und Zeit.

Da die Lösung der auftretenden Gleichungen, von einfachsten und oft nicht relevanten Fällen abgesehen, nicht exakt zu bestimmen ist, ist meist der Einsatz numerischer Methoden unabdingbar. Dabei wird üblicherweise das *kontinuierliche Ausgangsproblem* mit einem endlich dimensionalen Problem, dem sogenannten *diskreten Problem*, approximiert und dieses dann mit Hilfe eines Rechners gelöst. Sowohl die Approximationsgüte des diskreten Problems als auch die Möglichkeit, das resultierende algebraische Problem effizient auf einem Computer zu lösen, sind die Kerneigenschaften eines derartigen Diskretisierungsverfahrens. Ihre Studie erfordert eine vielschichtige Analyse, die sich von der rein analytischen Untersuchung der gegebenen partiellen Differentialgleichungen bis hin zu Diskretisierungsaspekten und deren algorithmischer Umsetzung erstreckt. Gegenstand dieses Manuskripts ist es, diesen Bogen einführend zu spannen. Besonderes Gewicht wird auf Galerkin-Verfahren - insbesondere die *Finite Elemente Methode* (FEM) - gelegt, deren theoretischen und praktischen Aspekte im Detail behandelt werden.

1.1. Notation

Wir betrachten offene, beschränkte Gebiete $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, mit Rand $\partial\Omega$. Im Allgemeinen heißt die gesuchte Funktion $u(t)$, $u(x)$ oder $u(x, t)$ für Argumente $x \in \mathbb{R}^d$ bzw. $t \in (0, T)$. Die Variable t steht typischerweise für eine Zeitvariable und $x = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$ für eine Ortsvariable. Für Funktionen $u(t)$, $u(x)$ oder $u(t, x)$ werden totale bzw. partielle Ableitungen geschrieben als

$$\mathrm{d}_t u := \frac{\mathrm{d}u}{\mathrm{d}t}, \quad \partial_t u := \frac{\partial u}{\partial t}, \quad \partial_i u := \frac{\partial u}{\partial x_i}.$$

Der *Gradient einer skalaren Funktion* sowie die *Divergenz einer Vektorfunktion* werden folgendermaßen bezeichnet:

$$\begin{aligned} \operatorname{grad} u := \quad & \nabla u := \quad (\partial_1 u, \ldots, \partial_d u)^\top, \\ \operatorname{div} u := \quad & \nabla \cdot u := \quad \partial_1 u_1 + \cdots + \partial_d u_d. \end{aligned}$$

Der Operator $\nabla = (\partial_1, \ldots, \partial_d)^\top$ ist der sogenannte *Nabla-Operator*. Die Kombination von Divergenz und Gradientenbildung ergibt den Laplace-Operator

$$\Delta u := \nabla \cdot (\nabla u) = \partial_1^2 u + \cdots + \partial_d^2 u.$$

Weiterhin wird die Ableitung in Richtung $n \in \mathbb{R}^d$ mit $\partial_n u := n \cdot \nabla u$ angegeben.

1.2. Aspekte der Modellbildung

1.2.1. Erhaltungsgleichungen

Die Grundlage einer Vielzahl von mathematischen Modellen beruht auf Erhaltungsgleichungen. Der Einfachheit halber beschränken wir uns auf das Prinzip der Massenerhaltung in einem zeitlich festgehaltenen Kontrollvolumen $V \subset \Omega$. Die zeitliche Änderung der Masse einer Substanz ergibt sich aus der

- durch ∂V eintretenden Masse,
- durch ∂V austretenden Masse,
- durch Quellen erzeugten Masse und der
- durch Senken vernichteten Masse.

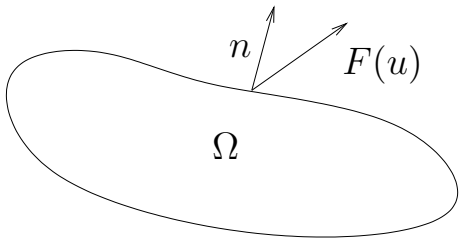


Abbildung 1.1.

Wir verwenden die folgenden physikalische Größen:

Variable	Einheit	Bezeichnung
$\rho(t,x)$	$[\mathrm{kg}\, m^{-3}]$	Dichte des Stoffes am Ort x zur Zeit t
$v(t,x)$	$[ms^{-1}]$	Geschwindigkeit
$F(\rho)$	$[\mathrm{kg}\, m^{-2}s^{-1}]$	Teilchenstromdichte (Fluss durch die Oberfläche des Volumens)
$q(t,x)$	$[\mathrm{kg}\, m^{-3}s^{-1}]$	Quelldichte

Die gesamte Änderung der Dichte ρ in V pro Zeiteinheit ist gleich dem *Massenfluss* durch die gesamte Oberfläche ∂V zuzüglich Quellen und/oder Senken der *Quelldichte* q :

$$\frac{\partial}{\partial t}\int_V \rho(x,t)\mathrm{d}x \;=\; -\int_{\partial V} F(\rho)\cdot n\mathrm{d}S + \int_V q(x,t)\mathrm{d}x, \tag{1.1}$$

wobei n den nach außen gerichteten Normalenvektor beschreibt. Daraus folgt nach Anwendung des Gauß'schen Integralsatzes

$$\frac{\partial}{\partial t}\int_V \rho(x,t)\mathrm{d}x \;=\; \int_V -\nabla\cdot F(\rho) + q(x,t)\mathrm{d}x. \tag{1.2}$$

Da diese Gleichung für beliebige zusammenhängende Kontrollvolumina $V \subset \Omega$ gelten soll, gilt

$$\frac{\partial \rho}{\partial t}(x,t) = -\nabla\cdot F(\rho) + q(x,t) \quad (x\in\Omega, t\in(0,T)). \tag{1.3}$$

Diese Gleichung ist für Erhaltungsgesetze generisch. Sei nun allgemein u eine Erhaltungsgröße. Der vorherigen Notation entsprechend gilt dann

$$\frac{\partial u}{\partial t}(x,t) = -\nabla\cdot F(u) + q(x,t) \quad (x\in\Omega, t\in(0,T)). \tag{1.4}$$

Der verallgemeinerte *Flussvektor* $F(u)$ fasst die Wirkung der Spannungen und Flüsse zusammen. Er wird üblicherweise gemäß

$$F(u) = F_D(u) + F_T(u) \tag{1.5}$$

in einen *Diffusionsstromanteil* (Spannung) F_D und einen *Transportstromanteil* (Fluss) F_T zerlegt. Für den Diffusionsanteil postuliere man

$$F_D(u) := -K(x,t)\nabla_x u(t,x), \tag{1.6}$$

wobei die Matrix $K(t,x)$ die Diffusionskoeffizienten darstellt. Im Fall der Massenerhaltung entspricht diese Annahme dem *1. Fickschen Gesetz*. Die Diffusionsströme fließen in Richtung des größten Konzentrationsgefälles und sind in ihrer Stärke proportional zur Konzentrationsdifferenz. Der Transportanteil ist definiert durch

$$F_T(u) := v_T(t,x)u(t,x). \tag{1.7}$$

Aus (1.4), (1.6) und (1.7) ergibt sich

$$\frac{\partial u}{\partial t}(x,t) = \nabla \cdot [K(x,t)\nabla_x u] - (\nabla \cdot v_T)u - v_T \cdot \nabla u + q(x,t), \quad (x \in \Omega, t \in (0,T)). \tag{1.8}$$

Zur vollständigen Beschreibung des physikalischen Vorgangs fehlt noch die Vorgabe der Randbedingungen. Dies betrachten wir im Folgenden im Zusammenhang mit Spezialfällen der Gleichung (1.8).

1.2.2. Kontinuitätsgleichung

Unter der Annahme, dass es keine Massenquellen oder -senken in Ω gibt und der Diffusionsanteil verschwindet ($F_D = 0$), gilt:

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho v), \quad (x \in \Omega, t \in (0,T)).$$

Diese Gleichung heißt *Kontinuitätsgleichung*.

1.2.3. Die Wärmeleitungsgleichung als Prototyp einer parabolischen Gleichung

Die Alltagserfahrung legt nahe, Körpern einen *Wärmeinhalt* bzw. eine *Wärmeenergie* zuzuordnen, die sowohl proportional zu ihrer Temperatur als auch zu ihrer Masse ist. Die Proportionalitätskonstante c ist die sogenannte *spezifische Wärme* und besitzt die Einheit $[J\,\mathrm{kg}^{-1}\,K^{-1}]$. Zum Zeitpunkt t in einem Gebiet Ω ist diese Wärmeenergie gegeben durch

$$W_\Omega(t) = c \int_\Omega \rho(x)T(t,x)\mathrm{d}x, \tag{1.9}$$

wobei $\rho(x)$ die Massedichte $[\mathrm{kg}\,m^{-3}]$ des Materials beschreibt, und $T(t,x)$ die Temperatur bezeichnet.

Das *Fouriersche Gesetz* ist von entscheidender Bedeutung zur Beschreibung von Wärmeaustauschvorgängen. Dieses besagt, dass ein durch Temperaturunterschied zustande kommender Wärmestrom durch ein Flächenstück von einem wärmeren in ein kälteres Gebiet proportional zu dem Temperaturunterschied ist. Bezeichne mit j den Wärmestrom, dann gilt

$$j(t,x) = -k \cdot \nabla T(x,t), \quad (x \in \Omega, t \geq 0), \tag{1.10}$$

wobei k $[Jm^{-1}s^{-1}K^{-1}]$ die materialabhängige *Wärmeleitfähigkeit* beschreibt. Dieses Gesetz entspricht der Gleichung (1.6) und ist äquivalent zum 1. Fickschen Gesetz im Zusammenhang mit der Massenerhaltung.

In einem ruhenden Medium ($v = 0$) ergibt sich aus (1.8)

$$c\rho(x)\frac{\partial T}{\partial t}(x,t) \quad = \quad \nabla \cdot (k\nabla T) + q(x,t), \quad (x \in \Omega, t \geq 0). \tag{1.11}$$

Von nun an wird immer vorausgesetzt, dass das Material im Gebiet Ω homogen ist, d.h. $\rho = \rho(x)$ und dass der Körper keine Energiequellen oder -senken besitzt. Daraus ergibt sich

$$\frac{\partial T}{\partial t}(x,t) \quad = \quad \nu \Delta T(x,t), \quad (x \in \Omega, t \geq 0), \tag{1.12}$$

wobei

$$\nu = \frac{k}{c\rho} \tag{1.13}$$

die *Temperaturleitfähigkeit* beschreibt. Diese Gleichung ist die so genannte *Wärmeleitungsgleichung* und gilt als Prototyp für die *parabolischen Gleichungen*.

Zur vollständigen Beschreibung des Wärmeaustausches müssen Randbedingungen vorgegeben werden. Es gibt eine Reihe unterschiedlicher Randbedingungen, die verschiedene physikalische Situationen beschreiben. Im Folgenden werden wir uns auf zwei Klassen von Randbedingungen beschränken. Die *Dirichlet-Randbedingung*

$$T(t,x) = T_0(t,x), \quad (x \in \partial\Omega,\ t \geq 0), \tag{1.14}$$

bedeutet, dass die Temperatur des Körpers am Rand auf dem vorgegebenen Wert T_0 etwa durch ständige Kühlung bzw. Heizung gehalten wird. Die *Neumann-Randbedingung*

$$\nabla T(t,x) \cdot n = T_0, \quad (x \in \partial\Omega,\ t \geq 0), \tag{1.15}$$

bedeutet, dass der Wärmefluss am Rand auf dem vorgegebenen Wert T_0 gehalten wird. Falls $T_0 = 0$ ist, bedeutet das, dass der Körper Ω thermisch isoliert wird.

Beispiel 1.1 Es wird angenommen, dass $\Omega = (0, \pi)$, $\nu = 1$ und

$$u(t,0) \quad = \quad u(t,\pi) = 0, \quad t \geq 0, \tag{1.16}$$

$$u(0,x) \quad = \quad \sum_{k=1}^{\infty} a_k \sin(kx), \quad (x \in \Omega). \tag{1.17}$$

Zu diesen Anfangswerten ist

$$u(t,x) = \sum_{k=1}^{\infty} a_k e^{-k^2 t} \sin(kx), \quad (x \in \Omega, t \geq 0), \tag{1.18}$$

offensichtlich eine Lösung der Wärmeleitungsgleichung (1.12). Im Falle eines unendlich langen Stabes $\Omega = (-\infty, +\infty)$ kann die Lösung mit Fourier-Integralen anstatt Fourier-Reihen dargestellt werden (siehe z.B. [12], S. 47)

$$u(x,t) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} e^{-|x-y|^2/(4t)} u(y,0) \mathrm{d}y, \quad (x \in \mathbb{R},\ t > 0). \tag{1.19}$$

In diesem Fall entfallen die Randwerte (1.16). Es ist bemerkenswert zu sehen, dass die Lösung im Punkt (t,x) von den Anfangswerten auf dem ganzen Gebiet abhängt. Die Informationsausbreitungsgeschwindigkeit ist unendlich. Diese Eigenschaft hat entscheidende Folgen für die numerische Behandlung derartiger Gleichungen.

1.2.4. Die Wellengleichung als Prototyp hyperbolischer Differentialgleichungen

Die *Wellengleichung* fungiert als vereinfachtes Modell für die Beschreibung einer schwingenden Saite ($d = 1$), einer Membran ($d = 2$) oder eines elastischen Körpers ($d = 3$). Die gesuchte Größe ist die *Verschiebung* $u(t, x)$ in einer vorgegebenen Richtung am Ort x und zur Zeit t. Das *2. Newtonsche Gesetz*, das die Impulserhaltung beschreibt, liefert

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} \int_V u \mathrm{d}x = - \int_{\partial V} F \cdot n \mathrm{d}S, \tag{1.20}$$

wobei F die Kraft, die am Rand $\partial\Omega$ auf Ω ausgeübt wird, bezeichnet und n den nach außen gerichteten Normalenvektor beschreibt. Zur Vereinfachung wurde in (1.20) die Dichte konstant als $\rho \equiv 1$ angenommen.

Man beachte, dass diese Gleichung ein Erhaltungsgesetz beschreibt. Entsprechend kann die Herleitung vom Paragraph 1.2.1 verwendet werden. Für kleine Verschiebungen wird angenommen, dass die Kraft, die auf den Körper Ω ausgeübt wird, linear von dem Gradienten der Verschiebung abhängt

$$F(t, x) = -a \nabla u(t, x), \quad (x \in \Omega, t \geq 0). \tag{1.21}$$

Daraus folgt

$$\begin{aligned} \int_\Omega \partial_{\mathrm{tt}}^2 u(t, x) \mathrm{d}x &= \int_\Omega \nabla \cdot F(t, x) \mathrm{d}x, \quad (x \in \Omega,\ t > 0), & (1.22) \\ &= -a \int_\Omega \Delta u(t, x) \mathrm{d}x, \quad (x \in \Omega,\ t > 0). & (1.23) \end{aligned}$$

Da diese Gleichung für beliebige zusammenhängende Kontrollvolumina $V \subset \Omega$ gelten soll, gilt

$$\partial_{\mathrm{tt}}^2 u(t, x) = -a \Delta u(t, x), \quad (x \in \Omega, t > 0). \tag{1.24}$$

Diese Gleichung ist die sogenannte *Wellengleichung* und gilt als Prototyp für die *hyperbolischen Gleichungen*. Für die Anfangsbedingungen suggeriert die physikalische Interpretation, dass man sowohl die Verschiebung u als auch die Geschwindigkeit u_t zum Zeitpunkt $t = 0$ zur sinnvollen Problemstellung vorgibt

$$\begin{aligned} u(0, x) &= f(x) \quad (x \in \Omega), & (1.25) \\ \partial_t u(0, x) &= g(x) \quad (x \in \Omega). & (1.26) \end{aligned}$$

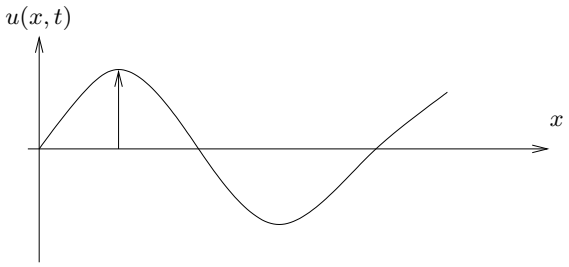


Abbildung 1.2.: Schwingende Saite mit Auslenkung

Beispiel 1.2 Wir nehmen an, dass $\Omega = \mathbb{R}$ und $a = 1$. Es sei die folgende Variablentransformation

$$\begin{aligned} \xi &:= x + t, \\ \eta &:= x - t. \end{aligned}$$

gegeben . Dann gilt

$$\begin{aligned} u_x &= u_\xi + u_\eta, \\ u_t &= u_\xi - u_\eta, \\ u_{xx} &= u_{\xi\xi} + 2u_{\xi\eta} + u_{\eta\eta}, \\ u_{tt} &= u_{\xi\xi} - 2u_{\xi\eta} + u_{\eta\eta}. \end{aligned}$$

Das Einsetzen dieser Formeln in der Wellengleichung (1.24) liefert

$$4u_{\xi\eta} = 0,$$

woraus folgt, dass

$$u(x,t) = \varphi(\xi) + \psi(\eta) = \varphi(x+t) + \psi(x-t). \tag{1.27}$$

Die Anfangsbedingungen (1.25-1.26) ergeben

$$\begin{aligned} f(x) &= \varphi(x) + \psi(x), \\ g(x) &= \varphi'(x) - \psi'(x), \end{aligned}$$

bzw.

$$\varphi' = \frac{1}{2}(f' + g), \tag{1.28}$$

$$\psi' = \frac{1}{2}(f' - g). \tag{1.29}$$

Daraus erhalten wir

$$\varphi(\xi) = \frac{1}{2}f(\xi) + \int_{x_0}^{\xi} g(x)\mathrm{d}x, \tag{1.30}$$

$$\psi(\eta) = \frac{1}{2}f(\eta) - \int_{x_0}^{\eta} g(x)\mathrm{d}x. \tag{1.31}$$

Aus (1.27) folgt schließlich

$$u(x,t) = \frac{1}{2}\left[f(x+t) + f(x-t)\right] + \frac{1}{2}\int_{x-t}^{x+t} g(s)\mathrm{d}s.$$

Man beachte, dass im Gegensatz zum parabolischen Fall (siehe Paragraph 1.2.3) die Lösung $u(t,x)$ nur von den Anfangswerten zwischen den Punkten $x - t$ und $x + t$ abhängt. Für hyperbolische Probleme verbreitet sich die Information nur mit endlicher Geschwindigkeit. Diese Eigenschaft spielt eine große Rolle für die numerische Behandlung derartiger Gleichungen.

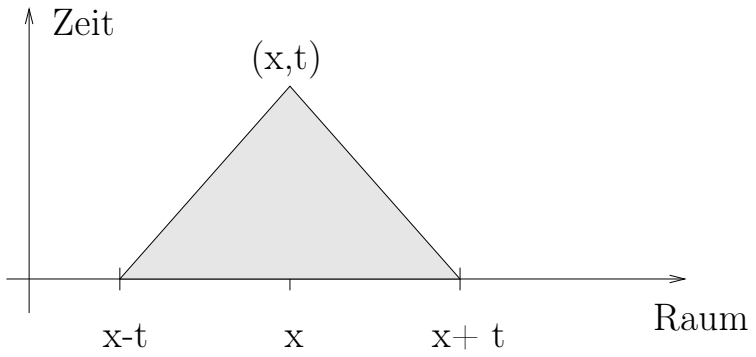


Abbildung 1.3.: Abhängigkeitsbereich bei der Wellengleichung.

1.2.5. Die Laplace-Gleichung als Prototyp elliptischer Differentialgleichungen

Eine Vielfalt von physikalischen Vorgängen kann über die *Laplace-Gleichung* modelliert werden. Eine typische physikalische Interpretation dieser Gleichung kann in Verbindung mit der Untersuchung der Dichte $u(t,x)$ einer chemischen Spezies in einem Gebiet Ω gegeben werden. Dabei wird vorausgesetzt, dass sich die Dichte im Gleichgewicht befindet, d.h. $\partial_t u(t,x) = 0$, und dass Transporteffekte nicht vorhanden sind, d.h. der Transportstromanteil $F_T(t,x) = 0$ (siehe (1.7)). Weiter postuliere man im Sinne des Fourier'schen Gesetzes (siehe (1.10))

$$F_D(x) := -\nu \nabla u(x), \quad (x \in \Omega), \tag{1.32}$$

wobei $\nu > 0$. Mit (1.8) folgt schließlich

$$-\nu \Delta u(x) = 0, \quad (x \in \Omega). \tag{1.33}$$

Diese Gleichung wird *Laplace-Gleichung* oder auch *Potentialgleichung* genannt und gilt als Prototyp für die *elliptischen Gleichungen*. Man beachte, dass für $f : \Omega \to \mathbb{R}$ die Gleichung

$$-\nu \Delta u(x) = f(x), \quad (x \in \Omega), \tag{1.34}$$

Poisson-Gleichung genannt wird.

1.2.6. Zusammenfassung und Ausblick

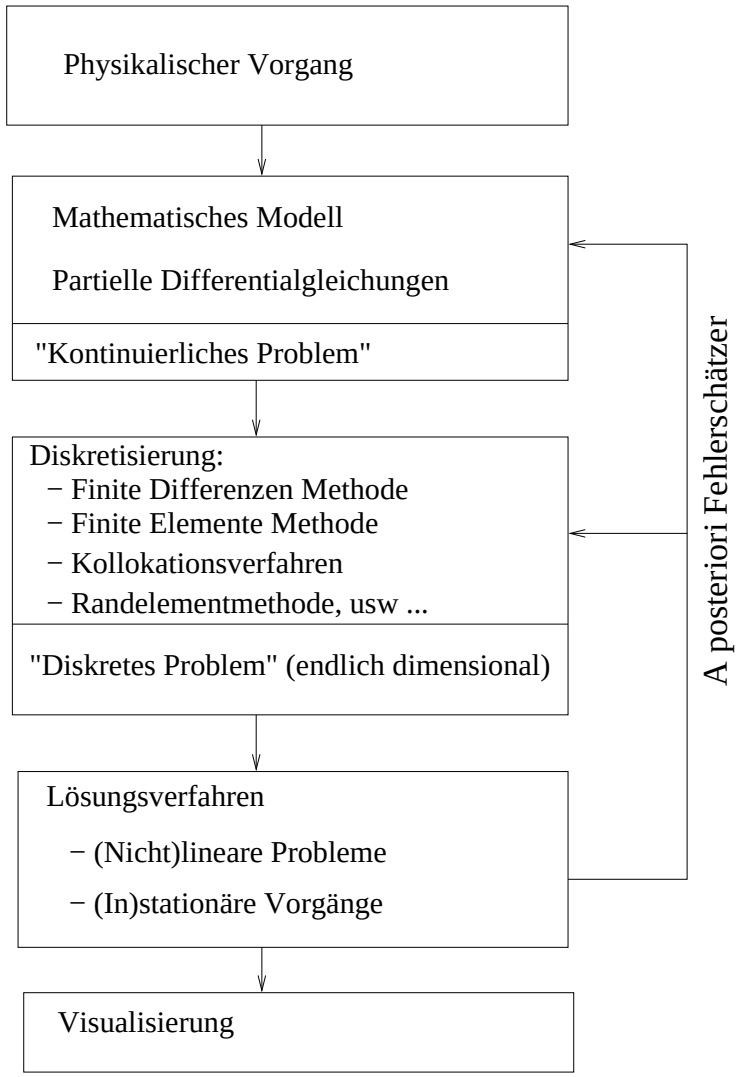


Abbildung 1.4.: Flussdiagramm eines typischen Lösungsprozesses.

1.3. Klassifizierung und Charakteristiken

Wir betrachten *partielle Differentialgleichungen zweiter Ordnung*, d.h.

$$F(x, u, \partial_i u, \partial_{ij}^2 u) = 0, \quad (x \in \Omega) \tag{1.35}$$

wobei $\Omega \subset \mathbb{R}^d$ ist und $i, j \in \{1, \ldots, d\}$. Typischerweise ist die Vorgabe der Werte von u oder Ableitungen von u auf dem Rande $\partial\Omega$ oder einem Teil davon notwendig, um das Problem vollständig zu formulieren. Im allgemeinen Fall stellt die Gleichung (1.35) eine nichtlineare Verknüpfung der gesuchten Funktion mit ihren partiellen Ableitungen bis zur Ordnung zwei dar.

1.3.1. Lineare, halblineare und quasilineare Gleichungen

Definition 1.3 *Eine Differentialgleichung der zweiten Ordnung nennt man*

a) linear*, falls sie die folgende Form hat*

$$\sum_{i,j=1}^{d} a_{ij}(x)\partial_{ij}^2 u(x) + \sum_{i=1}^{d} a_i(x)\partial_i u(x) + a(x)u(x) = f(x), \quad (x \in \Omega), \tag{1.36}$$

b) halblinear*, falls sie die folgende Form hat*

$$\sum_{i,j=1}^{d} a_{ij}(x)\partial_{ij}^2 u(x) + a_0(\partial_1 u(x), \ldots, \partial_d u(x), u(x), x) = 0, \quad (x \in \Omega), \tag{1.37}$$

c) quasilinear*, falls sie die folgende Form hat*

$$\sum_{i,j=1}^{d} a_{ij}(\partial_1 u, \ldots, \partial_d u, u, x)\partial_{ij}^2 u(x) + a_0(\partial_1 u, \ldots, \partial_d u, u, x) = 0, \quad (x \in \Omega). \tag{1.38}$$

Bemerkung 1.4 Quasilineare, halblineare und lineare Differentialgleichungen haben gemeinsam, dass die höchsten Ableitungen linear auftreten.

1.3.2. Typeinteilung

Aus der Definition 1.3 folgt, dass eine lineare Differentialgleichung zweiter Ordnung die folgende Gestalt hat:

$$\sum_{i,j=1}^{d} a_{ij}(x)\partial_{ij}^2 u(x) + \sum_{i=1}^{d} a_i(x)\partial_i u(x) + a(x)u(x) = f(x), \quad (x \in \Omega). \tag{1.39}$$

Wir betrachten die zugeordnete Matrix $A(x) := \{a_{ij}(x)\}_{1 \le i,j \le d}$ der Koeffizientenfunktionen der partiellen Ableitungen zweiter Ordnung. Unter der Annahme der Vertauschbarkeit gemischter partieller Ableitungen $\partial_{ij}^2 u = \partial_{ji}^2 u$ darf die Matrix A als symmetrisch vorausgesetzt werden (andernfalls bildet man neue Koeffizienten $\tilde{a}_{ij} := (a_{ij} + a_{ji})/2$ und die zugeordnete Matrix $\tilde{A}$).

Definition 1.5

1. *Die Gleichung (1.39) heißt* elliptisch *im Punkt x, wenn $A(x)$ negativ definit ist.*
2. *Die Gleichung (1.39) heißt* hyperbolisch *im Punkt x, wenn $A(x)$ einen positiven und $d-1$ negative Eigenwerte hat.*
3. *Die Gleichung (1.39) heißt* parabolisch *im Punkt x, wenn $A(x)$ $d-1$ negative Eigenwerte und einen Eigenwert gleich 0 hat.*

Bemerkung 1.6

1. Eine partielle Differentialgleichung zweiter Ordnung mit *konstanten* Koeffizienten auf ganz Ω ist immer überall in Ω elliptisch, parabolisch oder hyperbolisch.
2. Man beachte, dass der Typ einer Differentialgleichung ortsabhängig sein kann. Zum Beispiel ist die Differentialgleichung in $\mathbb{R}^2$
$$-\partial_{11}u(x_1,x_2) + x_2\partial_{22}u(x_1,x_2) = 0, \tag{1.40}$$
hyperbolisch für $x_2 > 0$, parabolisch für $x_2 = 0$ und elliptisch für $x_2 < 0$.
3. Die Vorzeichenkonvention in Definition 1.5 kann auch umgedreht werden. Dazu multipliziert man (1.39) mit -1 und definiert $\tilde{f} := -f$ als neue rechte Seite. Die charakteristischen Eigenschaften des jeweiligen Gleichungstyps bleiben dadurch erhalten, was im folgenden Unterabschnitt über die Einteilung mittels Charakteristiken deutlich wird. Für die Behandlung der verschiedenen Gleichungstypen im Folgenden ist es jedoch für die Übersichtlichkeit und Klarheit der Darstellung vorteilhaft, sich auf eine Vorzeichenkonvention festzulegen.

Beispiel 1.7

1. Wir betrachten die Poisson-Gleichung
$$-\Delta u = f$$
auf einem Gebiet $\Omega \subset \mathbb{R}^3$. Die zugehörige Koeffizientenmatrix $A(x)$ ist
$$A(x) = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$
Diese Matrix ist offensichtlich für alle $x \in \Omega$ negativ definit und die Poisson-Gleichung somit *elliptisch*.
2. Wir betrachten die Wärmeleitungsgleichung
$$\partial_t u - \Delta u = f$$
auf dem Zeit-Raum-Zylinder (man stelle sich den Raum in 2D als Grundfläche und die Zeit als Höhe des Zylinders vor)
$$\Omega_t := (t_0, T) \times \Omega \subset \mathbb{R}^4, \quad \Omega \subset \mathbb{R}^3,$$
d.h. wir haben eine Zeit- und drei Raumvariablen. Die zugehörige Koeffizientenmatrix $A(x)$ lautet
$$A(x) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

Diese Matrix ist offensichtlich negativ semidefinit mit 3 negativen Eigenwerten und einem Eigenwert gleich 0 für alle $x \in \Omega_t$. Somit ist die Wärmeleitungsgleichung *parabolisch* auf ganz Ω_t.

3. Wir betrachten die Wellengleichung

$$\partial_t^2 u - \Delta u = f$$

auf dem Zeit-Raum-Zylinder

$$\Omega_t := (t_0, T) \times \Omega \subset \mathbb{R}^4, \quad \Omega \subset \mathbb{R}^3,$$

d.h. wir haben eine Zeit- und drei Raumvariablen. Die zugehörige Koeffizientenmatrix $A(x)$ lautet

$$A(x) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

Diese Matrix hat offensichtlich einen positiven und drei negative Eigenwerte für alle $x \in \Omega_t$. Somit ist die Wellengleichung *hyperbolisch* auf ganz Ω_t.

Definition 1.8 *Ein elliptischer Operator der Form (1.39) heißt* gleichmäßig elliptisch, *wenn mit einer Zahl* $\alpha > 0$

$$-\xi^\top A(x)\xi \geq \alpha \|\xi\|^2, \quad (\xi \in \mathbb{R}^d,\ x \in \Omega),$$

gilt. Die Zahl α *wird als* Elliptizitätskonstante *bezeichnet.*

1.3.3. Typeinteilung mittels Charakteristiken

Der allgemeinen Definition 1.5 liegt das Prinzip der sogenannten *Charakteristiken* zugrunde, das wir im Folgenden für den Fall einer Differentialgleichung in zwei Variablen veranschaulichen wollen.

Ausgangspunkt für die Typeinteilung ist ein direkter Lösungsansatz, wie er bei gewöhnlichen Differentialgleichungen angewendet werden kann. Aus

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

erhält man durch sukzessives Differenzieren von $f(t, x)$ Formeln für alle Ableitungen von u:

$$u^{(i)}(0) = \frac{d^{i-1}}{dt^{(i-1)}} f(0, u(0)) =: f^{(i-1)}(0, u^0), \quad i = 1, 2, 3, \ldots,$$

woraus

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

folgt, falls $\sum_{i=1}^{\infty} \frac{t^i}{i!} f^{(i-1)}(0, u^0)$ konvergiert.

Sei $\Omega \subset \mathbb{R}^2$ ein Gebiet. Entlang des parametrisierten Jordan-Kurvenstücks

$$\Gamma = \{(x(\tau), y(\tau)), \tau \in [0, 1]\} \subset \Omega,$$

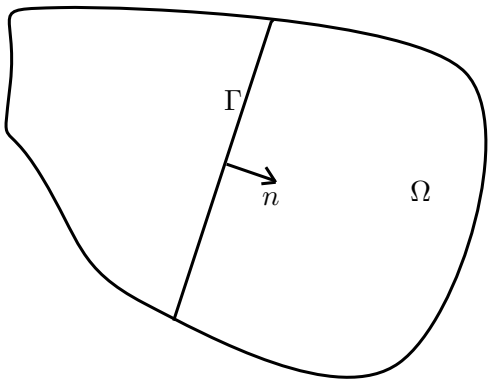


Abbildung 1.5.: Gebiet Ω, Kurve Γ und Normalenvektor n für Methode der Charakteristiken in 2D.

dessen Parametrisierung beliebig oft differenzierbar sei, seien für die Lösung $u(x, y)$ einer Differentialgleichung

$$Lu = a_{11}\partial_x^2 u + 2a_{12}\partial_x\partial_y u + a_{22}\partial_y^2 u + a_1\partial_x u + a_2\partial_y u + au = f$$

die Funktionswerte u sowie ihre Ableitungen $\partial_n u = \nabla u \cdot \vec{n}$ in Normalrichtung $\vec{n}$ zu Γ vorgegeben. Dabei seien die Koeffizienten a_{ij} als konstant vorausgesetzt. Weiterhin sollen nicht alle drei Koeffizienten a_{11}, a_{12} und a_{22} der Ableitungen zweiter Ordnung gleichzeitig Null sein.

Die Vorgabe der Funktionswerte entlang Γ entspricht der Tatsache, dass wir es mit einer Differentialgleichung zweiter Ordnung zu tun haben. Durch diese Vorgaben ist der ganze Gradient $\nabla u = (\partial_x u, \partial_y u)^\top$ entlang Γ bekannt. Wir wollen nun untersuchen, wie sich aus diesen Vorgaben alle weiteren Ableitungen von u entlang Γ bestimmen lassen, um damit, falls dieses Vorhaben erfolgreich ist, wieder einen Taylor-Reihenansatz für u in einer Umgebung von Γ zu machen.

Wir setzen

$$p = \partial_x u, \quad q = \partial_y u, \quad r = \partial_x^2 u, \quad s = \partial_x\partial_y u, \quad t = \partial_y^2 u.$$

Die Differentiation von p und q entlang Γ bezüglich des Parameters τ ergibt:

$$\begin{aligned}\partial_\tau p &= \partial_x p\partial_\tau x + \partial_y p\partial_\tau y = r\partial_\tau x + s\partial_\tau y\\ \partial_\tau q &= \partial_x q\partial_\tau x + \partial_y q\partial_\tau y = s\partial_\tau x + t\partial_\tau y\end{aligned}$$

Zusammen mit der Differentialgleichung $Lu = f$ ergibt dies ein 3×3 Gleichungssystem

$$\begin{aligned}a_{11}r + 2a_{12}s + a_{22}t &= f - a_1 p - a_2 q - au\\ \partial_\tau x r + \partial_\tau y s &= \partial_\tau p\\ \partial_\tau x s + \partial_\tau y t &= \partial_\tau q.\end{aligned}$$

Daraus erhalten wir die Koeffizientenmatrix B:

$$B = \begin{pmatrix} a_{11} & 2a_{12} & a_{22}\\ \partial_\tau x & \partial_\tau y & 0\\ 0 & \partial_\tau x & \partial_\tau y\end{pmatrix}$$

mit der Determinante der Koeffizientenmatrix B

$$\det B = a_{11}\partial_\tau y^2 - 2a_{12}\partial_\tau x\partial_\tau y + a_{22}\partial_\tau x^2$$

i) Fall: $\det B \neq 0$ entlang Γ
In diesem Fall sind alle zweiten Ableitungen r, s, t von u durch Vorgabe von u, $\partial_n u = \nabla u \cdot n$ entlang Γ *eindeutig* bestimmbar.

Durch weitere Differentiation des Gleichungssystems nach x und y erhält man wieder ein System für die dritten Ableitungen $\partial_x r, \partial_x s, \partial_x t$ sowie $\partial_y r, \partial_y s, \partial_y t$ jeweils *mit derselben Koeffizientenmatrix*. Durch weiteres Differenzieren lassen sich alle höheren Ableitungen von u entlang Γ bestimmen. Durch den Reihenansatz

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

bezüglich eines Punktes $(x_0, y_0) \in \Gamma$ erhält man dann in einer Umgebung der Kurve Γ eine Lösung der Differentialgleichung, die auf Γ die vorgegebenen Werte annimmt. Diese nennt man *Lösung der Cauchyschen Anfangswertaufgabe bezüglich der Anfangskurve* Γ.
ii) Fall: $|B| = 0$ in einem Punkt $(x_0, y_0) \in \Gamma$
Die quadratische Gleichung

$$a_{11} \partial_\tau y^2 - 2a_{12} \partial_\tau x \partial_\tau y + a_{22} \partial_\tau x^2 = 0$$

bestimmt gewisse Richtungen $\frac{\partial_\tau y}{\partial_\tau x} = \frac{\mathrm{d}y}{\mathrm{d}x}$ bzw. $\frac{\partial_\tau x}{\partial_\tau y} = \frac{\mathrm{d}x}{\mathrm{d}y}$ von Kurven (mit Graph $y = y(x)$ oder $x = x(y)$) durch den Punkt (x_0, y_0). Zu deren Bestimmung sei etwa angenommen, dass $a_{11} \neq 0$ und $\partial_\tau x \neq 0$. Dann besitzt die Gleichung

$$\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2 - 2\frac{a_{12}}{a_{11}} \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right) + \frac{a_{22}}{a_{11}} = 0$$

die Lösungen

$$\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)_{+/-} = \frac{a_{12}}{a_{11}} \pm \frac{1}{a_{11}} \sqrt{a_{12}^2 - a_{11} a_{22}}$$

Diese entsprechen Steigungen von Kurven durch den Punkt (x_0, y_0), entlang welcher die höheren Ableitungen von u sich ***nicht*** aus den Vorgaben entlang Γ bestimmen lassen. Entlang dieser kritischen bzw. charakteristischen Kurven (*Charakteristiken des Differentialoperators* L) lässt sich die Lösung nicht aus den obigen Vorgaben konstruieren. Die Existenz von Charakteristiken hängt allein von den Koeffizienten der höchsten Ableitungen des Operators L, d.h. seinem *Hauptteil* $a_{11}\partial_x^2 u + 2a_{12}\partial_x \partial_y u + a_{22}\partial_y^2 u$ ab.

Diesem wird die quadratische Form

$$q(x,y) = a_{11}x^2 + 2a_{12}xy + a_{22}y^2$$

zugeordnet. Die Gleichung $q(x,y) = 0$ beschreibt Kegelschnitte in der (x,y)-Ebene:

$$a_{12}^2 - a_{11}a_{22} \begin{cases} < 0: & \text{Ellipse} & \text{elliptische Gleichung} \\ = 0: & \text{Parabel} & \text{parabolische Gleichung} \\ > 0: & \text{Hyperbel} & \text{hyperbolische Gleichung} \end{cases}$$

1.4. Sachgemäß gestellte Probleme

1.4.1. Forderungen von Hadamard

Definition 1.9 *Ein Problem heißt* sachgemäß gestellt *(engl. well-posed), wenn die folgenden drei Forderungen von Hadamard erfüllt sind:*

i. Existenz der Lösung zu gegebenen Randwerten,

ii. Eindeutigkeit der Lösung,

iii. Stabilität, d.h. stetige Abhängigkeit der Lösung von den vorgegebenen Daten.

Andernfalls bezeichnet man das Problem als schlecht gestellt *(engl. improperly posed).*

Beispiel 1.10 Sei $\Omega := \{(x, y) \in \mathbb{R}^2 : y > 0\}$. Man betrachte $\Delta u = 0$ auf Ω mit den Randbedingungen

$$u(x, 0) = \frac{1}{n} \sin(nx), \quad x \in \mathbb{R}, \tag{1.41}$$

$$u_y(x, 0) = 0, \quad x \in \mathbb{R}. \tag{1.42}$$

Wie man leicht prüfen kann, liefert

$$u(x, y) = \frac{1}{n} \cosh(ny) \sin(nx), \quad x \in \Omega, \tag{1.43}$$

eine Lösung, die wie e^{ny} anwächst. Bei $y = 1$ gibt es beliebig kleine Anfangswerte, zu denen beliebig große Lösungen gehören. Die Lösungen sind nicht stabil gegenüber Störungen der Anfangswerte. Es zeigt sich, dass es in dieser Situation nicht sinnvoll ist, sowohl Werte für u als auch für $\partial_y u$ als Randbedingung vorzugeben.

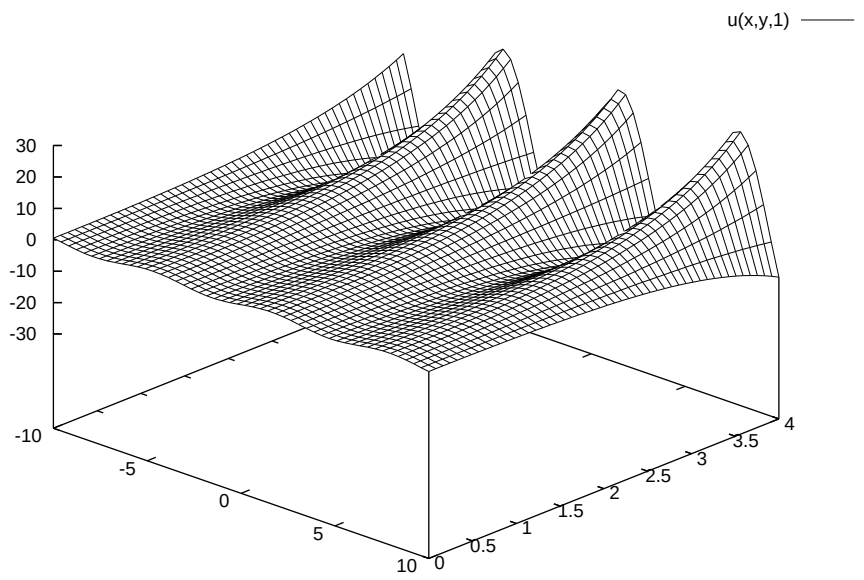


Abbildung 1.6.: Plot der Funktion u für $n = 1$ aus Beispiel 1.10.

Bemerkung 1.11 Man beachte, dass die sachgemäße Wohlgestelltheit eines Problems von der Wahl der Normen der betreffenden Funktionenräume, die für den Stabilitätsbeweis angewendet werden, abhängen kann.

1.5. Wohlgestelltheit der Poisson-Gleichung

Sei $\Omega \subset \mathbb{R}^d$ ein beschränktes Gebiet mit hinreichend glattem Rand $\partial\Omega$. Als prototypischen Modellfall betrachten wir die Poisson-Gleichung.

$$-\Delta u = f, \quad x \in \Omega. \tag{1.44}$$

Es gibt drei Typen von Randbedingungen und zugehörige Randwertaufgaben:

a) *Dirichlet-Randbedingungen*:

$$u = g, \quad x \in \partial\Omega, \tag{1.45}$$

b) *Neumann-Randbedingungen*:

$$\partial_n u = g, \quad x \in \partial\Omega, \tag{1.46}$$

c) *Robin-Randbedingungen*:

$$\alpha u + \partial_n u = g, \quad x \in \partial\Omega. \tag{1.47}$$

Die Randfunktion g wird i.a. als glatt und $\alpha \geq 0$ angenommen. Das Randwertproblem (1.44), (1.45) bezeichnet man auch als das *Dirichlet-Problem* für die Poisson-Gleichung.

