

Numerik für Differenzialgleichungen
Mitschrift zur Vorlesung bei Prof. B. Haasdonk
im Wintersemester 21/22

ZHUOYAO ZENG

INHALTSVERZEICHNIS

VORWORT	7
0. ORGANISATORISCHES	9
1. ODEs: ANFANGSWERTSPROBLEME	11
1.1. Grundlagen	12
1.2. Einschrittverfahren	25
1.3. Mehrschrittverfahren	68
1.4. Randwertprobleme für ODEs	72
2. PDEs: KLASISCHE LÖSUNGSTHEORIE & FINITE-DIFFERENZEN-VERFAHREN	81
2.1. Grundlagen / Notationen	82
2.2. Poisson-Gleichung	95
2.3. Finite Differenzen für Poisson-Gleichung	107
2.4. FD für allgemeine elliptische PDEs 2. Ordnung	116
3. PDEs: SCHWACHE LÖSUNGSTHEORIE & FINITE-ELEMENTE-METHODE	131
3.1. Schwache Ableitung & Sobolev-Räume	132
3.2. Schwache Lösungen für elliptische PDEs	144
3.3. Galerkin-Verfahren	158
3.4. Finite Elemente Methode	164
3.5. FEM-Konvergenz/Fehleranalyse	180

VORWORT

Diese Mitschrift entsteht während meiner Teilnahme an der Vorlesung und wird mit der Open-Source-Software *TeXmacs* erstellt. Ich habe neben den Inhalten aus dem Vorlesungshandbuch noch eigene Verständnis und eigene Interpretationen ergänzt. Es kann gut sein, dass ich mich irgendwo geirrt habe... Es müsste auch sein, dass es Tippfehler oder ähnliches gibt. Außerdem habe ich auf einige Sachen verzichtet, hauptsächlich Skizzen / Bilder oder numerische Demos, denn die Bilder habe ich nicht in hoher Auflösung, Screenshots aus Videos einzufügen ist eine schlechte Idee, und selber solche Bilder zu produzieren ist zu viel Aufwand...

Für Feedback / Anmerkung jederart kann man mich gerne unter zhuoyao.zeng@gmail.com erreichen und darauf freue ich mich sehr. Ich bedanke mich noch bei meiner Frau Jiahui Wang und meinem Kommilitonen Herrn Moritz Sigg für die Verbesserungsvorschläge.

KAPITEL 0

ORGANISATORISCHES

Kontakt

- Dozent (V+Ü): Prof. Dr. Bernhard Haasdonk (haasdonk@mathematik.uni-stuttgart.de),
- Sprechstunde: Di. 10:00 per WebEx, nach Vereinbarung in Präsenz.

Ilias

- Bitte in den Kurs beitreten,
- Aktuelle Informationen, Ankündigungen, Rundmails,
- Material: Übungsblätter, Lücken-Skript.

Übungen

- Wöchentliche Übungsblätter, ggf. mehrwöchige Programmierblätter,
- Etwa 3/4 Theorie, 1/4 Programmierung,
- Programmiersprache „egal“, u.a. MatLab,
- Einzelabgabe, Di. 18:00 in Ilias.

Scheinkriterien

- Zulassungsbedingung:
 - $\geq 50\%$ Theorie & Programmieraufgaben bearbeitet & abgegeben,
 - ≥ 2 mal Vorrechnen.
- Prüfung: mündlich (30 min), ggf. schriftlich falls ≥ 30 Anmeldungen.

Inhalt

Momentan geplante sind folgende Kapitel:

1. ODEs: Anfangswertprobleme,
2. PDEs: Klassische Lösungstheorie & Finite Differenzen-Verfahren,
3. PDEs: Schwache Lösungstheorie & Finite Elemente-Verfahren.

Einbettung in Numerik Profillinie

Vorlesungsreihe in Bachelor besteht aus:

- NUM I,
- NUM II,
- **NUMDGL**.

Danach sind Vorlesungen im Master-Bereich:

- NUMPDE (Einführung),
- NUMPDE (weiterführende Aspekte),
- Spezielle Aspekte der Numerik.

Letzteres in individueller Ausprägung vom Dozenten (z.B. „Approximation with Kernel Methods“ im SS22).

Literatur

- Plato: Numerische Mathematik kompakt: Grundlagenwissen für Studium und Praxis, Vieweg 2006 (für Kapitel 1 des Kurses)
- Stoer-Bulirsch: Numerische Mathematik 2, Springer 2007 (für Kapitel 1 des Kurses)
- Großmann & Roos: Numerische Behandlung partieller Differenzialgleichungen, Teubner 2005 (für Kapitel 2 des Kurses)
- Alt: Lineare Funktionalanalysis, Springer 2007 (für Kapitel 2 & 3 des Kurses)
- Braess: Finite Elemente Methode, Springer 2007 (für Kapitel 3 des Kurses)
- Evans: Partial Differential Equations, Springer 1998
- Eck, Garcke, Knabner: Mathematische Modellierung, Springer 2008

KAPITEL 1

ODEs: ANFANGSWERTSPROBLEME

Motivation

- Viele Phänomene in Natur und Technik sind durch Systeme von gewöhnlichen Differentialgleichungen (DGLn), auf Englisch „ordinary differential equations (ODEs)“, charakterisierbar.
- ODE bedeutet: Gleichung enthält Ableitungen der Funktion bzgl. einer Variablen (z.B. Zeit t oder Ort x).
- Gleichungen, die Ableitungen einer Funktion nach mehreren Variablen enthalten, nennt man partielle DGLn (auf Englisch „partial differential equations“, PDEs).
- Viele solcher ODEs / PDEs sind nicht analytisch lösbar, daher sind numerische Verfahren erforderlich.

Beispiel. (In-)Stationäre Wärmeleitung

Gegeben ist ein Stab der Länge 1, und eine brennende Kerze unter dem Mittelpunkt des Stabs.

Gesucht ist die Temperatur $y(x)$ von Position $x \in [0, 1]$ auf dem Stab.

Intuitiv würde man erwarten, dass die Temperatur im Mittelpunkt des Stabs am höchsten ist und sie in beide Richtungen sinkt.

In Mathematischer Modellierung lernt man, dass man das System durch einen „Diffusionsprozess“ beschreiben kann, genauer gesagt, durch die ODE

$$\forall x \in [0, 1]: -y''(x) = f(x) \quad \wedge \quad y(0) = y(1) = 0,$$

mit Querform $f: [0, 1] \rightarrow [0, 1]$ und Randbedingung $y(0) = y(1) = 0$.

Obiges System wird als stationäre Wärmeleitung genannt.

Eine Variante davon ist die instationäre Wärmeleitung, die durch die PDE

$$\frac{\partial}{\partial t} u(x, t) - \frac{\partial^2}{\partial x^2} u(x, t) \cdot \kappa = f(x, t)$$

mit Wärmeleitungscoefficient κ beschrieben werden kann.

Beide Fälle werden wir im Kapitel zu PDEs behandeln.

Beispiel. Teilchensysteme

Gegeben ist ein Teilchen, und gesucht ist die Position $y(t)$ und damit die Geschwindigkeit $y'(t)$ des Teilchens in einem Kraftfeld.

Das System lässt sich beschreiben mit der ODE

$$m \cdot y''(t) = f(t), \quad y(0) = y_0, \quad y'(0) = v_0,$$

wobei m die Masse des Teilchens, f die Funktion des Kraftfeldes, $y(0) = y_0$ die Anfangsposition und $y'(0) = v_0$ die Anfangsgeschwindigkeit bezeichnen.

Dies ist ein Anfangswertproblem und wird in Kapitel 1 behandelt.

1.1. GRUNDLAGEN

Zunächst klären wir die grundlegenden Begriffe und wir versuchen, die Lösbarkeit von analytischen Problemen sowie analytische Eigenschaften der Lösungen zu untersuchen.

Definition 1.1. (ODE-System n -ter Ordnung)

Seien $d, k, n \in \mathbb{N}$, $m := 1 + d(n+1)$, $D \subseteq \mathbb{R}^m$ ein Gebiet (d.h. offen und zusammenhängend) und $F: D \rightarrow \mathbb{R}^k$ stetig. Wir suchen ein Intervall $J \subset \mathbb{R}$ sowie eine Funktion $y: J \rightarrow \mathbb{R}^d$ mit

$$\forall t \in I: (t, y(t), y'(t), \dots, y^{(n)}(t)) \in D \quad \wedge \quad F(t, y(t), y'(t), \dots, y^{(n)}(t)) = 0 \quad (1.1)$$

und nennen dieses y eine Lösung der gewöhnlichen DGL n -ter Ordnung.

Bemerkung.

- Für ein $t \in J$ kann man $(t, y(t), y'(t), \dots, y^{(n)}(t)) \in \mathbb{R}^m$ als „Zustandsvektor“ nennen, und m entsprechend „die Dimension des Zustandsvektors“. Ein konkretes Beispiel dafür ist der Fall wenn $d = n = 1$ und y die Position eines Teilchens beschreibt.
- Im Fall $d = 1$ handelt es sich um eine „skalare DGL“, z.B. die stabile Wärmeleitungsgleichung; sonst „System von DGLn“.
- Zur Vereinfachung kürzt man die DGL (1.1) oft ab als $F(t, y, y', \dots, y^{(n)}) = 0$.
- Falls die DGL (1.1) auflösbar nach der höchsten Ableitung ist, d.h. es existiert eine Funktion $\exists f: \mathbb{R}^{m-d} \rightarrow \mathbb{R}^d$, so dass:

$$y^{(n)} = f(t, y, y^{(1)}, \dots, y^{(n-1)}), \quad (1.2)$$

dann heißt die DGL explizit, bspw. die stabile Wärmeleitungsgleichung, sonst heißt sie implizit.

- Falls f in (1.2) unabhängig von t ist, nennen wir es ein autonomes System.
- Explizite Systeme höherer Ordnung können durch Einführung neuer Variablen zu Systemen erster Ordnung umgeschrieben werden:

$$y^{(n)} = f(t, y, y', \dots, y^{(n-1)}) \quad (1.3)$$

\Leftrightarrow

$$\bar{y} := \begin{pmatrix} y \\ y' \\ \vdots \\ y^{(n-1)} \end{pmatrix} \wedge \bar{y}'(t) = \begin{pmatrix} y' \\ y'' \\ \vdots \\ f(t, y, y', \dots, y^{(n-1)}) \end{pmatrix} =: \bar{f}(t, \bar{y}). \quad (1.4)$$

- Eine Anfangsbedingung für (1.4) entspricht also Anfangsbedingung für alle Ableitungen von y bis $y^{(n-1)}$, also

$$\bar{y}(0) = (v_0 \ v_1 \ \dots \ v_{n-1})^T \Leftrightarrow y(0) = v_0 \wedge \dots \wedge y^{(n-1)}(0) = v_{n-1}.$$

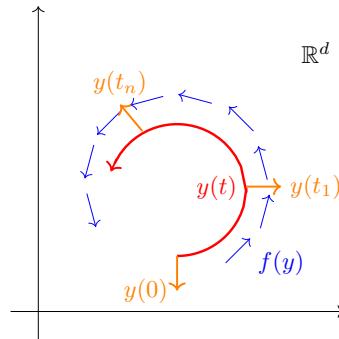
- Obige Überlegung besagt:

Bei expliziten DGLn reicht es, Systeme erster Ordnung zu betrachten.

Bemerkung. (Phasenraum)

Man kann die rechte Seite eines Systems erster Ordnung $f(t, y)$ als zeitabhängiges Vektorfeld und $y: I \rightarrow \mathbb{R}^d$ als Trajektorie eines Punktes im Vektorfeld interpretieren.

DGL besagt also, dass der Punkt sich in Richtung f bewegt — anschaulich steht y tangential an Vektorfeld f , z.B.: für ein zeitunabhängiges f (also autonomes System $y' = f(y)$) erhalten wir folgende Skizze:



Bemerkung. (Translationsinvarianz)

Bei einem autonomen System $y' = f(y)$ mit Lösung $y: I \rightarrow \mathbb{R}^d$ ist jede zeitverzögerte Funktion ebenfalls eine Lösung des Systems. Konkret heißt es:

Für $\Delta_t \in \mathbb{R}$, $\bar{I} := \{t - \Delta_t | t \in I\}$ ein verschobenes Zeitintervall setze

$$\forall t \in \bar{I}: \bar{y}(t) := y(t + \Delta_t).$$

$$\Rightarrow \bar{y}'(t) = y'(t + \Delta_t) = f(y(t + \Delta_t)) = f(\bar{y}(t)).$$

D.h. i.A. ist keine Eindeutigkeit der Lösungen von DGLn zu erwarten.

Die Eigenschaft der Translationsinvarianz motiviert auch die Einführung der Anfangswertbedingung. Bevor dies tatsächlich geschieht, schauen wir uns weitere Beispiele an:

Beispiel. (Populationsmodelle)

Unter einer unbekannten Populationsgröße $p(t)$ bzgl. Zeitparameter t und Startzeit t_0 versetzt man eine Lösung von

$$p(t_0) = p_0 \in \mathbb{R}_+ \quad \wedge \quad \forall t \in [t_0, \infty): p'(t) = r(t, p(t)) \cdot p(t)$$

mit $r(t, p)$ der relativen Änderungsrate.

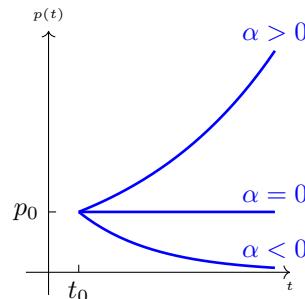
- i. Eine konstante Wachstumsrate $r(t, p) = \alpha \in \mathbb{R}$ liefert ein exponentielles Wachstum mit Lösung

$$p(t) = p_0 \cdot e^{\alpha(t-t_0)}.$$

D.h. wir erhalten

$$\lim_{t \rightarrow \infty} p(t) = \begin{cases} \infty, & \alpha > 0, \\ p_0, & \alpha = 0, \\ 0, & \alpha < 0. \end{cases}$$

Graphisch sieht es dann aus wie



Später bezeichnen wir den Fall $\alpha > 0$ instabil und den Fall $\alpha \leq 0$ stabil.

Im Allgemeinen ist konstante Wachstumsrate in biologischen Systemen eher unrealistisch.

- ii. Das Modell mit einem beschränkten Wachstums ist eine Modifikation und dabei verwendet man den Ansatz

$$r(t, p) = \beta \cdot (\xi - p) \quad \text{mit } \beta, \xi \in \mathbb{R}_+.$$

Die Motivation des Ansatzes besteht darin, dass die Population mit der Zeit gegen einen Schwellwert ξ konvergieren sollte. Um dies zu sehen, betrachten wir die DGL unter diesem Ansatz:

$$p' = r \cdot p = \beta(\xi - p) \cdot p = \beta\xi p - \beta p^2.$$

Mit $\alpha := \beta\xi$ erhalten wir die sogenannte logistische DGL

$$p' = \alpha p - \beta p^2.$$

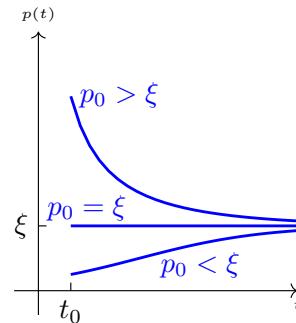
Die Lösung der logistische DGL erweist sich als

$$p(t) = \xi \frac{p_0}{p_0 + (\xi - p_0) e^{-\alpha(t-t_0)}} \quad (1.5)$$

und somit sieht man leicht, dass

$$\lim_{t \rightarrow \infty} p(t) = \xi$$

gilt, bzw. graphisch erhält man



Man kann auch leicht ausrechnen, dass $\operatorname{sgn}(p') = \operatorname{sgn}(\xi - p_0)$ konstant ist, und das zeigt nochmal die monotone Konvergenz von p gegen ξ .

Obige beiden Modelle beschreiben biologische Systeme mit einer Spezies, aber typischerweise sind biologische Systeme komplizierter. Das folgende Modell betrachtet die Populationen zweier Spezies mit einem gewissen Zusammenhang:

Beispiel. (Räuber-Beute-Modelle)

Sei $x(t)$ die Populationsgröße der Beutespezies und $y(t)$ die Populationsgröße der Räuberspezies.

Intuitiv würde man schon erwarten, dass die relativen Änderungsraten beider Spezies durch jeweils die Andere beeinflusst werden.

Diese Situation kann man mit den Lotka-Volterra-Gleichungen bzgl. Parameter $\alpha, \beta, \gamma, \delta > 0$ modellieren, also

$$\begin{cases} x' = (\alpha - \beta y) x, \\ y' = (-\gamma + \delta x) y, \end{cases}$$

mit den folgenden Interpretationen:

α : Wachstumsrate von x ohne Räuber y ,

βy : Sterberate von x durch Räuber y ,

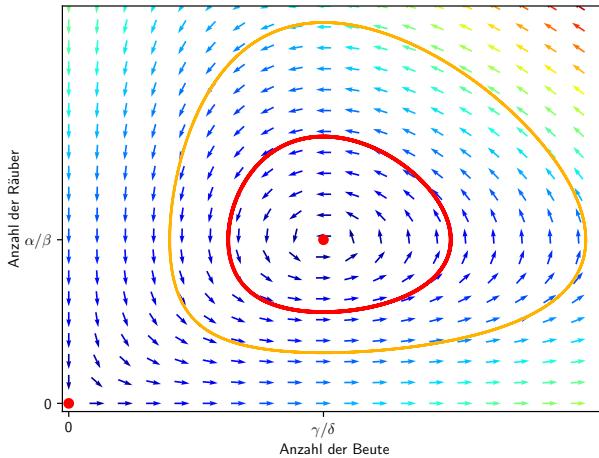
γ : Sterberate von y ohne Beute x ,

δx : Wachstumsrate von y bei Vorhandensein von Beute x .

Für solche Systeme sind ihre stationären Punkte interessant, d.h. Punkte, wo die Ableitungen verschwinden. Dazu kann man sich leicht überlegen:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Leftrightarrow \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} \equiv \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ oder } \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} \gamma/\delta \\ \alpha/\beta \end{pmatrix}.$$

Die möglichen Lösungskurven kann man im Phasenraum visualisieren:



Im Bild werden die Kurven mit negativen x oder y Werten nicht gezeichnet. Solche Kurven sind zwar mathematisch möglich, aber wenig realistisch.

Die Kurven sind geschlossen, denn sie sind Höhenlinien von der Funktion

$$J: (\mathbb{R}^+)^2 \rightarrow \mathbb{R}, (x, y) \mapsto \gamma \ln(x) - \delta x + \alpha \ln(y) - \beta y.$$

Genauer gesagt kann man zeigen, dass für jede Lösung $\begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$ mit $t \in \mathbb{R}$ gilt

$$\frac{d}{dt} J(x(t), y(t)) = 0.$$

Das bedeutet: Eine Lösung ist periodisch, wenn ein Punkt zum zweiten Mal erreicht wird, also:

$$\begin{pmatrix} x(\tau) \\ y(\tau) \end{pmatrix} = \begin{pmatrix} x(0) \\ y(0) \end{pmatrix} \Rightarrow \forall t \in \mathbb{R} \forall k \in \mathbb{Z}: \begin{pmatrix} x(t+k\tau) \\ y(t+k\tau) \end{pmatrix} = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}.$$

Wir sehen: Ohne „Startpunkt“ ist keine eindeutige Lösung zu erwarten.

Für den Rest des Kapitels betrachten wir nur noch Anfangswertprobleme:

Definition 1.2. (AWP)

Sei $d \in \mathbb{N}$, $D \subseteq \mathbb{R}^{d+1}$ ein Gebiet, $(t_0, y_0) \in D$ und $f: D \rightarrow \mathbb{R}^d$ stetig.

Gesucht ist ein Intervall $I \subseteq \mathbb{R}$ mit $t_0 \in I$ und ein $y \in C^1(I, \mathbb{R}^d)$, so dass

$$y(t_0) = y_0 \quad \wedge \quad \forall t \in I: y'(t) = f(t, y) \tag{1.6}$$

und man nennt y die Lösung vom Anfangswertproblem (1.6).

Beachte: Der Fall $I = \mathbb{R}$ ist auch möglich, falls es D passt.

Der folgende Satz besagt, dass man ein Anfangswertproblem äquivalent umformen kann:

Satz 1.3. (Volterra'sche Integralgleichung)

Wir definieren die Volterra'sche Integralgleichung (VI) als

$$y(t) = y_0 + \int_{t_0}^t f(s, y(s))ds \quad (1.7)$$

und stellen fest, dass die folgende zwei Bedingungen äquivalent sind:

- i. Eine Lösung $y \in C^1(I, \mathbb{R}^d)$ des AWP (1.6) erfüllt (1.7) für alle $t \in I$;
- ii. Eine Funktion $y \in C(I, \mathbb{R}^d)$, welche (1.7) für alle $t \in I$ erfüllt, ist eine Lösung des AWP (1.6), insbesondere ist y stetig differenzierbar.

Beweis. „i. \Rightarrow ii.“

Sei $y: I \rightarrow \mathbb{R}^d$ eine Lösung des AWPs.

Die Fundamentalsatz der Differenzial-&Integralrechnung liefert dann

$$\forall t \in I: \quad y(t) = y(t_0) + \int_{t_0}^t y'(s)ds = y(t_0) + \int_{t_0}^t f(s, y(s))ds.$$

„ii. \Rightarrow i.“

Sei $y: I \rightarrow \mathbb{R}^d$ stetig mit $t_0 \in I$ und y erfülle VI (1.7).

y erfüllt die Anfangswertbedingung, denn mit $t = t_0$ gilt

$$y(t_0) = y_0 + \int_{t_0}^{t_0} f(s, y(s))ds = y_0.$$

Da $f(s, y(s))$ stetig bzgl. s ist, ist $y(t) = \int_{t_0}^t f(s, y(s))ds$ als Integral stetiger Funktion differenzierbar, und somit

$$y'(t) = \frac{d}{dt} \left(y(t_0) + \int_{t_0}^t f(s, y(s))ds \right) = 0 + f(t, y(t))$$

also ist y eine Lösung vom AWP (1.6). □

Besitzt ein AWP (1.6) immer eine Lösung? Ja, da f als stetig vorausgesetzt ist:

Satz 1.4. (Existenzsatz von Peano)

Jedes AWP (1.6) mit stetigem $f: D \rightarrow \mathbb{R}^d$ besitzt mindestens eine lokale Lösung $y: I_0 \rightarrow \mathbb{R}^d$, welche sich auf den Rand von D fortsetzen lässt.

Wir verzichten hier auf den Beweis dieses Satzes.

Bemerkung. (zur Fortsetzbarkeit einer Lösung)

Sei $I_0 = (t_1, t_2)$. Dann ist die Fortsetzung eventuell sowohl für $t > t_2$ als auch für $t < t_1$ möglich, bis Rand ∂D erreicht wird.

Zu einer Lösung $y: I_0 \rightarrow \mathbb{R}^d$ existiert also ein maximales Existenzintervall I_{\max} , welches wir im folgenden oft einfach als I bezeichnen.

Dieses maximale Existenzintervall $I_{\max} = I$ kann aber „alle möglichen Formen“ haben: vielleicht beschränkt, oder vielleicht unbeschränkt, oder vielleicht abgeschlossen, halb-offen... Wir schauen uns einige Beispiele an:

- Lösung nur auf Halbachsen, d.h. für endliche Zeit:

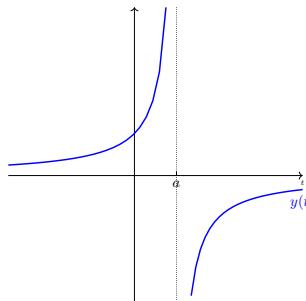
$$y' = y^2 \quad \wedge \quad D = \mathbb{R}^2.$$

Für $a \in \mathbb{R}$ (von der Anfangswertbedingung abhängig) ist die Funktion

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

eine Lösung dieser DGL, denn

$$y'(t) = -\frac{1}{(a-t)^2}(-1) = \frac{1}{(a-t)^2} = y(t)^2.$$



Offensichtlich ist y nur auf $(-\infty, a)$ oder (a, ∞) sinnvoll definiert, und welches Intervall der Beiden als I_{\max} genommen wird, hängt davon ab, in welchem Intervall t_0 liegt.

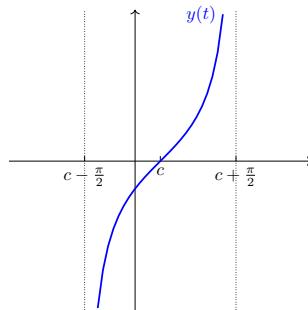
- Lösung nur auf endlichem Intervall:

$$y' = 1 + y^2.$$

Für $c \in \mathbb{R}$ (von der Anfangswertbedingung abhängig) ist die Funktion

$$y(t) = \tan(t - c)$$

eine Lösung der DGL und wegen der Polstellen von \tan kann I_{\max} höchstens $(c - \frac{\pi}{2}, c + \frac{\pi}{2})$ sein.



Satz von Peano besagt, jedes AWP mit stetigem f besitzt eine Lösung, aber diese muss nicht eindeutig sein:

Beispiel. (Mehrdeutigkeit von Lösung eines AWPs)

Für $d=1, D=\mathbb{R} \times \mathbb{R}^+$ betrachte

$$y' = \sqrt{|y(t)|} \quad \wedge \quad y_0 := y(0) = 0.$$

Das AWP besitzt zumindest zwei Lösungen $y(t) \equiv 0$ und $y(t) = \frac{1}{4}t^2$.

Also, für die (Existenz &) Eindeutigkeit einer Lösung beim AWP sind stärkere Bedingungen erforderlich:

Satz 1.5. (Existenz & Eindeutigkeit für AWP, Picard-Lindelöf)

Sei ein AWP (1.6) gegeben, also insbesondere ist f stetig.

Sei außerdem f Lipschitz-stetig bzgl. 2. Argument, d.h.

$$\exists L \in \mathbb{R}_+ \quad \forall (t, y), (t, \bar{y}) \in D: \|f(t, y) - f(t, \bar{y})\| \leq L \|y - \bar{y}\|.$$

Dann existiert eine Lösung $y: I_0 \rightarrow \mathbb{R}^d$, welche sich auf maximales Existenzintervall I_{\max} fortsetzen lässt und eindeutig ist.

Bemerkung.

- Im Fall von steitgem, stückweise differenzierbarem f gibt es eine Charakterisierung für die Lipschitz-Stetigkeit:

$$f \text{ Lipschitz-stetig bzgl. } y \Leftrightarrow \frac{\partial}{\partial y} f \text{ beschränkt durch Lipschitz-Konstante } L.$$

Dies liefert dann auch einen einfachen und in der Praxis umsetzbaren Test.

- Die Funktion $f(t, y) = \sqrt{|y|}$ aus Beispiel zu Mehrdeutigkeit ist nicht Lipschitz-stetig auf D , denn die partielle Ableitung

$$\frac{\partial}{\partial y} \sqrt{|y|} = \frac{1}{2\sqrt{|y|}} \operatorname{sgn}(y)$$

ist unbeschränkt für $y \rightarrow 0$. Somit ist dabei der Satz 1.5 nicht anwendbar.

Beweis. (von 1.5)

Existenz von Lösung $y: I \rightarrow \mathbb{R}^d$ für I mit $I_0 \subseteq I \subseteq I_{\max}$ folgt aus Peano & Fortsetzbarkeit.

Interessant ist die Eindeutigkeit:

OBdA sei I kompakt (sonst folgt Eindeutigkeit durch Ausschöpfen von I mit kompakten Intervallen).

Wir zeigen die Eindeutigkeit mit dem Banach'schen Fixpunktsatzes (BFS).

- Definiere auf $C(I, \mathbb{R}^d)$ eine neue „gewichtete“ Norm $\|\bullet\|_{\alpha, t_0}$ zu Parametern $\alpha > 0, t_0 \in I$:

$$\forall u \in C(I, \mathbb{R}^d): \quad \|u\|_{\alpha, t_0} := \sup_{t \in I} |u(t)| e^{-\alpha(t-t_0)}.$$

Man kann nachrechnen, dass dies tatsächlich eine Norm ist (Beweis ähnlich zu $\|\bullet\|_\infty$) und Cauchy-Folgen in $C(I, \mathbb{R}^d)$ bzgl. dieser Norm gleichmäßig konvergieren, d.h. $C(I, \mathbb{R}^d)$ ist vollständig.

- b) Wähle von $C(I, \mathbb{R}^d)$ eine Teilmenge (Definitionsbereich von Kontraktion)

$$M := \{y \in C(I, \mathbb{R}^d) \mid y(t_0) = y_0\}.$$

M ist abgeschlossen bzgl. $\|\bullet\|_{\alpha, t_0}$, denn für $(u_k)_{k \in \mathbb{N}} \in \text{CF}(C(I, \mathbb{R}^d), \|\bullet\|_{\alpha, t_0})$ existiert ein $u \in C(I, \mathbb{R}^d)$ so dass $u_k \xrightarrow{\text{glm.}} u$, da $(C(I, \mathbb{R}^d), \|\bullet\|_{\alpha, t_0})$ vollständig ist. Damit gilt $u(t_0) = \lim_{k \rightarrow \infty} u_k(t_0) = y_0$, also ist $u \in M$.

- c) Definiere die Abbildung Φ durch die Volterra'sche Integralgleichung (VI), d.h. für $u \in M$ setze

$$\Phi(u)(t) := y_0 + \int_{t_0}^t f(s, u(s)) ds, \quad t \in I.$$

Es gilt $\Phi(u) \in C(I, \mathbb{R}^d)$, denn mit Stetigkeit von u und f ist das Integrand stetig, und somit $\Phi(u)$ als Integral stetiger Funktion stetig.

Zudem ist $\Phi(u)(t_0) = y_0$, also ist $\Phi: M \rightarrow M$ eine Selbstabbildung.

Es bleibt noch die Kontraktionseigenschaft zu zeigen.

- d) Führe eine punktweise Abschätzung von $\Phi(-)(t)$ durch, also für $u, \bar{u} \in M$ und $t \geq t_0$ gilt:

$$\begin{aligned} |\Phi(u)(t) - \Phi(\bar{u})(t)| &= \left| \int_{t_0}^t f(s, u(s)) - f(s, \bar{u}(s)) ds \right| \\ &\leq \int_{t_0}^t |f(s, u(s)) - f(s, \bar{u}(s))| ds \\ &\leq L \int_{t_0}^t |u(s) - \bar{u}(s)| ds \\ &= L \int_{t_0}^t |u(s) - \bar{u}(s)| e^{-\alpha|s-t_0|} e^{\alpha|s-t_0|} ds \\ &\leq L \int_{t_0}^t \|u - \bar{u}\|_{\alpha, t_0} e^{\alpha|s-t_0|} ds \\ &= L \|u - \bar{u}\|_{\alpha, t_0} \int_{t_0}^t e^{\alpha|s-t_0|} ds \\ &= L \|u - \bar{u}\|_{\alpha, t_0} \frac{1}{\alpha} (e^{\alpha|t-t_0|} - 1) \\ &\leq \frac{L}{\alpha} \|u - \bar{u}\|_{\alpha, t_0} e^{\alpha|t-t_0|}. \end{aligned}$$

Die 3. Zeile folgt aus Lipschitz-Stetigkeit von f bzgl. 2. Arguments; bei 4. Zeile haben wir „mit 1 multipliziert“; andere Stellen sind selbsterklärend.

Die Abschätzung gilt analog auch für $t < t_0$, und somit gilt

$$\forall u, \bar{u} \in M \forall t \in I: |\Phi(u)(t) - \Phi(\bar{u})(t)| e^{-\alpha|t-t_0|} \leq \frac{L}{\alpha} \|u - \bar{u}\|_{\alpha,t_0}.$$

e) Dank d) erhalten wir durch Supremumsbildung

$$\forall u, \bar{u} \in M: \|\Phi(u)(t) - \Phi(\bar{u})(t)\|_{\alpha,t_0} \leq \frac{L}{\alpha} \|u - \bar{u}\|_{\alpha,t_0}.$$

Wähle $\alpha > L$, dann ist der Faktor $q := \frac{L}{\alpha} < 1$ und somit ist Φ eine Kontraktion.

f) Nach BFS besitzt Φ einen eindeutigen Fixpunkt $y \in M$, also $\Phi(y) = y$.

Somit löst y eindeutig die Volterra'sche Integralgleichung, und mit 1.3 ii. ist y auch die eindeutige Lösung des AWPs. \square

Die Stabilität von AWP ist sehr relevant, also es stellt sich die Frage:

Wie unterscheiden sich Lösungen zu leicht verändertem y_0 bzw. f ?

Um diese Frage zu beantworten, benötigen wir einige Hilfssätze:

Lemma 1.6. (Grönwall)

Seien h, w, k stetige, nicht-negative Funktionen auf $[a, b]$ und es gelte

$$\forall t \in [a, b]: h(t) \leq w(t) + \int_a^t k(s)h(s)ds.$$

Dann gilt

$$\forall t \in [a, b]: h(t) \leq w(t) + \int_a^t K(s, t)k(s)w(s)ds$$

mit

$$K(s, t) = \exp\left(\int_s^t k(\tau)d\tau\right).$$

Eine Interpretation für diesen Satz wäre:

Die Funktion h kann man als die „Fehlerfunktion“ sehen, und wenn der Fehler $h(t)$ zum Zeitpunkt t durch „den bisher gesammelten Fehler“ $\int_a^t k(s)h(s)ds$ abgeschätzt werden kann, dann kann man $h(t)$ unabhängig vom aktuellen Zeitpunkt abschätzen.

Beweis.

i. Setze eine Hilfsfunktion

$$H(t) := \int_a^t k(s)h(s)ds$$

$$\Rightarrow H(a) = 0 \text{ und } \forall t \in [a, b]: h(t) \leq w(t) + H(t).$$

$$\Rightarrow \forall t \in [a, b]: H'(t) = k(t)h(t) \leq k(t)(w(t) + H(t)), \text{ also insbesondere gilt}$$

$$H'(t) - k(t)H(t) \leq k(t)w(t). \quad (*)$$

ii. Untersuche K , also

$$\begin{aligned} K(a, t)K(s, a) &= \exp\left(\int_a^t k(\tau)d\tau\right)\exp\left(\int_s^a k(\tau)d\tau\right) \\ &= \exp\left(\int_a^t k(\tau)d\tau + \int_s^a k(\tau)d\tau\right) \\ &= \exp\left(\int_s^t k(\tau)d\tau\right) = K(s, t). \end{aligned}$$

D.h. $K(t, a)K(a, t) = K(t, t) = 1$.

Außerdem gilt

$$\frac{d}{dt}K(t, a) = \frac{d}{dt}\exp\left(-\int_a^t k(\tau)d\tau\right) = -K(t, a)k(t). \quad (**)$$

iii. Führe Ergebnisse zusammen und erhalte eine Ungleichung:

$$\begin{aligned} \frac{d}{dt}(K(t, a)H(t)) &= K(t, a)H'(t) + \left(\frac{d}{dt}K(t, a)\right)'H(t) \\ &= K(t, a)(H'(t) - k(t)H(t)) \\ &\leq K(t, a)k(t)w(t), \end{aligned}$$

wobei die 2. Zeile aus $(**)$ und 3. Zeile aus $(*)$ folgt.

iv. Integriere die Ungleichung aus 3. und erhalte

$$\begin{aligned} \int_a^t K(s, a)k(s)w(s)ds &\geq \int_a^t \frac{d}{ds}(K(s, a)H(s))ds \\ &= K(t, a)H(t) - K(a, a)H(a) \\ &= K(t, a)H(t). \end{aligned}$$

Damit gilt

$$\begin{aligned} H(t) = 1 \cdot H(t) &= K(t, a)K(a, t)H(t) \\ &\leq \int_a^t K(a, t)K(s, a)k(s)w(s)ds \\ &= \int_a^t K(s, t)k(s)w(s)ds. \end{aligned}$$

Insgesamt also

$$h(t) \leq w(t) + H(t) \leq w(t) + \int_a^t K(s, t)k(s)w(s)ds. \quad \square$$

Der Fall $k(t) \equiv c \in \mathbb{R}_+$ liefert eine vereinfachte Version des obigen Lemmas:

Folgerung 1.7. (Vereinfachtes Grönwall Lemma)

Seien h, w stetige, nichtnegative Funktionen auf $[a, b]$ und $c \in \mathbb{R}_+$ s.d. es gilt

$$\forall t \in [a, b]: \quad h(t) \leq w(t) + c \int_a^t h(s)ds.$$

Dann gilt

$$h(t) \leq w(t) + c \int_a^t e^{c(t-s)} w(s) ds \leq \left(\max_{s \in [a,t]} w(s) \right) e^{c(t-a)}. \quad (1.8)$$

Beweis. $k(t) \equiv c$ liefert die erste Ungleichung in (1.8).

Die 2. Ungleichung folgt aus der Rechnung

$$\begin{aligned} w(t) + c \int_a^t e^{c(t-s)} w(s) ds &\leq \max_{s \in [a,t]} w(s) \left(1 + c \int_a^t e^{c(t-s)} ds \right) \\ &= \max_{s \in [a,t]} w(s) e^{c(t-a)}. \end{aligned}$$

□

Nun zur Stabilitätsaussage:

Satz 1.8. (Stabilitätssatz AWP)

Seien $f, g: D \rightarrow \mathbb{R}^d$ stetig, wobei D ein Gebiet ist, und $(t_0, y_0), (t_0, z_0) \in D$.

Sei f noch Lipschitz-stetig bzgl. 2. Argument mit Lipschitz-Konstante $L \in \mathbb{R}_+$.

Seien $\varepsilon_1, \varepsilon_2 \in \mathbb{R}_+$ so dass

$$\forall (t, y) \in D: \|y_0 - z_0\| \leq \varepsilon_1 \quad \wedge \quad \|f(t, y) - g(t, y)\| \leq \varepsilon_2.$$

Seien $y, z: I \rightarrow \mathbb{R}^d$ zwei Lösungen der AWP

$$\begin{aligned} y' &= f(t, y), \quad y(t_0) = y_0 \\ z' &= g(t, z), \quad z(t_0) = z_0 \end{aligned}$$

wobei I der Schnitt der maximalen Intervalle beider Lösungen ist.

Dann gilt

$$\forall t \in I: \|y(t) - z(t)\| \leq \left(\varepsilon_1 + \varepsilon_2 \int_0^{|t-t_0|} e^{-Ls} ds \right) e^{L|t-t_0|} \quad (1.9)$$

sowie die gröbere Abschätzung

$$\forall t \in I: \|y(t) - z(t)\| \leq (\varepsilon_1 + \varepsilon_2 |t - t_0|) e^{-L|t-t_0|}. \quad (1.10)$$

Bemerkung.

- Der Stabilitätssatz lässt sich auch als „Stetigkeitssatz“ oder „Störungssatz“ nennen, denn die Lösungen hängen stetig von Daten (y_0, z_0, f, g) ab bzw. g, z_0 können als Störungen von f bzw. y_0 aufgefasst werden.
- Es existieren Beispiele, wo die Schranke in (1.9) scharf ist, d.h. „ $=$ “ anstatt „ \leq “; aber auch Beispiele, wo $y(t) \xrightarrow{t \rightarrow \infty} 0$ und $z(t) \xrightarrow{t \rightarrow \infty} 0$ aber die Schranke beliebig exponentiell wächst, also wo der Satz nutzlos wird (Blatt1 Aufg.4).

Beweis. Sei $t \in I$. Wir zeigen nur den Fall $t \geq t_0$, denn der andere Fall folgt mit analoger Argumentation und einiger Anpassungen geeigneter Vorzeichen.

i. Schätzt den punktweisen Fehler ab:

$$\begin{aligned}
& \|y(t) - z(t)\| \\
&= \|y_0 + \int_{t_0}^t f(s, y(s)) ds - z_0 - \int_{t_0}^t g(s, z(s)) ds\| \\
&\leq \|y_0 - z_0\| + \int_{t_0}^t \|f(s, y(s)) - g(s, z(s))\| ds \\
&\leq \varepsilon_1 + \int_{t_0}^t \|f(s, y(s)) - g(s, z(s))\| ds \\
&= \varepsilon_1 + \int_{t_0}^t \|f(s, y(s)) - f(s, z(s)) + f(s, z(s)) - g(s, z(s))\| ds \\
&\leq \varepsilon_1 + \int_{t_0}^t \|f(s, y(s)) - f(s, z(s))\| + \|f(s, z(s)) - g(s, z(s))\| ds \\
&\leq \varepsilon_1 + \int_{t_0}^t L \|y(s) - z(s)\| + \varepsilon_2 ds \\
&= \varepsilon_1 + \int_{t_0}^t L \|y(s) - z(s)\| ds + \int_{t_0}^t \varepsilon_2 ds \\
&= \varepsilon_1 + \varepsilon_2(t - t_0) + L \int_{t_0}^t \|y(s) - z(s)\| ds.
\end{aligned}$$

Das heißt, wir erhalten

$$\|y(t) - z(t)\| \leq \varepsilon_1 + \varepsilon_2(t - t_0) + L \int_{t_0}^t \|y(s) - z(s)\| ds.$$

ii. Wende das vereinfachte Grönwall Lemma an:

Mit $h(t) := \|y(t) - z(t)\|$, $w(t) := \varepsilon_1 + \varepsilon_2(t - t_0)$, $c := L$, $a := t_0$ und b der Art, dass $[a, b] \subset I$ ist, liefert das vereinfachte Grönwall Lemma:

$$\|y(t) - z(t)\| \leq \varepsilon_1 + \varepsilon_2(t - t_0) + L \int_{t_0}^t e^{L|t-s|} (\varepsilon_1 + \varepsilon_2(s - t_0)) ds. \quad (1.11)$$

iii. Rechne der Teilintegrale von ii. aus (beachte $t \geq t_0$):

$$\begin{aligned}
\int_{t_0}^t e^{L|t-s|} ds &= -\frac{1}{L}(e^{L(t-t)} - e^{L(t-t_0)}) = \frac{1}{L}(e^{L(t-t_0)} - 1). \\
\int_{t_0}^t e^{L|t-s|} s ds &= \left(-\frac{1}{L} e^{L(t-s)} \right)_{s=t_0}^{s=t} - \int_{t_0}^t \frac{-1}{L} e^{L(t-s)} ds \\
&= \frac{-1}{L} t + \frac{1}{L} e^{L(t-t_0)} + \frac{1}{L} \int_{t_0}^t e^{L(t-s)} ds.
\end{aligned}$$

Beachte: Das zweite Integral wird mit partieller Integration weitergerechnet.

iv. Setze bisherige Ergebnisse zusammen und schließe den Beweis ab:

Durch Einsetzen von iii. zu (1.11) und sehr langes Rechnen erhalten wir

$$\|y(t) - z(t)\| \leq \varepsilon_1 e^{L(t-t_0)} + \varepsilon_2 \int_{t_0}^t e^{L(t-s)} ds. \quad (1.12)$$

Mit Substitution der Integralvariable $\xi := s - t_0$ gilt

$$\int_{t_0}^t e^{L(t-s)} ds = \int_0^{t-t_0} e^{L(t-(\xi+t_0))} d\xi = \left(\int_0^{t-t_0} e^{-L\xi} d\xi \right) e^{L(t-t_0)}$$

und somit folgt die erste Ungleichung (1.9) aus (1.12).

Die Zweite Ungleichung (1.10) folgt aus (1.12) mit

$$\int_{t_0}^t e^{L(t-s)} ds \leq e^{L(t-t_0)} (t - t_0). \quad \square$$

Somit beenden wir den Theorieteil des Kapitels und als nächstes schauen wir uns konkrete numerische Verfahren an.

1.2. EINSCHRITTVERFAHREN

Die **Motivation** des Abschnitts besteht darin:

- Sei oBdA $t_0 = 0$, $T \in \mathbb{R}_+$. Gesucht ist eine Lösung von AWP aus Definition 1.2 auf Intervall $[0, T]$.
- Idee: Volterra-Integralgleichung und (zusammengesetzte) Quadratur des Integrals
- Wähle die „Anzahl der Schritte“ $K \in \mathbb{N}$, bestimme die Schrittweite $\tau := \frac{T}{K}$ und setze die „äquidistante Zeitschritte“ $t_k := k\tau$ für $k \in \{0, \dots, K\}$.
- Für ein $k \in \{0, \dots, K\}$ sei eine Approximation $y_k := y(t_k)$ gegeben.
- Wähle eine Quadratur auf $[0, 1]$, also: $\hat{Q}_n(\hat{\varphi}) = \sum_{i=0}^n \hat{w}_i \hat{\varphi}_i(\hat{s}_i)$.
- Transformiere \hat{Q}_n auf $[t_k, t_{k+1}]$, also: $Q_n(\varphi) = \sum_{i=0}^n \hat{w}_i \varphi_i(t_k + \hat{s}_i \tau)$ und schreibe $s_i := t_k + \hat{s}_i \tau$.
- Dann liefert die Volterra-Integralgleichung:

$$\begin{aligned} y(t_{k+1}) &= y(t_k) + \int_{t_k}^{t_{k+1}} f(s, y(s)) ds \\ &\approx y_k + \tau \sum_{i=0}^n \hat{w}_i f(s_i, y(s_i)). \end{aligned}$$

Wir müssen also nur geeignet $y(s_i)$ approximieren und erhalten dann die Approximation $y_{k+1} \approx y(t_{k+1})$.