Definition 1.12 (Klassische Lösung) *Seien $f \in C^0(\Omega)$ und $g \in C^0(\partial\Omega)$. Eine Lösung $u \in C^2(\Omega) \cap C^0\left(\overline{\Omega}\right)$ des Dirichlet-Problems für die Poisson-Gleichung heißt* klassische Lösung *von (1.44), (1.45).*

Im Folgenden untersuchen wir die Wohlgestelltheit des Dirichlet-Problems für die Poisson-Gleichung.

1.5.1. Eindeutigkeit

Satz 1.13 *Man postuliere die Existenz einer Lösung für das Dirichlet-Problem (1.44-1.45). Dann ist diese Lösung eindeutig.*

Beweis Es seien $u^{(1)}, u^{(2)}$ zwei Lösungen des Randwertproblems (1.44)-(1.45). Offensichtlich gilt:

$$\begin{aligned} -\Delta w &= 0, \quad (x \in \Omega), \\ w &= 0, \quad (x \in \partial\Omega), \end{aligned}$$

für $w := u^{(1)} - u^{(2)}$. Durch einfaches Ausrechnen bekommt man

$$0 = \int_\Omega -\Delta w w = \int_\Omega \nabla w \nabla w - \int_{\partial W} \frac{\partial w}{\partial n} w = \int_\Omega \nabla w \nabla w = \|\nabla w\|_\Omega^2 .$$

Damit $\nabla w = 0 \Rightarrow w = \text{const} \Rightarrow w = 0$. □

1.5.2. Existenz

Der Existenzbeweis für das Dirichlet-Problem ist wesentlich involvierter als der Eindeutigkeitsbeweis. Da eine vollständige Herleitung den Rahmen sprengen würde, wird im Folgenden der Existenzbeweis nur in groben Zügen erörtert. Man beachte, dass wir uns hier auf Techniken zum Existenzbeweis klassischer Lösungen beschränken. Wir verweisen auf [18, 12, 13] für eine vollständige Beweisherleitung. Existenzbeweise im Sinne von so genannten schwachen Lösungen sind Gegenstand des Kapitels 2.

Definition 1.14 (Fundamentallösung) *Die durch*

$$\phi(x) = \begin{cases} -\frac{1}{2\pi}\log(|x|), & d = 2, \\ \frac{1}{(d-2)\alpha_d}|x|^{2-d}, & d \geq 3, \end{cases} \tag{1.48}$$

definierte Funktion $\phi : \mathbb{R}^d\backslash\{0\} \to \mathbb{R}$ *heißt* Fundamentallösung. *Mit* α_d *bezeichnen wir den Flächeninhalt des Randes der Einheitskugel im* $\mathbb{R}^d$ *d.h.* $\alpha_d = |\partial B(0;1)|$.

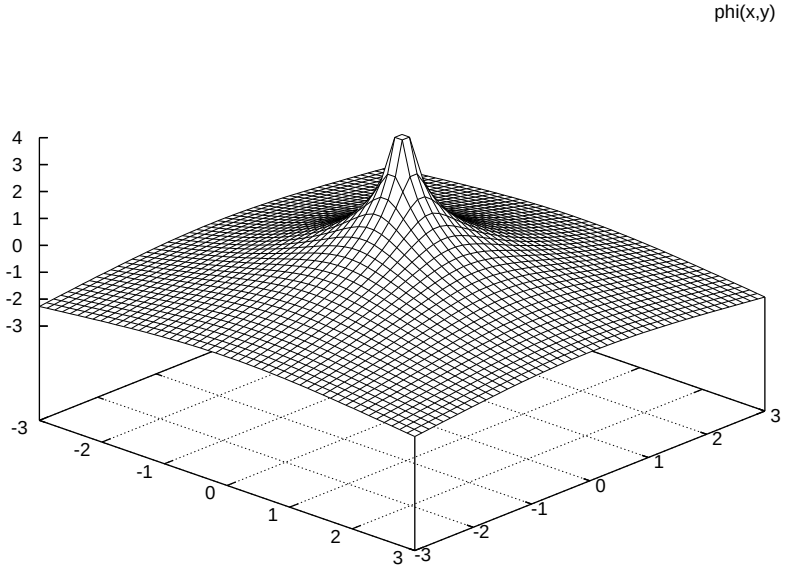


Abbildung 1.7.: Plot der Fundamentallösung $\phi(x)$ in 2D.

Die Fundamentallösung hat eine Singularität im Nullpunkt. Nach Konstruktion ist ϕ harmonisch auf $\mathbb{R}^d\backslash\{0\}$.

Die Fundamentallösung ϕ kann dazu verwendet werden, Lösungen der Laplace-Gleichung in Integralform darzustellen. Seien $u, v \in C^2(\Omega) \cap C^1(\overline{\Omega})$. Die erste Greensche Formel ergibt sich, wenn wir den Gaußschen Satz auf $v\nabla u$ anwenden,

$$\int_\Omega v(x)\Delta u(x)\mathrm{d}x + \int_\Omega \nabla v(x)\nabla u(x)\mathrm{d}x = \int_{\partial\Omega} v(\xi)\partial_n u(\xi)\mathrm{d}S(\xi). \tag{1.49}$$

Die zweite Greensche Formel ergibt sich, indem wir die Rollen von u und v vertauschen und subtrahieren

$$\int_\Omega v(x)\Delta u(x) - u(x)\Delta v(x)\mathrm{d}x = \int_{\partial\Omega} v(\xi)\partial_n u(\xi) - u(\xi)\partial_n v(\xi)\mathrm{d}S(\xi). \tag{1.50}$$

Satz 1.15 (Greensche Darstellungsformel) *Sei $\Omega \subset \mathbb{R}^d$ ein beschränktes Gebiet mit C^1-Rand, sei $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$. Dann gilt für $x \in \Omega$*

$$u(x) = \int_{\partial\Omega} \phi(\xi - x)\partial_n u(\xi) - u(\xi)\partial_n \phi(\xi - x) \mathrm{d}S(\xi) - \int_\Omega \phi(y - x)\Delta u(y) \mathrm{d}y. \tag{1.51}$$

Da $\partial_n u$ auf $\partial\Omega$ zunächst unbekannt ist, eignet sich die direkte Anwendung dieser Darstellungsformel nicht, um u aus den Daten f und g zu berechnen. Wir postulieren, dass es zu jedem $x \in \Omega$ eine Funktion $h^x \in C^2(\Omega) \cap C^1(\overline{\Omega})$ gibt mit

$$-\Delta h^x(y) = 0, \quad y \in \Omega, \tag{1.52}$$
$$h^x(y) = \phi(y - x), \quad y \in \partial\Omega. \tag{1.53}$$

Definition 1.16 (Greensche Funktion) *Die durch*

$$G(x, y) := \phi(y - x) - h^x(y), \quad x, y \in \overline{\Omega} \text{ und } x \neq y, \tag{1.54}$$

definierte Funktion heißt die Greensche Funktion *des Laplace-Operators in Ω.*

Man kann leicht prüfen, dass die Greensche Funktion folgende Eigenschaften besitzt:

i. $G(x, y)$ ist harmonisch in $\Omega \backslash \{x\}$,

ii. $G(x, y)$ ist symmetrisch, d.h. $G(x, y) = G(y, x)$,

iii. Es gilt
$$G(x, y) = 0, \quad (x \in \Omega, y \in \partial\Omega). \tag{1.55}$$

Satz 1.17 *Sei $\Omega \in \mathbb{R}^d$ ein beschränktes Gebiet mit C^1-Rand, sei $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$. Es existiere eine Greensche Funktion in Ω. Dann existiert eine klassische Lösung des Dirichlet-Problems (1.44), (1.45) mit folgender Integraldarstellung*

$$u(x) = -\int_{\partial\Omega} g(\xi)\partial_n G(x, \xi) \mathrm{d}S(\xi) + \int_\Omega G(x, y) f(y) \mathrm{d}y, \quad x \in \Omega. \tag{1.56}$$

Beweis Aus der zweiten Greenschen Formel (1.50) ergibt sich mit $v = h^x$

$$\int_\Omega h^x(y)\Delta u(y) dy = \int_{\partial\Omega} h^x(\xi)\partial_n u(\xi) - u(\xi)\partial_n h^x(\xi) dS(\xi). \tag{1.57}$$

Indem wir (1.57) von der Greenschen Darstellungsformel (1.51) subtrahieren, bekommen wir

$$u(x) = -\int_{\partial\Omega} u(\xi)\partial_n G(x, \xi) dS(\xi) + \int_\Omega G(x, y)\Delta u(y) dy, \quad (x \in \Omega). \tag{1.58}$$

Die Integraldarstellung folgt unmittelbar aus (1.58), (1.44) und (1.45).

□

Im Satz 1.17 wird der Beweis der Existenz einer klassischen Lösung auf den Beweis der Existenz einer Greenschen Funktion zurückgeführt. Auf der Einheitskugel in $\mathbb{R}^d$ kann die Greensche Funktion explizit definiert werden, nämlich

$$G(x, y) := \phi(y - x) - \phi\left(|x|\left(y - \frac{x}{|x|^2}\right)\right). \tag{1.59}$$

Für sie gilt

$$\partial_n G(x,y) = -\frac{1}{\alpha_d}\frac{1-|x|^2}{|x-y|^d}. \tag{1.60}$$

Die Lösung des Dirichlet-Problems (1.44-1.45) lautet,

$$u(x) = \frac{1-|x|^2}{\alpha_d}\int_{\partial B(0,1)} \frac{g(\xi)}{|x-\xi|^d}\,\mathrm{dS}(\xi), \quad (x \in \Omega), \tag{1.61}$$

wobei wir einfachheitshalber $f = 0$ vorausgesetzt haben.

1.5.3. Stetige Abhängigkeit der Lösung von den Daten

Satz 1.18 (Maximumsprinzip) *Für $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$ sei $Lu \le 0$ in Ω, wobei*

$$Lu := -\sum_{i,j=1}^{n} a_{ij}(x)\partial^2_{ij} u,$$

1. *Dann nimmt u sein Maximum auf dem Rand $\partial\Omega$ an.*
2. *Wenn darüber hinaus u sein Maximum an einem inneren Punkt annimmt und Ω zusammenhängend ist, ist u auf Ω konstant.*

Beweis *Schritt 1.* Spezialfall: $Lu < 0$ in Ω. Angenommen, für $x_0 \in \Omega$

$$u(x_0) = \sup_{x\in\Omega} u(x) > \sup_{x\in\partial\Omega} u(x)$$

Bei einer linearen Koordinatentransformation $x \to \xi = Ux$ lautet der Differentialoperator in den neuen Koordinaten

$$Lu = -\sum_{i,k=1}^{d} [U^\top A(x) U]_{i,k}\partial^2_{ij} u.$$

Aufgrund der Symmetrie von $A(x_0)$ lässt sich U so wählen, dass

$$U^\top A(x_0) U = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{pmatrix}, \quad \text{mit } \lambda_k > 0, \quad k = 1,\dots,d.$$

$u_{\xi_i}(x_0) = 0$, $u_{\xi_i\xi_i}(x_0) \le 0 \Rightarrow Lu(x_0) \ge 0$. Damit ist $Lu(x_0) = 0$. Dies führt zum Widerspruch mit der Voraussetzung.

Schritt 2. Der allgemeine Fall: $Lu \le 0$. $\exists \bar{x} \in \Omega : u(\bar{x}) = \sup_{x\in\partial\Omega} u(x)$. Wir definieren für $\varepsilon > 0$

$$h(x) := (x_1 - \bar{x}_1)^2 + \cdots + (x_d - \bar{x}_d)^2$$

auf $\partial\Omega$ beschränkt. Für ein hinreichend kleines ε nimmt $w := u + \varepsilon h$ sein Maximum immer noch im Innern an. Dann haben wir

$$\partial^2_{ij} h = 2\delta_{i,j} = 0$$

mit dem *Kronecker-Symbol*

$$\delta_{i,k} := \begin{cases} 1 & i = k \\ 0 & i \ne k \end{cases}.$$

Dann

$$Lw(x_0) = Lu(x_0) + \varepsilon Lh(x_0) = f(x_0) - 2\varepsilon \sum_i a_{i,i}(x_0) < 0.$$

Damit folgt aber mit Schritt 1, dass w ihr Maximum auf dem Rand annimmt. Dies führt zum Widerspruch mit der Voraussetzung. □

Satz 1.19

1. *Die Lösung der linearen Gleichung $Lu = f$ mit Dirichlet-Randbedingungen hängt stetig von den Randwerten ab. Seien u_1 und u_2 Lösungen der linearen Gleichung $Lu = f$ zu verschieden Randwerten, so ist*
$$\sup_{x\in\Omega} |u_1(x) - u_2(x)| = \sup_{z\in\partial\Omega} |u_1(z) - u_2(z)|.$$

2. *Sei L gleichmäßig elliptisch in Ω. Dann gibt es eine nur von Ω und der Elliptizitätskonstanten α abhängende Zahl c, so dass für jedes $u \in C^2(\Omega) \cap C^2(\bar{\Omega})$*
$$|u(z)| \le \sup_{z\in\partial\Omega} |u(z)| + c \sup_{z\in\Omega} |Lu(z)|.$$

Beweis
Schritt 1. Sei Ω in einem Kreis vom Radius R enthalten. Mittelpunkt des Kreises sei der Nullpunkt (o.B.d.A.). Setze

$$w(x) := R^2 - \sum_i x_i^2.$$

und berechne

$$\partial_{ij}^2 w = -2\delta_{i,j}.$$

Es ergibt sich $Lw \ge 2\alpha$ und $0 \le w \le R^2$ in Ω.

$$v(x) := \sup_{z\in\partial\Omega} |u(z)| + w(x)\frac{1}{2\alpha} \sup_{z\in\partial\Omega} |Lu(z)|.$$

Nach Konstruktion ist $Lv \ge |Lu|$ in Ω.

Schritt 2. $w := u_1 - u_2$. $Lw = Lu_1 - Lu_2 = 0$. $Lu_i = f$ $(i = 1, 2)$. Nach dem Maximumsprinzip

$$w(x) \le \sup_{z\in\partial\Omega} w(z) \le \sup_{z\in\partial\Omega} |w(z)|$$

Ähnlich gilt (Übung!)

$$w(x) \ge -\sup_{z\in\partial\Omega} |w(z)|.$$

Schritt 3. Aus $v \ge |u|$ auf $\partial\Omega$ folgt $-v(x) \le u(x) \le v(x)$ in Ω. Wegen $w \le R^2$ erhalten wir die Behauptung 2. □

1.6. Ergänzende Anmerkungen; Literatur

- Zu der Thematik Modellbildung und physikalische Bedeutung der partiellen Differentialgleichungen im Rahmen der Erhaltungsgesetze bieten die Bücher [4] und [9] eine sehr ausführliche Darstellung.
- Im Hinblick auf die Typeinteilung der partiellen Differentialgleichungen findet sich z.B. in [12] eine ausführliche Beschreibung.
- Die Existenztheorie für klassische Lösungen kann man detailliert in den Büchern [18], [12] und [13] finden.

2. Variationsformulierung elliptischer Randwertaufgaben 2. Ordnung

Die Betrachtungen zur *klassischen Lösungtheorie* im letzten Kapitel haben gezeigt, dass die bisherige Auffassung davon, was eine Lösung einer partiellen Differentiagleichung ist, im Allgemeinen zu einschränkend ist. Aus dieser Erkenntnis hat sich ein abgeschwächter Lösungsbegriff entwickelt, für welchen mit Hilfmitteln der Funktionalanalysis weitreichende Resultate bezüglich Existenz und Eindeutigkeit sowie Regularität der gewonnenen Lösungen erzielt werden können.

2.1. Das Dirichletsche Prinzip

Wir betrachten das Problem

$$\begin{aligned} -\Delta u &= 0 \quad \text{in } \Omega, \\ u &= g \quad \text{auf } \partial\Omega \end{aligned}$$

mit $\Omega \subset \mathbb{R}^d$. Das *Dirichletsche Prinzip* beruht auf der folgenden Beobachtung: Es sei $u \in C^2(\Omega)$ eine Funktion mit $u = g$ auf $\partial\Omega$ und

$$\underbrace{\int_\Omega |\nabla u|^2 \mathrm{d}x}_{D(u):=} = \min\left\{ \int_\Omega |\nabla v|^2 \mathrm{d}x : v : \Omega \to \mathbb{R}, v = g \text{ auf } \partial\Omega \right\}.$$

Behauptung: u ist Lösung unseres Randwertproblems.

Beweis Sei $\eta \in C_0^\infty(\Omega) := \{\varphi \in C^\infty(\Omega) : \mathrm{clos}(\mathrm{supp}(\varphi)) \text{ kompakt in } \Omega \text{ enthalten}\}$, wobei $\mathrm{supp}(\varphi) := \{x : \varphi(x) \neq 0\}$ der *Träger* von φ genannt wird.

Die Funktion

$$\alpha(t) := \int_\Omega |\nabla(u + t\eta)|^2 \mathrm{d}x$$

hat ein Minimum bei $t = 0$, oder, mit anderen Worten, $\alpha'(0) = 0$.

$$\alpha'(0) = 0 \Leftrightarrow \int_\Omega \nabla u(x) \nabla \eta(x) \mathrm{d}x = 0 \quad \forall \eta \in C_0^\infty(\Omega).$$

□

Lemma 2.1 *$g \in C^0(\Omega)$ erfülle*

$$\int_\Omega g(x)\eta(x)\mathrm{d}x = 0 \quad \forall \eta \in C_0^\infty(\Omega).$$

Dann ist $g(x) = 0$ für $x \in \Omega$.

Beweis Übung. □

Falls $u \in C^2(\Omega)$, so gilt $\Delta u(x) = 0$ für $x \in \Omega$. Beachte, dass in der obigen Herleitung zwei gähnende Löcher klaffen:

- In welchen Funktionenräumen nimmt das Dirichletintegral das Infimum an?
- Wenn ein Minium existiert, ist es glatt genug, um Lösung des Randwertproblems zu sein?

Sei $(u_n)_{n\in\mathbb{N}}$ eine Minimalfolge für D

$$\lim_{n\to\infty} D(u_n) = \inf\{D(v) : v : \Omega \to \mathbb{R}, v = g \text{ auf } \partial\Omega\} =: \kappa$$

Frage: Ist $(u_n)_{n\in\mathbb{N}}$ Cauchy-Folge? Beachte dabei, dass $D(u) \geq 0$.

Lemma 2.2 *Das Dirichletintegral D ist konvex, d.h.*

$$D(tu + (1-t)v) \leq tD(u) + (1-t)D(v)$$

für alle u, v und alle $t \in [0,1]$.

Beweis

$$\begin{aligned} D(tu + (1-t)v) &= \int_\Omega |t\nabla u + (1-t)\nabla v|^2 \mathrm{d}x \\ &\overset{(*)}{\leq} \int_\Omega |t\nabla u|^2 + |(1-t)\nabla v|^2 \mathrm{d}x \\ &\leq tD(u) + (1-t)D(u). \end{aligned}$$

Der Schritt $(*)$ nutzt dabei aus, dass die Funktion $w \mapsto |w|^2$ konvex ist, im darauffolgenden geht wesentlich ein, dass $t^2 \leq t$ für $|t| \leq 1$. □

Somit folgt

$$\begin{aligned} D(u_n - u_m) &= \int_\Omega |\nabla(u_n - u_m)|^2 \mathrm{d}x \\ &= 2\int_\Omega |\nabla u_n|^2 + 2\int_\Omega |\nabla u_m|^2 - 4\int_\Omega \left|\frac{\nabla u_n + \nabla u_m}{2}\right|^2 \\ &= 2D(u_n) + 2D(u_m) - 4D\left(\frac{u_n + u_m}{2}\right) \to 0 \end{aligned}$$

Hierbei wurde benutzt, dass

$$\kappa \leq D\left(\frac{u_n + u_m}{2}\right) \leq \frac{1}{2}D(u_n) + \frac{1}{2}D(u_m) \to \kappa \quad (n, m \to \infty).$$

Das bedeutet, dass $(\nabla u_n)_{n\in\mathbb{N}}$ eine Cauchyfolge in der Topologie des $L^2(\Omega)$ ist.

2.2. Sobolev-Räume

Definition 2.3 *Es sei $u \in L^2(\Omega)$. $v \in L^2(\Omega)$ heißt* schwache Ableitung *von u in Richtung x_i (mit $x = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$), falls*

$$\int_\Omega \varphi v = -\int_\Omega u \frac{\partial \varphi}{\partial x_i} \quad \forall \varphi \in C_0^\infty(\Omega)$$

gilt. Wir schreiben $v = D_i u$.

Bemerkung. Ist $u \in C^1(\Omega)$, so ist u schwach differenzierbar und die schwachen Ableitungen sind gerade die gewöhnlichen Ableitungen.

Definition 2.4 (Sobolevraum) *Der* Sobolevraum $W^{1,2}(\Omega)$ *ist definiert als der Raum derjenigen $u \in L^2(\Omega)$, die in jeder Richtung x_i $(i = 1, \ldots, n)$ schwache Ableitungen aus $L^2(\Omega)$ besitzen. In $W^{1,2}(\Omega)$ definieren wir durch*

$$\begin{aligned} (u,v)_{W^{1,2}(\Omega)} &:= \int_\Omega uv + \sum_{i=1}^d \int_\Omega D_i u D_i v, \\ \|u\|_{W^{1,2}(\Omega)} &:= \sqrt{(u,u)_{W^{1,2}(\Omega)}}, \\ |u|_{W^{1,2}(\Omega)} &:= \sqrt{\sum_{i=1}^d \int_\Omega (D_i u)^2} \end{aligned}$$

ein Skalarprodukt, eine Norm und eine Seminorm. Wir definieren $H^{1,2}(\Omega)$ als den Abschluss von $C^\infty(\Omega) \cap W^{1,2}(\Omega)$ bezüglich der $W^{1,2}(\Omega)$-Norm und $H_0^{1,2}(\Omega)$ als den Abschluss von $C_0^\infty(\Omega)$ bezüglich dieser Norm.

Literaturhinweis: Adams, Sobolev Spaces.

Satz 2.5 *$W^{1,2}(\Omega)$ ist vollständig bezüglich $\|\cdot\|_{W^{1,2}(\Omega)}$, also ein Hilbertraum. Es gilt $W^{1,2}(\Omega) = H^{1,2}(\Omega)$.*

Beispiel. $\Omega := [-1, 1]$.

1. $u(x) := |x|$. $u \in W^{1,2}(\Omega)$ mit

   $$Du(x) = \begin{cases} 1 & 0 < x < 1 \\ -1 & -1 < x < 0 \end{cases},$$

   denn

   $$\int_0^1 \varphi(x)\mathrm{d}x + \int_{-1}^0 -\varphi(x)\mathrm{d}x = \int_{-1}^1 \varphi'(x)|x|\mathrm{d}x.$$

2. Definiere

   $$u(x) := \begin{cases} 1 & 0 \le x < 1 \\ 0 & -1 < x < 0 \end{cases}$$

   ist nicht schwach differenzierbar, da $Du = 0$:

   $$0 = \int_{-1}^1 \varphi(x) \cdot 0 \mathrm{d}x = -\int_{-1}^1 \varphi'(x)u(x)\mathrm{d}x = -\int_0^1 \varphi'(x)\mathrm{d}x = \varphi(0) \tag{2.1}$$

Satz 2.6 (Poincaré-Ungleichung) *Für $u \in H_0^{1,2}(\Omega)$ gilt*

$$\|u\|_{L^2(\Omega)} \leq \left(\frac{|\Omega|}{\omega_d}\right)^{1/d} \|Du\|_{L^2(\Omega)}\,,$$

wobei $|\Omega|$ das (Lebesguesche) Maß von Ω und ω_d das Maß der Einheitskugel in $\mathbb{R}^d$ bezeichnet. Insbesondere ist für $u \in H_0^{1,2}(\Omega)$ die $W^{1,2}$-Norm durch die L^2-Norm von Du kontrolliert:

$$\|u\|_{W^{1,2}(\Omega)} \leq \left(1 + \left(\frac{|\Omega|}{\omega_d}\right)^{1/d}\right) \|Du\|_{L^2(\Omega)}\,.$$

Beweis später. □

2.3. Schwache Lösung der Poissongleichung

$-\Delta u = 0$ in Ω, $u = g$ auf $\partial\Omega$.

$$\kappa = \inf\left\{\int_\Omega |\nabla v|^2 : v \in H^{1,2}(\Omega), v - g \in H_0^{1,2}(\Omega)\right\}$$

$(u_n)_{n\in\mathbb{N}}$ Minimalfolge $\Rightarrow$ Cauchy-Folge bzgl. $\|\nabla\cdot\|$. Mittels der Poincaré-Ungleichung

$$\|u_n - u_m\|_{L^2(\Omega)} \quad \leq \quad \|Du_n - Du_m\|_{L^2(\Omega)}$$

sieht man, dass diese auch Cauchy-Folge bzgl. $\|\cdot\|_{W^{1,2}(\Omega)}$ ist. $\Rightarrow$ Existenz eines Minimums.

Die variationelle Formulierung unseres Problems ist dann: Suche $u \in H^{1,2}(\Omega)$, so dass

$$\int_\Omega \nabla u \cdot \nabla\varphi = 0 \quad \forall\varphi \in H_0^{1,2}(\Omega).$$

Definition 2.7 *Sei $u \in H^{1,2}(\Omega)$. u heißt* schwache Lösung *der Laplacegleichung, falls*

$$(\nabla u, \nabla\varphi) = 0 \quad \forall\varphi \in H_0^{1,2}(\Omega).$$

Definition 2.8 *Sei $f \in L^2(\Omega)$. $u \in H^{1,2}(\Omega)$ heißt* schwache Lösung *von*

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

falls

$$(\nabla u, \nabla\varphi) = (f, \varphi) \quad \forall\varphi \in H_0^{1,2}(\Omega).$$

Lemma 2.9 (Stabilitätslemma) *Es seien $u_{i=1,2}$ schwache Lösungen von $-\Delta u = f_i$ mit $u_2 - u_1 \in H^{1,2}(\Omega)$. Dann gilt*

$$\|u_2 - u_1\|_{H^{1,2}(\Omega)} \leq \text{Const}\,.\,\|f_2 - f_1\|_{L^2(\Omega)}\,.$$

Insbesondere ist eine solche schwache Lösung eindeutig bestimmt.

Beweis Betrachte

$$\begin{aligned}
\int_\Omega \nabla(u_2-u_1)\nabla\varphi &= \int (f_2-f_1)\varphi \quad \forall\varphi\in H^{1,2}(\Omega)\\
\Rightarrow \int_\Omega |\nabla(u_2-u_1)|^2 &= \int (f_2-f_1)(u_2-u_1)\\
\Leftrightarrow \|\nabla(u_1-u_2)\|_{L^2}^2 &\le \|f_2-f_1\|_{L^2}\,\|u_1-u_2\|_{L^2}\\
&\le \text{Const}\,.\,\|f_2-f_1\|_{L^2}\,\|\nabla(u_1-u_2)\|_{L^2}\,.\\
\Leftrightarrow \|\nabla(u_1-u_2)\|_{L^2} &\le \text{Const}\,.\,\|f_2-f_1\|_{L^2}
\end{aligned}$$

Mittels einer weiteren Anwendung der Poincaré-Ungleichung folgt die Behauptung. □

2.3.1. Bemerkungen zur Regularitätstheorie

Als Ergänzung von Abschnitt 2.2:

Definition 2.10 *$u\in L^2(\Omega)$ besitzt in $L^2(\Omega)$ die schwache Ableitung $v=\partial^\alpha u$, falls $v\in L^2(\Omega)$ und*

$$(\varphi,v)_{L^2(\Omega)} = (-1)^\alpha\,(\partial^\alpha\varphi,u)_{L^2(\Omega)} \quad \forall\varphi\in C_0^\infty(\Omega).$$

Hierbei ist $\alpha\in\mathbb{N}_0^d$ ein Multiindex (siehe Definition B.11).

Definition 2.11 *Für ganzzahliges $m\ge 0$ bezeichne $H^{m,2}(\Omega)$ die Menge aller Funktionen $u\in L^2(\Omega)$, die schwache Ableitungen $\partial^\alpha u$ für $|\alpha|\le m$ besitzen. Skalarprodukt und Norm in $H^{m,2}(\Omega)$ werden wie folgt festgelegt:*

$$\begin{aligned}
(u,v)_m &:= \sum_{|\alpha|\le m} (\partial^\alpha u,\partial^\alpha v)\,,\\
\|u\|_m &:= \sqrt{(u,u)_m}.
\end{aligned}$$

Satz 2.12 *$H^{m,2}(\Omega)$ ist vollständig bzgl. $\|\cdot\|_m$, also ein Hilbertraum.*

Definition 2.13 *Das Variationsproblem*

$$(\nabla u,\nabla\varphi)_0 = (f,\varphi)_0 \quad \forall\varphi\in H^{1,2}(\Omega)$$

heißt H^s-regulär ($s\ge 2$), wenn es zu jedem $f\in H^{s-2m}(\Omega)$ eine Lösung $u\in H^s(\Omega)$ gibt und mit einer Zahl $c=c(\Omega,s)$ gilt

$$\|u\|_s \le c\,\|f\|_{s-2}\,.$$

Satz 2.14

1. *Wenn Ω konvex ist, dann ist das Dirichlet-Problem H^2-regulär.*
2. *Sei $s\ge 2$. Wenn Ω einen C^s-Rand besitzt, ist das Dirichlet-Problem H^s-regulär.*

Beispiel. $\Omega=\{(x,y)\in\mathbb{R}^2 : x^2+y^2<1, x>0 \text{ oder } y<0\}$.

$w(z):=z^{2/3}$ ist analytisch in Ω. Der Imaginärteil $u(z):=\text{Im}(w(z))$ ist Lösung der Randwertaufgabe

$$\begin{aligned}
\Delta u &= 0 && \text{in } \Omega,\\
u(e^{i\varphi}) &= \sin\left(\tfrac{2}{3}\varphi\right) && 0\le\varphi\le\tfrac{3\pi}{2},\\
u &= 0 && \text{sonst auf } \partial\Omega.
\end{aligned}$$

Wegen $w'(z)=2/3z^{-1/3}$ sind nicht einmal die ersten Ableitungen von u für $z\to\infty$ beschränkt.

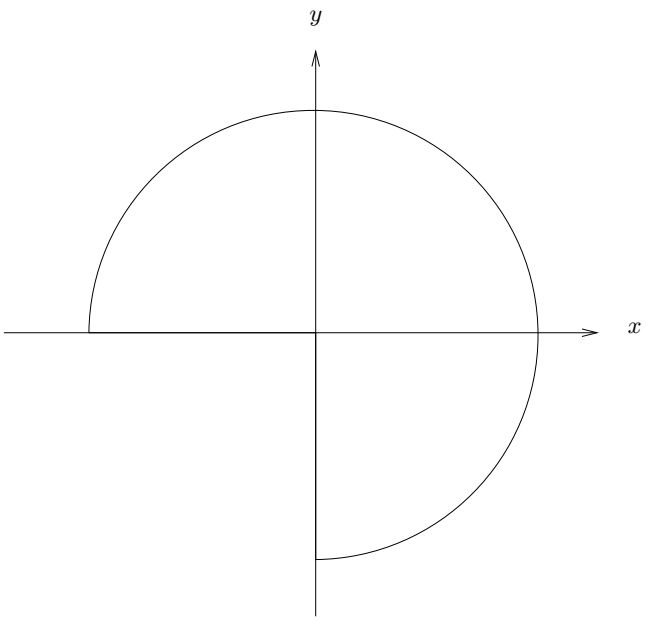


Abbildung 2.1.: Das Gebiet Ω.

2.4. Satz von Lax-Milgram

Definition 2.15 $(H, (\cdot,\cdot))$ *sei ein Hilbertraum mit zugehöriger Norm* $\|\cdot\|$, $A : H \times H \to \mathbb{R}$ *eine stetige, symmetrische Bilinearform. Hierbei bedeutet die* Stetigkeit, *dass*

$$A(u,v) \leq C \|u\| \|v\| \quad \forall u, v \in H$$

gilt. Die Symmetrie *bedeutet, dass für alle* $u, v \in H$

$$A(u,v) = A(v,u).$$

A *heißt* koerziv *bzw.* elliptisch, *falls*

$$A(v,v) \geq \lambda \|v\|^2 \quad \forall v \in H$$

mit einem $\lambda > 0$ *gilt und* stark koerziv *bzw.* stark elliptisch, *falls*

$$A(v,v) \geq \lambda \|v\|^2 \geq \lambda_0 \quad \forall v \in H$$

mit $\lambda, \lambda_0 > 0$ *gilt.*

Beispiel. Mit $H := H^{1,2}(\Omega)$ betrachte

$$A(u,v) := \int \nabla u \cdot \nabla v.$$

i. symmetrisch!

ii. Stetigkeit ($\rightarrow$Höldersche Ungleichung)

iii. Elliptizität ($\rightarrow$Hölder und Poincaré-Friedrichs)

Satz 2.16 (Lax-Milgram, Fassung für konvexe Mengen) *Sei $(H, (\cdot,\cdot))$ ein Hilbertraum mit Norm $\|\cdot\|$, $V \subset H$ konvex und abgeschlossen. $A : H \times H \to \mathbb{R}$ sei eine stetige, symmetrische und elliptische Bilinearform. Dann wird für jede stetige, lineare Abbildung $L : H \to \mathbb{R}$ das Funktional*

$$J(v) := \frac{1}{2}A(v,v) - L(v)$$

durch genau ein $u \in V$ minimiert.

Beweis J ist nach unten beschränkt, denn es ist

$$\begin{aligned} J(v) &\overset{A \text{ ellipt.}}{\geq} \frac{1}{2}\lambda \|v\|^2 - \|L\| \, \|v\| \\ &= \frac{1}{2\lambda} \left(\lambda \|v\| - \|L\|\right)^2 - \frac{\|L\|^2}{2\lambda} \\ &\geq -\frac{\|L\|^2}{2\lambda}. \end{aligned}$$

Setze $\kappa := \inf_{v \in V} J(v)$, und es sei nun $\{u_n\}_{n\in\mathbb{N}}$ eine Minimalfolge, also $\kappa = \lim_{n\to\infty} J(u_n)$. Dann ist

$$\begin{aligned} \lambda \|u_n - u_m\|^2 &\leq A\left(u_n - u_m, u_n - u_m\right) \\ &= 2A(u_n,u_n) + 2A(u_m,u_m) - A(u_n+u_m, u_n+u_m) \\ &= 4J(u_n) + 4J(u_m) - 8J\left(\frac{u_n+u_m}{2}\right) \\ &\leq 4J(u_n) + 4J(u_m) - 9\kappa, \end{aligned}$$

da V konvex und deshalb $\frac{u_n+u_m}{2} \in V$ ist. Dabei wurde in der zweiten Gleichung die Parallelogramm-Gleichung Satz B.8 mit dem Skalarprodukt $A(\cdot,\cdot)$ benutzt. Wegen der Minimalfolgeneigenschaft ist

$$\lim_{n\to\infty} J(u_n) = \lim_{m\to\infty} J(u_m) = \kappa$$

und es folgt

$$\|u_n - u_m\| \to 0 \quad \text{für} \quad n,m \to \infty.$$

Also ist $\{u_n\}_{n\in\mathbb{N}}$ eine Cauchy-Folge in H und es existiert $u = \lim_{n\to\infty} u_n$. Da V abgeschlossen ist, gilt auch $u \in V$. Aus der Stetigkeit von J folgt

$$J(u) = \lim_{n\to\infty} J(u_n) = \inf_{v\in V} J(v),$$

also existiert ein Minimum.

Die Lösung ist eindeutig: Seien u_1 und u_2 zwei Minima. Dann ist offensichtlich

$$u_1, u_2, u_1, u_2, \ldots$$

eine Minimalfolge. Wie oben gezeigt wurde, ist jede Minimalfolge eine Cauchyfolge, was nur für $u_1 = u_2$ möglich ist. □

Satz 2.17 (Lax-Milgram) *Sei $(H, (\cdot,\cdot))$ ein Hilbertraum, $V \subset H$ ein abgeschlossener, linearer Unterraum von H. Weiterhin seien für A und L die Voraussetzungen von Satz 2.16 erfüllt. Dann existiert genau ein $u \in V$, welches die Gleichung*

$$A(u,\varphi) = L(\varphi) \quad \forall \varphi \in V \tag{2.2}$$

löst.

Beweis Es wird zunächst gezeigt, dass u genau dann ist ein Minimum des Funktionals

$$J(v) := \frac{1}{2}A(v,v) - L(v)$$

in V ist, wenn $A(u,\varphi) = L(\varphi)$ für alle $\varphi \in V$ gilt.

Sei u ein Minimum des Funktionals J. Dann ist u insbesondere ein kritischer Punkt von J. Dass u ein kritischer Punkt ist, bedeutet, dass

$$\frac{\mathrm{d}}{\mathrm{d}t}J(u+t\varphi)|_{t=0} = 0 \quad \forall \varphi \in V,$$

also

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}A(u+t\varphi, u+t\varphi) - L(u+t\varphi)\right)|_{t=0} \\ &= A(u,\varphi) - L(\varphi) \end{aligned}$$

für alle $\varphi \in V$, also gilt (2.2). Gilt umgekehrt die Gleichung (2.2), so ist

$$J(u+t\varphi) = J(u) + t\underbrace{(A(u,\varphi) - L(\varphi))}_{=0} + t^2\underbrace{A(\varphi,\varphi)}_{\geq 0} \quad \geq \quad J(u)$$

für alle $\varphi \in V$, also ist u ein Minimum von J.

Da V als linearer Unterraum insbesondere konvex ist, folgt die Existenz und Eindeutigkeit einer Lösung von (2.2) aus Satz 2.16. □

Bemerkung.

1. Es ist wichtig, zu beachten, dass man (insbesondere bei Dirichlet-Randbedingungen) nur $H_0^{1,2}(\Omega)$-Funktionen als Testfunktionen verwendet – mit Betonung auf dem Index 0. Dies kommt daher, dass man *u nur* mit Funktionen stören darf, die die Randwerte nicht ändern!
2. Der Satz kann auch für den Spezialfall $V = H$ angewendet werden.

Lemma 2.18 (Céa-Lemma) *$A : H \times H \to \mathbb{R}$ sei eine stetige, symmetrische, elliptische Bilinearform, $L : H \to \mathbb{R}$ linear stetig. Wir betrachten die Probleme*

$$\begin{aligned} (1): \quad A(u,\varphi) + L(\varphi) &= 0 \quad \forall \varphi \in H, \\ (2): \quad A(u_V,\varphi) + L(\varphi) &= 0 \quad \forall \varphi \in V \subset H. \end{aligned}$$

Dann gilt

$$\|u - u_V\|_H \leq \frac{C}{\lambda}\inf_{v\in V}\|u - v\|_H,$$

wobei C und λ die Konstanten aus Definition 2.15 sind.

Bemerkung. V ist für unsere (Rechen-)Zwecke normalerweise ein endlichdimensionaler Unterraum von H. Das Céa-Lemma beschränkt dann den Fehler in unserer Lösung, falls die Bestapproximation an u in V bekannt ist (oder abgeschätzt werden kann (!)).

Beweis Subtrahieren der Gleichungen (1) und (2) ergibt die *Galerkin-Orthogonalität*:

$$A(u - u_V, \varphi) = 0 \quad \forall \varphi \in V,$$

die besagt, dass der Fehler senkrecht auf allen Testfunktionen steht. Wegen der Elliptizität gilt

$$\begin{aligned}\|u-u_V\|^2 &\leq \frac{1}{\lambda}A(u-u_V,u-u_V)\\ &= \frac{1}{\lambda}A(u-u_V,u-v)+\frac{1}{\lambda}\underbrace{A(u-u_V,\underbrace{v-u_V}_{\in V})}_{=0}\\ &= \frac{1}{\lambda}A(u-u_V,u-v)\leq\frac{C}{\lambda}\|u-u_V\|\,\|u-v\|\,.\end{aligned}$$

Die Behauptung folgt durch Division durch $\|u-u_V\|$. □

3. Die Finite-Elemente-Methode: ein 1D-Beispiel

Bevor im nächsten Kapitel die Finite-Elemente-Methode in voller Allgemeinheit eingeführt wird, sollen die fundamentalen Konzepte dieser Methode anhand eines eindimensionalen Beispiels erläutert werden.

$\Omega := [0,1]$. Suche $u \in H^{1,2}(\Omega)$, so dass $a(u,\varphi) = (f,\varphi)$ für alle $\varphi \in H^{1,2}(\Omega)$. Wähle nun $V_h = \text{span}\{\varphi_1, \ldots, \varphi_n\}$. Entwickle

$$u_h = \sum_{i=1}^{n} u_h^i \varphi_i \in V_h.$$

Dann

$$a\left(\sum_{i=1}^{n} u_h^i \varphi_i, \varphi\right) = (f,\varphi) \quad \forall \varphi \in V_h.$$

Als „Vertreter“ aller $\varphi \in V_h$ nimmt man die Basisfunktionen φ_i:

$$a\left(\sum_{i=1}^{n} u_h^i \varphi_i, \varphi_j\right) = (f,\varphi_j) \quad \forall j \in \{1,\ldots,n\}.$$

Dieses Vorgehen ist gerechtfertig, da aufgrund der Vektorraumstruktur von V_h jede Funktion in V_h als Linearkombination der φ_i $(i = 1,\ldots,n)$ darstellbar ist und sich folglich die obige Gleichung für die Basisfunktionen auf *alle* Funktionen in V_h überträgt (*Superpositionsprinzip*).

Dies führt auf ein lineares System $Au = b$, wobei $A = \{a_{i,j}\} \in \mathbb{R}^{n\times n}$ mit $a_{i,j} = a(\varphi_j, \varphi_i)$, $u = \{u_h^i\}$, $b_j = (f,\varphi_j)$.

Es ergeben sich *lediglich* die Einträge $a_{i,i}, a_{i,i+1}, a_{i,i-1} \neq 0$ in Zeile i von A. Diese Matrix ist somit dünn besetzt. Sei $\Omega := (a,b)$ mit $a < b$. $I_i := [a_i, b_i]$, so dass

$$\bar{\Omega} = \bigcup_{i=0}^{N} I_i$$

und $\overset{\circ}{I}_i \cap I_j = \emptyset$ für $i \neq j$.

$$a = x_0 < \cdots < x_N < x_{N+1} = b,$$

d.h. $a_i = x_i$, $b_i = x_{i+1}$ für $0 \leq i \leq N$. Die $\{x_i\}$ heißen die *Knoten* des Gitters. $h_i := x_{i+1} - x_i$ für $0 \leq i \leq N$ und $h := \max_{0\leq i\leq N} h_i$. Wir betrachten den Vektorraum

$$P_h^1 := \{v_h \in C^0(\bar{\Omega}) : \forall i \in \{0,\ldots,N\}, v_h|_{I_i} \in \mathbb{P}_1\}.$$

Für $i \in \{1,\ldots,N\}$ setzen wir

$$\varphi_i(x) := \begin{cases} \frac{1}{h_{i-1}}(x - x_{i-1}) & x \in I_{i-1}, \\ \frac{1}{h_i}(x_{i+1} - x) & x \in I_i, \\ 0 & \text{sonst} \end{cases} \in P_h^1,$$

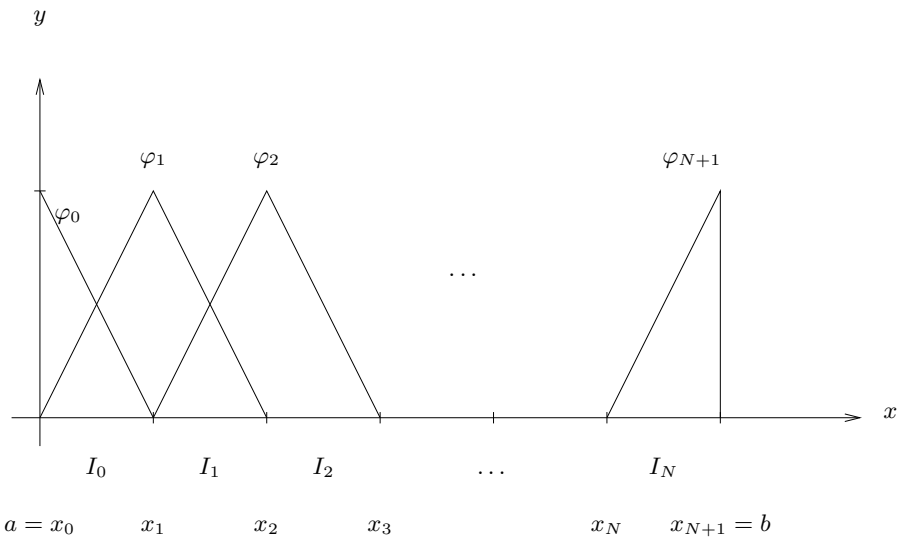


Abbildung 3.1.: Hutfunktionen ψ_i auf einer Zerlegung des Intervalls $[a, b]$.

sowie

$$\varphi_0(x) := \begin{cases} \frac{1}{h_0}(x_1 - x) & x \in I_0, \\ 0 & \text{sonst} \end{cases} \in P_h^1$$

und

$$\varphi_{N+1}(x) := \begin{cases} \frac{1}{h_N}(x - x_N) & x \in I_N, \\ 0 & \text{sonst} \end{cases} \in P_h^1.$$

Satz 3.1 *Die Menge* $\{\varphi_i\}_{0 \le i \le N+1}$ *bildet eine Basis von* P_h^1.

Beweis Offensichtlich gilt $\varphi_i(x_j) = \delta_{i,j}$. Seien $(\alpha_0, \alpha_1, \ldots, \alpha_{N+1})^\top \in \mathbb{R}^{N+2}$ die Koeffizienten zu einer Funktion $w \in P_h^1$ bezüglich der Basis $\{\varphi_i\}_i$. Dann gilt an den Knoten des Gitters

$$w(x_j) = \sum_{i=0}^{N+1} \alpha_i \varphi_i(x_j) = \sum_{i=0}^{N+1} \alpha_i \delta_{i,j} = \alpha_j = 0$$

Somit sind die $\{\varphi_i\}_{0 \le i \le N+1}$ linear unabhängig. Trivialerweise gilt

$$v_h \in P_h^1 \Rightarrow v_h = \sum_{i=0}^{N+1} v_h(x_i)\varphi_i.$$

□

Für $i \in \{0, \ldots, N+1\}$ seien die Funktionale

$$\begin{aligned} \gamma_i : C(\bar{\Omega}) &\to \mathbb{R} \\ v &\mapsto \gamma_i(v) := v(x_i) \in \mathbb{R} \end{aligned}$$

definiert. $\{\gamma_i\}_{0 \le i \le N+1}$ sind *globale Freiheitsgrade* in P_h^1. $\{\gamma_i\}_{0 \le i \le N+1}$ bildet eine Basis des Dualraums $(P_h^1)^*$. Suche

$$u_h = \sum_{i=0}^{N+1} u_h^i \varphi_i,$$

so dass

$$\begin{aligned} a(u_h, \varphi_j) &= (f, \varphi_j) \quad \text{für } j = \{0, \ldots, N+1\}, \\ \Leftrightarrow \sum_{i=0}^{N+1} u_h^i a(\varphi_i, \varphi_j) &= b_h \\ \Leftrightarrow A_h u_h &= b_h. \end{aligned}$$

A_h wird *Steifigkeitsmatrix* genannt.

Um den Fehler der Finite-Elemente-Lösung abschätzen zu können, legt das Céa-Lemma nahe, den Interpolationsfehler in V_h zu betrachten.

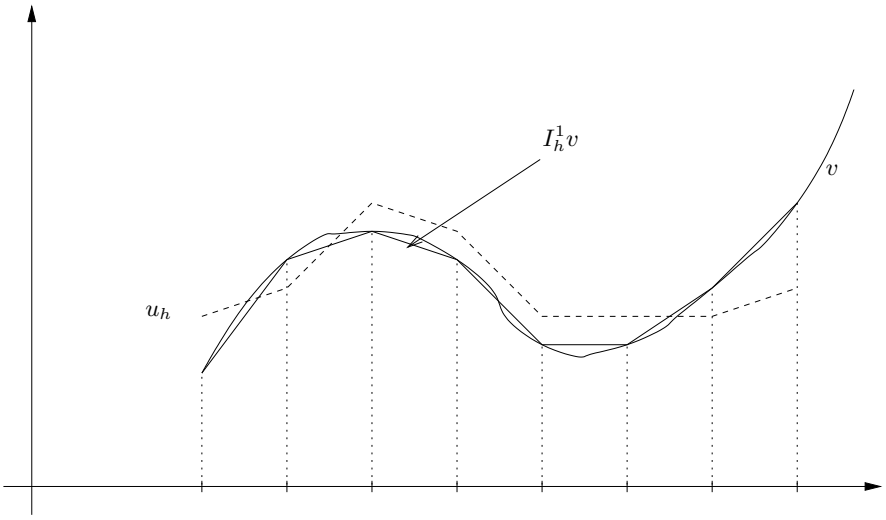


Abbildung 3.2.: Interpolierte $I_h^1 v$ und exakte Lösung u_h (schematisch).

Wir betrachten den Interpolationsoperator

$$\begin{aligned} I_h^1 : C^0(\bar{\Omega}) &\to P_h^1, \\ v &\mapsto \sum_{i=0}^{N+1} \gamma_i(v)\varphi_i \in P_h^1. \end{aligned}$$

Satz 3.2 *Für $v \in H^2(\Omega)$ gilt*

$$\begin{aligned} \left\| v - I_h^1 v \right\|_{0,\Omega} &\leq h^2 \left| v \right|_{2,\Omega} \quad \forall h > 0, \\ \left\| \nabla(v - I_h^1 v) \right\|_{0,\Omega} &\leq h \left| v \right|_{2,\Omega} \quad \forall h > 0. \end{aligned}$$

Bemerkung. Falls $v \in H^1(\Omega) \setminus H^2(\Omega)$,

$$\begin{aligned} \left\| v - I_h^1 v \right\|_{0,\Omega} &\leq h \left| v \right|_{1,\Omega} \quad \forall h > 0, \\ \lim_{h \to 0} \left| v - I_h^1 v \right|_{1,\Omega} &= 0. \end{aligned}$$

Wir erhalten die folgende *A-priori-Abschätzung*:

$$\left\| u - u_h \right\|_{1,\Omega} \leq \frac{C}{\lambda} \inf_{v \in P_h^1} \left\| u - v \right\|_{1,\Omega} \leq \frac{C}{\lambda} \left\| u - I_h^1 u \right\|_{1,\Omega} \leq \frac{C}{\lambda} h \left| u \right|_{1,\Omega}.$$

Beachte: Die Interpolierte $I_h^1 u$ (deren Benutzung uns das Céa-Lemma vorschlägt) hat *nichts* mit der berechneten Lösung u_h zu tun.

Die Basisfunktionen $\{\varphi_i\}_{0\le i\le N+1}$ lassen sich als Komposition von einer Funktion, die gitterunabhängig definiert wird, und einer geometrischen Transformation definieren. Seien $\hat{\kappa} = [0,1]$ das Referenzintervall und die affinen Transformationen $T_i : \hat{x} \in \hat{\kappa} \mapsto x = x_i + \hat{x}h_i$ für $0 \le i \le N$. Es seien die Funktionen

$$\begin{aligned} \hat{\varphi}_0(\hat{x}) &:= 1 - \hat{x} && \forall \hat{x} \in \hat{\kappa}, \\ \hat{\varphi}_1(\hat{x}) &:= \hat{x} && \forall \hat{x} \in \hat{\kappa} \end{aligned}$$

definiert. Diese Funktionen bilden eine Basis von $P_1(\hat{\kappa})$. Dann gilt

$$\begin{aligned} \varphi_i(x) &= \begin{cases} (\hat{\varphi}_1 \circ T_{i-1}^{-1})(x) & x \in [x_{i-1}, x_i], \\ (\hat{\varphi}_0 \circ T_i^{-1})(x) & x \in [x_i, x_{i+1}]. \end{cases} \qquad (i = 1, \ldots, N), \\ \varphi_0(x) &= (\hat{\varphi}_0 \circ T_0^{-1})(x) \quad x \in [x_0, x_1] \ , \\ \varphi_{N+1}(x) &= (\hat{\varphi}_1 \circ T_N^{-1})(x) \quad x \in [x_N, x_{N+1}] \ . \end{aligned}$$

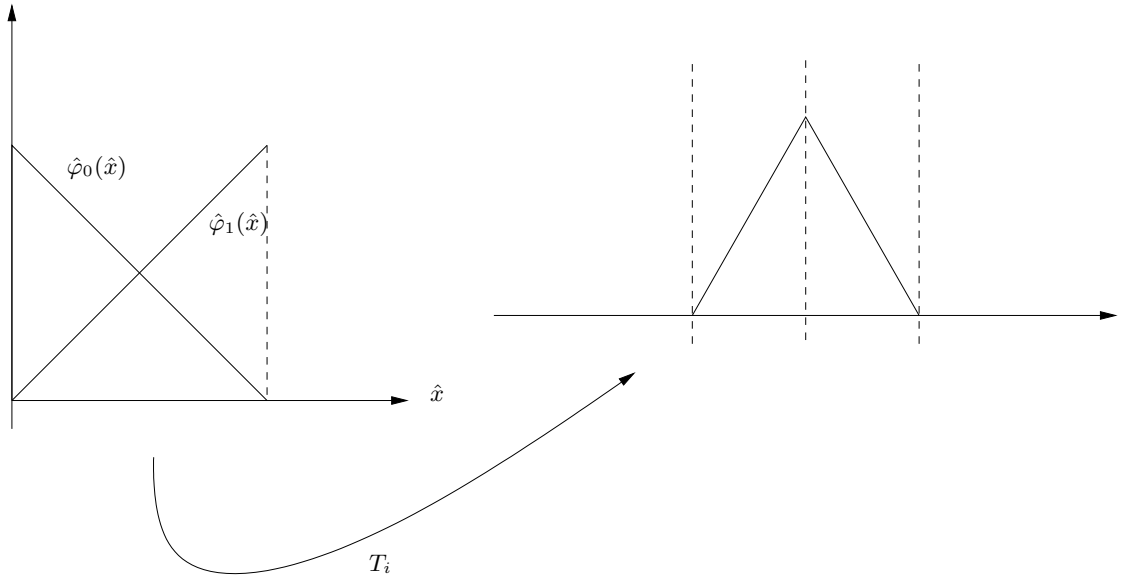


Abbildung 3.3.: Abbildung von lokalen Formfunktionen auf globale mit Hilfe der geometrischen Transformation T_i.

Definition 3.3 (Freiheitsgrad, Formfunktion) *Es seien $\{\gamma_j\}_{0\le j\le N+1} \in (V_h^1)^*$ die linearen Funkionale, so dass*

$$\gamma_j(v_h) = v_h(x_j) \quad \forall v_h \in V_h^1$$

In Anlehnung an die Mechanik werden diese Funktionale als (globale) Freiheitsgrade *bezeichnet. Die Funktionen $\{\varphi_i\}_{0\le i\le N+1}$, die so definiert sind, dass*

$$\gamma_j(\varphi_i) = \delta_{i,j} \quad (0 \le i, j \le N+1, \varphi_i \in V_h^1)$$

gilt, werden (globale) Formfunktionen *(engl.* shape functions*) genannt. Entsprechend werden die Funktionen $\{\hat{\varphi}_i\}_{i=0,1}$* (lokale) Formfunktionen *genannt.*

Eine mögliche Erweiterung der obigen Methode mit *linearen* Elementen ist es, Polynome höheren Grades zu betrachten

$$P_h^k := \{v_h \in C^0(\bar{\Omega}) : \forall i \in \{0, \ldots, N\}, v_h|_{I_i} \in \mathbb{P}_k\}.$$

Satz 3.4 $0 \leq l \leq k$. *Dann gilt für* $v \in H^{l+1}(\Omega)$

$$\left\| v - I_h^k v \right\|_{0,\Omega} + h \left| v - I_h v \right|_{1,\Omega} \leq C h^{l+1} \left| v \right|_{l+1,\Omega}.$$

Für die lokalen Formfunktionen werden üblicherweise die Lagrangeschen Basispolynome herangezogen:

$$\hat{\varphi}_i^k(\hat{x}) = \frac{\prod_{j=0, j\neq i}^{k} (\hat{x} - \hat{x}_j)}{\prod_{j=0, j\neq i}^{k} (\hat{x}_i - \hat{x}_j)},$$

wobei $\hat{x}_j = j/k$ für $0 \leq j \leq k$.

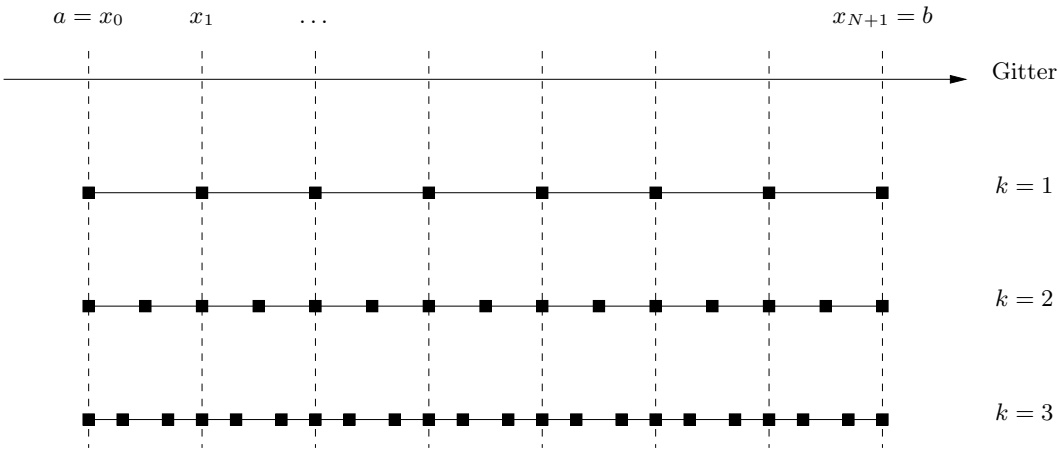


Abbildung 3.4.: Gitterpunkte einer Diskretisierung höherer Ordnung in einer Dimension.

$x_{i,j} := x_i + (j/k)h_i = x_i + \hat{x}_j h_i$ für $0 \leq i \leq N$ und $0 \leq j \leq k-1$, außerdem $x_{N+1,0} = x_0$. Es gilt

$$\dim(V_h^k) = k(N+1) + 1.$$

Man definiert

$$\varphi_{i,0}(x) := \begin{cases} \hat{\varphi}_k^k \circ T_{i-1}^{-1}(x) & x \in [x_{i-1}, x_i], \\ \hat{\varphi}_0^k \circ T_i^{-1}(x) & x \in [x_i, x_{i+1}], \\ 0 & \text{sonst.} \end{cases}$$

und

$$\varphi_{i,j}(x) := \begin{cases} \hat{\varphi}_j^k \circ T_i^{-1}(x) & x \in [x_i, x_{i+1}], \\ 0 & \text{sonst.} \end{cases}$$

für $1 \leq j \leq k-1$ und $1 \leq i \leq N$. Die Definitionen für die Basisfunktionen zu x_0 und x_{N+1} gehen analog.

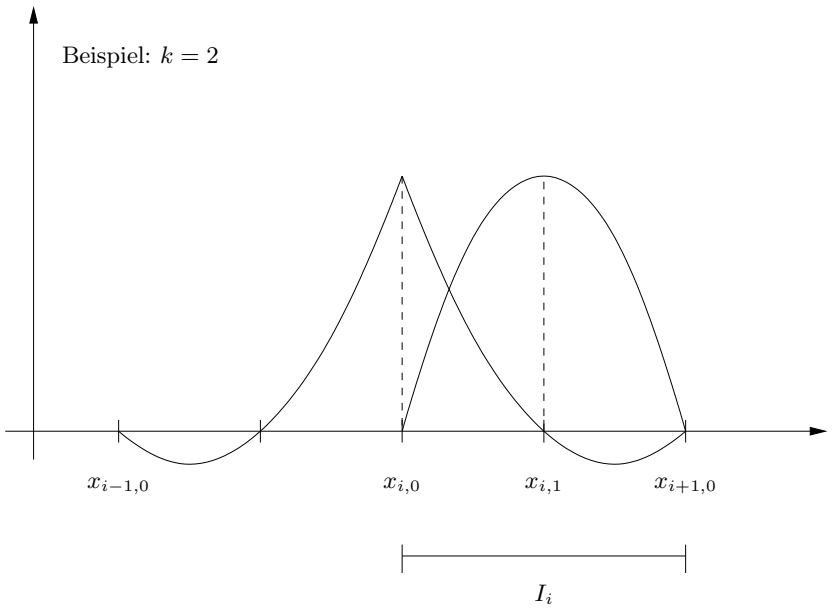


Abbildung 3.5.: Formfunktionen zweiter Ordnung in 1D.

4. Interpolation mit Finiten Elementen

Die Finite-Elemente-Methode baut auf dem schwachen Lösungsbegriff aus Kapitel 2 auf. Das Kernziel dabei ist es, geeignete *endlichdimensionale Unterräume* der beteiligten Funktionenräume, beispielsweise H^1 oder L^2, zu konstruieren.

Abschnitt 4.1 befasst sich mit der allgemeinen Definition von Finiten Elementen. Weiterhin werden einige wichtige Typen von Finiten Elementen vorgestellt. Die übrigen Abschnitte dieses Kapitels beschäftigen sich mit der Frage nach der Approximationsgüte von Finiten Elementen. Dabei wird im wesentlichen der Fehler zwischen der kontinuierlichen (schwachen) Lösung und der diskreten Finite-Elemente-Lösung in verschiedenen Normen — beispielsweise H^m- oder L^2-Norm — abgeschätzt. Die Güte der Approxmation wird dann nach Konvergenzordnung bezüglich des wesentlichen Diskretisierungsparameters h, welcher die Feinheit des diskreten Gitters beschreibt, beurteilt.

4.1. Definition von Finiten Elementen

Definition 4.1 *Ein* finites Element *ist ein Tripel* $(\hat{K}, \hat{\Pi}, \hat{\Sigma})$ *mit den folgenden Eigenschaften:*

i. $\hat{K}$ *ist eine nicht leere, kompakte und einfach zusammenhängende Teilmenge von* $\mathbb{R}^d$*,*

ii. $\hat{\Pi}$ *ist ein endlichdimensionaler Vektorraum von Funktionen, die auf* $\hat{K}$ *definiert sind,*

iii. $\hat{\Sigma}$ *ist eine Menge von* s *linear unabhängigen Funktionalen* $\{\hat{\sigma}_i\}_{1\leq i\leq s}$*, so dass die Abbildung*

$$\hat{p} \in \hat{\Pi} \mapsto (\hat{\sigma}_1(\hat{p}), \ldots, \hat{\sigma}_s(\hat{p}))^\top \in \mathbb{R}^s$$

ein Isomorphismus ist. Die Funktionale $\{\hat{\sigma}_i\}_{1\leq i\leq s}$ *werden* lokale Freiheitsgrade *genannt.*

Bemerkung.

1. Für alle $(\alpha_1, \ldots, \alpha_s)^\top \in \mathbb{R}^s$ existiert genau ein $\hat{p} \in \hat{\Pi}$, so dass $\hat{\sigma}_i(\hat{p}) = \alpha_i$ $(1 \leq i \leq s)$. Diese Eigenschaft wird *Unisolvenz* genannt.

2. Aus *iii.* folgt weiterhin, dass $\dim\left(\hat{\Pi}\right) = s$.

Für die folgenden Konstruktionen von Finiten Elementen verwenden wir die Polynom-Vektorräume $\mathbb{P}_k$ und $\mathbb{Q}_k$ in $\mathbb{R}^d$. $\mathbb{P}_k$ ist der Raum der Polynome in den Variablen $(x_1, \ldots, x_d)$ mit reellen

Koeffizienten und deren globaler Grad kleiner oder gleich k ist:

$$\begin{aligned}
\mathbb{P}_k &= \left\{ p(x) = \sum_{0\le i\le k} \alpha_i x^i : \ \alpha_i \in \mathbb{R} \right\}, \text{falls } d = 1. \\
\mathbb{P}_k &= \left\{ p(x_1,x_2) = \sum_{0\le i+j\le k} \alpha_{i,j} x_1^i x_2^j : \ \alpha_{i,j} \in \mathbb{R} \right\}, \text{falls } d = 2. \\
\mathbb{P}_k &= \left\{ p(x_1,x_2,x_3) = \sum_{0\le i+j+l\le k} \alpha_{i,j,l} x_1^i x_2^j x_3^l : \ \alpha_{i,j,l} \in \mathbb{R} \right\}, \text{falls } d = 3, \\
\mathbb{P}_k &= \left\{ p(x) = \sum_{|\beta|\le k} \alpha_\beta x^\beta : \ \alpha_\beta \in \mathbb{R},\ \beta \in \mathbb{N}_0^d,\ x \in \mathbb{R}^d \right\}.
\end{aligned}$$

Man kann leicht zeigen, dass

$$\dim \mathbb{P}_k = \begin{pmatrix} d+k \\ k \end{pmatrix} = \begin{cases} k+1 & d=1 \\ \frac{1}{2}(k+1)(k+2) & d=2 \\ \frac{1}{6}(k+1)(k+2)(k+3) & d=3 \end{cases}$$

Wenn beim Aufbau der Finiten Elemente alle Polynome vom Grad kleiner gleich k vorkommen, spricht man von Finiten Elementen mit vollständigen Polynomen.

$\mathbb{Q}_k$ ist der Raum der Polynome in den Variablen $(x_1, x_2, \ldots, x_d)$ mit reellen Koeffizienten, deren Grad bezüglich der *einzelnen Variablen* kleiner oder gleich k ist.

$$\begin{aligned}
\mathbb{Q}_k &= \mathbb{P}_k, \text{ falls } d = 1. \\
\mathbb{Q}_k &= \left\{ p(x_1,x_2) = \sum_{0\le i,j\le k} \alpha_{i,j} x_1^i x_2^j : \ \alpha_{i,j} \in \mathbb{R} \right\}, \text{ falls } d = 2. \\
\mathbb{Q}_k &= \left\{ p(x_1,x_2,x_3) = \sum_{0\le i,j,l\le k} \alpha_{i,j,l} x_1^i x_2^j x_3^l : \ \alpha_{i,j,l} \in \mathbb{R} \right\}, \text{ falls } d = 3.
\end{aligned}$$

Man kann leicht zeigen, dass

$$\dim(\mathbb{Q}_k) = (k+1)^d$$

ist. Weiterhin gilt $\mathbb{P}_k \subset \mathbb{Q}_k \subset \mathbb{P}_{kd}$.

4.1.1. Lagrangesche Finite Elemente auf Simplizes

Definition 4.2 *Ein* Simplex $S \subset \mathbb{R}^d$ *ist die konvexe Hülle von* $d+1$ *Punkten* $\{a_i\}_{0\le i\le d}$*, die nicht alle in einer Hyperebene liegen:*

$$S := \left\{ \mathbb{R}^d \ni x := \sum_{j=0}^{d} \lambda_j a_j : \ \sum_{j=0}^{d} \lambda_j = 1, \lambda_j \ge 0 \ (j = 0, \ldots, d) \right\}.$$

Das *Einheitssimplex* in $\mathbb{R}^d$ ist die Menge

$$\hat{S} = \left\{ \hat{x} \in \mathbb{R}^d : \hat{x}_i \ge 0, 1 \le i \le d, \sum_{i=1}^{d} \hat{x}_i \le 1 \right\}.$$

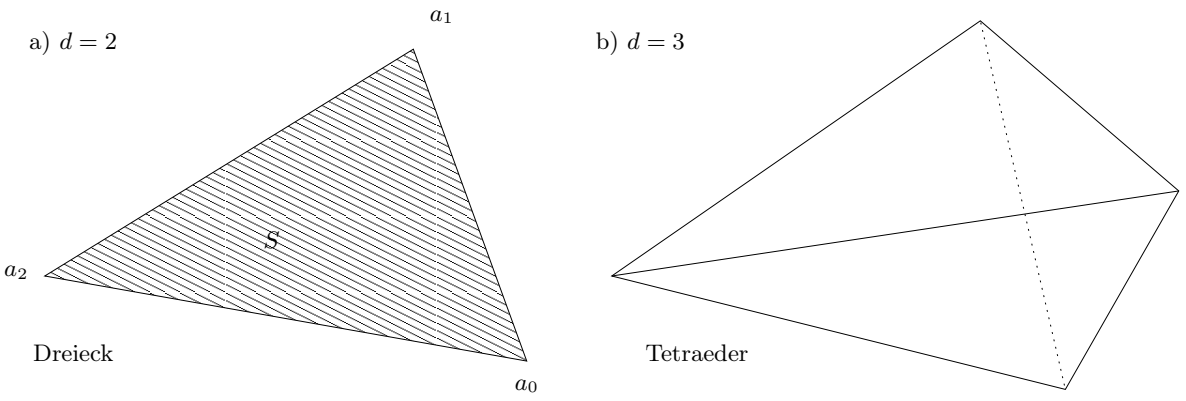


Abbildung 4.1.: Beispiele von Simplizes in verschiedenen Dimensionen.

Ein Simplex kann eindeutig definiert werden als die Abbildung des Einheitssimplex mittels einer bijektiven affinen Transformation, d.h.

$$x = T\hat{x} + b, \quad T \in \mathbb{R}^{d\times d}, \quad \det T \neq 0, \quad b \in \mathbb{R}^d.$$

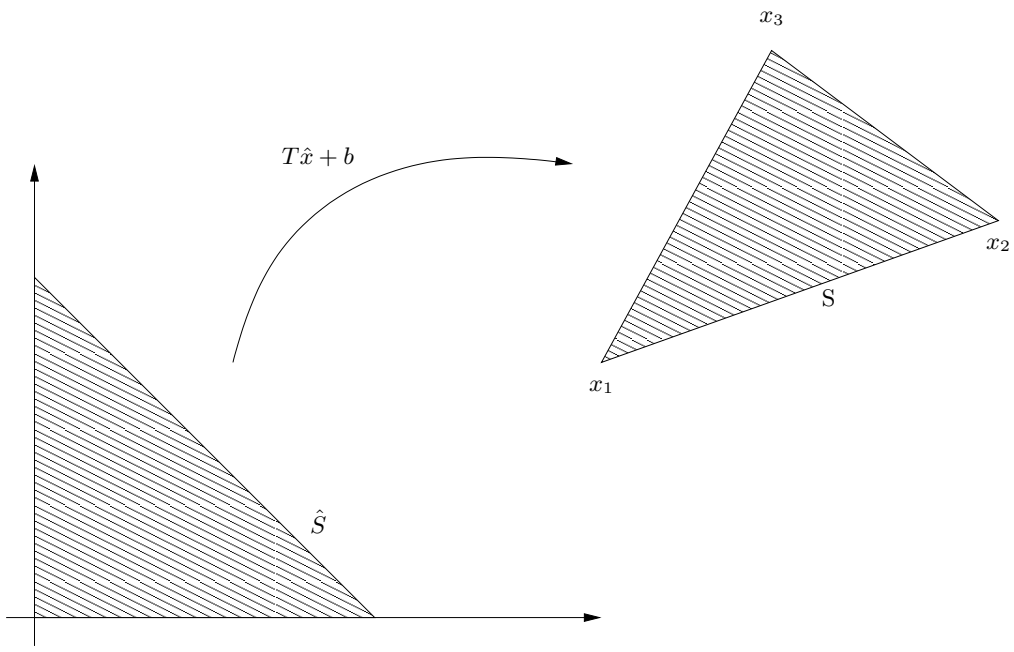


Abbildung 4.2.: Eine affine Transformation des Einheitssimplex in zwei Dimensionen.

Die Darstellung des Gebietes S als Konvexkombination von Eckpunkten $\{a_i\}_{0\leq i\leq d}$ liefert gleichzeitig eine Parametrisierung des Simplex S mittels der Koordinaten $\{\lambda_j\}_{0\leq j\leq d}$ in

$$x = \sum_{j=0}^{d} \lambda_j a_j.$$

Diese werden als *baryzentrische Koordinaten* bezeichnet. Gemeinsame Kanten benachbarter Dreiecke besizten in den baryzentrischen Koordinaten die gleiche Darstellung.

Satz 4.3 *Es sei K ein Simplex in $\mathbb{R}^d$ und $n = \dim \mathbb{P}_k$. Man bezeichne mit $\{x_i\}_{1\le i\le n}$ die Knoten, die die folgenden baryzentrischen Koordinaten besitzen:*

$$\left(\frac{i_0}{k}, \ldots, \frac{i_d}{k}\right), \quad 0 \le i_0, \ldots, i_d \le k, \quad i_0 + i_1 + \cdots + i_d = k.$$

Sei $\Sigma = \{\sigma_1, \ldots, \sigma_n\}$, die Menge der linearen Funktionale, so definiert, dass $\sigma_i(p) = p(x_i)$ $(1 \le i \le n)$ für $p \in \mathbb{P}_k$. Dann ist $(K, \mathbb{P}_k, \Sigma)$ ein (Lagrangesches) Finites Element.

Beweis

1. Zuerst zeigen wir, dass die Kardinalität der Menge $\Sigma_k := \{x_i\}_{1\le i\le n}$ gleich der Dimension des Vektorraums $\mathbb{P}_k$ ist. Durch die Normierungsbedingung

   $$i_0 + i_1 + \cdots + i_d = k$$

   können wir

   $$\frac{i_0}{k} = 1 - \sum_{j=1}^{d} \frac{i_j}{k}$$

   folgern. Somit sind für jedes $x_i \in \Sigma_k$ d Koeffizienten i_j zu bestimmen. Dies ist äquivalent zu der Anzahl der Multiindizes $\alpha = (\alpha_1, \ldots, \alpha_d)^\top$ mit den Eigenschaften $\alpha_i \ge 0$ $(i = 1, \ldots, d)$ und $\sum_{i=1}^{d} \alpha_i \le k$, was wiederum äquivalent zu den Bedingungen an die x^α in der Definition von $\mathbb{P}_k$ ist. Folglich ist $\dim \mathbb{P}_k = \operatorname{card}(\Sigma_k)$.

2. Die Abbildungen $\sigma_i : \mathbb{P}_k \to \mathbb{R}$ sind linear, da

   $$\sigma_i(p_1 + p_2) = p_1(x_i) + p_2(x_i) = \sigma_i(p_1) + \sigma_i(p_2)$$

   für $p_1, p_2 \in \mathbb{P}_k$. Daher genügt es zu zeigen, dass die Abbildung $\sigma := (\sigma_1, \ldots, \sigma_n) : \mathbb{P}_k \to \mathbb{R}^n$ injektiv ist, wegen der Gleichheit der Dimension von $\mathbb{P}_k$ und Kardinalität von Σ_k sowie der Linearität folgt die Bijektivität.

3. Wir zeigen durch Induktion nach d, dass aus $p \in \mathbb{P}_k$ mit $p(x) = 0$ für alle $x \in \Sigma_k$ folgt, dass $p = 0$ auf ganz $\mathbb{R}^d$.

   Ein Polynom, definiert auf $\mathbb{R}$, vom Grad k ist identisch 0, falls es $k+1$ Nullstellen besitzt. (Induktionsanfang)

   Angenommen, dies ist auch für ein Polynom auf $\mathbb{R}^d$ wahr (Induktionshypothese).

4. Den Induktionsschritt führen wir mit Induktion nach k.

   Für $k = 1$: Eine affin lineare Funktion, die an allen Ecken des nicht-degenerierten Simplex K verschwindet, ist identisch 0 (Induktionsanfang). Angenommen, dies ist für alle Polynome vom Grad $k-1$ richtig (Induktionshypothese). Sei $p \in \mathbb{P}_k$ mit $p = 0$ auf Σ_k. Σ_k enthält die Teilmenge

   $$\Sigma_k' = \{x \in \Sigma_k : \ i_0(x) = 0\},$$

   was Σ_k auf dem $(d-1)$-dimensionalen Simplex mit den Ecken $(a_1, \ldots, a_d)$ entspricht. Die Einschränkung von p auf die Hyperebene, welche durch die Punkte $(a_1, \ldots, a_d)$ definiert ist, ist ein Polynom vom Grad k in $(d-1)$ Variablen. Somit folgt aus der Induktionshypothese (für die Induktion nach d), dass $p = 0$ auf dieser Hyperebene.

   Wir drücken p nun in den Variablen $(x_1, \ldots, x_d)$ aus, so dass die Hyperebene durch $x_d = 0$ charakterisiert ist. Dann gilt

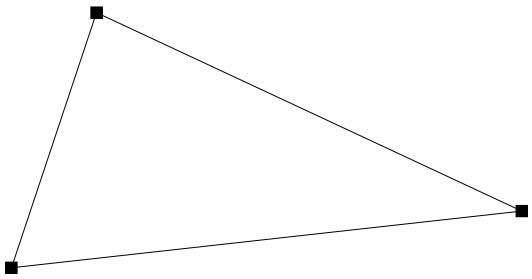
   $$p(x_1, \ldots, x_d) = x_d q(x_1, \ldots, x_{d-1})$$

mit einem Polynom q vom Grad $k-1$, das auf $\Sigma_k - \Sigma_k'$ identisch 0 sein muss, da $x_d \neq 0$ auf dieser Menge. $\Sigma_k - \Sigma_k'$ entspricht dem Gitter Σ_{k-1} für q und aus der Induktionshypothese (für die Induktion nach k) folgt $q = 0$ auf $\mathbb{R}^d$, woraus $p = 0$ auf $\mathbb{R}^d$ folgt.

□

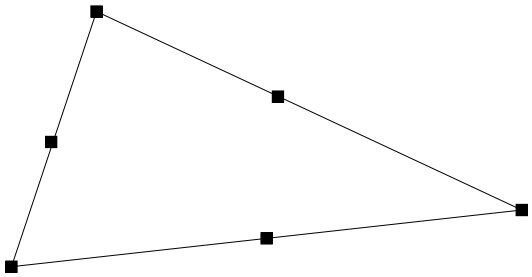
Bemerkung. Die Menge Σ_k wird auch *principal lattice* genannt.

■ ← Freiheitsgrad/Vorgabe des Funktionswertes



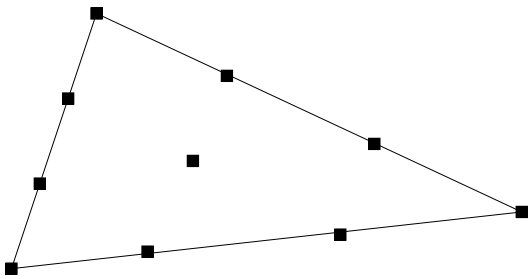
$\hat{\Pi} = \mathbb{P}_1$

Lineares Dreieckselement



$\hat{\Pi} = \mathbb{P}_2$

Quadratisches Dreieckselement



$\hat{\Pi} = \mathbb{P}_3$

Kubisches Dreieckselement

Abbildung 4.3.: Dreieckselemente.

Das Bedeutsame an den simplizialen Elementen ist, dass sie sich für $\mathbb{R}^d$ formulieren lassen.

4.1.2. Argyris-Elemente

Beim Argyris-Element wird $\hat{\Pi} = \mathbb{P}_5$ als zugrundeliegender Polynomraum verwendet. An jeder der drei Ecken hat man $1 + 2 + 3 = 6$ Freiheitsgrade (einen für die Funktion selbst, zwei

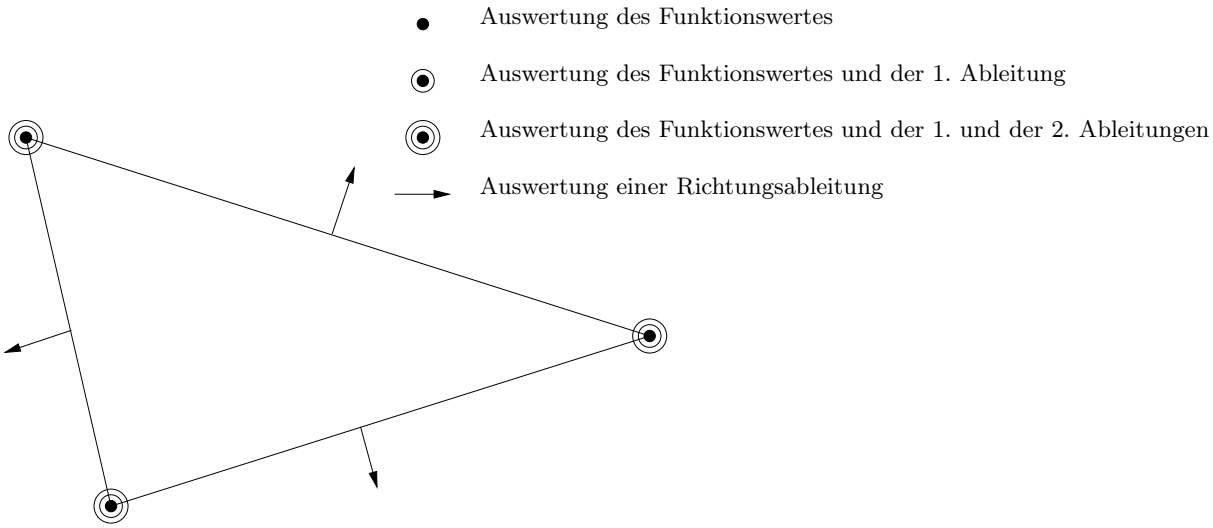


Abbildung 4.4.: Das Argyris-Element.

für die ersten partiellen Ableitungen nach x und y sowie drei für die partiellen Ableitungen zweiter Ordnung), zusätzlich die 3 Richtungsableitungen mit je 1 Freiheitsgrad, also insgesamt 21 Freiheitsgrade, was mit der Dimension des $\mathbb{P}_5$ übereinstimmt.

Lemma 4.4 *Sei $c_1 \in \mathbb{R}$ und ein p d-variates Polynom vom Grad $n \geq 1$ und $p(c_1, \cdot) \equiv 0$. Dann ist p von der Form $p(x) = (x_1 - c_1)\tilde{p}(x)$ mit einem d-variaten Polynom $\tilde{p}$ vom Grad $n-1$.*

Beweis Wir schreiben p in der Form

$$p(x_1, \tilde{x}) = \sum_{j=0}^{n} (x_1 - c_1)^j Q_{n-j}(\tilde{x})$$

mit $\tilde{x} = (x_2, \ldots, x_d)$ und $(d-1)$-variaten Polynomen Q_{n-j} vom Grad $\leq n-j$. Auf Grund der Voraussetzung $p(c_1, \cdot) = 0$ gilt $Q_n \equiv 0$. Daraus folgt die Behauptung mit

$$\tilde{p}(x_1, \tilde{x}) = \sum_{j=1}^{n} (x_1 - c_1)^{j-1} Q_{n-j}(\tilde{x}).$$

□

Lemma 4.5

i. $\mathbb{P}_5$ ist unisolvent bezüglich der Menge Σ der Knotenfunktionale des Argyris-Elements.

ii. Für eine Triangulierung des Gebietes Ω mit Argyris-Elementen gilt $V_h \subset C^1(\Omega)$.

Beweis $\dim \mathbb{P}_5 = 21$. Sei q_5 ein Polynom vom Grad ≤ 5 mit $\sigma_i(q_5) = 0$ für alle $i \in \{1, \ldots, 21\}$. Dann ist die Beschränkung von q_5 auf die Kante $[x_1, x_2]$ ein Polynom vom Grad ≤ 5 mit dreifachen Nullstellen bei x_1 und x_2, und somit identisch 0. Daher folgt aus Lemma 4.4, dass

$$q_5\left(\sum_{j=0}^{2} \lambda_j x_j\right) = \lambda_0 q_4\left(\sum_{j=0}^{2} \lambda_j x_j\right)$$

mit $q_4 \in \mathbb{P}_4$, da $\lambda_0 = 0$ auf $[x_1, x_2]$. Durch entsprechende Argumentation für die anderen beiden Kanten erhält man

$$q_5\left(\sum_{j=0}^{2}\lambda_j x_j\right) = \lambda_0\lambda_1\lambda_2 q_2\left(\sum_{j=0}^{2}\lambda_j x_j\right).$$

Wir führen die Darstellung

$$\begin{aligned} q_2^{(0)}(\lambda_1,\lambda_2) &:= q_2((1-\lambda_1-\lambda_2)x_0 + \lambda_1 x_1 + \lambda_2 x_2) \\ q_5^{(0)}(\lambda_1,\lambda_2) &:= \underbrace{(1-\lambda_1-\lambda_2)}_{=\lambda_0}\lambda_1\lambda_2 q_2^{(0)}(\lambda_1,\lambda_2) \end{aligned}$$

von q_2 und q_5 in den baryzentrischen Koordinaten λ_1 und λ_2 ein. Dann

$$\frac{\partial q_5^{(0)}}{\partial\lambda_1}(\lambda_1,\lambda_2) = \lambda_1\lambda_2\left\{(1-\lambda_1-\lambda_2)\frac{\partial q_2^{(0)}}{\partial\lambda_1} - q_2^{(0)}(\lambda_1,\lambda_2)\right\} + (1-\lambda_1-\lambda_2)\lambda_2 q_2^{(0)}(\lambda_1,\lambda_2)$$

und folglich

$$\begin{aligned} \frac{\partial^2 q_5^{(0)}}{\partial\lambda_1\partial\lambda_2} &= \lambda_1\left\{(1-\lambda_1-\lambda_2)\partial_{\lambda_1}q_2^{(0)} - q_2^{(0)}\right\} + \lambda_1\lambda_2\left\{(1-\lambda_1-\lambda_2)\partial_{\lambda_1}\partial_{\lambda_2}q_2^{(0)} - \partial_{\lambda_1}q_2^{(0)} - \partial_{\lambda_2}q_2^{(0)}\right\} \\ &\quad + (1-\lambda_1-\lambda_2)q_2^{(0)} + (1-\lambda_1-\lambda_2)\lambda_2 q_2^{(0)} - \lambda_2 q_2^{(0)}. \end{aligned}$$

Also gilt aufgrund von $\sigma_i(q_5) = 0$

$$0 = \frac{\partial^2 q_5^{(0)}}{\partial\lambda_1\partial\lambda_2} = q_2^{(0)}(0,0) = q_2^{(0)}(x_0).$$

Entsprechend gilt auch $q_2^{(0)}(x_1) = q_2^{(0)}(x_2) = 0$. Die Normalableitung von q_5 im Mittelpunkt $x_{011} = (x_1+x_2)/2$ verschwindet ebenfalls wegen $\sigma_i(q_5) = 0$. Daraus folgt $q_2(x_{011}) = 0$. Entsprechend gilt auch $q_2(x_{101}) = q_2(x_{110}) = 0$. Somit folgt die Unisolvenz von $\mathbb{P}_2$, da wir 6 Nullstellen für q_2, welches 6 Freiheitsgrade besitzt, gefunden haben. $\Rightarrow q_2 \equiv 0 \Rightarrow q_5 \equiv 0$. □

4.1.3. Crouzeix-Raviart-Elemente

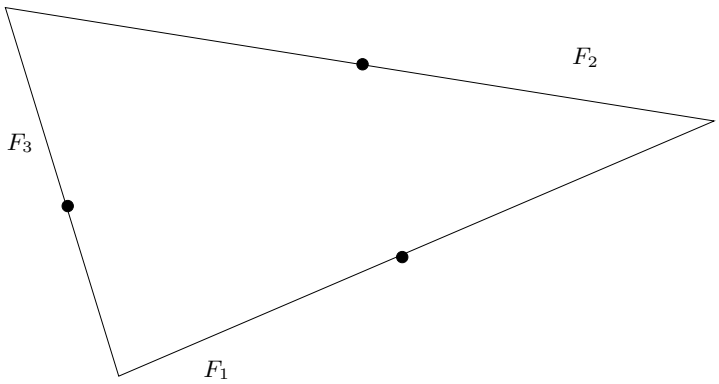


Abbildung 4.5.: Crouzeix-Raviart-Element.