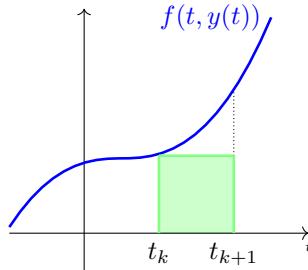
- Einen solchen Ansatz nennen wir Einschrittverfahren (ESV), da nur y_k als vorige Lösung erforderlich ist, um y_{k+1} zu berechnen. Dies steht im Gegensatz zu Mehrschrittverfahren (MSV), welche weitere Iteration y_{k-1}, y_{k-2} , etc. benötigen und im nächsten Abschnitt betrachtet werden.

Einfache ESV erhalten wir durch Rückwärts bzw. Vorwärts-Rechtecksregel:

Definition 1.9. (Euler-Verfahren)

- i. Explizites Euler-Verfahren:

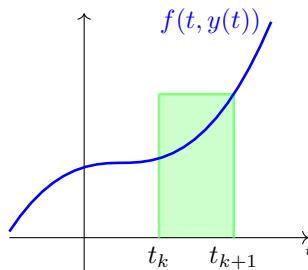
Für $k \in \{0, \dots, K-1\}$ setze $y_{k+1} := y_k + \tau f(t_k, y_k)$.



- ii. Implizites Euler-Verfahren:

Für $k \in \{0, \dots, K-1\}$ löse die folgenden nicht-lineare Gleichungssystem nach y_{k+1} :

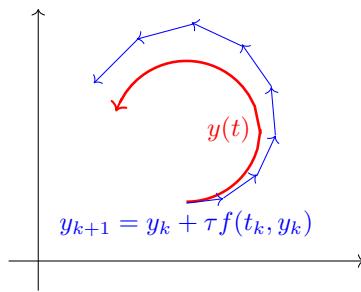
$$y_{k+1} = y_k + \tau f(t_{k+1}, y_{k+1}).$$



Bemerkung.

- Geometrische Interpretation explizites Euler-Verfahrens:

Das explizite Euler-Verfahren kann als Polygonzugverfahren interpretiert werden, denn $f(t_k, y_k)$ ist aktuelle „Geschwindigkeit“ im Phasenraum. Die approximierende Lösungsmenge geht also von aktueller Position y_k um τ in Richtung aktueller Geschwindigkeit:



- Rechenaufwand implizites Euler-Verfahrens:

Das implizite Euler-Verfahren erfordert Lösung eines nicht-linearen Gleichungssystems in jedem Schritt, typischerweise z.B. mittels Newton-Verfahren. Dies ist aufwendiger als einzelne f -Auswertung im expliziten Euler-Verfahren. Als Startwert für Newton kann z.B. alter Wert y_k verwendet werden. Wegen der i.A. lokalen Konvergenz von Newton sollte τ nicht zu groß gewählt werden.

Beispiel. (Euler-Verfahren für Räuber-Beute-Modell)

hier kommt noch ein Bild aus Uebungsaufgabe PB1A2

Aus dem Bild sehen wir:

- Explizites Euler: Größeres τ liefert schnelleres Verfahren, aber Spiralen divergieren nach außen und keine Periozität beobachtbar.
- Implizites Euler liefert auch Spiralen, die im Fixpunkt laufen, aber auch keine Periozität beobachtbar.

Verbesserungen von Euler-Verfahren sind möglich durch bessere Quadraturen, z.B. mit der Trapezregel

$$Q_1(f) = \frac{\tau}{2}[f(t_k, y(t_k)) + f(t_{k+1}, y(t_{k+1}))]. \quad (1.13)$$

Eine Wahl von $f(t_{k+1}, y(t_{k+1}))$ durch explizites Euler Verfahren liefert:

Definition 1.10. (Explizites Verfahren von Heun)

Für $k \in \{0, \dots, K-1\}$ setze

$$y_{k+1} := y_k + \frac{\tau}{2}(f_0 + f_1)$$

mit $f_0 := f(t_k, y_k)$ und $f_1 := f(t_{k+1}, y_k + \tau f(t_k, y_k))$.

Eine Wahl $f(t_{k+1}, y(t_{k+1})) \approx f(t_{k+1}, y_{k+1})$ liefert in (1.13):

Definition 1.11. (Crank-Nicolson-Verfahren)

Für $k \in \{0, \dots, K-1\}$ löse das nicht-lineare Gleichungssystem

$$y_{k+1} = y_k + \frac{\tau}{2}(f(t_k, y_k) + f(t_{k+1}, y_{k+1})).$$

Das Crank-Nicolson-Verfahren, sowie das explizite und implizite Euler-Verfahren, sind Spezialfälle des θ -Verfahrens, bei dem $\theta \in [0, 1]$ eine Gewichtung des expliziten und impliziten Anteils mit Gewichten θ bzw. $(1 - \theta)$ gewählt werden.

Während das explizite Euler oder Heun-Verfahren für jedes $\tau > 0$ wohldefinierte Iterierte besitzen (solange die Argumente von f in D liegen), müssen bei impliziten Verfahren in der Regel Zeitschrittweiten-Beschränkungen erfolgen, damit das Verfahren wohldefiniert ist.

Wir zeigen dies für das implizite Euler-Verfahren nach zwei Hilfsaussagen:

Satz 1.12. (Brouwer'scher Fixpunktsatz)

Sei $B := \overline{B_1(0)} \subseteq \mathbb{R}^d$ die abgeschlossene Einheitskugel und $h: B \rightarrow B$ stetig.

Dann besitzt h einen Fixpunkt in B , d.h. $\exists x^* \in B: h(x^*) = x^*$.

Der Brouwer'sche Fixpunktsatz verlangt schwächere Bedingung von der Funktion als Banach'scher Fixpunktsatz. Damit kann man zwar für eine größere Klasse von Funktionen die Existenz eines Fixpunkts nachweisen, aber man hat das iterative Konvergenzverfahren zur Bestimmung des Fixpunkts nicht mehr.

Beweis. (aus Topologie von Prof. Eisermann)

Gegenannahme: Es gilt $f(x) \neq x$ für alle $x \in B$.

Zu jedem $x \in B$ existiert genau ein $t \in \mathbb{R}_+$ so dass $r(x) = f(x) + t(x - f(x))$ in ∂B liegt. Dies definiert eine stetige Abbildung $r: B \rightarrow \partial B$. Für jedes $x \in \partial B$ gilt $r(x) = x$, also ist r eine Retraktion von B auf ∂B . Das ist aber unmöglich, also muss f mindestens einen Fixpunkt haben. \square

Folgerung 1.13. (Existenz von Nullstelle)

Sei $B_\delta := \overline{B_\delta(0)} \subset \mathbb{R}^d$ der abgeschlossene Ball mit Radius $\delta \in \mathbb{R}_+$ um 0.

Sei $g: B_\delta \rightarrow \mathbb{R}^d$ stetig mit

$$\forall x \in \partial B_\delta: \langle g(x), x \rangle \geq 0.$$

Dann besitzt g (mindestens) eine Nullstelle in B_δ .

Beweis. wieder durch Widerspruch:

Angenommen, $\forall x \in B_\delta: g(x) \neq 0$.

Definiere eine Normalisierungsabbildung auf Einheitskugel:

$$h: B_1 \rightarrow B_1, y \mapsto -\frac{g(\delta y)}{\|g(\delta y)\|}.$$

h ist wohldefiniert und stetig.

Mit Brouwer'schem Fixpunktsatz existiert ein $y^* \in B_1$ mit $y^* = h(y^*)$ und somit

$$x^* := \delta y^* = -\frac{\delta g(x^*)}{\|g(x^*)\|} \in \partial B_\delta.$$

Aber

$$\langle g(x^*), x^* \rangle = -\delta \frac{\langle g(x^*), g(x^*) \rangle}{\|g(x^*)\|} = -\delta \|g(x^*)\| < 0$$

ist ein Widerspruch zur Voraussetzung. \square

Mit obigen Vorbereitungen kommen wir nun zur Wohldefiniertheit des impliziten Euler-Verfahrens:

Satz 1.14. (Existenz und Eindeutigkeit der impliziten Euler-Iterierten)

Sei $D = I \times \mathbb{R}^d$ ein Gebiet.

Sei $f: D \rightarrow \mathbb{R}^d$ stetig und erfülle die „einseitige Lipschitz-Bedingung“

$$\exists L \in \mathbb{R} \quad \forall (t, x), (t, y) \in D: \quad \langle f(t, x) - f(t, y), x - y \rangle \leq L \|x - y\|^2.$$

Falls $\tau L < 1$, dann hat die nicht-lineare Gleichung

$$y_{k+1} = y_k + \tau f(t_{k+1}, y_{k+1})$$

genau eine Lösung y_{k+1} zu gegebenem y_k .

Beweis. Zur Existenz der Lösung:

Gesucht ist also eine Nullstelle y_{k+1} von stetiger Funktion

$$g(y) := y - (y_k + \tau f(t_{k+1}, y)) \quad \text{für } y \in \mathbb{R}^d.$$

Die einseitige Lipschitz-Bedingung liefert

$$\begin{aligned} \langle g(x) - g(y), x - y \rangle &= \langle x - y_k - \tau f(t_{k+1}, x) - y + y_k + \tau f(t_{k+1}, y), x - y \rangle \\ &= \|x - y\|^2 - \tau \langle f(t_{k+1}, x) - f(t_{k+1}, y), x - y \rangle \\ &\geq \|x - y\|^2 - \tau L \|x - y\|^2 \\ &= (1 - \tau L) \|x - y\|^2. \end{aligned}$$

Wegen $\tau L < 1$ ist $c := 1 - \tau L > 0$, und wir erhalten:

$$\langle g(x) - g(y), x - y \rangle \geq c \|x - y\|^2. \tag{1.14}$$

Setze $y=0$ in (1.14) und erhalte:

$$\begin{aligned}\langle g(x), x \rangle &= \langle g(x) - g(0), x - 0 \rangle + \langle g(0), x - 0 \rangle \\ &\geq \langle g(x) - g(0), x - 0 \rangle + (-\|g(0)\| \cdot \|x\|) \\ &\geq c \|x - 0\|^2 - \|g(0)\| \cdot \|x\| \\ &= \|x\|(c \|x\| - \|g(0)\|)\end{aligned}$$

wobei die 2. Zeile aus Cauchy-Schwarz-Ungleichung und 3. Zeile aus (1.14) folgen.

Für jedes $\delta \geq \frac{\|g(0)\|}{c}$ und jedes $x \in \partial B_\delta$ gilt dann $\langle g(x), x \rangle \geq 0$.

Dank 1.13 existiert mindestens eine Nullstelle $x \in B_\delta$.

Zur Eindeutigkeit der Lösung:

Seien $x_1, x_2 \in B_\delta$ Nullstellen von g , dann gilt

$$0 = \langle 0, x_1 - x_2 \rangle = \langle g(x_1) - g(x_2), x_1 - x_2 \rangle \geq c \|x_1 - x_2\|^2$$

und damit muss $x_1 = x_2$ sein. \square

Bemerkung. (Einseitige Lipschitz-Bedingung)

- Einseitige Lipschitz-Bedingung ist Abschwächung der Lipschitz-Stetigkeit:
 f Lipschitz-stetig im 2. Arg. $\Rightarrow f$ erfüllt einseitige Lipschitz-Bedingung:
 $\langle f(t, x) - f(t, y), x - y \rangle \leq \|f(t, x) - f(t, y)\| \cdot \|x - y\| \leq L \|x - y\|^2$.
- Es gibt Funktionen, welche die einseitige Lipschitz-Bedingung erfüllen, aber Lipschitz-stetig bzgl. einem größeren/anderen L oder sogar nicht Lipschitz-stetig sind:
 - $f(t, y) = -y$ erfüllt die einseitige Lipschitz-Bedingung mit $L = -1$, aber f ist Lipschitz-stetig bzgl. y mit Lipschitz-Konstante 1.
 - $f(t, y) = -y^3$ erfüllt die einseitige Lipschitz-Bedingung mit $L = 0$, aber f ist nicht Lipschitz-stetig als Funktion auf $I \times \mathbb{R}$.

Als nächstes verallgemeinern wir das Einschritt-Verfahren:

Definition 1.15. (Allgemeines Einschritt-Verfahren)

Sei ein AWP gemäß Definition 1.2 gegeben und oBdA $I = [0, T]$ für ein $T \in \mathbb{R}_+$.

Für ein $K \in \mathbb{N}$ seien Zeitpunkte $0 := t_0 < t_1 < \dots < t_K := T$ zu Gitter $\Delta := \{t_0, \dots, t_K\}$ mit lokalen Schrittweiten $\tau_k := t_{k+1} - t_k$ mit $k \in \{0, \dots, K-1\}$ gegeben.

Ein Verfahren der Form

$$\forall k \in \{0, \dots, K-1\}: \quad y_{k+1} := y_k + \tau_k \phi(t_k, \tau_k, y_k, y_{k+1})$$

heißt Einschrittverfahren (ESV) mit Verfahrensfunktion ϕ , die als stetig vorausgesetzt wird.

Falls ϕ nicht von y_{k+1} abhängt, heißt das Verfahren explizit, und wir schreiben abkürzend $\phi(t_k, \tau_k, y_k)$; sonst heißt das Verfahren implizit.

Also ist explizites / implizites Euler-Verfahren aus Definition 1.9 Spezialart vom Einschrittverfahren.

Wir vereinbaren noch folgende **Notationen**:

- $\{y_k\}_{k=0}^K$ können als Gitterfunktion $y_\Delta: \Delta \rightarrow \mathbb{R}^d$ interpretiert werden, in dem man $y_\Delta(t_k) := y_k$ setzt.
- Wir definieren die Gitterweite eines Gitters $\Delta := \{t_0, \dots, t_K\}$ als

$$\tau_\Delta := \max_{k \in \{0, \dots, K-1\}} \tau_k.$$

- Sei $y: I \rightarrow \mathbb{R}^d$ eine Lösung des AWP. Unser Ziel ist eine Approximation $y(t_k) \approx y_k$, daher definieren wir zu einem Gitter $\Delta := \{t_0, \dots, t_K\}$ die Fehlerfunktion $e_\Delta: \Delta \rightarrow \mathbb{R}^d$ mit

$$\forall k \in \{0, \dots, K\}: e_\Delta(t_k) = e_k := y_k - y(t_k)$$

sowie eine Gitternorm

$$\|e_\Delta\|_\Delta := \max_{k \in \{0, \dots, K\}} \|e_k\|_2,$$

damit wir über Fehler und Konvergenz sprechen können.

Definition 1.16. (Konvergenz, Konvergenzordnung)

Ein ESV heißt konvergent

: \Leftrightarrow Für alle Folgen $(\Delta_K)_{K \in \mathbb{N}}$ mit $\tau_{\Delta_K} \xrightarrow{K \rightarrow \infty} 0$ gilt

$$\lim_{K \rightarrow \infty} \|e_{\Delta_K}\|_\Delta = 0.$$

Das ESV hat (mindestens) Konvergenzordnung $p \in \mathbb{R}_+$

: $\Leftrightarrow \exists C \in \mathbb{R}_+ \exists \tau^* \in \mathbb{R}_+ \forall \Delta \text{ mit } \tau_\Delta \leq \tau^*:$

$$\|e_\Delta\|_\Delta \leq C \tau_\Delta^p.$$

Beispiel. (Konv. Ord. des expliziten Euler-Verfahren beim skalaren AWP)

Wir wollen zeigen, dass das explizite Euler-Verfahren für lineares skalares AWP mit mind. Ordnung 1 konvergiert.

Betrachte der Einfachheit halber äquidistante Gitter Δ (Für allgemeine Gitter gilt die Aussage auch, nur wird die Notation lästiger...).

Ein lineares skalares AWP

$$y' = \lambda y, \quad y(0) = y_0$$

hat eine Lösung $y: \mathbb{R} \rightarrow \mathbb{R}, t \mapsto y_0 e^{\lambda t}$.

Auf äquidistantem Gitter $\Delta := \{0 = t_0, \dots, t_K\}$, also

$$\tau := \tau_0 = \tau_1 = \dots = \tau_{K-1},$$

hat das explizite Euler-Verfahren die Verfahrensfunktion

$$\phi(t_k, \tau_k, y_k, y_{k+1}) = f(y_k, t) = \lambda y_k$$

und somit liefert es für alle $k \in \{0, \dots, K-1\}$:

$$y_{k+1} = y_k + \tau \lambda y_k = (1 + \tau \lambda) y_k = \dots = (1 + \tau \lambda)^{k+1} y_0.$$

Mit Taylor-Entwicklung 1. Grades auf die Lösungsfunktion $y(t) = e^{\lambda t}$ um 0 folgt

$$e^{\lambda t} = 1 + \tau \lambda + \mathcal{O}(\tau^2).$$

Also gilt für alle $k \in \{0, \dots, K-1\}$:

$$y_k = (1 + \tau \lambda)^k y_0 = (e^{\lambda \tau} + \mathcal{O}(\tau^2))^k y_0 = (e^{\lambda k \tau} + \mathcal{O}(k \tau^2)) y_0$$

wobei die zweite Gleichheit aus binomischer Formel folgt.

Da $\Delta := \{0 = t_0, \dots, t_K\}$ äquidistant ist, gilt

$$e^{\lambda k \tau} = e^{\lambda t_k},$$

und somit

$$e_k = y_k - e^{\lambda t_k} y_0 = \mathcal{O}(k \tau^2) y_0,$$

welches zusammen mit $K = T/t$ besagt

$$\|e_\Delta\|_\Delta = \max_{k \in \{0, \dots, K\}} \|e_k\|_2 = \mathcal{O}(K \tau^2) \|y_0\| = T \|y_0\| \mathcal{O}(\tau).$$

\Rightarrow Man erhält Konvergenzordnung mindestens $p = 1$.

Für allgemeine ESV hängt Konvergenz mit dem Begriff der Konsistenz zusammen:

Definition 1.17. (Konsistenz, Konsistenzordnung)

Sei ϕ eine Verfahrensfunktion eines ESV zur Approximation von Lösung eines AWP $y' = f(t, y)$ (mit geeigneter AWB), d.h. für jedes $k \in \{0, \dots, K-1\}$ gilt

$$y_{k+1} = y_k + \tau_k \phi(t_k, \tau_k, y_k, y_{k+1}).$$

Das Verfahren heißt konsistent mit dem AWP, g.d.w.

$$\forall (t, y) \in D: \quad \lim_{\tau \rightarrow \infty} \phi(t, \tau, y, y + \tau) = f(t, y).$$

Für t, τ s.d. $[t, t + \tau] \subseteq I$, $y: I \rightarrow \mathbb{R}^d$ eine Lösung des AWP und $z \in \mathbb{R}^d$ eine Lösung von

$$z = y(t) + \tau \phi(t, \tau, y(t), z)$$

definieren wir den Iterationsfehler

$$\eta(t, \tau) := z - y(t + \tau) = y(t) + \tau\phi(t, \tau, y(t), z) - y(t + \tau).$$

Das Verfahren ϕ heißt konsistent mit (mindestens) Ordnung $p \in \mathbb{R}_+$: $\Leftrightarrow \exists C_p \in \mathbb{R}_+ \forall t, \tau \text{ s.d. } [t, t + \tau] \subseteq I:$

$$\|\eta(t, \tau)\| \leq C_p \tau^{p+1}.$$

Bemerkung. (zu Definition 1.17)

- Die Größe z kann man interpretieren als die Lösung des Verfahrens von der exakten Lösung y zum Zeitpunkt t und Schrittweite τ .
- η misst somit, wie sehr ein einzelner Schritt des Verfahrens bei exakten Anfangswerten $y(t_k)$ die Lösung $y(t_{k+1})$ verpasst.
- Es sind äquivalent

$$\eta(t, \tau) = \mathcal{O}(\tau^{p+1}) \Leftrightarrow \frac{y(t + \tau) - y(t)}{\tau} - \phi(t, \tau, y(t), z) = \mathcal{O}(\tau^p).$$

Die Größe an der rechten Seite nennt man Abschneidefehler, oder auf Englisch „truncation error“.

- Konsistenzordnung eines ESV ist unter Annahme gewisser Differenzierbarkeit von f durch Taylor-Approximation bestimmbar. Die Differenzierbarkeit von f überträgt sich auf die Lösung y , denn $y' = f(t, y)$.

Beispiel. (Konsistenzordnung der beiden Euler-Verfahren für $d = 1$)

Sei $d = 1$ und $f \in C^1$, d.h. $y \in C^2$.

- i. Das explizite Euler-Verfahren lautet

$$\phi(t, \tau, y_k, y_{k+1}) = f(t, y_k).$$

$\Rightarrow \phi(t, 0, y, y) = f(t, y)$, also ist das explizite Euler-Verfahren konsistent.

Mit Taylor-Entwicklung 1. Grades auf $y(t + \tau)$ nach τ folgt

$$y(t + \tau) = y(t) + \tau y'(t) + \frac{1}{2} \tau^2 y''(\xi) \quad \text{mit } \xi \in [t, t + \tau].$$

Wegen $y'(t) = f(t, y)$ ist

$$\eta(t, \tau) = y(t) + \tau f(t, y(t)) - y(t) - \tau f(t, y(t)) - \frac{1}{2} \tau^2 y''(\xi) = -\frac{1}{2} \tau^2 y''(\xi).$$

Das bedeutet

$$|\eta(t, \tau)| = \left| \frac{\tau^2}{2} y''(\xi) \right| \leq \frac{1}{2} \|y''\|_\infty \tau^2,$$

also beträgt die Konsistenzordnung von mindestens $p = 1$.

ii. Das implizite Euler-Verfahren lautet

$$\phi(t, \tau, y_k, y_{k+1}) = f(t + \tau, y_{k+1}).$$

$\Rightarrow \phi(t, 0, y, y) = f(t, y)$, also ist das implizite Euler-Verfahren konsistent.

Der Iterationsfehler lautet dann

$$\eta(t, \tau) := z - y(t + \tau) = y(t) + \tau f(t + \tau, z) - y(t + \tau).$$

Entwicklung von $f(t + \tau, z)$ durch Taylor nach z um $y(t + \tau)$ liefert

$$f(t + \tau, z) = f(t + \tau, y(t + \tau)) + (z - y(t + \tau)) \frac{\partial}{\partial y} f(t + \tau, \bar{\xi}) \text{ mit } \bar{\xi} \in [y(t + \tau), z].$$

Beachte $f(t + \tau, y(t + \tau)) = y'(t + \tau)$ und $z - y(t + \tau) = \eta(t, \tau)$, also

$$\begin{aligned} \eta(t, \tau) &= y(t) + \tau y'(t + \tau) + \tau \eta(t, \tau) \frac{\partial}{\partial y} f(t + \tau, \bar{\xi}) - y(t + \tau) \\ \Leftrightarrow \left(1 - \tau \frac{\partial}{\partial y} f(t + \tau, \bar{\xi})\right) \eta(t, \tau) &= y(t) + \tau y'(t + \tau) - y(t + \tau) \end{aligned} \quad (*)$$

Nochmal Taylor-Entwicklung von $y(t)$ um $t + \tau$ liefert:

$$y(t) = y(t + \tau) - \tau y'(t + \tau) + \frac{1}{2} \tau^2 y''(\xi) \quad \text{mit geeignetem } \xi.$$

Einsetzen in $(*)$:

$$\left(1 - \tau \frac{\partial}{\partial y} f(t + \tau, \bar{\xi})\right) \eta(t, \tau) = \frac{1}{2} \tau^2 y''(\xi).$$

Falls es zusätzlich ein $C_0 \in \mathbb{R}_+$ gibt s.d.

$$\left|1 - \tau \frac{\partial}{\partial y} f(t + \tau, \bar{\xi})\right| > C_0 > 0 \quad (1.15)$$

gilt, dann erhält man

$$\frac{1}{2} \tau^2 \|y''\|_\infty \geq \frac{1}{2} \tau^2 |y''(\xi)| = \left|1 - \tau \frac{\partial}{\partial y} f(t + \tau, \bar{\xi})\right| \cdot |\eta(t, \tau)| \geq C_0 |\eta(t, \tau)|$$

und das bedeutet

$$|\eta(t, \tau)| \leq \frac{1}{2C_0} \|y''\|_\infty \tau^2$$

also Konsistenzordnung von mindestens $p = 1$.

Bemerkung.

- Zu obigem Beispiel, aber im Fall $d > 1$:

Die Konsistenzordnung 1 bei beiden Euler-Verfahren gilt auch für Systeme, also $d > 1$, jedoch darf in Taylor-Darstellung

$$y(t + \tau) = \sum_{j=0}^n \frac{1}{j!} y^{(j)}(t) \tau^j + R_{n+1}(t, \tau)$$

keine Restglieddarstellung

$$R_{n+1}(t, \tau) = \frac{\tau^{n+1}}{(n+1)!} y^{(n+1)}(\xi) \quad \text{mit } \xi \in (t, t+\tau)$$

verwendet werden, denn dies gilt nicht für vektorwertige Funktionen!

Dabei kann man aber komponentenweise betrachten:

$$\forall i \in \{1, \dots, d\}: \quad (R_{n+1}(t, \tau))_i = \frac{\tau^{n+1}}{(n+1)!} (y^{(n+1)})_i(\xi_i) \quad \text{mit } \xi_i \in (t, t+\tau).$$

- (1.15) ist wieder eine Zeitschrittweitebeschränkung, diesmal für Konsistenz; in Satz 1.14 für die Wohldefiniertheit.
- Das Crank-Nicolson-Verfahren ist konsistent mit mindestens der Ordnung $p=2$.

Wie oft in der Numerik implizieren Konsistenz und Stabilität die Konvergenz, also unser Ziel ist jetzt quasi

Konsistenz + Stabilität \Rightarrow Konvergenz von ESV.

Im vorliegenden Fall von ESV erweist sich die Lipschitz-Stetigkeit als der geeignete Stabilitätsbegriff. Damit können wir etwas über die Konvergenz von ESV sagen:

Satz 1.18. (Fehlerschranke und Konvergenz von ESV)

Sei ESV mit Verfahrensfunktion ϕ gegeben.

Das ESV habe mindestens eine Konsistenzordnung p und sei die Funktion

$$\forall k \in \{0, \dots, K\}: \quad \phi_k(y, z) := \phi(t_k, \tau_k, y, z)$$

Lipschitz-stetig bzgl y und z , d.h. es existieren Konstante L, C_0 , s.d. es für geeignete y, y_1, y_2, z, z_1, z_2 gilt

$$\begin{aligned} \|\phi_k(y_1, z) - \phi_k(y_2, z)\| &\leq L \|y_1 - y_2\| \\ \|\phi_k(y, z_1) - \phi_k(y, z_2)\| &\leq C_0 \|z_1 - z_2\| \end{aligned}$$

mit L, C_0 unabhängig von $y, y_1, y_2, z, z_1, z_2, k$.

Falls $C_0 > 0$ und es für die Gitterweite $\tau := \tau_\Delta$ eine Konstante C_1 gebe, s.d.

$$1 - \tau C_0 \geq C_1 > 0, \tag{1.16}$$

dann gilt für $k \in \{0, \dots, K\}$ die Fehlerschranke

$$\|y_k - y(t_k)\| \leq \frac{C_p}{C_2} (\mathbf{e}^{C_2 \tau k} - 1) \tau^p \tag{1.17}$$

mit C_p der Konsistenzkonstante (vgl. Def. 1.17) und $C_2 := \frac{L+C_0}{2}$.

Insbesondere konvergiert das ESV mit Ordnung mindestens p .

Beweis. Per Induktion zeigen wir

$$\|y_k - y(t_k)\| \leq C_p \frac{(1 + C_2 \tau)^k - 1}{C_2} \tau^p, \quad (1.18)$$

und daraus folgt die gewünschte Fehlerschranke, denn

$$(1 + C_2 \tau)^k = \left(1 + \frac{k C_2 \tau}{k}\right)^k \leq e^{k C_2 \tau}.$$

Induktionsanfang $k=1$:

Dank der Konsistenzbedingung ist $\|\eta(t_0, \tau_0)\| \leq C_p \tau_0^{p+1}$, und somit

$$\|y_1 - y(t_1)\| = \|\eta(t_0, \tau_0)\| \leq C_p \tau_0^{p+1} \leq C_p \tau^{p+1} = C_p \frac{(1 + C_2 \tau)^1 - 1}{C_2} \tau^p.$$

Induktionsschritt $k \rightarrow k+1$:

Sei z_{k+1} ein Ergebnis eines Schrittes mit exaktem Wert zu t_k , also

$$z_{k+1} = y(t_k) + \tau \phi_k(y(t_k), z_{k+1}).$$

Mit z_{k+1} kann man eine „Fehleraufspaltung“ für $y_{k+1} - y(t_{k+1})$ durchführen:

$$y_{k+1} - y(t_{k+1}) = y_{k+1} - z_{k+1} + z_{k+1} - y(t_{k+1}).$$

- $z_{k+1} - y(t_{k+1})$ nennt man „neuen lokalen Fehler“ und bei seiner Norm gilt wegen der Konsistenz des Verfahrens

$$\|z_{k+1} - y(t_{k+1})\| \leq C_p \tau^{p+1}.$$

- $y_{k+1} - z_{k+1}$ nennt man „Fehlerfortpflanzung“, und die Abschätzung dieses Terms erfolgt durch mehrere Schritte:

i. Einsetzen von Definition und Umschreiben:

$$\begin{aligned} y_{k+1} - z_{k+1} &= (y_k + \tau_k \phi_k(y_k, y_{k+1})) - (y(t_k) + \tau_k \phi_k(y(t_k), z_{k+1})) \\ &= y_k - y(t_k) + \tau_k (\phi_k(y_k, y_{k+1}) - \phi_k(y(t_k), z_{k+1})). \end{aligned}$$

Den Term $\phi_k(y_k, y_{k+1}) - \phi_k(y(t_k), z_{k+1})$ schreiben wir wieder um:

$$\begin{aligned} &\phi_k(y_k, y_{k+1}) - \phi_k(y(t_k), z_{k+1}) \\ &= \phi_k(y_k, y_{k+1}) - \phi_k(y(t_k), y_{k+1}) + \phi_k(y(t_k), y_{k+1}) - \phi_k(y(t_k), z_{k+1}). \end{aligned}$$

ii. Erste Abschätzung mittels Lipschitz-Stetigkeit von ϕ_k :

$$\begin{aligned} \phi_k(y_k, y_{k+1}) - \phi_k(y(t_k), y_{k+1}) &\leq L \|y_k - y(t_k)\| \\ \phi_k(y(t_k), y_{k+1}) - \phi_k(y(t_k), z_{k+1}) &\leq C_0 \|y_{k+1} - z_{k+1}\|. \end{aligned}$$

Somit gilt

$$\begin{aligned} &\|y_{k+1} - z_{k+1}\| \\ &= \|y_k - y(t_k) + \tau_k (\phi_k(y_k, y_{k+1}) - \phi_k(y(t_k), z_{k+1}))\| \\ &\leq \|y_k - y(t_k)\| + |\tau_k| \|\tau_k (\phi_k(y_k, y_{k+1}) - \phi_k(y(t_k), z_{k+1}))\| \\ &\leq \|y_k - y(t_k)\| + |\tau_k| (L \|y_k - y(t_k)\| + C_0 \|y_{k+1} - z_{k+1}\|) \end{aligned}$$

und mit Umformung erhalten wir

$$\|y_{k+1} - z_{k+1}\| \leq \frac{1 + \tau L}{1 - C_0 \tau} \|y_k - y(t_k)\|.$$

iii. Nachweis von $\frac{1 + \tau L}{1 - C_0 \tau} \leq 1 + C_2 \tau$:

Es ist

$$\begin{aligned} \frac{1 + \tau L}{1 - C_0 \tau} &\leq 1 + C_2 \tau \\ \Leftrightarrow 1 + \tau L &\leq 1 + (C_2 - C_0) \tau - C_2 C_0 \tau^2 \\ \Leftrightarrow L &\leq C_2 - C_0 - C_2 C_0 \tau \\ \Leftrightarrow L + C_0 &\leq C_2 (1 - C_0 \tau). \end{aligned}$$

Mit $C_2 := \frac{L + C_0}{C_1}$ und $1 - C_0 \tau \geq C_1$ ist die letzte Umgleichung erfüllt:

$$C_2 (1 - C_0 \tau) \geq \frac{L + C_0}{C_1} (1 - C_0 \tau) \geq \frac{L + C_0}{C_1} C_1 = L + C_0.$$

iv. Einsetzen von iii. in ii. und Nutzen von Induktionsvoraussetzung:

$$\begin{aligned} \|y_{k+1} - z_{k+1}\| &\leq \frac{1 + \tau L}{1 - C_0 \tau} \|y_k - y(t_k)\| && \text{(ii.)} \\ &\leq (1 + C_2 \tau) \|y_k - y(t_k)\| && \text{(iii.)} \\ &\leq (1 + C_2 \tau) C_p \frac{(1 + C_2 \tau)^k - 1}{C_2} \tau^p. && \text{(I.V.)} \end{aligned}$$

Für den Gesamtfehler erhalten wir dann:

$$\begin{aligned} \|y_{k+1} - y(t_{k+1})\| &\leq \|y(t_{k+1}) - z_{k+1}\| + \|z_{k+1} - y_{k+1}\| \\ &\leq C_p \tau^{p+1} + (1 + C_2 \tau) C_p \frac{(1 + C_2 \tau)^k - 1}{C_2} \tau^p \\ &= C_p \left(\frac{(1 + C_2 \tau)^{k+1} - (1 + C_2 \tau)}{C_2} + \tau \right) \tau^p \\ &= C_p \left(\frac{(1 + C_2 \tau)^{k+1} - 1}{C_2} \right) \tau^p. \end{aligned}$$

Also wird (1.18) nachgewiesen. □

Bemerkung.

- Der Satz 1.18 nimmt keine Voraussetzung für das Gitter an. Damit haben wir insbesondere die Konvergenz vom expliziten und impliziten Euler-Verfahren für allgemeine Gitter (nicht unbedingt äquidistant) und für allgemeine rechte Seite (nicht unbedingt skalare Funktion) gezeigt.
- Falls ESV explizit ist, hängt ϕ_k nicht von z ab, also kann man $C_0 = 0$, $C_1 = 1$ und $C_2 = L$ wählen.

- Falls die Konstanten C_2, C_p in (1.17) bekannt sind, kann dies zur Steuerung von τ verwendet werden, um gewünschte Genauigkeit zu erreichen, also:
Zu $\delta > 0$ wähle

$$\tau \leq \sqrt{\frac{\delta C_2}{C_p} (\mathbf{e}^{C_2 T} - 1)^{-1}},$$

dann gilt für alle $k \in \{0, \dots, K\}$:

$$\|y_k - y(t_k)\| \leq \delta.$$

Dies ist eine theoretische Schranke und Rundungsfehler werden dabei nicht betrachtet.

- Im Fall von Rundungsfehlern kann man die Rechengenauigkeit ε in Konsistenzfehler erfassen

$$\|z_{k+1} - y(t_{k+1})\| \leq C_p \tau^{p+1} + \varepsilon \quad (1.19)$$

und die Fehlerschranke entsprechend anpassen. Aus (1.19) folgt, dass es keinen Sinn macht, τ wesentlich kleiner als $\sqrt[p+1]{\frac{\varepsilon}{C_p}}$ zu wählen, denn dann wird Rechenaufwand groß aber da ε bei Konsistenzfehler dominiert, wird kann der gesamte Konsistenzfehler nicht weiter verringert werden.

- Die Ungleichung (1.16) ist wieder eine Zeitschrittweitenbedingung und dies ist erfüllt, g.d.w. $\tau C_0 < 1$. Dies ist hinreichend für die Wohldefiniertheit eines allgemeinen impliziten ESV.

Für numerische Verfahren ist es oft vorteilhaft, qualitative Eigenschaften der exakten Lösung in der numerischen Approximation wiederzuspiegeln. Unter anderem interessieren wir uns für die **Beschränktheit der Lösung**. Dies nennt man auch die **Stabilität von ESV** und werden wir im folgenden untersuchen.

Definition 1.19. (Testgleichung)

Wir nennen skalare ODE

$$y'(t) = \lambda y(t), \quad y(0) = y_0 \quad (1.20)$$

für $\lambda \in \mathbb{C}$, $y_0 \in \mathbb{C}$, $t \in \mathbb{R}_{\geq 0}$ Testgleichung mit exakter Lösung

$$y(t) = y_0 \mathbf{e}^{\lambda t}.$$

Eine Lösung der Testgleichung (analytisch oder erhalten durch numerische Verfahren) heißt instabil, g.d.w. gilt

$$\lim_{t \rightarrow \infty} |y(t)| = \infty.$$

Entsprechend heißt eine Lösung der Testgleichung (analytisch oder erhalten durch numerische Verfahren) stabil g.d.w. gilt

$$\lim_{t \rightarrow \infty} |y(t)| < \infty.$$

Bemerkung.

- Man sieht leicht: Die analytische Lösung der Testgleichung ist stabil g.d.w. $\operatorname{Re}(\lambda) \leq 0$ gilt, und instabil g.d.w. $\operatorname{Re}(\lambda) > 0$ gilt.
- Ziel ist also für ein ESV, im Fall $\operatorname{Re}(\lambda) \leq 0$, beschränkte Lösungen bei Anwendung auf (1.20) zu produzieren.
- Explizites Euler-Verfahren mit konstantem τ angewandt auf (1.20) liefert

$$y_{k+1} = y_k + \tau \lambda y_k = (1 + \tau \lambda) y_k = (1 + \tau \lambda)^{k+1} y_0.$$

Mit $R(z) := 1 + z$ ist

$$y_{k+1} = R(\tau \lambda)^{k+1} y_0$$

und zudem gilt

$$\{y_k\}_{k=1}^K \text{ beschränkt} \Leftrightarrow |1 + \tau \lambda| \leq 1 \Leftrightarrow |R(\tau \lambda)| \leq 1.$$

D.h. für $z = \lambda \tau \in \overline{B_1(-1)}$ ist die Lösung der Testgleichung stabil, sonst instabil.

Definition 1.20. (Stabilitätsfunktion, Stabilitätsgebiet)

Sei ein ESV für AWP (1.20) in der Form

$$y_{k+1} = R(\tau_k \lambda) y_k$$

gegeben mit einer Funktion

$$R: \mathbb{C} \rightarrow \mathbb{C} \cup \{\infty\}.$$

Die Funktion R nennen wir die Stabilitätsfunktion und die Menge

$$S := \{z \in \mathbb{C} : |R(z)| \leq 1\}$$

nennen wir das Stabilitätsgebiet des ESV.

Beispiel.

- Explizites Euler-Verfahren: $R(z) = 1 + z$ sowie $S = \overline{B_1(-1)}$.
- Implizites Euler-Verfahren:

$$y_{k+1} = y_k + \tau_k \lambda y_{k+1} \Leftrightarrow (1 - \tau_k \lambda) y_{k+1} = y_k \Leftrightarrow y_{k+1} = (1 - \tau_k \lambda)^{-1} y_k$$

also ist $R(z) = (1 - z)^{-1}$, und somit gilt

$$\{y_k\}_{k=1}^K \text{ beschränkt} \Leftrightarrow |R(\tau_k \lambda)| \leq 1 \Leftrightarrow 1 \leq |1 - \tau_k \lambda|$$

und d.h. $S = \mathbb{C} \setminus B_1(1)$.

Bei den obigen Beispiele sieht man, dass es beim impliziten Euler-Verfahren jede Wahl von Schrittweiten $\tau > 0$ zu stabiler Lösung der Testgleichung führt. Diese schöne Eigenschaft verdient einen eigenen Namen:

Definition 1.21. (A-Stabilität)

Ein ESV mit Stabilitätsgebiet S heißt A-stabil (absolut stabil)

$$\Leftrightarrow \{z \in \mathbb{C} \mid \operatorname{Re}(z) \leq 0\} \subseteq S.$$

Bemerkung.

- Per Definition ist ein ESV A-stabil, g.d.w. das ESV bei einer Testgleichung (1.20) mit $\operatorname{Re}(\lambda) < 0$ für alle Schrittweiten $\tau > 0$ immer eine beschränkte Folge liefert.
- Implizites Euler-Verfahren ist A-stabil.
- Explizites Euler-Verfahren ist hingegen nicht A-stabil.

Für $\operatorname{Re}(\lambda) \leq 0$ liefert das Explizite Euler-Verfahren nur beschränkte Folge, falls $\tau \lambda \in S = \overline{B_1(-1)}$, d.h. $|1 + \tau \lambda| \leq 1$ als Zeitschrittweitenbedingung, und dies ist z.B. verletzt, falls $\tau > \frac{2}{|\lambda|}$.

Im Bezug auf A-Stabilität ist der folgende Begriff nahliegend:

Definition 1.22. (Isometrieerhaltung)

Ein ESV heißt isometrieerhaltend, g.d.w. die Stabilitätsfunktion R jeden Punkt der imaginären Achse auf dem Einheitskreis abbildet, also

$$\forall z \in \mathbb{C} \text{ mit } \operatorname{Re}(z) = 0: |R(z)| = 1.$$

Bemerkung.

- Die anschauliche Idee bei Def. 1.22 ist Längenerhaltung:
Für $\lambda = i$, d.h. $y(t) = y_0 e^{it}$ die Lösung von Testgleichung $y' = iy$, gilt $|y(t)| = |y_0|$ und bei isometrieerhaltenden Verhalten

$$y_{k+1} = R(\tau_k \lambda) y_k \Rightarrow |y_{k+1}| = |y_k| = \dots = |y_0|$$

also wird die Länge von y_0 erhalten.

- Gegenbeispiel mit explizitem Euler-Verfahren:

Sei $\operatorname{Re}(z) = 0$, z.B. $z = i\omega$ mit $\omega \in \mathbb{R}$, dann gilt

$$|R(z)| = |1 + i\omega| = \sqrt{1 + \omega^2} \neq 1 \text{ für } \omega \neq 0$$

also nicht isometriehaltend.

- Positive Beispiele folgen später bei speziellen „Runge-Kutta-Verfahren“ und bei „Gauß-Kollokationsverfahren“.

Bemerkung. (Rechtfertigung der Einschränkung auf Testgleichung)

Die Einschränkung auf Testproblem bzw. Testgleichung ist sinnvoll, denn man kann alle nicht-lineare Probleme mit $y: I \rightarrow \mathbb{R}^d$ und $y' = f(t, y)$ wobei f differenzierbar lokal linearisieren.

Wir betrachten ein Beispiel eines autonomen Systems:

$$y'(t) = f(y(t)) \approx f(y_0) + D_y f(y_0)(y(t) - y(0))$$

mit $D_y f(y_0) \in \mathbb{R}^{d \times d}$. Dies verhält sich lokal um t_0 wie Linearisierung

$$y'(t) = A y(t) + b \quad \text{mit} \quad A = D_y f(y_0), b = f(y_0) - D_y f(y_0) y_0.$$

Unter Annahmen $b=0$ und $A=U\Lambda U^{-1}$ diagonalisierbar mit $\Lambda=\text{diag}(\lambda_1, \dots, \lambda_d)$ folgt mit Lösung des homogenen LGS

$$\tilde{y}' = A \tilde{y}, \quad \tilde{y}(t_0) = y_0$$

und das bedeutet

$$\begin{aligned} y(t) \approx \tilde{y}(t) &= \exp(A(t-t_0)) y_0 \\ &= U \text{diag}(\mathbf{e}^{\lambda_1(t-t_0)}, \dots, \mathbf{e}^{\lambda_d(t-t_0)}) U^{-1} y_0. \end{aligned}$$

Mit einer Transformation $z := U^{-1} \tilde{y}$ gilt

$$z' = \Lambda z,$$

also $z'_i = \lambda_i z_i$ für alle $i \in \{1, \dots, d\}$, d.h. d separate Gleichungen von Typ der Testgleichung.

Als nächstes behandeln wir eine allgemeine Klasse von Verfahren, das sogenannte Runge-Kutta-Verfahren, mit dem wir die bisherigen Verfahren (Explizites/Implizites Euler, Heun, ...) zusammenfassen können.

Die **Motivation des Runge-Kutta-Verfahren** besteht darin:

- Wir wollen DGLn mit Quadratur der Volterra'schen Integralgleichung

$$\begin{aligned} y_{k+1} \approx y(t_{k+1}) &= y(t_k) + \int_{t_k}^{t_{k+1}} f(s, y(s)) ds \\ &\approx y_k + \tau_k \sum_{i=1}^s b_i f(t_k + c_i \tau_k, v_i) \end{aligned}$$

mit $c_i \in [0, 1]$, $i \in \{1, \dots, s\}$ und Unbekannten $v_i \approx y(t_k + c_i \tau_k)$ approximieren.

- Idee für v_i : Wieder mittels Quadratur finden, so dass viele f -Auswertungen wieder verwendet werden können:

$$v_i \approx y_k + \int_{t_k}^{t_k + c_i \tau_k} f(s, y(s)) ds \approx y_k + \tau_k \sum_{j=1}^s a_{ij} f(t_k + c_j \tau_k, v_j).$$

- Damit konstante Funktionen auch exakt integriert werden können, scheint es sinnvoll zu verlangen

$$\sum_{i=1}^s b_i = 1 \quad \text{sowie} \quad \forall i \in \{1, \dots, s\}: \sum_{j=1}^s a_{ij} = c_i.$$

Definition 1.23. (Runge-Kutta-Verfahren)

Sei AWP auf $I = [0, T]$, $K \in \mathbb{N}$, $\{t_k\}_{k=0}^K$, $\{\tau_k\}_{k=0}^{K-1}$ wie in Definition 1.15.

Zu Stufenzahl $s \in \mathbb{N}$ seien Knoten $c_i \in \mathbb{R}$, Gewichte $b_i \in \mathbb{R}$ sowie $a_{ij} \in \mathbb{R}$ gegeben für $i, j \in \{1, \dots, s\}$.