Man nimmt $\hat{\Pi} = \mathbb{P}_1$ und definiert

$$\sigma_i(p) = \frac{1}{|F_i|}\int_{F_i} p_i \quad \text{für } i = 1,2,3,$$

wobei F_i die Kanten des Dreiecks bezeichnen. In V_h gilt praktischerweise

$$\sigma_i(p) = p(x_i).$$

4.1.4. Raviart-Thomas-Elemente

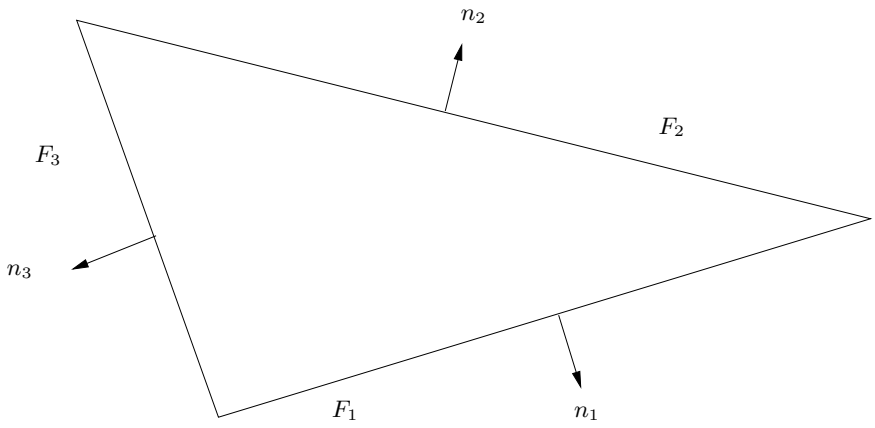


Abbildung 4.6.: Raviart-Thomas-Element.

Man nimmt $\hat{\Pi} = [\mathbb{P}_0]^2 \oplus \mathbb{P}_0$. Die Freiheitsgrade sind

$$\sigma_i(p) = \int_{F_i} p \cdot n_i.$$

Raviart-Thomas-Elemente eignen sich besonders zur Diskretisierung von Erhaltungsgesetzen (s. Kapitel 1.) Diese Elemente sind *konservativ*.

4.2. Transformationsformel und H^m-Fehlerabschätzung

Lemma 4.6 (Transformationsformel) *Seien $\hat{K}$, K abgeschlossen und sei*

$$\begin{aligned} \psi : \hat{K} &\to K \\ \hat{x} &\mapsto x_0 + A\hat{x} \end{aligned}$$

eine bijektive affine Abbildung mit einem Vektor $x_0 \in \mathbb{R}^d$ und einer regulären Matrix $A \in \mathbb{R}^{d\times d}$. Dann liegt $v = u \circ \psi$ in $H^k(\hat{K})$ für alle $u \in H^k(K)$, und es gilt

$$|v|_{H^k(\hat{K})} \leq \|A\|_2^k \, |\det A|^{-1/2} \, |u|_{H^k(K)} .$$

Bemerkung.

- affine Abbildung ψ: keine Einschränkung in $\mathbb{P}_1$
- in $\mathbb{P}_2$ Verbesserung möglich, aber nicht mit ψ realisierbar.

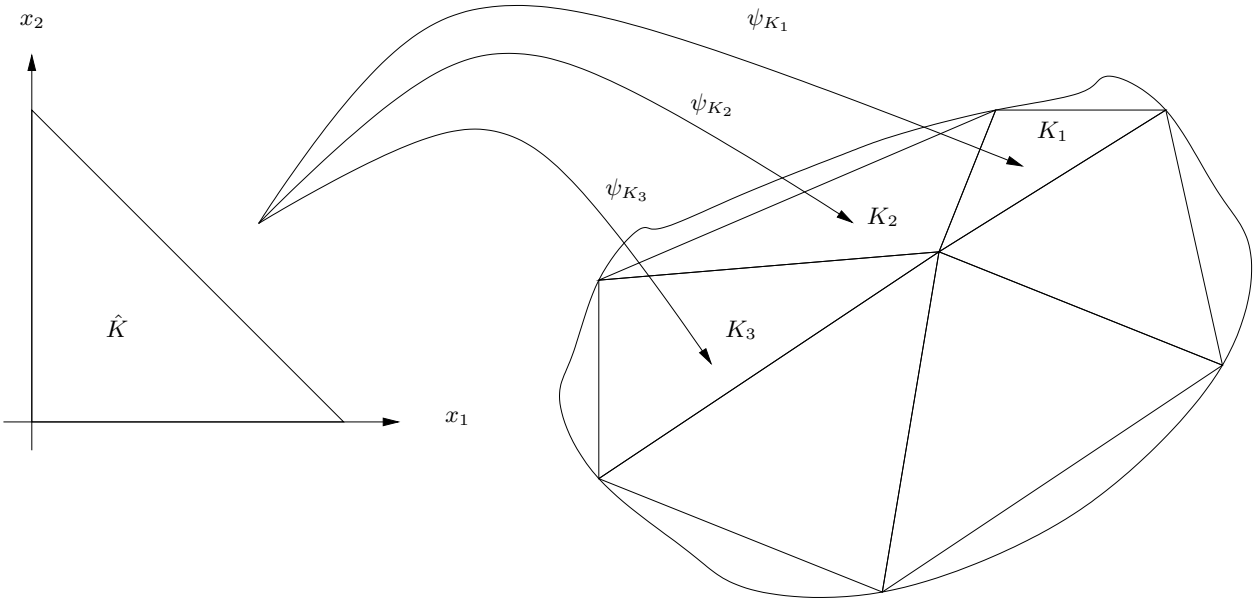


Abbildung 4.7.: Abbildung des Referenzelements $\hat{K}$ in die Zerlegung.

Beweis Sei $\hat{x} \in \hat{K}$ und $x = \psi(\hat{x})$. Für die partiellen Ableitungen der Ordnung k gilt nach der Kettenregel

$$\frac{\partial}{\partial x_{i_k}} \cdots \frac{\partial}{\partial x_{i_1}} v(\hat{x}) = \sum_{j_k=1}^{d} \cdots \sum_{j_1=1}^{d} \left(\frac{\partial}{\partial x_{j_k}} \cdots \frac{\partial}{\partial x_{j_1}} u \right)(x) a_{j_k,i_k} \cdots a_{j_1,i_1} \tag{4.1}$$

für $i_1, \ldots, i_k \in \{1, \ldots, d\}$. Durch k-malige Anwendung der Ungleichung

$$\sum_{i=1}^{d} \left| \sum_{j=1}^{d} a_{j,i} c_i \right|^2 \leq \left\| A^\top \right\|_2^2 \sum_{j=1}^{d} |c_j|^2 \quad \text{sowie} \quad \left\| A^\top \right\|_2 = \|A\|_2$$

erhalten wir

$$\begin{aligned} \sum_{|\alpha| \leq k} \frac{k!}{\alpha!} |D^\alpha v(\hat{x})|^2 &= \sum_{i_k=1}^{d} \cdots \sum_{i_1=1}^{d} \left| \frac{\partial}{\partial \hat{x}_{i_k}} \cdots \frac{\partial}{\partial \hat{x}_{i_1}} v(\hat{x}) \right|^2 \\ &\leq \|A\|_2^{2k} \sum_{j_k=1}^{d} \cdots \sum_{j_1=1}^{d} \left| \frac{\partial}{\partial \hat{x}_{j_k}} \cdots \frac{\partial}{\partial \hat{x}_{j_1}} u(x) \right|^2 \\ &= \|A\|_2^{2k} \sum_{|\beta|=k} \frac{k!}{\beta!} |(D^\beta u)(\psi(\hat{x}))|^2. \end{aligned}$$

Damit

$$\int_{\hat{K}} \sum_{|\alpha|=k} |D^\alpha v(\hat{x})|^2 |\det A| d\hat{x} \leq \|A\|_2^{2k} \int_K \sum_{|\beta|=k} \frac{k!}{\beta!} |(D^\beta u)(x)|^2.$$

Durch Wurzelziehen folgt die Behauptung. □

Lemma 4.7 (Inkreis, Umkreis) *Mit der Bezeichung* $B(a,r) := \{x \in \mathbb{R}^d : |x-a| < r\}$ *und unter den Voraussetzungen*

$$\begin{aligned} B(a^{\hat{K}}, \rho^{\hat{K}}) \subset \quad \hat{K} \quad \subset B(b^{\hat{K}}, h^{\hat{K}}) \\ B(a^{K}, \rho^{K}) \subset \quad K \quad \subset B(b^{K}, h^{K}) \end{aligned}$$

gilt

$$\|A\|_2 \leq \frac{h^K}{\rho^{\hat{K}}}, \quad \left\|A^{-1}\right\|_2 \leq \frac{h^{\hat{K}}}{\rho^K}, \quad \mathrm{cond}_2(A) \leq \frac{h^K h^{\hat{K}}}{\rho^K \rho^{\hat{K}}}.$$

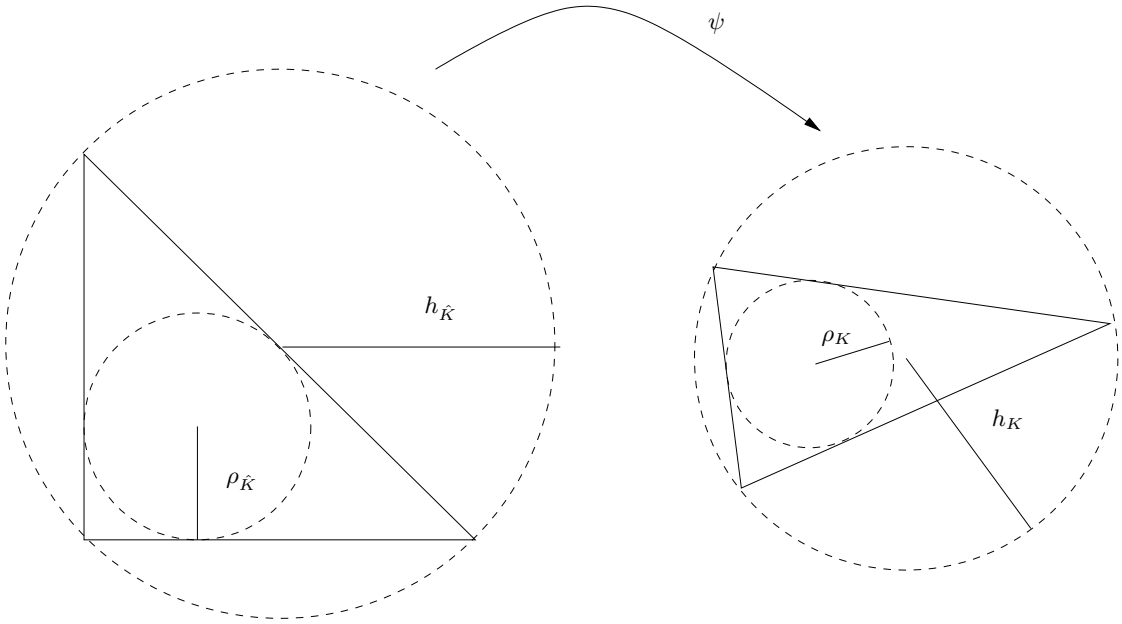


Abbildung 4.8.: Illustration von Lemma 4.7.

Beweis Zu jedem $\hat{x} \in \mathbb{R}^d$ mit $|\hat{x}| \leq 2\rho^{\hat{K}}$ gibt es Punkte $\hat{x}_1, \hat{x}_2 \in \hat{K}$ mit $\hat{x}_1 - \hat{x}_2 = \hat{x}$. Wegen $\psi(\hat{x}_1), \psi(\hat{x}_2) \in K \subset B(b^K, h^K)$ folgt daraus

$$\begin{aligned} |A\hat{x}|_2 &= |\psi(\hat{x}_1) - \psi(\hat{x}_2)| \leq 2h^K \\ \|A\|_2 &= \frac{1}{2\rho^{\hat{K}}} \sup_{|\hat{x}|_2 \leq 2\rho^{\hat{K}}} |A\hat{x}|_2 \leq \frac{h^K}{\rho^{\hat{K}}}. \end{aligned}$$

Durch Vertauschung der Rollen von K und $\hat{K}$ erhalten wir die Ungleichung für $\left\|A^{-1}\right\|$ und durch die Kombination beider Ungleichungen die Abschätzung für $\mathrm{cond}_2(A)$. □

Satz 4.8 (Approximationssatz für Finite Elemente) *Seien folgende Voraussetzungen gegeben:*

1. $(\hat{K}, \hat{\Pi}, \hat{\Sigma})$ *sei ein Referenzelement mit* $\mathbb{P}_{t-1} \subset \hat{\Pi}$ *und die Funktionale* $\hat{\sigma}_j^{\hat{K}} : H^t(\hat{K}) \to \mathbb{R}$ *seien stetig.*

2. *Sei* $\mathcal{T}$ *eine zulässige Zerlegung des Gebietes* Ω *und jedes* $K \in \mathcal{T}$ *gehöre zu einem Finiten Element, das affin interpolationsäquivalent zum Referenzelement* $\hat{K}$ *ist. Sei weiter*

$$h := \sup_{K \in \mathcal{T}} h^K, \quad \kappa := \sup_{K \in \mathcal{T}} \frac{h^K}{\rho^K}, \quad \textit{und} \quad \tilde{\kappa} := \kappa \frac{h^{\hat{K}}}{\rho^{\hat{K}}}$$

und es gelte $h \leq \rho^{\hat{K}}$.

Dann gilt

$$\|u - I_K u\|_{H^m(\Omega)} \leq C_{\hat{K},t} \sqrt{m+1} \tilde{\kappa} \left(\frac{h}{\rho^{\hat{K}}}\right)^{t-m} |u|_{H^t(\Omega)} \tag{4.2}$$

für $m \in \{0, \ldots, t\}$.

Beweis Sei $K \in \mathcal{T}$ und sei

$$\begin{aligned} \psi^K : \hat{K} &\to K \\ \hat{x} &\mapsto x_0^K + A_K \hat{x} \end{aligned}$$

Dann gilt (i., ii. und iii. bezieht sich auf Bemerkung 4.10)

$$\begin{aligned} |u - I_K u|_{H^m(K)} &\underset{\text{ii.}}{\leq} \left\|A_K^{-1}\right\|_2^m |\det A_K^{-1}|^{-1/2} \left|v - I_{\hat{K}} v\right|_{H^m(\hat{K})} \\ &\leq \left\|A_K^{-1}\right\|_2^m |\det A_K^{-1}|^{-1/2} \left\|v - I_{\hat{K}} v\right\|_{H^t(\hat{K})} \\ &\underset{\text{i.}}{\leq} C_{\hat{K},t} \left\|A_K^{-1}\right\|_2^m |\det A_K^{-1}|^{-1/2} |v|_{H^t(\hat{K})} \\ &\underset{\text{ii.}}{\leq} C_{\hat{K},t} \left\|A_K^{-1}\right\|_2^m |\det A_K^{-1}|^{-1/2} \|A_K\|_2^t |\det A_K|^{-1/2} |u|_{H^t(K)} \\ &= C_{\hat{K},t} \operatorname{cond}_2(A_K)^m \|A_K\|^{t-m} |u|_{H^t(K)} \\ &\underset{\text{iii.}}{\leq} C_{\hat{K},t} \tilde{\kappa}^m \left(\frac{h}{\rho^{\hat{K}}}\right)^{t-m} |u|_{H^t(K)} . \end{aligned}$$

Quadriert man diese Ungleichung und summiert man über alle Halbnormen, so folgt (4.2). □

Bemerkung 4.9

1. Das Céa-Lemma macht eine Aussage über die realisierbare Approximationsgüte in dem Raum V, in dem wir die Lösung zu unserer PDE suchen. Da wir elliptische Probleme in $V = H_0^1$ lösen wollen, ist die Aussage von Satz 4.8 in diesem Fall

   $$\|u - I_K u\|_{H^1(\Omega)} \leq C_{\hat{K},t} \sqrt{2} \tilde{\kappa} \left(\frac{h}{\rho^{\hat{K}}}\right)^{t-1} |u|_{H^t(\Omega)}$$

   bzw. in Kombination mit dem Céa-Lemma

   $$\|u - u_h\|_{H^1(\Omega)} \leq \frac{C}{\alpha} C_{\hat{K},t} \sqrt{2} \tilde{\kappa} \left(\frac{h}{\rho^{\hat{K}}}\right)^{t-1} |u|_{H^t(\Omega)} .$$

2. In diesem Fall ist es offensichtlich sinnvoll, für eine gehaltvolle Aussage (d.h. eine sinnvolle Approximation, bei welcher der Approximationsfehler mit $h \to 0$ ebenfalls gegen Null geht) $t \geq 2$ anzunehmen. Wir haben (bisher) jedoch keine Garantie dafür, dass $u \in H^t$ für $t \geq 2$.

   Unter der Annahme $u \in H^2(\Omega)$ gilt also

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

3. Dies ist eine Analogie zur Taylor-Reihenentwicklung: Dort benötigen wir für Aussagen über den Abschneidefehler ebenfalls eine Differenzierbarkeitsstufe mehr, als wir in der Entwicklung selbst berücksichtigt haben.

Bemerkung 4.10 (Zusammenfassung)

i. $\left\|u - I_{\hat{K}} u\right\|_{H^t(\hat{K})} \leq C_{\hat{K},t} \left|u\right|_{H^t(\hat{K})}$,

ii. $\left|v\right|_{H^k(\hat{K})} \leq \left\|A\right\|_2^k \left|\det A\right|^{-1/2} \left|u\right|_{H^k(K)}$ (vgl. Lemma 4.6),

iii. $\left\|A\right\|_2 \leq \frac{h^K}{\rho^{\hat{K}}}$, $\left\|A^{-1}\right\|_2 \leq \frac{h^{\hat{K}}}{\rho^K}$, $\mathrm{cond}_2(A) \leq \frac{h^K h^{\hat{K}}}{\rho^K \rho^{\hat{K}}}$ (vgl. Lemma 4.7).

Wir hätten gerne eine ähnliche Abschätzung des Approximationsfehlers bezüglich der L^2-Norm. Diese lässt sich jedoch nicht auf Basis des Céa-Lemmas gewinnen, da in L^2 die Voraussetzung der Elliptizität der zur PDE gehörenden Bilinearform nicht erfüllt ist. Es ist dennoch möglich, eine Abschätzung in L^2 zu beweisen, siehe Abschnitt 4.4.

Definition 4.11 (Quasi-Uniform) *Für ein Gebiet K sei ρ^K der Radius einer größten Kugel, die in K enthalten ist und h^K der Radius einer kleinsten Kugel, die K enthält. Eine Familie $\{\mathcal{T}_h\}$ von Zerlegungen eines Gebietes Ω heißt* quasi-uniform *(engl.* shape-regular*), falls es eine von h unabhängige Konstante C gibt mit*

$$\sup_{K \in \mathcal{T}_h} \frac{h^K}{\rho^K} < C.$$

Beispiel 4.12 Lagrange-Elemente auf Dreiecken und Tetraedern. Die Menge der Formfunktionen ist hier ein vollständiger Polynomraum mit $p = 1, 2, \ldots$

$$\left\|u - Iu\right\|_{H^m(\Omega)} \leq C h^{t-m} \left|u\right|_{H^t(\Omega)} \quad \text{für} \quad m \in \{0, 1\}, t \in \{1, \ldots, p+1\}.$$

Aber: I enthält Punktauswertungen, die (z.B.) in H^0 nicht definiert sind! Wir müssen sicherstellen, dass die Funktionale $\Sigma = \{\sigma_1, \ldots, \sigma_s\}$, $\sigma_j : H^t(\hat{K}) \to \mathbb{R}$ stetig sind!

Es folgt also ein kleiner Exkurs in die Funktionalanalysis.

Definition 4.13 (Kompakter metrischer Raum) *Ein metrischer Raum X heißt* kompakt*, wenn aus jeder vorgegebenen offenen Überdeckung $\{U_i\}_{i \in I}$ von X endlich viele $U_{i_1}, \ldots, U_{i_n}$ ausgewählt werden können, so dass diese X noch überdecken:*

$$X \subset \bigcup_{i=1}^{N} U_{j_i}.$$

Definition 4.14 (Kompakter Operator) *Es seien X, Y Banachräume und $T : X \to Y$ eine lineare Abbildung. T heißt* kompakt*, wenn folgende gleichwertige Bedingungen erfüllt sind:*

i. Das Bild jeder beschränkten Menke unter T ist relativ kompakt*.*

ii. Das Bild der offenen Einheitskugel ist relativ kompakt*.*

iii. Ist $\{x_k\}_{k \in \mathbb{N}} \subset X$ eine beschränkte Folge, so enthält $\{Tx_k\}_{k \in \mathbb{N}}$ eine konvergente Teilfolge.

(A relativ kompakt*:$\Leftrightarrow \bar{A}$ kompakt) Aus der Kompaktheit einer Abbildung folgt automatisch ihre Stetigkeit.*

Definition 4.15 (Kompakte/stetige Einbettung) *Gilt $X \subset Y$ und die Inklusion* $\mathrm{id} : X \hookrightarrow Y$ *ist kompakt/stetig, so heißt X* kompakt/stetig *in Y* eingebettet*. Im Falle einer stetigen Einbettung gilt*

$$\left\|u\right\|_Y \leq C \left\|u\right\|_X.$$

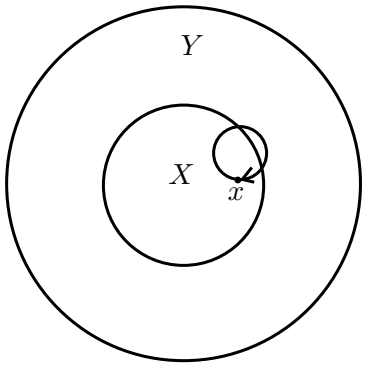


Abbildung 4.9.: Kompakte/stetige Einbettung

Satz 4.16 (Sobolevscher Einbettungssatz) *Sei $\Omega \subset \mathbb{R}^d$ ein Gebiet.*

i. $H^t(\Omega) \hookrightarrow C(\Omega)$ *($C(\Omega)$ versehen mit der Supremumsnorm) ist eine stetige Einbettung, falls $t > \frac{d}{2}$. Insbesondere:*

$\underline{d = 1:}$ $H^1(\Omega) \hookrightarrow C(\Omega) \Rightarrow \sup_{x\in\Omega} |u(x)| \leq C \|u\|_{H^1(\Omega)}$.

$\underline{d \in \{2, 3\}:}$ $H^2(\Omega) \hookrightarrow C(\Omega) \Rightarrow \sup_{x\in\Omega} |u(x)| \leq C \|u\|_{H^2(\Omega)}$.

ii. $H^t(\Omega) \hookrightarrow C^2(\Omega)$ *ist stetig für* $t > 2 + \frac{d}{2}$.

Nach dem Sobolevschen Einbettungssatz ist die Einbettung $H^t(\Omega) \to C(\Omega)$ für $t > \frac{d}{2}$ stetig. Daher muss $t \geq 2$ für $d = 2, 3$ gewählt werden, damit die Knotenfunktionale auf H^t stetig sind.

Bemerkung. $H^t(\Omega) \hookrightarrow C(\Omega)$ für $t > \frac{d}{2}$. Betrachte auf $\Omega := B(0, 1/2) \subset \mathbb{R}^2$ die Funktion

$$u(x_1, x_2) = \log(-\log(x_1^2 + x_2^2)) \in H^1(\Omega).$$

Diese ist weder stetig noch beschränkt!

Beispiel 4.17 (Argyris-Elemente) Hier ist der Ansatzraum $\mathbb{P}_5$ und die Knotenfunktionale beinhalten Ableitungen bis zur Ordnung 2. Da $H^t(\Omega) \hookrightarrow C^2$ für $t > 2+d/2$, muss $t \geq 4$ gewählt werden:

$$\|u - Iu\|_{H^m(\Omega)} \leq Ch^{t-m} |u|_{H^t(\Omega)} \quad \text{für} \quad m \in \{0, 1, 2\}, t \in \{4, 5, 6\}.$$

Nochmal von vorne: Wenn wir Punktauswertungen verwenden wollen, so verlagert sich unser Problem, und wir suchen also ein $u \in \underline{H^2}(\Omega)$, so dass

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

Dann haben wir dank Céa die Abschätzung

$$\|u - u_h\| \leq C \inf_{v\in V_h} \|u - v_h\| \leq C \|u - Iu\| .$$

Das ist aber problematisch. Statt dessen wollen wir lieber unseren Interpolationsoperator umdefinieren. Stichworte hierzu sind der *Clément-* und der *Scott-Zhang-Interpolationsoperator*.

4.3. Fehlerschätzungen für elliptische Probleme

Definition 4.18 *Sei $m \geq 1$, $H_0^m(\Omega) \subset V \subset H^m(\Omega)$ und $a(\cdot,\cdot)$ eine V-elliptische Bilinearform. Das Variationsproblem*

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

*heißt H^s-*regulär*, wenn es zu jedem $f \in H^{s-2m}$ eine Lösung $u \in H^s(\Omega)$ gibt und mit einer Zahl $C(\Omega,a,s)$*

$$\|u\|_s \leq C(\Omega,a,s)\,\|f\|_{s-2m}$$

gilt.

Satz 4.19 *Sei a eine H^1-elliptische Bilinearform mit hinreichend glatten Koeffizientenfunktionen.*

i. Wenn Ω konvex ist, ist das Dirichlet-Problem H^2-regulär.

ii. Sei $s \geq 2$. Wenn Ω einen C^s-Rand besitzt, ist das Dirichlet-Problem H^s-regulär.

Satz 4.20 (Fehlerabschätzung in der Energienorm) *Sei Ω konvex. Ferner sei eine Familie quasiuniformer Triangulierungen $\mathcal{T}_h$ von Ω gegeben. Dann gilt für die FE-Näherung $u_h \in V_h$ bei linearen Dreieckselementen*

$$\|u-u_h\|_1 \leq ch\,\|u\|_2 \leq ch\,\|f\|_0\,.$$

Beweis Es gilt

$$\|u-u_h\|_1 \underset{\text{Céa}}{\leq} C \inf_{v_h \in V_h} \|u-v_h\|_1 \leq C\,\|u-I_h u\|_1 \leq C_1 h\,\|u\|_2 \underset{H^2\text{-Reg.}}{\leq} c_2 h\,\|f\|_0\,.$$

□

4.4. L^2-Abschätzung

Satz 4.21 (Aubin-Nitsche) *Sei H ein Hilbertraum mit der Norm $|\cdot|$ und dem Skalarprodukt $(\cdot,\cdot)$. Weiter sei V ein Unterraum, der durch die Norm $\|\cdot\|$ zum Hilbertraum wird. Ferner sei $V \hookrightarrow H$ stetig. Dann gilt für die Finite-Elemente-Lösung $u_h \in V_h \subset V$*

$$|u-u_h| \leq C\,\|u-u_h\| \sup_{g \in H, g \neq 0} \left[\frac{1}{|g|} \inf_{v \in V_h} \|\varphi_g - v\|\right],$$

wenn jedem $g \in H$ die eindeutige (schwache) Lösung der Gleichung

$$a(w,\varphi_g) = (g,w) \quad \forall w \in V$$

zugeordnet wird, die auch das duale Problem *genannt wird.*

Beweis Die Norm eines Elementes in einem Hilbertraum lässt sich mittels eines Dualitätsargumentes bestimmen (für Details siehe Paragraph B.1.3.1):

$$|w| = \sup_{g\in H, g\neq 0} \frac{(g,w)}{|g|}.$$

Also

$$\begin{aligned}(g, u-u_h) &= a(u-u_h, \varphi_g) \underset{\text{Galerkin-Orth.}}{=} a(u-u_h, \varphi_g - v_h) \\ &\leq C\,\|u-u_h\|\,\|\varphi_g - v_h\|\,.\end{aligned}$$

Da dieses Argument für jedes $v_h \in V_h$ gilt, erhalten wir

$$(g, u-u_h) \leq C\,\|u-u_h\| \inf_{v_h\in V_h} \|\varphi_g - v_h\|\,.$$

Das Dualitätsargument liefert

$$|u-u_h| = \sup_{g\in H, g\neq 0} \frac{(g,u-u_h)}{|g|} \leq C\,\|u-u_h\| \sup_{g\in H, g\neq 0} \left[\frac{1}{|g|} \inf_{v\in V_h} \|\varphi_g - v\|\right].$$

□

Satz 4.22 *Sei Ω konvex. Es gilt*

$$\|u-u_h\|_0 \leq Ch\,\|u-u_h\|_1\,,$$

wenn $u \in H^1(\Omega)$. Gilt außerdem $f \in L^2(\Omega)$ und damit $u \in H^2(\Omega)$ (H^2-Regularität), so ist

$$\|u-u_h\|_0 \leq Ch^2\,\|f\|_0\,.$$

Beweis Direkte Folge von Satz 4.21. □

4.5. Inverse Abschätzung

Motivation: Wir hätten gerne

$$\|u\|_t \leq C\,\|u\|_m$$

für $m > t$.

Definition 4.23 *Sei $t \in \mathbb{N}_0$ und $p \in [1,\infty)$. Wir führen den Raum $H^{1,p}(\mathcal{T}_h)$ aller Funktionen $v : \Omega \to \mathbb{R}$ ein, für die $v|_K \in H^{1,p}(\Omega)$ für alle $K \in \mathcal{T}_h$. Auf diesem Raum definieren wir die* gitterunabhängige Norm

$$\|v\|_{H^{1,p}(\mathcal{T}_h)} := \sum_{K\in\mathcal{T}_h} \|v\|_{H^{1,p}(K)}\,.$$

Satz 4.24 (Inverse Abschätzung) *Sei $\mathcal{T}_h$ eine uniforme Familie von Triangulierungen eines Gebietes $\Omega \subset \mathbb{R}^d$. (D.h. $\mathcal{T}_h$ ist quasi-uniform und es gibt ein $C > 0$ mit*

$$\min_{K\in\mathcal{T}_h} h_K \geq ch$$

für alle h.) Sei (V_h) eine zugehörige Familie von Finite-Elemente-Räumen, wobei alle Finiten Elemente affin äquivalent zu einem Referenz-Element $(\hat{K}, \hat{\Pi}, \hat{\Sigma})$ seien. Dann gibt es für alle $m, t \in \mathbb{N}_0$ mit $m \leq t$ eine Konstante $C > 0$, so dass für alle $h > 0$ die Ungleichung

$$\|v\|_{H^t(\mathcal{T}_h)} \leq Ch^{m-t}\,\|v\|_{H^m(\mathcal{T}_h)}$$

für alle $v \in V_h$ gilt.

Beweis *Schritt 1: auf einem einzigen Element.* Wir zeigen zunächst, dass es eine Konstante $C > 0$ gibt, so dass

$$|v|_{H^t(K)} \leq C\,|v|_{H^m(K)}$$

für alle $v \in \Pi^K$. Sei $\{\varphi_1, \ldots, \varphi_r\}$ eine Orthonormalbasis von $\mathbb{P}_{m-1}$ in $\mathbb{P}_m$. Dann ist nach Satz

$$\|v\|_* := \left(\sum_{i=1}^{r} |\,(v, \varphi_i)_{H^m(K)}\,|^2 + |v|^2_{H^m(K)} \right)^{1/2}$$

eine äquivalente Norm auf $H^m(K)$. Für die orthogonale Projektion

$$Pv := \sum_{i=1}^{r} (v, \varphi_i)_{H^m(K)}\, \varphi_i$$

auf $\mathbb{P}_{m-1}$ gilt

$$|Pv|_{H^m(K)} = |Pv|_{H^t(K)} = 0$$

Außerdem sind auf dem endlichdimensionalen Vektorraum $\Pi^K + \mathbb{P}_{m-1}$ die Normen $\|\cdot\|_{H^t(K)}$ und $\|\cdot\|_*$ äquivalent. Also:

$$|v|_{H^t(K)} \;=\; |v - Pv|_{H^t(K)} \leq \|v - Pv\|_{H^t(K)} \leq C\,|v|_{H^m(K)}$$

für alle $v \in \Pi^K$.

Schritt 2: auf der ganzen Triangulierung. Durch ein *Skalierungsargument* wie im Beweis des Approximationssatzes gilt für alle $0 \leq m \leq t$ die Ungleichung

$$|v|_{H^k(\mathcal{T}_h)} \leq C h^{t-m}\,|v|_{H^m(\mathcal{T}_h)}\,.$$

(Beachte: die h-Potenz kommt durch die Determinante der Transformation ins Spiel!) $\square$

5. Realisierung der Finite-Elemente-Methode

Die praktische Realisierung der Finite-Elemente-Methode erfordert einigen Implementierungsaufwand, da etliche Komponenten zur Durchführung einer Finite-Elemente-Berechung von Nöten sind. Zu allererst wird ein *Gitter* bzw. eine *Triangulierung* des Gebietes, auf dem die partielle Differentialgleichung gelöst werden soll, benötigt. Dessen Beschreibung und grobe Gewinnung wird in Abschnitt 5.1 betrachtet, wobei die Erzeugung des Gitters oft nicht Teil der Finite-Elemente-Software selbst ist, sondern mit externen Programmen als *Preprocessing*-Schritt durchgeführt wird. Aufbauend auf dem Gitter muss der Finite-Elemente-Raum repräsentiert werden, wobei hier der Komplexität je nach Anforderungsprofil — gemischte Elementtypen, Anzahl an gesuchten Funktionen (für Systeme von Differentialgleichungen) — nach oben keine Grenzen gesetzt sind. In jedem Fall führt die Finite-Elemente-Methode auf Gleichungssysteme — diese können je nach Gleichung linear oder nichtlinear sein —, die diskret repräsentiert und gelöst werden müssen. Dafür werden insbesondere Datenstrukturen für Matrizen und Vektoren benötigt. Diese müssen in geeigneter Weise mit der diskreten Repräsentation der (kontinuierlichen) Bilinearform und rechten Seite im Sinne des Finite-Elemente-Raumes befüllt, d.h. *assembliert*, werden. Dieser Prozess ist Gegenstand von Abschnitt 5.2. Die Lösung der Gleichungssysteme wird, zumindest für den linearen Fall, im nächsten Kapitel diskutiert. Schließlich werden für die Betrachtung und Begutachtung der Resultate der Berechnung noch Routinen benötigt, welche die Ergebnisse in geeignete Dateiformate zur Visualisierung, beispielsweise VTK (`http://www.vtk.org/`), schreiben. Dies ist Teil des sogenannten *Postprocessing*.

5.1. Gittererzeugung

Beschreibung des Gitters: Ein Dreiecksgitter im $\mathbb{R}^2$ wird im Allgemeinen durch folgende Daten beschrieben:

a) eine Liste mit den Koordinaten aller Eckpunkte

b) eine Liste von Dreiecken mit den Nummern der Eckpunkte jedes Dreiecks

a)

	1	2	3	$\cdots$	9
x	0	0	0	$\cdots$	$\frac{3}{2}$
y	3	2	1	$\cdots$	$\frac{1}{2}$

b)

	1	2	3	$\cdots$
#	1	1	2	$\cdots$
#	7	2	4	$\cdots$
#	8	7	7	$\cdots$

Tabelle 5.1.: Datenstrukturen zur Darstellung der Triangulierung aus Abbildung 5.1.

Qualitätskriterien: Für Finite Elemente Rechnungen wünscht man sich im Allgemeinen Gitter, für die

- die Anzahl der Eckpunkte bei vorgegebener Maximalgröße h eines Elements klein ist
- der kleinste vorkommende Innenwinkel möglichst groß ist.

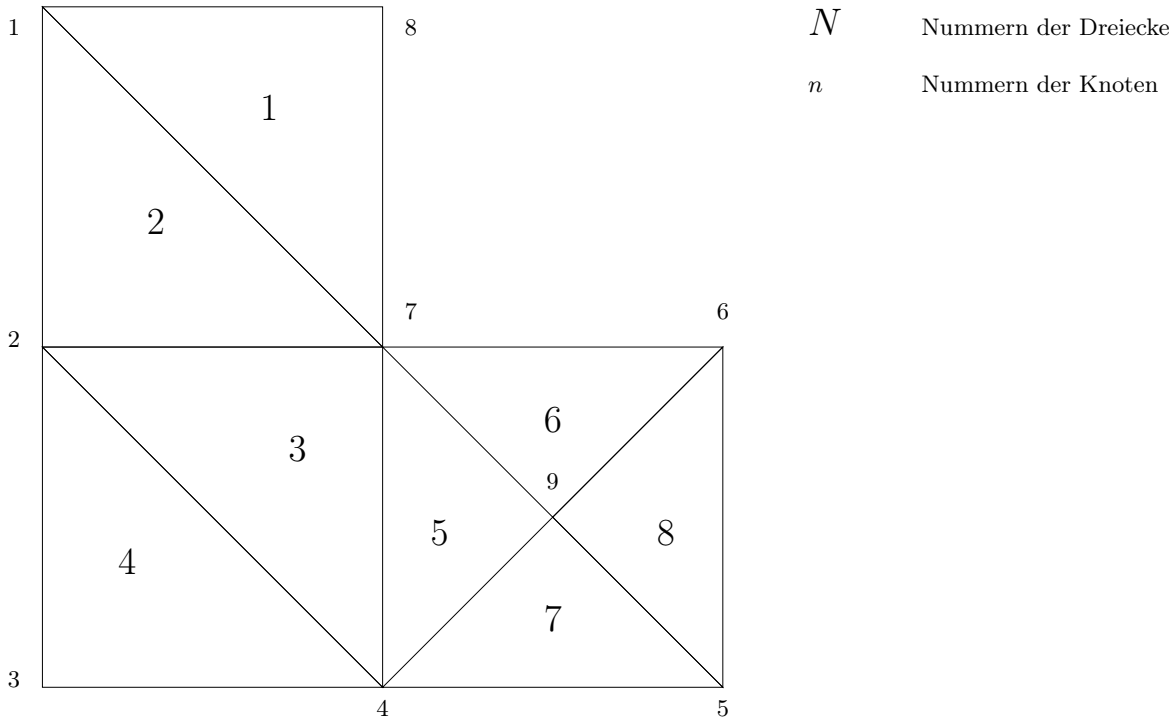


Abbildung 5.1.: Beispiel einer Triangulierung.

Als Bemerkung sei hier eingeflochten, dass die obige Datenstruktur durch objektorientiertes Programmieren (z.B. in C++ oder Python) erheblich eleganter programmiert werden kann.

5.1.1. Delaunay-Triangulierung

Problem: Gegeben sei eine endliche Menge S von Punkten im $\mathbb{R}^2$. Gesucht ist eine Triangulierung der konvexen Hülle dieser Punkte, in der genau diese Punkte als Eckpunkte vorkommen und für die der kleinste vorkommende Innenwinkel möglichst groß ist.

Dazu definieren wir die *Voronoi-Umgebung* eines Punktes $p \in S$ als Menge aller Punkte in $\mathbb{R}^2$, die mindestens so dicht an p liegen wie an irgendeinem anderen Punkt in S.

$$V_p := \{x \in \mathbb{R}^2 : |x - p|_2 \leq |x - q|_2 \text{ für alle } q \in S\}.$$

Wir verbinden je zwei Punkte von S genau dann durch eine gerade Linie, wenn ihre Voronoi-Umgebungen eine gemeinsame Kante besitzen. Die so entstehende Zerlegung der konvexen Hülle von S heißt *Delaunay-Triangulierung*.

Satz 5.1 *Sei S eine endliche Menge von Punkten im $\mathbb{R}^2$ in allgemeiner Lage und sei $\mathcal{T}$ eine Triangulierung der konvexen Hülle von S, deren Eckpunkte genau die Punkte von S sind. Dann sind folgende Aussagen äquivalent:*

1. *$\mathcal{T}$ ist* die *Delaunay-Triangulierung von S.*
2. *(Kreiskriterium) Das Innere des Umkreises jedes Dreiecks $K \in \mathcal{T}$ enthält keine Punkte von S.*

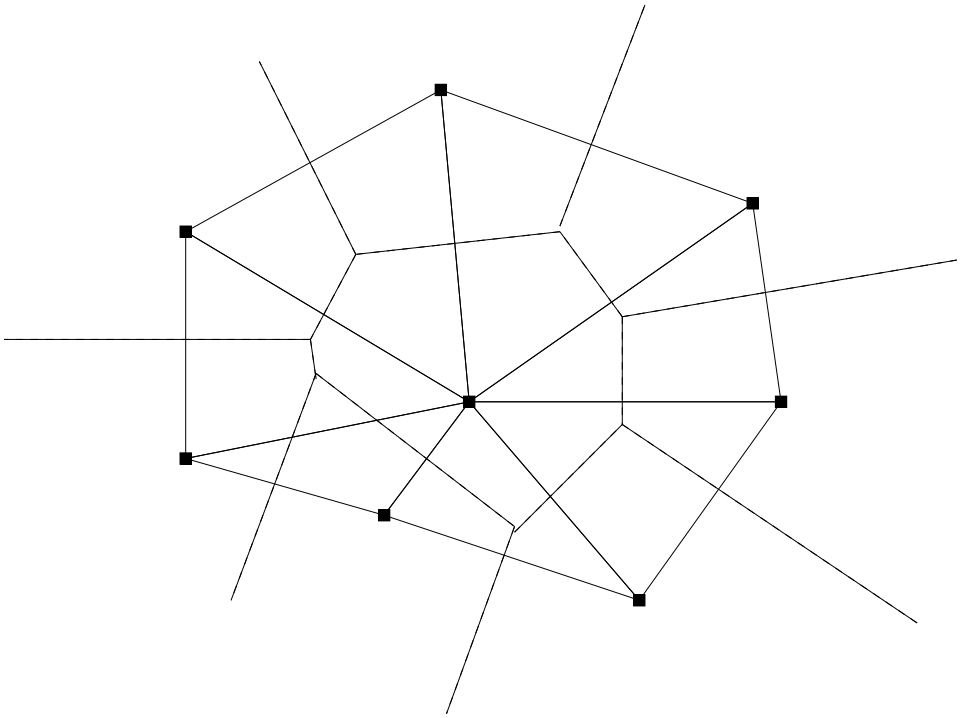


Abbildung 5.2.: Entsprechung zwischen einem Voronoi-Diagramm und der Delaunay-Triangulierung.

3. *(Winkelkriterium) Der kleinste Innenwinkel aller Dreiecke $K \in \mathcal{T}$ ist maximal unter allen möglichen Triangulierungen der betrachteten Art.*

Inkrementeller Delaunay-Algorithmus:

- Initialisierung: $S \leftarrow \{x^{(1)}, \ldots, x^{(n)}\}$.
- Schritt 1: Wähle ein Dreieck, dessen Inneres die konvexe Hülle von S enthält:

  $$T_0 := \Delta((3m, 0), (-3m, -3m), (0, 3m)), \quad \text{wobei} \quad m > \max_i\{|x_1^{(i)}|, |x_2^{(i)}|\}.$$

  Setze $\mathcal{T} \leftarrow \{T_0\}$
- Schritt 2: Für $i = 1, \ldots, n$:
  - Bestimme ein Dreieck $K \in \mathcal{T}$ mit $x^{(i)} \in K$.
  - Verbinde $x^{(i)}$ mit den Eckpunkten von K und ersetze K in $\mathcal{T}$ durch die so definierten neuen Dreiecke.
  - Solange eine unzulässige (d.h. nicht das Kreiskriterium bzgl $x^{(i+1)}$ erfüllende) Kante $[p, q]$ in $\mathcal{T}$ existiert:
    - Drehe die Kante $[p, q]$. (vgl. Abbildung 5.3)
- Schritt 3: Schneide die überflüssigen Stücke, die durch das große Dreieck T_0 dazugekommen sind, einfach ab.

Vgl. `http://www-users.informatik.rwth-aachen.de/~roberts/software.html`.

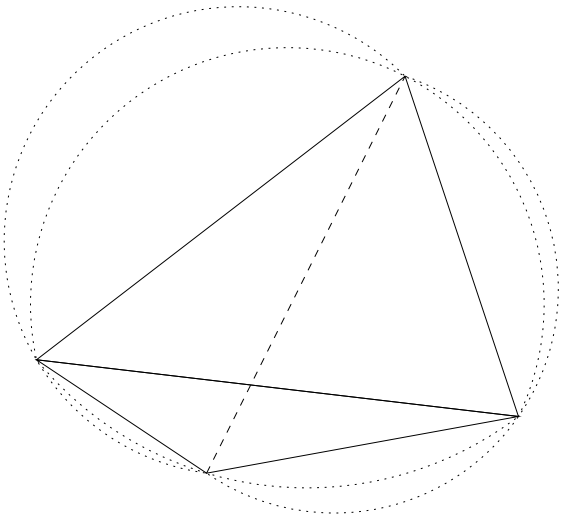


Abbildung 5.3.: Drehen einer Kante im Delaunay-Algorithmus.

5.2. Assemblierung der Finite-Elemente-Matrizen

5.2.1. Besetzungsstruktur und Gesamtalgorithmus

Bei der Diskretisierung mit Finiten Elementen ist eine approximierende Lösung $u_h \in V_h$ gesucht, wobei V_h einen Finite-Elemente-Raum mit $\dim V_h < \infty$ bezeichnet. Da V_h ein Vektorraum ist, wird für die Lösung u_h der Ansatz

$$u_h = \sum_{j=1}^{n} u_h^{(j)} \psi_h^{(j)}$$

mit der gegebenen Finite-Elemente-Basis $\{\psi_h^{(j)}\}_{j=1}^n$ gemacht. Dabei ist n die Anzahl der Freiheitsgrade im gesamten Rechengitter und $u_h^{(j)}$ der gesuchte Wert von u_h am j-ten Freiheitsgrad ($j \in \{1, \ldots, n\}$).

Die Variationsformulierung linearer partieller Differentialgleichungen führt zu einer *Bilinearform*

$$a : V \times V \to \mathbb{R}.$$

Im diskreten Fall wird der Raum der Testfunktionen V ebenfalls auf den endlichdimensionalen Raum V_h eingeschränkt, so dass für eine Testfunktion $\varphi_h \in V_h$ ebenfalls ein Ansatz der Form

$$\varphi_h = \sum_{i=1}^{n} \varphi_h^{(i)} \psi_h^{(i)}$$

gemacht werden kann. Da in V_h aufgrund der Vektorraum-Struktur das Superpositionsprinzip gilt und a *linear* bezüglich der Testfunktion ist, ist die Forderung

$$a(u_h, \varphi_h) = (f, \varphi_h) \quad \text{für alle} \quad \varphi_h \in V_h$$

äquivalent dazu,

$$a(u_h, \psi_h^{(i)}) = \left(f, \psi_h^{(i)}\right) \quad \text{für alle} \quad \psi_h^{(i)},\ i \in \{1, \ldots, n\}$$

zu fordern. Setzt man schließlich den Ansatz für u_h ein, so folgt aufgrund der Linearität von a bezüglich der gesuchten Lösung

$$a(u_h, \psi_h^{(i)}) = a\left(\sum_{j=1}^{n} u_h^{(j)} \psi_h^{(j)}, \psi_h^{(i)}\right) = \sum_{j=1}^{n} a\left(u_h^{(j)} \psi_h^{(j)}, \psi_h^{(i)}\right) = \sum_{j=1}^{n} u_h^{(j)} a\left(\psi_h^{(j)}, \psi_h^{(i)}\right) = \left(f, \psi_h^{(i)}\right)$$

für alle $\psi_h^{(i)}$, $i \in \{1, \ldots, n\}$. Dies ist nichts anderes als ein lineares Gleichungssystem

$$Au = b,$$

wobei die Matrix A und die rechte Seite b durch

$$A := (a_{i,j})_{i,j=1}^{n}, \qquad a_{i,j} := a\left(\psi_h^{(j)}, \psi_h^{(i)}\right),$$
$$b := (b_i)_{i=1}^{n}, \qquad b_i := \left(f, \psi_h^{(i)}\right)$$

gegeben sind und

$$u := \left(u_h^{(j)}\right)_{j=1}^{n}$$

der gesuchte Lösungsvektor ist. Man kann sich also merken, dass eine *Testfunktion* eine *Zeile* und eine *Ansatzfunktion* eine *Spalte* der Matrix A definiert. Ebenso definiert eine Testfunktion eine Komponente der rechten Seite b und eine Ansatzfunktion eine Komponente des Lösungsvektors u.

Beispiel 5.2 Mit $a(u, \varphi) := (\nabla u, \nabla \varphi)$ ergibt sich das lineare Gleichungssystem

$$\sum_{j=1}^{n} u_h^{(j)} \left(\nabla \psi_h^{(j)}, \nabla \psi_h^{(i)}\right) = \left(f, \psi_h^{(i)}\right) \quad \text{für alle} \quad i \in \{1, \ldots, n\}.$$

Diese spezielle Matrix A heißt *Steifigkeitsmatrix*.

Die Bilinearform a ist durch ein Integral über das Gebiet Ω definiert. Folglich können nur diejenigen Einträge $a_{i,j} \neq 0$ sein, für deren definierende Ansatz- und Testfunktionen

$$\operatorname{supp} \psi_h^{(i)} \cap \operatorname{supp} \psi_h^{(j)} \neq \emptyset$$

erfüllt ist. Typischerweise gibt es daher sehr viel mehr Einträge, die gleich Null sind, als solche, die ungleich Null sind, da die Funktionen ψ_h bei den meisten Wahlen für V_h sehr kleine, lokal begrenzte Träger haben. Diese Eigenschaft nennt man *dünnbesetzt* (*engl:* sparse).

Im diskreten Fall wird statt über Ω über dessen Triangulierung $\mathcal{T}_h$ integriert, so dass das Integral über Ω gemäß

$$\begin{aligned} a_{i,j} &= \int_\Omega g\left(\psi_h^{(j)}, \psi_h^{(i)}\right) \mathrm{d}x \\ &= \sum_{K \in \mathcal{T}_h} \int_K g\left(\psi_h^{(j)}, \psi_h^{(i)}\right) \mathrm{d}x \\ &=: \sum_{K \in \mathcal{T}_h} \hat{a}_{i,j} \end{aligned}$$

berechnet werden kann, wobei g den Integranden in der Definition von a bezeichnet und angenommen wurde, dass

$$\bigcup_{K \in \mathcal{T}_h} = \Omega, \quad \overset{\circ}{K}_i \cap \overset{\circ}{K}_j = \emptyset \ (i \neq j).$$

Die *lokalen Beiträge* $\hat{a}_{i,j}$ können folglich nur dann von Null verschieden sein, wenn $\psi_h^{(j)}, \psi_h^{(i)}$ die Eigenschaften

$$\operatorname{supp}\psi_h^{(i)} \cap \operatorname{supp}\psi_h^{(j)} \neq \emptyset, \quad \operatorname{supp}\psi_h^{(i)} \cap K \neq \emptyset, \quad \operatorname{supp}\psi_h^{(j)} \cap K \neq \emptyset$$

erfüllen.

Der Algorithmus zur Assemblierung einer Finite-Elemente-Matrix ist also der folgende:

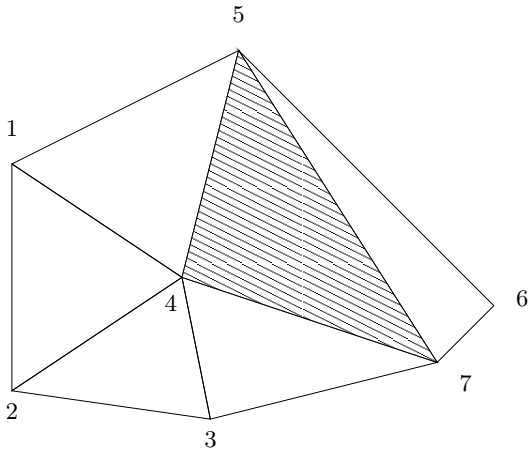


Abbildung 5.4.

Algorithmus 5.3 (Assemblierung einer Finite-Elemente-Matrix) Für jedes $K \in \mathcal{T}_h$:

1. Berechne die *lokale* Finite-Elemente-Matrix

$$\hat{a}_{i,j} = \int_K g\left(\psi_h^{(j)}, \psi_h^{(i)}\right) \mathrm{d}x,$$

so dass

$$\operatorname{supp}\psi_h^{(i)} \cap \operatorname{supp}\psi_h^{(j)} \neq \emptyset, \quad \operatorname{supp}\psi_h^{(i)} \cap K \neq \emptyset, \quad \operatorname{supp}\psi_h^{(j)} \cap K \neq \emptyset.$$

Beispiel: Für ein Dreieck mit $\mathbb{P}_1$ als Ansatzraum ergibt sich eine 3×3-Matrix als lokale Finite-Elemente-Matrix.

2. Verteile die *lokalen Beiträge* $\hat{a}_{i,j}$ additiv in die *globale Matrix* A. Verwende dazu eine globale Nummerierung der Freiheitsgrade von 1 bis n.

Man verwendet hierzu eine so genannte *connectivity matrix* P:

$$\underbrace{\begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}}_{\text{lokale Nummerierung}} \xrightarrow{P} \underbrace{\begin{pmatrix} 5 \\ 4 \\ 7 \end{pmatrix}}_{\text{globale Nummerierung}}$$

Also führt man

$$a_{P(i),P(j)} = a_{P(i),P(j)} + \hat{a}_{i,j}$$

aus.

Bemerkung 5.4 1. Die Iteration über alle $K \in \mathcal{T}_h$ und die additive Verteilung der lokalen Beiträge in die globale Matrix in Schritt 2 nennt man auch den *globalen Teil der Assemblierung*. Dieser Algorithmus ist generisch für Finite-Elemente-Verfahren und muss nur einmal implementiert werden!

2. Schritt 1 zur Assemblierung der lokalen Beiträge wird analog auch *lokaler Teil der Assemblierung* genannt. Innerhalb des Assemblierungsprozesses ist dies der einzige Teil, der spezifisch für die zu lösende partielle Differentialgleichung ist! Daher muss die entsprechende Routine zur Berechnung für jede Differentialgleichung individuell geschrieben werden. Dass man auch dies mit sehr generischen Mitteln bewerkstelligen kann, wird im Folgenden dargestellt werden.

3. Algorithmus 5.3, dessen Herleitung und die ersten beiden Bemerkungen gelten sinngemäß natürlich auch für Vektoren! Hier sind die lokalen Beiträge entsprechend kleine Vektoren, die ebenfalls additiv in einen globalen Vektor zusammengeführt werden müssen.

5.2.2. Berechnung der lokalen Beiträge: Basisfunktionen und Transformation auf das Referenzelement

Wie im letzten Abschnitt 5.2.1 dargestellt wurde, lässt sich die Diskretisierungsmatrix A durch *lokale* Beiträge effizient berechnen. Bei der tatsächlichen Berechnung dieser Beiträge bedient man sich des *Transformationssatzes* für Integrale: Die Idee dabei ist es, die Integration nicht über das Element K auszuführen, sondern dieses Integral auf das Referenzelement $\hat{K}$ zu transformieren. Diese Vorgehensweise bietet den wesentlichen Vorteil, dass man nur eine Quadraturformel für das Referenzelement benötigt und somit die lokalen Beiträge einheitlich und weitestgehend unabhängig von der tatsächlichen Gestalt des Elementes K behandeln kann.

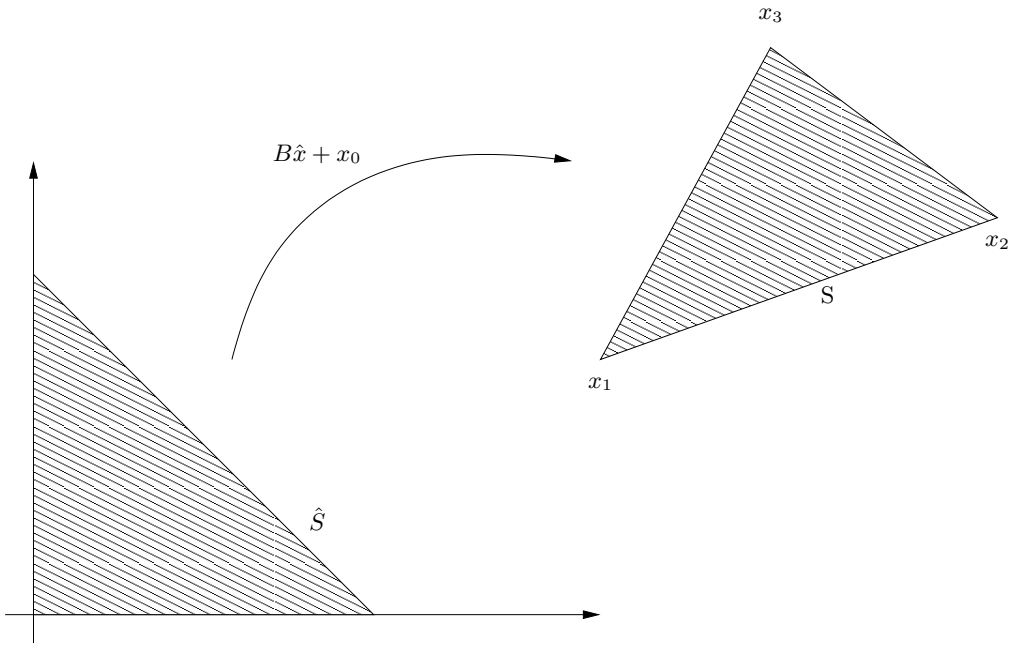


Abbildung 5.5.: Erinnerung: Eine affine Transformation des Einheitssimplex in zwei Dimensionen.

Sei Φ eine bijektive, affin lineare Abbildung mit $\Phi(\hat{K}) = K$. Dann gilt

$$\int_K g\left(\psi_h^{(j)}, \psi_h^{(i)}\right)(x)\mathrm{d}x = \int_{\hat{K}} g\left(\psi_h^{(j)}, \psi_h^{(i)}\right)(\Phi(\hat{x}))\left|\det\left(\hat{\mathrm{D}}\Phi(\hat{x})\right)\right|\mathrm{d}\hat{x},$$

wobei $\hat{\mathrm{D}}\Phi(\hat{x})$ die Jacobi-Matrix von Φ im Punkt $\hat{x}$ bezeichnet.

In der Definition von g können neben den ψ_h auch deren Ableitungen vorkommen. Dabei ist es wünschenswert, dass die ψ_h und deren Ableitungen auf K mittels der Formfunktionen $\hat{\psi}_h \in \hat{\Pi}$ auf $\hat{K}$ berechnet werden können. Dabei gilt mit $\Phi(\hat{x}) := B\hat{x} + x_0$

$$\begin{aligned} \psi_h(x) &= \hat{\psi}_h(\hat{x}), & x &:= \Phi(\hat{x}), \\ \nabla\psi_h(x) &= B^{-\top}\hat{\nabla}\hat{\psi}_h(\hat{x}), & x &:= \Phi(\hat{x}), \end{aligned}$$

d.h. im Falle von Ableitungen müssen die Ableitungsoperatoren mittransformiert werden (Übung!). Da Φ als bijektiv und affin linear vorausgesetzt wurde, ist die Matrix $B^{-\top}$ wohldefiniert. Weiterhin wurden stillschweigend ψ_h und $\hat{\psi}_h$ gemäß der Zuordnung von Freiheitsgraden auf K zu Freiheitsgraden auf $\hat{K}$ miteinander identifiziert.

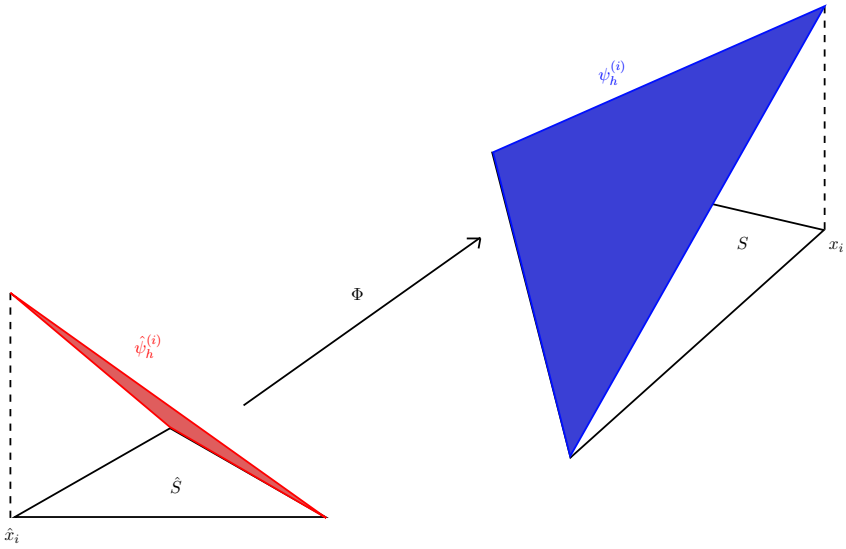


Abbildung 5.6.: Transformation einer Formfunktion $\hat{\psi}_h^{(i)} \in \hat{\Pi}$ von $\hat{S}$ nach S.

Beispiel 5.5 Für die Poisson-Gleichung ergibt sich

$$\begin{aligned} \hat{a}_{i,j} &= \int_K \nabla\psi_h^{(j)}(x) \cdot \nabla\psi_h^{(i)}(x)\mathrm{d}x \\ &= \int_{\hat{K}} \left(B^{-\top}\hat{\nabla}\hat{\psi}_h^{(j)}(\hat{x})\right) \cdot \left(B^{-\top}\hat{\nabla}\hat{\psi}_h^{(i)}(\hat{x})\right) |\det(B)|\,\mathrm{d}\hat{x} \end{aligned}$$

Das folgende Beispiel soll die Identifizierung von ψ_h mit $\hat{\psi}_h$ anhand von Dreiecken mit linearem Lagrange-Ansatz veranschaulichen:

Beispiel 5.6 (Lagrangsche Finite-Elemente auf Dreiecken) In Abbildung 5.5 ist $\hat{S}$ durch die Eckpunkte

$$\hat{x}_1 = (0,0)^\top, \quad \hat{x}_2 = (1,0)^\top, \quad \hat{x}_3 = (0,1)^\top$$

gegeben und Φ erfülle

$$\Phi(\hat{x}_i) = x_i, \quad \forall i \in \{1, 2, 3\}.$$

Bei Verwendung eines $\mathbb{P}_1$-Ansatzes für Lagrangesche Finite Elemente sind die drei lokalen Formfunktionen durch

$$\hat{\psi}_h^{(i)}(\hat{x}_j) = \delta_{i,j}$$

eindeutig definiert. Gemäß Satz 4.3 können die lokalen Freiheitsgrade auf $\hat{S}$ mit den Werten der lokalen Formfunktionen an den Punkten $\hat{x}_i$ identifiziert werden. Man ordnet daher eine Formfunktion $\hat{\psi}_h^{(i)}$ auch dem Punkt $\hat{x}_i$ zu, an welchem sie dem Wert Eins annimmt. Eine Formfunktion $\psi_h^{(i)}$ wird nun mit $\hat{\psi}_h^{(i)}$ über

$$\psi_h^{(i)}(x_j) = \psi_h^{(i)}(\Phi(\hat{x}_j)) = \delta_{i,j} = \hat{\psi}_h^{(i)}(\hat{x}_j)$$

identifiziert.

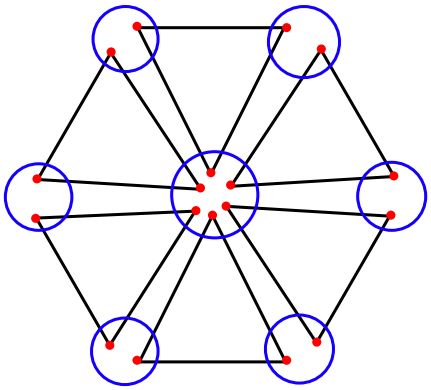


Abbildung 5.7.: Identifizierung *lokaler* Freiheitsgrade (rot) benachbarbarter Dreiecke zu *globalen* Freiheitsgraden (blau umrandet).

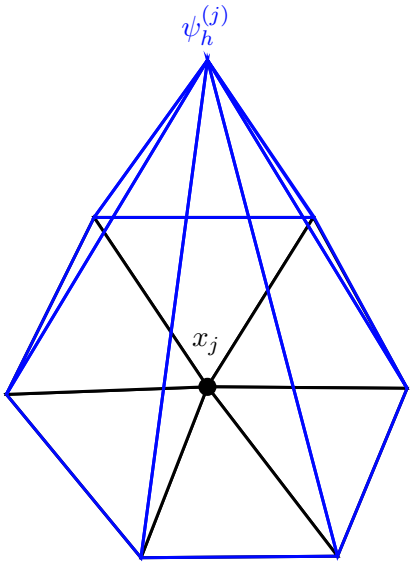


Abbildung 5.8.: Basisfunktion nach der Identifizierung gemeinsamer Freiheitsgrade auf benachbarten Dreiecken

Es sei an dieser Stelle betont, dass mehrere Freiheitsgrade von verschiedenen Elementen $K \in \mathcal{T}_h$ geometrisch auf derselben Entität (z.B. Punkt, Kante, Fläche) definiert sein können. Die Definition 4.1 von Finiten Elementen erlaubt es, diese unabhängig voneinander zu behandeln. Dies führt auf die sogenannte *unstetige (Galerkin-)Finite-Elemente-Methode* (*engl.:* Discontinuous Galerkin-FEM bzw. DG-FEM), die nicht Gegenstand dieser Vorlesung ist.

In dieser Vorlesung wird die *stetige (Galerkin-)Finite-Elemente-Methode* betrachtet. Dabei werden zueinander korrespondierende Freiheitsgrade von verschiedenen $K \in \mathcal{T}_h$, die geometrisch auf derselben Entität definiert sind, miteinander identifiziert und zu einem *globalen* Freiheitsgrad mit *einer* zugehörigen Basisfunktion zusammengefasst, siehe Abbildungen 5.7 und 5.8. Dies führt dazu, dass die Finite-Elemente-Lösung u_h bezüglich der Freiheitsgrade *stetig* über die verschiedenen Elemente $K \in \mathcal{T}_h$ hinweg ist.

Bemerkung 5.7

- Bei der DG-FE-Methode müssen zusätzliche, von der partiellen Differentialgleichung abhängige Terme, assembliert werden, um eine Wohlgestelltheit des diskreten Problems zu gewährleisten. Dies sind die sogenannten *Sprungterme* (*engl.:* jump terms), deren Konstruktion alles andere als trivial ist. Durch die Forderung der Stetigkeit einer Basisfunktion über die Elemente $K \in \mathcal{T}_h$, deren Schnitt mit dem Träger der Basisfunktion nichtleer ist, entfällt bei der stetigen FE-Methode die Notwendigkeit solcher zusätzlichen Terme.
- Da beispielsweise auch Ableitungen der Formfunktionen als Freiheitsgrade definiert werden dürfen, entstehen durch die Identifizierung der Freiheitsgrade Basisfunktionen, deren entsprechenden Ableitungen dann ebenfalls stetig sind.
- Alle Aussagen bezüglich der Assemblierung von Matrizen gelten natürlich auch wieder sinngemäß für Vektoren.

Zur tatsächlichen Berechung der Integrale auf den Referenzelementen werden entsprechende Quadraturregeln benötigt, deren Herleitung und Diskussion jedoch nicht Gegenstand dieser Vorlesung ist.

6. Löser für große dünnbesetzte lineare Gleichungssysteme

Wie in Abschnitt 5.2.1 dargelegt wurde, führt die Diskretisierung von partiellen Differentialgleichungen mit der Finite-Elemente-Methode auf sogenannte *dünnbesetzte* Matrizen, bei denen es sehr viel mehr Einträge gibt, die gleich Null sind, als solche, die von Null verschieden sind. Dieser Eigenschaft muss in der Praxis Rechnung getragen werden, sowohl in Bezug auf die Speicherung der Matrizen als auch in der Wahl der Löser für die linearen Gleichungssysteme.

Beispiel 6.1 (Dünn- versus vollbesetzt) Angenommen, es wird die Poisson-Gleichung auf Dreieckselementen mit linearem Lagrange-Ansatz diskretisiert. Unter der Annahme, dass höchsten sechs Dreiecke zu einem Knoten im Gitter adjazent sind, gibt es pro Zeile höchstens 7 Einträge, die von Null verschieden sind. Wenn es im gesamten Gitter nun 10.000 Knoten gibt, dann hat die Matrix, wenn sie *vollbesetzt* oder *dicht* abgespeichert wird

$$10.000 \cdot 10.000 = 100.000.000$$

Einträge, von denen höchstens $7 \cdot 10.000 = 70.000$ von Null verschieden sind. Somit müssen nur maximal $0,07\%$ der Einträge wirklich gespeichert werden! Unter der Annahme, dass ein Eintrag 8 Byte Speicher braucht, ergibt sich für die vollbesetzte Matrix ein Gesamtspeicherbedarf von ungefähr ≈ 762.94 MByte, während für die Nicht-Null-Einträge ≈ 0.53 MByte benötigt werden.

Die gewählten Löser sollten ebenfalls von dieser Struktur profitieren. Somit sind z.B. LU- oder QR-Zerlegungen weniger geeignet, da durch die Elimination bzw. Orthogonalisierung vormalige Null-Einträge zu Nicht-Null-Einträgen werden. Diese nennt man daher auch *fill-ins*.

Im Folgenden werden vornehmlich Projektionsmethoden betrachtet. Diese Klasse von Lösern hat die Eigenschaft, dass die Finite-Elemente-Matrix lediglich in Form von Matrix-Vektor-Produkten benötigt wird - eine Operation, die ebenfalls von der Struktur der Matrix profitiert, da bei geeigneter Speicherung unnötige Multiplikationen mit Null-Einträgen vermieden werden, was zu einer erheblichen Beschleunigung der Rechenzeit der Verfahren beiträgt.

6.1. Eigenschaften der Finite-Elemente-Matrizen zu stetigen, koerziven Bilinearformen

Sei $a : V \times V \to \mathbb{R}$ die stetige und koerzive Bilinearform, die zu einer schwachen Formulierung einer elliptischen Differentialgleichung zweiter Ordnung gehört, sowie V ein Hilbertraum. Sei $V_h \subset V$ ein endlichdimensionaler Finite-Elemente-Raum der Dimension $n < \infty$ und $\{\psi_i\}_{i=1}^n \subset V_h$ eine Finite-Elemente-Basis von V_h und $A \in \mathbb{R}^{n \times n}$ die zugehörige Finite-Elemente-Matrix

zur Bilinearform a. Dann gilt für ein beliebiges $x \in \mathbb{R}^n \backslash \{0\}$

$$\begin{aligned} x^\top Ax &= \sum_{i,j=1}^{n} x_i a_{ij} x_j \\ &= \sum_{i,j=1}^{n} x_i a(\psi_j, \psi_i) x_j \\ &= \sum_{i,j=1}^{n} a(x_j \psi_j, x_i \psi_i) \\ &= a\left(\underbrace{\sum_{j=1}^{n} x_j \psi_j}_{=:\tilde{x} \neq 0}, \underbrace{\sum_{i=1}^{n} x_i \psi_i}_{=\tilde{x}} \right) \\ &\geq \lambda \left\| \tilde{x} \right\|_V^2 \\ &> 0 \end{aligned}$$

aufgrund der Koerzivität der Bilinearform a, das heißt, die Matrix A ist positiv definit.

Ist die Bilinearform a zusätzlich symmetrisch, so folgt wegen

$$a_{ij} = a(\psi_j, \psi_i) = a(\psi_i, \psi_j) = a_{ji}$$

die Symmetrie der Matrix A.

6.2. Allgemeine Eigenschaften von Projektionsverfahren

Betrachte das lineare Gleichungssytem

$$Ax = b \tag{6.1}$$

mit einer reellen und regulären $n \times n$-Matrix A, einer rechten Seite $b \in \mathbb{R}^n$ und einem gesuchten Lösungsvektor $x \in \mathbb{R}^n$. Mit A werden sowohl die Matrix selbst als auch die lineare Abbildung bezeichnet, die sie repräsentiert.

Die grundlegende Idee von *Projektionsverfahren* ist es, Näherungslösungen an die Lösung von (6.1) aus einem Unterraum von $\mathbb{R}^n$ zu bestimmen. Wenn $\mathcal{K}$ diesen Raum der *approximierenden Kandidaten* oder *Suchunterraum* bezeichnet und m seine Dimension, dann müssen, im Allgemeinen, m Bedingungen gestellt werden, um so eine Approximation zu gewinnen. Ein typischer Weg, diese Bedingungen zu beschreiben, ist es, m (unabhängige) Orthogonalitätsbedingungen zu stellen. Genauer: Das Residuum

$$b - Ax \tag{6.2}$$

wird mit Nebenbedingungen derart versehen, dass es orthogonal zu m linear unabhängigen Vektoren ist. Dies definiert einen zweiten Unterraum $\mathcal{L} \subset \mathbb{R}^n$ der Dimension m, der *Raum der Nebenbedingungen* oder auch *linker Unterraum* genannt wird (der Name erklärt sich später). Dies sind sogenannte *Petrov-Galerkin-Bedingungen*.

Es gibt zwei große Klassen von Projektionsmethoden: *orthogonale*, bei denen $\mathcal{K}$ und $\mathcal{L}$ identisch sind, was die Petrov-Galerkin- zu Galerkin-Bedingungen vereinfacht, und *schiefe*, bei denen sich $\mathcal{L}$ und $\mathcal{K}$ voneinander unterscheiden (Petrov-Galerkin-Bedingungen).

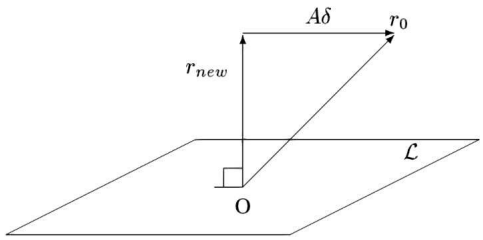


Abbildung 6.1.: Interpretation der Orthogonalitätsbedingung

6.2.1. Allgemeine Projektionsverfahren

Ein Projektionsverfahren *auf einen Unterraum $\mathcal{K}$ und orthogonal zu $\mathcal{L}$* ist ein Prozess, der eine approximierende Lösung $\bar{x}$ an (6.1) findet, indem die Bedingungen gestellt werden, dass einerseits $\bar{x} \in \mathcal{K}$ und andererseits das Residuum (6.2) orthogonal zu $\mathcal{L}$ ist:

$$\text{Finde } \bar{x} \in \mathcal{K}, \text{ so dass } b - A\bar{x} \perp \mathcal{L}.$$

Wenn eine Startlösung x_0 (ein sogenannter *initial guess*) zu $\bar{x}$ vorliegt, dann muss die Lösung im affin linearen Raum $x_0 + \mathcal{K}$ gesucht werden. In diesem Fall wird die Approximationsaufgabe formuliert als:

$$\text{Finde } \bar{x} \in x_0 + \mathcal{K}, \text{ so dass } b - A\bar{x} \perp \mathcal{L}.$$

In diesem Fall kann $\bar{x} = x_0 + \delta$, $\delta \in \mathcal{K}$, geschrieben werden und das Anfangsresiduum r_0 ist definiert als

$$r_0 = b - Ax_0.$$

Die Bedingung $b - A\bar{x} \perp \mathcal{L}$ formt sich zu $b - A(x_0 + \delta) \perp \mathcal{L}$ bzw. $r_0 - A\delta \perp \mathcal{L}$ um. Mit anderen Worten kann die approximierende Lösung definiert werden als:

$$\text{Finde } \bar{x} = x_0 + \delta, \ \delta \in \mathcal{K}, \text{ so dass } (r_0 - A\delta, w) = 0 \text{ für alle } w \in \mathcal{L}.$$

Die Orthogonalitätsbedingung, die an das neue Residuum $r_{new} = r_0 - A\delta$ gestellt wird, ist in Abbildung 6.1 veranschaulicht.

Dies ist ein grundlegender Projektionsschritt in seiner allgemeinsten Form. Typischerweise wird in jedem Projektionsschritt ein neues Paar von Unterräumen $\mathcal{K}$ und $\mathcal{L}$ gewählt und der Startwert x_0 als die neueste Approximation aus dem letzten Projektionsschritt. Auf diese Weise ergibt sich auf natürliche Weise ein Iterationsverfahren zur Lösung des linearen Gleichungssystems (6.1).

Bemerkung 6.2 Die Orthogonalitätsbedingung stellt eine Analogie zur *Galerkin-Orthogonalität*

$$a(u - u_V, \varphi) = 0 \quad \forall \varphi \in V$$

des Fehlers bei der Variationsformulierung elliptischer Differentialgleichungen zweiter Ordnung dar. Das dortige Analogon zu $\mathbb{R}^n$ war ein Hilbertraum H und $\mathcal{K} = \mathcal{L}$ kann mit $V \subset H$ identifiziert werden, wobei V ein abgeschlossener, linearer Unterraum von H ist.

Man kann obiges Projektionsverfahren auch mit Hilfe von Matrizen darstellen. Sei $\mathcal{K}_m = \text{span}\{v_1, \ldots, v_m\}$ mit $v_i \in \mathbb{R}^n$, $1 \leq i \leq m$, $V_m := [v_1, \ldots, v_m] \in \mathbb{R}^{n \times m}$, sowie $\mathcal{L}_m = \text{span}\{w_1, \ldots, w_m\}$, $W_m = [w_1, \ldots, w_m] \in \mathbb{R}^{n \times m}$, $i \leq i \leq m$. Damit lautet die Orthogonalitätsbedingung:

$$\text{Suche } \bar{x} \in \mathcal{K}_m \text{ mit } (b - A\bar{x}, w) = 0 \text{ für alle } w \in \mathcal{L}_m.$$

$\bar{x}$ ist als eine Linearkombination der v_i darstellbar, d.h. $\bar{x} = V_m y$ mit einem $y \in \mathbb{R}^m$. Da $\dim \mathcal{L}_m < \infty$, genügt es, die Orthogonalitätsbedingung bezüglich aller Basisvektoren von $\mathcal{L}_m$ zu prüfen und man erhält:

$$\text{Finde } \bar{x} \in \mathcal{K}_m, \text{ so dass}$$

$$W_m^\top b - W_m^\top A V_m y = 0 \quad \Leftrightarrow \quad H_m y = W_m^\top b,$$

$$\text{wobei } H_m = W_m^\top A V_m.$$

Somit besteht die Aufgabe bei der Konstruktion von Projektionsverfahren darin, geeignete Räume $\mathcal{K}_m$ und $\mathcal{L}_m$ zu definieren, so dass das obige Gleichungssystem einfach und schnell lösbar ist.

Im Folgenden werden Verfahren betrachtet, denen eine spezielle Konstruktion des Unterraums $\mathcal{K}$ zugrunde liegt.

6.3. Krylov-Unterraum-Verfahren

Eine *Krylov-Unterraum-Methode* ist ein (Projektions-)Verfahren, bei dem im m-ten Projektionsschritt der Unterraum $\mathcal{K}_m$ der Dimension m als der *Krylov-Unterraum*

$$\mathcal{K}_m(A, r_0) = \text{span}\left\{r_0, Ar_0, A^2 r_0, \ldots, A^{m-1} r_0\right\} \tag{6.3}$$

gewählt wird. Wenn es zu keinen Missverständnissen führen kann, schreiben wir auch kurz $\mathcal{K}_m$ statt $\mathcal{K}_m(A, r_0)$. Die Größe des Krylov-Unterraums wächst also in jedem Schritt des Approximationsprozesses. Die verschiedenen Krylov-Unterraum-Verfahren unterscheiden sich nun in der Wahl des Raumes $\mathcal{L}_m$, der ebenfalls die Dimension m hat und somit mit jedem Approximationsschritt wächst.

Bemerkung 6.3 Somit hat jedes $x \in \mathcal{K}_m(A, r_0)$ die Form

$$x = \sum_{i=1}^{m} \alpha_i \left(A^{i-1} r_0\right), \quad \alpha_i \in \mathbb{R}, \quad 1 \leq i \leq m.$$

Dies ist äquivalent und etwas abstrakter darstellbar als

$$x = p_m(A) r_0, \quad p \in \mathbb{P}_{m-1},$$

d.h.: Die Lösung ist darstellbar als das Produkt eines Polynoms einer Matrix mit einem Vektor. Somit besteht die Aufgabe bei der Konstruktion von Krylov-Unterraum-Verfahren darin, geeignete Polynome p_m zu finden, die die inverse Matrix $A^{-1} \approx p_m(A)$ in geeigneter Weise approximieren.

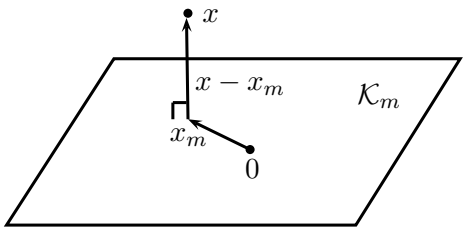


Abbildung 6.2.: Interpretation der optimalen Approximation für symmetrische, positiv definite Gleichungssysteme

6.3.1. Das Verfahren der Konjugierten Gradienten

Zunächst wird der wichtige Spezialfall betrachtet, dass A symmetrisch und positiv definit ist und $\mathcal{K}_m = \mathcal{L}_m$.

Bemerkung 6.4 $H_m = V_m^\top A V_m$ ist symmetrisch.

Satz 6.5 *Die Matrix H_m ist* tridiagonal.

Beweis Aufgrund der Struktur von $H_m = W_m^\top A V_m$ ist H_m eine *Hessenberg-Matrix*, d.h. H_m ist eine untere Dreiecksmatrix mit einer Nebendiagonalen oberhalb der Diagonalen. Da H_m symmetrisch ist, ist H_m offenkundig tridiagonal. □

Das reduzierte Problem $H_m y = V_m^\top b$ führt zu einem linearen System, für welches eine Tridiagonalmatrix invertiert werden muss.

$$H_m = \begin{bmatrix} \alpha_1 & \beta_1 & & 0 \\ \beta_1 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_{m-1} \\ 0 & & \beta_{m-1} & \alpha_m \end{bmatrix}$$

Die LU-Zerlegung für eine solche Matrix ist in $\mathcal{O}(m)$ berechenbar.

Die Orthogonalitätsbedingung für einen Projektionsschritt

$$0 = (b - Ax_m, w) = (x - x_m, w)_A \quad \forall w \in \mathcal{K}_m \tag{6.4}$$

bedeutet in diesem Fall, dass der Defekt $x - x_m$ A-orthogonal zu $\mathcal{K}_m$ ist, d.h. bezüglich des durch A induzierten Skalarproduktes ist x_m die *optimale* Approximation an die Lösung x, die in $\mathcal{K}_m$ gefunden werden kann (Abbildung 6.2).

Nach Konstruktion ist stets

$$\begin{aligned} r_m &= b - Ax_m \\ &= r_0 - r_0 + b - Ax_m \\ &= r_0 + A(x_0 - x_m) \\ &\in r_0 + A\mathcal{K}_m(A, r_0) \\ &\subset \mathcal{K}_{m+1}(A, r_0). \end{aligned}$$

Da nach Konstruktion $r_m \perp \mathcal{K}_m$, ist also stets

$$(r_m, r_i) = 0, \quad i = 0, \ldots, m-1.$$

Weiterhin folgt im Fall $A^m r_0 \in \mathcal{K}_m(A, r_0)$, d.h. $A^m r_0$ ist als Linearkombination von

$$\{r_0, Ar_0, A^2 r_0, \ldots, A^{m-1} r_0\}$$

darstellbar, notwendigerweise, dass $r_m = 0$ bzw.

$$Ax_m = b$$

ist.

Das *Verfahren der Konjugierten Gradienten* (*engl.:* conjugate gradient method, CG method) erzeugt ausgehend von (6.4) A-orthogonale Projektionsrichtungen, die eine Basis des Krylov-Unterraums $\mathcal{K}_m(A, r_0)$ bilden.

Algorithmus 6.6 (CG-Verfahren)

1. Startwert $x_0 \in \mathbb{R}^n$, Anfangsresiduum $r_0 = b - Ax_0$, $d_0 = r_0$.
2. Für $k \geq 0$:

   a)
   $$\alpha_k = \frac{\|r_k\|_2^2}{(Ad_k, d_k)}$$

   b)
   $$x_{k+1} = x_k + \alpha_k d_k$$

   c)
   $$r_{k+1} = r_k - \alpha_k A d_k$$

   d)
   $$\beta_k = \frac{\|r_{k+1}\|_2^2}{\|r_k\|_2^2}$$

   e)
   $$d_{k+1} = r_{k+1} + \beta_k d_k$$

Bemerkung 6.7

1. Bei genauer Betrachtung sieht man, dass das CG-Verfahren nichts anderes macht, als eine Gauß-Elimination der Matrix H_m durchzuführen.
2. In der Terminologie der Polynome betrachtet ist dies wiederum nichts anderes als das Orthogonalisieren der Polynome p_m, was für orthogonale Polynome *immer* mit einer 3-er Rekursion durchgeführt werden kann.