Wir definieren das zugehörige Runge-Kutta-Verfahren (RK-Verfahren) durch

$$\forall k \in \{0, \dots, K-1\}: \quad y_{k+1} := y_k + \tau_k \sum_{i=1}^s b_i f(t_k + c_i \tau_k, v_i) \quad (1.21)$$

mittels der Stufen $v_1, \dots, v_s \in \mathbb{R}^d$, also

$$\forall i \in \{1, \dots, s\}: \quad v_i := y_k + \tau_k \sum_{j=1}^s a_{ij} f(t_k + c_j \tau_k, v_j). \quad (1.22)$$

Hierbei sei vorausgesetzt, dass

$$\sum_{i=1}^s b_i = 1 \quad \text{und} \quad \forall i \in \{1, \dots, s\}: \sum_{j=1}^s a_{ij} = c_i. \quad (1.23)$$

Beispiel.

- Im Fall $s = 1$, $b_1 = 1$, $c_1 = 0$ und $a_{11} = 0$ erhalten wir

$$y_{k+1} = y_k + \tau_k f(t_k, v_1) \quad \text{mit} \quad v_1 = y_k + \tau_k \cdot 0 \cdot f(t_k, v_1) = y_k$$

also das explizite Euler-Verfahren.

- Im Fall $s = 1$, $b_1 = 1$, $c_1 = 1$ und $a_{11} = 1$ erhalten wir

$$y_{k+1} = y_k + \tau_k f(t_k + \tau_k, v_1) \quad \text{mit} \quad v_1 = y_k + \tau_k \cdot 1 \cdot f(t_k + \tau_k, v_1) = y_{k+1}$$

also das implizite Euler-Verfahren.

- Ähnlich für Verfahren von Heun, Crank-Nicolson-Verfahren, θ -Verfahren als Spezialfälle.

Bemerkung.

- Falls $\{v_i\}_{i=1}^s$ wohldefiniert und stetig von y_k abhängig sind, so ist ein RK-Verfahren ein ESV mit stetiger Verfahrensfunktion

$$\phi(t_k, \tau_k, y_k, y_{k+1}) = \sum_{i=1}^s b_i f(t_k + c_i \tau_k, v_i)$$

unabhängig von y_{k+1} , also im Sinne von Definition 1.15 ein „explizites ESV“ und schreiben $\phi(t_k, \tau_k, y_k)$. Wir nennen das RK-Verfahren trotzdem implizit, falls Gleichungssysteme für v_i gelöst werden müssen.

- Bedingung an b_i in (1.23) impliziert Konsistenz:

Für $\tau_k = 0$, $y_k = y$ gilt:

$$\phi(t, 0, y, y) = \sum_{i=1}^s b_i f(t + c_i \cdot 0, v_i)$$

mit

$$v_i = y + 0 \cdot \sum_{j=1}^s a_{ij} f(t + c_i \cdot 0, v_j) = y.$$

Daher gilt

$$\phi(t, 0, y, y) = \sum_{i=1}^s b_i f(t + c_i \cdot 0, v_i) = \sum_{i=1}^s b_i f(t, y) = f(t, y)$$

da $\sum_{i=1}^s b_i = 1$ nach (1.23).

Bemerkung. (Butcher-Tableau)

- Zu RK-Verfahren setze

$$A := (a_{ij})_{i,j=1}^s \in \mathbb{R}^{s \times s}, \quad c := (c_i)_{i=1}^s \in \mathbb{R}^s, \quad b := (b_i)_{i=1}^s \in \mathbb{R}^s.$$

Dann schreiben wir die Koeffizienten als Butcher Tableau

$$\begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array} \quad \text{bzw. kompakter} \quad \frac{c}{b^T} \mid A.$$

- Falls A (oBdA nach Zeilen-/Spaltenpermutation) untere Dreiecksmatrix mit Nulldiagonale ist, dann ist das Verfahren explizit, sonst implizit. Falls A (oBdA nach Zeilen-/Spaltenpermutation) untere Dreiecksmatrix $\neq 0$, so nennen wir es diagonal-implizites RK-Verfahren („DIRK-Verfahren“).
- Für explizites RK-Verfahren werden einfach v_1, \dots, v_s der Reihe nach berechnet durch f -Auswertung & (1.22).
- Für implizites RK-Verfahren sind alle Stufen $\{v_i\}_{i=1}^s$ unbekannt, diese werden durch ein Nullstellenproblem in $\mathbb{R}^{d \cdot s}$ bestimmt, dann wird y_{k+1} mit (1.21) berechnet. Details dazu kommen später.

- Für DIRK-Verfahren werden die Reihe nach v_1, \dots, v_s durch s Nullstellenprobleme in \mathbb{R}^d bestimmt, anschließend y_{k+1} aus (1.21).
- Bei expliziten RK-Verfahren werden in Butcher-Tableau kann man die Diagonale sowie oberen Dreiecksteil weglassen.

Beispiel. (Bekannte ESV als RK-Verfahren)

- $s = 1$: Explizites Euler Implizites Euler

$$\begin{array}{c|c} 0 & 0 \\ \hline 1 & \end{array} \quad \begin{array}{c|c} 1 & 1 \\ \hline 1 & \end{array}$$

- $s = 2$: Das explizite Verfahren von Heun (Def. 1.10) lautet

$$y_{k+1} = y_k + \tau_k \frac{1}{2} (f_0 + f_1) \quad \text{mit} \quad f_0 = f(t_k, y_k), \quad f_1 = f(t_k + \tau_k, y_k + \tau_k f_0).$$

Durch Beobachtung und Vergleich mit (1.21) erhalten wir

$$b_1 = b_2 = \frac{1}{2}, \quad v_1 = y_k, \quad v_2 = y_k + \tau_k f_0, \quad c_1 = 0, \quad c_2 = 1.$$

$$v_1 = y_k \Rightarrow a_{11} = 0, a_{12} = 0.$$

$$v_2 = y_k + \tau_k f_0 \Rightarrow a_{21} = 1, a_{22} = 0.$$

Somit erhalten wir als Butcher-Tableau

$$\begin{array}{c|cc} 0 & & \\ \hline 1 & 1 & \\ \hline & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

- $s = 2$: Das Crank-Nicolson (Def. 1.11) lautet

$$y_{k+1} = y_k + \tau_k \frac{1}{2} (f(t_k, y_k) + f(t_k + \tau_k, y_{k+1})).$$

Durch Beobachtung und Vergleich mit (1.21) erhalten wir

$$b_1 = b_2 = \frac{1}{2}, \quad v_1 = y_k, \quad v_2 = y_{k+1}, \quad c_1 = 0, \quad c_2 = 1.$$

$$v_1 = y_k \Rightarrow a_{11} = 0, a_{12} = 0.$$

$$v_2 = y_{k+1} = y_k + \tau_k \frac{1}{2} (f(t_k, y_k) + f(t_k + \tau_k, y_{k+1})) \Rightarrow a_{21} = a_{22} = \frac{1}{2}.$$

Somit erhalten wir als Butcher-Tableau

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \hline 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

- In allen Beispielen sind die Bedingungen (1.23) erfüllt, also $\sum_{i=1}^s b_i = 1$ und $\forall i \in \{1, \dots, s\}: \sum_{j=1}^s a_{ij} = c_i$.
- Später werden wir noch bessere Verfahren kennenlernen.

Nächste Frage: Wie „gut“ können RK-Verfahren sein?

Dazu bemerken wir zunächst, dass die Konsistenzordnung von RK-Verfahren eng mit Quadraturen zusammenhängt:

Satz 1.24. (Konsistenzordnung & Quadratur)

Falls ein RK-Verfahren der Form von (1.21), also

$$y_{k+1} = y_k + \tau_k \sum_{i=1}^s b_i f(t_k, c_i \tau_k, v_i)$$

mit $c_1, \dots, c_s \in [0, 1]$, eine Konsistenzordnung p besitzt, so ist die Quadratur

$$Q(g) := \sum_{i=1}^s b_i g(c_i) \quad \text{für } g \in C([0, 1], \mathbb{R})$$

auf \mathbb{P}_{p-1} exakt, d.h.

$$\forall g \in \mathbb{P}_{p-1}: \quad Q(g) = \int_0^1 g(t) dt.$$

Beweis. Wir zeigen die Exaktheit der Quadratur für alle Monome t^0, \dots, t^{p-1} :

Wähle ein $m \in \{0, \dots, p-1\}$ und betrachte ein AWP

$$y'(t) = t^m, \quad y(0) = 0$$

mit Lösung $y(t) = \frac{1}{m+1} t^{m+1}$.

Für erste Iterierte y_1 gilt (mit $\tau := \tau_1$) wegen Konsistenz

$$|y(\tau) - y_1| = \mathcal{O}(\tau^{p+1}) \tag{\#}$$

für $\tau \rightarrow 0$.

Andererseits mit $y(t) = \frac{1}{m+1} t^{m+1}$ und Verfahrensfunktion $f(t, y) = t^m$ gilt

$$\begin{aligned} |y(\tau) - y_1| &= \left| \frac{\tau^{m+1}}{m+1} - \left(0 + \tau \sum_{i=1}^s f(0 + c_i \tau, v_i) \right) \right| \\ &= \tau^{m+1} \left| \frac{1}{m+1} - \sum_{i=1}^s b_i c_i^m \right| \\ &= \tau^{m+1} \left| \int_0^1 t^m dt - Q(t^m) \right|. \end{aligned}$$

Also gilt mit (#)

$$\left| \int_0^1 t^m dt - Q(t^m) \right| = |y(\tau) - y_1| \tau^{-(m+1)} = \mathcal{O}(\tau^{p+1-(m+1)}) = \mathcal{O}(\tau^{p-m}).$$

Wegen $p > m$ geht die rechte Seite für $\tau \rightarrow 0$ gegen 0, und die linke Seite ist unabhängig von τ , also gilt

$$\int_0^1 t^m dt - Q(t^m) = 0 \quad \Leftrightarrow \quad \int_0^1 t^m dt = Q(t^m). \quad \square$$

Wir bemerken hier, dass die Umkehrung vom Satz 1.24 hier (noch) nicht gilt, weil die Koeffizienten v_i unspezifiziert sind. Diese Aspekte werden wir später bei Kollokationsverfahren nochmal betrachten.

Aus 1.24 und Eigenschaften von Quadraturen folgt sofort:

Folgerung 1.25. (Grenze der Konsistenzordnung)

Kein RK-Verfahren mit s Stufen kann Konsistenzordnung $p > 2s$ besitzen.

Beweis. Sonst wäre mit 1.24 Q exakt auf \mathbb{P}_{2s} , aber dies ist unmöglich, denn nach NUM I ist Q mit s Punkten höchstens auf \mathbb{P}_{2s-1} exakt. \square

Die Frage nach Existenz eines RK-Verfahrens mit Konsistenzordnung $2s$ werden wir später durch Konstruktion solches Verfahrens positiv beantworten.

Satz 1.26. (Autonome / Nicht-autonome AWP)

Wenn ein RK-Verfahren für alle autonomen DGLn Konsistenzordnung p besitzt, dann ist das Verfahren auch konsistent mit Ordnung p für nicht-autonome DGLn.

Der Satz 1.26 gilt, weil man jede nicht-autonome DGL zu einer autonomen DGL umwandeln kann (vgl. Aufgabe 2 vom Blatt 2).

Als nächstes charakterisieren wir Konsistenzordnung ≤ 3 . Dazu führen wir das „Hadamard Produkt“, also eintragsweise Multiplikation von Vektoren, ein:

$$\forall c \in \mathbb{R}^s: \quad c \odot c := (c_i \cdot c_i)_{i=1}^s = \text{diag}(c_1, \dots, c_s)c,$$

sowie eine Notation für den Vektor mit konstantem Eintrag 1:

$$\mathbf{1}_s := (1, \dots, 1)^T \in \mathbb{R}^s.$$

Satz 1.27. (Konsistenzbedingungen)

Sei ein RK-Verfahren bzgl. $A \in \mathbb{R}^{s \times s}$ und $b, c \in \mathbb{R}^s$ für ein AWP $y' = f(t, y)$, $y(t_0) = y_0$ gegeben.

Dann hat das Verfahren mindestens Konsistenzordnung

i. $p = 1$.

ii. $p = 2$ g.d.w. gilt

$$\langle b, c \rangle = \sum_{i=1}^s b_i c_i = \frac{1}{2}. \quad (1.24)$$

iii. $p=3$ g.d.w. zusätzlich zu (1.24) noch zwei Bedingungen gellten:

$$\langle b, c \odot c \rangle = \sum_{i=1}^s b_i c_i^2 = \frac{1}{3} \quad \text{und} \quad \langle b, Ac \rangle = \sum_{i=1}^s b_i \sum_{j=1}^s a_{ij} c_j = \frac{1}{6},$$

jeweils unter Annahme ausreichender Differenzierbarkeit von $f \in C^p$.

Beweis. Dank Satz 1.26 reicht es, nur autonome DGL $y'(t) = f(y(t))$ zu betrachten.

Außerdem betrachten wir nur den Fall skalarer Funktionen (sonst funktioniert es auch analog, nur technischer).

Sei oBdA $t_0 := 0$, $T < \infty$, sowie $f \in C^3$ (sonst gleiche Argumentation, nur ohne höhere Ableitungen) also $y \in C^4$, und f auf D beschränkt.

Beweisstrategie: Wir formen die Lösung y sowie die Näherungslösung y_1 geschickt um und machen Koeffizientenvergleich.

i. Untersuchung von y :

Ableiten von $y' = f(y)$ liefert:

$$y'' = f'(y)y' = f'(y)f(y),$$

$$y''' = (f''(y)y')f(y) + f'(y)(f'(y)y') = f''(y)f(y)^2 + f'(y)^2f(y).$$

Damit können wir die Taylor-Entwicklung von Lösung des AWP y nach τ um 0 umschreiben als

$$\begin{aligned} y(\tau) &= y(0) + y'(0)\tau + \frac{1}{2}y''(0)\tau^2 + \frac{1}{6}y'''(0)\tau^3 + \mathcal{O}(\tau^4) \\ &= y_0 + f(y_0)\tau + \frac{1}{2}f'(y_0)f(y_0)\tau^2 \\ &\quad + \frac{1}{6}(f''(y_0)f(y_0)^2 + f'(y_0)^2f(y_0))\tau^3 + \mathcal{O}(\tau^4). \end{aligned}$$

ii. Untersuchung von $f(v_i)$ als Vorbereitung für Untersuchung von y_1 :

Taylor-Entwicklung 2. Grades von $f(v_i)$ um y_0 ergibt:

$$f(v_i) = f(y_0) + f'(y_0)(v_i - y_0) + \frac{1}{2}f''(y_0)(v_i - y_0)^2 + \mathcal{O}((v_i - y_0)^3).$$

Mit der Definition $v_i = y_0 + \tau \sum_{j=1}^s a_{ij} f(v_j)$ gilt

$$v_i - y_0 = \tau \sum_{j=1}^s a_{ij} f(v_j) = \mathcal{O}(\tau)$$

da f beschränkt ist, und wir setzen dies in die obige Entwicklung ein:

$$\begin{aligned} f(v_i) &= f(y_0) + f'(y_0)(v_i - y_0) + \frac{1}{2}f''(y_0)(v_i - y_0)^2 + \mathcal{O}((v_i - y_0)^3) \\ &= f(y_0) + f'(y_0)\tau \sum_{j=1}^s a_{ij} f(v_j) + \frac{1}{2}f''(y_0) \left(\tau \sum_{j=1}^s a_{ij} f(v_j) \right)^2 + \mathcal{O}(\tau^3). \end{aligned}$$

Die $f(v_j)$ auf der rechten Seite lösen wir nochmal mit Taylor-Entwicklung auf: $f(v_j)$ zu $f'(y_0)$ werden durch Taylor 1. Grades um y_0 ersetzt

$$\begin{aligned} f(v_j) &= f(y_0) + f'(y_0)(v_j - y_0) + \mathcal{O}((v_j - y_0)^2) \\ &= f(y_0) + f'(y_0)\tau \sum_{l=1}^s a_{jl}f(y_0) + \mathcal{O}(\tau^2), \end{aligned}$$

und $f(v_j)$ zu $f''(y_0)$ werden durch Taylor 0. Grades um y_0 ersetzt

$$f(v_j) = f(y_0) + \mathcal{O}(v_j - y_0) = f(y_0) + \mathcal{O}(\tau).$$

Taylor-Ergebnisse 0. & 1. Grades von $f(v_j)$ zurück in Taylor-Ergebnis 2. Grades von $f(v_i)$ einzusetzen ergibt

$$\begin{aligned} f(v_i) &= f(y_0) + f'(y_0)\tau \sum_{j=1}^s a_{ij}f(v_j) + \frac{1}{2}f''(y_0) \left(\tau \sum_{j=1}^s a_{ij}f(v_j) \right)^2 + \mathcal{O}(\tau^3) \\ &= f(y_0) + f'(y_0)\tau \sum_{j=1}^s a_{ij} \left(f(y_0) + f'(y_0)\tau \sum_{l=1}^s a_{jl}f(y_0) + \mathcal{O}(\tau^2) \right) \\ &\quad + \frac{1}{2}f''(y_0) \left(\tau \sum_{j=1}^s a_{ij}(f(y_0) + \mathcal{O}(\tau)) \right)^2 + \mathcal{O}(\tau^3) \\ &= f(y_0) + f'(y_0) \sum_{j=1}^s a_{ij}f(y_0)\tau + f'(y_0) \sum_{j=1}^s a_{ij}f'(y_0) \sum_{l=1}^s a_{jl}f(y_0)\tau^2 \\ &\quad + \frac{1}{2}f''(y_0) \left(\sum_{j=1}^s a_{ij}f(y_0) \right)^2 \tau^2 + \mathcal{O}(\tau^3). \end{aligned}$$

Beachte $c_i = \sum_{j=1}^s a_{ij}$ bzw. $c_j = \sum_{l=1}^s a_{jl}$, und somit

$$\begin{aligned} f(v_i) &= f(y_0) + f'(y_0) \sum_{j=1}^s a_{ij}f(y_0)\tau + f'(y_0) \sum_{j=1}^s a_{ij}f'(y_0) \sum_{l=1}^s a_{jl}f(y_0)\tau^2 \\ &\quad + \frac{1}{2}f''(y_0) \left(\sum_{j=1}^s a_{ij}f(y_0) \right)^2 \tau^2 + \mathcal{O}(\tau^3) \\ &= f(y_0) + f'(y_0)c_i f(y_0)\tau + f'(y_0) \sum_{j=1}^s a_{ij}f'(y_0)c_j f(y_0)\tau^2 \\ &\quad + \frac{1}{2}f''(y_0)(c_i f(y_0))^2 \tau^2 + \mathcal{O}(\tau^3) \\ &= f(y_0) + c_i f'(y_0)f(y_0)\tau \\ &\quad + \sum_{j=1}^s a_{ij}c_j f'(y_0)^2 f(y_0)\tau^2 \\ &\quad + c_i^2 \frac{1}{2}f''(y_0)f(y_0)^2 \tau^2 + \mathcal{O}(\tau^3). \end{aligned}$$

iii. Untersuchung von y_1 :

Mit ii. gilt daher

$$\begin{aligned}
 y_1 &= y_0 + \tau \sum_{i=1}^s b_i f(v_i) \\
 &= y_0 + \sum_{i=1}^s b_i f(y_0) \tau \\
 &\quad + \sum_{i=1}^s b_i c_i f'(y_0) f(y_0) \tau^2 \\
 &\quad + \sum_{i=1}^s b_i \sum_{j=1}^s a_{ij} c_j f'(y_0)^2 f(y_0) \tau^3 \\
 &\quad + \sum_{i=1}^s b_i c_i^2 \frac{1}{2} f''(y_0) f(y_0)^2 \tau^3 + \mathcal{O}(\tau^4).
 \end{aligned}$$

iv. Koeffizientenvergleich von iii. mit i.:

Man vergleicht das Ergebnis aus i.

$$\begin{aligned}
 y(\tau) &= y_0 + f(y_0) \tau \\
 &\quad + \frac{1}{2} f'(y_0) f(y_0) \tau^2 \\
 &\quad + \frac{1}{6} (f''(y_0) f(y_0)^2 + f'(y_0)^2 f(y_0)) \tau^3 + \mathcal{O}(\tau^4)
 \end{aligned}$$

mit dem Ergebnis von iii. und stellt fest, dass:

- $|y(\tau) - y_1| = \mathcal{O}(\tau^2)$, also Konsistenzordnung $p = 1$ falls $\sum_{i=1}^s b_i = 1$.
- $|y(\tau) - y_1| = \mathcal{O}(\tau^3)$, also Konsistenzordnung $p = 2$ falls zusätzlich noch $\langle b, c \rangle = \frac{1}{2}$.
- $|y(\tau) - y_1| = \mathcal{O}(\tau^3)$, also Konsistenzordnung $p = 2$ falls zusätzlich noch $\langle b, Ac \rangle = \frac{1}{6}$ und $\langle b, c \odot c \rangle = \frac{1}{3}$. \square

Bemerkung.

- Obiges ist im Prinzip konstruktives Ergebnis, d.h. wenn man solche Koeffizienten finden, dann erhält man ein Verfahren mit gewünschter Konsistenzordnung.
- Aber dabei sind i.A. nicht lineare algebraische Gleichungen zu lösen, und solche Lösungen sind i.A. nicht eindeutig.

Beispiel. (Expl. RK-Verfahren höherer Ordnung)

- i. Verfahren von Runge: $s = 2$

0	
$\frac{1}{2}$	$\frac{1}{2}$
$\frac{2}{3}$	$\frac{2}{3}$
0	1

Also gilt

$$\sum_{j=1}^2 a_{ij} = c_i, \quad \sum_{i=1}^2 b_i = 1, \quad \langle b, c \rangle = \frac{1}{2}, \quad \langle b, c \odot c \rangle = \frac{1}{4} \neq \frac{1}{3}$$

und d.h. Konsistenzordnung $p = 2$.

ii. Verfahren von Heun dritter Ordnung: $s = 3$

$$\begin{array}{c|cc} 0 & & \\ 1 & 1 \\ \hline \frac{1}{3} & \frac{1}{3} \\ 2 & 0 \\ \hline \frac{2}{3} & \frac{2}{3} \\ \hline 1 & 0 \\ \hline \frac{1}{4} & \frac{3}{4} \end{array}$$

Also gilt

$$\sum_{j=1}^2 a_{ij} = c_i, \quad \sum_{i=1}^2 b_i = 1, \quad \langle b, c \rangle = \frac{1}{2}, \quad \langle b, c \odot c \rangle = \frac{1}{3}, \quad \langle b, AC \rangle = \frac{1}{6}$$

und d.h. Konsistenzordnung $p = 3$.

iii. Klassisches RK-Verfahren: $s = p = 4$

$$\begin{array}{c|ccc} 0 & & & \\ 1 & 1 & & \\ \hline \frac{1}{2} & 2 & & \\ 1 & 0 & 1 \\ \hline \frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 0 & 0 & 1 \\ \hline 1 & 2 & 2 & 1 \\ \hline \frac{1}{6} & \frac{6}{6} & \frac{6}{6} & \frac{1}{6} \end{array}$$

mit Konsistenzordnung $p = 4$. Satz 1.27 liefert, dass das Verfahren mind. Konsistenzordnung 3 hat. Die Begründung für $p = 4$ kommt später.

In den obigen Beispielen gilt immer $p \leq s$ für explizite RK-Verfahren, und dies ist kein Zufall:

Satz 1.28. (Grenze der Konsistenzordnung für explizite RK-Verfahren)

Ein explizites RK-Verfahren mit s Stufen hat höchstens die Konsistenzordnung $p = s$.

Diese Aussage folgt leicht mit dem folgenden Satz:

Satz 1.29. (Stabilitätsfunktion von RK-Verfahren)

Sei RK-Verfahren mit s Stufen durch (A, b, c) gegeben.

Dann hat zugehörige Stabilitätsfunktion R die Gestalt

$$R(z) = 1 + z \langle b, (I - zA)^{-1} \mathbf{1}_s \rangle$$

und die Polstellen von R sind Kehrwerte der Eigenwerte von A .

Falls Verfahren explizit, so ist $R \in \mathbb{P}_s$ also Polynom.

Falls Verfahren implizit, so ist R rationale Funktion.

Beweis. Betrachte skalare Testgleichung

$$y' = \lambda y, \quad y(0) = 1.$$

RK-Verfahren liefert:

$$\forall i \in \{1, \dots, s\}: \quad v_i := y_k + \tau_k \sum_{j=1}^s a_{ij} \lambda v_j.$$

Mit $\vec{y} := y_k \cdot \mathbf{1}$ und $v := (v_1, \dots, v_s)^T$ folgt

$$v = \vec{y} + \lambda \tau_k A v \quad \Leftrightarrow \quad (I - \lambda \tau_k A)v = \vec{y}.$$

Sei $z := \lambda \tau_k$ mit τ_k derart, dass $\frac{1}{z} \notin \sigma(A)$ gilt.

Dann ist $(I - \lambda \tau_k A)$ regulär und man erhält

$$v = (I - \lambda \tau_k A)^{-1} \vec{y}.$$

Somit gilt

$$\begin{aligned} y_{k+1} &= y_k + \tau_k \sum_{i=1}^s b_i \lambda v_i \\ &= y_k + \tau_k \lambda \langle b, v \rangle \\ &= y_k + \tau_k \lambda \langle b, (I - \lambda \tau_k A)^{-1} \vec{y} \rangle \\ &= y_k + \tau_k \lambda \langle b, (I - \lambda \tau_k A)^{-1} \mathbf{1}_s \rangle y_k \\ &= (1 + \tau_k \lambda \langle b, (I - \lambda \tau_k A)^{-1} \mathbf{1}_s \rangle) y_k \\ &= R(\lambda \tau_k) y_k \\ &= R(z) y_k \end{aligned}$$

also ist R eine rationale Funktion in z , und die Pollstellen / Singularitäten tauchen vor, g.d.w. $I - zA$ nicht regulär ist, also g.d.w. $z^{-1} \in \sigma(A)$.

Falls das Verfahren explizit ist, dann hat die Matrix A eine untere Dreieksmatrixgestalt mit Nulldiagonale.

$\Rightarrow \chi_A(X) = \det(A - IX) = X^s$ und mit Cayley-Hamilton ist $A^s = 0$.

Somit erhalten wir

$$\begin{aligned} (I - zA)(I + zA + \dots + z^{s-1}A^{s-1}) &= I + zA + \dots + z^{s-1}A^{s-1} \\ &\quad - zA - \dots - z^{s-1}A^{s-1} - z^s A^s \\ &= I. \end{aligned}$$

$\Rightarrow (I - zA)^{-1} = I + zA + \dots + z^{s-1}A^{s-1}$ und das bedeutet

$$\langle b, (I - zA)^{-1} \mathbf{1}_s \rangle \in \mathbb{P}_s.$$

$\Rightarrow R(z) = 1 + z \langle b, (I - zA)^{-1} \mathbf{1}_s \rangle$ ist ein Polynom s -ten Grades. \square

Beweis. (vom Satz 1.28, $p \leq s$ für explizite RK-Verfahren)

Betrachte die Testgleichung $y' = \lambda y$, $y(0) = 1$ mit Lösung $y(t) = e^{\lambda t}$.

Mit $q(z) := \langle b, (I - zA)^{-1} \mathbf{1}_s \rangle \in \mathbb{P}_s$ ist

$$|y(\tau) - R(\lambda\tau)y_0| = |e^{\lambda\tau} - q(\tau)y_0|.$$

Da man $e^{\lambda\tau}$ durch Polynom vom Grad s höchstens bis auf Fehler $\mathcal{O}(\tau^{s+1})$ approximieren kann, gilt also $p \leq s$. \square

Der Beweis von Satz 1.28 schlägt vor, dass die „besonders guten“ expliziten RK-Verfahren im Sinne von maximaler Konsistenzordnung genau diejenigen sind, die die Exponentialfunktion approximieren. Diese Idee konkretisieren wir mit der folgenden Bemerkung:

Bemerkung. (Realisierung $p = s$ für kleines s und explizites RK)

Für einige explizite RK-Verfahren ist $R(z)$ tatsächlich führende Terme der Exponentialreihe

$$e^z = \sum_{k=0}^{\infty} \frac{1}{k!} z^k = 1 + \frac{1}{1}z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \dots$$

Dies erklärt eine Konsistenzordnung $p = s$:

- Explizites Euler, $s = 1$:

$$R(z) = 1 + z \quad \Rightarrow p = 1.$$

- Klassisches RK-Verfahren, $s = 4$:

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} \frac{1}{6} \\ \frac{2}{6} \\ \frac{2}{6} \\ \frac{1}{6} \end{pmatrix}, \quad c = \begin{pmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{pmatrix}.$$

Damit gilt

$$(I - zA)^{-1} = \sum_{k=0}^3 (zA)^k = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{z}{2} & 1 & 0 & 0 \\ \frac{z^2}{4} & \frac{z}{2} & 1 & 0 \\ \frac{z^3}{4} & \frac{z^2}{2} & 2 & 1 \end{pmatrix}, \quad (I - zA)^{-1} \mathbf{1}_s = \begin{pmatrix} 1 \\ \frac{z}{2} + 1 \\ \frac{z^2}{4} + \frac{z}{2} + 1 \\ \frac{z^3}{4} + \frac{z^2}{2} + z + 1 \end{pmatrix},$$

und somit ist

$$\begin{aligned} R(z) &= 1 + \langle b, (I - zA)^{-1} \mathbf{1}_s \rangle z \\ &= 1 + z + z^2 \left(\frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{6} \cdot 1 \right) + z^3 \left(\frac{1}{3} \cdot \frac{1}{4} + \frac{1}{6} \cdot \frac{1}{2} \right) + z^4 \frac{1}{6} \cdot \frac{1}{4} \\ &= \sum_{k=0}^4 \frac{1}{k!} z^k, \end{aligned}$$

also Konsistenzordnung $p = s = 4$.

Bemerkung. (Maximale Ordnung von expliziten RK-Verfahren)

Man könnte anhand der Beispiele vermuten, dass $p = s$ also maximale Konsistenzordnung für jedes s und geeignet gewählte Koeffizienten für explizite RK-Verfahren realisiert werden kann.

Dies ist leider nicht so!

Für maximale realisierbare Konsistenzordnung $p^*(s)$ gilt (hier ohne Beweis):

s	1	2	3	4	5	6	7	8	9	$s \geq 10$
$p^*(s)$	1	2	3	4	4	5	6	7	8	$p^*(s) \leq s - 2$

Die nächste Frage, mit der wir uns beschäftigen, ist die Wohldefiniertheit von RK-Verfahren:

Satz 1.30. (Wohldefiniertheit & Lipschitz-Stetigkeit von RK-Verfahren)

Sei $D = I \times \mathbb{R}^d$, $f: D \rightarrow \mathbb{R}^d$ Lipschitz-stetig bzgl. y mit Lipschitz-Konstante $L \geq 0$.

Sei ein RK-Verfahren (A, b, c) der Stufe s gegeben.

Dann gilt:

- i. Die zugehörige Verfahrensfunktion

$$\phi(t, \tau, y) := \sum_{i=1}^s b_i f(t + c_i \tau, v_i)$$

ist wohldefiniert für alle $\tau < (\|A\|_\infty L)^{-1}$, d.h. zu $t, t + \tau \in I$ und $y \in \mathbb{R}^d$ sind alle Stufen v_1, \dots, v_s eindeutig bestimmt.

- ii. Die Verfahrensfunktion ϕ ist Lipschitz-stetig bzgl. y mit

$$\forall y, \bar{y} \in \mathbb{R}^d: \quad \|\phi(t, \tau, y) - \phi(t, \tau, \bar{y})\| \leq \frac{L \|b\|_1}{1 - \tau L \|A\|_\infty} \|y - \bar{y}\|. \quad (1.25)$$

- iii. Falls $\tau < (2 \|A\|_\infty L)^{-1}$ und alle Gewichte b_i nicht negativ sind, dann gilt

$$\|\phi(t, \tau, y) - \phi(t, \tau, \bar{y})\| \leq 2L \|y - \bar{y}\|. \quad (1.26)$$

Vor dem Beweis bemerken wir:

- $\|b\|_1 = \sum_{i=1}^s |b_i|$ 1-Norm von Vektoren.
- $\|A\|_\infty = \max_{i \in \{1, \dots, s\}} \sum_{j=1}^s |a_{ij}|$ Zeilensummennorm von Matrizen.
- Wie in Satz 1.14 bzw. Übung, nehmen wir an, dass der Ortsgebiet für y das ganze \mathbb{R}^d also unbeschränkt ist, insbesondere kann y beliebig groß gewählt werden.

Beweis.

- i. Wohldefiniertheit der Stufen:

Sei $y \in \mathbb{R}^d$ und $t, t + \tau \in I$.

Zu $\{v_i\}_{i=1}^s \subseteq \mathbb{R}^d$ definieren wir

$$v := \begin{pmatrix} v_1 \\ \vdots \\ v_s \end{pmatrix} \in \mathbb{R}^{sd}, \quad F(v) := \begin{pmatrix} F_1(v) \\ \vdots \\ F_s(v) \end{pmatrix} \in \mathbb{R}^{sd}$$

sowie

$$F_i(v) := y + \tau \sum_{i=1}^s a_{ij} f(t + c_j \tau, v_j).$$

Also nach Definition von RK-Verfahren ist $F_i(v) = v_i$.

Somit sind die Stufen wohldefiniert, wenn F einen eindeutigen Fixpunkt $v = F(v)$ besitzt.

Dies zeigen wir wie üblich mit dem Banach'schen Fixpunktsatz.

Auf \mathbb{R}^{ds} definieren wir eine neue Norm

$$\|w\| := \max_{i \in \{1, \dots, s\}} \|w_i\| \quad \text{für } w = \begin{pmatrix} w_1 \\ \vdots \\ w_s \end{pmatrix} \in \mathbb{R}^{ds}, \quad w_1, \dots, w_s \in \mathbb{R}^d.$$

Beachte $\|\bullet\| \neq \|\bullet\|$, sondern es ist das Maximum der 2-Norm von Zeilen.

Außerdem gilt für $v, w \in \mathbb{R}^{ds}$:

$$\begin{aligned} \|F_i(v) - F_i(w)\| &\leq \tau \sum_{j=1}^s |a_{ij}| \cdot \|f(t + c_j \tau, v_j) - f(t + c_j \tau, w_j)\| \\ &\leq \tau \sum_{j=1}^s |a_{ij}| L \|v_j - w_j\| \\ &\leq \tau \left(\sum_{j=1}^s |a_{ij}| \right) L \max_{j \in \{1, \dots, s\}} \|v_j - w_j\| \end{aligned}$$

wobei die 2. Zeile aus Lipschitz-Stetigkeit von f folgt.

Somit gilt

$$\begin{aligned} \|F(v) - F(w)\| &= \max_{i \in \{1, \dots, s\}} \|F_i(v) - F_i(w)\| \\ &\leq \max_{i \in \{1, \dots, s\}} \tau \left(\sum_{j=1}^s |a_{ij}| \right) L \max_{j \in \{1, \dots, s\}} \|v_j - w_j\| \\ &= \tau \max_{i \in \{1, \dots, s\}} \left(\sum_{j=1}^s |a_{ij}| \right) L \max_{j \in \{1, \dots, s\}} \|v_j - w_j\| \\ &\leq \tau \|A\|_\infty L \|v - w\|. \end{aligned}$$

Wegen $q := \tau \|A\|_\infty L < 1$ ist F eine Kontraktion auf \mathbb{R}^d , also es existiert einen eindeutigen Fixpunkt nach Banach'scher Fixpunktsatz.

Damit sind alle Stufen wohldefiniert.

ii. Lipschitz-Stetigkeit von ϕ :

Seien $y, \bar{y} \in \mathbb{R}^d$ mit $y \neq \bar{y}$ und die zugehörigen Stufen

$$v_i = y + \tau \sum_{j=1}^s a_{ij} f(t + c_j \tau, v_j) \quad \text{sowie} \quad \bar{v}_i = \bar{y} + \tau \sum_{j=1}^s a_{ij} f(t + c_j \tau, \bar{v}_j).$$

Somit gilt es für jedes $i \in \{1, \dots, s\}$:

$$\begin{aligned} \|v_i - \bar{v}_i\| &\leq \|y - \bar{y}\| + \tau \sum_{j=1}^s |a_{ij}| \cdot \|f(t + c_j \tau, v_j) - f(t + c_j \tau, \bar{v}_j)\| \\ &\leq \|y - \bar{y}\| + \tau \sum_{j=1}^s |a_{ij}| \cdot L \|v_j - \bar{v}_j\| \\ &\leq \|y - \bar{y}\| + \tau \sum_{j=1}^s |a_{ij}| \cdot L \|v - \bar{v}\| \\ &\leq \|y - \bar{y}\| + \tau \|A\|_\infty \cdot L \|v - \bar{v}\|, \end{aligned}$$

wobei die 2. Zeile aus der Lipschitz-Stetigkeit von f im 2. Argument folgt, die 3. Zeile aus der Definition von $\|\bullet\|$, und die letzte Zeile aus der Definition von $\|\bullet\|_\infty$ von Matrizen.

Mit $q := \tau \|A\|_\infty \cdot L$ gilt dann

$$\|v - \bar{v}\| = \max_{i \in \{1, \dots, s\}} \|v_i - \bar{v}_i\| \leq \|y - \bar{y}\| + q \|v - \bar{v}\|$$

also

$$(1 - q) \|v - \bar{v}\| \leq \|y - \bar{y}\|$$

und wegen $\tau < (\|A\|_\infty L)^{-1}$ ist $(1 - q) < 1$ und somit

$$\|v - \bar{v}\| \leq \frac{1}{1 - q} \|y - \bar{y}\| = \frac{1}{1 - \tau \|A\|_\infty L} \|y - \bar{y}\|.$$

Damit folgt (1.25):

$$\begin{aligned} \|\phi(t, \tau, y) - \phi(t, \tau, \bar{y})\| &\leq \sum_{i=1}^s |b_i| \cdot \|f(t + c_i \tau, v_i) - f(t + c_i \tau, \bar{v}_i)\| \\ &\leq L \sum_{i=1}^s |b_i| \|v_i - \bar{v}_i\| \\ &\leq L \|b\|_1 \|v - \bar{v}\| \\ &\leq \frac{L \|b\|_1}{1 - \tau L \|A\|_\infty} \|y - \bar{y}\|. \end{aligned}$$

iii. Wir zeigen die vereinfachten Schranken (1.26):

- Falls alle b_i nicht negativ sind, gilt

$$\|b\|_1 = \sum_{i=1}^s |b_i| = \sum_{i=1}^s b_i = 1.$$

- Falls $\tau < (2 \| A \|_\infty L)^{-1}$ gilt, folgt

$$\frac{L \| b \|_1}{1 - \tau L \| A \|_\infty} \leq \frac{L}{1 - \frac{1}{2}} = 2L.$$

Und somit lässt sich (1.25) zu (1.26) vereinfachen. \square

Folgerung 1.31. (Konvergenz)

Seien Voraussetzungen von 1.30 für ein RK-Verfahren mit Konsistenzordnung p erfüllt, d.h. insbesondere sind alle $b_i \geq 0$ sowie $\tau < (2 \| A \|_\infty L)^{-1}$.

Dann gilt die Fehlerschranke

$$\| y_k - y(t_k) \| \leq \frac{C_p}{2L} (e^{2L\tau k} - 1) \tau^p$$

also insbesondere hat das RK-Verfahren die Konvergenzordnung p .

Beweis. Zu Erinnerung: RK-Verfahren sind explizite ESV im Sinne von Def. 1.15, also ist ϕ unabhängig von y_{k+1} ist.

ϕ ist Lipschitz-stetig mit Lipschitz-Konstante $2L$ nach 1.30.

Damit folgt die Behauptung aus Satz 1.18 mit $C_0 = 0$ und $C_2 = 2L$. \square

Beispiel. (Implizite RK-Verfahren)

Implizite RK-Verfahren sind im Sinne von der Bemerkung nach Def. 1.23, also dabei sollten Gleichungssysteme für v_i gelöst werden.

- Implizite Mittelpunktsformel: $s = 1, p = 2$

$$y_{k+1} = y_k + \tau_k f\left(t_k + \frac{1}{2}\tau_k, \frac{1}{2}(y_k + y_{k+1})\right)$$

und somit

$$\begin{aligned} v_1 := \frac{1}{2}(y_k + y_{k+1}) &= \frac{1}{2}y_k + \frac{1}{2}\left(y_k + \tau_k f\left(t_k + \frac{1}{2}\tau_k, v_1\right), v_1\right) \\ &= y_k + \frac{1}{2}\tau_k f\left(t_k + \frac{1}{2}\tau_k, v_1\right) \end{aligned}$$

also Butcher-Tableau

$$\begin{array}{c|c} 1 & 1 \\ \hline \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

daher

$$\sum_{i=1}^s b_i = 1, \quad \forall i \in \{1, \dots, s\}: \sum_{j=1}^s a_{ij} = c_j, \quad \langle b, c \rangle = \frac{1}{2}, \quad \langle b, c \odot c \rangle = \frac{1}{4} \neq \frac{1}{3}$$

und das bedeutet Konsistenzordnung $p = 2$.

- "2-Punkt-Gauß-Verfahren": $s = 2, p = 4$

$$\begin{array}{c|cc} \frac{1}{2} - \frac{1}{\sqrt{12}} & \frac{1}{4} & \frac{1}{4} - \frac{1}{\sqrt{12}} \\ \frac{1}{2} + \frac{1}{\sqrt{12}} & \frac{1}{4} + \frac{1}{\sqrt{12}} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

also

$$\sum_{i=1}^s b_i = 1, \quad \forall i \in \{1, \dots, s\}: \sum_{j=1}^s a_{ij} = c_j, \quad \langle b, c \rangle = \frac{1}{2},$$

$$\langle b, c \odot c \rangle = \frac{1}{3}, \quad \langle b, Ac \rangle = \frac{1}{6}$$

und d.h. Konsistenzordnung $p \geq 3$ nach 1.27.

$p = 4$ folgt aus später mit Satz 1.36.

Bemerkung. (Implementierung von impliziten RK-Verfahren)

- Statt nichtlineares System

$$\forall i \in \{1, \dots, s\}: v_i = y_k + \tau_k \sum_{j=1}^s a_{ij} f(t_k + c_j \tau_k, v_j)$$

in \mathbb{R}^{sd} zu lösen, werden Rundungsfehler reduziert, in dem Korrekturen

$$w_i := v_i - y_k$$

berechnet werden durch

$$\forall i \in \{1, \dots, s\}: w_i = \tau_k \sum_{j=1}^s a_{ij} f(t_k + c_j \tau_k, y_k + w_j).$$

Mit der Notation

$$w := \begin{pmatrix} w_1 \\ \vdots \\ w_s \end{pmatrix} \in \mathbb{R}^{ds}$$

wird das Problem umgeformt zur Suche nach Nullstellen von $G: \mathbb{R}^{sd} \rightarrow \mathbb{R}^{sd}$ mit

$$G(w) := w - \tau_k \begin{pmatrix} \sum_{j=1}^s a_{1j} f(t_k + c_j \tau_k, y_k + w_j) \\ \vdots \\ \sum_{j=1}^s a_{sj} f(t_k + c_j \tau_k, y_k + w_j) \end{pmatrix}.$$

- Lösung des Nullstellenproblems bei G mittels Newton-Verfahren verwendet Ableitung $DG \in \mathbb{R}^{ds \times ds}$, welche mit partieller Ableitung $D_y f$ ausgedrückt werden kann

$$DG = I - \tau_k \begin{pmatrix} a_{11} D_y f(t_k + c_1 \tau_k, y_k + w_1) & \cdots & a_{1s} D_y f(t_k + c_s \tau_k, y_k + w_s) \\ \vdots & \ddots & \vdots \\ a_{s1} D_y f(t_k + c_1 \tau_k, y_k + w_1) & \cdots & a_{ss} D_y f(t_k + c_s \tau_k, y_k + w_s) \end{pmatrix}.$$

- Approximation von DG durch vereinfachtes Newton-Verfahren, in dem $D_y f(t_k + c_j \tau_k, y_k + w_j)$ durch $J := D_y f(t_k, y_k)$ ersetzt wird für alle $j \in \{1, \dots, s\}$. Somit gilt

$$DG \approx I - \tau_k A \otimes J$$

wobei Tensorprodukt im Matrizenfall gerade das Kronecker-Produkt ist. Diese Formulierung ermöglicht, große $ds \times ds$ Matrizen zu umgehen, stattdessen Matrixmultiplikation mit DG (in iterativen LGS-Lösern) als Doppelschleife über s^2 vielen Blöcken zu realisieren.

- Man kann sogar J in mehreren Zeitschritten identisch lassen und J erst erneut berechnen, wenn man Probleme mit Konvergenz oder Genauigkeit trifft, z.B. wenn die Anzahl der Newton-Iteration zu groß wird.

In Folgerung 1.25 haben wir gesehen, dass kein RK-Verfahren mit s Stufen eine Konsistenzordnung $p > 2s$ besitzen kann.

Man stellt sich die Frage, ob es RK-Verfahren mit Konsistenzordnung $p = 2s$ existieren?

Die Antwort ist positiv und es sind die sogenannten „Gauß-Kollokationsverfahren“, welche als nächstes behandelt werden. Zunächst klären wir, was sind überhaupt „Kollokationsverfahren“:

Definition 1.32. (Kollokationsverfahren)

Seien $t_k, \tau_k, y_k \in \mathbb{R}$ sowie Stützstellen $c_1 < \dots < c_s \in [0, 1]$ gegeben.