Satz 6.8 (CG-Konvergenz) *Das CG-Verfahren bricht für jeden Startwert $x_0 \in \mathbb{R}^n$ nach spätestens $n-1$ Schritten mit $x_k = x$ ab. Für $0 \le k \le n-1$ gilt die Fehlerabschätzung*

$$\left\| e^{(k)} \right\|_A \le 2 \left(\frac{1 - \frac{1}{\sqrt{\kappa}}}{1 + \frac{1}{\sqrt{\kappa}}} \right)^k \left\| e^{(0)} \right\|_A, \quad k \ge 0,$$

mit der Spektralkondition

$$\kappa := \kappa_2(A) := \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

Zur Reduzierung des Anfangsfehlers um den Faktor ε sind höchstens

$$t(\varepsilon) \le \frac{1}{2}\sqrt{\kappa} \ln\left(\frac{2}{\varepsilon}\right) + 1$$

Iterationsschritte erforderlich.

Beweis Die Optimalitätsbedingung (6.4) kann äquivalent formuliert werden als

$$\begin{aligned} \|x_k - x\|_A &= \min_{y \in x_0 + \mathcal{K}_k(A, r_0)} \|y - x\|_A \\ &= \min_{p \in \mathbb{P}_{k-1}} \|x_0 - x + p(A) r_0\|_A . \end{aligned}$$

Wegen $r_0 = b - Ax_0 = A(x - x_0)$ folgt weiterhin, dass

$$\begin{aligned} \|x_k - x\|_A &= \min_{p \in \mathbb{P}_{k-1}} \|(I - p(A)A)(x_0 - x)\|_A \\ &\le \min_{p \in \mathbb{P}_{k-1}} \|I + Ap(A)\|_A \, \|x_0 - x\|_A \\ &\le \min_{p \in \mathbb{P}_k,\ p(0)=1} \|p(A)\|_A \left\| e^{(0)} \right\|_A \end{aligned}$$

mit dem (durch das vorgegebene x_0 konstanten) Anfangsfehler $e^{(0)}$ gilt, wobei

$$\|p(A)\|_A = \sup_{y \in \mathbb{R}^n,\ y \neq 0} \frac{\|p(A)y\|_A}{\|y\|_A}$$

die durch A induzierte Norm des (matrixwertigen) Polynoms p bezeichnet (siehe dazu auch Abschnitt B.1.3.1). Diese Norm soll zur Abschätzung des Approximationsfehlers im folgenden nach oben abgeschätzt werden.

Seien $\{w_1, \ldots, w_n\}$ die (orthonormalen) Eigenvektoren von A. Diese bilden eine Basis des $\mathbb{R}^n$. Dann hat jedes $y \in \mathbb{R}^n$ eine Darstellung

$$y = \sum_{i=1}^{n} \gamma_i w_i \quad \text{mit} \quad \gamma_i = (y, w_i) .$$

Es folgt, dass

$$\begin{aligned}
\|p(A)y\|_A^2 &= \left\| p(A) \sum_{i=1}^{n} \gamma_i w_i \right\|_A^2 \\
&= \left(p(A) \sum_{i=1}^{n} \gamma_i w_i \right)^\top A \left(p(A) \sum_{j=1}^{n} \gamma_j w_j \right) \\
&= \left(\sum_{i=1}^{n} \gamma_i p(A) w_i \right)^\top A \left(\sum_{j=1}^{n} \gamma_j p(A) w_j \right) \\
(Aw_j = \lambda_j w_j) \Rightarrow \quad &= \left(\sum_{i=1}^{n} \gamma_i p(\lambda_i) w_i \right)^\top A \left(\sum_{j=1}^{n} \gamma_j p(\lambda_j) w_j \right) \\
&= \left(\sum_{i=1}^{n} \gamma_i p(\lambda_i) w_i \right)^\top \left(\sum_{j=1}^{n} \gamma_j p(\lambda_j) A w_j \right) \\
&= \left(\sum_{i=1}^{n} \gamma_i p(\lambda_i) w_i \right)^\top \left(\sum_{j=1}^{n} \lambda_j \gamma_j p(\lambda_j) w_j \right) \\
&= \sum_{i,j=1}^{n} \lambda_j \gamma_i \gamma_j p(\lambda_i) p(\lambda_j) \left(w_i, w_j \right)_{\mathbb{R}^2} \\
((w_i, w_j)_{\mathbb{R}^2} = \delta_{i,j}) \Rightarrow \quad &= \sum_{i=1}^{n} \lambda_i p(\lambda_i)^2 \gamma_i^2 \\
&\leq M^2 \sum_{i=1}^{n} \lambda_i \gamma_i^2 \\
&\leq M^2 \|y\|_A^2 ,
\end{aligned}$$

wobei

$$M := \sup_{\lambda \leq \mu \leq \Lambda} \|p(\mu)\| , \quad \lambda = \lambda_{\min}(A), \quad \Lambda = \lambda_{\max}(A).$$

Daraus folgt die Abschätzung

$$\|p(A)\|_A \leq M.$$

Somit haben wir bewiesen, dass

$$\left\| e^{(k)} \right\|_A \leq \min_{p \in \mathbb{P}_k,\ p(0)=1} \left\{ \sup_{\lambda \leq \mu \leq \Lambda} |p(\mu)| \right\} \left\| e^{(0)} \right\|_A .$$

Die Lösung dieses "min-max"-Problems ist gegeben durch

$$\bar{p}(\mu) = T_k \left(\frac{\Lambda + \lambda - 2\mu}{\Lambda - \lambda} \right) \left[T_k \left(\frac{\Lambda + \lambda}{\Lambda - \lambda} \right) \right]^{-1}$$

mit dem k-ten Tschebyscheff-Polynom T_k auf $[-1, 1]$. Dabei ist

$$\sup_{\lambda \leq \mu \leq \Lambda} \bar{p}(\mu) = \left[T_k \left(\frac{\Lambda + \lambda}{\Lambda - \lambda} \right) \right]^{-1} .$$

Aus der Darstellung

$$T_k(\mu) = \frac{1}{2}\left[\left(\mu + \sqrt{\mu^2-1}\right)^k + \left(\mu - \sqrt{\mu^2-1}\right)^k\right], \quad \mu \in [-\infty, \infty],$$

für die Tschbyscheff-Polynome folgt über die Identität

$$\frac{\kappa+1}{\kappa-1} \pm \sqrt{\left(\frac{\kappa+1}{\kappa-1}\right)^2 - 1} = \frac{\kappa+1}{\kappa-1} \pm \frac{2\sqrt{\kappa}}{\kappa-1} = \frac{\left(\sqrt{\kappa}\pm 1\right)^2}{\kappa-1} = \frac{\sqrt{\kappa}\pm 1}{\sqrt{\kappa}\mp 1}$$

die Abschätzung nach unten

$$T_k\left(\frac{\Lambda+\lambda}{\Lambda-\lambda}\right) = T_k\left(\frac{\kappa+1}{\kappa-1}\right) = \frac{1}{2}\left[\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k\right] \geq \frac{1}{2}\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k.$$

Also wird

$$\sup_{\lambda\leq\mu\leq\Lambda} \bar{p}(\mu) \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k.$$

Daraus folgt die Abschätzung der benötigten Iterationen für eine gegebene relative Genauigkeit folgendermaßen:

$$2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{t(\varepsilon)} < \varepsilon \quad \Rightarrow \quad t(\varepsilon) > \ln\left(\frac{2}{\varepsilon}\right)\ln\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^{-1}.$$

Wegen

$$\ln\left(\frac{x+1}{x-1}\right) = 2\left\{\frac{1}{x} + \frac{1}{3}\frac{1}{x^3} + \frac{1}{5}\frac{1}{x^5} + \ldots\right\} \geq \frac{2}{x}$$

ist dies erfüllt für

$$t(\varepsilon) \geq \frac{1}{2}\sqrt{\kappa}\ln\left(\frac{2}{\varepsilon}\right).$$

□

Bemerkung 6.9 Die Anzahl der Iterationen des CG-Verfahrens wächst exponentiell bei Gitterverfeinerung:

Gitter	**Diskretisierungsparameter**	$\kappa(A_h)$	**Iterationen (beispielhaft)**
1	h	$\mathcal{O}(h^{-2})$	100
2	$\frac{h}{2}$	$4\cdot\mathcal{O}(h^{-2})$	≈ 200
3	$\frac{h}{4}$	$16\cdot\mathcal{O}(h^{-2})$	≈ 400

Exkurs [orthogonale Polynome, Tschebyscheff-Polynome, Legendre-Polynome] Wir betrachten die Monome (Polynome)

$$p_0(x) = 1, \quad p_1(x) = x, \quad p_2(x) = x^2, \quad \ldots$$

Orthogonalisiert man diese Polynome bezüglich des Skalarproduktes

$$(p,q) := \int_{-1}^{1} p(x)q(x)\left(\sqrt{1-x^2}\right)^{-1}\mathrm{d}x,$$

erhält man die Tschebyscheff-Polynome (siehe Abbildung 6.3), die somit nach Konstruktion die Eigenschaft

$$(T_k, T_j) = \delta_{k,j}$$

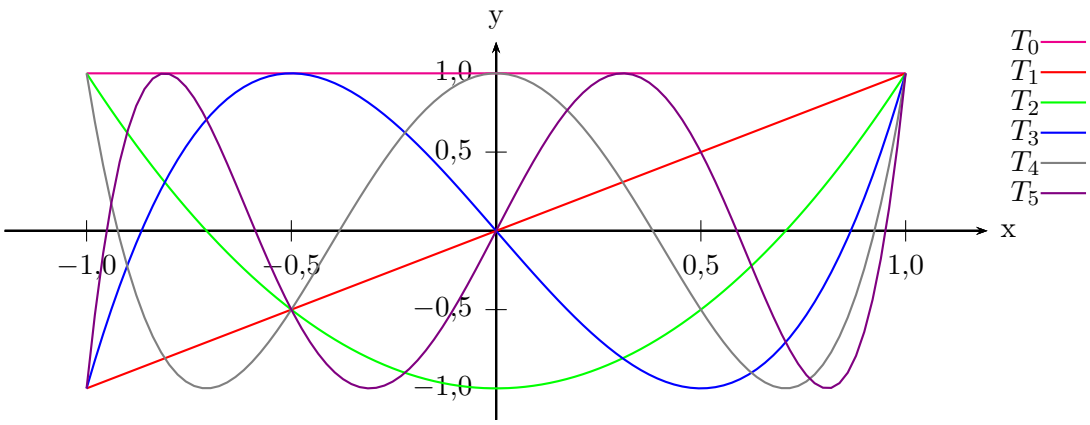


Abbildung 6.3.: Tschebyscheff-Polynome T_0 bis T_5

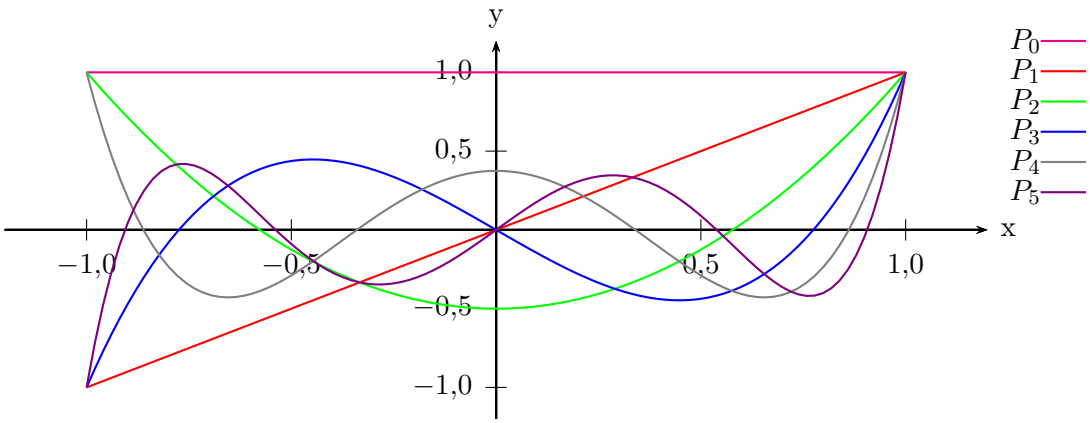


Abbildung 6.4.: Legendre-Polynome P_0 bis P_5

haben.

Bei der Orthogonalisierung bezüglich des Skalarproduktes

$$(p, q) := \int_{-1}^{1} p(x)q(x)\mathrm{d}x$$

erhält man die Legendre-Polynome (siehe Abbildung 6.4).

□

Beispiel 6.10

$$\kappa(A_h) \approx \mathcal{O}(h^{-2}),$$

wobei A_h die Steifigkeitsmatrix des Laplace-Operators zur Gitterweite h bezeichnet.

Exkurs Eingabe x, Ausgabe y.

$$x \mapsto y = f(x)$$
$$x + \delta x \mapsto y + \delta y = f(x + \delta x)$$

Bei der praktischen Lösung von partiellen Differentialgleichungen geht man um mit

- $\|u - u_h\|$, dem Diskretisierungsfehler, und

- $\|u_h - \tilde{u}_h\|$, dem Verfahrensfehler.

Es gilt $\|u - \tilde{u}_h\| \leq \|u - u_h\| + \|u_h - \tilde{u}_h\|$.

Wozu braucht man die Konditionszahl üblicherweise? Wie kann man die Kondition verbessern? Dies beantwortet der folgende

Satz 6.11 (Allgemeiner Störungssatz) *Seien Störungen δA der Matrix A und δb der rechten Seite b gegeben, so dass*

$$\mu := \mathrm{cond}_2(A)\frac{\|\delta A\|}{\|A\|} < 1.$$

Dann gilt die Fehlerabschätzung

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\mathrm{cond}_2(A)}{1-\mu}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right).$$

Keiner der relativen Fehler wird besser als die übliche Maschinengenauigkeit sein, die im Moment typischerweise bei 10^{-16} liegt.

Satz 6.12 *Auf einer Folge von (gleichmäßig) regulären Zerlegungen $\mathcal{T}_h$ gilt für die Spektralkonditionen der symmetrischen und positiv definiten Steifigkeitsmatrizen A_h und der Massematrizen M_h*

$$\begin{aligned} \mathrm{cond}_2(A_h) &= \mathcal{O}(h^{-2}) \text{ und} \\ \mathrm{cond}_2(M_h) &= \mathcal{O}(1). \end{aligned}$$

□

6.4. Mehrgitter-Verfahren

Seien $A_h \in \mathbb{R}^{n\times n}$ die Diskretisierungsmatrix des Laplace-Operators und $\{w_{h,1}, \ldots, w_{h,n}\}$ die (orthonormalen) Eigenvektoren der Matrix A_h. Unser *Ziel* ist es,

$$A_h x_h = b_h$$

zu lösen, wobei $x_h, b_h \in \mathbb{R}^n$. Wir *nehmen an*, dass $\tilde{x}$ die derzeitige Approximation von x_h ist. Wir definieren den *Fehler* als

$$e := x_h - \tilde{x} = \sum_{i=1}^{n} \alpha_i w_{h,i},$$

wobei $\alpha_i \in \mathbb{R}$, $1 \leq i \leq n$. Die Eigenvektoren werden so angeordnet, dass

$$0 < \lambda_i \leq \lambda_j, \quad \text{für} \quad i < j,$$

für die zugehörigen Eigenwerte gilt.

Mit dem Fehler e ist die Lösung x_h darstellbar als

$$x_h = \underbrace{x_h - \tilde{x}}_{=e} + \tilde{x} = \tilde{x} + e,$$

womit das Gleichungssystem $A_h x_h = b_h$ äquivalent in

$$A_h x_h = b_h \quad \Leftrightarrow \quad A_h(\tilde{x} + e) = b_h \quad \Leftrightarrow \quad A_h e = \underbrace{b_h - A\tilde{x}}_{=d_h}$$

transformiert werden kann mit dem *Residuum* d_h. Man kann also genauso gut das lineare Gleichungssystem $A_h e = b_h - A\tilde{x}$ lösen und $x_h = \tilde{x} + \omega e$ berechnen. Im Mehrgitterverfahren wird zur Korrektur einer gegebenen (Start-)Lösung eine approximative Korrektur bzw. ein approximativer Fehler e berechnet!

Bei exakter Rechnung ist $\omega = 1$, aufgrund der Rundungsfehler setzt man in der Regel $\omega \approx 0,9$.

Bemerkung 6.13 (Jacobi-Verfahren als Glätter) $Ax = b \Leftrightarrow (D + U + L)x = b$. Dann ist die Jacobi-Iteration gegeben durch

$$x_k = D^{-1}(b - Lx_{k-1} - Ux_{k-1}).$$

Diese Iteration dämpft schnelle Oszillationen des Fehlers, die mit lokalen Anteilen des Fehlers identifiziert werden können, sehr effektiv, da sie nur Informationen "in der Nachbarschaft" eines Freiheitsgrades berücksichtigt und diese mittels der Inversen der Diagonale der Matrix mittelt.

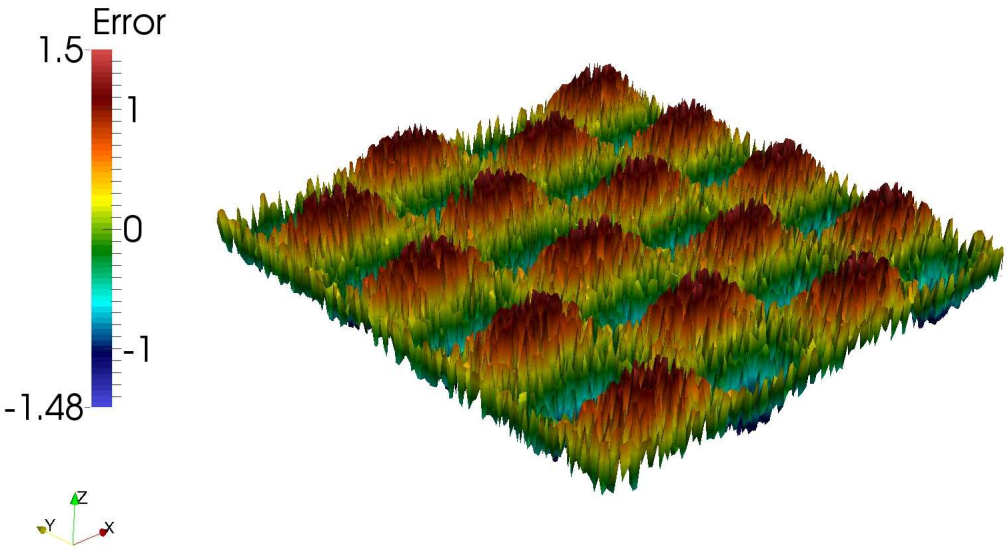
Somit eignet sich das Jacobi-Verfahren sehr gut als *Glätter* (*engl.:* smoother), siehe Abbildung 6.5.

Das Mehrgitter-Verfahren basiert im Wesentlichen auf den folgenden beiden Ideen:

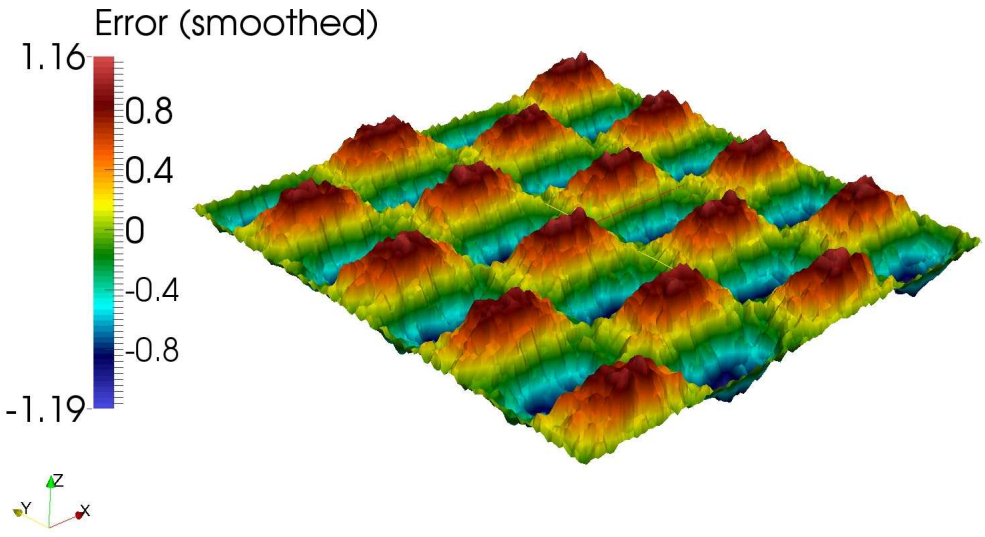
1. Idee:
Klassische iterative Verfahren sind allgemein sehr gute Glätter. Damit lässt sich ein auf einem gegebenen Gitter *bezüglich der Gitterweite h hochfrequenter* Fehler soweit glätten, dass dieser — wiederum bezüglich h — auf diesem Gitter *niederfrequent* wird, vgl. Abbildung 6.5.

2. Idee:
Der auf dem gegebenen Gitter nun glatte bzw. niederfrequente Fehler wird auf ein gröberes Gitter *restringiert*. Dies geschieht nicht mit dem Fehler direkt, da dieser unbekannt ist, sondern das zugehörige Residuum wird zur *Berechnung des (approximativen) Fehlers* auf das gröbere Gitter restringiert. Durch die nun größere Gitterweite, beispielsweise $H := 2h$ bei gleichmäßig verfeinerter Gitterhierarchie, ist der restringierte Fehler bzw. das restringierte Residuum nun *bezüglich H* wiederum hochfrequent, vgl. Abbildung 6.6, so dass das Verfahren mit der 1. Idee nun rekursiv bis zu einem gröbsten Gitter fortgeführt werden kann. Löst man die Gleichung für den *dortigen* Fehler/die *dortige* Korrektur e_{grob} *exakt* oder in geeigneter Weise *approximativ*, so kann diese Korrektur mittels *Prolongation* wieder auf die feineren Gitter übertragen werden und die (approximative) Lösung $\tilde{x}$ auf dem feinsten Gitter schließlich korrigiert werden. Durch die Prolongation entspricht diese Korrektur allerdings nicht der analytisch exakten Korrektur e, so dass dieses Verfahren iterativ mehrmals angewendet werden muss, bis ein hinreichend kleiner Fehler — kontrolliert über das Residuum — erreicht wird.

Bemerkung 6.14 Aus der 2. Idee wird die Effizienz des Mehrgitterverfahrens deutlich. Zum einen findet durch die Grobgitterkorrektur auf dem gröbsten Level eine schnelle *Informationsausbreitung* über die gesamte Triangulierung $\mathcal{T}_h$ hinweg statt. Zum anderen ist auf den gröberen Gittern der Aufwand für sämtliche auftretenden Operationen deutlich geringer als auf dem feinsten Gitter. Da die meiste "Arbeit" durch die rekursiven Aufrufe auf den gröberen Gittern ablaufen, vgl. Abbildung 6.7, ist der Gesamtaufwand im Vergleich zu der Anzahl der mathematisch notwendigen Schritte (Glättung, Restriktion, Prolongation auf den verschiedenen Leveln, Lösen auf dem gröbsten Gitter) gemessen in Rechenoperationen und -zeit gering.

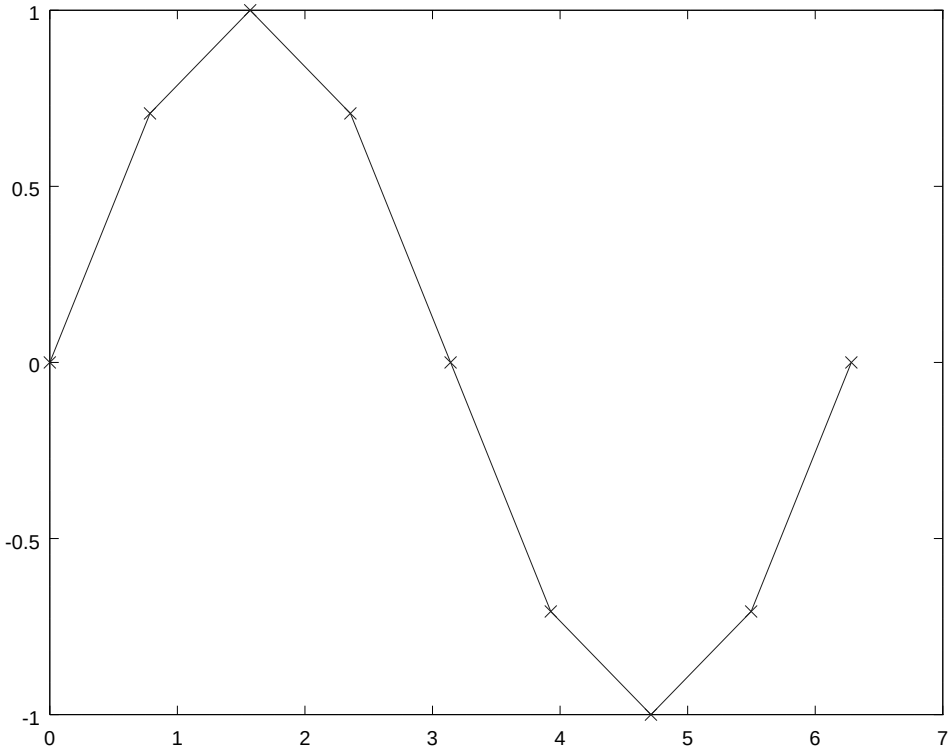


(a) Fehler vor der Glättung

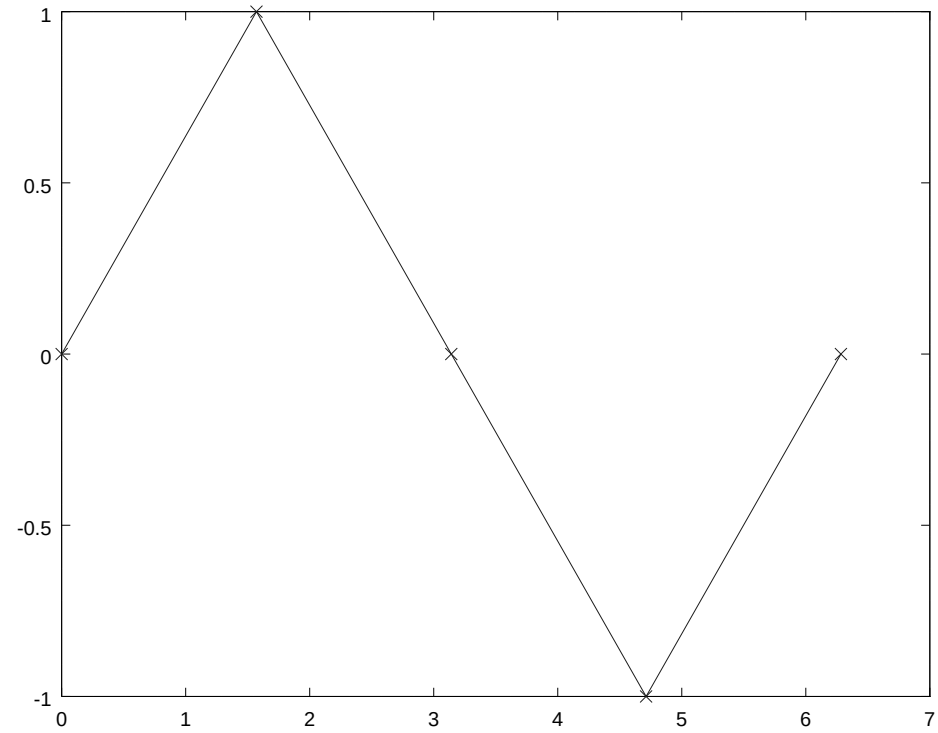


(b) Fehler nach der Glättung

Abbildung 6.5.: Fehler vor Anwendung des Glätters und nach drei Glättungsschritten mit dem gedämpften Jacobi-Verfahren



(a) Niederfrequente Funktion auf einem feinen Gitter



(b) Dieselbe Feunktion restringiert auf ein gröberes Gitter. Die Funktion ist hier hochfrequent.

Abbildung 6.6.: Funktion auf einem feinen und einem gröberen Gitter in 1D

Mathematisch präzise kann der Mehrgitteralgorithmus mit Hilfe von entsprechenden Operatoren geschrieben werden. Der Einfachheit und Lesbarkeit halber wird im Folgenden zunächst das *Zweigitterverfahren* beschrieben, das sich auf ein feines und ein gröberes Gitter beschränkt. Der allgemeine Fall wird im Anschluss diskutiert.

Wir betrachten eine Folge von Gittern $\mathcal{T}_l = \mathcal{T}_{h,l}$, $l = 0, \ldots, L$, zunehmender Feinheit $h_0 > h_1 > \cdots > h_l > \cdots > h_L$ sowie zugehörige FE-Räume $V_l = V_{h,l}$. Der Einfachheit halber sei angenommen, dass die FE-Räume hierarchisch geordnet sind: $V_0 \subset V_1 \subset \cdots \subset V_L$. Zwischen den Funktionen $v_l \in V_l$ und den zugehörigen Knotenwertvektoren $y_l \in \mathbb{R}^{N_l}$ gilt der übliche Zusammenhang $v_l(a_n) = y_{l,n}$, $n = 1, \ldots, N_l$, mit den Koordinaten a_n des n-ten Freiheitsgrades. Das kontinuierliche Problem und sein FE-Analogon schreiben wir in kompakter Form

$$\begin{aligned} a(u, \varphi) &= (f, \varphi) \quad \forall \varphi \in V \\ a(u_l, \varphi_l) &= (f, \varphi_l) \quad \forall \varphi_l \in V_l \quad \text{(auf Gitter } \mathcal{T}_l\text{)} \end{aligned}$$

Zu den Matrizen $A_l = A_{h_l}$ auf den Gittern $\mathcal{T}_l$ sind Operatoren $\mathcal{A}_l : V_l \to V_l$ assoziiert durch

$$(\mathcal{A}_l v_l, w_l) = a(v_l, w_l) = (A_l y_l, z_l) \quad \forall v_l, w_l \in V_l.$$

Wir führen noch Transferoperatoren

$$\begin{aligned} r_l^{l-1} &: V_l \to V_{l-1} \qquad \text{(Restriktion)}, \\ P_{l-1}^{l} &: V_{l-1} \to V_l \qquad \text{(Prolongation)}, \end{aligned}$$

sowie einen Glättungsoperator

$$S_l : \ V_l \to V_l$$

ein. Die Schreibweise S_l^ν bedeutet, dass die Glättung ν-mal durchgeführt wird. Die schematische Darstellung des Mehrgitterschrittes $u_L^{(k)} \to u_L^{(k+1)}$ lautet

$$\begin{array}{lcll} \text{MG}(L,\, u_L^{(k)})\text{: } u_L^{(k)} & \to & \bar{u}_L^{(k)} = S_L^\nu(u_L^{(k)}) & \text{Glättungsschritt} \\ & \to & d_L = f_L - \mathcal{A}_L \bar{u}_L^{(k)} & \text{Residuum} \\ & \to & \tilde{d}_{L-1} = r_L^{L-1} d_L & \text{Restriktion} \\ & \to & q_{L-1} \cong \mathcal{A}_{L-1}^{-1} \tilde{d}_{L-1} & \text{Lösen auf Level } L-1 \\ & \to & \tilde{q}_L = P_{L-1}^{L} q_{L-1} & \text{Prolongation} \\ & \to & u_L^{(k+1)} = \bar{u}_L^{(k)} + \omega_L \tilde{q}_L & \end{array}$$

Bemerkung 6.15

1. Das obige Schema beschränkt sich auf zwei Level (L und $L-1$); natürlich kann man es auch auf mehr Gitter verallgemeinern: Dazu ersetzt man $q_{L-1} \cong \mathcal{A}_{L-1}^{-1} d_{L-1}$ durch $q_{L-1} = \text{MG}(L-1,\, \tilde{d}_{L-1})$.

2. Bei einem rekursiven Aufruf von MG$(\cdot,\cdot)$ darf man natürlich ein geeignetes Abbruchkriterium (z.B. $L \geq 0$) nicht vergessen!

3. Es ist in der Literatur oftmals auch üblich, nach der Korrektur der Lösung $u_L^{(k+1)} = \bar{u}_L^{(k)} + \omega_L \tilde{q}_L$ einen weiteren Glättungsschritt durchzuführen. Dies ist insbesondere bei mehr als zwei Gitterleveln sinnvoll, um durch die Prolongation verursachte hochfrequente Fehler zu glätten, was im Allgemeinen zu einer verbesserten Konvergenz des Verfahrens in der Praxis führt. Ansonsten würde die nächste Glättung erst im nächsten Mehrgitterschritt erfolgen, wodurch sich in der Zwischenzeit Prolongationsfehler möglicherweise fortpflanzen können.

6.4.1. Konvergenz und Aufwandsanalyse

Der Zweigitterprozess lässt sich in der folgenden Form schreiben:

$$\begin{aligned} u_L^{(k+1)} &= \underbrace{S_L^\nu\left(u_L^{(k)}\right)}_{=\bar{u}_L^{(k)}} + p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}\left(f_L - \mathcal{A}_L \underbrace{S_L^\nu\left(u_L^{(k)}\right)}_{=\bar{u}_L^{(k)}}\right) \\ &= S_L^\nu\left(u_L^{(k)}\right) + p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L\left(u_L - S_L^\nu\left(u_L^{(k)}\right)\right) \end{aligned}$$

Exkurs [Beweisidee für iterative Verfahren] Gegeben seien Näherungslösungen $x_0, x_1, \ldots$. Wir wollen

$$\|x_{k+1} - x_k\| \to 0 \quad (k \to \infty)$$

zeigen. Dies lässt sich oftmals beweisen, indem man das iterative Verfahren in Fixpunktform transformiert und den Beweis mittels der Anwendung eines Fixpunktsatzes führt:

$$\begin{aligned} x_{k+1} &= x_k + B(f - Ax_k) \\ (x - x_{k+1}) &= (x - x_k - BA(x - x_k)) \\ e_{k+1} &= e_k - BAe_k \\ e_{k+1} &= (I - BA)e_k = (I - BA)^{k+1} e_0 \\ \|e_{k+1}\| &\leq \|(I - BA)\|^{k+1} \|e_0\| \end{aligned}$$

Im Allgemeinen wählt man $B \approx A^{-1}$. Konvergenz folgt also insbesondere in dem Fall, dass $\|(I - BA)\|^k < 1$. Dazu bietet sich die *Spektralnorm*

$$\|A\| = \max_{i=1,\ldots,n} \sqrt{|\lambda_i(A)|}$$

an. □

Daraus folgt für den Fehler e_L^{k+1}

$$e_L^{k+1} = \left(I_L - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L\right)\left(S_L^\nu\left(u_L^{(k)}\right) - u_L\right)$$

Die Glättungsoperation ist gegeben in der affin-linearen Form

$$S_L(v_L) := S_L v_L + g_L$$

und erfüllt als Fixpunktiteration die Bedingung

$$S_L(u_L) = u_L.$$

Daraus erschließt man rekursiv, dass

$$S_L^\nu\left(e_L^{(k)}\right) - u_L = S_L\left(S_L^{\nu-1}\left(u_L^{(k)}\right) - u_L\right) = \ldots = S_L^\nu e_L^{(k)}.$$

Mit dem sogenannten Zweigitteroperator

$$ZG_L(\nu) := \left(I_L - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L\right) S_L^\nu$$

gilt daher

$$e_L^{(k+1)} = ZG_L(\nu) e_L^{(k)}.$$

Exkurs [Jacobi-Verfahren als affin-lineare Fixpunktiteration] $A = L+D+U$, $(L+D+U)x = b$. Dann lautet die Jacobi-Iteration

$$x_{k+1} = \underbrace{D^{-1}b}_{=g_L} - \underbrace{D^{-1}(L+U)}_{=S_L} x_k.$$

Das Jacobi-Verfahren ist also eine Fixpunktiteration in affin-linearer Form. □

Satz 6.16 *Für hinreichend häufige Glättung $\nu > 0$ ist der Zweigitteralgorithmus konvergent mit einer bezüglich L* gleichmäßigen *L^2-Konvegenzrate*

$$\|ZG_L(\nu)\| \leq \rho_{ZG(\nu)} := \frac{C}{\nu} < 1.$$

Beweis Wir schreiben

$$ZG_L(\nu) = \left(\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}\right) \mathcal{A}_L S_L^\nu$$

und schätzen ab:

$$\|ZG_L(\nu)\| \leq \left\|\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}\right\| \|\mathcal{A}_L S_L^\nu\|$$

Die erste Norm auf der rechten Seite beschreibt die Qualität der Approximation der Feingitterlösung auf dem gröberen Gitter.

Zu zeigen:

1. Glättungseigenschaft:

$$\|\mathcal{A}_L S_L^\nu v_L\| \leq C_S \nu^{-1} h_L^{-2} \|v_L\| \quad \forall v_L \in V_L$$

2. Approximationseigenschaft:

$$\left\|\left(\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}\right) v_L\right\| \leq C_a h_L^2 \|v_L\| \quad \forall v_L \in V_L$$

Die beiden Aussagen werden im Folgenden bewiesen:

1. Glättungseigenschaft:
   $\mathcal{A}_L$ besitzt reelle, positive Eigenwerte $0 < \lambda_1 \leq \ldots \leq \lambda_i \leq \ldots \leq \lambda_{N_L} =: \Lambda_L$. Mit den zugehörigen L^2-orthonormalen Eigenfunktionen w_i kann man schreiben:

$$v_L = \sum_{i=1}^{N_L} \gamma_i w_i, \quad \gamma_i := (v_L, w_i), \qquad \forall v_L \in V_L.$$

   Für den Richardson-Iterationsoperator

$$S_L = I_L - \theta_L \mathcal{A}_L : \; V_L \to V_L,$$

   wobei

$$\theta_L = \frac{1}{\Lambda_L},$$

gilt dann

$$\mathcal{A}_L S_L^\nu v_L = \sum_{i=1}^{N_L} \gamma_i \lambda_i \left(1 - \frac{\lambda_i}{\Lambda_L}\right)^\nu w_i$$

und folglich

$$\begin{aligned}\|\mathcal{A}_L S_L^\nu v_L\|^2 &= \sum_{i=1}^{N_L} \gamma_i^2 \lambda_i^2 \left(1 - \frac{\lambda_i}{\Lambda_L}\right)^{2\nu} \\ &\leq \Lambda_L^2 \max_{1\leq i\leq N_L} \left\{ \left(\frac{\lambda_i}{\Lambda_L}\right)^2 \left(1 - \frac{\lambda_i}{\Lambda_L}\right)^{2\nu} \right\} \underbrace{\sum_{i=1}^{N_L} \gamma_i^2}_{=\|v_L\|^2}.\end{aligned}$$

Mit Hilfe der Beziehung

$$\max_{0\leq x\leq 1} \left\{x^2(1-x)^{2\nu}\right\} \leq (1+\nu)^{-2}$$

ergibt sich

$$\|\mathcal{A}_L S_L^\nu v_L\|^2 \leq \Lambda_L^2 (1+\nu)^{-2} \|v_L\|^2.$$

Da $\Lambda_L \leq c h_L^{-2}$, liefert diese Ungleichung

$$\|\mathcal{A}_L S_L^\nu\| \leq C_S \nu^{-1} h_L^{-2} \quad \text{für} \quad \nu \geq 1.$$

2. Approximationseigenschaft:
Restriktion: L^2-Projektion.

$$\text{Suche} \quad u \in V_{L-1}: \quad (u,\varphi)_{L^2(\Omega)} = (u_L,\varphi)_{L^2(\Omega)} \ \forall \varphi \in V_{L-1}.$$

$L^2(\Omega)$ ist ein Hilbertraum. Ferner gilt

$$\begin{aligned}&(u,u)_{L^2(\Omega)} = \|u\|^2_{L^2(\Omega)} > 0, \quad \forall u \neq 0, \\ &(u,v)_{L^2(\Omega)} \leq \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \quad \forall u,v \in L^2(\Omega).\end{aligned}$$

Somit erfüllt die L^2-Projektion die Voraussetzungen des Lemmas von Lax-Milgram und stellt folglich ein wohlgestelltes Problem dar.
Prolongation:

$$p_{L-1}^L = Id.$$

Ferner erfüllt der Operator $\mathcal{A}_L : V_L \to V_L$ definitionsgemäß

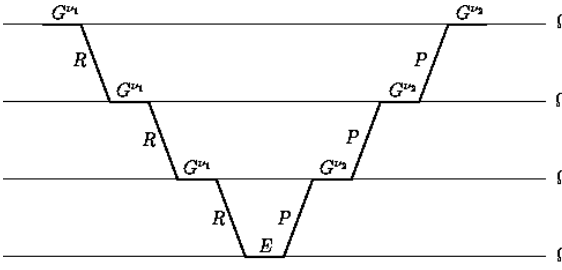
$$(\mathcal{A}_L v_L, \varphi_L) = a(v_L,\varphi_L) \quad v_L, \varphi_L \in V_L.$$

Für ein beliebiges, aber fest gewähltes $f_L \in V_L$ gilt demnach für die Funktionen $v_L = \mathcal{A}_L^{-1} f_L$ und $v_{L-1} = \mathcal{A}_{L-1}^{-1} r_L^{L-1} f_L$

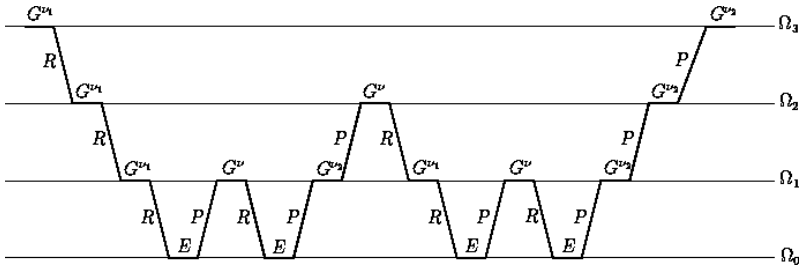
$$\begin{aligned}a(v_L,\varphi_L) &= (f_L,\varphi_L) \quad \forall \varphi_L \in V_L, \\ a(v_{L-1},\varphi_{L-1}) &= \left(r_L^{L-1} f_L, \varphi_{L-1}\right) = (f_L,\varphi_{L-1}) \quad \forall \varphi_{L-1} \in V_{L-1}.\end{aligned}$$

Der Funktion $v_L \in V_L$ wird eine Funktion $v \in H_0^1(\Omega) \cap H^2(\Omega)$ als Lösung der Aufgabe

$$a(v,\varphi) = (f_L,\varphi) \quad \text{für alle} \quad \varphi \in V$$



(a) V-Zyklus



(b) W-Zyklus

Abbildung 6.7.: Verschiedene Mehrgitterzyklen (Quelle: Wikipedia)

zugeordnet. Für diese gelten die Abschätzungen

$$\begin{aligned}
\left\|\nabla^2 v\right\| &\leq C \left\|f_L\right\| \quad \text{(a-priori-Fehlerabschätzung)} \\
\left\|v - v_L\right\| &\leq C h_L^2 \left\|\nabla^2 v\right\| \\
\left\|v - v_{L-1}\right\| &\leq C h_{L-1}^2 \left\|\nabla^2 v\right\|
\end{aligned}$$

Damit erhalten wir wegen $h_{L-1} \leq 2h_L$

$$\begin{aligned}
\left\|v_L - v_{L-1}\right\| &\leq C h_L^2 \left\|\nabla^2 v\right\| \\
&\leq C h_L^2 \left\|f_L\right\|, \\
\left\|\left(\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}\right) f_L\right\| &\leq C h_L^2 \left\|f_L\right\|, \\
\left\|\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}\right\| &\leq C h_L^2.
\end{aligned}$$

□

Für das Mehrgitterverfahren erhalten wir das folgende Resultat:

Satz 6.17 *Es sei angenommen, dass der Zweigitteralgorithmus mit einer L^2-Konvergenzrate $\rho_{ZG} \to 0$ für $\nu \to \infty$ gleichmäßig bezüglich L konvergiert. Dann konvergiert für hinreichend häufige Glättung der Mehrgitteralgorithmus (W-Zyklus, siehe Abbildung 6.7) mit einer von L unabhängigen L^2-Konvergenzrate $\rho_{MG} < 1$,*

$$\left\|u_L - MG\left(L, u_L^{(k)}, f_L\right)\right\| \leq \rho_{MG} \left\|u_L - u_L^{(k)}\right\|.$$

Beweis Der Beweis wird durch Induktion nach dem Gitterlevel L geführt. Es wird nur der W-Zyklus betrachtet un der Einfachheit halber nicht versucht, die auftretenden Konstanten zu optimieren. Sei ν so groß, dass die Konvergenzrate $\rho_{ZG} \leq \frac{1}{8}$. Wir wollen zeigen, dass die

Konvergenzrate des Mehrgitteralgorithmus $\rho_{MG} \leq \frac{1}{4}$, gleichmäßig bezüglich L. Für $L = 1$ ist dies dann offenbar richtig (Zweigitteralgorithmus!). Sei nun auch $\rho_{MG} \leq \frac{1}{4}$ für Gitterlevel $L-1$. Auf Gitterlevel L gilt dann ausgehend von der Iterierten $u_L^{(k)}$ mit der approximativen Lösung q_{L-1}^2 (d.h. nach zweimaliger Anwendung der Grobgitterkorrektur) und der exakten Lösung der Defektgleichung $\hat{q}_{L-1}$ auf Level $L-1$:

$$\begin{aligned} u_L^{(k+1)} &= MG\left(L, u_L^{(k)}, f_L\right) \\ &= ZG\left(L, u_L^{(k)}, f_L\right) + p_{L-1}^L\left(q_{L-1}^{(2)} - \hat{q}_{L-1}\right). \end{aligned}$$

Nach Induktionsvoraussetzung ist (man beachte, dass der Startwert der Mehrgitteriteration auf Level $L-1$ gleich Null ist und $\hat{q}_{L-1} = \mathcal{A}_{L-1}^{-1} r_L^{L-1} d_L$)

$$\begin{aligned} \left\|\hat{q}_{L-1} - q_{L-1}^{(2)}\right\| &\leq \rho_{MG}^2 \left\|\hat{q}_{L-1}\right\| \\ &= \rho_{MG}^2 \left\|\mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L S_L^\nu \left(u_L - u_L^{(k)}\right)\right\|. \end{aligned}$$

Kombination der letzten Beziehungen ergibt für den Iterationsfehler $e_L^{(k)} := u_L^{(k)} - u_L$

$$\left\|e_L^{(k+1)}\right\| \leq \left(\rho_{ZG} + \rho_{MG}^2 \left\|\mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L S_L^\nu\right\|\right) \left\|e_L^{(k)}\right\|.$$

Die Norm rechts ist bereits im Zusammenhang mit der Konvergenz des Zweigitteralgorithmus abgeschätzt worden. Mit dem Zweigitteroperator

$$ZG_L(\nu) = \left(\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}\right) \mathcal{A}_L S_L^\nu$$

gilt

$$\begin{aligned} \mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L S_L^\nu &= S_L^\nu - \left(\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}\right) \mathcal{A}_L S_L^\nu \\ &= S_L^\nu - ZG_L \end{aligned}$$

und somit

$$\begin{aligned} \left\|\mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L S_L^\nu\right\| &\leq \left\|S_L^\nu\right\| + \left\|ZG_L\right\| \\ &\leq 1 + \rho_{ZG} \\ &\leq 2 \end{aligned}$$

Daraus folgt mit Hilfe der Annahme über ρ_{ZG} und der Induktionsannahme, dass

$$\begin{aligned} \left\|e_L^{(k+1)}\right\| &\leq \left(\rho_{ZG} + \rho_{MG}^2 \left\|\mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L S_L^\nu\right\|\right) \left\|e_L^{(k)}\right\| \\ &\leq \left(\rho_{ZG} + 2\rho_{MG}^2\right) \left\|e_L^{(k)}\right\| \\ &\leq \left(\frac{1}{8} + 2\frac{1}{16}\right) \left\|e_L^{(k)}\right\| \\ &\leq \frac{1}{4} \left\|e_L^{(k)}\right\|. \end{aligned}$$

□

7. Parabolische Gleichungen

Ein parabolische partielle Differentialgleichungen mit homogenen Dirichlet-Randbedingungen hat die allgemeine Form

$$\begin{cases} \partial_t u + Lu & = f, \quad \text{in } \Omega \times (0,T) \\ u_{|\partial\Omega} & = 0 \\ u_{|t=0} & = u^0 \end{cases}$$

Dabei ist L ein elliptischer Differentialoperator 2. Ordnung, z.B. $L = -\Delta$, bezüglich der Ortsvariable $x \in \Omega$.

Eine mögliche variationelle Formulierung dieser Gleichung lautet

$$\begin{cases} (\partial_t u, \varphi) + a(u, \varphi) & = (f, \varphi) \quad \forall \varphi \in V, t > 0, \\ u_{|t=0} & = u^0. \end{cases}$$

Dabei wurde angenommen, dass Ω sich nicht mit der Zeit ändert! Andernfalls müsste der Raum V ebenfalls mit der Zeit variieren.

7.1. Diskretisierungsansätze

Bei der Diskretisierung von parabolischen Differentialgleichungen muss der Tatsache Rechnung getragen werden, dass es eine gegenüber den übrigen Variablen ausgezeichnete Variable gibt. In dem Fall der obigen allgemeinen parabolischen Gleichung ist die Zeitvariable t gegenüber der Ortsvariablen x ausgezeichnet in dem Sinne, dass eine Ableitung bezüglich t nur von 1. Ordnung in die Gleichung eingeht, die Ableitungen bezüglich x jedoch auch von 2. Ordnung. Aufgrund der Elliptizität von L erfüllt die obige Gleichung somit die Definition 1.5 einer parabolischen Gleichung.

Die ersten beiden vorgestellten Diskretisierungsansätze tragen dieser Besonderheit Rechnung, indem bei der Diskretisierung Ort und Zeit getrennt behandelt werden — einmal wird zuerst im Ort und dann in der Zeit diskretisiert, einmal genau umgekehrt. Der dritte Ansatz führt die Diskretisierung bezüglich beider Variablen gleichzeitig durch.

7.1.1. Linienmethode

Zunächst wird eine Diskretisierung bezüglich der Ortsvariable vorgenommen., d.h. mit Hilfe eines Finite-Differenzen- oder Finite-Elemente-Ansatzes werden diskrete Funktionen $u_h(t) = u_h(\cdot, t)$ bestimmt aus der Gleichung

$$\begin{cases} \partial_t u_h(t) + \mathcal{A}_h u_h(t) & = f_h(t), \\ u_h(0) & = u_h^0. \end{cases}$$

Dabei ist $\mathcal{A}_h$ eine im Ort diskrete Repräsentierung des Operators L. Diese Aufgabe lautet gemäß der oben vorgeschlagenen variationellen Formulierung in variationeller Form wie folgt:

$$\begin{cases} (\partial_t u_h(t), \varphi_h) + a(u_h(t), \varphi_h) & = (f, \varphi_h), \quad \forall \varphi_h \in V_h \subset V, t \in I \\ u_h(0) & = P_h u^0, \end{cases}$$

wobei $a(\cdot, \cdot)$ eine dem Operator L zugeordnete Bilinearform und P_h ein Projektionsoperator nach V_h ist. Nach Einführung einer Knotenbasis $\{\varphi_h^{(n)}, n = 1, \ldots, N = \dim(V_h)\}$ kann für die Lösung $u_h(t)$ der Ansatz

$$u_h(t) = \sum_{i=1}^{N} U_h^{(i)}(t) \varphi_h^{(i)}$$

gemacht werden und dieses Problem geht in ein System für den Vektor

$$U_h(t) = \left(U_h^{(n)}(t) \right)_{n=1}^{N}$$

der zugehörigen Knoten- bzw. Freiheitsgradwerte $U_h^{(n)}$, $n = 1, \ldots, N$,

$$\begin{cases} M_h \partial_t U_h(t) + A_h U_h(t) & = b_h(t), \quad t \geq 0 \\ U_h(0) & = U_h^0 \end{cases}$$

mit der *Steifigkeitsmatrix*

$$A_h = \left(a(\varphi_h^{(n)}, \varphi_h^{(m)}) \right)_{m,n=1}^{N}$$

und der Massematrix

$$M_h = \left(\left(\varphi_h^{(n)}, \varphi_h^{(m)} \right) \right)_{m,n=1}^{N}$$

über. Dieses Problem nennt man das *semi-diskrete Problem*.

Das semidiskrete Problem stellt ein System gewöhnlicher Differentialgleichungen der Ordnung 1 dar, so dass für dessen Diskretisierung bezüglich der Zeit t theoretisch jedes Verfahren, dass zur Lösung solcher Systeme geeignet ist, in Frage kommt. Beispiele für solche Verfahren sind das explizite und implizite Euler-Verfahren, die Runge-Kutta-Verfahren oder Mehrschritt-Methoden.

Exkurs Für ein abstraktes System von gewöhnlichen Differentialgleichungen 1. Ordnung

$$u'(t) = f(t, u(t))$$

lautet das explizite Euler-Verfahren

$$\frac{u_{k+1} - u_k}{\Delta t} = f(t_k, u_k)$$

und das implizite Euler-Verfahren

$$\frac{u_{k+1} - u_k}{\Delta t} = f(t_{k+1}, u_{k+1}).$$

□

Mit dem expliziten Euler-Verfahren erhält man das diskrete System

$$\begin{aligned} M_h \left(\frac{U_h^{k+1} - U_h^k}{\Delta t} \right) &= -A_h U_h^k + b_h^k \\ \Leftrightarrow \quad M_h U_h^{k+1} &= (M_h - \Delta t A_h) U_h^k + \Delta t b_h^k \end{aligned}$$

Ein solches Problem, das sowohl bezüglich des Ortes als auch der Zeit diskretisiert ist, heißt *(vollständig) diskretes Problem*.

Mit dem impliziten Euler-Verfahren erhält man das diskrete System

$$\begin{aligned} & M_h \left(\frac{U_h^{k+1} - U_h^k}{\Delta t} \right) = -A_h U_h^{k+1} + b_h^{k+1} \\ \Leftrightarrow \quad & \left(M_h + \Delta t A_h \right) U_h^{k+1} = M_h U_h^k + \Delta t b_h^{k+1} \end{aligned}$$

Für Existenz- und Eindeutigkeitsbeweise der Lösungen des semi- und vollständig diskreten Problems sei hier auf die Literatur, beispielsweise [20], verwiesen.

Exkurs

$$\begin{aligned} \kappa(M_h) &= \mathcal{O}(1) \\ \kappa(A_h) &= \mathcal{O}\left(h^{-2}\right) \end{aligned}$$

□

Zur Durchführung einer nicht expliziten, d.h. "impliziten", Methode müssen in jedem Zeitschritt Gleichungssysteme gelöst werden. Die hohe Dimension des Systems N in der Größenordnung $10^3 - 10^8$ impliziert im Hinblick auf die Lösungsökonomie Einschränkungen bei der Wahl der Zeitschrittverfahren. Es kommen in der Regel nur Schemata einfacher Struktur, d.h. mit wenigen Matrix-Vektor-Multiplikationen, und niedriger Ordnung (ungefähr 1 bis 4) in Frage, da sonst die Speicherung aller Matrizen und Vektoren zu teuer wird.

Eine weitere wesentliche Einschränkung besteht in der generischen Steifigkeit des Systems. Die Systemmatrix A_h hat in Abhängigkeit von der Gitterfeinheit h die Kondition

$$\kappa(A_h) = \mathcal{O}\left(h^{-2}\right).$$

Bei *expliziten* Verfahren sind also einschneidende Schrittweitenrestriktionen einzuhalten, welche deren Verwendung in der Regel verbietet. Der (formale) Vorteil der expliziten Verfahren, dass in den einzelnen Zeitschritten keine impliziten Gleichungssysteme zu lösen sind (was bei Finite-Elemente-Diskretisierungen auch nicht ganz stimmt, da wenigsten ein Gleichungssystem mit der Massematrix gelöst werden muss, was aber aufgrund der guten Kondition effizient möglich ist), wird besonders in höheren Raumdimensionen durch die hohe Zahl der durchzuführenden Zeitschritte (besonders bei der Wahl verfeinerter Ortsgitter) schnell amortisiert.

7.1.2. Rothe-Methode

Bei der Rothe-Methode wird die Differentialgleichung als gewöhnliche Differentialgleichung für eine Hilbertraum-wertige Funktion $U(t) \in V$ aufgefasst und zunächst mit einem A-stabilen Verfahren in der Zeit diskretisiert. Bei Verwendung z.B. des impliziten Euler-Schemas ergibt sich eine Folge von speziellen Randwertaufgaben.

$$U^m + kLU^m = U^{m-1} + kf^m \quad (m \geq 1), \qquad U^0(x) = u^0(x).$$

Dabei bezeichnet $k = \Delta t$ die Zeitschrittweite. Diese Probleme werden nun nacheinander auf möglicherweise wechselnden, dem Lösungsverlauf angepassten, Ortsgittern diskretisiert. Das Problem dabei ist der adäquate Transfer der jeweiligen Startlösung U^{m-1} in jedem Zeitschritt vom alten auf das neue Ortsgitter.

Bemerkung 7.1 Hier zeigt sich der systematische Vorteil einer Finite-Elemente-Galerkin-Methode, bei der sich ganz automatisch als *richtige* Wahl die L^2-Projektion von U^{m-1} auf das neue Gitter ergibt. Bei anderen Diskretisierungsverfahren im Ort, beispielsweise Finite-Differenzen- oder Finite-Volumen-Verfahren, gibt es keine derartige *natürliche* und *richtige* Wahl.

7.1.3. Globale Orts-Zeit-Diskretisierung

Eine simultane Diskretisierung auf einem (unstrukturierten) Gitter der ganzen (x,t)-Ebene wird verwendet, z.B. mittels einer Finite-Elemente-Galerkin-Methode in Raum *und* Zeit. Dieser theoretisch durchaus attraktive Ansatz wird bei höher-dimensionalen Problemen wegen der globalen Kopplung aller Unbekannten im Allgemeinen sehr rechenaufwändig.

Bemerkung 7.2 Der hohe Rechenaufwand kann jedoch reduziert werden, indem man Test- und Ansatzräume bezüglich der Zeit geschickt wählt, beispielsweise als stückweise konstante Funktionen. Auf diese Weise wird die Diskretisierung in der Zeit wieder in ein *iteratives* Zeitschrittverfahren überführt. Auf diese Weise lassen sich etliche "klassische" Verfahren rekonstruieren, wobei der Galerkin-Ansatz in der Zeit neue Perspektiven für die theroretische Analyse dieser Verfahren eröffnet.

7.2. Zeitschrittverfahren: Konsistenz und Konvergenz

Exkurs [Satelliten-Problem] Um den Einfluss von Konsistenz, Konvergenzrate und Stabilität eines Zeitschrittverfahrens auf die numerische Lösung von zeitabhängigen Problemen zu verdeutlichen, wird das folgende System von gewöhnlichen Differentialgleichungen betrachtet, das die periodische Bahn eines Satelliten beschreibt, der um die Erde und ihren Mond kreist:

Die Position $(x,y) \in \mathbb{R}^2$ eines Massepunktes (Satelliten) in dem von Erde und Mond aufgebauten Gravitationsfeld genügt dem Differentialgleichungssystem

$$\begin{cases} x'' &= x + 2y' - \hat{\mu}\frac{x+\mu}{[(x+\mu)^2+y^2]^{\frac{3}{2}}} - \mu\frac{x-\hat{\mu}}{[(x-\hat{\mu})^2+y^2]^{\frac{3}{2}}} \\ y'' &= y - 2x' - \hat{\mu}\frac{y}{[(x+\mu)^2+y^2]^{\frac{3}{2}}} - \mu\frac{y}{[(x-\hat{\mu})^2+y^2]^{\frac{3}{2}}} \end{cases}$$

Dabei bezeichnet $\mu = \frac{1}{82.45}$ das Massenverhältnis vom Mond zur Erde und $\hat{\mu} := 1 - \mu$. Die Skalierung ist so gewählt, dass der Abstand 1 in der x-y-Ebene gerade dem (als konstant angenommenen) Abstand von der Erde zum Mond entspricht.

Dieses Modell wird mit den Anfangswerten

$$x(0) = 1.2, \quad y(0) = 0, \quad x'(0) = 0, \quad y'(0) = 1.049357510$$

versehen. Mit diesen Daten erhält man eine periodische Satellitenbahn mit einer Periode von

$$T = 6.192169331.$$

Die Abbildungen 7.1-7.4 zeigen die numerische Lösung dieses Systems mit verschiedenen Zeitschrittverfahren. Die Zeitschrittweite ist bei allen Verfahren $\Delta t = \frac{T}{M}$, wobei M die Anzahl der Zeitschritte bezeichnet.

□

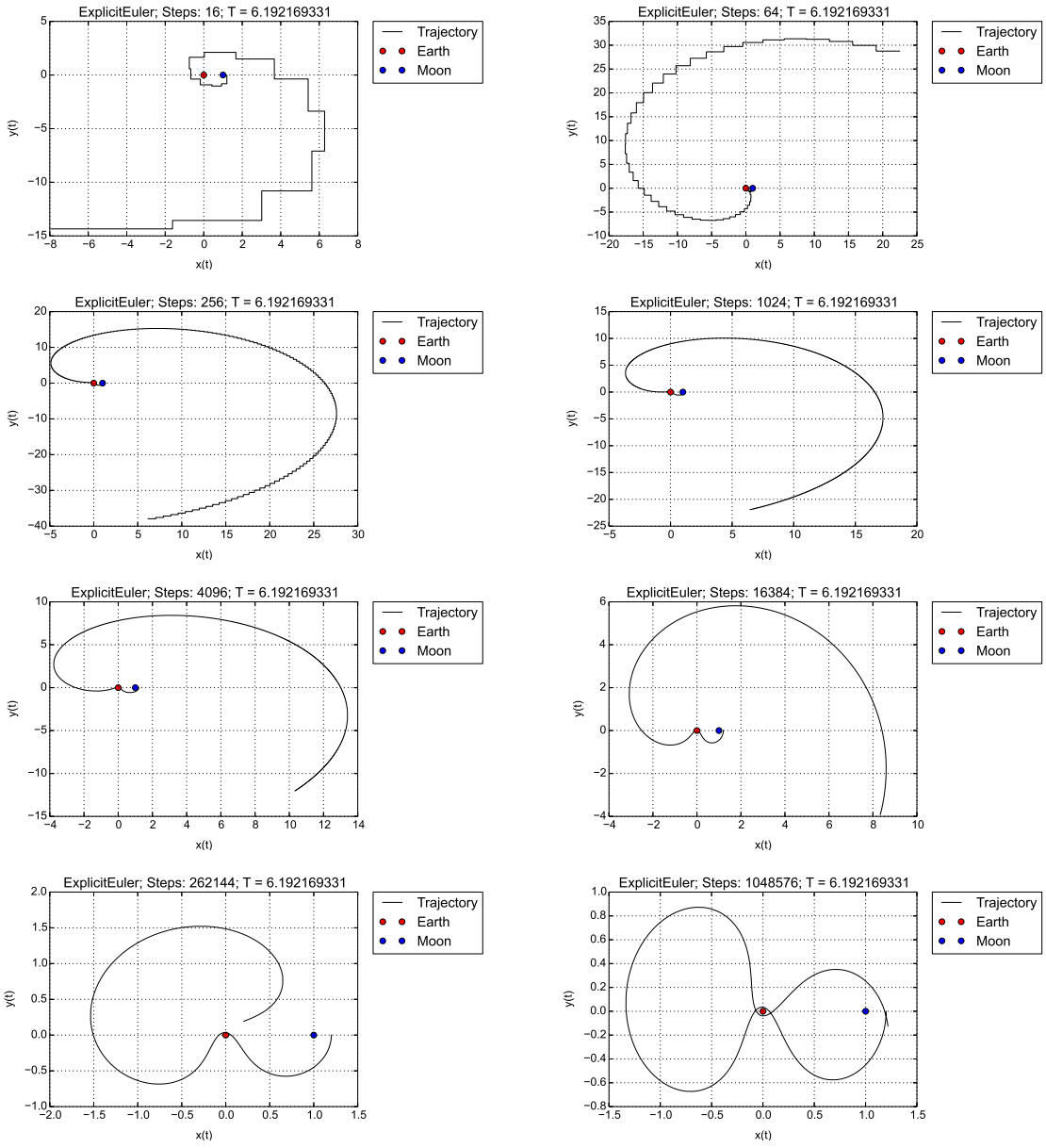


Abbildung 7.1.: Lösung des Satellitenproblems mit dem expliziten Euler-Verfahren (1. Ordnung)

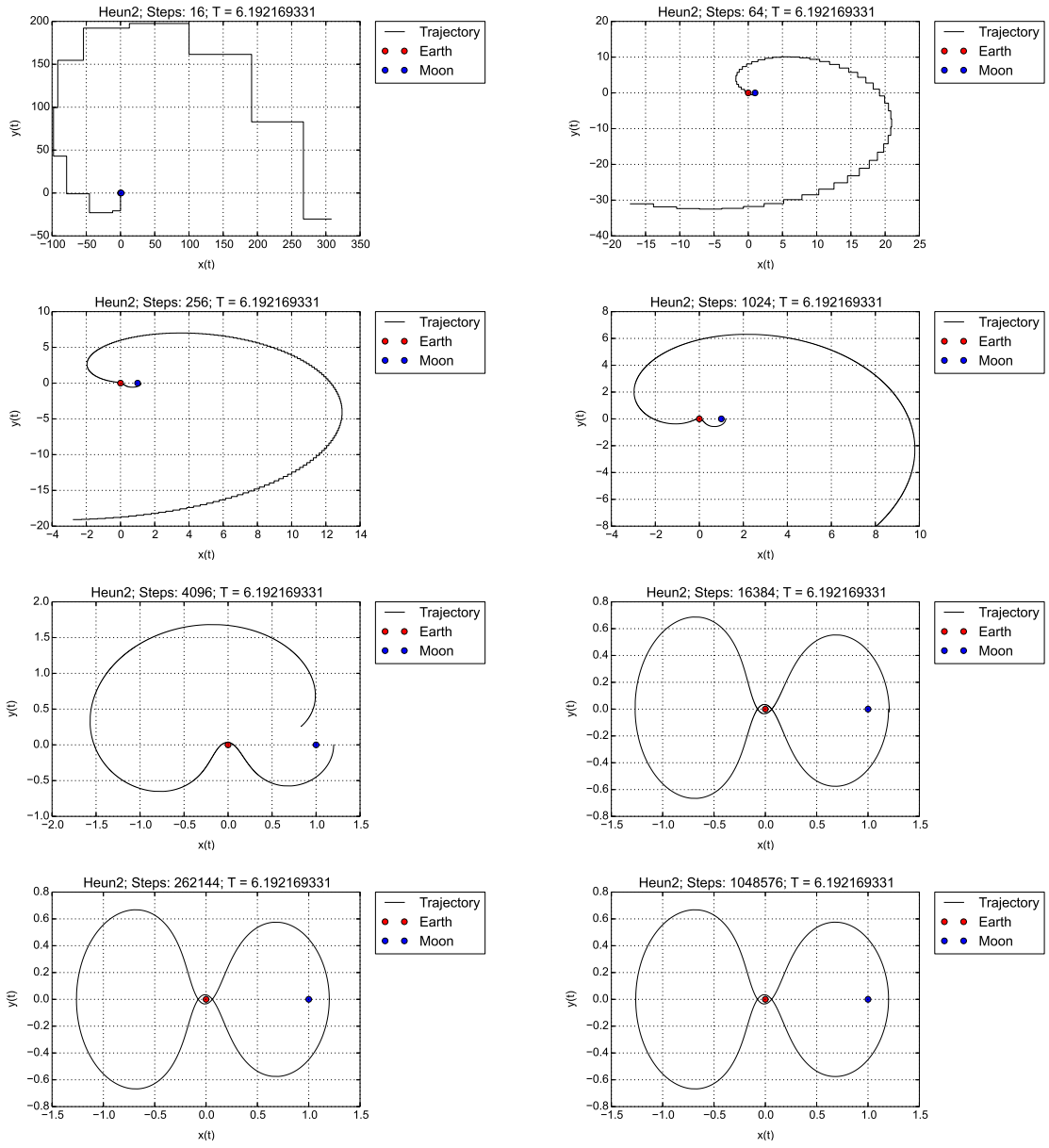


Abbildung 7.2.: Lösung des Satellitenproblems mit dem (expliziten) Verfahren von Heun (2. Ordnung)

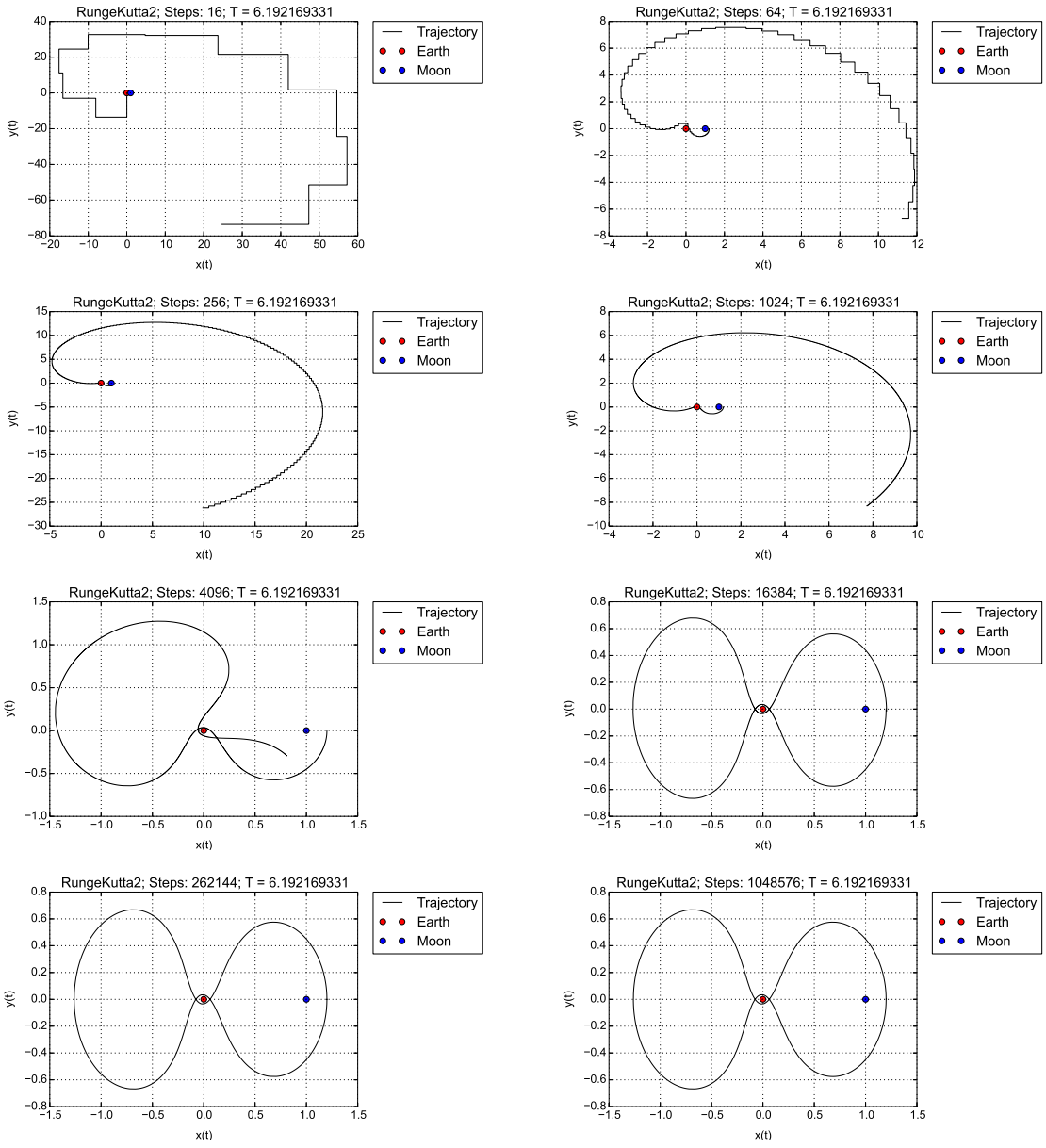


Abbildung 7.3.: Lösung des Satellitenproblems mit dem (expliziten) Runge-Kutta-Verfahren (2. Ordnung)

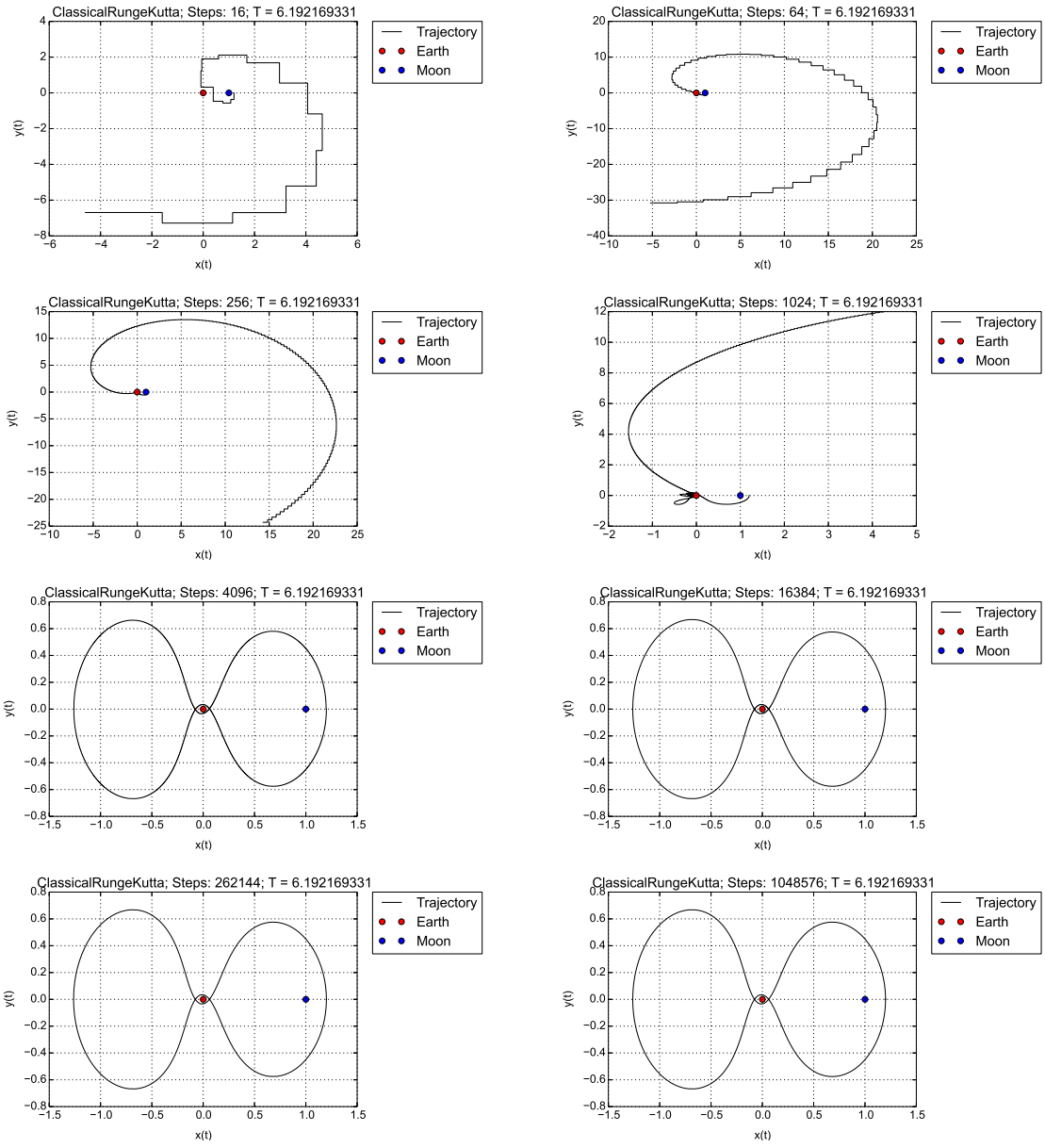


Abbildung 7.4.: Lösung des Satellitenproblems mit dem klassischen (expliziten) Runge-Kutta-Verfahren (4. Ordnung)

7.2.1. Konsistenz

In diesem Abschnitt werden parabolische partielle Differentialgleichungen mit homogenen Dirichlet-Randbedingungen von der Form

$$\begin{cases} \partial_t u + Lu &= f, \quad \text{in } Q_T := \Omega \times I \\ u_{|\partial\Omega} &= 0 \\ u_{|t=0} &= u^0 \end{cases} \tag{7.1}$$

mit $L = -a\Delta$, $a > 0$, $\Omega \subset \mathbb{R}^2$ beschränkt, $I = (0, T)$, betrachtet. Für die Diskretisierung wird die Linien-Methode gewählt und es werden folgende Zeitschrittverfahren diskutiert:

1. Explizites Euler-Verfahren

$$\frac{1}{\Delta t}\left\{U_h^m - U_h^{m-1}\right\} + A_h U_h^{m-1} = f^{m-1}, \quad m \geq 1$$

2. Implizites Euler-Verfahren

$$\frac{1}{\Delta t}\left\{U_h^m - U_h^{m-1}\right\} + A_h U_h^{m} = f^{m}, \quad m \geq 1$$

3. Crank-Nicolson-Verfahren (Trapezregel)

$$\frac{1}{\Delta t}\left\{U_h^m - U_h^{m-1}\right\} + \frac{1}{2}A_h\left(U_h^m + U_h^{m-1}\right) = \frac{1}{2}\Delta t\left(f^m + f^{m-1}\right), \quad m \geq 1$$

Wie bei der Analyse von Differenzenverfahren verwenden wir den Abschneidefehler $\tau_{h,k}^m = (\tau_n^m)_{n=1}^N$ der Differenzenformeln. Diesen erhält man durch formales Auswerten der Differenzenformeln auf der exakten Lösung

$$\Delta t \tau_{h,k}^m := u^m - F\left(u^m, u^{m-1}, u^{m-2}\right).$$

Die Funktion F heißt *Verfahrensfunktion* der Differenzenformel. Bei einer Ortsdiskretisierung der Ordnung p verhält sich der Abschneidefehler dann gemäß

$$\tau_{h,k}^m = \mathcal{O}\left(h^p + (\Delta t)^q\right),$$

wobei q die Ordnung des Zeitschrittverfahrens ist.

Beispiel 7.3 Für das explizite Euler-Verfahren ist die Verfahrensfunktion gegeben durch

$$F\left(u^m, u^{m-1}, u^{m-2}\right) = u^{m-1} - \Delta t A_h u^{m-1} + \Delta t f^{m-1}$$

Satz 7.4 (Konsistenz) *Für die Anfangswertaufgabe (AWA)* (7.1) *genügen die Abschneidefehler den folgenden scharfen Abschätzungen:*

1. *Explizites und implizites Euler-Verfahren*

$$\max_{\bar{Q}_T}\left|\tau_{h,k}^m\right| \leq \max_{\bar{Q}_T}\left|\tau_h^m\right| + \frac{1}{2}\Delta t \max_{\bar{Q}_T}\left|\partial_t^2 u\right|$$

2. *Crank-Nicolson-Verfahren:*

$$\max_{\bar{Q}_T}\left|\tau_{h,k}^m\right| \leq \max_{\bar{Q}_T}\left|\tau_h^m\right| + \frac{1}{12}\left(\Delta t\right)^2 \max_{\bar{Q}_T}\left|\partial_t^3 u\right|$$

Dabei ist $\tau_h^m = \mathcal{O}(h^2)$ der Abschneidefehler der Ortsdiskretisierung.

Beweis Der Abschneidefehler der Ortsdiskretisierung genügt im allgemeinen der Abschätzung

$$|\tau_h^m| = |Lu^m - L_h u^m| \leq Ch^2 M_4^m(u),$$

wobei L_h der Ortsdifferenzenoperator ist und

$$M_4(u) = \max_{\bar{\Omega}} \left|\nabla^4 u^m\right|.$$

Speziell in einer Raumdimension mit $\Omega = (0,1)$ gilt (mit Taylor)

$$|\tau_h^m| = \left|\partial_x^2 u^m - L_h u^m\right| \leq \frac{1}{12} h^2 \max_{[0,1]} \left|\partial_x^4 u^m\right|.$$

Für die explizite Euler-Formel gilt

$$\begin{aligned}
\left|\frac{1}{\Delta t}\left(u^m - u^{m-1}\right) + L_h u^{m-1}\right| &= \left|\frac{1}{\Delta t}\int_{t_{m-1}}^{t_m} \partial_t u \mathrm{d}t + L_h u^{m-1}\right| \\
&= \left|\frac{1}{\Delta t}\int_{t_{m-1}}^{t_m} \partial_t u \mathrm{d}t \underbrace{-\partial_t u^{m-1} - Lu^{m-1}}_{=0} + L_h u^{m-1}\right| \\
&\leq \frac{1}{\Delta t}\left|\int_{t_{m-1}}^{t_m} \left\{\partial_t u - \partial_t u^{m-1}\right\} \mathrm{d}t\right| + \left|Lu^{m-1} - L_h u^{m-1}\right| \\
&\leq \frac{1}{\Delta t}\int_{t_{m-1}}^{t_m} (t - t_{m-1})\,\mathrm{d}t \max_{[t_{m-1},t_m]} \left|\partial_t^2 u\right| + \left|\tau_h^{m-1}\right|,
\end{aligned}$$

woraus

$$\max_{\bar{Q}_T} \left|\tau_{h,k}^m\right| \leq \frac{1}{2}\Delta t \max_{[t_{m-1}tm]} \left|\partial_t^2 u\right| + \left|\tau_h^{m-1}\right|$$

folgt. Der Beweis für die implizite Euler-Formel geht analog.

Für die Crank-Nicolson-Formel gilt

$$\begin{aligned}
&\left|\frac{1}{\Delta t}\left((u^m - u^{m-1})\right) + \frac{1}{2} L_h \left(u^m - u^{m-1}\right)\right| \\
=&\left|\frac{1}{\Delta t}\int_{t_{m-1}}^{t_m} \partial_t u \mathrm{d}t - \frac{1}{2}\left(\partial_t u^m + \partial_t u^{m-1}\right) + \frac{1}{2}\left(Lu^m - L_h u^m\right) + \frac{1}{2}\left(Lu^{m-1} - L_h^{m-1}\right)\right| \\
\leq&\frac{1}{\Delta t}\left|\int_{t_{m-1}}^{t_m} \frac{1}{2}(t - t_m)(t - t_{m-1})\,\mathrm{d}t\right| \max_{[t_{m-1},t_m]} \left|\partial_t^3 u\right| + \frac{1}{2}\left(|\tau_h^m| + \left|\tau_h^{m-1}\right|\right),
\end{aligned}$$

woraus

$$\max_{\bar{Q}_T} \left|\tau_{h,k}^m\right| \leq \max_{\bar{Q}_T} |\tau_h^m| + \frac{1}{12}(\Delta t)^2 \max_{\bar{Q}_T} \left|\partial_t^3 u\right|$$

folgt. □

Im Folgenden sollen allgemeine Ansätze für die Konstruktion von Zeitschrittverfahren für parabolische Probleme diskutiert werden.