Ein Polynomvektor $P_s \in (\mathbb{P}_s)^d$ heißt ein Kollokationspolynom auf $[t_k, t_k + \tau_k]$, g.d.w. wenn es für jedes $j \in \{1, \dots, s\}$ gilt

$$P_s(t_k) = y_k \quad \text{sowie} \quad P'_s(t_k + c_j \tau_k) = f(t_k + c_j \tau_k, P_s(t_k + c_j \tau_k)).$$

Dann definiert

$$y_{k+1} := P_s(t_{k+1})$$

ein Kollokationsverfahren.

Bemerkung.

- „Kollokationsverfahren“ bedeutet allgemeiner, dass eine DGL an speziellen Punkten („Kollokationspunkte“) ausgewertet bzw. punktweise erfüllt wird.
- Hier ist P_s eine approximative Lösung eines AWP auf $[t_k, t_k + \tau_k]$.

- Die Existenz und Eindeutigkeit eines Kollokationsverfahren folgt aus dem folgenden Satz.

Satz 1.33. (Äquivalenz zwischen Kollokation- & impliziten RK-Verfahren)

Ein Kollokationsverfahren zu Stützstellen $0 \leq c_1 < \dots < c_s \leq 1$ und Kollokationspolynom $P_s \in (\mathbb{P}_s)^d$ ist äquivalent zu einem impliziten RK-Verfahren (A, b, c) bzgl. Parametern $A := (a_{ij})_{i,j=1}^s$, $b := (b_i)_{i=1}^s$ und $c := (c_i)_{i=1}^s$, wobei $0 \leq c_1 < \dots < c_s \leq 1$ und für $i, j \in \{1, \dots, s\}$ gilt

$$a_{ij} := \int_0^{c_i} L_j(t) dt, \quad b_j := \int_0^1 L_j(t) dt$$

mit $L_j \in \mathbb{P}_{s-1}$ dem j -ten Lagrange Polynom zu Knoten c_1, \dots, c_s , d.h.

$$L_j(t) := \prod_{i=1, i \neq j}^s \frac{t - c_i}{c_j - c_i}.$$

Beweis. „Kollokationspolynom \Rightarrow impl. RK-Verfahren“

Sei $P_s \in (\mathbb{P}_s)^d$ ein Kollokationspolynom zu Stützstellen $c_1 < \dots < c_s \in [0, 1]$ und (A, b, c) so definiert wie oben.

Damit (A, b, c) ein RK-Verfahren wird, müssen wir dazu die Stufen v_1, \dots, v_s definieren, und dann weisen wir nach, dass das gegebene Kollokationsverfahren gerade diesem RK-Verfahren entspricht, also sind die Bedingungen (1.21) – (1.23) aus der Definition von RK-Verfahren erfüllt.

i. Definition von Stufen v_1, \dots, v_s :

Aus der Definition vom Kollokationspolynom und vom RK-Verfahren würden wir für $i \in \{1, \dots, s\}$ setzen:

$$v_i := P_s(t_k + c_i \tau_k)$$

Um zu sehen, dass diese Stufendefinition tatsächlich der beim RK-Verfahren passt, und um den Zusammenhang von P_s und v_i zu den a_{ij} und b_j zusehen, benötigen wir zunächst eine Hilfsbeobachtung:

ii. Hilfsbeobachtung:

$$P_s \in (\mathbb{P}_s)^d \Rightarrow P'_s \in (\mathbb{P}_{s-1})^d.$$

Die Lagrange-Polynome L_1, \dots, L_s bilden eine Basis von \mathbb{P}_{s-1} .

\Rightarrow Wir erhalten eine Darstellung von P'_s mittels L_1, \dots, L_s , also insb.

$$\begin{aligned} P'_s(t_k + t \tau_k) &= \sum_{j=1}^s L_j(t) P'_s(t_k + c_j \tau_k) \\ &= \sum_{j=1}^s L_j(t) f(t_k + c_j \tau_k, P_s(t_k + c_j \tau_k)) \end{aligned}$$

wobei die 2. Zeile aus der Definition des Kollokationspolynoms folgt.

iii. Zu Bedingung (1.22):

Nach Definition vom Kollokationspolynom gilt $P_s(t_k) = y_k$, und zusammen mit der Integraldarstellung von Funktionen erhalten wir für jedes $i \in \{1, \dots, s\}$:

$$\begin{aligned} v_i &= P_s(t_k + c_i \tau_k) \\ &= y_k + \int_{t_k}^{t_k + c_i \tau_k} P'_s(t) dt \\ &= y_k + \tau_k \int_0^{c_i} P'_s(t_k + t \tau_k) dt \\ &= y_k + \tau_k \int_0^{c_i} \sum_{j=1}^s L_j(t) f(t_k + c_j \tau_k, P_s(t_k + c_j \tau_k)) dt \\ &= y_k + \tau_k \sum_{j=1}^s \int_0^{c_i} L_j(t) dt f(t_k + c_j \tau_k, P_s(t_k + c_j \tau_k)) \\ &= y_k + \tau_k \sum_{j=1}^s a_{ij} f(t_k + c_j \tau_k, v_i), \end{aligned}$$

wobei die 3. Zeile aus Integraltransformation folgt, die 4. Zeile aus ii.

Damit sehen wir auch „woher die Formel für a_{ij} kommen sollte“.

iv. Zu Bedingung (1.21):

$$\begin{aligned} y_{k+1} &= P_s(t_k + \tau_k) \\ &= P_s(t_k) + \int_{t_k}^{t_k + \tau_k} P'_s(t) dt \\ &= y_k + \tau_k \int_0^1 P'_s(t_k + t \tau_k) dt \\ &= y_k + \tau_k \int_0^1 \sum_{j=1}^s L_j(t) f(t_k + c_j \tau_k, P_s(t_k + c_j \tau_k)) dt \\ &= y_k + \tau_k \sum_{j=1}^s \int_0^1 L_j(t) dt f(t_k + c_j \tau_k, P_s(t_k + c_j \tau_k)) \\ &= y_k + \tau_k \sum_{j=1}^s b_j f(t_k + c_j \tau_k, v_j) \end{aligned}$$

wobei die 3. Zeile aus Integraltransformation folgt, die 4. Zeile aus ii..

Damit sehen wir auch „woher die Formel für b_j kommen sollte“.

v. Zu Bedingung (1.23):

Die Lagrange-Polynome L_1, \dots, L_s haben die Eigenschaft

$$\sum_{j=1}^s L_j = 1$$

und somit gilt

$$\sum_{i=1}^s b_i = \sum_{j=1}^s \int_0^1 L_j(t) dt = \int_0^1 \sum_{j=1}^s L_j(t) dt = \int_0^1 1 dt = 1$$

sowie für $\forall i \in \{1, \dots, s\}$:

$$\sum_{j=1}^s a_{ij} = \sum_{j=1}^s \int_0^{c_i} L_j(t) dt = \int_0^{c_i} \sum_{j=1}^s L_j(t) dt = \int_0^{c_i} 1 dt = c_i.$$

„impl. RK-Verfahren \Rightarrow Kollokationspolynom“

Sei nun ein RK-Verfahren (A, b, c) gegeben mit Stufen $v_1, \dots, v_s \in \mathbb{R}^d$ wobei die Stützstellen $0 \leq c_1 < \dots < c_s \leq 1$ und A, b eine Integraldarstellung mittels Lagrange-Polynome erlauben. Wir konstruieren ein zugehöriges Kollokationspolynom P_s und zeigen, dass es die gewünschten Eigenschaften erfüllt.

a) Konstruktion von P_s :

Zu den Werten $f(t_k + c_1 \tau_k, v_1), \dots, f(t_k + c_s \tau_k, v_s) \in \mathbb{R}^d$ existiert ein eindeutiges $q \in (\mathbb{P}_{s-1})^d$, s.d. es für jedes $j \in \{1, \dots, s\}$:

$$q(t_k + c_j \tau_k) = f(t_k + c_s \tau_k, v_j)$$

gilt, denn q ist die eindeutige Lösung von d Interpolationsproblemen bzgl. den Werten $f(t_k + c_s \tau_k, v_j)$.

Sei $P_s \in (\mathbb{P}_s)^d$ die komponentenweise Stammfunktion von q mit eindeutiger Integrationskonstante s.d.

$$P_s(t_k) = y_k$$

gilt.

b) Nachweis der Eigenschaften von Kollokationspolynom bei P_s :

Die Rechnungen für $v_i = P_s(t_k + c_i \tau_k)$ und $y_{k+1} = P_s(t_k + \tau_k)$ sind genau die Rechnungen wie bei iii. & iv. aus letztem Teil, allerdings jeweils in umgekehrten Reihenfolgen.

Damit liefert die Konstruktion in a) tatsächlich ein Kollokationspolynom. \square

Folgerung 1.34. (Existenz & Eindeutigkeit von Kollokationspolynomen)

Sei L die Lipschitz-Konstante von f bzgl. 2. Arguments y und $\tau_\Delta < (\|A\|_\infty L)^{-1}$.

Dann gibt es zu beliebigem t_k, t_{k+1} mit $\tau_k := t_{k+1} - t_k \leq \tau_\Delta$, beliebigem $y \in \mathbb{R}^d$ und beliebigem $c_1 < \dots < c_2 \in [0, 1]$ genau ein Kollokationspolynom.

Beweis. Nach Satz 1.30 implizieren die Voraussetzungen die Wohldefiniertheit des RK-Verfahrens, also Existenz & Eindeutigkeit des RK-Verfahrens.

Mit Satz 1.33 folgt also Existenz & Eindeutigkeit von P_s . \square

Bemerkung. (Berechnung der Gewichte)

Zu gegebenen $c_1 < \dots < c_s \in [0, 1]$ lassen sich $(a_{ij})_{i,j=1}^s$ und $(b_i)_{i=1}^s$ mittels der Hilfsmatrizen

$$C := \begin{pmatrix} 1 & c_1 & c_1^2 & \cdots & c_1^{s-1} \\ 1 & c_2 & c_2^2 & \cdots & c_2^{s-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & c_s & c_s^2 & \cdots & c_s^{s-1} \end{pmatrix}, \quad d := \begin{pmatrix} 1 \\ 1 \\ \frac{1}{2} \\ \vdots \\ \frac{1}{s-1} \end{pmatrix} \quad \text{sowie} \quad D := \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & \frac{1}{2} & & \\ & & & \ddots & \\ & & & & \frac{1}{s-1} \end{pmatrix}$$

berechnen, also

$$A = (a_{ij})_{i,j=1}^s = CDC^{-1} \quad \text{sowie} \quad b = (b_i)_{i=1}^s = (C^T)^{-1}d.$$

Die Formeln gelten, denn mit den Koeffizienten von L_j als $\Lambda_j = (\lambda_i^{(j)})_{i=0}^{s-1}$ und dem Vektor $V(x) := (1, x, x^2, \dots, x^{s-1})$ gilt $\int_0^x L_j(t) dt = V(x)^T D \Lambda_j$ und zusammen mit den Interpolationsbedingungen gilt $C \Lambda_j = e_j$ wobei e_j der j -te Einheitsvektor ist. Daher ist $\Lambda_j = C^{-1} e_j$ und somit $\int_0^x L_j(t) dt = V(x)^T D C^{-1} e_j$, also insbesondere $a_{ij} = V(c_i)^T D C^{-1} e_j$ und $b_j = V(1)^T D C^{-1} e_j = e_j^T (C^T)^{-1} d$. Der Rest ist Umformung.

Wir erinnern uns daran, dass wir beim Satz 1.24 gezeigt haben:

- Falls ein RK-Verfahren $(A, b = \{b_i\}_{i=1}^s, c = \{c_i\}_{i=1}^s)$ Konsistenzordnung p hat, dann ist die Quadratur bzgl. Gewichten $\{b_i\}_{i=1}^s$ und Stützstellen $\{c_i\}_{i=1}^s$ exakt auf \mathbb{P}_{p-1} .

Die Umgekehrte Richtung ging i.A. nicht, aber dies können wir nun mit Kollokationsverfahren nachholen:

Satz 1.35. (Konsistenzordnung von Kollokationsverfahren)

Sei ein Kollokationsverfahren gegeben mit Kollokationspolynom $P_s \in (\mathbb{P}_s)^d$ und Stützstellen $0 \leq c_1 < \dots < c_s \leq 1$, und das zugehörige RK-Verfahren (A, b, c) wie in 1.33 definiert.

Sei die Quadratur

$$Q: \text{Abb}([0, 1], \mathbb{R}) \rightarrow \mathbb{R}, \quad g \mapsto \sum_{i=1}^s b_i g(c_i)$$

exakt auf Polynomen aus \mathbb{P}_{p-1} , d.h.

$$\forall \psi \in \mathbb{P}_{p-1}: \quad Q(\psi) = \int_0^1 \psi(t) dt.$$

Dann ist das Kollokationsverfahren bzw. das zugehörige RK-Verfahren konsistent mit Ordnung mind. p .

Beweis. Mit Satz 1.26 reicht es, autonome AWP

$$y'(t) = f(y(t)), \quad y(t_k) = y_k \quad \text{mit } y \text{ Lösung auf } [0, T]$$

zu betrachten.

Nach Definition vom Kollokationspolynom gilt $P'_s(t_k + c_j \tau_k) = f(P_s(t_k + c_j \tau_k))$ für jedes $j \in \{1, \dots, s\}$, und in anderen Stellen $t \in [0, T]$ stellt $P_s(t)$ eine Approximation von $y(t)$ dar. Daher gibt es eine Funktion $\delta: [0, T] \rightarrow \mathbb{R}^d$, s.d.

$$\forall j \in \{1, \dots, s\}: \quad \delta(t_k + c_j \tau_k) = 0 \quad \text{und} \quad P'_s(t) = f(P_s(t)) + \delta(t).$$

Daraus definieren wir ein neues parametrisches AWP auf $[t_k, t_{k+1}]$, also setze für $\theta \in [0, 1]$:

$$u'(t, \theta) := \frac{\partial}{\partial t} u(t, \theta) = f(u(t, \theta)) + \theta \delta(t), \quad u(t_k, \theta) = y_k. \quad (1.27)$$

Nach Konstruktion von u gilt

$$u(t, 0) = y(t) \quad \text{und} \quad u(t, 1) = P_s(t).$$

Somit gilt wegen Satz von Newton-Leipnitz

$$P_s(t) - y(t) = u(t, 1) - u(t, 0) = \int_0^1 \frac{\partial}{\partial \theta} u(t, \theta) d\theta. \quad (1.28)$$

Der Term $\frac{\partial}{\partial \theta} u(t, \theta)$ untersuchen wir durch Ableiten von (1.27), also

$$\frac{\partial}{\partial \theta} u'(t, \theta) = D_\theta f(u(t, \theta)) \cdot \frac{\partial}{\partial \theta} u(t, \theta) + \delta(t)$$

wobei man beobachtet, dass $\frac{\partial}{\partial \theta} u'(t, \theta) = \frac{\partial^2}{\partial \theta \partial t} u(t, \theta) = \frac{\partial}{\partial t} \left(\frac{\partial}{\partial \theta} u(t, \theta) \right)$.

Setze $w(t) := \frac{\partial}{\partial \theta} u(t, \theta)$, dann ist $w(t)$ Lösung einer inhomogenen linearen DGL

$$w'(t) = B(t, \theta) w(t) + \delta(t), \quad w(t_k) = 0$$

mit $B(t, \theta) := D_\theta f(u(t, \theta))$.

Sei $\Phi(t): [t_k, t_{k+1}] \rightarrow \mathbb{R}^{d \times d}$ die Matrix eines Fundamentalssystems der zugehörigen homogenen linearen DGL $w'(t) = B(t, \theta) w(t)$, d.h. jede Spalte von Φ löst die homogenen DGL & die Spalten von Φ sind linear unabhängig.

Aus Analysis 3 wissen wir, dass es dann gilt

$$w(t) = \Phi(t) \int_{t_k}^t \Phi^{-1}(s) \delta(s) ds = \int_{t_k}^t \Phi(t) \Phi^{-1}(s) \delta(s) ds.$$

Einsetzen in (1.28) ergibt für $t = t_{k+1}$

$$\begin{aligned} P_s(t_{k+1}) - y(t_{k+1}) &= \int_0^1 \frac{\partial}{\partial \theta} u(t_{k+1}, \theta) d\theta \\ &= \int_0^1 w(t_{k+1}) d\theta \\ &= \int_0^1 \int_{t_k}^t \Phi(t_{k+1}) \Phi^{-1}(s) \delta(s) ds d\theta \\ &= \int_{t_k}^t \int_0^1 \Phi(t_{k+1}) \Phi^{-1}(s) \delta(s) d\theta ds \end{aligned}$$

Setze $g(s) := \int_0^1 \Phi(t_{k+1}) \Phi^{-1}(s) \delta(s) d\theta$ und wir schätzen den Quadraturfehler von g mit einem Ergebnis aus NUM 1 ab:

- Bei einer auf $\mathbb{P}_{p-1}[a, b]$ exakten Quadratur gilt

$$\forall f \in C^p[a, b]: \quad \left| \int_a^b f(t) dt - Q(f) \right| \leq \frac{(b-a)^{p+1}}{p!} \|f^{(p)}\|_\infty.$$

Hier ist $[a, b] = [t_k, t_{k+1}]$ also $b - a = \tau_k$, und somit

$$P_s(t_{k+1}) - y(t_{k+1}) = \int_{t_k}^{t_{k+1}} g(s) ds = Q(g) + \mathcal{O}(\tau_k^{p+1}).$$

Beachte $P_s(t_{k+1}) = y_{k+1}$ und $Q(g) = \sum_{j=1}^s b_j g(t_k + c_j \tau) = 0$ da $\delta(t_k + c_j \tau_k) = 0$. Somit ist $P_s(t_{k+1}) - y(t_{k+1}) = Q(g) + \mathcal{O}(\tau_k^{p+1}) = \mathcal{O}(\tau_k^{p+1})$. \square

Wir fassen kurz die bisherigen Ergebnisse zusammen:

- Wir interessieren uns für ein „besonders gutes RK-Verfahren“ im Sinne von „möglichst hoher Konsistenzordnung“.
- Folgerung 1.25 besagt, dass ein RK-Verfahren mit s Stufen höchstens die Konsistenzordnung $2s$ haben kann.
- Satz 1.28 besagt, dass ein explizites RK-Verfahren mit s Stufen höchstens die Konsistenzordnung s besitzt, und d.h. wir sollen eher die impliziten RK-Verfahren untersuchen.
- Nach Satz 1.35 entspricht die Konsistenzordnung der durch Kollokationspolynome definierten impliziten RK-Verfahren dem Exaktheitsgrad der Quadratur, die durch $(b_i)_{i=1}^s$ und $(c_i)_{i=1}^s$ definiert wird, wobei die Koeffizienten b_1, \dots, b_s von den Stützstellen c_1, \dots, c_s bestimmt werden.

Deswegen sollen wir, um das „bestmögliche RK-Verfahren“ zu finden, die Stützstellen c_1, \dots, c_s geschickt wählen, so dass die daraus entstehende Quadratur möglichst hohen Exaktheitsgrad besitzt.

Dazu bietet sich ein nützliches Resultat aus NUM 1 an — Gauß-Quadratur, und somit erhalten wir

Satz 1.36. (Gauß-Kollokationsverfahren)

Sei $s \in \mathbb{N}$ und $\{c_j\}_{j=1}^s$ die eindeutigen Stützstellen der Gauß-Quadratur auf Intervall $[0, 1]$.

Dann ist das zugehörige Kollokationsverfahren („Gauß-Kollokationsverfahren“) konsistent mit Ordnung $p = 2s$.

Beweis. In NUM 1 wird die Existenz & Eindeutigkeit der Gauß-Legendre-Quadratur gezeigt und diese ist exakt auf \mathbb{P}_{2s-1} .

Mit Satz 1.35 folgt die Konsistenzordnung $p = 2s$ für das zugehörige Kollokationsverfahren. \square

Beispiel.

- $s = 1$: Zugehörige Gauß-Quadratur ist Mittelpunktsformel

$$Q(f) = f\left(\frac{1}{2}\right),$$

also

$$c_1 = \frac{1}{2}, \quad L_1(t) = 1 \in \mathbb{P}_0, \quad b_1 = \int_0^1 L_1(t) dt, \quad a_{11} = \int_0^{c_1} L_1(t) dt = \frac{1}{2},$$

d.h. Butcher-Tableau

$$\begin{array}{c|c} 1 & 1 \\ \hline \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

und früher haben wir schon die Konsistenzordnung $p = 2$ von impliziter Mittelpunktsformel nachgerechnet, hier nochmal mit Satz 1.36 bestätigt.

- $s = 2$: Dies führt auf „2-Punkt-Gauß-Verfahren“, welches wir früher schon gesehen haben, und jetzt können wir die damals behauptete Konsistenzordnung $p = 4$ begründen.

Jetzt untersuchen wir ein anderes Thema: **Stabilität von RK-Verfahren**.

Wir haben schon gesehen: Explizites Euler-Verfahren ist nicht A-stabil und nicht isometrieerhaltend.

Dies kann man verallgemeinern:

Folgerung 1.37. (Expl. RK-Verfahren weder A-stabil noch isom.erhaltend)

Kein explizites RK-Verfahren ist A-stabil oder isometrieerhaltend.

Beweis. Explizite RK-Verfahren haben nach Satz 1.29 Polynome als Stabilitätsfunktion $R(z)$, und d.h.

$$\forall z \in \mathbb{C} \setminus \{0\}: \quad \lim_{\lambda \rightarrow \infty} |R(\lambda z)| = \infty$$

also

$$\mathbb{C}^- \not\subseteq S := \{z \in \mathbb{C}: |R(z)| \leq 1\}$$

und somit nicht A-stabil.

Für $z \in i\mathbb{R}$ und $|z|$ genügend groß ist $|R(z)| \neq 1$, also Verfahren nicht isometrieerhaltend. \square

Beispiel. (Positivbeispiel: Implizite Mittelpunktsregel)

Aus Butcher-Tableau

1	1
$\frac{1}{2}$	$\frac{1}{2}$
—	—
1	

folgt $b = (1)$ und $A = \begin{pmatrix} 1 & 0 \\ \frac{1}{2} & 1 \end{pmatrix}$, also lautet die Stabilitätsfunktion nach Satz 1.29

$$\begin{aligned} R(z) &= 1 + z \langle b, (I - zA)^{-1} \mathbf{1}_s \rangle \\ &= 1 + z \left(1 - \frac{1}{2}z \right)^{-1} \\ &= \frac{2+z}{2-z}. \end{aligned}$$

Mit $\alpha, \beta \in \mathbb{R}$ und $z = \alpha + \beta i$ ist dann

$$R(z) = R(\alpha + \beta i) = \frac{2 + \alpha + \beta i}{2 - \alpha - \beta i}.$$

Man sieht leicht:

- Für $\alpha = 0$ ist $|2 + \beta i| = |2 - \beta i|$, daher $|R(z)| = 1$, also isometrieerhaltend.
- Für $\alpha \leq 0$ ist $|2 + \alpha + \beta i| \leq |2 - \alpha - \beta i|$, daher $|R(z)| \leq 1$, also A-stabil.

Bemerkung.

- 2-Punkt-Gauß-Verfahren ist A-stabil und isometrieerhaltend.
- I.A. sind alle Gauß-Kollokationsverfahren A-stabil. Auf den Beweis dafür verzichten wir hier.

Beispiel. (Numerische Beispiele)

Dies steht auf Seite 59 vom Handskript bzw. beim Vorlesungsvideo 10.2.

Bemerkung. (Adaptive Zeitschrittweiten-Steuerung)

- Erinnerung aus NUM I bzgl. zusammengesetzte Quadraturen:
Die Adaptivität kann man realisieren durch 2 Quadraturen unterschiedlicher Ordnung, dann Vergleich ihrer Ergebnisse für Teilintervalle, ggf. Verfeinerung von Intervallen und Neuberechnung der Teil-Integrale.
- Analoges ist für AWP möglich:
Zu jedem t_k kann mittels Schrittweiten-Steuerung optimales τ_k bestimmt werden.
- Ansatz mit Extrapolation:
Bei bekannter Ordnung eines Verfahrens gilt für $t \in \Delta$, $\tau := \tau_\Delta$:

Bei bekannter Ordnung eines Verfahrens gilt für $t \in \Delta$, $\tau := \tau_\Delta$:

$$y(t) - y_\tau(t) = \tau^p e_p(t) + \mathcal{O}(\tau^{p+1}) \quad (1.29)$$

mit unbekannter Fehlerfunktion $e_p(t)$.

Falls kleinere Schritte $\bar{\tau} := \frac{1}{2}\tau$ gewählt wird, dann gilt

$$y(t) - y_{\bar{\tau}}(t) = \bar{\tau}^p e_p(t) + \mathcal{O}(\bar{\tau}^{p+1}). \quad (1.30)$$

Auflösen von (1.29) und (1.30) nach $y(t)$ ergibt

$$\bar{\tau}^p e_p(t) + \mathcal{O}(\bar{\tau}^{p+1}) + y_{\bar{\tau}}(t) = y(t) = \tau^p e_p(t) + \mathcal{O}(\tau^{p+1}) + y_{\tau}(t)$$

wobei nach Definition von $\bar{\tau}$ gilt

$$\bar{\tau}^p = \frac{\tau^p}{2^{p+1}}, \quad \mathcal{O}(\bar{\tau}^{p+1}) = \mathcal{O}\left(\frac{\tau^{p+1}}{2^{p+1}}\right) = \mathcal{O}(\tau^{p+1})$$

also erhalten wir bei der obiger Gleichung

$$\tau^p e_p(t) \left(\frac{1}{2^p} - 1 \right) = y_{\tau}(t) - y_{\bar{\tau}}(t) + \mathcal{O}(\tau^{p+1}).$$

Auflösen nach $e_p(t)$ liefert:

$$e_p(t) = \frac{y_{\tau}(t) - y_{\bar{\tau}}(t)}{\tau^p \left(\frac{1}{2^p} - 1 \right)} + \mathcal{O}(\tau) =: \hat{e}_p(t) + \mathcal{O}(\tau)$$

Beachte: Die Größe $\hat{e}_p(t) = \frac{y_{\tau}(t) - y_{\bar{\tau}}(t)}{\tau^p \left(\frac{1}{2^p} - 1 \right)}$ ist eine berechenbare Schätzung für den Fehler e_p .

Einsetzen von $e_p(t) = \hat{e}_p(t) + \mathcal{O}(\tau)$ in (1.29) ergibt

$$y(t) - y_{\tau}(t) = \tau^p e_p(t) + \mathcal{O}(\tau^{p+1}) = \tau^p \hat{e}_p(t) + \mathcal{O}(\tau^{p+1}).$$

Setze noch $\hat{e}_{\tau}(t) := \tau^p \hat{e}_p(t) = y(t) - y_{\tau}(t) + \mathcal{O}(\tau^{p+1})$, dann ist

$$\hat{e}_{\tau}(t) = \frac{y_{\tau}(t) - y_{\bar{\tau}}(t)}{\left(\frac{1}{2^p} - 1 \right)} \quad (1.31)$$

ein Schätzer für den Fehler der Ordnung $p+1$.

- Kriterium für optmales τ_{opt} bei gewisser Zielgenauigkeit $\varepsilon \in \mathbb{R}_{>0}$:

Setze

$$\varepsilon = \|\hat{e}_{\tau_{\text{opt}}}(t)\| = \tau_{\text{opt}}^p \|\hat{e}_{\tau_p}(t)\| = \frac{\tau_{\text{opt}}^p}{\tau_p} \|\hat{e}_{\tau}(t)\|$$

dann erhalten wir

$$\tau_{\text{opt}} = \sqrt[p]{\frac{\varepsilon}{\|\hat{e}_{\tau}(t)\|}} \quad (1.32)$$

und falls $\|\hat{e}_{\tau}(t)\| < \varepsilon$, dann gilt $\tau_{\text{opt}} > \tau$, sonst $\tau_{\text{opt}} \leq \tau$.

- Anwendung auf einzelnen Zeitschritt:

Zu y_k und aktueller Schrittweite τ :

- berechne 1 Schritt des Verfahrens mit Schrittweite τ : $y_{\tau}(t + \tau)$
- berechne 2 Schritte des Verfahrens mit Schrittweite $\frac{\tau}{2}$: $y_{\frac{\tau}{2}}(t + \tau)$

- berechne Schätzer $\|\hat{e}_\tau(t)\|$ mit (1.31) und τ_{opt} mit (1.32)
- wähle τ_{opt} als Schrittweite für den nächsten Zeitschritt oder berechne $y_{\tau_{\text{opt}}}(t + \tau_{\text{opt}})$, d.h. aktuellen Zeitschritt noch einmal mit Schrittweite τ_{opt} .
- Durch Auflösen von (1.30) nach $e_p(t)$ und Einsetzen in (1.29) folgt

$$\frac{1}{\bar{\tau}^p}(y(t) - y_{\bar{\tau}}(t)) + \mathcal{O}(\bar{\tau}) = e_p(t)$$

also

$$\begin{aligned} y(t) - y_\tau(t) &= \tau^p \left(\frac{1}{\bar{\tau}^p}(y(t) - y_{\bar{\tau}}(t)) + \mathcal{O}(\bar{\tau}) \right) + \mathcal{O}(\tau^{p+1}) \\ &= 2^p(y(t) - y_{\bar{\tau}}(t)) + \mathcal{O}(\tau^{p+1}) \end{aligned}$$

und somit

$$y(t) = \frac{y_\tau(t) - 2^p(y_{\bar{\tau}}(t))}{1 - 2^p} + \mathcal{O}(\tau^{p+1}).$$

Also ist der Bruch eine Approximation der Ordnung $p+1$, und das ist besser als y_τ oder $y_{\frac{\tau}{2}}$, und diesen Bruch kann man „umsonst“ berechnen.

1.3. MEHRSCHRITTVERFAHREN

Motivation:

- Verwende $m \in \mathbb{N}$ Iterierten y_k, \dots, y_{k-m+1} , um $y_{k+1} \approx y(t_{k+1})$ zu approximieren.
- Ansatz: VI, Polynominterpolation für $t \mapsto f(t, y(t))$ und Integration:

$$\begin{aligned} y(t_{k+1}) &= y_k + \int_{t_k}^{t_{k+1}} f(s, y(s)) ds \\ &\approx y_k + \int_{t_k}^{t_{k+1}} p(s) ds \\ &=: y_{k+1} \end{aligned}$$

mit $p \in \mathbb{P}_{m-1}$ Interpolationspolynom zu Stützstellen $\{t_i\}_{i=k-m+1}^k$ und Zielwerten $\{f(t_i, y_i)\}_{i=k-m+1}^k =: \{f_i\}_{i=k-m+1}^k$.

Es bleibt noch die Frage, wie man das Integral konkret ausrechnen kann. Dazu stellen wir p über Lagrange-Polynome dar:

$$p(t) = \sum_{i=k-m+1}^k f_i L_i(t) \quad \text{mit} \quad L_i(t) = \prod_{l=k-m+1, l \neq i}^k \frac{t - t_l}{t_i - t_l}.$$

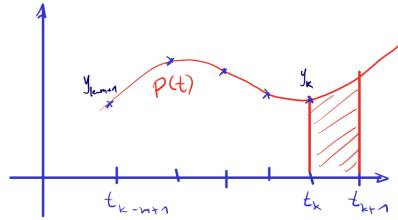
Annahme: Aquidistantes Gitter mit Schrittweite τ , d.h. $t_i = t_0 + i\tau$.

Dann erhalten wir

$$\begin{aligned}
 y_{k+1} - y_k &= \int_{t_k}^{t_{k+1}} p(t) dt \\
 &= \int_{t_k}^{t_{k+1}} \sum_{i=k-m+1}^k f_i \prod_{l=k-m+1, l \neq i}^k \frac{t - t_l}{t_i - t_l} dt \\
 &= \tau \int_0^1 \sum_{i=k-m+1}^k f_i \prod_{l=k-m+1, l \neq i}^k \frac{(t_k - s\tau) - t_l}{t_i - t_l} ds \\
 &= \tau \int_0^1 \sum_{j=1}^m f_{j+k-m} \prod_{n=1, n \neq j}^m \frac{k + s - (n + k - m)}{(j + k - m) - (n + k - m)} ds \\
 &= \tau \int_0^1 \sum_{j=1}^m f_{j+k-m} \prod_{n=1, n \neq j}^m \frac{m + s - n}{j - n} ds \\
 &= \tau \sum_{j=1}^m f_{j+k-m} \int_0^1 \prod_{n=1, n \neq j}^m \frac{m + s - n}{j - n} ds
 \end{aligned}$$

wobei die 3. Zeile aus Integraltransformation folgt, die 4. Zeile wegen $t_i = t_0 + i\tau$.

Eine Illustration der Situation:



Definition 1.38. (Adams-Basforth-Verfahren)

Zu AWP gemäß Def. 1.2 und äquidistantem Zeitgitter wählt man ein $m \in \mathbb{N}$.

Seien $y_0, \dots, y_{m-1} \in \mathbb{R}^d$ geeignet gegeben.

Setze $f_k := f(t_k, y_k)$ für $k \in \{0, \dots, m-1\}$ und dazu $f_{k+1} := f(t_{k+1}, y_{k+1})$.

Dann definiere für $k \geq m-1$:

$$y_{k+1} := y_k + \tau \sum_{j=1}^m \beta_j f_{k-m+j} \quad \text{mit} \quad \beta_j := \int_0^1 \prod_{n=1, n \neq j}^m \frac{m + s - n}{j - n} ds. \quad (1.33)$$

Bemerkung.

- Alle β_j sind von τ und k unabhängig, also kann man die β_j für gegebenes $m \in \mathbb{N}$ schon berechnen.
- Adams-Basforth-Verfahren (AB-Verfahren) ist explizit.

Beispiel.

- $m = 1$:

Unter der Annahme, dass ein leeres Produkt 1 liefert, gilt

$$\beta_1 = \int_0^1 \prod_{n=1, n \neq 1}^1 \frac{1+s-n}{j-n} ds = \int_0^1 1 ds = 1$$

und damit

$$y_{k+1} = y_k + \tau f_{k-1+1} = y_k + \tau f_k$$

also erhält man das explizite Euler-Verfahren.

- $m = 2$:

Es ist

$$\beta_1 = \int_0^1 \frac{2+s-2}{1-2} ds = -\frac{1}{2}, \quad \beta_2 = \int_0^1 \frac{2+s-1}{2-1} ds = \frac{3}{2}$$

und somit

$$y_{k+1} = y_k + \tau \left(\frac{3}{2} f_k - \frac{1}{2} f_{k-1} \right).$$

- $m = 4$:

Es handelt sich um das „eigentliche Adams-Bashforth-Verfahren“, also

$$y_{k+1} = y_k + \frac{\tau}{24} (55 f_k - 59 f_{k-1} + 37 f_{k+1} - 9 f_{k-3}).$$

Bemerkung. (Weitere auf Integration beruhende Methoden)

- Implizite Version von AB-Verfahren, also "Adams-Moulton-Verfahren": Verwende eine weitere Stützstelle, also einschließlich nächstem Zeitpunkt und erhalte eine Fixpunktgleichung:

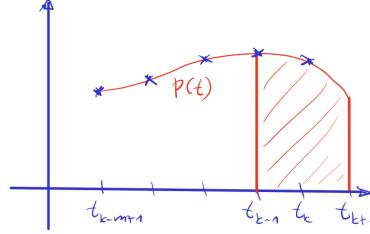
$$\begin{aligned} y_{k+1} &= y_k + \tau \sum_{j=1}^{m+1} \beta_j f(t_{k-m+j}, y_{k-m+j}) \\ &= y_k + \tau \sum_{j=1}^m \beta_j f(t_{k-m+j}, y_{k-m+j}) + \tau \beta_{m+1} f(t_{k+1}, y_{k+1}) \\ &= y_k + \tau \sum_{j=1}^m \beta_j f_{k-m+j} + \tau \beta_{m+1} f(t_{k+1}, y_{k+1}) \end{aligned}$$

mit $\{\beta_j\}_{j=1}^{m+1}$ gerechnet gemäß (1.33), also aus $m+1$ Punkten gerechnet.

Eine Möglichkeit zur Lösung dieser Fixpunktgleichung ist das sogenannte „Prädikator-Korrektor-Verfahren“, bei dem man das explizite AB-Verfahren zur Prädikation von \hat{y}_{k+1} verwendet, und dann dies als Startwert für einen Korrektor-Schritt nimmt, d.h. 1 bis 2 Iteration eines Fixpunktverfahrens.

- Erweiterung auf Zeitpunkte vor t_k , also

$$y_{k+1} - y_{k-1} = \int_{t_{k-1}}^{t_{k+1}} p(s) ds.$$



- Mit $p \in \mathbb{P}_{m-1}$ Interpolation von $t \mapsto f(t, y(t))$ über $\{t_i\}_{i=k-m+1}^k$ ergibt

$$y_{k+1} = y_{k-1} + \tau \sum_{j=1}^m \beta_j f_{k-m+j}$$

und geeignete $\{\beta_j\}_{j=1}^m$, und das ist die „Nystrom-Methoden“.

- Eine entsprechende implizite Version mit $p \in \mathbb{P}_m$ Interpolation von $t \mapsto f(t, y(t))$ über $\{t_i\}_{i=k-m+1}^{k+1}$ ist die „Milne-Simpson-Methoden“.

Bemerkung. (Auf Differentiation beruhende Verfahren)

Die Idee dabei besteht in Polynominterpolation $q \in \mathbb{P}_m$ der Werte $\{y_{k-m+j}\}_{j=1}^{m+1}$ mit unbekanntem Wert y_{k+1} , also

$$q(t) = \sum_{i=k-m+1}^{k+1} y_i L_i(t) \quad \text{mit} \quad L_i(t) = \prod_{l=k-m+1, l \neq i}^{k+1} \frac{t - t_l}{t_i - t_l}$$

und somit gilt analog wie bei AB-Verfahren

$$q(t_k + s\tau) = \sum_{i=k-m+1}^{k+1} y_i \prod_{l=k-m+1, l \neq i}^{k+1} \frac{k+s-l}{i-l} = \sum_{j=1}^{m+1} y_{k-m+j} \prod_{n=1, n \neq j}^{m+1} \frac{m+n+s}{j-n}.$$

Bei der Bedingung an dem unbekannten y_{k+1} gibt es zwei Möglichkeiten

- Bestimme y_{k+1} s.d. $q'(t_k) = f(t_k, y_k)$, d.h. q' erfüllt die DGL in t_k , also

$$q'(t_k) = \frac{1}{\tau} \frac{\partial}{\partial s} q(t_k + s\tau) \Big|_{s=0} = \frac{1}{\tau} \sum_{j=1}^{m+1} y_{k-m+j} \left(\frac{\partial}{\partial s} \prod_{n=1, n \neq j}^{m+1} \frac{m-n+s}{j-n} \right) \Big|_{s=0}. \quad (1.34)$$

Der Term $\alpha_j := \left(\frac{\partial}{\partial s} \prod_{n=1, n \neq j}^{m+1} \frac{m-n+s}{j-n} \right) \Big|_{s=0}$ lässt sich unabhängig von den y_k im Voraus berechnen.

Das Verfahren hat also die Gestalt

$$\sum_{j=1}^{m+1} \alpha_j y_{k-m+j} = \tau f(t_k, y_k)$$

und falls $\alpha_{m+1} \neq 0$, ist dies explizit nach y_{k+1} auflösbar.

- b) Bestimme y_{k+1} s.d. $q'(t_{k+1}) = f(t_{k+1}, y_{k+1})$, d.h. q' erfüllt die DGL in t_{k+1} , also ergibt sich ein Verfahren der Gestalt

$$\sum_{j=1}^{m+1} \tilde{\alpha}_j y_{k-m+j} = \tau f(t_{k+1}, y_{k+1})$$

$$\text{mit } \tilde{\alpha}_j := \left(\frac{\partial}{\partial s} \prod_{n=1, n \neq j}^{m+1} \frac{m-n+s}{j-n} \right) \Big|_{s=1}.$$

Diese Variante nennt man „Backward Difference Formulae“, oder auch „BDF-Formeln“.

Bemerkung.

- Im Prinzip kann man MSV beliebiger Ordnung definieren.
- Die Anlaufwerte (Startwerte) y_1, \dots, y_{m-1} kann man mit einem ESV bestimmen, dessen Ordnung sinnvollerweise mindestens der Ordnung des MSV entsprechen sollte.
- Ein Vorteil von MSV gegenüber ESV besteht darin, dass man bei MSV nur eine f -Auswertung pro Iteration. Dies ist insbesondere vorteilhaft, wenn die Auswertungen von f teuer sind.

1.4. RANDWERTPROBLEME FÜR ODES

Wir geben einen sehr knappen Einblick in Randwertprobleme (RWP) für lineare ODEs ohne (ausführliche) Beweise.

Definition 1.39. (Inhomogenes lineares RWP)

Sei $I = [a, b]$, $B_a, B_b \in \mathbb{R}^{d \times d}$, $A: I \rightarrow \mathbb{R}^{d \times d}$, $f \in C(I, \mathbb{R}^d)$, sowie $g \in \mathbb{R}^d$.

Gesucht ist eine Funktion $y: I \rightarrow \mathbb{R}^d$ als Lösung von

$$\forall t \in I: \quad y'(t) - A(t)y(t) = f(t), \tag{1.35}$$

$$B_a y(a) + B_b y(b) = g.$$

Die Gleichung (1.35) heißt ein Inhomogenes lineares RWP mit Randbedingung (RB) $B_a y(a) + B_b y(b) = g$.

Beachte: Für $B_a = I$, $B_b = 0$ erhalten wir wieder ein lineares AWP.

Definition 1.40. (Sturm-Liouville-Probleme)

Sei $I = [a, b]$, $p \in C^1(I, \mathbb{R}_{>0})$, $q, r, f \in C(I, \mathbb{R})$, sowie $\alpha_0, \alpha_1, \beta_0, \beta_1, g_a, g_b \in \mathbb{R}$.

Gesucht ist eine Funktion $y: I \rightarrow \mathbb{R}$ s.d. es gilt

$$\forall t \in I: -(p(t)y'(t))' + q(t)y'(t) + r(t)y(t) = f(t), \quad (1.36)$$

$$\alpha_1y'(a) + \alpha_0y(a) = g_a, \quad \beta_1y'(b) + \beta_0y(b) = g_b.$$

Bemerkung.

- Im Fall $q=0, p=1, r=0$ und $x := t$ erhalten wir die Diffusionsgleichung

$$\forall x \in [a, b]: -y''(x) = f(x),$$

die beschreibt, wie sich Stoffkonzentration y mit Quellterm f im ruhenden Medium örtlich ausbreitet, z.B. Tinte in Wasser, Wärme in Stab, etc.

- Die Funktion p heißt Diffusionsterm, q Transportterm oder Konvektionsterm, r Reaktionsterm und f Quellterm. Eine allgemeine Bezeichnung für (1.36) lautet Konvektions-Reaktions-Diffusions-Gleichung.
- Im Fall $\alpha_1 = \beta_1 = 0$ und $\alpha_0 \neq 0 \neq \beta_0$ spricht man von der Dirichlet-Randbedingung. Im Zusammenhang zur stationären Wärmeleitung (vgl. Anfang des Kapitels) gibt die Dirichlet-Randbedingung die Temperaturen an Randpunkte vor.
- Im Fall $\alpha_0 = \beta_0 = 0$ und $\alpha_1 \neq 0 \neq \beta_1$ spricht man von der Neumann-Randbedingung, insbesondere spricht man bei $y'(a) = 0 = y'(b)$ von isolierender Randbedingung.
- SL-Problem ist ein Spezialfall vom inhomogenen linearen RWP.

Dies sieht man, wenn man das SL-Problem als System erster Ordnung schreibt:

y löst (1.36), d.h.

$$-p'y' - py'' + qy' + ry = f \quad \text{mit RB} \quad \begin{cases} \alpha_1y'(a) + \alpha_0y(a) = g_a \\ \beta_1y'(b) + \beta_0y(b) = g_b \end{cases}$$

\Leftrightarrow

$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} := \begin{pmatrix} y \\ y' \end{pmatrix}$ löst

$$\begin{pmatrix} z'_1 \\ z'_2 \end{pmatrix} - \begin{pmatrix} 0 & 1 \\ \frac{r}{p} & \frac{q-p'}{p} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{f}{p} \end{pmatrix}$$

$$\text{mit RB} \quad \begin{pmatrix} \alpha_0 & \alpha_1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} z_1(a) \\ z_2(a) \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ \beta_0 & \beta_1 \end{pmatrix} \begin{pmatrix} z_1(b) \\ z_2(b) \end{pmatrix} = \begin{pmatrix} g(a) \\ g(b) \end{pmatrix},$$

also ist $B_a := \begin{pmatrix} \alpha_0 & \alpha_1 \\ 0 & 0 \end{pmatrix}$, $B_b := \begin{pmatrix} 0 & 0 \\ \beta_0 & \beta_1 \end{pmatrix}$ und $g := \begin{pmatrix} g(a) \\ g(b) \end{pmatrix}$.

Es stellt sich natürlich die Frage, ob die Lösungen zu inh. lin. RWP bzw. zu SL-Problemen existieren oder im Fall Existenz eindeutig sind.

Wir klären zunächst die Eindeutigkeit von Lösung zu SL-Problem, denn die Struktur des Beweises wird später oft auftauchen.

Satz 1.41. (Eindeutigkeit von Lösung eines SL-Problems)

Das SL-Problem gemäß Def. 1.40 mit Dirichlet-Randbedingung ($\alpha_1 = \beta_1 = 0$ und $\alpha_0 \neq 0 \neq \beta_0$) hat höchstens eine Lösung $y \in C^2((a, b), \mathbb{R}) \cap C([a, b], \mathbb{R})$, falls die folgenden drei Bedingungen erfüllt sind:

- p ist „echt-positiv“, d.h. $\exists p_0 \in \mathbb{R} \forall t \in [a, b]: p(t) \geq p_0 > 0$,
- $p, q \in C^1$
- Die Daten erfüllen noch eine Ungleichung

$$p_0 + (b-a)^2 \min_{t \in [a, b]} \left(r(t) - \frac{1}{2} q'(t) \right) > 0.$$

Beweis. Es reicht zu zeigen, dass das zugehörige homogene Problem ($f = 0$, $g_a = g_b = 0$) nur triviale Lösung hat, d.h. sei y eine Lösung von

$$-(py')' + qy' + ry = 0 \quad (1.37)$$

mit $y(a) = y(b) = 0$, und wir zeigen $y = 0$.

(Um Notation zu vereinfachen, schreiben wir bei Integranden p statt $p(t)$, sofern es im Kontext klar ist.)

- i. Multiplizieren (1.37) mit y und Integrieren über $[a, b]$ ergibt

$$-\int_a^b (py')' y dt + \int_a^b qy'y dt + \int_a^b ry^2 dt = 0.$$

Partielle Integration beim 1. & 2. Term liefert

$$\begin{aligned} -\int_a^b (py')' y dt &= \int_a^b py' y' dt - [py'y]_a^b \\ \int_a^b qy'y dt &= -\int_a^b \frac{1}{2} q'y^2 dt + \left[\frac{1}{2} qy^2 \right]_a^b \end{aligned}$$

wobei die beiden Auswertungsterme 0 sind, da $y(a) = y(b) = 0$.

Wir fassen die drei übrigen Integralterme zusammen und erhalten

$$\int_a^b p(y')^2 dt + \int_a^b \left(r - \frac{1}{2} q' \right) y^2 dt = 0.$$

ii. Abschätzen der linken Seite nach Unten liefert

$$\begin{aligned} 0 &= \int_a^b p(y')^2 dt + \int_a^b \left(r - \frac{1}{2} q' \right) y^2 dt \\ &\geq \int_a^b p(y')^2 dt + \min_{t \in [a,b]} \left(r - \frac{1}{2} q' \right) \int_a^b y^2 dt \\ &\geq p_0 \int_a^b (y')^2 dt + \min_{t \in [a,b]} \left(r - \frac{1}{2} q' \right) \int_a^b y^2 dt \end{aligned}$$

wobei die 3. Zeile die „echte Positivität“ von p genutzt hat.

Also wir erhalten

$$p_0 \int_a^b (y')^2 dt + \min_{t \in [a,b]} \left(r - \frac{1}{2} q' \right) \int_a^b y^2 dt \leq 0. \quad (1.38)$$

iii. Hilfswerkzeug: (1 dimensionale) Poincaré-Ungleichung

$$\forall u \in C^1([a,b], \mathbb{R}) \text{ mit } u(a) = 0: \quad \int_a^b u^2 dt \leq (b-a)^2 \int_a^b (u')^2 dt.$$

Beweis dazu:

Es ist

$$\begin{aligned} u(t)^2 &= \left(u(a) + \int_a^t u'(s) ds \right)^2 \\ &= \left(0 + \int_a^t 1 \cdot u'(s) ds \right)^2 \\ &\leq \left(\left(\int_a^t 1^2 ds \right)^{\frac{1}{2}} \cdot \left(\int_a^t (u')^2 ds \right)^{\frac{1}{2}} \right)^2 \\ &= \int_a^t 1 ds \cdot \int_a^t (u')^2 ds \\ &\leq \int_a^b 1 ds \cdot \int_a^b (u')^2 ds \\ &= (b-a) \cdot \int_a^b (u')^2 ds \end{aligned}$$

wobei die 3. Zeile aus Cauchy-Schwarz-Ungleichung folgt, und somit gilt

$$\int_a^b u^2 dt \leq (b-a) \max_{t \in [a,b]} u(t)^2 \leq (b-a)^2 \int_a^b (u')^2 ds.$$

iv. Anwendung von iii. auf $\int_a^b (y')^2 dt$ ergibt

$$\int_a^b (y')^2 dt \geq (b-a)^{-2} \int_a^b y^2 dt$$

und somit wird (1.38) zu

$$\left(p_0(b-a)^{-2} + \min_{t \in [a,b]} \left(r(t) - \frac{1}{2} q'(t) \right) \right) \int_a^b y^2 dt \leq 0.$$

Nach Voraussetzung ist der komplizierte Term positiv, und das Integral ist nicht negativ da Integrand nicht negativ, also muss $\int_a^b y^2 dt = 0$ sein, und das bedeutet $y \equiv 0$. \square

Wir bemerken hier, dass *Multiplikation mit Testfunktion und partielle Integration werden wesentliche Zutate für Finite Elemente* im Kapitel 3 sein.

Die Existenzaussagen für SL-Probleme betrachten wir nicht separat, sondern wir untersuchen jetzt die Existenz- und Eindeutigkeitsaussage für allgemeine inhomogenen lineare RWP:

Bemerkung. (Existenz & Eindeutigkeit für allgemeine lineare RWP)

Sei ein inhomogenes lineares RWP gemäß Def. 1.39 gegeben, d.h. gesucht ist eine Lösung $y: I \rightarrow \mathbb{R}^d$ s.d.

$$\forall t \in I: \quad y'(t) - A(t)y(t) = f(t) \quad \wedge \quad B_a y(a) + B_b y(b) = g$$

mit $I = [a, b]$, $B_a, B_b \in \mathbb{R}^{d \times d}$, $A: I \rightarrow \mathbb{R}^{d \times d}$, $f \in C(I, \mathbb{R}^d)$, sowie $g \in \mathbb{R}^d$.

1. Betrachte hierzu $d+1$ AWP für Funktionen $u_0, \dots, u_d: I \rightarrow \mathbb{R}^d$

$$\begin{aligned} u'_0(t) - A(t)u_0(t) &= f(t), & u_0(a) &= 0 \\ u'_i(t) - A(t)u_i(t) &= 0, & u_i(a) &= e_i \end{aligned} \tag{1.39}$$

mit $i \in \{1, \dots, d\}$ und $\{e_i\}_{i=1}^d$ den Einheitsvektoren.

2. Jedes AWP in (1.39) ist eindeutig lösbar nach Picard-Lindelöf und wir bilden mit den d Lösungen u_1, \dots, u_d die Fundamentalmatrix

$$U(t) := (u_1(t) \quad \cdots \quad u_d(t)) \in \mathbb{R}^{d \times d}.$$

3. Die Lösungen zu $y'(t) - A(t)y(t) = f(t)$ bilden nach Analysis 3 einen d -dimensionalen affinen Lösungsraum.

Elemente davon sind genau in der Form

$$y(t) = u_0(t) + \sum_{i=1}^d s_i u_i(t) =: u_0(t) + U(t) \cdot s \tag{1.40}$$

mit Koeffizientenvektor $s := (s_i)_{i=1}^d \in \mathbb{R}^d$, denn diese lösen die DGL

$$\begin{aligned} y'(t) &= u'_0(t) + U'(t)s \\ &= f(t) + A(t)u_0(t) + A(t)U(t)s \\ &= f(t) + A(t)(u_0(t) + U(t)s) \\ &= f(t) + A(t)y(t). \end{aligned}$$

4. Die Randbedingungen liefert Gleichung für s :

$$\begin{aligned} B_a y(a) + B_b y(b) &= g \\ \Leftrightarrow B_a u_0(a) + B_a U(a)s + B_b u_0(b) + B_b U(b)s &= g \\ \Leftrightarrow B_a \quad 0 \quad + B_a \quad I \quad s + B_b u_0(b) + B_b U(b)s &= g \\ \Leftrightarrow (B_a + B_b U(b))s &= g - B_b u_0(b). \end{aligned}$$

Falls $B_a + B_b U(b)$ invertierbar ist, dann erhalten wir

$$s = (B_a + B_b U(b))^{-1}(g - B_b u_0(b)).$$

Das obige Ergebnis fassen wir in einem Satz zusammen:

Satz 1.42. (Existenz & Eindeutigkeit von Lösung zu inh. lin. RWP)

Ein inhomogenes lineares RWP gemäß Def. 1.39 besitzt eine eindeutige Lösung, g.d.w. die Matrix

$$Q := B_a + B_b U(b)$$

invertierbar ist.

Unsere Überlegungen von vorhin sind zwar konstruktiv, aber nicht numerisch durchführbar.

Um dies nachzuholen, führen wir die Schießverfahren ein:

Definition 1.43. (Einfach-Schießverfahren)

Sei ein allgemeines inhomogenes lineares RWP gemäß Def. 1.39 gegeben.

Wähle ein Gitter Δ auf $I = [a, b]$.

Für $j \in \{0, \dots, d\}$ berechne mit ESV oder MSV approximative Lösung

$$\forall k \in \{0, \dots, K\}: \quad u_{j,k} \approx u_j(t_k)$$

zu den $d+1$ AWP in (1.39).

Definiere die diskrete Fundamentalmatrix zur Endzeit

$$U_\tau := (u_{1,K} \ \cdots \ u_{d,K}) \in \mathbb{R}^{d \times d}$$

und damit die diskrete Schießmatrix

$$Q_\tau := B_a + B_b U_\tau.$$

Falls Q_τ invertierbar ist, löse das LGS nach $s_\tau := (s_{\tau,j})_{j=1}^d \in \mathbb{R}^d$, also löse

$$Q_\tau s_\tau = g - B_b u_{0,K}$$

und erhalte eine approximative Lösung des RWP

$$\forall k \in \{0, \dots, K\}: \quad y_k := u_{0,k} + \sum_{j=1}^d s_{\tau,j} u_{j,k}.$$

Satz 1.44. (Konvergenz vom Schießverfahren)

Falls RWP wohldefiniert ist, d.h. Q invertierbar, so ist für genügend kleine $\tau \leq \tau_{\max}$ die diskrete Schießmatrix Q_τ invertierbar.

Falls ESV / MSV für $\{u_i\}_{i=0}^d$ besitzt Ordnung p nach Satz 1.18, dann konvergiert das Einfach-Schießverfahren mit Ordnung p .

Beweis. (Skizze)

Beim ESV haben wir gesehen, dass es für geeignete $C, \tilde{C} \in \mathbb{R}$ gilt

$$\|u_{j,k} - u_j(t_k)\| \leq C e^{\tilde{C}(t_k-a)} \tau^p. \quad (1.41)$$

Damit zeigt man, dass $\|U_\tau - U(t_k)\|$ und $\|Q - Q_\tau\|$ klein sind.

Dann folgt mit LGS-Stabilität, dass $\|s - s_\tau\|$ ebenfalls klein ist und mit Stetigkeit von Norm auch $\|y_k - y(t_k)\|$ mit Ordnung p konvergiert. \square

Wir bemerken hier, dass der Begriff „Schießverfahren“ auf ballistischen Anwendungen beruht.

Bemerkung. (Erweiterungen)

- Mehrfach-Schießverfahren

Fehler in (1.41) könnte durch große Intervalle I ggf. sehr groß sein.

Idee: Intervalle I wird in M grobe Intervalle zerlegt und das RWP in Form von M gekoppelten RWP formuliert.

Dies führt auf ein LGS für Koeffizienten $s_m \in \mathbb{R}^d$ auf Teilintervallen mit $m \in \{1, \dots, M\}$.

Dieses LGS ist größer als bei Einfach-Schießverfahren, aber bessere Genauigkeit wird damit erreichbar.

- Nichtlineare RWP

Suche $y: I \rightarrow \mathbb{R}^d$ s.d.

$$\forall t \in I: \quad y'(t) = f(t, y(t)) \quad \wedge \quad r(y(a), y(b)) = 0.$$

Ansatz: Formuliere ein parametrisches AWP bzgl. ein $s \in \mathbb{R}^d$, also

$$\forall t \in I: \quad y'(t, s) = f(t, y(t, s)) \quad \wedge \quad y(a, s) = s.$$

und bestimme s der Art, s.d. $F(s) := r(y(a, s), y(b, s)) = 0$ wird.

Also entsteht dadurch ein Nullstellenproblem für $F(s)$.

Dies kann man z.B. mit Fixpunkt-Schießverfahren lösen, d.h. wähle geeignetes $s^{(0)} \in \mathbb{R}^d$ und setze

$$s^{(m+1)} := s^{(m)} - CF(s^{(m)})$$

mit einer invertierbaren Matrix C .

Dabei sind die Iterationen teuer, weil Auswertung von F Lösung eines AWP erfordert, um $y(b, s)$ zu bestimmen.

Speziell mit $C := DF$ erhalten wir das Newton-Schießverfahren.

KAPITEL 2

PDEs: KLASISCHE LÖSUNGSTHEORIE & FINITE-DIFFERENZEN-VERFAHREN

Wichtige Beispiele für PDEs

Seien $\Omega \subseteq \mathbb{R}^2$, $\Omega_T \subseteq \mathbb{R} \times \mathbb{R}^+$ Gebiete und $u \in C^2(\Omega \rightarrow \mathbb{R})$.

Man erhalten

- Poisson-Gleichung

Für ein $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ und $\forall (x, y) \in \Omega$:

$$-\frac{\partial^2}{\partial x^2}u(x, y) - \frac{\partial^2}{\partial y^2}u(x, y) = f(x, y)$$

wobei $-\frac{\partial^2}{\partial x^2}u(x, y) - \frac{\partial^2}{\partial y^2}u(x, y)$ mit $\Delta(\bullet) = \nabla^2(\bullet) := \langle \nabla, \nabla(\bullet) \rangle$ dem Laplace-Operator vereinfacht wird, also $-\frac{\partial^2}{\partial x^2}u(x, y) - \frac{\partial^2}{\partial y^2}u(x, y) = -\Delta u$, und somit

$$-\Delta u = f.$$

- Instationäre Wärmeleitungsgleichung / Diffusionsgleichung

Für ein $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ und $\forall (x, t) \in \Omega_T$:

$$\frac{\partial}{\partial t}u(x, t) - \frac{\partial^2}{\partial x^2}u(x, t) = f(x, t).$$

- Wellengleichung

Für $\forall (x, t) \in \Omega_T$:

$$\frac{\partial^2}{\partial t^2}u(x, t) = c \frac{\partial^2}{\partial x^2}u(x, t).$$

Fragen in Kapitel 2 & 3

- Gegeben eine PDE, unter welchen zusätzlichen Bedingungen und in welchen Funktionenräumen kann Existenz und Eindeutigkeit der Lösung garantiert werden?
- Kann für Spezialfälle die Lösung explizit angegeben werden?
- Welche Eigenschaften haben die Lösungen? Welche Regularität? Gilt Beschränktheit? Gibt es Invarianzen?
- Durch welche numerische Verfahren können wir die Lösung approximieren?

- Konvergiert die Lösung des numerischen Verfahrens gegen die Lösung der PDE? Mit welcher Rate?
- Kann der Fehler der numerischen Lösung quantifiziert werden?
- Wie können numerische Verfahren effizient im Rechner umgesetzt werden?

2.1. GRUNDLAGEN / NOTATIONEN

Definition 2.1. (Multiindex & Partielle Ableitung)

Sei $u: \mathbb{R}^d \rightarrow \mathbb{R}$ genügend oft differenzierbar.

Wir nennen $\beta = (\beta_i)_{i=1}^d \in \mathbb{N}_0^d$ mit Länge $|\beta| := \sum_{i=1}^d \beta_i =: k$ einen Multiindex der Ordnung k .

Wir definieren die partielle Ableitung von u zum Index β als

$$\partial^\beta u := \left(\frac{\partial}{\partial x_1} \right)^{\beta_1} \cdots \left(\frac{\partial}{\partial x_d} \right)^{\beta_d} u.$$

Dazu schreiben wir $\mathbb{B}_k := \{ \beta \in \mathbb{N}_0^d : |\beta| = k \}$ als die Menge der Multiindices der Ordnung k und setze

$$D^k u := (\partial^\beta u)_{\beta \in \mathbb{B}_k}$$

als den Vektor aller partiellen Ableitungen der Ordnung k (mit beliebiger Reihenfolge).

Wir spezifizieren noch einige Funktionsräume:

Definition 2.2. (Räume stetig differenzierbarer Funktionen)

Sei $\Omega \subseteq \mathbb{R}^d$ offen und beschränkt.

Für $m \in \mathbb{N}_0$ bezeichnen wir mit $C^m(\bar{\Omega}, \mathbb{R}^n)$ den Raum der m -mal stetig differenzierbaren Funktionen von $\bar{\Omega}$ nach \mathbb{R}^n , d.h. für $i \in \{1, \dots, m\}$ lässt sich die i -te Ableitung auf $\bar{\Omega}$ stetig fortsetzen, wobei $C^0(\bar{\Omega}, \mathbb{R}^n)$ den Raum der stetigen Funktionen bedeutet.

Für $n = 1$ schreiben wir auch kurz $C^m(\bar{\Omega}) := C^m(\bar{\Omega}, \mathbb{R})$.

Auf $C^0(\bar{\Omega})$ definieren wir die Supremumsnorm

$$\forall u \in C^0(\bar{\Omega}): \quad \|u\|_\infty := \sup_{x \in \bar{\Omega}} |u(x)|$$

und damit eine Norm auf $C^m(\bar{\Omega})$ durch

$$\forall u \in C^m(\bar{\Omega}): \quad \|u\|_{C^m(\bar{\Omega})} := \sum_{\forall k \in \{0, \dots, m\} \forall \beta \in \mathbb{B}_k} \|\partial^\beta u\|_\infty.$$

Bemerkung.

- $C^m(\bar{\Omega})$ ist ein Banachraum, d.h. ein vollständiger normierter Raum.

- Für $u \in C^m(\bar{\Omega})$ ist also $\partial^\beta u \in C^{m-|\beta|}(\bar{\Omega})$ und für $l \in \{0, \dots, m\}$ gilt es für die l -te Ableitung $D^l u \in (C^{m-l}(\bar{\Omega}))^{|\mathbb{B}_l|}$.
- Für $u \in C^1(\bar{\Omega})$ vereinbaren wir für den Gradient / die Jacobi-Matrix

$$D^1 u := \nabla u = \begin{pmatrix} \frac{\partial}{\partial x_1} u & \cdots & \frac{\partial}{\partial x_d} u \end{pmatrix}^T = (Du)^T.$$

- Für $m = \infty$ setzen wir

$$C^\infty(\bar{\Omega}) := \bigcap_{k \in \mathbb{N}} C^k(\bar{\Omega}).$$

- Wir werden später auch Räume $C^m(\Omega)$ für offenes und potentiell unbeschränktes $\Omega \subseteq \mathbb{R}^d$ zulassen.

Dabei ist eine Norm schwer definierbar, aber man kann den Raum $C^m(\Omega)$ mit einer Frechét-Metrik versehen, bzgl. deren $C^m(\Omega)$ vollständig wird (siehe Alt, Abschnitt 1.6).

In $C^m(\bar{\Omega})$ gibt es viele Elemente, die zwar nicht in $C^{m+1}(\bar{\Omega})$ liegen, aber trotzdem schöne Eigenschaften haben, wie z.B. differenzierbar an sehr vielen Stellen. Daher ist es sinnvoll, neue Funktionsräume zwischen $C^m(\bar{\Omega})$ und $C^{m+1}(\bar{\Omega})$ zu definieren:

Definition 2.3. (Hölderräume)

Für $\alpha \in [0, 1]$ und $u \in C^0(\bar{\Omega})$ mit $\Omega \subseteq \mathbb{R}^d$ offen & beschränkt definieren wir die Hölderkonstante

$$\text{hoel}_\alpha(u, \bar{\Omega}) := \sup_{x, y \in \bar{\Omega}, x \neq y} \frac{|u(x) - u(y)|}{\|x - y\|^\alpha}$$

und damit die Hölderräume durch

$$C^{m, \alpha}(\bar{\Omega}) := \{u \in C^m(\bar{\Omega}) \mid \forall \beta \in \mathbb{B}_m: \text{hoel}_\alpha(\partial^\beta u, \bar{\Omega}) < \infty\}.$$

Hölderräume $C^{m, \alpha}(\bar{\Omega})$ versehen mit der Norm

$$\|u\|_{C^{m, \alpha}(\bar{\Omega})} := \|u\|_{C^m(\bar{\Omega})} + \sum_{\beta \in \mathbb{B}_m} \text{hoel}_\alpha(\partial^\beta u, \bar{\Omega})$$

sind Banachräume.

Bemerkung.

- Der Nachweis für die Vollständigkeit der Hölderräume findet man bei Alt, Lemma 1.8.
- Funktionen $u \in C^{0,1}(\bar{\Omega})$ sind also Lipschitz-stetig mit Lipschitz-Konstante $L := \text{hoel}_1(u, \bar{\Omega})$.

Satz 2.4. (Schachtelung der Hölderräume)

Für $m \in \mathbb{N}_0$ und $0 \leq \hat{\alpha} \leq \alpha \leq 1$ gilt

$$C^m(\bar{\Omega}) \supseteq C^{m,\hat{\alpha}}(\bar{\Omega}) \supseteq C^{m,\alpha}(\bar{\Omega}) \supseteq C^{m+1}(\bar{\Omega}).$$

Beweis. Die erste Inklusion $C^m(\bar{\Omega}) \supseteq C^{m,\hat{\alpha}}(\bar{\Omega})$ ist klar nach Definition.

Verbleibendes zeigen wir für $m = 0$ (für $m > 0$ analog mit Ableitungen).

- Zweite Inklusion:

Für $u \in C^{0,\alpha}(\bar{\Omega})$ setze

$$K := \text{hoel}_\alpha(u, \bar{\Omega}) = \sup_{x,y \in \bar{\Omega}, x \neq y} \frac{|u(x) - u(y)|}{\|x - y\|^\alpha} < \infty$$

sowie

$$\text{diam}(\bar{\Omega}) := \sup_{x,x' \in \bar{\Omega}} \|x - x'\| < \infty$$

wobei $\text{diam}(\bar{\Omega}) < \infty$ wegen Beschränktheit von $\bar{\Omega}$.

Dazu setze

$$M := \sup_{s \in (0, \text{diam}(\bar{\Omega}))} \frac{s^\alpha}{s^{\hat{\alpha}}} = \sup_{s \in (0, \text{diam}(\bar{\Omega}))} s^{\alpha - \hat{\alpha}}.$$

Mit $\hat{\alpha} \leq \alpha$ und $\text{diam}(\bar{\Omega}) < \infty$ ist M beschränkt.

Somit gilt für jedes $x, y \in \bar{\Omega}$ mit $x \neq y$:

$$\frac{|u(x) - u(y)|}{\|x - y\|^{\hat{\alpha}}} = \frac{|u(x) - u(y)|}{\|x - y\|^\alpha} \cdot \frac{\|x - y\|^\alpha}{\|x - y\|^{\hat{\alpha}}} \leq KM$$

und daher

$$\text{hoel}_{\hat{\alpha}}(u, \bar{\Omega}) = \sup_{x,y \in \bar{\Omega}, x \neq y} \frac{|u(x) - u(y)|}{\|x - y\|^{\hat{\alpha}}} \leq KM < \infty$$

also $u \in C^{0,\hat{\alpha}}(\bar{\Omega})$.

- Dritte Inklusion:

Analog wie oben gilt für jedes $x, y \in \bar{\Omega}$ mit $x \neq y$

$$\frac{|u(x) - u(y)|}{\|x - y\|^\alpha} = \frac{|u(x) - u(y)|}{\|x - y\|} \cdot \frac{\|x - y\|}{\|x - y\|^\alpha} = \frac{|u(x) - u(y)|}{\|x - y\|} \|x - y\|^{1-\alpha}.$$

Da $\alpha \leq 1$ und $\bar{\Omega}$ beschränkt, ist $\|x - y\|^{1-\alpha}$ beschränkt durch ein $\zeta \in \mathbb{R}^+$.

Daher gilt

$$\text{hoel}_\alpha(u, \bar{\Omega}) = \sup_{x,y \in \bar{\Omega}, x \neq y} \frac{|u(x) - u(y)|}{\|x - y\|^\alpha} \leq \sup_{x,y \in \bar{\Omega}, x \neq y} \frac{|u(x) - u(y)|}{\|x - y\|} \zeta \leq \|u'\|_\infty \zeta$$

wobei die letzte Ungleichung aus $u \in C^1(\bar{\Omega})$ und dem Mittelwertsatz der Differenziation folgt.

Also ist $\text{hoel}_\alpha(u, \bar{\Omega}) < \infty$ und somit $u \in C^{0,\alpha}(\bar{\Omega})$. □

Beispiel.

Für $\Omega = (0, 1)$ betrachten wir $u(x) = \sqrt{x}$.

- Es ist $u \in C^0(\bar{\Omega})$ aber $u \notin C^1(\bar{\Omega})$, denn u' ist nicht stetig im Punkt $x=0$ fortsetzbar.
- Für $\alpha \leq \frac{1}{2}$ ist $u \in C^{0,\alpha}(\bar{\Omega})$, denn für $x, y \in (0, 1)$ mit o.B.d.A. $y < x$ gilt

$$\begin{aligned}\frac{|u(x) - u(y)|}{\|x - y\|^\alpha} &\leqslant \frac{\sqrt{x} - \sqrt{y}}{\sqrt{x - y}} \\ &\leqslant \frac{(\sqrt{x} - \sqrt{y})(\sqrt{x} + \sqrt{y})}{\sqrt{x - y}(\sqrt{x} + \sqrt{y})} \\ &= \frac{x - y}{\sqrt{x - y}(\sqrt{x} + \sqrt{y})} \\ &= \frac{\sqrt{x - y}}{\sqrt{x} + \sqrt{y}} \\ &\leqslant \frac{\sqrt{x}}{\sqrt{x}} = 1 < \infty.\end{aligned}$$

- Für $\alpha > \frac{1}{2}$ ist $u \notin C^{0,\alpha}(\bar{\Omega})$, also auch nicht Lipschitz-stetig, denn

$$\text{hoel}_\alpha(u, \bar{\Omega}) = \sup_{x, y \in \bar{\Omega}, x \neq y} \frac{|\sqrt{x} - \sqrt{y}|}{|x - y|^\alpha} \geqslant \sup_{x \in \bar{\Omega}} \frac{\sqrt{x}}{x^\alpha} = \sup_{x \in \bar{\Omega}} x^{\frac{1}{2} - \alpha} = \infty.$$

Wir definieren hier noch die für später nützlichen Räume:

Definition 2.5. (L^p -Räume, Teil 1)

Sei $\Omega \subseteq \mathbb{R}^d$ offen, beschränkt und versehen mit der gewöhnlichen Borel- σ -Algebra und dem d -dimensionalen Lebesgue-Maß λ^d .

Für $p \in [1, \infty)$ definieren wir

$$\tilde{L}^p(\Omega) := \left\{ u: \Omega \rightarrow \mathbb{R} \mid u \text{ ist } \lambda\text{-messbar} \wedge \int_{\Omega} |u|^p d\lambda^d < \infty \right\}$$

und darauf eine Seminorm $\|\bullet\|_p$ durch

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

Für $p = \infty$ definieren wir

$$\tilde{L}^\infty(\Omega) := \left\{ u: \Omega \rightarrow \mathbb{R} \mid u \text{ ist } \lambda\text{-messbar} \wedge \text{ess sup}_{x \in \Omega} |u(x)| < \infty \right\}$$

und darauf eine Norm $\|\bullet\|_\infty$ durch

$$\|u\|_\infty = \|u\|_{L^\infty} := \text{ess sup}_{x \in \Omega} |u(x)|.$$

Bemerkung.

- Für u messbar ist $\text{ess sup}_{x \in \Omega} |u(x)| := \sup_{N \subseteq \Omega, \lambda(N)=0} \sup_{x \in \Omega \setminus N} |u(x)|$.
- Für $p \in [1, \infty)$ ist $\|\bullet\|_p$ keine Norm auf $\tilde{L}^p(\Omega)$, denn positive Definitheit ist verletzt: Für eine nicht leere Nullmenge $N \subseteq \Omega$ ist ihre charakteristische Funktion $\chi_N \neq 0$ aber $\|\chi_N\|_p = 0$.
- Für $u \in C^0(\bar{\Omega})$ ist $u \in \tilde{L}^\infty(\Omega)$ und beide Definitionen von $\|\bullet\|_\infty$ stimmen dann überein.

Definition. (L^p -Räume, Teil 2)

Sei \sim eine Äquivalenzrelation auf $\tilde{L}^p(\Omega)$ via

$$u \sim v \iff \exists N \subseteq \Omega \text{ mit } \lambda(N) = 0 \forall x \in \Omega \setminus N: u(x) = v(x).$$

Dann definiert die Menge der Äquivalenzrelation

$$L^p(\Omega) := \tilde{L}^p(\Omega) / \sim$$

den **L^p -Raum auf Ω** , wobei $\|\bullet\|_p$ kostant auf jeder Äquivalenzklasse ist, also kann $\|\bullet\|_p$ sinnvoll auf $L^p(\Omega)$ erweitert werden.

Bemerkung.

- $L^p(\Omega)$ ist normierter Raum mit Norm $\|\bullet\|_p$ und sogar vollständig, also ein Banachraum (Lemma 1.11, Satz 1.14 in Alt).
- Elemente von $L^p(\Omega)$ sind also Äquivalenzklassen von Funktionen, die sich nur auf Nullmengen unterscheiden. Konsequente Unterscheidung zwischen Funktionen / Äquivalenzklassen wäre sehr mühsam / umständlich: „ $u \in L^p(\Omega)$ “ soll eigentlich „ $u \in U \in L^p(\Omega)$ für geeignete Äquivalenzklassen U mit u als Repräsentant“ heißen.

- Praktische Konvention

Wir nennen $L^p(\Omega)$ trotzdem einen Funktionenraum.

Beim Arbeiten mit solchen „ L^p -Funktionen“ muss jedoch immer bedacht / hinterfragt werden, ob Operationen sinnvoll auf $L^p(\Omega)$ definiert sind, d.h. ob es unabhängig von der Wahl des Repräsentanten ist.

Beispiele:

- Der Integraloperator

$$T: L^1(\Omega) \rightarrow \mathbb{R}, u \mapsto \int_{\Omega} u \, d\lambda^d$$

ist ein wohldefiniertes stetiges lineares Funktional.

- Punktauswertung bzgl. eines festen $\bar{x} \in \Omega$, also

$$S: L^p(\Omega) \rightarrow \mathbb{R}, u \mapsto u(\bar{x})$$

ist nicht sinnvoll definiert.

- Auf $L^2(\Omega)$ ist ein Skalarprodukt (\bullet, \bullet) definiert durch

$$(u, v) = \langle u, v \rangle_{L^2} := \int_{\Omega} uv \, d\lambda^d$$

und dies induziert auch die L^2 -Norm durch $\|u\| = \sqrt{\langle u, u \rangle}$. Somit besitzt $L^2(\Omega)$ sogar eine Hilbertraum-Struktur.

- Zu Banachraum V ist der Dualraum V' definiert durch

$$V' := \{\varphi: V \rightarrow \mathbb{R} \mid \varphi \text{ ist linear und stetig}\}$$

und mit der induzierten Norm

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

ist V' wieder ein Banachraum.

- Für $1 < p, q < \infty$ mit $\frac{1}{p} + \frac{1}{q} = 1$ ist $L^q(\Omega)$ isomorph zu $(L^p(\Omega))'$.
- $L^2(\Omega)$ ist also wegen $\frac{1}{2} + \frac{1}{2} = 1$ isomorph zu seinem eigenen Dualraum.

Satz 2.6. (Young'sche Ungleichung)

Seien $a, b \in \mathbb{R}_+$ und $p, q \in (1, \infty)$ mit $\frac{1}{p} + \frac{1}{q} = 1$.

Dann gilt

$$ab \leq \frac{1}{p}a^p + \frac{1}{q}b^q.$$

Beweis. Für $a = b = 0$ ist die Behauptung klar, daher seien $a \neq 0 \neq b$.

Die Konkavität des Logarithmus liefert

$$\ln(ab) = \ln(a) + \ln(b) = \frac{1}{p} \ln(a^p) + \frac{1}{q} \ln(b^q) \leq \ln\left(\frac{1}{p}a^p + \frac{1}{q}b^q\right).$$

Anwendung von \exp auf beiden Seiten liefert die Behauptung. □

Folgerung 2.7. (Young'sche Ungleichung für $p = q = 2$)

Für $a, b \in \mathbb{R}_+$ und beliebiges $\varepsilon \in \mathbb{R}_+$ gilt

$$ab \leq \frac{\varepsilon}{2}a^2 + \frac{1}{2\varepsilon}b^2.$$

Bemerkung.

- Beweis von 2.7 durch $ab = (\sqrt{\varepsilon}a)\left(\frac{1}{\sqrt{\varepsilon}}b\right)$ und dann Anwendung von 2.6.

- 2.7 wird häufig verwendet um „Produkte zu Spalten“ und einen Term beliebig klein zu machen durch geeignetes ε .

Satz 2.8. (Hölder-Ungleichung)

Seien $p, q \in [1, \infty]$ mit $\frac{1}{p} + \frac{1}{q} = 1$, $u \in L^p(\Omega)$ und $v \in L^q(\Omega)$ beliebig.
Dann ist das Produkt $uv \in L^1(\Omega)$ und es gilt

$$\|uv\|_1 \leq \|u\|_p \|v\|_q.$$

Bemerkung.

- Den Beweis von 2.8 überlassen wir als Übung.
- Für $p = q = 2$ folgt die Cauchy-Schwarz-Ungleichung

$$|(u, v)| = \left| \int_{\Omega} uv \, d\lambda^d \right| \leq \int_{\Omega} |uv| \, d\lambda^d = \|uv\|_1 \leq \|u\|_2 \|v\|_2 = \sqrt{(u, u)} \sqrt{(v, v)}.$$

Definition 2.9. (Lokal integrierbare Funktionen)

Wir definieren den Raum der lokal integrierbaren Funktionen

$$L^1_{\text{loc}}(\Omega) := \left\{ u: \Omega \rightarrow \mathbb{R} \mid u \text{ Lebesgue-messbar} \wedge \forall K \subseteq \Omega \text{ kompakt}: \int_K |u| \, d\lambda^d < \infty \right\}.$$

Beispiel.

- $u \in L^1(\Omega) \Rightarrow u \in L^1_{\text{loc}}(\Omega)$.
- $u \in L^1_{\text{loc}}(\Omega) \not\Rightarrow u \in L^1(\Omega)$, z.B. $u \equiv 1$ liegt in $L^1_{\text{loc}}(\Omega)$ aber nicht in $L^1(\Omega)$.

Definition 2.10. (Funktionen mit kompaktem Träger)

Für $\Omega \subseteq \mathbb{R}^d$ offen (aber möglicherweise unbeschränkt) und $m \in \mathbb{N}_0 \cup \{\infty\}$ definieren wir

$$C_0^m(\Omega) := \{u \in C^m(\Omega) \mid \text{supp}(u) \subseteq \Omega \text{ kompakt}\}$$

wobei $\text{supp}(u) := \overline{\{x \in \Omega \mid u(x) \neq 0\}}$ den Träger von u bedeutet.

Satz 2.11. (Fundamentalsatz der Variationsrechnung)

Für $\Omega \subseteq \mathbb{R}^d$ offen und $u \in L^1_{\text{loc}}(\Omega)$ gilt

$$\forall v \in C_0^\infty(\Omega): \int_{\Omega} uv \, d\lambda^d = 0 \Leftrightarrow u = 0 \text{ fast überall.}$$

Wir verweisen auf 2.11 in Alt für den Beweis von 2.11.

Definition 2.12. (Ableitungsoperatoren)

Sei $\Omega \subseteq \mathbb{R}^d$. Für $i \in \{1, \dots, d\}$ schreiben wir $\partial_{x_i} := \frac{\partial}{\partial x_i}$.

Für $u \in C^1(\Omega)$ definieren wir den Gradienten von u durch

$$\operatorname{grad} u = \nabla u := \begin{pmatrix} \partial_{x_1} u \\ \vdots \\ \partial_{x_d} u \end{pmatrix}.$$

Für Vektorfeld $v = (v_i)_{i=1}^d \in C^1(\Omega, \mathbb{R}^d)$ definieren wir die Divergenz

$$\operatorname{div} v := \nabla \bullet v = \sum_{i=1}^d \partial_{x_i} v_i$$

und für $d=3$ die Rotation

$$\operatorname{rot} v := \nabla \times v = \begin{pmatrix} \partial_{x_2} v_3 - \partial_{x_3} v_2 \\ \partial_{x_3} v_1 - \partial_{x_1} v_3 \\ \partial_{x_1} v_2 - \partial_{x_2} v_1 \end{pmatrix}.$$

Für $u \in C^2(\Omega)$ definieren wir den Laplace-Operator durch

$$\Delta u := \nabla \bullet (\nabla u) = \operatorname{div} \operatorname{grad} u = \sum_{i=1}^d \partial_{x_i}^2 u = \operatorname{Spur}(Hu)$$

wobei Hu die Hesse-Matrix von u bezeichnet.

Definition 2.13. (Lipschitz-Gebiet)

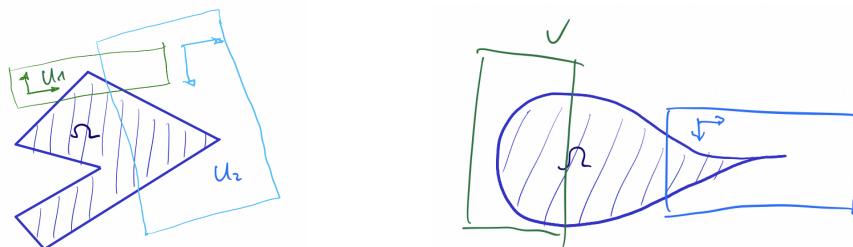
Sei $\Omega \subseteq \mathbb{R}^d$ offen und beschränkt.

Ω heißt Lipschitz-Gebiet g.d.w. es endlich viele offene Menge $U_1, \dots, U_n \subseteq \mathbb{R}^d$ existieren, s.d. die folgenden zwei Bedingungen erfüllt werden:

- Der Rand von Ω wird durch U_1, \dots, U_d überdeckt, also $\partial\Omega \subseteq \bigcup_{i=1}^n U_i$, und
- $\forall i \in \{1, \dots, n\}: \partial\Omega \cap U_i$ lässt sich in geeigneter Richtung als Graph einer Lipschitz-stetigen Funktion schreiben und Ω liegt auf einer Seite des Graphen.

Beispiel.

Gegeben z.B. folgende zwei Gebiete.



Das linke Gebiet ist ein Lipschitz-Gebiet.

Das rechte Gebiet ist kein Lipschitz-Gebiet, denn dabei lässt sich die „Spitze“ nicht als Lipschitz-stetige Funktion darstellen.

Satz 2.14. (Satz von Gauß für Lipschitz-Gebiet)

Sei $\Omega \subseteq \mathbb{R}^d$ Lipschitz-Gebiet und $v \in C^0(\bar{\Omega}, \mathbb{R}^d) \cap C^1(\Omega, \mathbb{R}^d)$ ein Vektorfeld mit $\operatorname{div} v \in L^1(\Omega)$, d.h. auf $\partial\Omega$ ist v zwar nicht differenzierbar, aber v bleibt beschränkt.

Dann gilt

$$\int_{\Omega} \operatorname{div} v \, d\lambda^d = \int_{\partial\Omega} v \bullet n \, d\lambda^{d-1}$$

wobei $n: \mathbb{R}^d \rightarrow \mathbb{R}^d$ äußere Einheitsnormale an den Rand von Ω bezeichnet.

Wir verweisen auf A.5.9. in Alt für den Beweis von 2.14, und nutzen dies direkt um eine vektorielle Form von partieller Integration zu formulieren:

Folgerung 2.15. (Partielle Integration)

Für $u \in C^1(\bar{\Omega})$ und $v \in C^1(\bar{\Omega}, \mathbb{R}^d)$ gilt

$$\int_{\Omega} \nabla u \bullet v \, d\lambda^d = - \int_{\Omega} u \operatorname{div} v \, d\lambda^d + \int_{\partial\Omega} uv \bullet n \, d\lambda^{d-1}.$$

Beweis. Satz von Gauß 2.14 & Produktregel liefern:

$$\int_{\partial\Omega} uv \bullet n \, d\lambda^{d-1} = \int_{\Omega} \operatorname{div}(uv) \, d\lambda^d = \int_{\Omega} \nabla u \bullet v \, d\lambda^d + \int_{\Omega} u \operatorname{div} v \, d\lambda^d \quad \square$$

Nachdem wir die funktionaltheoretischen Grundlagen gesammelt haben, formulieren wir jetzt die Gegenstände, für die wir uns eigentlich interessieren:

Definition 2.16. (Skalare PDE)

Sei $k \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$ und $F: \mathbb{R}^{|\mathbb{B}_k|} \times \mathbb{R}^{|\mathbb{B}_{k-1}|} \times \dots \times \mathbb{R}^d \times \mathbb{R} \times \Omega \rightarrow \mathbb{R}$.

Wir nennen

$$\forall x \in \Omega: \quad F(D^k u(x), D^{k-1} u(x), \dots, D^1 u(x), u(x), x) = 0 \quad (2.1)$$

eine skalare PDE der Ordnung k für die unbekannte Lösung $u: \Omega \rightarrow \mathbb{R}$.

Es gibt auch Systeme von PDEs für vektorielle Funktionen, aber wir betrachten sie hier nicht.

Definition 2.17. (Klassische Lösung)

Sei eine skalare PDE gemäß (2.1) gegeben.

Falls eine Funktion $u \in C^k(\Omega)$ die Bedingung (2.1) erfüllt, dann nennen wir u eine klassische Lösung der PDE.

Im Kapitel 2 beschäftigen wir uns mit klassischer Lösung und im Kapitel 3 werden wir sehen, warum dieser Begriff in vielen Fällen zu eingeschränkt ist.

Als nächstes führen wir noch einige wichtige Arten von PDEs ein, und diese Begriffe bezeichnet man oft als **Klassifikation linearer PDEs zweiter Ordnung**.

Definition 2.18. (Linearer Differenzialoperator 2. Ordnung)

Sei $\Omega \subset \mathbb{R}^d$ offen, $A = (a_{ij})_{i,j=1}^d \in C^0(\Omega, \mathbb{R}^{d \times d})$, $b = (b_i)_{i=1}^d \in C^0(\Omega, \mathbb{R}^d)$ & $c \in C^0(\Omega)$.

Dann nennen wir $\mathcal{L}: C^2(\Omega) \rightarrow C^0(\Omega)$ mit

$$(\mathcal{L}u)(x) = - \sum_{i,j=1}^d a_{ij}(x) \partial_{x_i} \partial_{x_j} u(x) + \sum_{i=1}^d b_i(x) \partial_{x_i} u(x) + c(x)u(x)$$

einen linearen Differenzialoperator zweiter Ordnung.

Bemerkung.

- Mit \mathcal{L} werden u sowie die möglichen 1. & 2. partiellen Ableitungen von u mit skalarwertigen Funktionen multipliziert und dann addiert.
- Wir nennen $-\sum_{i,j=1}^d a_{ij}(x) \partial_{x_i} \partial_{x_j} u(x)$ den Hauptteil von \mathcal{L} .
- OBdA kann A symmetrisch gewählt werden, denn $\partial_{x_i} \partial_{x_j} = \partial_{x_j} \partial_{x_i}$, also falls A nicht symmetrisch ist, ergibt $\tilde{A} := \frac{1}{2}(A + A^T)$ identisches \mathcal{L} .
- Da A oBdA symmetrisch ist, kann man annehmen, dass für jedes $x \in \Omega$, $A(x)$ nur reelle Eigenwerte hat.
- Zu $f \in C^0(\Omega)$ ergibt sich entsprechende PDE

$$\mathcal{L}u = f. \quad (2.2)$$

Definition 2.19. (Klassifikation)

Sei \mathcal{L} ein linearer Differentialoperator zweiter Ordnung bzgl. (A, b, c) wobei A symmetrisch ist und $x \in \Omega$.

Der Operator \mathcal{L} heißt

- elliptisch in x , falls alle Eigenwerte von $A(x)$ gleichzeitig positiv oder gleichzeitig negativ sind,
- parabolisch in x , falls $d - 1$ Eigenwerte von $A(x)$ gleichzeitig positiv oder gleichzeitig negativ sind und einen Eigenwert 0 ist, aber die Matrix $(A(x) b(x)) \in \mathbb{R}^{d+1 \times d}$ vollen Rang besitzt, also $\text{Rang}(A(x) b(x)) = d$,

- hyperbolisch in x , falls $d - 1$ Eigenwerte von $A(x)$ positiv (oder alle negativ) sind und einen Eigenwert negatives (oder positives) Vorzeichen besitzt.

Der Operator \mathcal{L} heißt elliptisch / parabolisch / hyperbolisch in Ω , falls er elliptisch / parabolisch / hyperbolisch in allen Punkten von Ω ist.

Die PDE gemäß (2.2) ist elliptisch / parabolisch / hyperbolisch, falls der zugehörige Operator \mathcal{L} dies in Ω ist.

Bemerkung.