Die Lösung von (7.1) besitzt eine explizite Darstellung

$$u(x,t) = \sum_{n=1}^{\infty} u_n^0 v^{(n)}(x) e^{-\lambda_n t}, \quad (x,t) \in Q_T,$$

mit den Eigenwerten und orthonormierten Eigenfunktionen des regulären elliptischen Operators $-a\Delta : V \subset L^2(\Omega) \to L^2(\Omega)$,

$$\begin{aligned} &0 < \lambda_1 \leq \ldots \leq \lambda_n \leq \ldots \quad (n \in \mathbb{N}) \\ &v^{(n)}(x) \in V : \quad -a\Delta v^{(n)} = \lambda_n v^{(n)} \end{aligned}$$

und den Entwicklungskoeffizienten der Startwerte

$$u^{(0)}(x) := \sum_{n=0}^{\infty} u_n^0 v^{(n)}(x), \quad u_n^0 = \left(u^0, v^{(n)}\right)_\Omega .$$

Exkurs [Eigenwerte und Eigenfunktionen eines Operators] Für quadratische Matrizen besteht die Aufgabe der Berechnung der Eigenwerte darin, diejenigen λ zu finden, für welche die Matrix

$$L - \lambda I$$

singulär ist. Analog kann man für Operatoren zwischen Hilbert-Räumen vorgehen. □

Wenn man in die obige Darstellung der Lösung die Reihenentwicklung der Exponentialfunktion einsetzt, erhält man wegen der gleichmäßigen Konvergenz der Reihen

$$\begin{aligned} u(x,t) &= \sum_{n=1}^{\infty} u_n^0 v^{(n)}(x) \left(\sum_{i=0}^{\infty} (-1)^i \frac{\lambda_n^i t^i}{i!} \right) \\ &= \sum_{i=0}^{\infty} (-1)^i \frac{t^i}{i!} \left(\sum_{n=1}^{\infty} u_n^0 \lambda_n^i v^{(n)}(x) \right) \\ &= \sum_{i=0}^{\infty} (-1)^i \frac{t^i}{i!} \left(\sum_{n=1}^{\infty} u_n^0 (-a\Delta)^i v^{(n)}(x) \right) \\ &= \sum_{i=0}^{\infty} (-1)^i \frac{t^i}{i!} (-a\Delta)^i \left(\sum_{n=1}^{\infty} u_n^0 v^{(n)}(x) \right) \\ &= \sum_{i=0}^{\infty} \frac{1}{i!} (at\Delta)^i u^0(x) \\ &= e^{at\Delta} u^0(x). \end{aligned}$$

Exkurs Für Matrizen gilt

$$e^{tA} = I + tA + \frac{(tA)^2}{2!} + \frac{(tA)^3}{3!} + \ldots$$

Auf die gleiche Weise kann man, mit Δ statt A, die Exponentialfunktion des Operators Δ definieren. □

Die Definition der Operatorfunktion $e^{at\Delta}$ über eine konvergente Taylor-Reihe lässt sich auf beliebige analytische Funktionen übertragen, beispielsweise auf Sinus- und Kosinusfunktion. Es

sei betont, dass eine solche kompakte Lösungsdarstellung nur im Fall zeitlich konstanter Koeffizienten a möglich ist. Daraus ergibt sich auf dem diskreten Zeitgitter, mit der Zeitschrittweite k, folgendes Iterationsverfahren:

$$u(\cdot, m) = e^{ak\Delta} u(\cdot, t_{m-1}), \quad m \in \mathbb{N}.$$

Beispiel 7.5 (Berechnung der Exponentialfunktion für Matrizen) Sei A zunächst eine Diagonalmatrix,

$$A = diag\,(\lambda_1, \ldots, \lambda_n)\,.$$

Dann ist

$$\begin{aligned} e^{tA} &= I + diag\,(t\lambda_1, \ldots, t\lambda_n) + diag\left(\frac{t^2\lambda_1^2}{2}, \ldots, \frac{t^2\lambda_n^2}{2}\right) + \ldots \\ &= diag\left(e^{t\lambda_1}, \ldots, e^{t\lambda_n}\right). \end{aligned}$$

Falls A keine Diagonalmatrix ist, kann man versuchen, A zu diagonalisieren. Dies ist auf jeden Fall möglich, falls A symmetrisch ist,

$$A = UDU^\top, \quad U^\top U = I,$$

mit der Diagonalmatrix D der Eigenwerte von A. Dann ist

$$e^{tA} = Ue^{tD}U^\top.$$

Das oben gewonnene Iterationsverfahren legt es nun nahe, den Zeitschritt $t_{m-1} \to t_m$ mit Hilfe einer rationalen Approximation $R(z) \approx e^z$ der Exponentialfunktion der Ordnung $q+1$ anzusetzen,

$$R(z) = \frac{P(z)}{Q(z)} = e^z + \mathcal{O}\left(|z|^{q+1}\right), \quad z \le 0,$$

mit geeigneten Polynomen $P \in \Pi_r$ und $Q \in \Pi_s$, wobei Q auf $z \in \mathbb{R}_-$ keine Nullstelle haben darf. Das Diskretisierungsschema lautet dann

$$U_h^m = R(-kA_h)U_h^{m-1}$$

bzw.

$$Q(-kA_h)U_h^m = P(-kA_h)U_h^{m-1}.$$

Beispiel 7.6 Für die bisher betrachteten Verfahren lautet $R(z)$ wie folgt:

- Explizites Euler-Verfahren:
$$R(z) = 1 + z.$$
- Implizites Euler-Verfahren:
$$R(z) = (1-z)^{-1}.$$
- Crank-Nicolson-Verfahren:
$$R(z) = \left(1 + \frac{1}{2}z\right)\left(1 - \frac{1}{2}z\right)^{-1}.$$

Durch die Ordnungsbedingung

$$e^z Q_{rs}(z) - P_{rs}(z) = \mathcal{O}\left(|z|^{r+s+1}\right), \quad z \leq 0,$$

für den Ansatz $P_{rs} \in \Pi_r$, $Q_{rs} \in \Pi_s$ wird man auf die sogenannten *Padé-Schemata* geführt. Diese sind eindeutig bestimmt und werden gewöhnlich in der sogenannten *Padé-Tafel* dargestellt.

$$\left|\begin{array}{cccc} \frac{1}{1} & \frac{1+z}{1} & \frac{1+z+\frac{1}{2}z^2}{1} & \frac{1+z+\frac{1}{2}z^2+\frac{1}{3!}z^3}{1} & \cdots \\ \frac{1}{1-z} & \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z} & \frac{1+\frac{2}{3}z+\frac{1}{3!}z^2}{1-\frac{1}{3}z} & \frac{1+\frac{3}{4}z+\frac{1}{4}z^2+\frac{1}{24}z^3}{1-\frac{1}{4}z} & \cdots \\ \frac{1}{1-z+\frac{z^2}{2}} & \frac{1+\frac{1}{3}z}{1-\frac{2}{3}z+\frac{1}{6}z^2} & \cdots & & \end{array}\right|$$

Offensichtlich sind alle bisher betrachteten Einschrittverfahren Padé-Formeln und damit in diesem Sinne ordnungsoptimal. Aus der Padé-Tafel erhält man nun weitere Zeitschrittverfahren höherer Ordnung. Dabei kommen aus Effizienzgründen nur die "diagonalen" oder "subdiagonalen" Padé-Schemata in Frage.

Beispiel 7.7

$$\left(I + \frac{1}{3}\Delta t A_h\right) U_h^{m-1} = \left(I - \frac{2}{3}\Delta t A_h + \frac{1}{6}\left(\Delta t\right)^2 A_h^2\right) U_h^m.$$

Hier ist $q = 3$.

Für die weitere Analyse sei angemerkt, dass eine rationale Approximation $R(z)$ der Exponentialfunktion (der Ordnung $r \geq 1$) die Eigenschaft

$$|R(z)| \leq e^{\delta z}, \quad -1 \leq z \leq 0,$$

hat.

Exkurs

$$\partial_t T = a\Delta T.$$

Die rechte Seite ist symmetrisch (symmetrischer Operator!) in der schwachen Formulierung bezüglich der Test- und Ansatzfunktionen.

$$\partial_t T = a\Delta T + \mathbf{v} \cdot \nabla T.$$

Duch den Advektions- bzw. Transportterm ist die rechte Seite nun im Allgemeinen nicht symmetrisch!!! Damit sind die Eigenwerte des Operators im Allgemeinen nicht mehr reell! □

Die Wirkung der Zeitschrittschemata lässt sich mit Hilfe der Spektralzerlegung der Matrix A_h wieder beschreiben durch

$$U_h^m = \sum_{n=1}^{N} U_n^0 R(-k\lambda_n)^m v^{(n)}, \quad m \geq 1,$$

bzw. (mit der Euklidischen Vektornorm)

$$\|U_h^m\|^2 = \sum_{n=1}^{N} \left\|U_n^0\right\|^2 |R(-k\lambda_n)|^{2m}.$$

Definition 7.8 (Stabilitätsbegriffe)

1. A-Stabilität
$$|R(z)| \leq 1 \quad (z \leq 0)$$
sichert die Stabilität der Zeititeration
$$\sup_{m \geq 0} |U_h^m| < \infty.$$

2. Strenge A-Stabilität
$$|R(z)| \leq 1 - c\Delta t \quad (z \leq -1)$$
sichert die Beschränktheit der diskreten Lösung auch im Fall inhomogener rechter Seiten
$$\sup_{m \geq 0} |U_h^m| < c \sup_{m \geq 0} |f^m| .$$

3. *Die* starke A-Stabilität
$$|R(z)| \leq \kappa < 1 \quad (z \leq -1)$$
sichert die exponentielle Dämpfung hochfrequenter Lösungsanteile.

4. *Zur korrekten Wiedergabe von Schwingungsprozessen sollte*
$$R(\pm i) \approx 1$$
sein.

Beispiel 7.9

1. Explizites Euler-Verfahren: Keine Stabilität !
2. Implizites Euler-Verfahren: Stark A-stabil.
3. Crank-Nicolson-Verfahren: A-stabil und $R(\pm i) \approx 1$

Bemerkung 7.10 Alle Algorithmen unterhalb der Diagonale in der Padé-Tafel sind stark A-stabil.

A. Einige spezielle Klassen von Matrizen

A.1. Irreduzible Matrizen

Definition A.1 *Eine Matrix $A = (a_{i,j}) \in \mathbb{K}^{n\times n}$ heißt* reduzibel*, falls Mengen $I, J \subset \{1, \ldots, n\}$ mit folgenden Eigenschaften existieren:*

$$I \neq \emptyset, J \neq \emptyset, I \cap J = \emptyset, I \cup J = \{1, \ldots, n\},$$
$$a_{i,j} = 0 \quad \forall i \in I, j \in J.$$

Andernfalls heißt die Matrix irreduzibel.

Satz A.2 *Für $n \geq 2$ ist eine Matrix $A = (a_{i,j}) \in \mathbb{K}^{n\times n}$ genau dann reduzibel, wenn man die Indizes so anordnen kann, dass A die Blockgestalt*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}$$

mit quadratischen Matrizen A_{11}, A_{22} und nichtleeren Indexteilmengen I_2, J_2 annimmt.

Beweis Übung. □

Bemerkung A.3 Die Lösung eines gegebenen, nichtsingulären Gleichungssystem $Ax = b$ mit einer reduziblen Matrix $A = (a_{i,j})$ lässt sich in zwei kleinere Teilaufgaben zerlegen:

i. Man bestimmt zunächst die Unbekannten x_i $(i \in I)$ des linearen Gleichungssystems

$$\sum_{j=1}^{n} a_{i,j}x_j = \sum_{j\in J} a_{i,j}x_j = b_i \quad (i \in I).$$

ii. Erst dann löst man

$$\sum_{j=1}^{n} a_{i,j}x_j = b_i - \sum_{j\in I} a_{i,j}x_i \quad (i \in I).$$

Beachte: A regulär$\Leftrightarrow$ A_{11} und A_{22} regulär.

Satz A.4 *Eine Tridiagonalmatrix ist irreduzibel genau dann, wenn jedes ihrer Nebendiagonalelemente von Null verschieden ist.*

Beweis Übung. □

Definition A.5 *Eine Matrix $A = (a_{i,j}) \in \mathbb{C}^{n\times n}$ heißt* irreduzibel diagonaldominant*, falls A irreduzibel ist und folgendes gilt:*

$$\sum_{1\le j\le n, j\neq i} |a_{i,j}| \le |a_{i,i}| \quad \forall i = 1,\dots,n$$

und

$$\exists i \in \{1,\dots,n\} : \sum_{1\le j\le n, j\neq i} |a_{i,j}| < |a_{i,i} < .$$

Satz A.6 *Eine irreduzibel diagonaldominante Matrix $A = (a_{i,j}) \in \mathbb{K}^{n\times n}$ ist regulär.*

Beweis Annahme: A singulär. $\exists v \in \mathbb{K}^n \setminus \{0\} : Av = 0$. Dann setze

$$\begin{aligned} I &:= \{i : |v_i| = \|v\|_\infty\} \\ J &:= \{j : |v_j| < \|v\|_\infty\} \end{aligned}$$

Beachte: $I \neq \emptyset$, $I \cap J = \emptyset$ und $I \cup J = \{1,\dots,n\}$.

Zusätzliche Annahme: $J = \emptyset$, also $|v_j| = \|v\|_\infty$ für alle $j \in \{1,\dots,n\}$. Dann

$$\begin{aligned} 0 &= \sum_{j=1}^{n} a_{i,j} v_j \\ -a_{i,i} v_i &= \sum_{1\le j\le n, j\neq i} a_{i,j} v_j \\ |a_{i,i}||v_i| &= \left| \sum_{1\le j\le n, j\neq i} |a_{i,j}||v_j| \right| \\ &\le \sum_{1\le j\le n, j\neq i} |a_{i,j}||v_j|. \end{aligned}$$

Widerspruch zur Def. von „irreduzibel diagonaldominant“. Also $J \neq \emptyset$.

Da die Matrix A irreduzibel ist, existieren Indizes $i_* \in I$ und $j_* \in J$ mit $a_{i_*,j_*} \neq 0$.

$$\begin{aligned} |a_{i_*,i_*}| &\le \sum_{1\le j\le n, j\neq i} |a_{i_*,j}| \underbrace{\frac{|v_j|}{|v_{i_*}|}}_{<1}, \text{ da } Av = 0 \\ &\le \sum_{1\le j\le n} |a_{i_*,j}| \text{ und } a_{i_*,j_*} \neq 0 \end{aligned}$$

Widerspruch zur Def. von „irreduzibel diagonaldominant“.$\Rightarrow$ A regulär. □

Beispiel A.7

$$L_h^{1\mathrm{D}} = \begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix} \text{ ist regulär.}$$

Definition A.8 *Eine Matrix $A = (a_{i,j}) \in \mathbb{C}^{n\times n}$ heißt* diagonaldominant, *falls*

$$\sum_{1\le j\le n, j\neq i} |a_{i,j}| < |a_{i,i}| \quad \forall i = 1,\dots,n.$$

Sie heißt schwach diagonaldominant, *falls*

$$\sum_{1\le j\le n, j\neq i} |a_{i,j}| \le |a_{i,i}| \quad \forall i = 1,\dots,n.$$

Satz A.9 *Sei $A \in \mathbb{C}^{n\times n}$ diagonaldominant oder irreduzibel diagonaldominant. Dann gilt*

$$\rho(D^{-1}B) < 1$$

mit der Zerlegung $A = D - B$, wobei D die Diagonalsubmatrix von A beschreibt. ($\rho(A)$ ist der Spektralradius von A.)

Beweis Hier einfügen. □

Definition A.10 *Für zwei Matrizen $A = (a_{i,j}), B = (b_{i,j}) \in \mathbb{R}^{n\times n}$ schreibt man $A \le B :\Leftrightarrow a_{i,j} \le b_{i,j}$ für $i,j = 1,\dots,n$, bzw. analog $A \ge B$, $A < B$, $A > B$. Eine Matrix heißt nichtnegativ, wenn $A \ge 0$ gilt.*

Lemma A.11 *Seien $A, B \in \mathbb{R}^{n\times n}$ Matrizen und $v, w \in \mathbb{R}^n$ Vektoren. Dann gelten die Implikationen*

$$\begin{aligned} A \ge 0, v \le w &\Rightarrow Av \le Aw, \\ A \ge 0 &\Rightarrow |Av| \le A|v|, \end{aligned}$$

wobei

$$|v| := \begin{pmatrix} |v_1| \\ \vdots \\ |v_n| \end{pmatrix}.$$

Definition A.12 *Eine Matrix $A \in \mathbb{K}^{n\times n}$ heißt* M-Matrix, *falls gilt:*

a) Die Matrix A ist regulär und besitzt eine nichtnegative Inverse $A^{-1} \ge 0$.

b) Alle Einträge der Matrix A, außer denen auf der Diagonalen, sind nicht positiv, d.h. $a_{i,j} \le 0$ für $i \neq j$ aus $i,j = 1,\dots,n$.

Satz A.13 *Eine Matrix $A \in \mathbb{K}^{n\times n}$ ist eine M-Matrix genau dann, wenn*

i. $a_{i,i} > 0$ für $i = 1,\dots,n$,

ii. $a_{i,j} \le 0$ für alle $i,j \in \{1,\dots,n\}$ mit $i \neq j$,

iii. $\rho(D^{-1}B) < 1$ mit der Zerlegung $A = D - B$, wobei D der Diagonalanteil von A ist.

Beweis „⇐“: Sei $\rho(C) < 1$, mit $C = D^{-1}B$. Dann konvergiert die geometrische Reihe

$$S := \sum_{k=0}^{\infty} C^k.$$

Da $D^{-1} \geq 0$ und $B \geq 0$, ergibt sich $C \geq 0$, also $C^k \geq 0$, also $S \geq 0$. Aus

$$I = (I-C)^{-1}(I-C) = \sum_{k=0}^{\infty} C^k(I-C) = S(I-C) = SD^{-1}(D-B) = SD^{-1}A$$

folgt, dass

$$A^{-1} = SD^{-1} \geq 0,$$

also A eine M-Matrix ist.

„⇒“: Sei A eine M-Matrix, $\lambda \in \mathbb{C}$ ein Eigenwert von $D^{-1}B$ und $u \in \mathbb{C}^n \setminus \{0\}$ ein zugehöriger Eigenvektor. Da $D^{-1} \geq 0$ und $B \geq 0$, gilt

$$|\lambda||u| = |\lambda u| = |D^{-1}Bu| \leq D^{-1}B|u|.$$

Daraus folgt, dass

$$-(A^{-1}D)D^{-1}B|u| \leq -(A^{-1}D)|\lambda||u|,$$

da $A^{-1}D \geq 0$. Der Vektor $|u|$ kann folgenderweise abgeschätzt werden:

$$\begin{aligned} |u| &= A^{-1}(D-B)|u| = A^{-1}D(I-D^{-1}B)|u| \\ &\leq A^{-1}D|u| - A^{-1}D|\lambda||u| = (1-|\lambda|)A^{-1}D|u|. \end{aligned}$$

Für den Fall, dass $|\lambda| \geq 1$ gilt, folgt $|u| \leq 0$, d.h. $u = 0$. W! zur Tatsache, dass u Eigenvektor ist. Die Aussage gilt für alle Eigenwerte λ von $D^{-1}B$. Somit ist gezeigt, dass $\rho(D^{-1}B) < 1$. □

Satz A.14 *Sei $A = (a_{i,j}) \in \mathbb{R}^{n\times n}$ eine M-Matrix und $u \in \mathbb{R}^n$ ein Vektor, so dass $Au \geq (1,\ldots,1)^\top$. Dann gilt*

$$\left\|A^{-1}\right\|_\infty \leq \|u\|_\infty .$$

Beweis Sei $v \in \mathbb{R}^n, |v| \leq \|v\|_\infty (1,\ldots,1)^\top \leq \|v\|_\infty Au$. Da A M-Matrix ist, gilt $A^{-1} \geq 0$.

$$|A^{-1}v| \leq A^{-1}|v| \leq \|v\|_\infty A^{-1}Au = \|v\|_\infty u.$$

□

B. Funktionalanalytische Grundlagen

B.1. Normierte, Banach- und Hilbert-Räume

B.1.1. Normierte Räume

Definition B.1 (Normierter linearer Raum) *Sei X ein Vektorraum definiert über einem Körper $\mathbb{K}$ ($= \mathbb{R}$ oder $\mathbb{C}$) und $\|\cdot\|_X : X \to \mathbb{R}$ eine Abbildung. Das Paar $(X, \|\cdot\|_X)$ ist ein* normierter linearer Raum *(und die Abbildung $\|\cdot\|_X$ ist eine* Norm*), falls $\|\cdot\|_X$ die folgenden Eigenschaften für alle $x, y \in X$ und $\alpha \in \mathbb{R}$ erfüllt:*

1. $\|x\|_X \geq 0$ *and* $\|x\|_X = 0 \;\Leftrightarrow\; x = 0$ *(Definitheit)*
2. $\|\alpha x\|_X = |\alpha| \, \|x\|_X$ *(Homogenität)*
3. $\|x + y\|_X \leq \|x\|_X + \|y\|_X$ *(Dreiecksungleichung)*

Beispiel B.2

1. $\mathbb{R}^n$, $\|x\|_{\mathbb{R}^n} = \sqrt{\sum_{i=1}^n x_i^2}$, $x \in \mathbb{R}^n$
2. $L^p(\Omega)$, $\Omega \subset \mathbb{R}^n$, $\|f\|_{L^p(\Omega)} = \left(\int_\Omega |f(x)|^p \, \mathrm{d}x\right)^{\frac{1}{p}}$, $f : \Omega \to \mathbb{R}$, $x \in \Omega$

B.1.2. Vollständigkeit, Banach-Raum

Definition B.3 (Cauchy-Folge, Vollständigkeit, Banach-Raum) *Sei $(X, \|\cdot\|_X)$ ein normierter linearer Raum.*

1. *Eine Folge $(x_k)_{k \in \mathbb{N}}$ in X heißt* Cauchy-Folge*, falls $\|x_k - x_l\|_X \to 0$ für $k, l \to \infty$.*
2. *x heißt* Grenzwert *von $(x_k)_{k \in \mathbb{N}}$, falls $\lim_{k \to \infty} \|x_k - x\|_X = 0$.*
3. *$(X, \|\cdot\|_X)$ heißt* vollständig *oder* Banach-Raum*, falls jede Cauchy-Folge in X einen Grenzwert in X hat.*

Beispiel B.4

1. $\mathbb{R}^n$, $\|x\|_{\mathbb{R}^n} = \sqrt{\sum_{i=1}^n x_i^2}$, $x \in \mathbb{R}^n$
2. $L^p(\Omega)$, $\Omega \subset \mathbb{R}^n$, $\|f\|_{L^p(\Omega)} = \left(\int_\Omega |f(x)|^p \, \mathrm{d}x\right)^{\frac{1}{p}}$, $f : \Omega \to \mathbb{R}$, $x \in \Omega$

B.1.3. (Prä-) Hilbert-Raum

Definition B.5 ((Prä-) Hilbert-Raum)

1. *Sei X ein $\mathbb{K}$-Vektorraum. Eine Abbildung $(\cdot,\cdot) : X \times X \to \mathbb{K}$ heißt* Hermitesche Sesquilinearform*, falls*

   a) $(x,y) = \overline{(y,x)}\ \forall x,y \in X$ *(Hermitesch)*

   b) $(\alpha x, y) = \alpha\,(x,y)\ \forall x,y \in X,\ \alpha \in \mathbb{K}$

   c) $(x, y_1 + y_2) = (x,y_1) + (x,y_2)\ \forall x, y_1, y_2 \in X$

2. *Die Sesquilinearform heißt* positiv semidefinit*, falls $(x,x) \geq 0\ \forall x \in X$, und* positiv definit*, falls $(x,x) \geq 0$ und $(x,x) = 0 \ \Leftrightarrow x = 0$. Eine positiv definite Sesquilinearform heißt* Skalarprodukt*.*

3. *Das Paar $(X, (\cdot,\cdot)$ heißt* Prä-Hilbert-Raum*, falls $(\cdot,\cdot)$ ein Skalarprodukt ist.*

4. *Falls X vollständig bezÃijglich der durch das Skalarprodukt induzierten Norm $\|\cdot\|_X := \sqrt{(\cdot,\cdot)}$ ist, heißt X* Hilbert-Raum*.*

Beispiel B.6

1. $\mathbb{R}^n$, $(x,y)_{\mathbb{R}^n} = \sum_{i=1}^n x_i y_i,\ x,y \in \mathbb{R}^n$
2. $L^2(\Omega),\ \Omega \subset \mathbb{R}^n$, $(f,g)_{L^2(\Omega)} = \int_\Omega f(x)g(x)\mathrm{d}x,\ f,g : \Omega \to \mathbb{R},\ x \in \Omega$

Satz B.7 (Cauchy-Schwarz-Ungleichung) *Sei $(X,(\cdot,\cdot))$ ein Prä-Hilbert-Raum und $\|\cdot\|$ die induzierte Norm. Dann gilt die* Cauchy-Schwarz-Ungleichung *für alle $x,y \in X$:*

$$|(x,y)| \leq \|x\| \cdot \|y\| .$$

Satz B.8 (Parallelogramm-Gleichung) *Sei $(X, \|\cdot\|_X)$ ein normierter linearer Raum.*

1. *Die Norm $\|\cdot\|_X$ wird genau dann von einem Skalarprodukt induziert, falls die* Parallelogramm-Gleichung *für alle $x,y \in X$ gilt:*

   $$\|x+y\|_X^2 + \|x-y\|_X^2 = 2\,\|x\|_X^2 + 2\,\|y\|_X^2 .$$

   Folglich ist ein normierter linearer Raum genau dann ein Prä-Hilbert-Raum, wenn die Parallelogramm-Gleichung gilt.

2. *Ein Banach-Raum $(X, \|\cdot\|_X)$ ist genau dann ein Hilbert-Raum, wenn die Parallelogramm-Gleichung gilt..*

Satz B.9 (Riesz'scher Darstellungssatz) *Sei X ein Hilbertraum, $J : X \to X'$ (X' der* Dualraum *von X) sei definiert als*

$$J(x)(y) = (y,x)_X \quad \textit{für} \quad x,y \in X.$$

J ist ein isometrischer, konjugiert linearer Isomorphismus (d.h. $J(\alpha x))\overline{\alpha}J(x)$).

Bemerkung B.10 Satz B.9 besagt, dass sich jedes Element aus X' durch ein Element aus X darstellen lässt (und umgekehrt).

B.1.3.1. Exkurs: Bedeutung und Interpretation von Normen

Für eine Matrix $A \in \mathbb{R}^{n\times n}$ wird die Norm durch die zugrunde gelegte Norm $\|\cdot\|$ auf $\mathbb{R}^n$ induziert:

$$\begin{aligned}\|A\| &= \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\|Ax\|}{\|x\|} \\ &= \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\sqrt{(Ax, Ax)}}{\|x\|} \\ &= \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\sqrt{(A^\top Ax, x)}}{\|x\|} \\ &= \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\sqrt{(x, x)_A}}{\|x\|},\end{aligned}$$

wobei wir angenommen haben, dass $\|x\|^2 = (x, x)$ und $(x, y)_A := \left(A^\top Ax, y\right)$. Die Matrix A definiert eine Abbildung $A : \mathbb{R}^n \to \mathbb{R}^n$, $x \mapsto Ax$. Die Norm der Abbildung A gibt also an, um welchen Faktor ein Element $x \in \mathbb{R}^n$ *maximal* durch die Anwendung von A auf x gestreckt/ausgedehnt wird.

Folglich entspricht die Norm der Matrix A ihrer *Lipschitz-Konstante*!

Dieses Konzept der Definition einer Norm lässt sich allgemein auf Abbildungen zwischen normierten und Hilbert-Räumen übertragen:

- Im Falle normierter Räume X, Y ist die Norm einer Abbildung $A : X \to Y$ definiert als

$$\|A\| := \sup_{x\in X, x\neq 0} \frac{\|A(x)\|_Y}{\|x\|_X}.$$

- Im Falle eines Hilbertraumes H kann man den Hilbertraum mit seinem eigenen Dualraum identifizieren, siehe Satz B.9. Folglich lässt sich die Norm eines Elementes x in H im Sinne der Dualitätspaarung von H mit sich selbst definieren (dies ist nichts anderes als eine Abbildung von H nach $\mathbb{R}$), denn die Norm wird durch das Skalarprodukt in H induziert:

$$\|x\|_H := \sup_{g\in H, g\neq 0} \frac{(x, g)_H}{\|g\|_H}.$$

Für weitere Details zur Bedeutung und Interpretation des Begriffes einer Norm sei an dieser Stelle auf die Literatur, z.B. [1] oder [27], verwiesen.

B.2. Multiindex-Schreibweise für Ableitungen und Polynome

Definition B.11 (Multiindex)

1. *Ein Vektor der Form $\alpha = (\alpha_1, \ldots, \alpha_n)$, $\alpha_i \in \mathbb{N}_0$, heißt* Multiindex *der* Ordnung

$$|\alpha| := \alpha_1 + \ldots + \alpha_n.$$

2. *Gegeben sei ein Multiindex α. Dann definieren wir*

$$D^\alpha u(x) := \frac{\partial^{|\alpha|} u(x)}{\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n}} = \partial_{x_1}^{\alpha_1} \ldots \partial_{x_n}^{\alpha_n} u(x).$$

3. *Falls k eine nicht-negative ganze Zahl ist, definieren wir die Menge der partiellen Ableitungen der Ordnung k durch* $D^k u(x) := \{D^\alpha u(x) : \quad |\alpha| = k\}$.

4. *Falls $k = 1$ ist, ordnen wir die Elemente von Du in einem Vektor $Du = (\partial_{x_1} u, \ldots, \partial_{x_n} u)$, dem* Gradienten, *an. Falls $k = 2$ ist, ordnen wir die Elemente von $D^2 u$ in einer Matrix $D^2 u = (\partial_{x_i} \partial_{x_j} u)_{i,j=1}^n$, der* Hesse-Matrix, *an.*

Mit der Definition

$$x^\alpha := x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$$

für $x \in \mathbb{R}^d$ und $\alpha \in \mathbb{N}_0^d$ können weiterhin Polynome in d Variablen kompakt und elegant definiert werden.

B.3. Sobolev-Räume

B.3.1. Schwache Ableitungen

Definition B.12 (Schwache Ableitung) *Sei $\Omega \subset \mathbb{R}^n$ offen, $u, v \in L^1_{loc}(\Omega)$ und α ein Multiindex. Wir nennen v die* α-te schwache partielle Ableitung von u, *geschrieben*

$$D^\alpha u = v,$$

falls

$$\int_\Omega u D^\alpha \varphi \mathrm{d}x = (-1)^{|\alpha|} \int_\Omega v \varphi \mathrm{d}x$$

für alle Testfunktionen $\varphi \in C_c^\infty(\Omega)$ gilt.

Bemerkung B.13

- $C_c^\infty(\Omega)$ bezeichnet den Funktionenraum der unendlich oft differenzierbaren Funktionen $\varphi : \Omega \to \mathbb{R}$, die kompakten Träger in Ω haben.
- $L^p(\Omega) = \left\{ u : \Omega \to \mathbb{R} : \quad \|u\|_{L^p(\Omega)} < \infty \right\}$
- $L^p_{loc}(\Omega) = \{u : \Omega \to \mathbb{R} : \quad u \in L^p(V),\ V \subset\subset \Omega\}$ (d.h. V ist eine kompakte Teilmenge von Ω)

Lemma B.14 (Eindeutigkeit schwacher Ableitungen) *Eine schwache α-te partielle Ableitung von u ist, falls sie existiert, fast überall eindeutig definiert.*

B.3.2. Definition von Sobolev-Räumen und elementare Eigenschaften

Definition B.15 (Sobolev-Raum) *Der Sobolev-Raum*

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

besteht aus allen lokal summierbaren Funktionen $u : \Omega \to \mathbb{R}$, so dass für jeden Multiindex α mit $|\alpha| \leq k$, $D^\alpha u$ im schwachen Sinne existiert und zu $L^p(\Omega)$ gehört.

Bemerkung B.16

1. Falls $p = 2$ ist, schreiben wir üblicherweise

$$H^k(\Omega) = W^{k,2}(\Omega) \quad (k = 0, 1, \ldots).$$

Der Buchstabe H wird benutzt, da (wie wir sehen werden) $H^k(\Omega)$ ein Hilbert-Raum ist. Beachte, dass $H^0(\Omega) = L^2(\Omega)$.

2. Ab sofort identifizieren wir Funktionen in $W^{k,p}(\Omega)$ miteinander, die fast überall übereinstimmen, d.h. die Elemente von $W^{k,p}$ sind Äquivalenzklassen.

Definition B.17 (Sobolev-Norm) *Falls $u \in W^{k,p}(\Omega)$, definieren wir dessen* Norm *als*

$$\|u\|_{W^{k,p}(\Omega)} := \begin{cases} \left(\sum_{|\alpha|\leq k} \int_\Omega |D^\alpha u|^p \,\mathrm{d}x\right)^{\frac{1}{p}} & (1 \leq p < \infty) \\ \sum_{|\alpha|\leq k} \operatorname{ess\,sup}_\Omega |D^\alpha u| & (p = \infty) \end{cases}$$

Bemerkung B.18

1. Man kann leicht nachprüfen, dass dadurch tatsächlich eine Norm definiert wird.

2. Der Begriff der Konvergenz in $W^{k,p}$ bezüglich der Sobolev-Norm ist derjenige für normierte lineare Räume.

Satz B.19 (Eigenschaften schwacher Ableitungen) *Seien $u, v \in W^{k,p}(\Omega)$, $|\alpha| \leq k$. Dann gilt:*

1. *$D^\alpha u \in W^{k-|\alpha|,p}(\Omega)$ and $D^\beta\left(D^\alpha u\right) = D^\alpha\left(D^\beta u\right) = D^{\alpha+\beta}u$ für alle Multiindizes α, β mit $|\alpha| + |\beta| \leq k$.*

2. *Für alle $\lambda, \mu \in \mathbb{R}$ ist $\lambda u + \mu v \in W^{k,p}(\Omega)$ und*

$$D^\alpha(\lambda u + \mu v) = \lambda D^\alpha u + \mu D^\alpha v.$$

3. *Falls V eine offene Teilmenge von Ω ist, dann ist $u \in W^{k,p}(V)$.*

Bemerkung B.20 Diese Regeln sind offensichtlich richtig für glatte Funktionen, aber Funktionen in Sobolev-Räumen sind nicht notwendigerweise glatt: wir müssen uns immer einzig und allein auf die Definition schwacher Ableitungen zum Beweis stützen!

Satz B.21 (Sobolev-Räume als Funktionenräume) *Für alle $k = 1, 2, 3, \ldots$ und $1 \leq p \leq \infty$ ist der Sobolev-Raum $W^{k,p}(\Omega)$ ein Banach-Raum.*

Bemerkung B.22 Die Bedeutung und Tragweite dieses Satzes kann nicht überschätzt werden!

Satz B.23 (Sobolev-Räume als Hilbert-Räume) *Für alle $k = 1, 2, 3, \ldots$ ist der Sobolev-Raum $H^k(\Omega)$ ein Hilbert-Raum versehen mit dem Skalarprodukt*

$$(u, v)_{H^k(\Omega)} := \sum_{|\alpha|\leq k} \int_\Omega (D^\alpha u)(D^\alpha v)\mathrm{d}x \quad \forall u, v \in H^k(\Omega).$$

B.3.3. Approximation durch glatte Funktionen

Satz B.24 (Globale Approximation durch glatte Funktionen) *Angenommen, Ωist beschränkt, und weiterhin sei $u \in W^{k,p}(\Omega)$ für ein $1 \leq p < \infty$. Dann existieren Funktionen $u_m \in C^\infty(\Omega) \cap W^{k,p}(\Omega)$, so dass*

$$u_m \to u \quad \text{in } W^{k,p}(\Omega).$$

Bemerkung B.25 Die Bedeutung und Tragweite dieses Satzes kann ebenfalls nicht überschätzt werden!

- Wir haben keine Annahmen über die Glattheit von $\partial\Omega$ getroffen.
- Wir beobachten bei genauer Betrachtung, dass wir *nicht* $u_m \in C^\infty(\overline{\Omega})$ postuliert haben.
- Die Aussage des Satzes ist, dass $C^\infty(\Omega)$ *dicht* in $W^{k,p}(\Omega)$ ist. Somit können wir Sätze und Aussagen für glatte Funktionen beweisen und die Resultate durch ein Dichtheitsargument auf Sobolev-Räume übertragen! (natürlich muss eine Aussage in diesem Sinne übertragbar sein)

Satz B.26 (Global Approximation durch glatte Funktionen bis zum Rand) *Sei Ω beschränkt und $\partial\Omega$ in C^1. Weiterhin nehmen wir an, dass $u \in W^{k,p}(\Omega)$ für ein $1 \leq p < \infty$. Dann existieren Funktionen $u_m \in C^\infty(\overline{\Omega})$, so dass*

$$u_m \to u \quad \text{in } W^{k,p}(\Omega).$$

B.3.4. Spuren

- Wir benötigen die Möglichkeit, "Randwerte" entlang $\partial\Omega$ einer Funktion $u \in W^{k,p}(\Omega)$ zuzuweisen, wobei wir annehmen, dass $\partial\Omega$ in C^1 ist.
- Dies stellt für $u \in C(\overline{\Omega})$ kein Problem dar.
- Problem in $W^{k,p}(\Omega)$:
  - $u \in W^{k,p}(\Omega)$ ist im Allgemeinen *nicht glatt*
  - Nur fast überall in Ω definiert.
  - $\partial\Omega$ hat n-dimensionales Lebesgue-Maß Null $\Rightarrow$ wir können dem Ausdruck "u eingeschränkt auf $\partial\Omega$" keine direkte Bedeutung zuweisen.
- Lösung: Begriff des Spur-Operators.

Satz B.27 (Spursatz) *Sei Ω beschränkt mit $\partial\Omega$ in C^1, sowie $1 \leq p < \infty$. Dann existiert ein beschränkter linearer Operator*

$$T : W^{1,p}(\Omega) \to L^p(\partial\Omega),$$

so dass

1. *$Tu = u_{|\partial\Omega}$ if $u \in W^{1,p}(\Omega) \cap C(\overline{\Omega})$ und*
2. $$\|Tu\|_{L^p(\partial\Omega)} \leq C \|u\|_{W^{1,p}(\Omega)}$$

für alle $u \in W^{1,p}(\Omega)$, wobei die Konstante C nur von p und Ω abhängt.

Definition B.28 (Spur) *Wir nennen Tu die* Spur *von u auf $\partial\Omega$.*

Bemerkung B.29 Der Spuroperator gibt uns gemäß der 1. Eigenschaft im Spursatz genau die Werte einer Funktion u im klassischen Sinne, falls diese Funktion glatt genug, d.h. stetig, bis zum Rand ist. Die "Hoffnung" ist also, dass dieser Operator für alle anderen Funktionen in $W^{1,p}(\Omega)$ ebenfalls eine sinnvolle Interpretation von "Randwerten" gibt.

Satz B.30 (Funktionen mit Spur Null in $W^{1,p}(\Omega)$) *Sei Ω beschränkt mit $\partial\Omega$ in C^1. Wir nehmen weiterhin an, dass $u \in W^{1,p}(\Omega)$. Dann ist $u \in W_0^{1,p}(\Omega)$ genau dann, wenn $Tu = 0$ auf $\partial\Omega$.*

Bemerkung B.31 Wir arbeiten hauptsächlich in $H_0^1(\Omega) = W_0^{1,2}(\Omega)$.

B.3.5. Ungleichungen

1. Es gibt tonnenweise wichtige Ungleichungen im Zusammenhang mit Sobolev-Räumen.
2. Sobolev-Ungleichungen,Ungleichungen vom Gagliardo-Nirenberg-Typ, Morrey's-Ungleichung, ...
3. Wir diskutieren nur kurz ein wichtiges Beispiel, das im Wesentlichen besagt, dass es in jeder Äquivalenzklasse eines Sobolev-Raums einen (in gewissem Sinne) stetigen Repräsentanten gibt.

Definition B.32 *Wir sagen, u^* ist eine* Version *einer gegebenen Funktion u, falls $u = u^*$ fast überall.*

Definition B.33 (Hölder-Raum) *Der* Hölder-Raum $C^{k,\gamma}(\overline{\Omega})$, $0 < \gamma \leq 1$, *besteht aus allen Funktionen $u \in C^k(\overline{\Omega})$, $u : \Omega \to \mathbb{R}$, für welche die Norm*

$$\|u\|_{C^{k,\gamma}(\overline{\Omega})} := \sum_{|\alpha|\leq k} \|D^\alpha u\|_{C(\overline{\Omega})} + \sum_{|\alpha|=k} [D^\alpha u]_{C^{0,\gamma}(\overline{\Omega})}$$

endlich ist, wobei

$$[u]_{C^{0,\gamma}(\overline{\Omega})} := \sup_{x,y\in\Omega,\ x\neq y} \left\{ \frac{|u(x) - u(y)|}{|x-y|^\gamma} \right\}$$

und

$$\|u\|_{C(\overline{\Omega})} := \sup_{x\in\Omega} |u(x)| .$$

Satz B.34 *Sei Ω eine offene und beschränkte Teilmenge des $\mathbb{R}^n$ mit $\partial\Omega$ in C^1. Weiterhin nehmen wir an, dass $n < p \leq \infty$ und $u \in W^{1,p}(\Omega)$. Dann besitzt u eine Version $u^* \in C^{0,\gamma}(\overline{\Omega})$ für $\gamma = 1 - \frac{n}{p}$, für welche die Abschätzung*

$$\|u^*\|_{C^{0,\gamma}(\overline{\Omega})} \leq C \|u\|_{W^{1,p}(\Omega)}$$

gilt. Die Konstante C hängt nur von p, n und Ω.

Satz B.35 *Sei $k \geq 1$ und Ω beschränkt. Eine Funktion $v : \overline{\Omega} \to \mathbb{R}$, die stückweise unendlich oft differenzierbar ist, gehört zu $H^k(\Omega)$ genau dann, wenn $v \in C^{k-1}(\overline{\Omega})$.*

Bemerkung B.36

1. Wir identifizieren daher eine Funktion in $W^{1,p}(\Omega)$ $(p > n)$ mit ihrer stetigen Version.

2. Beispielsweise gilt der letzte Satz für Funktionen, die stückweise durch Polynome definiert und hinreichend glatt sind. (Vergleiche dazu auch die Definition 4.1 von Finiten Elementen.)

Literaturverzeichnis

[1] H. W. Alt. *Lineare Funktionalanalysis, 3. Auflage.* Springer-Verlag, 1999.

[2] L. Angermann and P. Knabner. *Numerik partieller Differentialgleichungen.* Springer, 2000.

[3] W. Auzinger. *Numerik partieller Differentialgleichungen.* Vorlesungsskriptum, 2004.

[4] R.B. Bird, W.E. Stewart, and E.N. Lightfoot. *Transport phenomena.* John Wiley Sons, 1960.

[5] D. Braess. *Finite Elemente. Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie (3. korrigierte und ergänzte Aufl.).* Springer Verlag, 2003.

[6] S.C. Brenner and L.R. Scott. *The mathematical theory of finite element methods.* Springer, 1994.

[7] P. Ciarlet. *The finite element method for elliptic problems.* North-Holland, Amsterdam, 1978.

[8] B. Dacorogna. *Direct methods in the calculus of variations.* Springer-Verlag, 1989.

[9] Ch. Eck and H. Garcke and P. Knabner. *Mathematische Modellierung.* Springer, 2008.

[10] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. *Computational Differential Equations.* Cambridge University Press, 1996.

[11] A. Ern and J.-L. Guermond. *Theory and practice of finite elements.* Springer, 2004.

[12] L. C. Evans. *Partial differential equations.* American Mathematical Society, 1998.

[13] D. Gilbarg and N.S. Trudinger. *Elliptic partial differential equations of second order. Reprint of the 1998 ed.* Classics in Mathematics. Springer, 2001.

[14] C. Großmann and H.-G. Roos. *Numerik partieller Differentialgleichungen.* Teubner Studienbücher Mathematik, 1992.

[15] W. Hackbusch. *Theorie und Numerik elliptischer Differentialgleichungen.* Teubner Studienbücher: Mathematik, 1986.

[16] F. Hirzebruch and W. Scharlau. *Einführung in die Funktionalanalysis.* Spektrum Akademischer Verlag, 1996.

[17] C. Johnson. *Numerical solution of partial differential equations by the finite element method.* Cambridge University Press, 1987.

[18] J. Jost. *Partielle Differentialgleichungen.* Springer, 1998.

[19] R.J. LeVeque. *Numerical methods for conservation laws.* Birkhäuser Verlag, 1992.

[20] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations.* Springer Verlag, 1996.

[21] R. Rannacher. *Numerische Mathematik 2 (Numerik partieller Differentialgleichungen).* Vorlesungsskriptum, 2004.

[22] Ch. Schwab. *p− and hp− Finite Element Methods. Theory and Applications in Solid and Fluid Mechanics.* Oxford University Press, 1998.

[23] H.R. Schwarz. *Methode der finiten Elemente. Eine Einführung unter besonderer Berücksichtigung der Rechenpraxis. 3., neubearb. Aufl.* Teubner Studienbücher Mathematik, 1991.

[24] G. Strang and J. Fix. *An analysis of the Finite Element Method.* Prentice-Hall Series in Automatic Computation, 1973.

[25] W. Törnig and P. Spellucci. *Numerische Mathematik für Ingenieure und Physiker.* Springer-Verlag, 1990.

[26] A. Tveito and R. Winther. *Einführung in partielle Differentialgleichungen.* Springer Verlag, 2002.

[27] D. Werner. *Funktionalanalysis, 2. überarb. Aufl.* Springer-Verlag, 1997.

[28] J. Wloka. *Partielle Differentialgleichungen. Sobolevräume und Randwertaufgaben.* Mathematische Leitfäden. Stuttgart: B. G. Teubner., 1982.

Index