- Wir werden uns auf elliptische Operatoren konzentrieren und o.B.d.A. nehmen wir bei einem elliptischen Operator an, dass die Vorzeichen der Eigenwerte von $A(x)$ für jedes $x \in \Omega$ alle positiv sind (Wenn alle Eigenwerte negativ sind, multiplizieren wir den Operator mit -1).
- Die Klassifikation ist nicht vollständig bzw. nicht erschöpfend, aber ausreichend für die meisten praktischen Zwecke.
- In der Praxis ist auch Änderung des Typs in Teilgebieten von Ω möglich, und solche PDEs nennt man „vom gemischten Typ“.
- Begriffe sind motiviert aus Kegelschnitten / Quadriken

$$z^T A(x) z = 1,$$

welche unter den geannten Bedingungen ein Ellipsoid, Paraboloid bzw. Hyperboloid beschreiben.

- In Def. 2.18 wird Zeitvariable t , falls vorhanden, als eine der Variablen x_i interpretiert, d.h. auch Zeitableitung erster oder zweiter Ordnung erlaubt.

Beispiel.

- i. Laplace / Poisson-Gleichung ($-\Delta u = f$):

Dabei ist

$$\mathcal{L}u = -\Delta u$$

und d.h.

$$A = I, \quad b = 0, \quad c = 0.$$

A hat nur positive Eigenwerte und somit ist \mathcal{L} elliptisch.

- ii. Instationäre Wärmeleitung ($\partial_t u - \partial_x^2 u = f$):

Dabei ist

$$\mathcal{L}u = \partial_t u - \partial_x^2 u$$

und d.h.

$$A = \begin{pmatrix} I & \\ & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}, \quad c = 0.$$

A hat $d - 1$ positive Eigenwerte und $\text{Rang}(A, b) = d$, also \mathcal{L} parabolisch.

iii. Wellengleichung ($\partial_t^2 u - \Delta u = f$):

Dabei ist

$$\mathcal{L}u = \partial_t^2 u - \Delta u$$

und d.h.

$$A = \begin{pmatrix} I & \\ & -1 \end{pmatrix}, \quad b = 0, \quad c = 0.$$

A hat $d - 1$ positive Eigenwerte und einen negativen Eigenwert, also ist \mathcal{L} hyperbolisch.

iv. Tricomi-Gleichung ($x_2 \partial_{x_1}^2 u + \partial_{x_2}^2 u = 0$ für $\Omega = \mathbb{R}^2$):

Dabei ist

$$\mathcal{L}u = x_2 \partial_{x_1}^2 u + \partial_{x_2}^2 u$$

und d.h.

$$A = \begin{pmatrix} x_2 & \\ & 1 \end{pmatrix}, \quad b = 0, \quad c = 0.$$

Also ist die Gleichung vom gemischten Typ:

- elliptisch in $(x_1, x_2) \in \mathbb{R} \times \mathbb{R}_+$,
- hyperbolisch in $(x_1, x_2) \in \mathbb{R} \times \mathbb{R}_-$.

Aber für $x_2 = 0$ ist die Gleichung nicht parabolisch, denn dann ist $\text{Rang}(A, b) = 1 \neq 2$.

Bemerkung. (Charakterisierung)

Unterscheidung in Typen ist sinnvoll wegen wesentlich unterschiedlichen Lösungseigenschaften:

- Elliptische PDEs:
 - Lösungen sind oft sehr glatt und meistens sind die Randbedingungen vorgegeben.
 - Lösungen erfüllen häufig „Maximumsprinzip“ (dies wird später erklärt).

- Parabolische PDEs:

- „Ausgezeichnete“ Achse ist meistens die Zeitachse. Die PDE kann oft umgeschrieben werden als

$$\partial_t u + \tilde{\mathcal{L}} u = \tilde{f}$$

wobei $\tilde{\mathcal{L}}$ ein elliptischer Operator auf dem Raum erzeugt von den anderen $d - 1$ Eigenvektoren.

- Häufige Anfangswerte für u und ggf. Randwerte für u vorgegeben.
- Operator hat regularisierenden Effekt, d.h. Lösungen sind häufig glatter als Anfangsdaten.
- „Unendliche Ausbreitungsgeschwindigkeit“ kann man dabei beobachten.

- Hyperbolische PDEs:

- „Ausgezeichnete“ Achse ist meistens die Zeitachse. Die PDE kann oft umgeschrieben werden als

$$\partial_t^2 u + \tilde{\mathcal{L}} u = \tilde{f}$$

wobei $\tilde{\mathcal{L}}$ ein elliptischer Operator (bzgl. x) mit negativen Eigenwerten ist.

- Solche PDEs beschreiben Schwingungsvorgänge.
- Es werden Anfangsbedingungen für u und für $\partial_t u$ vorgegeben, ggf. noch Randbedingungen für u .
- „Endliche Ausbreitungsgeschwindigkeit“ liegt meist vor.
- Kein regularisierender Effekt lässt sich beobachten.

Bemerkung. (Eigenschaften allgemeiner elliptischer Operatoren)

- Falls \mathcal{L} rotationsinvariant ist, dann ist $\mathcal{L}u = -a\Delta u + cu$ für ein $a \in \mathbb{R}_+$, d.h. $A = aI$ und $b = 0$.
- Wir nennen \mathcal{L} gleichmäßig elliptisch, falls $\exists \alpha \in \mathbb{R}_+$ s.d.

$$\forall z \in \mathbb{R}^d \quad \forall x \in \Omega: \quad \langle z, A(x)z \rangle \geq \alpha \|z\|^2,$$

also sind alle Eigenwerte von $A(x)$ mind. so groß wie α . Dieses α nennt man Elliptizitätskonstante.

- Für gleichmäßige elliptische PDEs folgen auch Maximums, Minimums & Vergleichsprinzip. Dies wird bald für Poisson-Gleichung betrachtet.

Im Folgenden beschränken wir unsere Betrachtung auf eine spezielle PDE:

2.2. POISSON-GLEICHUNG

Definition 2.20. (Poisson / Laplace-Gleichung)

Sei $\Omega \subseteq \mathbb{R}^d$.

Wir nennen die PDE

$$-\Delta u = 0$$

Laplace-Gleichung.

Für ein $f: \Omega \rightarrow \mathbb{R}$ nennen wir die PDE

$$-\Delta u = f$$

die Poisson-Gleichung.

Bemerkung.

- Das Vorzeichen ist eigentlich willkürlich, aber so taucht es in Def. 2.18 auf.
- Wir erwarten i.A. keine Eindeutigkeit von Lösungen:
Mit u einer Lösung ist auch $u + c$ für jedes $c \in \mathbb{R}$ auch eine Lösung von Laplace- & von Poisson-Gleichung.
 \rightsquigarrow Weitere Bedingung erforderlich, also Randbedingung.
- Klassische Lösungen von $-\Delta u = 0$ nennen wir auch harmonisch.

Bemerkung. (Rotations- & Translationsinvianz)

- Falls f, Ω rotationssymmetrisch sind, d.h. für $\forall T \in O_d(\mathbb{R})$ gilt $\Omega = T\Omega$ und für $\forall x \in \Omega: f(x) = f(Tx)$, dann liefert jede klassische Lösung der Poisson-Gleichung $u \in C^2(\Omega)$ mit T eine neue klassische Lösung, nämlich $v := u \circ T$.
- Falls f, Ω translationsinvariant bzgl. $t \in \mathbb{R}^d$ sind, d.h. $\Omega = \Omega + t$ sowie $\forall x \in \Omega: f(x) = f(x + t)$, dann ist mit u klassischer Lösung auch $v(x) := u(x + t)$ eine klassische Lösung der Poisson-Gleichung.

Unser **Ziel** besteht darin, Lösungen der Poisson-Gleichung zu finden und ihre Eigenschaften zu untersuchen.

Die Lösungen erhalten wir in zwei Schritten:

- i. Konstruiere für Laplace-Gleichung ($f = 0$) auf $\Omega := \mathbb{R}^d \setminus \{0\}$ spezielle rotationssymmetrische Lösungen, also die sogenannten „Fundamentallösungen“;

- ii. Konstruiere für die Poisson-Gleichung (i.A. $f \neq 0$) Lösungen durch Faltung mit Fundamentallösungen.

Bemerkung. (Ansatz für Fundamentallösung der Laplace-Gleichung)

Annahme von Rotationssymmetrie & Umschreiben der PDE als ODE:

- Sei $u \in C^2(\Omega)$ klassische Lösung der Laplace-Gleichung für $\Omega = \mathbb{R}^d \setminus \{0\}$ rotationsinvariant, d.h. es existiert ein $v \in C^2((0, \infty))$ mit $u(x) = v(\|x\|_2)$.
- Für $x \in \mathbb{R}^d$ schreiben wir $r(x) := \|x\|_2 = (\langle x, x \rangle)^{\frac{1}{2}}$. Es ist also

$$u(x) = v(r(x))$$

und wir werden, wenn es keine Mehrdeutigkeit gibt, $r(x)$ mit r abkürzen.

- Wir sehen zunächst

$$\nabla_x r = \left((\langle x, x \rangle)^{\frac{1}{2}} \right)' = \frac{1}{2} (\langle x, x \rangle)^{-\frac{1}{2}} (\langle x, x \rangle)' = \frac{1}{2} (\langle x, x \rangle)^{-\frac{1}{2}} 2x = \frac{x}{r}$$

wobei man $(\langle x, x \rangle)' = 2x$ aus

$$\langle x + h, x + h \rangle = \langle x, x \rangle + 2\langle x, h \rangle + \langle h, h \rangle$$

für $h \rightarrow 0$ sieht, also insbesondere

$$\partial_{x_i} r(x) = \frac{x_i}{r(x)},$$

und damit

$$\partial_{x_i} u = v'(r) \partial_{x_i} r = v'(r) \frac{x_i}{r}$$

sowie

$$\partial_{x_i}^2 u = v''(r) \left(\frac{x_i}{r} \right)^2 + v'(r) \left(\frac{x_i}{r} \right)' = v''(r) \left(\frac{x_i}{r} \right)^2 + v'(r) \frac{r - x_i \frac{x_i}{r}}{r^2}.$$

- Das bedeutet

$$\begin{aligned} \Delta u = \sum_{i=1}^d \partial_{x_i} u &= v''(r) \frac{1}{r^2} \sum_{i=1}^d (x_i)^2 + v'(r) \frac{1}{r^2} \sum_{i=1}^d \left(r - \frac{1}{r} x_i^2 \right) \\ &= v''(r) \frac{1}{r^2} r^2 + v'(r) \frac{1}{r^2} \left(dr - \frac{1}{r} r^2 \right) \\ &= v''(r) + v'(r) \frac{1}{r} (d-1) \end{aligned}$$

- Als notwendige Bedingung muss v die folgende ODE lösen, also

$$0 = \Delta u = v''(r) + v'(r) \frac{1}{r} (d-1).$$

- Wir nehmen zusätzlich an, dass v streng monoton ist, d.h. $v' > 0$, dann ist

$$\frac{v''(r)}{v'(r)} = \frac{1-d}{r}.$$

- Beachte: $\ln(v'(r))' = \frac{1}{v'(r)} v''(r)$.
- Wähle ein $r_0 \in (0, \infty)$, dann gilt es bei der obigen ODE:

$$\begin{aligned} \ln(v'(r)) - \ln(v'(r_0)) &= \int_{r_0}^r \frac{v''(s)}{v'(s)} d\lambda^1(s) \\ &= \int_{r_0}^r \frac{1-d}{s} d\lambda^1(s) = (1-d)\ln(r) - (1-d)\ln(r_0), \end{aligned}$$

also nach Umformung

$$\ln(v'(r)) = (1-d)\ln(r) - (1-d)\ln(r_0) + \ln(v'(r_0)) = \ln(r^{1-d} \cdot v'(r_0) \cdot r_0^{d-1}).$$

- Mit $a := v'(r_0) \cdot r_0^{d-1}$ und Anwendung von \exp auf beider Seiten liefert

$$v'(r) = ar^{1-d}$$

und das bedeutet, dass es ein $b \in \mathbb{R}$ gibt, s.d.

$$v(r) = \begin{cases} ar + b, & d = 1 \\ a\ln(r) + b, & d = 2 \\ \frac{a}{(2-d)r^{d-2}} + b, & d \geq 3 \end{cases}$$

- Zwar scheint die Wahl von a abhängig von r_0 und $v'(r_0)$ zu sein, aber man kann nachrechnen, dass v mit jedem $a \in \mathbb{R}$ die Laplace-Gleichung löst.

Definition 2.21. (Fundamentallösungen)

Sei $\Omega := \mathbb{R}^d \setminus \{0\}$ und $d > 1$.

Wir schreiben $\omega_d = |\partial B_1(0)| := \lambda^{d-1}(\partial B_1(0))$ als die Oberfläche der Einheits sphäre in \mathbb{R}^d .

Die Funktion $\Phi \in C^\infty(\Omega)$ definiert durch

$$\Phi(x) := \begin{cases} -\frac{1}{2\pi} \ln(\|x\|), & d = 2 \\ \frac{1}{(d-2)\omega_d} \frac{1}{\|x\|^{d-2}}, & d \geq 3 \end{cases} \quad (2.3)$$

ist eine klassische Lösung der Laplace-Gleichung und wird oft als Fundamentallösung der Laplace-Gleichung bezeichnet.

Bemerkung.

- Φ kann man salopp aber kompakt schreiben als $\int \frac{-1}{\omega_d} \|x\|^{1-d} d\|x\|$.
- Φ hat Singularität bei $x = 0$, aber wir werden gleich sehen, dass diese Singularität „nicht schlimm“ ist.
- Der Skalierungsfaktor ist so gewählt, dass die kommende Eigenschaft v „schön“ wird:

Lemma 2.22. (Eigenschaften von Φ)

Sei $\Omega := \mathbb{R}^d \setminus \{0\}$ für ein $d \in \mathbb{N}_{\geq 2}$.

Eine Fundamentallösung der Laplace-Gleichung Φ gemäß Def. 2.21 besitzt die folgenden Eigenschaften:

i. Zu jedem $\varepsilon \in \mathbb{R}_+$ ist $|\Phi|$ integrierbar in $B_\varepsilon(0)$, also

$$\forall \varepsilon \in \mathbb{R}_+: \quad \int_{B_\varepsilon(0)} |\Phi| d\lambda^d < \infty.$$

ii. Für $\varepsilon \rightarrow 0$ konvergiert auch das Integral von $|\Phi|$ in $B_\varepsilon(0)$ gegen 0, also

$$\lim_{\varepsilon \rightarrow 0} \int_{B_\varepsilon(0)} |\Phi| d\lambda^d = 0.$$

iii. $\Phi \in L^1_{\text{loc}}(\mathbb{R}^d)$, d.h. Φ ist Lebesgue-messbar und für jedes $K \subseteq \mathbb{R}^d$ kompakt $\Phi \in L^1(K)$.

iv. In jeder Richtung wächst Φ langsamer als ε^{1-d} , also für $\forall e \in \mathbb{R}^d$ mit $\|e\|_2 = 1$:

$$\lim_{\varepsilon \rightarrow 0} \Phi(\varepsilon e) \varepsilon^{d-1} = 0.$$

v. Der Ausfluss von $\nabla \Phi$ auf $\partial B_\varepsilon(0)$ ergibt -1 , also

$$\int_{\partial B_\varepsilon(0)} \nabla \Phi \bullet n d\lambda^d = -1.$$

Beweis.

i. Wegen Rotationsinvarianz von Φ und $B_\varepsilon(0)$ gilt

$$\begin{aligned} \int_{B_\varepsilon(0)} |\Phi| d\lambda^d &= \int_0^\varepsilon \int_{\partial B_r(0)} |\Phi(re)| d\lambda^{d-1} d\lambda(r) \\ &= \int_0^\varepsilon |\Phi(re)| \left(\int_{\partial B_r(0)} d\lambda^{d-1} \right) d\lambda(r) \\ &= \int_0^\varepsilon |\Phi(re)| \omega_d r^{d-1} d\lambda(r). \end{aligned}$$

• Falls $d = 2$:

Es ist $\Phi(re) = -\frac{1}{2\pi} \ln(r)$ und $\omega_d = 2\pi$ und somit

$$\begin{aligned} \int_{B_\varepsilon(0)} |\Phi| d\lambda^d &= \int_0^\varepsilon |\Phi(re)| \omega_d r^{d-1} d\lambda(r) \\ &= \int_0^\varepsilon \frac{1}{2\pi} |\ln(r)| 2\pi r d\lambda(r) \\ &= \int_0^\varepsilon \ln(r) r d\lambda(r). \end{aligned}$$

Dank L'Hospital gilt

$$\lim_{r \rightarrow 0} \ln(r)r = \lim_{r \rightarrow 0} \frac{\ln(r)}{r^{-1}} = \lim_{r \rightarrow 0} \frac{r^{-1}}{-r^{-2}} = \lim_{r \rightarrow 0} -r = 0$$

und daher ist $\ln(r)r \in C([0, \varepsilon])$, also obiges Integral endlich.

- Falls $d \geq 3$:

Es ist $\Phi(re) = \frac{1}{(d-2)\omega_d} r^{2-d}$ und damit

$$\begin{aligned} \int_{B_\varepsilon(0)} |\Phi| d\lambda^d &= \int_0^\varepsilon |\Phi(re)| \omega_d r^{d-1} dr \\ &= \int_0^\varepsilon \frac{1}{(d-2)\omega_d} r^{2-d} \omega_d r^{d-1} dr \\ &= \frac{1}{d-2} \int_0^\varepsilon r dr \\ &= \frac{\varepsilon^2}{2(d-2)} < \infty. \end{aligned}$$

ii. Folgt unmittelbar aus dem Beweis von i.

iii. Sei $K \subseteq \mathbb{R}^d$ kompakt.

a) Falls $0 \notin K$, dann ist Φ stetig auf K , also nimmt Maximum an, daher

$$\int_K |\Phi| d\lambda^d \leq \lambda^d(K) \sup_{x \in K} |\Phi(x)| < \infty.$$

b) Falls $0 \in K^\circ$, dann existiert ein $\varepsilon \in \mathbb{R}_+$ s.d. $B_\varepsilon(0) \subseteq K$, daher

$$\int_K |\Phi| d\lambda^d = \int_{K \setminus B_\varepsilon(0)} |\Phi| d\lambda^d + \int_{B_\varepsilon(0)} |\Phi| d\lambda^d$$

wobei erster Term $< \infty$ wegen a) und zweiter Term $< \infty$ wegen i.

c) Falls $0 \in \partial K$, so setze $K' := K \cup \overline{B_\varepsilon(0)}$, und mit b) folgt dann

$$\int_K |\Phi| d\lambda^d \leq \int_{K'} |\Phi| d\lambda^d < \infty.$$

iv. $d = 2$: $\Phi(\varepsilon e) \varepsilon^{d-1} = -\frac{1}{2\pi} |\ln(\varepsilon)| \varepsilon^{2-1} \xrightarrow{\varepsilon \rightarrow 0} 0$.

$$d = 3: \Phi(\varepsilon e) \varepsilon^{d-1} = \frac{1}{(d-2)\omega_d} \varepsilon^{2-d} \varepsilon^{d-1} = \frac{1}{(d-2)\omega_d} \varepsilon \xrightarrow{\varepsilon \rightarrow 0} 0.$$

v. Dank Rotationsinvarianz von Φ und Ω ist $\nabla \Phi$ auch rotationsinvariant, und somit $\nabla \Phi \bullet n$ auch. D.h. $\nabla \Phi \bullet n$ nimmt auf $\partial B_\varepsilon(0)$ einen konstanten Wert, also gilt

$$\forall x \in \partial B_\varepsilon(0) \quad \forall e \in \mathbb{R}^d \text{ mit } \|e\|_2 = 1: \nabla \Phi(x) \bullet n(x) = \nabla \Phi(\varepsilon e) \bullet n(\varepsilon e) = \nabla \Phi(\varepsilon e) \bullet e,$$

und daher

$$\int_{\partial B_\varepsilon(0)} \nabla \Phi \bullet n d\lambda^d = \nabla \Phi(\varepsilon e) \bullet e \int_{\partial B_\varepsilon(0)} d\lambda^d = \nabla \Phi(\varepsilon e) \bullet e |\partial B_\varepsilon(0)|.$$

Beachte: $\nabla\Phi(\varepsilon e) \bullet e$ ist die Richtungsableitung von Φ in Richtung e an der Stelle εe , also $\nabla\Phi(\varepsilon e) \bullet e = \frac{d}{d\varepsilon}\Phi(\varepsilon e)$, und mit $|\partial B_\varepsilon(0)| = \omega_d \varepsilon^{d-1}$ folgt daher

$$\int_{\partial B_\varepsilon(0)} \nabla\Phi \bullet n \, d\lambda^d = \nabla\Phi(\varepsilon e) \bullet e |\partial B_\varepsilon(0)| = \frac{d}{d\varepsilon} \nabla\Phi(\varepsilon e) \omega_d \varepsilon^{d-1}.$$

- $d=2$:

$$\begin{aligned} \Phi(\varepsilon e) &= -\frac{1}{2\pi} \ln(\varepsilon) \quad \Rightarrow \frac{d}{d\varepsilon} \nabla\Phi(\varepsilon e) = -\frac{1}{2\pi} \frac{1}{\varepsilon} \\ &\Rightarrow \frac{d}{d\varepsilon} \nabla\Phi(\varepsilon e) \omega_d \varepsilon^{d-1} = -\frac{1}{2\pi} \frac{1}{\varepsilon} \omega_d \varepsilon^{d-1} = -1. \end{aligned}$$

- $d \geq 3$:

$$\begin{aligned} \Phi(\varepsilon e) &= \frac{1}{(d-2)\omega_d} \varepsilon^{2-d} \quad \Rightarrow \frac{d}{d\varepsilon} \nabla\Phi(\varepsilon e) = \frac{2-d}{(d-2)\omega_d} \varepsilon^{1-d} = -\frac{1}{\omega_d} \varepsilon^{1-d} \\ &\Rightarrow \frac{d}{d\varepsilon} \nabla\Phi(\varepsilon e) \omega_d \varepsilon^{d-1} = -\frac{1}{\omega_d} \varepsilon^{1-d} \omega_d \varepsilon^{d-1} = -1. \end{aligned} \quad \square$$

Wir haben gesagt, dass wir eine Lösung der Poisson-Gleichung durch Faltung erhalten, und bevor wir dies tun, sollen wir den Zusammenhang zwischen Faltung und Differentiation herstellen:

Satz 2.23. (Faltung & Differentiation)

Zu $u \in L^1_{loc}(\mathbb{R}^d)$ und $\phi \in C_0^\infty(\mathbb{R}^d)$ definieren wir die Faltung $u * \phi: \mathbb{R}^d \rightarrow \mathbb{R}$ durch

$$(u * \phi)(x) := \int_{\mathbb{R}^d} u(x-y) \phi(y) \, d\lambda^{d-1}(y).$$

Dabei gilt:

- $u * \phi \in C^m(\mathbb{R}^d)$.
- $\forall \beta \in \mathbb{B}_m: \partial^\beta(u * \phi) = u * \partial^\beta \phi$.

Beweis. Hier wird nur der Fall $m=1$ betrachtet (vollständige Version siehe Alt, Abschnitt 2.8.):

- Das Integral ist endlich, also Faltung wohldefiniert, denn:

Für ein $x \in \mathbb{R}^d$ setze $M_x := \{x-z \mid z \in \text{supp}(\phi)\} \subseteq \mathbb{R}^d$, damit gilt

$$\begin{aligned} \left| \int_{\mathbb{R}^d} u(x-y) \phi(y) \, d\lambda^{d-1}(y) \right| &\leq \|\phi\|_\infty \int_{\text{supp}(\phi)} |u(x-y)| \, d\lambda^{d-1}(y) \\ &= \|\phi\|_\infty \int_{M_x} |u(y)| \, d\lambda^{d-1}(y) \\ &< \infty \end{aligned}$$

da $\phi \in C_0^1(\mathbb{R}^d)$ und $u \in L^1_{loc}(\mathbb{R}^d)$.

- Dank Variablentransformation gilt die Symmetrie in der Integration, also

$$(u * \phi)(x) = \int_{\mathbb{R}^d} u(y) \phi(x - y) d\lambda^d(y).$$

- Damit gilt es für die Ableitung:

$$\begin{aligned} \partial_{x_i}(u * \phi)(x) &= \lim_{h \rightarrow 0} \frac{(u * \phi)(x + h e_i) - (u * \phi)(x)}{h} \\ &= \lim_{h \rightarrow 0} \int_{\mathbb{R}^d} u(y) \frac{\phi(x + h e_i - y) - \phi(x, y)}{h} d\lambda^d(y). \end{aligned}$$

Da $\phi \in C_0^1(\mathbb{R}^d)$, also insbesondere ϕ Lipschitzstetig (Satz 2.4), gilt für $h \rightarrow 0$

$$\frac{\phi(x + h e_i - \bullet) - \phi(x, \bullet)}{h} \xrightarrow{\text{gleichmäßig}} \partial_{x_i} \phi(x - \bullet).$$

Somit sind die obigen Grenzwerte vertauschbar, also

$$\begin{aligned} \partial_{x_i}(u * \phi)(x) &= \lim_{h \rightarrow 0} \int_{\mathbb{R}^d} u(y) \frac{\phi(x + h e_i - y) - \phi(x, y)}{h} d\lambda^d(y) \\ &= \int_{\mathbb{R}^d} u(y) \lim_{h \rightarrow 0} \frac{\phi(x + h e_i - y) - \phi(x, y)}{h} d\lambda^d(y) \\ &= \int_{\mathbb{R}^d} u(y) \partial_{x_i} \phi(x - y) d\lambda^d(y) \\ &= \partial_{x_i}(u * \partial_{x_i} \phi)(x). \end{aligned}$$

Analog für höhere Ableitungen. □

Satz 2.24. (Faltungslösung der Poisson-Gleichung)

Sei $\Omega := \mathbb{R}^d$ für $d \in \mathbb{N}_{\geq 2}$, $f \in C_0^2(\Omega)$ und Φ eine Fundamentallösung der Laplace-Gleichung gemäß Def. 2.21.

Dann besitzt die Poisongleichung $-\Delta u = f$ eine klassische Lösung $u := \Phi * f$.

Beweis.

a) Nach 2.22 iii) ist $\Phi \in L^1_{\text{loc}}(\mathbb{R}^d)$ und daher erhalten wir mit 2.23

$$\partial_{x_i}^2 u = \partial_{x_i}^2 \Phi * f = \Phi * \partial_{x_i}^2 f,$$

somit gilt

$$-\Delta u(x) = - \int_{\mathbb{R}^d} \Phi(y) \Delta_x f(x - y) d\lambda^d(y) = \int_{\mathbb{R}^d} \Phi(y) \Delta_y f(x - y) d\lambda^d(y).$$

Für ein $\varepsilon \in \mathbb{R}_+$ zerlegen wir das Integral in zwei Teile:

$$\begin{aligned} -\Delta u(x) &= \int_{\mathbb{R}^d} \Phi(y) \Delta_y f(x - y) d\lambda^d(y) \\ &= \int_{B_\varepsilon(0)} \Phi(y) \Delta_y f(x - y) d\lambda^d(y) + \int_{\mathbb{R}^d \setminus B_\varepsilon(0)} \Phi(y) \Delta_y f(x - y) d\lambda^d(y) \\ &=: I_\varepsilon(x) + J_\varepsilon(x) \end{aligned}$$

und wir werden die beiden Teile weiter untersuchen.

b) Zu I_ε :

Beschränktheit von $\Delta_y f(x - y) \in C_0(\Omega)$ und 2.22 ii) liefert

$$\forall x \in \Omega: |I_\varepsilon(x)| \leq \|\Delta_y f(x - \bullet)\|_\infty \int_{B_\varepsilon(0)} |\Phi(y)| d\lambda^d(y) \xrightarrow{\varepsilon \rightarrow 0} 0.$$

c) Zu J_ε :

Wir integrieren zunächst J_ε partiell gemäß Satz 2.15

$$\begin{aligned} J_\varepsilon(x) &= \int_{\mathbb{R}^d \setminus B_\varepsilon(0)} \Phi(y) \Delta_y f(x - y) d\lambda^d(y) \\ &= \int_{\mathbb{R}^d \setminus B_\varepsilon(0)} \Phi(y) \nabla_y \operatorname{div}_y(f(x - y)) d\lambda^d(y) \\ &= - \int_{\mathbb{R}^d \setminus B_\varepsilon(0)} \nabla_y \Phi(y) \bullet \nabla_y f(x - y) d\lambda^d(y) \\ &\quad + \int_{\partial B_\varepsilon(0)} \Phi(y) \nabla_y f(x - y) \bullet n d\lambda^{d-1}(y) \\ &=: T_0 + T_3 \end{aligned}$$

Dann führen wir nochmal partielle Integration für T_0 durch, also

$$\begin{aligned} T_0 &= - \int_{\mathbb{R}^d \setminus B_\varepsilon(0)} \nabla_y \Phi(y) \bullet \nabla_y f(x - y) d\lambda^d(y) \\ &= \int_{\mathbb{R}^d \setminus B_\varepsilon(0)} \Delta_y \Phi(y) f(x - y) d\lambda^d(y) - \int_{\partial B_\varepsilon(0)} (\nabla_y \Phi(y) \bullet n) f(x - y) d\lambda^{d-1}(y) \\ &=: T_1 + T_2. \end{aligned}$$

wobei $T_1 = 0$ gilt, denn mit Φ Fundamentallösung der Laplace-Gleichung ist $\Delta_y \Phi(y) = 0$.

d) Zu T_3 :

Mit $f \in C_0^2(\Omega)$ ist $\nabla_y f(x - y) \in C_0^1(\Omega, \mathbb{R}^d)$ also beschränkt, d.h.

$$\exists C \in \mathbb{R}_+ \forall x \in \Omega: |\nabla_y f(x - y)| \leq C.$$

Zudem ist Φ wegen Rotationssymmetrie auf $\partial B_\varepsilon(0)$ konstant, d.h. für ein $e \in \Omega$ mit $\|e\| = 1$ gilt

$$\int_{\partial B_\varepsilon(0)} |\Phi(y)| d\lambda^{d-1}(y) = \int_{\partial B_\varepsilon(0)} |\Phi(\varepsilon e)| d\lambda^{d-1}(y) = |\Phi(\varepsilon e)| \omega_d \varepsilon^{d-1}.$$

Zusammen mit 2.22 iv) gilt daher

$$\begin{aligned} |T_3| &= \int_{\partial B_\varepsilon(0)} \Phi(y) \nabla_y f(x - y) \bullet n d\lambda^{d-1}(y) \\ &\leq C \int_{\partial B_\varepsilon(0)} |\Phi(y)| d\lambda^{d-1}(y) \\ &= C |\Phi(\varepsilon e)| \omega_d \varepsilon^{d-1} \xrightarrow{\varepsilon \rightarrow 0} 0. \end{aligned}$$

e) Zu T_2 :

Mit dem Trick „Addieren mit 0“ schreiben wir T_2 geschickt um:

$$\begin{aligned} T_2 &= - \int_{\partial B_\varepsilon(0)} (\nabla_y \Phi(y) \bullet n) f(x-y) d\lambda^{d-1}(y) \\ &= - \int_{\partial B_\varepsilon(0)} (\nabla_y \Phi(y) \bullet n) (f(x-y) + f(x) - f(x)) d\lambda^{d-1}(y) \\ &= - f(x) \int_{\partial B_\varepsilon(0)} \nabla_y \Phi(y) \bullet n d\lambda^{d-1}(y) \\ &\quad - \int_{\partial B_\varepsilon(0)} (\nabla_y \Phi(y) \bullet n) (f(x-y) - f(x)) d\lambda^{d-1}(y) \\ &=: T_4 + T_5 \end{aligned}$$

und wir sehen leicht, dass mit 2.22 v) gilt

$$T_4 = -f(x) \int_{\partial B_\varepsilon(0)} \nabla_y \Phi(y) \bullet n d\lambda^{d-1}(y) = -f(x) \cdot (-1) = f(x).$$

f) Zu T_5 :

Da $f \in C_0^2(\Omega)$, ist f insbesondere Lipschitzstetig, und damit

$$\exists K \in \mathbb{R}_+ \forall x \in \Omega \forall y \in \partial B_\varepsilon(0): |f(x-y) - f(x)| \leq K|y| \leq K\varepsilon.$$

Zudem ist Φ wegen Rotationssymmetrie auf $\partial B_\varepsilon(0)$ konstant, also insbesondere auch $\nabla_y \Phi(y) \bullet n$ konstant auf $\partial B_\varepsilon(0)$.

Somit gilt

$$\begin{aligned} |T_5| &= \left| - \int_{\partial B_\varepsilon(0)} (\nabla_y \Phi(y) \bullet n) (f(x-y) - f(x)) d\lambda^{d-1}(y) \right| \\ &\leq \int_{\partial B_\varepsilon(0)} |(\nabla_y \Phi(y) \bullet n)| K\varepsilon d\lambda^{d-1}(y) \\ &= K\varepsilon \left| \int_{\partial B_\varepsilon(0)} (\nabla_y \Phi(y) \bullet n) d\lambda^{d-1}(y) \right| \\ &= K\varepsilon \quad 1 \end{aligned}$$

wobei die letzte Zeile aus 2.22 v) folgt, also wir erhalten

$$|T_5| \leq K\varepsilon \xrightarrow{\varepsilon \rightarrow 0} 0.$$

g) Zusammengefasst:

$$-\Delta u(x) = \lim_{\varepsilon \rightarrow 0} J_\varepsilon(x) = \lim_{\varepsilon \rightarrow 0} T_4 = f(x). \quad \square$$

Jetzt untersuchen wir weiter die Eigenschaften der Laplace-Gleichung:

Definition 2.25. (Mittelwerte)

Sei $K \subseteq \mathbb{R}^d$ mit $|K| := \lambda^d(K) < \infty$ und $|\partial K| := \lambda^{d-1}(\partial K) < \infty$.

Für $u \in L^1(K)$ und $u \in L^1(\partial K)$ definieren wir den Mittelwert von u auf K

$$\int_K u \, d\lambda^d := \frac{1}{|K|} \int_K u \, d\lambda^d$$

sowie den Randmittelwert von u auf K

$$\int_{\partial K} u \, d\lambda^{d-1} := \frac{1}{|\partial K|} \int_{\partial K} u \, d\lambda^{d-1}.$$

Bemerkung. Die Anforderungen $|\partial K| := \lambda^{d-1}(\partial K) < \infty$ bzw. $u \in L^1(\partial K)$ sind notwendig, um den Randmittelwert zu definieren, denn $\lambda^d(\partial K) = 0$ und somit könnte u auf ∂K nicht wohldefiniert sein.

Satz 2.26. (Mittelwerte harmonischer Funktionen)

Sei $u \in C^2(\Omega)$ harmonisch, d.h. $\Delta u = 0$.

Dann gilt für jedes $x \in \Omega$ und jedes $r \in \mathbb{R}_+$ s.d. $\overline{B_r(x)} \subseteq \Omega$:

$$u(x) = \int_{B_r(x)} u \, d\lambda^d = \int_{\partial B_r(x)} u \, d\lambda^{d-1}.$$

Beweis.

Siehe Blatt 7 Aufg 1. □

Satz 2.27. (Maximumsprinzip für harmonische Funktionen)

Sei $\Omega \subseteq \mathbb{R}^d$ offen und beschränkt, $u \in C^2(\bar{\Omega})$ mit $-\Delta u = 0$.

Dann gilt:

i. u nimmt Maximum auf dem Rand an, also

$$\max_{x \in \bar{\Omega}} u(x) = \max_{x \in \partial \Omega} u(x)$$

ii. Falls das Maximum in Ω angenommen wird und Ω zusammenhängend ist, dann ist u eine konstante Funktion.

Beweis.

- Zunächst ii.:

Sei $x \in \Omega$ mit $u(x) = \max_{y \in \bar{\Omega}} u(y)$.

Für jedes $r \in \mathbb{R}_+$ mit $B_r(x) \subseteq \Omega$ gilt wegen Satz 2.26

$$u(x) = \int_{B_r(x)} u \, d\lambda^d. \tag{2.4}$$

Weil $u(x)$ das Maximum ist, muss es sein, dass $\forall y \in B_r(x): u(y) = u(x)$.

Setze $M := \{y \in \Omega \mid u(y) = u(x)\}$.

M ist offen, denn für ein $y \in M$, mit Argumentation wie zuvor sehen wir, dass es für $\forall r \in \mathbb{R}_+$ mit $B_r(y) \subseteq \Omega$ gilt $u|_{B_r(y)} = u(y) = u(x)$, also $B_r(y) \subseteq M$.

M ist abgeschlossen, denn M ist Urbild von $\{0\}$ bzgl. der stetigen Abbildung $\varphi: \Omega \rightarrow \mathbb{R}, y \mapsto u(y) - u(x)$.

Weil Ω zusammenhängend ist, sind \emptyset und Ω die einzigen offen & abgeschlossenen Teilmengen, und da M offenbar nicht leer ist, gilt $M = \Omega$.

- Nun i.:

Sei $x_0 \in \bar{\Omega}$ mit $u(x_0) = \max_{y \in \bar{\Omega}} u(y)$.

Falls $x_0 \in \partial\Omega$, so ist die Aussage klar.

Falls $x_0 \in \Omega$, dann wählen wir $\Omega_0 \subseteq \Omega$ zusammenhängend mit $x_0 \in \Omega_0$ und

$$\partial\Omega_0 \cap \partial\Omega \neq \emptyset. \quad (2.5)$$

ii. besagt, dass u auf Ω_0 konstanten Wert annimmt, nämlich $u(x_0)$.

Dank Stetigkeit von u auf $\bar{\Omega}$ nimmt u auf $\partial\Omega_0$ auch den Wert $u(x_0)$ an.

Die Bedingung (2.5) liefert dann die Behauptung. \square

Bemerkung. (Verallgemeinerungen)

- Maximumsprinzip

Für $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$ mit $-\Delta u \leq 0$ gilt:

u nimmt Maximum auf dem Rand an.

- Minimumsprinzip

Für $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$ mit $-\Delta u \geq 0$ gilt:

u nimmt Minimum auf dem Rand an.

- Vergleichsprinzip

Für $u, v \in C^2(\Omega) \cap C^0(\bar{\Omega})$ gilt:

Falls $-\Delta u \leq -\Delta v$ in Ω und $u \leq v$ auf $\partial\Omega$, dann ist $u \leq v$ in Ω .

Beweis. Wir zeigen hier den Vergleichsprinzip:

Setze $w := u - v$, also $-\Delta w \leq 0$, und wegen des Maximumsprinzips nimmt w auf Rand das Maximum an.

Damit erhalten wir für $\forall x \in \bar{\Omega}$:

$$w(x) \leq \max_{x \in \bar{\Omega}} w(x) = \max_{x \in \partial\Omega} w(x) = \max_{x \in \partial\Omega} (u(x) - v(x)) \leq 0$$

also $u \leq v$ auf Ω . \square

Daraus können wir sofort folgern:

Folgerung 2.28. (Eindeutigkeit Poisson-RWP)

Sei $\Omega \subseteq \mathbb{R}^d$ offen und beschränkt, $g \in C^0(\partial\Omega)$ und $f \in C^0(\Omega)$.

Dann existiert höchstens eine Lösung $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$ des Poisson-RWP

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

Beweis. Seien u, \tilde{u} zwei klassische Lösungen des Poisson-RWP.

Dann ist $u - \tilde{u}$ harmonisch mit Nullrandwerten, denn

$$-\Delta(u - \tilde{u}) = -\Delta u + \Delta \tilde{u} = f - f = 0 \quad \text{in } \Omega,$$

$$u - \tilde{u} = g - g = 0 \quad \text{auf } \partial\Omega.$$

Mit Satz 2.27 nimmt $u - \tilde{u}$ Maximum auf Rand an, also

$$u - \tilde{u} \leq \max_{y \in \partial\Omega} (u(y) - \tilde{u}(y)) = 0,$$

und analog erhalten wir $\tilde{u} - u \leq 0$, also $u = \tilde{u}$. \square

Wir bemerken hier, dass die Existenz der Lösung noch nicht gezeigt ist. Dies wird im Kapitel 3 betrachtet.

Folgerung 2.29. (Stetige Abhängigkeit von Randdaten)

Sei $\Omega \subseteq \mathbb{R}^d$ offen und beschränkt.

Seien $u, \tilde{u} \in C^2(\Omega) \cap C^0(\bar{\Omega})$ klassische Lösungen von Poisson-RWP zu identischen $f \in C^0(\Omega)$, aber verschiedenen Randwerten $g, \tilde{g} \in C^0(\partial\Omega)$.

Dann gilt:

$$\|u - \tilde{u}\|_\infty = \max_{x \in \bar{\Omega}} |u(x) - \tilde{u}(x)| \leq \max_{x \in \partial\Omega} |g(x) - \tilde{g}(x)| = \|g - \tilde{g}\|_\infty.$$

Beweis. Die Funktion $w := u - \tilde{u}$.

Mit $-\Delta w = -\Delta(u - \tilde{u}) = -\Delta u + \Delta \tilde{u} = f - f = 0$ ist w harmonisch, und zwar mit Randwerten $g - \tilde{g}$.

Dank Maximumsprinzips gilt

$$\max_{x \in \bar{\Omega}} w(x) = \max_{x \in \partial\Omega} w(x) = \max_{x \in \partial\Omega} g(x) - \tilde{g}(x) \leq \max_{x \in \partial\Omega} |g(x) - \tilde{g}(x)|$$

und analog für $-w = \tilde{u} - u$, und somit gilt die Behauptung. \square

Folgerung 2.30. (Stetige Abhängigkeit von rechter Seite)

Sei $\Omega \subseteq \mathbb{R}^d$ offen und beschränkt.

Setze $R := \sup_{x \in \Omega} \|x\|$ sowie $C := R^2/2$.

Seien $u, \tilde{u} \in C^2(\Omega) \cap C^0(\bar{\Omega})$ klassische Lösungen von Poisson-RWP zu identischen Randwerten $g \in C^0(\partial\Omega)$, aber verschiedenen $f, \tilde{f} \in C^0(\Omega)$.

Dann gilt:

$$\|u - \tilde{u}\|_{\infty} \leq C \|f - \tilde{f}\|_{\infty}.$$

Beweis. Setze $w: \Omega \rightarrow \mathbb{R}, x \mapsto R^2 - \|x\|^2$.

Da Ω offen ist, gilt $0 \leq w < R^2$.

Zudem gilt $\partial_{x_i} \partial_{x_j} w(x) = -2\delta_{ij}$ und somit $-\Delta w = 2d \geq 2$.

Setze nun $\bar{u} := u - \tilde{u}$ und $\bar{f} := f - \tilde{f}$, dann ist $\bar{u} \in C^2(\Omega) \cap C^0(\bar{\Omega})$ mit

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

Setze noch $\bar{w} := \frac{1}{2} \|\bar{f}\|_{\infty} w$, dann gilt $\bar{w} \geq 0 = \pm \bar{u}$ auf $\partial\Omega$ sowie

$$-\Delta \bar{w} = \frac{1}{2} \|\bar{f}\|_{\infty} 2d = \|\bar{f}\|_{\infty} d \geq \|\bar{f}\|_{\infty} = \|-\Delta \bar{u}\|_{\infty} \geq -\Delta(\pm \bar{u}).$$

Das Vergleichsprinzip liefert dann $\bar{w} \geq \pm \bar{u}$ und somit

$$\|\bar{u}\|_{\infty} \leq \|\bar{w}\|_{\infty} = \frac{1}{2} \|\bar{f}\|_{\infty} \|w\|_{\infty} \leq \frac{1}{2} \|\bar{f}\|_{\infty} R^2 = \frac{1}{2} \|f - \tilde{f}\|_{\infty} R^2 = C \|f - \tilde{f}\|_{\infty}. \quad \square$$

Satz 2.31. (C^∞ -Regelarität)

Sei $\Omega = \mathbb{R}^d$ und $u \in C^2(\Omega)$ harmonisch. Dann ist $u \in C^\infty(\Omega)$.

Beweis.

Siehe Blatt 7 Aufg 2. □

2.3. FINITE DIFFERENZEN FÜR POISSON-GLEICHUNG

Wir beenden den Theorieteil und kommen nun zu numerischen Verfahren, nämlich das Finite-Differenzen-Verfahren (FD-Verfahren):

Definition 2.32. (Finite Differenzen)

Sei $h \in \mathbb{R}_+$, $e_1, \dots, e_d \in \mathbb{R}^d$ die Einheitsvektoren und $x \in \mathbb{R}^d$.

Sei dazu eine Funktionen gegeben

$$u: \{x, x \pm he_j \mid j \in \{1, \dots, d\}\} \rightarrow \mathbb{R}.$$

Dann definieren wir die Vorwärtsdifferenz (rechtseitige Differenz)

$$\partial_{x_j}^{+h} u(x) := \frac{u(x + he_j) - u(x)}{h},$$

die Rückwärtsdifferenz (rückseitige Differenz)

$$\partial_{x_j}^{-h} u(x) := \frac{u(x) - u(x - he_j)}{h}$$

sowie die symmetrische oder zentrale Differenz

$$\partial_{x_j}^{c,h} u(x) := \frac{u(x + h e_j) - u(x - h e_j)}{2h} = \frac{1}{2} (\partial_{x_j}^{+h} u + \partial_{x_j}^{-h} u)(x).$$

Satz 2.33. (Approximationsgüte von Finiten Differenzen)

Sei $u: \Omega \rightarrow \mathbb{R}$, $x \in \Omega \subseteq \mathbb{R}^d$, $r \in \mathbb{R}_+$ mit $B_r(x) \subseteq \Omega$. Für jedes $h \in \mathbb{R}_{<r}$ gilt dann

i. für jedes $u \in C^2(\Omega)$:

$$|\partial_{x_j} u(x) - \partial_{x_j}^{\pm h} u(x)| \leq \frac{h}{2} \|\partial_{x_j}^2 u\|_\infty.$$

ii. für jedes $u \in C^3(\Omega)$:

$$|\partial_{x_j} u(x) - \partial_{x_j}^{c,h} u(x)| \leq \frac{h^2}{6} \|\partial_{x_j}^3 u\|_\infty.$$

iii. für jedes $u \in C^4(\Omega)$:

$$|\partial_{x_j}^2 u(x) - \partial_{x_j}^{-h} \partial_{x_j}^{+h} u(x)| \leq \frac{h^2}{12} \|\partial_{x_j}^4 u\|_\infty.$$

Beweis. Wir beobachten, dass die Aussage für alle $d \in \mathbb{N}$ gilt, sobald sie für $d=1$ gilt, denn für $u \in C^k(\bar{\Omega})$, $j \in \{1, \dots, d\}$ und $v_j(t) := u(x + t e_j)$ ist $v_j \in C^k([r, -r])$ und es folgt, z.B. i) durch

$$|\partial_{x_j} u(x) - \partial_{x_j}^{\pm h} u(x)| = |\partial_t v_j(0) - \partial_t^{\pm h} v_j(0)| \leq \frac{h}{2} \|\partial_t^2 v_j\|_{C^0([-r, r])} \leq \frac{h}{2} \|\partial_{x_j}^2 u\|_{C^0(\bar{\Omega})}$$

also reicht es, die Aussage für $d=1$ zu zeigen:

i. Taylor-Entwicklung an x liefert:

$$u(x+h) = u(x) + h u'(x) + \frac{h^2}{2} u''(\xi) \quad \text{mit } \xi \in (x, x+h)$$

also

$$|u'(x) - \partial_x^{+h} u(x)| = \left| u'(x) - \frac{u(x+h) - u(x)}{h} \right| = \left| -\frac{h}{2} u''(\xi) \right| \leq \frac{h}{2} \|u''\|_\infty$$

und analog für ∂_x^{-h} .

ii. Subtraktion von Taylor-Entwicklungen an x

$$\begin{aligned} u(x+h) &= u(x) + h u'(x) + \frac{h^2}{2} u''(x) + \frac{h^3}{6} u^{(3)}(\xi) \quad \text{mit } \xi \in (x, x+h) \\ u(x-h) &= u(x) - h u'(x) + \frac{h^2}{2} u''(x) - \frac{h^3}{6} u^{(3)}(\zeta) \quad \text{mit } \zeta \in (x-h, x) \end{aligned}$$

liefert

$$u(x+h) - u(x-h) = 2h u'(x) + \frac{h^3}{6} (u^{(3)}(\xi) + u^{(3)}(\zeta))$$

also

$$\left| u'(x) - \frac{u(x+h) - u(x-h)}{2h} \right| = \left| \frac{h^2}{12} (u^{(3)}(\xi) + u^{(3)}(\zeta)) \right| \leq \frac{h^2}{12} 2 \|u^{(3)}\|_\infty.$$

iii. Addition von

$$\begin{aligned} u(x+h) &= u(x) + h u'(x) + \frac{h^2}{2} u''(x) + \frac{h^3}{6} u^{(3)}(x) + \frac{h^4}{24} u^{(4)}(\xi) \quad \text{mit } \xi \in (x, x+h) \\ -2u(x) &= -2u(x) \\ u(x-h) &= u(x) - h u'(x) + \frac{h^2}{2} u''(x) - \frac{h^3}{6} u^{(3)}(x) + \frac{h^4}{24} u^{(4)}(\zeta) \quad \text{mit } \zeta \in (x-h, x) \end{aligned}$$

liefert

$$\begin{aligned} \partial_x^{-h} \partial_x^{+h} u(x) &= \frac{1}{h} \left(\frac{u(x+h) - u(x)}{h} - \frac{u(x) - u(x-h)}{h} \right) \\ &= \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} \\ &= \frac{1}{h^2} \left(\frac{h^2}{2} + \frac{h^2}{2} \right) u''(x) + \frac{1}{h^2} \frac{h^4}{24} (u^{(4)}(\xi) - u^{(4)}(\zeta)) \\ &\leqslant u''(x) + \frac{h^2}{24} (2\|u^{(4)}\|_\infty) \end{aligned}$$

also folgt iii. sofort daraus. \square

Bemerkung.

- Die Approximation $\partial_{x_j}^{-h} \partial_{x_j}^{+h} u(x)$ in iii. ist also „zweite zentrale Differenz“ und dabei gilt

$$\partial_{x_j}^{-h} \partial_{x_j}^{+h} u(x) = \partial_{x_j}^{+h} \partial_{x_j}^{-h} u(x) = \partial_{x_j}^{c, \frac{h}{2}} \partial_{x_j}^{c, \frac{h}{2}} u(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}.$$

- Aus dem Beweis folgt, dass

$$u^{(4)} \equiv 0 \Rightarrow \partial_{x_j}^{-h} \partial_{x_j}^{+h} u = u''$$

z.B. im Fall $u \in \mathbb{P}_3$.

D.h. die zweite zentrale Differenz ist exakt auf \mathbb{P}_3 .

- Man kann zentrale Differenzen für höhere Ableitungen verallgemeinern

$$\forall m \in \mathbb{N}: \partial_{x_j}^{h,m} u(x) := \left(\partial_{x_j}^{c, \frac{h}{2}} \right)^m u(x).$$

Falls $u: \{x + (k - \frac{m}{2})h e_j \mid k \in \{0, \dots, m\}\} \rightarrow \mathbb{R}$, dann ist

$$\partial_{x_j}^{h,m} u(x) = \frac{1}{h^m} \sum_{k=0}^m \binom{m}{k} (-1)^{k+m} u\left(x + \left(k - \frac{m}{2}\right)h e_j\right).$$

Folgerung 2.34. (FD-Approximation für Laplace)

Sei $u: \{x, x \pm h e_j\} \rightarrow \mathbb{R}$.

Dann definieren wir

$$\Delta_h u(x) := \sum_{j=1}^d \partial_{x_j}^{-h} \partial_{x_j}^{+h} u(x) \tag{2.6}$$

und es gilt unter den Voraussetzung von Satz 2.33 für $u \in C^4(\bar{\Omega})$

$$|\Delta u(x) - \Delta_h u(x)| \leq \frac{d}{12} \|u\|_{C^4(\bar{\Omega})} h^2.$$

Beweis. Es ist

$$\begin{aligned} |\Delta u(x) - \Delta_h u(x)| &= \left| \sum_{j=1}^d \partial_{x_j}^2 u(x) - \partial_{x_j}^{-h} \partial_{x_j}^{+h} u(x) \right| \\ &\leq \sum_{j=1}^d |\partial_{x_j}^2 u(x) - \partial_{x_j}^{-h} \partial_{x_j}^{+h} u(x)| \\ &\leq \sum_{j=1}^d \frac{h^2}{12} \|\partial_{x_j}^4 u\|_\infty \\ &\leq \frac{d}{12} h^2 \|u\|_{C^4(\bar{\Omega})} \end{aligned}$$

wobei die dritte Zeile aus 2.33 iii. folgt. \square

Bemerkung.

Für $p(x) = \prod_{i=1}^d p_i(x_i)$ für $p_1, \dots, p_d \in \mathbb{P}_3$ ist Δ_h exakt an p , d.h. $\Delta_h p(x) = \Delta p(x)$.

Definition 2.35. (Würfelgebiet)

Sei $w := [0, 1]^d$ der Einheitswürfel und zu $\delta \in \mathbb{R}_+$, $x \in \mathbb{R}^d$ setzen wir

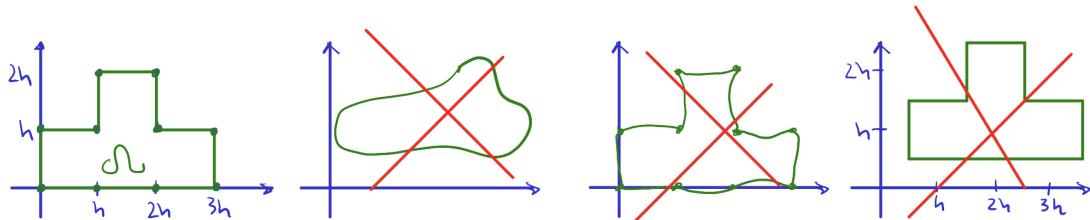
$$w_\delta(x) := \delta(x + w) = [x_1 \delta, (x_1 + 1)\delta] \times \cdots \times [x_d \delta, (x_d + 1)\delta].$$

Sei $\Omega \subseteq \mathbb{R}^d$ offen und beschränkt.

Ω heißt ein Würfelgebiet, g.d.w. Ω das Innere einer Vereinigung von „kleinen Würfeln“ ist, also $\exists h \in \mathbb{R}_+ \exists Z \subseteq \mathbb{Z}^d$: $\Omega = W^\circ$ mit

$$W := \bigcup_{z \in Z} w_h(z).$$

Beispiel.



Nur die erste Menge ist ein Würfelgebiet zu Wurfelseitenlänge h , aber die vierte Menge ist ein Würfelgebiet zu Wurfelseitenlänge $\frac{h}{2}$.

Man sieht leicht, dass ein Würfelgebiet zu $h \in \mathbb{R}_+$ auch ein Würfelgebiet zu h/n für jedes $n \in \mathbb{N}$ ist.

Definition 2.36. (FD-Gitter für Würfelgebiet)

Sei $\Omega \subseteq \mathbb{R}^d$ ein Würfelgebiet zu $h \in \mathbb{R}^d$, $\Gamma := \partial\Omega$ also $\bar{\Omega} := \Omega \cup \Gamma$.

Wir definieren das Gitter $\bar{\Omega}_h$ durch

$$\begin{aligned}\Omega_h &:= \{x \in \Omega \mid \exists z \in \mathbb{Z}^d: x = hz\} \\ \Gamma_h &:= \{x \in \Gamma \mid \exists z \in \mathbb{Z}^d: x = hz\} \\ \bar{\Omega}_h &:= \Omega_h \cup \Gamma_h.\end{aligned}$$

Bemerkung.

- Jeder Punkt in Ω_h hat genau $2d$ direkte Nachbaren mit Abstand h in $\bar{\Omega}_h$.
- Erweiterung für allgemeine Gebiete folgt später.

Definition 2.37. (Gitterfunktion)

Zu einem Gitter $\bar{\Omega}_h$ definieren wir den Raum der Gitterfunktionen

$$X_h := \{v: \bar{\Omega}_h \rightarrow \mathbb{R}\},$$

davon den Teilraum der Funktionen mit Nullrandwerten

$$X_h^\circ := \{v \in X_h \mid \forall x \in \Gamma_h: v(x) = 0\} \subseteq X_h,$$

und außerdem den Raum der Funktionen der inneren Punkte

$$Y_h := \{v: \Omega_h \rightarrow \mathbb{R}\}$$

jeweils versehen mit Maximumsnormen

$$\|v\|_{\bar{\Omega}_h} := \max_{x \in \bar{\Omega}_h} |v(x)| \quad \text{bzw.} \quad \|v\|_{\Omega_h} := \max_{x \in \Omega_h} |v(x)|.$$

Bemerkung. (Zusätzliche Raumstruktur)

- $(X_h, \|\bullet\|_{\bar{\Omega}_h})$, $(X_h^\circ, \|\bullet\|_{\bar{\Omega}_h})$, $(X_h, \|\bullet\|_{\Omega_h})$ sowie $(Y_h, \|\bullet\|_{\Omega_h})$ sind Banachräume, da endlich dimensional und damit vollständig.
- Man kann X_h auch mit einer Hilbertraumstruktur versehen, indem man das „diskrete l^2 -Skalarprodukt“ definiert, also

$$\langle u, v \rangle_{l^2} := h^d \sum_{x \in \bar{\Omega}_h} u(x)v(x)$$

welches die Norm $\|u\|_{l^2} := \sqrt{\langle u, u \rangle_{l^2}} = \sqrt{h^d \sum_{x \in \bar{\Omega}_h} u(x)u(x)}$ induziert.

- Bzgl. $\|\bullet\|_{l^2}$ ist X_h ebenfalls vollständig, und es gilt sogar

$$\forall u \in C^0(\bar{\Omega}): \quad \lim_{h \rightarrow 0} \|u\|_{l^2} = \|u\|_{L^2}.$$

- Man kann X_h auch mit einer Seminorm versehen, welche die Ableitungen mit einbezieht („diskrete H^1 -Seminorm“)

$$|u|_{h^1} := \sqrt{h^d \sum_{x \in \Omega_h} \sum_{j=1}^d (\partial_{x_j}^{+h} u(x))^2}.$$

Dies ist nur eine Seminorm auf X_h aber eine Norm auf X_h^0 . Damit erhält durch Kombination mit l^2 -Norm eine Norm auf X_h („diskrete H^1 -Norm“)

$$\|u\|_{h^1} := \sqrt{\|u\|_{l^2}^2 + |u|_{h^1}^2}$$

bzgl. welcher X_h auch Hilbertraum ist. Details dazu kommen im Kapitel 3.

Wir bemerken hier noch: Für $v \in X_h$ und $x \in \Omega_h$ ist mit (2.6) $(-\Delta_h v)(x)$ wohldefiniert, d.h. wir können $\Delta_h: X_h \rightarrow Y_h$ als linearen Operator interpretieren.

Definition 2.38. (FD-Approximation für Poisson-RWP)

Sei ein Gitter $\bar{\Omega}_h$ gegeben.

Dann nennen wir $u_h \in X_h$ eine FD-Approximation des Poisson-RWP aus Korollar 2.28, g.d.w. es gilt

$$\begin{aligned} \forall x \in \Omega_h: \quad -\Delta_h u_h(x) &= f(x), \\ \forall x \in \Gamma_h: \quad u_h(x) &= g(x). \end{aligned} \tag{2.7}$$

Bemerkung. (Berechnung via LGS)

- Sei eine Aufzählung $\{x_1, \dots, x_n\} = \Omega_h$ gegeben. Dann ist (2.7) äquivalent zu LGS für unbekannte $\vec{u}_h := \{u_i\}_{i=1}^n$ mit $u_i := u_h(x_i)$ für $i \in \{1, \dots, n\}$, denn für $x \in \Gamma_h$ ist $u_h(x)$ schon durch $g(x)$ festgelegt.
- Sei FD-Operator in $x_i \in \Omega_h$ gegeben durch

$$-\Delta_h u_h(x_i) = \sum_{j=1}^n \alpha_{ij} u(x_j) + \sum_{x \in \Gamma_h} \beta_{i,x} u(x)$$

wobei $\sum_{j=1}^n \alpha_{ij} u(x_j)$ den Anteil für die Punkte in Ω_h ist und $\sum_{x \in \Gamma_h} \beta_{i,x} u(x)$ für den Rand.

Dann ist ein LGS gegeben durch

$$A_h \vec{u}_h = b_h$$

mit $A_h = (\alpha_{ij})_{i,j=1}^n$ und $b_h = (f(x_i) - \sum_{x \in \Gamma_h} \beta_{i,x} u(x))_{i=1}^n$.

- Beachte: A_h ist dünn besetzt (sparse), da es nur sehr wenige Nichtnull-Einträge pro Zeile gibt, daher in Praxis Sparse-Matrix-Datenstruktur verwendet wird.

Beispiel 2.39. (FD-Approximation für Poisson-RWP in 1D)

Sei $d := 1$, $\Omega := (0, 1)$, $n \in \mathbb{N}$, $h := \frac{1}{n+1}$, $x_i := ih$ für $i \in \{0, \dots, n+1\}$, dann ist

$$\Omega_h := \{x_1, \dots, x_n\}, \quad \Gamma_h := \{x_0, x_{n+1}\}.$$

Poisson-Problem:

$$\begin{aligned} -u''(x) &= f(x) \quad \text{in } \Omega, \\ u(0) &= \alpha, \quad u(1) = \beta. \end{aligned}$$

Sei $u_i \approx u(x_i)$ für $i \in \{1, \dots, n\}$, $u_0 := \alpha$ und $u_{n+1} := \beta$.

Zugehörige Diskretisierung

$$\forall i \in \{1, \dots, n\}: \quad -\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} = f(x_i)$$

und somit LGS

$$\underbrace{\frac{1}{h^2} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & \ddots & & \\ & \ddots & \ddots & -1 & \\ & & -1 & 2 & \end{pmatrix}}_{A_h} \underbrace{\begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}}_{\vec{u}_h} = \underbrace{\begin{pmatrix} f(x_1) + \frac{\alpha}{h^2} \\ f(x_2) \\ \vdots \\ f(x_{n-1}) \\ f(x_n) + \frac{\beta}{h^2} \end{pmatrix}}_{b_h}.$$

Bemerkung.

- \tilde{A}_n und A_h sind tridiagonal und symmetrisch und klar ist $A_h := \frac{1}{h^2} \tilde{A}_n$.
- A_h ist regulär, denn per Induktion folgt $\det \tilde{A}_n = n + 1$:
Für $n = 1$ ist $\det(\tilde{A}_n) = \det(2) = n + 1$.

Im Induktionsschritt $n \rightarrow n + 1$ gilt mit Laplace-Entwicklung:

$$\begin{aligned} \det(\tilde{A}_{n+1}) &= \det \begin{pmatrix} & & 0 & & \\ & \tilde{A}_n & \vdots & & \\ & 0 & & -1 & \\ & 0 \cdots 0 & -1 & 2 & \end{pmatrix} \\ &= 2\det(\tilde{A}_n) - (-1)\det \begin{pmatrix} & & 0 & & \\ & \tilde{A}_{n-1} & \vdots & & \\ & 0 & & -1 & \\ & 0 \cdots 0 & -1 & -1 & \end{pmatrix} \\ &= 2\det(\tilde{A}_n) + (-1)\det(\tilde{A}_{n-1}) \\ &= 2n + (-1)(n-1). \end{aligned}$$

- A_h ist positiv definit, denn mit Satz von Gershgorin ist $\lambda_i(A_h) \in [0, 4]$ und wegen Regularität von A_h ist sogar $\lambda_i(A_h) \in (0, 4]$.
- Damit existiert eindeutige FD-Approximation u_h des Poisson-RWP in 1D.
- Wegen A_h positiv definit & symmetrisch kann CG oder PCG als iterative LGS-Löser verwendet werden.

- Für kleine h ist A_h und \tilde{A}_n schlecht konditioniert, also ist Vorkonditionierung zu empfehlen.
- Man kann für das diskrete Problem auch „stetige Abhängigkeit von den Daten“ beweisen (siehe Übung).

Beispiel 2.40. (FD-Approximation für Poisson-RWP in 2D)

Sei $d := 2$, $\Omega := (0, 1)^2$, $m \in \mathbb{N}$, $h := \frac{1}{m+1}$, $n := m^2$ und ein Poisson-RWP mit $g = 0$ gegeben.

Statt Einzelindex nutzen wir zur besseren Übersicht Doppelindex, also:

$$\forall i, j \in \{0, \dots, m+1\}: \quad x_{ij} := (ih, jh), \quad u_{i,j} \approx u(x_{ij}).$$

Diekretisierung des RWP für $\forall i, j \in \{0, \dots, m+1\}$:

$$\begin{aligned} \frac{1}{h^2}(4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}) &= f(ih, jh) \quad \text{für innere Punkte,} \\ u_{0,j} = u_{m+1,j} = u_{i,0} = u_{i,m+1} &= 0 \quad \text{für Randpunkte.} \end{aligned}$$

Für jede beliebige Wahl von Aufzählung der $u_{i,j}$ erhält man unter Anpassung der Umordnung ein System von LGS

$$A_h \vec{u}_h = b_h$$

wobei A_h symmetrisch ist und $\frac{4}{h^2}$ auf Diagonalen sowie $-\frac{1}{h^2}$ an bis zu 4 Stellen pro Zeilen und pro Spalte.

Mit lexikographischer Aufzählung, also $\vec{u}_h = (u_{11}, u_{12}, \dots, u_{1m}, u_{21}, \dots, u_{mm})$, bekommt A_h eine „Bandstruktur“, also „block-tridiagonal“, d.h.

$$A_h = \frac{1}{h^2} \begin{pmatrix} B & C & & & \\ C & B & \ddots & & \\ & \ddots & \ddots & C & \\ & & & C & B \end{pmatrix} \in \mathbb{R}^{m^2 \times m^2}$$

wobei

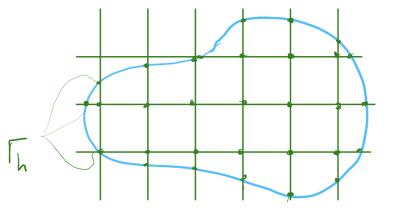
$$B = \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & \ddots & & \\ & \ddots & \ddots & -1 & \\ & & -1 & 4 & \end{pmatrix} \in \mathbb{R}^{m \times m}, \quad C = -I_m = \begin{pmatrix} -1 & & 0 \\ & \ddots & \\ 0 & & -1 \end{pmatrix} \in \mathbb{R}^{m \times m}.$$

Bemerkung. (Erweiterung auf beliebige Gebiete)

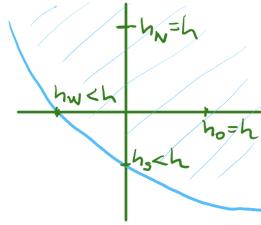
- Falls Ω beschränkt aber kein Würfelgebiet ist, muss das Gitter modifiziert werden, indem Schnittpunkte von Γ mit den von $\mathbb{Z}^d \cdot h$ erzeugten Würfekanten hinzugenommen werden:

$$\Gamma_h := \{x \in \Gamma \mid \exists j \in \{1, \dots, d\} \exists z \in \mathbb{Z}^d: x \in h(z + \mathbb{R}e_j)\}$$

dann wie gehabt $\bar{\Omega}_h := \Omega_h \cup \Gamma_h$ mit Ω_h aus Def. 2.36.



- Koeffizienten der FD-Diskretisierung werden angepasst.



Taylor-Entwicklung liefert:

1D:

$$\begin{aligned} u''(x) &= \frac{2}{h_W(h_O + h_W)} u(x - h_W) - \frac{2}{h_W h_O} u(x) \\ &\quad + \frac{2}{h_O(h_O + h_W)} u(x + h_O) + \mathcal{O}(h). \end{aligned}$$

2D:

$$\begin{aligned} \Delta u(x) &= \frac{2}{h_W(h_O + h_W)} u(x - e_1 h_W) + \frac{2}{h_O(h_O + h_W)} u(x + e_1 h_O) \\ &\quad + \frac{2}{h_S(h_S + h_N)} u(x - e_2 h_S) + \frac{2}{h_N(h_S + h_N)} u(x + e_2 h_N) \\ &\quad - \left(\frac{2}{h_O h_W} + \frac{2}{h_S h_N} \right) u(x) + \mathcal{O}(h). \end{aligned}$$

Dies ist die Shortley-Weller-Approximation.

- Man verliert also eine h -Potenz in der Approximationsgüte.
- Systemmatrix wird i.A. nicht symmetrisch.

Bemerkung. (Andere Randbedingungen)

Neben Dirichlet auch andere Randbedingungen möglich, z.B. Neumann-Randbedingung:

$$\forall x \in \Gamma_N \subseteq \Gamma \text{ wobei } \Gamma_N \text{ Neumann-Randteil: } (\nabla u(x)) \bullet n = g_N(x).$$

Wir nehmen zusätzlich an, dass x auf Kante und nicht auf Ecke des Würfelgebiets liegt, sonst ist keine Normale $n(x)$ definiert.

Hier fehlt eine Skizze....

Sei $n = \pm e_j$ äußerer Normalenvektor für ein $j \in \{1, \dots, d\}$.

Wir approximieren ∇u durch

$$(\nabla_h u)_i := \begin{cases} \partial_{x_i}^{c,h} u & \text{für } i \neq j \text{ (zentrale Diff.)} \\ \partial_{x_j}^{-h} u & \text{für } i = j \text{ (Rückwärts-Diff.)} \end{cases}$$

Für $x \in \Gamma_h \cap \Gamma_N$ ist nun $u_h(x)$ auch Unbekannt, denn man kennt nicht seinen Funktionswert, nur die Ableitung.

Also neue Gleichung für LGS:

$$\forall x \in \Gamma_h \cap \Gamma_N: (\nabla_h u_h(x)) \bullet n = g_N(x).$$

2.4. FD FÜR ALLGEMEINE ELLIPTISCHE PDEs 2. ORDNUNG

Definition 2.41. (Allgemeine Elliptisches RWP)

Zu einem beschränkten Gebiet $\Omega \subseteq \mathbb{R}^d$, einer $f \in C^0(\Omega)$ und einer $g \in C^0(\bar{\Omega})$ sei

$$(\mathcal{L}u)(x) = - \sum_{i,j=1}^d a_{ij}(x) \partial_{x_i} \partial_{x_j} u(x) + \sum_{i=1}^d b_i(x) \partial_{x_i} u(x) + c(x) u(x) \quad (2.8)$$

gleichmäßig elliptisch und $a_{ij}, b_i, c \in C^0(\bar{\Omega})$.

Gesucht ist $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$ mit den Eigenschaften

$$\begin{aligned} (\mathcal{L}u)(x) &= f(x) && \text{für } x \in \Omega, \\ u(x) &= g(x) && \text{für } x \in \Gamma. \end{aligned}$$

Definition 2.42. (FD-Approximation)

Für \mathcal{L} aus (2.8), $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$, $h \in \mathbb{R}_+$, $x \in \Omega$ mit $\overline{B_h(x)} \subseteq \bar{\Omega}$ definieren wir

$$\begin{aligned} (\mathcal{L}_h u)(x) &:= - \sum_{i=1}^d a_{ii}(x) \partial_{x_i}^{-h} \partial_{x_i}^{+h} u(x) - \sum_{i,j=1}^d a_{ij}(x) \partial_{x_i}^{c,h} \partial_{x_j}^{c,h} u(x) \\ &\quad + \sum_{i=1}^d b_i(x) \partial_{x_i}^{c,h} u(x) + c(x) u(x). \end{aligned}$$

Bemerkung. Wir werden sehen, dass

- Def. 2.42 nur unter zusätzlichen Annahmen an a_{ij}, b_i, c, h eine „stabile“ Diskretisierung ergibt;
- eine etwas sorgfältigere Diskretisierung des Hauptteils eine erweiterte Klasse von Funktionen a_{ij}, b_i, c „stabil“ diskretisiert.

Satz 2.43. (FD-Approximationsfehler für \mathcal{L}_h)

Sei $u \in C^4(\Omega)$, $x \in \Omega$ s.d. für $i \in \{1, \dots, d\}$ und $\sigma_i \in \{0, 1, -1\}$ gilt $x + \sum_{i=1}^d \sigma_i e_i h$.

Dann existiert ein C (unabhängig von x und h), s.d.

$$|(\mathcal{L}u)(x) - (\mathcal{L}_h u)(x)| \leq C \cdot h^2.$$

Beweis. Mittels Taylor, analog zu 2.33 & 2.34. \square

Bemerkung. (FD-Stern)

- Die Diskretisierung eines PDE-Operators \mathcal{L} kann man anschaulicher notieren, z.B. für $d=2$ und um (x_1, x_2) herum:

Falls $(\mathcal{L}_h u)(x_1, x_2) = \frac{1}{h^2} \sum_{i,j=-m}^m \alpha_{ij} u(x_1 + ih, x_2 + jh)$ für ein $m \in \mathbb{N}$ und geeignete Koeffizienten α_{ij} , so ist

$$\begin{pmatrix} \alpha_{-m,m} & \cdots & \alpha_{0,m} & \cdots & \alpha_{m,m} \\ \vdots & & \vdots & & \vdots \\ \alpha_{-m,0} & \cdots & \alpha_{0,0} & \cdots & \alpha_{m,0} \\ \vdots & & \vdots & & \vdots \\ \alpha_{-m,-m} & \cdots & \alpha_{0,-m} & \cdots & \alpha_{m,-m} \end{pmatrix}_* \quad (2.9)$$

der zugehörige FD-Stern.

- Für $\mathcal{L}_h u = -\Delta_h$ aus Fol. 2.34 ergibt sich also für $m=1$

$$\begin{pmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{pmatrix}_*$$

der 5-Punkte-Differenzen-Stern.

- Für \mathcal{L}_h aus Def. 2.42 ergibt sich der FD-Stern als (oBdA $a_{12} = a_{21}$)

$$\begin{aligned} & \frac{1}{2} \begin{pmatrix} a_{12}(x) & -2a_{22}(x) & -a_{12}(x) \\ -2a_{11}(x) & 4(a_{11}(x) + a_{22}(x)) & -2a_{11}(x) \\ -a_{12}(x) & -2a_{22}(x) & a_{12}(x) \end{pmatrix}_* \\ & + \frac{h}{2} \begin{pmatrix} 0 & b_2(x) & 0 \\ -b_1(x) & 0 & b_1(x) \\ 0 & -b_2(x) & 0 \end{pmatrix}_* + h^2 \begin{pmatrix} 0 & 0 & 0 \\ 0 & c(x) & 0 \\ 0 & 0 & 0 \end{pmatrix}_* \end{aligned} \quad (2.10)$$

- Für $m=1$ (also 3×3 Sterne) ist höchstens Approximationsordnung 2 erreichbar, wie in 2.43 & 2.34 für spezielle \mathcal{L}_h realisiert.
- Für $m > 1$ sind bessere Approximationsordnungen erreichbar.

Definition 2.44. (FD-Approximation für elliptisches RWP)

Sei Ω Würfelgebiet zu $h \in \mathbb{R}_+$ und $\bar{\Omega}_h$ das zugehörige Gitter.

Dann ist $u_h \in X_h$ FD-Approximation des elliptischen RWP, g.d.w. es gilt

$$\begin{aligned} (\mathcal{L}_h u_h)(x) &= f(x) \quad \text{für } x \in \Omega_h, \\ u_h(x) &= g(x) \quad \text{für } x \in \Gamma_h. \end{aligned}$$

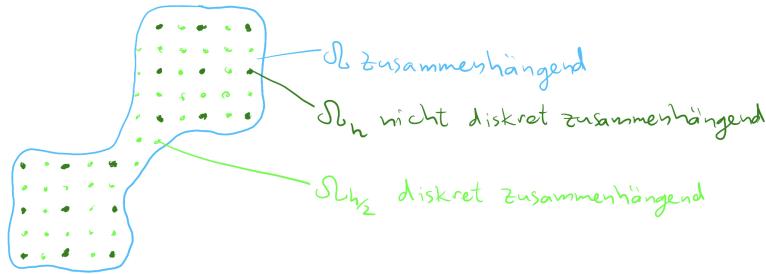
Definition 2.45. (Diskreter Zusammenhang)

Ein Gitter $\bar{\Omega}_h$ heißt diskrete zusammenhängend, g.d.w. es für alle $x, y \in \Omega_h$ (also innere Punkte) eine Punktfolge $z_1, \dots, z_k \in \Omega_h$ s.d.

$$z_0 = x, \quad z_k = y, \quad |z_2 - z_1| = \dots = |z_k - z_{k-1}| = h.$$

Beispiel.

Falls Ω zusammenhängend, ist für genügend kleines h auch $\bar{\Omega}_h$ diskret zusammenhängend, z.B.



Bei der Untersuchung von FD-Stern hilft uns folgendes technisches Lemma:

Lemma 2.46. (Sternlemma)

Sei $k \in \mathbb{N}$, $\{\alpha_i\}_{i=0}^k \subseteq \mathbb{R}$ und $\{p_i\}_{i=0}^k \subseteq \mathbb{R}$ gegeben s.d.

$$\alpha_0 > 0, \quad \alpha_1, \dots, \alpha_k < 0, \quad \sum_{i=0}^k \alpha_i = 0, \quad \sum_{i=0}^k \alpha_i p_i \leq 0, \quad p_0 = \max_{i \in \{1, \dots, k\}} p_i.$$

Dann folgt $p_0 = p_1 = \dots = p_k$.

Bei Anwendung des Lemma später werden die α_i die FD-Koeffizienten sein, und p_i die Lösungswerte in den Punkten. Das Lemma dient mehr oder weniger als ein Kriterium für die Konstantheit der Lösungswerten.

Beweis.

Mit $p_0 = \max_{i \in \{1, \dots, k\}} p_i$ und $\alpha_1, \dots, \alpha_k < 0$ ist $\alpha_i(p_i - p_0) \geq 0$ für jedes $i \in \{1, \dots, k\}$, und somit

$$0 \leq \sum_{i=0}^k \alpha_i(p_i - p_0) = \underbrace{\sum_{i=0}^k \alpha_i p_i}_{\leq 0} - p_0 \underbrace{\sum_{i=0}^k \alpha_i}_{= 0} \leq 0$$

also ist $\sum_{i=0}^k \alpha_i(p_i - p_0) = 0$, und da alle α_i von 0 verschieden sind, müssen alle p_i gleich sein. \square

Mit diesem Lemma kann für gewisse \mathcal{L}_h ein diskretes Analogon des Maximumsprinzips 2.27 gezeigt werden.

Dies impliziert dann analoge Aussagen wie im Kontinuierlichen.

Satz 2.47. (Diskretes Maximumsprinzip)

Sei $u_h \in X_h$ FD-Approximation des RWP 2.44 mit $f(x) \leq 0$ für $\forall x \in \Omega_h$.

Der Differenzenstern (2.9) zu $m=1$ (also 3×3 Stern) erfülle in allen Punkten

- $$\begin{array}{ll} i. \quad \sum_{i,j=-1}^1 \alpha_{ij} = 0 & iii. \quad \forall (i,j) \neq (0,0): \alpha_{ij} \leq 0 \\ ii. \quad \alpha_{00} > 0 & iv. \quad \alpha_{1,0} < 0, \quad \alpha_{0,1} < 0, \quad \alpha_{-1,0} < 0, \quad \alpha_{0,-1} < 0. \end{array}$$

Dann gilt

$$\max_{x \in \bar{\Omega}_h} u_h(x) = \max_{x \in \Gamma_h} u_h(x).$$

Beweis. Sei $x \in \bar{\Omega}_h$ mit $u_h(x) = \max_{x \in \Gamma_h} u_h(x)$.

Falls $x \in \Gamma_h$, ist die Behauptung klar.

Falls $x \in \Omega_h$, setze $p_0 := u_h(x)$, $(p_i)_{i=1}^k$ als Nachbarwerte von $u_h(x)$ und $\alpha_0 := \alpha_{00}$ und $(\alpha_i)_{i=1}^k$ als Nicht-Null-Koeffizienten des FD-Sterns.

Es gilt dann

$$\sum_{i=0}^k \alpha_i p_i = (\mathcal{L}_h u_h)(x) = f(x) \leq 0$$

also liefert das Sternlemma: $p_0 = p_1 = \dots = p_k$, d.h. u_h ist konstant auf x und seinen Nachbarn, welche im Differenzenstern auftreten.

Wiederholung dieses Argumentes in alle $2d$ Hauptrichtungen führt zum Rand wegen Beschränktheit von Ω . \square

Falls $\bar{\Omega}_h$ diskret zusammenhängend ist, führt das letzte Argument zu allen Punkten in Ω_h , also folgt:

Folgerung 2.48. (FD-Approximation u_h konstant)

Es gelten die Voraussetzungen von 2.47.

Falls $\bar{\Omega}_h$ diskret zusammenhängend ist und FD-Approximation u_h ihr Maximum im Inneren annimmt, so ist u_h konstant.

Bemerkung.

- Obiges diskretes Maximumsprinzip gilt für $\mathcal{L}_h = -\Delta_h$, weil Voraussetzungen erfüllt sind.
- Für \mathcal{L}_h aus (2.10) mit $a_{12} = 0$, $c = 0$, $b \neq 0$ sind Voraussetzungen von 2.47 für „genügend kleines“ h erfüllt:

Man nutzt die gleichmäßige Elliptizität aus und erhält mit $z = e_i$

$$0 < \alpha \leq \langle z, A(x)z \rangle = a_{ii}(x)$$

also $\alpha_{00} = 2a_{11} + 2a_{22} > 0$.

Falls es ein B gibt s.d. alle $|b_i| \leq B$ und $h \leq \frac{2}{B}\alpha$, dann gilt für $\alpha_{1,0}$

$$\alpha_{1,0} = -a_{11} + \frac{h}{2}b_1 < -\alpha + \frac{2}{B}\alpha \frac{1}{2}B = 0.$$

Analog folgt $\alpha_{-1,0} < 0$, $\alpha_{0,-1} < 0$ und $\alpha_{0,1} < 0$.

$\sum_{i,j=-1}^1 \alpha_{ij} = 0$ ist klar da $c = 0$.

- Falls $c(x) > 0$ in (2,5), so ist $\sum_{i,j=-1}^1 \alpha_{ij} > 0$. Für diesen Fall kann man Abschwächungen des Sternlemmas/diskretes Maximumsprinzip beweisen.
- Falls $a_{1,2} \neq 0$, so ist Modifikation der Diskretisierung notwendig. (Dies wird später betrachtet)

Folgerung 2.49. (Diskretes Vergleichsprinzip)

Seien $u_h, v_h \in X_h$ und es gelte

$$\begin{aligned} \mathcal{L}_h u_h &\leq \mathcal{L}_h v_h && \text{in } \Omega_h \\ u_h &\leq v_h && \text{auf } \Gamma_h \end{aligned}$$

und das diskrete Maximumsprinzip.

Dann gilt

$$u_h \leq v_h \quad \text{auf } \bar{\Omega}_h.$$

Beweis.

Für $w_h := u_h - v_h$ gilt

$$\begin{aligned} \mathcal{L}_h w_h &= \mathcal{L}_h u_h - \mathcal{L}_h v_h \leq 0 && \text{in } \Omega_h \\ w_h &\leq 0 && \text{auf } \Gamma_h. \end{aligned}$$

Aus disrekttem Maximumsprinzip folgt

$$\max_{x \in \Omega_h} w_h(x) = \max_{x \in \Gamma_h} w_h(x) \leq 0$$

also $u_h \leq v_h$ in $\bar{\Omega}_h$. □

Folgerung 2.50. (Existenz & Eindeutigkeit der FD-Approx. für ellip. RWP)

Sei ein diskretisiertes RWP gemäß 2.44 gegeben und es gelte das diskrete Maximumsprinzip.

Dann existiert eine eindeutige FD-Approximation $u_h \in X_h$.

Beweis. Zunächst die Eindeutigkeit:

Seien $u_h, \tilde{u}_h \in X_h$ zwei Lösungen gegeben.

Setze $v := u_h - \tilde{u}_h$, dann ist

$$\begin{aligned} \mathcal{L}_h v &= \mathcal{L}_h u_h - \mathcal{L}_h \tilde{u}_h = f|_{\Omega_h} - f|_{\Omega_h} = 0 && \text{in } \Omega_h \\ v &= u_h - \tilde{u}_h = g|_{\Gamma_h} - g|_{\Gamma_h} = 0 && \text{auf } \Gamma_h. \end{aligned}$$

Das diskrete Maximumsprinzip liefert

$$\forall x \in \bar{\Omega}_h: v(x) \leq \sup_{y \in \Gamma_h} v(y) = 0.$$

Analoge Argumentation für $-v$ liefert:

$$\forall x \in \bar{\Omega}_h: -v(x) \leq \sup_{y \in \Gamma_h} -v(y) = 0$$

also $v = 0$ auf $\bar{\Omega}_h$.

Zur Existenz:

FD-Diskretisierung führt auf $n \times n$ System

$$A_h \underline{u}_h = b_h$$

mit $\ker(A_h) = 0$ wegen Eindeutigkeit, also ist A_h invertierbar und somit ist $\underline{u}_h := A_h^{-1} b_h$ der eindeutiger Koeffizientenvektor für u_h . \square

Folgerung 2.51. (Stetige Abhängigkeit von Randdaten)

Seien $u_h, \tilde{u}_h \in X_h$ FD-Approximation zum RWP 2.44 mit identischem f aber unterschiedlichen Randdaten g, \tilde{g} und es gelte das diskrete Maximumsprinzip.

Dann gilt

$$\|u - \tilde{u}\|_{\bar{\Omega}_h} = \|g - \tilde{g}\|_{\Gamma_h} := \sup_{x \in \Gamma_h} |g(x) - \tilde{g}(x)|.$$

Beweis.

Setze $v := u_h - \tilde{u}_h$, dann ist $\mathcal{L}_h v = 0$ in Ω_h .

Mit dem diskreten Maximumsprinzip folgt für $\forall x \in \bar{\Omega}_h$

$$v \leq \max_{y \in \Gamma_h} v(y) \leq \max_{y \in \Gamma_h} |v(y)| = \max_{y \in \Gamma_h} |g(y) - \tilde{g}(y)|$$

und analog für $-v(x)$, und damit folgt die Behauptung. \square

Bemerkung.

- Anschauliche Bedeutung: Leichte Änderung in Daten ergibt nur leichte Änderung in der Lösung.
- Ähnlich kann man stetige Abhängigkeit von rechter Seite f formulieren.

Definition 2.52. (Stabilität, Konsistenz, Konvergenz)

Sei $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$ Lösung des elliptischen RWP 2.41 und $u_h \in X_h$ die FD-Approximation aus 2.44.

Das FD-Verfahren heißt

- i. konsistent mit Ordnung p , wenn für ein $C_c \in \mathbb{R}$ (abhängig von u aber unabhängig von h) gilt

$$\|\mathcal{L}_h u - \mathcal{L} u\|_{\Omega_h} \leq C_c h^p.$$

ii. stabil, oder genauer (X_h°, Y_h) -stabil, wenn für ein C_s (abhängig von u aber unabhängig von h) gilt

$$\forall v_h \in X_h^\circ: \|v_h\|_{\bar{\Omega}_h} \leq C_s \|\mathcal{L}_h v_h\|_{\Omega_h}.$$

iii. konvergent mit Ordnung p , wenn für ein C (abhängig von u aber unabhängig von h) gilt

$$\|u - u_h\|_{\Omega_h} \leq Ch^p.$$

Bemerkung. (Konsistenz der FD-Approximation)

- Punkweise Fehlerschranken aus 2.34 & 2.43 besagten:

$$|\mathcal{L}_h u(x) - \mathcal{L} u(x)| \leq Ch^2$$

mit C unabhängig von x , also uniforme Schranke bzgl. x falls $u \in C^4(\Omega)$.

Mit $C_c = C$ gilt also Konsistenz der FD-Approximation mit Ordnung 2 falls Ω_h ein Würfelgebiet ist.

- Für Ω ein allgemeines Nicht-Würfelgebiet ist i.A. die Konsistenzordnung der FD-Approximation geringer, z.B. von Ordnung 1 bei Shortley-Weller-Approximation.

Bemerkung. (Stabilität)

Stabilität im Sinne von Def. 2.52 ii) bedeutet anschaulich, dass die FD-Approximation durch die rechte Seite unabhängig von h beschränkt bleibt:

Sei $w_h: \Omega_h \rightarrow \mathbb{R}$ also $w_h \in Y_h$ und $v_h \in X_h^\circ$ Lösung von

$$\begin{aligned} \mathcal{L}_h v_h &= w_h && \text{in } \Omega_h \\ v_h &= 0 && \text{auf } \Gamma_h \end{aligned}$$

dann ist also

$$\|v_h\|_{\bar{\Omega}_h} \leq C_s \|\mathcal{L}_h v_h\| = C_s \|w_h\|_{\Omega_h}.$$

Satz 2.53. (Hinreichende Bedingung für Stabilität)

Sei $A_h \in \mathbb{R}^{n \times n}$ die FD-Systemmatrix.

Falls es eine Konstante C_s unabhängig von h existiert, s.d.

$$\|A_h^{-1}\|_\infty \leq C_s$$

gilt, dann ist das FD-Verfahren stabil.

Beweis.

Seien $v_h, w_h \in \mathbb{R}^n$ Vektoren der inneren Knotenwerte für $v_h \in X_h^\circ$ sowie $w_h \in Y_h$, also $A_h v_h = w_h$, bzw. $\mathcal{L}_h v_h = w_h$.

Dann erhalten wir

$$\|v_h\|_{\bar{\Omega}_h} = \|\underline{v}_h\|_\infty = \|A_h^{-1}\underline{w}_h\|_\infty \leq \|A_h^{-1}\|_\infty \|\underline{w}_h\|_\infty \leq C_s \|w_h\|_{\Omega_h} = C_s \|\mathcal{L}_h v_h\|_{\Omega_h}. \quad \square$$

Satz 2.54. (Stabilität für Poisson-RWP mit FD-Diskretisierung)

Sei $\Omega \subseteq \mathbb{R}^d$ beschränktes Gebiet mit $\Omega \subseteq B_R(0)$ für ein $R \in \mathbb{R}_+$.

Dann gilt für alle $v_h \in X_h^\circ$

$$\|v_h\|_{\bar{\Omega}_h} \leq \frac{R^2}{2d} \|\Delta_h v_h\|_{\Omega_h}$$

also ist das FD-Verfahren stabil mit $C_s := \frac{R^2}{2d}$.

Für den Beweis von 2.54 zeigen wir zunächst eine Hilfsaussage:

Lemma 2.55.

Für $w_h \in X_h$ eine Lösung von

$$\begin{aligned} -\Delta_h w_h &= 1 && \text{in } \Omega_h \\ w_h &= 0 && \text{auf } \Gamma_h \end{aligned}$$

gilt

$$\forall x \in \bar{\Omega}_h: 0 \leq w_h(x) \leq \frac{1}{2d}(R^2 - \|x\|_2^2). \quad (2.11)$$

Beweis.

Sei $w(x) := \frac{1}{2d}(R^2 - \|x\|_2^2)$, also kann man w als ein Produkt von mehreren Polynomen einer Variable sehen, und somit ist Δ_h exakt für w gemäß Kor. 2.34 bzw. die Bemerkung danach.

Daher gilt für jedes $x \in \Omega_h$

$$\begin{aligned} -\Delta_h w(x) &= -\Delta w(x) \\ &= -\sum_{i=1}^d \partial_{x_i}^2 (R^2 - \|x\|_2^2) \frac{1}{2d} \\ &= -\sum_{i=1}^d (-2) \frac{1}{2d} \\ &= 1 \\ &= -\Delta_h w_h(x). \end{aligned}$$

Weiter ist $w \geq 0 = w_h$ auf Γ_h nach Wahl von R .

Aus dem diskreten Vergleichsprinzip 2.49 folgt $w \geq w_h$ auf $\bar{\Omega}_h$, also gilt die zweite Ungleichung in (2.11).

Die erste Ungleichung in (2.11) folgt aus diskretem Maximumsprinzip für $-w_h$:

$$\begin{aligned} \forall x \in \Omega_h: -\Delta_h(-w_h) = -1 \leq 0 &\Rightarrow \max_{x \in \bar{\Omega}_h} -w_h(x) \leq \max_{x \in \Gamma_h} -w_h(x) = 0 \\ &\Rightarrow \forall x \in \bar{\Omega}_h: w_h \geq 0. \end{aligned}$$

und damit sind wir fertig mit dem Nachweis der Hilfsaussage. \square

Beweis. (von 2.54)

Sei $v_h \in X_h^\circ$ und w_h aus Lemma 2.55.

Dann gilt für $x \in \Omega_h$

$$-\frac{\Delta_h v_h(x)}{\|\Delta_h v_h\|_{\Omega_h}} \leq \frac{|\Delta_h v_h(x)|}{\|\Delta_h v_h\|_{\Omega_h}} \leq 1 = -\Delta_h w_h(x).$$

Für $x \in \Gamma_h$ gilt

$$\frac{-\Delta_h v_h(x)}{\|\Delta_h v_h\|_{\Omega_h}} = 0 = w_h(x).$$

Also folgt mit diskretem Vergleichsprinzip 2.49 für alle $x \in \bar{\Omega}_h$

$$\frac{v_h(x)}{\|\Delta_h v_h\|_{\Omega_h}} \leq w_h(x) \leq \frac{1}{2d}(R^2 - \|x\|_2^2) \leq \frac{R^2}{2d}$$

wobei das zweite Gleichheitszeichen aus 2.55 kommt.

Analoge Argumentation gilt für $-v_h$, also ist $\frac{-v_h(x)}{\|\Delta_h v_h\|_{\Omega_h}} \leq \frac{R^2}{2d}$.

Damit gilt insgesamt

$$\|v_h\|_{\bar{\Omega}_h} \leq \frac{R^2}{2d} \|\Delta_h v_h\|_{\Omega_h}.$$

\square

Satz 2.56. (Konvergenz)

Sei ein FD-Verfahren für elliptisches RWP 2.41 gegeben.

Falls das FD-Verfahren stabil & konsistent mit Ordnung p ist, dann ist es auch konvergenz mit Ordnung p .

Beweis.

Sei u exakte Lösung und $u_h \in X_h$ FD-Approximation des elliptischen RWP.

Dann hat $u - u_h$ Nullrandwerte auf Γ_h , also folgt wegen Stabilität

$$\|u - u_h\|_{\bar{\Omega}_h} \leq C_s \|\mathcal{L}_h(u - u_h)\|_{\Omega_h} = C_s \|\mathcal{L}_h u - \mathcal{L}_h u_h\|_{\Omega_h}.$$

Wegen $\mathcal{L}_h u_h(x) = f(x) = (\mathcal{L}u)(x)$ für $\forall x \in \Omega_h$ folgt mit Konsistenz

$$\|u - u_h\|_{\bar{\Omega}_h} \leq C_s \|\mathcal{L}_h u - \mathcal{L}_h u_h\|_{\Omega_h} \leq C_s \|\mathcal{L}_h u - \mathcal{L}u\|_{\Omega_h} \leq C_s C_c h^p.$$

\square

Satz 2.56 ist auf FD-Diskretisierung für Poisson-RWP anwendbar, denn wir haben Konsistenz (Bem. nach 2.52) und Stabilität im Satz 2.54 nachgewiesen, also folgt:

Folgerung 2.57. (Konvergenz für Poisson-RWP bei FD-Diskretisierung)

Sei $\Omega \subseteq \mathbb{R}^d$ beschränktes Gebiet und die Lösung u des Poisson-RWP erfülle $u \in C^4(\bar{\Omega})$.

Dann konvergiert das FD-Verfahren, d.h.

$$\|u - u_h\|_{\bar{\Omega}_h} \leq Ch^p$$

mit $p = 2$ für Würfelgebiet und $p = 1$ für allgemeine Gebiete.

Ein Weg, Stabilität eines FD-Verfahren zu zeigen, führt über das diskrete Maximumsprinzip, wie für Poisson-RWP durchgeführt.

Einen alternativen Weg bietet Satz 2.53: Es reicht, $\|A_h^{-1}\|_\infty \leq C_s$ für geeignete Konstante unabhängig von h zu zeigen.

Dies ist mit M -Matrix-Theorie möglich.

Wir vereinbaren zunächst zwei Notationen:

- Für eine Matrix $A \in \mathbb{R}^{m \times n}$ schreiben wir „ $A \geq 0$ “ für „ A eintragsweise nicht negativ“. Analog für $A > 0$, $A \leq 0$, $A < 0$.
- Für zwei Matrizen $A, B \in \mathbb{R}^{m \times n}$ schreiben wir „ $A \geq B$ “ für „ A eintragsweise größer gleich B “. Analog für $A > B$, $A \leq B$, $A < B$.

Definition 2.58. (L_0 , L , M -Matrix)

Eine quadratische Matrix $A = (a_{ij})_{i,j=1}^n \in \mathbb{R}^{n \times n}$ ist eine

- i. L_0 -Matrix, g.d.w. $\forall i \neq j: a_{ij} \leq 0$.
- ii. L -Matrix, g.d.w. A eine L_0 -Matrix ist und dazu $\forall i \in \{1, \dots, n\}: a_{ii} > 0$.
- iii. M -Matrix, g.d.w. A eine L_0 -Matrix ist und zudem $A \in \text{Gl}_n(\mathbb{R})$ sowie $A^{-1} \geq 0$.

Bemerkung.

- Wenn A^{-1} existiert und $A^{-1} \geq 0$ gilt, nennt man A auch inversmonoton, also ist eine M -Matrix eine inversmonotone L_0 -Matrix.
- Nun besteht das Ziel darin, Zusatzbedingungen zu finden, sodass wir aus einer L - oder L_0 -Matrix eine M -Matrix schließen und die Norm der Inversmatrix abschätzen können.

Satz 2.59. (M -Kriterien)

Sei $A \in \mathbb{R}^{n \times n}$ eine L_0 -Matrix.

- i. A ist inversmonoton (also M -Matrix), g.d.w. es ein $e \in \mathbb{R}^n$ mit den Eigenschaften $e > 0$ und $Ae > 0$ existiert.
- ii. Falls A eine M -Matrix ist, gilt für e aus i.

$$\|A^{-1}\|_\infty \leq \frac{\|e\|_\infty}{\min_{k \in \mathbb{N}_{\leq n}} (Ae)_k}.$$

Beweis. Wir schreiben im folgenden $(1 \cdots 1)^T =: \mathbf{1}_n$.

i. " \Rightarrow "

Für A inversmonoton, setze $e := A^{-1} \cdot \mathbf{1}_n$.

Damit ist $e > 0$ und $Ae = \mathbf{1}_n > 0$.

" \Leftarrow "

Sei $e \in \mathbb{R}^n$ mit $e > 0$ und $Ae > 0$, d.h.

$$\forall i \in \mathbb{N}_{\leq n}: \sum_{j=1}^n a_{ij} e_j > 0.$$

Da A eine L_0 -Matrix ist, also $\forall i \neq j: a_{ij} e_j \leq 0$, muss $\forall i \in \mathbb{N}_{\leq n}: a_{ii} e_i > 0$ sein, also alle Diagonaleinträge $a_{ii} > 0$, insbesondere ist A eine L -Matrix.

Damit ist $D := \text{diag}(a_{11}, \dots, a_{nn})$ invertierbar.

Setze dann $P := D^{-1}(D - A) = I - D^{-1}A \geq 0$ und auflösen nach A liefert

$$A = D(I - P).$$

Weiter ist $(I - P)e = D^{-1}Ae > 0$ und daher

$$e = Ie > Pe. \quad (2.12)$$

Wir definieren eine spezielle Norm

$$\|x\|_e := \max_{i \in \mathbb{N}_{\leq n}} \frac{x_i}{e_i}$$

sowie die zugehörige induzierte Matrixnorm

$$\|P\|_e := \sup_{\|x\|_e=1} \|Px\|_e.$$

Es gilt $\|e\|_e = \max_{i \in \{1, \dots, n\}} \frac{|e_i|}{e_i} = 1$ also $\|P\|_e \geq \|Pe\|_e$.

Zudem sehen wir: Für $y \in \mathbb{R}^n$ mit $\|y\|_e \max_{i \in \mathbb{N}_{\leq n}} \frac{|y_i|}{e_i} = 1$ ist $y \leq e$.

Wegen $P \geq 0$ gilt

$$\forall i \in \mathbb{N}_{\leq n}: (Py)_i = \sum_{j=1}^n (P)_{ij} y_j \leq \sum_{j=1}^n (P)_{ij} e_j = (Pe)_i$$

also es ist $Py \leq Pe$ und damit $\|P\|_e \leq \|Pe\|_e$.

Insgesamt ist $\|P\|_e = \sup_{\|x\|_e=1} \|Px\|_e = \|Pe\|_e = \max_{i \in \mathbb{N}_{\leq n}} \frac{(Pe)_i}{e_i}$.

Wegen $Pe < e$ aus (2.12) gilt

$$\|P\|_e = \|Pe\|_e < \|e\|_e = 1$$

und damit ist $I - P$ invertierbar wobei die Inverse mittels Neumann'sche Reihe darzustellen ist, also

$$(I - P)^{-1} = \sum_{j=0}^{\infty} P^j.$$

Wegen $A = D(I - P)$ existiert auch $A^{-1} = (I - P)^{-1}D^{-1}$.

Mit $P \geq 0$ ist $P^j \geq 0$ also $(I - P)^{-1} \geq 0$, und da noch $D^{-1} \geq 0$, ist das Produkt $A^{-1} = (I - P)^{-1}D^{-1} \geq 0$, also ist A inversmonoton.

ii. Sei A eine M -Matrix und $w, f \in \mathbb{R}^n$ mit $Aw = f$, d.h. $w = A^{-1}f$, und damit

$$\forall i \in \mathbb{N}_{\leq n}: w_i = (A^{-1}f)_i = \sum_{j=1}^n (A^{-1})_{ij}f_j \leq \|f\|_\infty \sum_{j=1}^n (A^{-1})_{ij}$$

also gilt

$$w \leq \|f\|_\infty A^{-1} \mathbf{1}_n \quad (2.13)$$

und analog für $-w \leq \|f\|_\infty A^{-1} \mathbf{1}_n$.

Es gilt $Ae \geq \min_{k \in \mathbb{N}_{\leq n}} (Ae)_k \mathbf{1}_n$, und mit $Ae > 0$ folgt

$$\frac{Ae}{\min_{k \in \mathbb{N}_{\leq n}} (Ae)_k} \geq \mathbf{1}_n.$$

Also mit (2.13) ist daher

$$\pm w \leq \|f\|_\infty A^{-1} \frac{Ae}{\min_{k \in \mathbb{N}_{\leq n}} (Ae)_k} = \|f\|_\infty \frac{e}{\min_{k \in \mathbb{N}_{\leq n}} (Ae)_k}$$

also gilt

$$\|w\|_\infty \leq \|f\|_\infty \frac{\|e\|_\infty}{\min_{k \in \mathbb{N}_{\leq n}} (Ae)_k}.$$

Daher erhalten wir

$$\|A^{-1}\|_\infty = \sup_{f \in \mathbb{R}^d \setminus \{0\}} \frac{\|A^{-1}f\|_\infty}{\|f\|_\infty} = \frac{\|w\|_\infty}{\|f\|_\infty} \leq \frac{\|e\|_\infty}{\min_{k \in \mathbb{N}_{\leq n}} (Ae)_k}. \quad \square$$

Beispiel. (FD Für Poisson-RWP, $d = 1$, Bsp. 2.39)

Für $h := \frac{1}{n+1}$ sowie

$$A_h := \frac{1}{h^2} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & \ddots & & \\ & \ddots & \ddots & -1 & \\ & & -1 & 2 & \end{pmatrix} \in \mathbb{R}^{n \times n}$$

kann man nachweisen, dass A_h eine M -Matrix ist.

Siehe Uebung...

Satz 2.60.

Sei A_h eine L -Matrix. Falls A_h stark diagonaldominant oder schwach diagonaldominant & unzerlegbar (irreduzibel, also gibt es keine Permutationsmatrix P sodass PA ein Nulldiagonalblock hat), dann ist A_h eine M -Matrix.

Für den Beweis des Satzes verweisen wir auf Großmann / Roos Satz 2.8.

Bemerkung.

Falls A_h strikt diagonaldominant, so ist $e = (1 \cdots 1)^T$ ein geeigneter Vektor für 2.59, denn $e > 0$ und $A_h e > 0$ also

$$\|A_h^{-1}\| \leq \frac{1}{\min_{k \in \mathbb{N}_{\leq n}} \sum_{j=1}^n a_{kj}}.$$

Bemerkung. (Gemischte Ableitungen)

- Wir erinnern uns daran, dass in 2.42

$$-\sum_{i,j=1}^d a_{ij}(x) \partial_{x_i}^{c,h} \partial_{x_j}^{c,h} u(x)$$

definiert wird, was zu

$$\frac{1}{2} \begin{pmatrix} a_{12} & 0 & -a_{12} \\ 0 & 0 & 0 \\ -a_{12} & 0 & a_{12} \end{pmatrix}_*$$

in (2.10) führte. Wegen wechselnder Vorzeichen im Fall $a_{12} \neq 0$ können weder das Maximumsprinzip noch M -Matrix-Eigenschaft mit unseren Techniken gezeigt werden.

- Man kann FD-Sterne 2. Ordnung mit nichtpositiven „Eck-Koeffizienten“ für $-2a_{12}\partial_{x_1}\partial_{x_2}u(x)$ konstruieren:

i. Falls $-a_{12} > 0$, dann

$$\begin{pmatrix} a_{12} & -a_{12} & 0 \\ -a_{12} & 2a_{12} & -a_{12} \\ 0 & -a_{12} & a_{12} \end{pmatrix}_*,$$

ii. Falls $-a_{12} < 0$, dann

$$\begin{pmatrix} 0 & a_{12} & -a_{12} \\ a_{12} & -2a_{12} & a_{12} \\ -a_{12} & a_{12} & 0 \end{pmatrix}_*.$$

Mit Konvention

$$a_{12}^+ := \max \{0, a_{12}\}, \quad a_{12}^- := \min \{a_{12}, 0\}$$

folgt für \mathcal{L}_h :

$$\begin{aligned} & \begin{pmatrix} a_{12}^- & -(a_{22} - |a_{12}|) & -a_{12}^+ \\ -(a_{11} - |a_{12}|) & -(a_{11} + a_{22} - |a_{12}|) & -(a_{11} - |a_{12}|) \\ -a_{12}^+ & -(a_{22} - |a_{12}|) & a_{12}^- \end{pmatrix}_* \\ & + \frac{h}{2} \begin{pmatrix} 0 & b_2 & 0 \\ -b_1 & 2hc & b_1 \\ 0 & -b_2 & 0 \end{pmatrix}_*. \end{aligned} \tag{2.14}$$

Folgerung 2.61.

Falls $a_{ii} > |a_{12}| + \frac{h}{2}|b_i|$ für $i \in \{1, 2\}$ und $c \geq 0$, so ist die FD-Systemmatrix A_h zur Diskretisierung (2.14) eine M -Matrix.

Beweis.

Nicht-Diagonal-Elemente von A_h sind nicht positiv; Diagonale sind echt positiv. A_h ist stark diagonaldominant (Summe alle FD-Koeffizienten ≥ 0) also ist A_h mit Satz 2.60 eine M -Matrix. \square

Bemerkung.

- Man kann zeigen, dass $\|A_h^{-1}\|_\infty$ uniform in h beschränkt ist.
- Bedingung $a_{ii} > |a_{12}| + \frac{h}{2}|b_i|$ liefert eine Bedingung für h , d.h. „genügend kleine Gitterweite“ ist erforderlich für Stabilität bei Advectionsterm $b_i \neq 0$. Falls b_i groß (sog. konvektionsdominanter Fall) ist, kann dies zu inpraktikabel kleinen Gitterweiten führen.
- Falls $A(x) = 0, c = 0$ (also reine Advektion) ist, kann man sogar analytisch leicht sehen, dass FD-Diskretisierung mit zentralen Differenzen nicht stabil ist, wie z.B. bei „hyperbolische Gleichung erster Ordnung“, also wenn $\Omega = (0, 1)^2, b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, g(x_1, x_2) = (x_1 - x_2)^2$ und

$$\begin{aligned} \partial_{x_1} u + \partial_{x_2} u &= 0 && \text{in } \Omega \\ u &= g && \text{auf } \Gamma. \end{aligned}$$

- Für konvektionsdominante oder hyperbolische PDE erster Ordnung sind sogenannte Diskretisierungen erforderlich \rightsquigarrow „Finite-Volumen(FV)-Verfahren“.

Beispiel. (Numerisches Beispiel)

siehe Vorlesungsvideo 20

Bemerkung. (Relevanz von FD-Verfahren)

- Bis Mitte 20. Jahrhundert wurden FD-Diskretisierungen als Allzweck-Werkzeug gesehen (Collatz-Zitat, siehe z.B. Großmann/Roos Seite 33), weil sie eine Vielzahl von Anwendungsproblemen sehr leicht mit ausreichender Genauigkeit approximieren. Allmählich kamen dann Finite Elemente Methoden auf, welche wesentlich aufwendigere Assemblierung der Systemmatrizen erfordern, aber bei gleicher Gitterfeinheit häufig präzisere Ergebnisse liefern.
- Bei Konvergenzanalyse von FD-Verfahren trifft man häufig starke (unrealistische) Annahmen an Glattheit der Lösung.

KAPITEL 3

PDEs: SCHWACHE LÖSUNGSTHEORIE & FINITE-ELEMENTE-METHODE

Motivation für schwache Lösungsbegriffe:

- i. Unstetiger Quellterm in Poisson-Gleichung

$$\forall x \in \Omega = (0, 1): -u''(x) = f(x).$$

Falls f unstetig, ist es nicht zu erwarten, dass $u \in C^2$ existiert.

- ii. Unstetige Koeffizient in Diffusionsproblem

$$\forall x \in \Omega = (0, 1): -(a(x)u'(x))' = 0.$$

Falls $a(x)$ unstetig, kann evtl. stetiges u existieren mit unstetiger Ableitung als eine „Lösung“ der DGL, wie z.B.

$$a(x) := \begin{cases} 1, & x \leq \frac{1}{2} \\ 2, & x > \frac{1}{2} \end{cases}, \quad u(x) := \begin{cases} x, & x \leq \frac{1}{2} \\ \frac{1}{4} + \frac{1}{2}x, & x > \frac{1}{2} \end{cases}.$$

Für $\forall x \neq \frac{1}{2}$ gilt $a(x)u'(x) = 1$, also ist $a(x)u'(x)$, sog. der „Fluss“, stetig fortsetzbar. Dabei ist u keine klassische Lösung, aber „schwache Lösung“.

- iii. Transport-Probleme haben manchmal „Travelling-Wave“ Lösung, z.B.: Die Wellengleichung

$$\partial_t^2 u - c^2 \partial_x^2 u = 0$$

auf $\Omega = \mathbb{R}^2$ hat als Lösung u.a.

$$u(x, t) = u_0(x - ct)$$

für alle $u_0 \in C^2(\mathbb{R})$. Diese Formel macht aber für $u_0 \notin C^0$ auch Sinn (wie im Fall von Shockfront oder Überschall-Knall). Für $u_0 \notin C^0$ ist u keine klassische Lösung, sondern eine „Distributionslösung“.

- iv. Bei nicht linearen PDEs können sich aus glatten Anfangsdaten nach endlicher Zeit Unstetigkeiten entwickeln, auch dort ist verallgemeinerter Lösungsbegriff erforderlich.

3.1. SCHWACHE ABLEITUNG & SOBOLEV-RÄUME

Für $d \in \mathbb{N}$ bezeichnen wir mit λ^d das d -dimensionale Lebesgue-Maß von \mathbb{R}^d .

Definition 3.1. (Schwache Ableitung)

Sei $\beta \in \mathbb{N}_0^d$ ein Multiindex, $\Omega \subseteq \mathbb{R}^d$ und $u \in L_{\text{loc}}^1(\Omega)$.

Eine Funktion $v^\beta \in L_{\text{loc}}^1(\Omega)$ heißt eine schwache Ableitung (bzgl. β) von u , g.d.w. es gilt

$$\forall \phi \in C_0^\infty(\Omega): \quad \int_{\Omega} u \partial^\beta \phi \, d\lambda^d = (-1)^{|\beta|} \int_{\Omega} v^\beta \phi \, d\lambda^d.$$

Eine Funktion, die eine schwache Ableitung (bzgl. β) hat, nennt man schwach differenzierbar (bzgl. β).

Bemerkung.

- Wir schreiben (noch) v^β statt $\partial^\beta u$, da wir noch nicht die Eindeutigkeit haben und noch nicht wissen, ob klassische Ableitungen verallgemeinert werden.
- Die Funktion $\phi \in C_0^\infty(\Omega)$ in der Definition wird als „Testfunktion“ genannt.

Beispiel.

Sei $\Omega := (-1, 1)$ und $u(x) := |x|$.

Wir behaupten, dass sgn die schwache Ableitung von u ist.

Dazu betrachten wir für eine beliebige Testfunktion $\phi \in C_0^\infty(\Omega)$:

$$\begin{aligned} \int_{(-1,1)} u \phi' \, d\lambda &= \int_{(-1,0]} u \phi' \, d\lambda + \int_{[0,1)} u \phi' \, d\lambda \\ &= \int_{(-1,0]} x \phi'(x) \, d\lambda(x) + \int_{[0,1)} -x \phi'(x) \, d\lambda(x) \\ &= \left(- \int_{(-1,0]} -\phi(x) \, d\lambda(x) + [-x\phi(x)]_{-1}^0 \right) \\ &\quad + \left(- \int_{[0,1)} \phi(x) \, d\lambda(x) + [x\phi(x)]_0^1 \right) \\ &= \int_{(-1,0]} \phi(x) \, d\lambda(x) - \int_{[0,1)} \phi(x) \, d\lambda(x) \\ &= - \int_{(-1,1)} \text{sgn}(x) \phi(x) \, d\lambda(x) \end{aligned}$$

wobei die dritte Zeile aus partieller Integration kommt, und die vierte Zeile wegen $\phi(-1) = \phi(1) = 0$ dank $\phi \in C_0^\infty(\Omega)$.

Also ist $\text{sgn}(x)$ die schwache Ableitung von u .

Satz 3.2. (Eindeutigkeit von schwacher Ableitung)

Für $\Omega \subseteq \mathbb{R}^d$ und ein $u \in L_{\text{loc}}^1(\Omega)$ existiert zu $\beta \in \mathbb{N}_0^d$ höchstens eine schwache Ableitung bzgl. β im Sinne der $L_{\text{loc}}^1(\Omega)$ -Norm.

Beweis.

Seien v^β, \tilde{v}^β zwei schwache Ableitungen von u bzgl. β , dann folgt

$$(-1)^{|\beta|} \int_{\Omega} \tilde{v}^\beta \phi \, d\lambda^d = \int_{\Omega} u \partial^\beta \phi \, d\lambda^d = (-1)^{|\beta|} \int_{\Omega} v^\beta \phi \, d\lambda^d$$

also gilt

$$\forall \phi \in C_0^\infty(\Omega): \quad \int_{\Omega} (v^\beta - \tilde{v}^\beta) \phi \, d\lambda^d = 0$$

und mit 2.11 Fundamentallemma der Variationsrechnung folgt $v^\beta - \tilde{v}^\beta = 0$ f.ü., also $v^\beta = \tilde{v}^\beta$ f.ü.. \square

Satz 3.3. (Klassische partielle Ableitung ist schwache Ableitung)

Sei $m \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, $u \in C^m(\Omega)$, v^β für ein $|\beta| \leq m$ die schwache Ableitung von u bzgl. β , $\partial^\beta u \in C^{m-|\beta|}(\Omega)$ die klassische partielle Ableitung bzw. β .

Dann ist $v^\beta = \partial^\beta u$ f.ü., und wir bezeichnen ab jetzt mit $\partial^\beta u$ auch die schwache Ableitung von u bzgl. β .

Beweis.

Wegen partieller Integration gilt

$$\forall \phi \in C_0^\infty: \quad \int_{\Omega} u \partial^\beta \phi \, d\lambda^d = (-1)^{|\beta|} \int_{\Omega} \partial^\beta u \phi \, d\lambda^d$$

also ist $\partial^\beta u$ wegen Eindeutigkeit die schwache Ableitung von u . \square

Bemerkung.

Satz 3.3 gilt auch auf Teilintervalle, d.h. falls u stückweise klassisch differenzierbar und schwach differenzierbar, dann stimmen die beiden Ableitungen überein.

Beispiel.

- $u(x) = |x|$ ist schwach differenzierbar mit $\text{sgn}(x)$ als stückweise klassische Ableitung.
- $u(x) = \text{sgn}(x)$ ist nicht schwach differenzierbar, denn:

Falls u schwach differenzierbar wäre, dann müsste die stückweise klassische Ableitung $\partial u = \begin{cases} 0, & x < 0 \\ 0, & x > 0 \end{cases}$ schon die schwache Ableitung sein.

Aber einerseits gilt für eine beliebige stetige gerade Testfunktion ϕ mit $\text{supp}(\phi) \subseteq [-1, 1]$ und $\phi(0) \neq 0$

$$(-1) \int_{\Omega} \partial u \phi d\lambda = 0$$

und andererseits

$$\begin{aligned} \int_{\Omega} u \partial \phi d\lambda &= \int_{[0,1]} u \partial \phi d\lambda + \int_{[-1,0]} u \partial \phi d\lambda \\ &= \int_{[0,1]} \partial \phi d\lambda - \int_{[-1,0]} \partial \phi d\lambda \\ &= (\phi(1) - \phi(0)) - (\phi(0) - \phi(-1)) \\ &= -2\phi(0) \\ &\neq 0 \end{aligned}$$

also kann obiges ∂u nicht die schache Ableitung sein.

Bemerkung. (Zusammenhang zwischen klassischer & schwacher Ableitung)

- Klassische Ableitung ist letztendlich eine punktweise Eigenschaft, aber schwache Ableitung bezieht sich auf der ganzen Funktion (bzw. auf dem ganzen Definitionsbereich der Funktion).
- Falls $u \in L^1_{\text{loc}}(\Omega)$ beide Ableitungen besitzt, dann stimmen sie überein, aber die Existenz von einer Ableitung muss nicht unbedingt die Existenz der Anderen implizieren, z.B. die Betragsfunktion ist nicht klassisch differenzierbar aber schwach differenzierbar, und die Funktion

$$f(x) = \begin{cases} x^2 \sin\left(\frac{1}{x^2}\right), & x \in (0, 1] \\ 0, & x = 0 \end{cases}$$

ist klassisch differenzierbar aber nicht schwach differenzierbar.

Definition 3.4. (Sobolev-Räume)

Sei $\Omega \subseteq \mathbb{R}^d$ offen, $m \in \mathbb{N}_0$, $p \in [1, \infty]$ und $u \in L^1_{\text{loc}}(\Omega)$.

Falls $\partial^\beta u$ die schwachen Ableitungen bzgl. aller Multiindices $|\beta| \leq m$ existieren, dann definieren wir die Sobolev-Norm durch

$$\|u\|_{H^{m,p}(\Omega)} := \left(\sum_{|\beta| \leq m} \|\partial^\beta u\|_{L^p(\Omega)}^p \right)^{\frac{1}{p}}$$

für $p \in [1, \infty)$ und

$$\|u\|_{H^{m,\infty}(\Omega)} := \max_{|\beta| \leq m} \|\partial^\beta u\|_{L^\infty(\Omega)}$$

für $p = \infty$.

Damit definieren wir die Sobolev-Räume durch

$$H^{m,p}(\Omega) := \{u \in L^1_{\text{loc}}(\Omega) : \|u\|_{H^{m,p}(\Omega)} < \infty\}.$$

Für $p=2$ schreiben wir auch $H^m(\Omega) := H^{m,2}(\Omega)$.

Schließlich definieren wir die Sobolev-Seminorm durch

$$|u|_{H^{m,p}(\Omega)} := \left(\sum_{|\beta|=m} \|\partial^\beta u\|_{L^p(\Omega)}^p \right)^{\frac{1}{p}}$$

für $p \in [1, \infty)$ und

$$|u|_{H^{m,\infty}(\Omega)} := \max_{|\beta|=m} \|\partial^\beta u\|_{L^\infty(\Omega)}$$

für $p=\infty$.

Bemerkung.

- Anstelle $H^{m,p}(\Omega)$ ist der Literatur auch oft $W^{m,p}(\Omega)$ verwendet.
- Aus Satz 3.3 ergibt sich

$$C_0^m(\Omega) \subseteq H^{m,p}(\Omega)$$

und falls Ω beschränkt ist, gilt auch

$$C^m(\Omega) \subseteq H^{m,p}(\Omega).$$

Satz 3.5. (Vollständigkeit von $H^{m,p}(\Omega)$)

Sei $\Omega \subseteq \mathbb{R}^d$ offen, $p \in [1, \infty]$ und $m \in \mathbb{N}_0$.

Dann ist $H^{m,p}(\Omega)$ vollständig, also ein Banachraum, insb. ist $H^m(\Omega)$ ein Hilbertraum mit Skalarprodukt

$$(u, v)_{H^m(\Omega)} := \sum_{|\beta| \leq m} \langle \partial^\beta u, \partial^\beta v \rangle_{L^2(\Omega)}.$$

Bemerkung.

- $C^m(\Omega)$ ist nicht vollständig bzw. $\|\bullet\|_{H^{m,p}(\Omega)}$, denn z.B. für $\Omega = (-1, 1)$, $m=0$, und $(u_n)_{n \in \mathbb{N}_{>1}} \subseteq C^0(\Omega)$ mit

$$u_n(x) := \begin{cases} 0 & , x \in (1, 0] \\ nx & , x \in (0, \frac{1}{n}] \\ 1 & , x \in (\frac{1}{n}, 1) \end{cases}$$

gilt

$$\lim_{n \rightarrow \infty} u_n = \chi_{(0,1)}$$

also $\lim_{n \rightarrow \infty} u_n = \chi_{(0,1)} \notin C^0(\Omega)$ aber $\chi_{(0,1)} \in L^p(\Omega)$.

- Alternativ kann man Sobolev-Räume durch Vervollständigung von $C^m(\Omega)$ definieren, also falls Ω beschränkt und $p \in [1, \infty)$ ist, gilt

$$H^{m,p}(\Omega) = \overline{C^m(\Omega)}^{\|\bullet\|_{H^{m,p}(\Omega)}}$$

(siehe Alt. 1.27).

Beweis. Hier nur für $p = 2$.

Sei $(v_n)_{n \in \mathbb{N}}$ eine Cauchy-Folge in $H^m(\Omega)$.

Für $\forall \beta \in \mathbb{N}_0^d$ mit $|\beta| \leq m$ ist daher $(\partial^\beta v_n)_{n \in \mathbb{N}}$ eine Cauchy-Folge in $L^2(\Omega)$ und wegen Vollständigkeit von L^2 existiert ein $v^\beta \in L^2(\Omega)$ s.d.

$$\|\partial^\beta v_n - v^\beta\|_{L^2(\Omega)} \xrightarrow{n \rightarrow \infty} 0.$$

Insbesondere existiert für $\beta = (0, \dots, 0)$ der Grenzwert

$$\lim_{n \rightarrow \infty} \partial^\beta v_n = \lim_{n \rightarrow \infty} v_n =: v^0$$

und wir zeigen, dass v^0 der Grenzwert von $(v_n)_{n \in \mathbb{N}}$ bzgl. Sobolev-Norm ist.

Mit obiger Übunglegung gilt für eine beliebige Testfunktion $\phi \in C_0^\infty(\Omega)$ und alle $\beta \in \mathbb{N}_0^d$ dann

$$\begin{aligned} \langle v^\beta, \phi \rangle_{L^2(\Omega)} &= \lim_{n \rightarrow \infty} \langle \partial^\beta v_n, \phi \rangle_{L^2(\Omega)} \\ &= \lim_{n \rightarrow \infty} (-1)^{|\beta|} \langle v_n, \partial^\beta \phi \rangle_{L^2(\Omega)} \\ &= (-1)^{|\beta|} \left\langle \lim_{n \rightarrow \infty} v_n, \partial^\beta \phi \right\rangle_{L^2(\Omega)} \\ &= (-1)^{|\beta|} \langle v^0, \partial^\beta \phi \rangle_{L^2(\Omega)} \end{aligned}$$

wobei die erste und dritte Zeile aus Stetigkeit vom Skalarprodukt folgt und die zweite Zeile wegen partieller Integration.

D.h., für alle $\beta \in \mathbb{N}_0^d$ ist v^β die schwache Ableitung von v^0 , also $\partial^\beta v^0 \in L^2(\Omega)$ mit $\|\partial^\beta v_n - \partial^\beta v^0\|_{L^2(\Omega)} \xrightarrow{n \rightarrow 0} 0$ und somit $v^0 \in H^m(\Omega)$ mit $\|v_n - v^0\|_{H^m(\Omega)} \xrightarrow{n \rightarrow 0} 0$. \square

Satz 3.6. (Meyers-Serrin, Approximierbarkeit durch C^∞ -Funktionen)

Für $p \in [1, \infty)$ ist $H^{m,p}(\Omega) \cap C^\infty(\Omega)$ dicht in $H^{m,p}(\Omega)$, d.h.

$$\forall p \in [1, \infty) \quad \forall f \in H^{m,p}(\Omega) \quad \exists (f_j)_{j \in \mathbb{N}} \subseteq H^{m,p}(\Omega) \cap C^\infty(\Omega): \quad \|f - f_j\|_{H^{m,p}(\Omega)} \xrightarrow{j \rightarrow \infty} 0.$$

Für den Beweis dieses Satzes verweisen wir auf Alt 1.28.

Bemerkung. (Rechenregeln für schwache Ableitungen)

Aufgrund der Approximierbarkeit mit C^∞ -Funktionen sieht man leicht, dass Regeln zum Umgang mit schwachen Ableitungen von C^∞ -Funktionen auf $H^{m,p}$ -Funktionen übertragen werden können, d.h. insb. Linearität, partielle Integration, Gauß'scher Integralsatz, Produkt-/Kettenregel, etc.

Bemerkung. (Randwerte)

Da L^p -Funktionen auf Nullmengen undefiniert sind bzw. beliebig abgeändert werden können, ist unklar, was man unter „Randwerten“ einer $H^{m,p}$ -Funktion verstehen soll. Tatsächlich hilft die zusätzliche Regularität für $m \geq 1$, die sogenannten „schwachen Randwerte“ zu definieren, welche man mit dem „Spuroperator“ extrahieren kann:

Definition 3.7. (Sobolev-Räume mit schwachen Nullrandwerten)

Für $p \in [1, \infty)$ und $m \in \mathbb{N}$ definieren wir Sobolevräume mit Nullrandwerten

$$H_0^{m,p}(\Omega) := \overline{C_0^m(\Omega)}^{\|\bullet\|_{H^{m,p}(\Omega)}}.$$

Bemerkung.

- Ω darf auch unbeschränkt sein.
- In Literatur findet man auch $W_0^{m,p}$, $H^{0m,p}$, etc. als Notation.

Satz 3.8. (Vollständigkeit von $H_0^{m,p}$)

Für $p \in [1, \infty)$ und $m \in \mathbb{N}$ ist $H_0^{m,p}(\Omega)$ ein abgeschlossener Teilraum von $H^{m,p}(\Omega)$, insbesondere ist $H_0^{m,p}(\Omega)$ ein Banachraum und $\|\bullet\|_{H^{m,p}(\Omega)}$ überträgt sich auf $H_0^{m,p}(\Omega)$.

Beweis.

Es ist $C_0^m(\Omega) \subseteq H^{m,p}(\Omega)$ und $H^{m,p}(\Omega)$ ist abgeschlossen nach 3.5, also ist $H_0^{m,p}(\Omega) = \overline{C_0^m(\Omega)} \subseteq H^{m,p}(\Omega)$ nach Konstruktion ein abgeschlossener Teilraum. \square

Bemerkung.

Mengentheoretisch gilt also

$$\begin{aligned} L^p(\Omega) &= H^{0,p}(\Omega) \supseteq H^{1,p}(\Omega) \supseteq \dots \supseteq H^{m,p}(\Omega) \\ &\quad \cup \qquad \qquad \qquad \cup \\ H_0^{1,p}(\Omega) &\supseteq \dots \supseteq H_0^{m,p}(\Omega) \\ &\quad \cup \qquad \qquad \qquad \cup \\ C_0^1(\Omega) &\supseteq \dots \supseteq C_0^m(\Omega). \end{aligned}$$

Satz 3.9. (Spursatz)

Sei $\Omega \subseteq \mathbb{R}^d$ ein Lipschitz-Gebiet und $p \in [1, \infty)$.

Dann existiert ein stetiger linearer Operator, der sogenannte Spuroperator

$$\gamma: H^{1,p}(\Omega) \rightarrow L^p(\partial\Omega)$$

mit der Eigenschaft

$$\forall u \in H^{1,p}(\Omega) \cap C^0(\bar{\Omega}): \quad \gamma(u) = u|_{\partial\Omega}.$$

Insbesondere gilt

$$\forall u \in H_0^{1,p}(\Omega): \quad \gamma(u) = 0$$

und wegen Stetigkeit existiert eine Konstante $C_\gamma \in \mathbb{R}_+$ s.d.

$$\forall u \in H^{1,p}(\Omega): \quad \|\gamma(u)\|_{L^p(\partial\Omega)} \leq C_\gamma \|u\|_{H^{1,p}(\Omega)}.$$

Bemerkung.

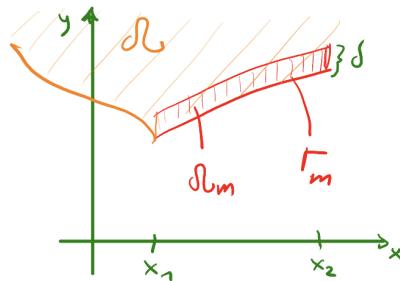
- Auf Nicht-Lipschitz-Gebieten ist der Satz 3.8 i.A. falsch.
- Wir werden nur einen Spezialfall von 3.8 zeigen. Der allgemeiner Fall ist beim Alt A.5.7. zu finden.

Beweis. Hier nur für $d=p=2$ und Ω mit stückweise glattem Rand, d.h. $\partial\Omega$ erlaubt eine endliche Zerlegung

$$\partial\Omega = \bigcup_{m=1}^n \Gamma_m$$

s.d. es auf jedem Γ_m nach geeigneter Rotation des Koordinatensystems gilt

- $\exists \phi_m \in C^1([x_1, x_2]): \quad \Gamma_m = \{(x, y) \in \mathbb{R}^2 \mid x \in [x_1, x_2], y = \phi_m(x)\}$
- Für ein $\delta \in \mathbb{R}_+$ ist $\Omega_m^\delta := \{(x, y) \in \mathbb{R}^2 \mid x \in [x_1, x_2], \phi_m(x) \leq y \leq \phi_m(x) + \delta\}$ in Ω enthalten.



Beweisstrategie:

- i. Wir definieren zunächst

$$\gamma: C^1(\bar{\Omega}) \rightarrow L^2(\partial\Omega), \quad v \mapsto v|_{\partial\Omega}.$$

γ ist klar linear, und wir sollte noch die Wohldefiniertheit (im Sinne von $\gamma(v) \in L^2(\partial\Omega)$) und die Beschränktheit (Stetigkeit) von γ bzgl. H^1 -Norm bei $C^1(\bar{\Omega})$ nachweisen.

- ii. Wir nutzen die Eigenschaft $H^1(\Omega) = \overline{C^1(\bar{\Omega})}^{\| \cdot \|_{H^1(\Omega)}}$ und erweitern γ aus i. auf $H^1(\Omega)$ durch

$$\gamma(v) := \lim_{n \rightarrow \infty} \gamma(v_n)$$

für $v \in H^1(\Omega)$ mit $(v_n)_{n \in \mathbb{N}} \subseteq C^1(\bar{\Omega})$ s.d.

$$\lim_{n \rightarrow \infty} \|v - v_n\|_{H^1(\Omega)} = 0.$$

Dank i. und gilt die Abschätzungen mit derselben Schranke

$$\|\gamma(v)\|_{L^2(\partial\Omega)} = \lim_{n \rightarrow \infty} \|\gamma(v_n)\|_{L^2(\partial\Omega)} \leq C_\gamma \lim_{n \rightarrow \infty} \|v_n\|_{H^1(\Omega)} = C_\gamma \|v\|_{H^1(\Omega)}$$

und dies zeigt die Wohldefiniertheit von γ .

Also bleibt es noch zu zeigen

$$\exists C_\gamma \in \mathbb{R}_+ \forall v \in C^1(\bar{\Omega}): \quad \|\gamma(v)\|_{L^2(\partial\Omega)} \leq C_\gamma \|v\|_{H^1(\Omega)}.$$

Dazu:

- a) Sei $v \in C^1(\bar{\Omega})$ und wir betrachten zunächst $v|_{\Gamma_m}$ für $\partial\Omega = \bigcup_{m=1}^n \Gamma_m$.
 Γ_m ist der Graph einer Funktion $\phi_m \in C^1([x_1, x_2])$ sowie $\Omega_m^\delta \subseteq \Omega$ (siehe Anfang des Beweises).

Für $x \in [x_1, x_2]$ erhalten wir mittels Integraldarstellung

$$\forall t \in [0, \delta]: \quad v(x, \phi_m(x)) = v(x, \phi_m(x) + t) - \int_{[0, t]} \partial_y v(x, \phi_m(x) + s) d\lambda(s).$$

Integration obiger Gleichung über t von 0 bis δ ergibt

$$\begin{aligned} \delta v(x, \phi_m(x)) &= \int_{[0, \delta]} v(x, \phi_m(x) + t) d\lambda(t) \\ &\quad - \int_{[0, \delta]} \int_{[0, t]} \partial_y v(x, \phi_m(x) + s) d\lambda(s) d\lambda(t). \end{aligned}$$

- b) Beim 2. Integral ist der Integrationsbereich ein Dreieck-Normalbereich, und d.h. wir können das 2. Integral umschreiben

$$\begin{aligned} &\int_{[0, \delta]} \int_{[0, t]} \partial_y v(x, \phi_m(x) + s) d\lambda(s) d\lambda(t) \\ &= \int_{[0, \delta]} \int_{[0, s]} \partial_y v(x, \phi_m(x) + s) d\lambda(t) d\lambda(s) \\ &= \int_{[0, \delta]} (\delta - s) \partial_y v(x, \phi_m(x) + s) d\lambda(s) \end{aligned}$$

also erhalten wir

$$\begin{aligned} \delta v(x, \phi_m(x)) &= \int_{[0, \delta]} v(x, \phi_m(x) + t) d\lambda(t) \\ &\quad - \int_{[0, \delta]} (\delta - s) \partial_y v(x, \phi_m(x) + s) d\lambda(s). \end{aligned} \tag{3.1}$$

- c) Wegen Young'sche Ungleichung mit $\varepsilon = 1$ für $p = 2$, also $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$, gilt

$$(a + b)^2 = a^2 + 2ab + b^2 \leq a^2 + 2\left(\frac{1}{2}a^2 + \frac{1}{2}b^2\right) + b^2 = 2a^2 + 2b^2$$

und dies angewandt auf Quadrat von (3.1) ergibt

$$\begin{aligned} \delta^2 v(x, \phi_m(x))^2 &\leq 2 \left(\int_{[0, \delta]} v(x, \phi_m(x) + t) d\lambda(t) \right)^2 \\ &\quad + 2 \left(\int_{[0, \delta]} (\delta - s) \partial_y v(x, \phi_m(x) + s) d\lambda(s) \right)^2. \end{aligned}$$

d) Anwendung von Cauchy-Schwarz auf beide rechte Terme in c) ergibt

$$\begin{aligned} &2 \left(\int_{[0, \delta]} v(x, \phi_m(x) + t) d\lambda(t) \right)^2 \\ &\leq 2 \left(\left(\int_{[0, \delta]} 1^2 d\lambda(t) \right)^{\frac{1}{2}} \left(\int_{[0, \delta]} v(x, \phi_m(x) + t)^2 d\lambda(t) \right)^{\frac{1}{2}} \right)^2 \\ &= 2 \delta \int_{[0, \delta]} v(x, \phi_m(x) + t)^2 d\lambda(t) \end{aligned}$$

sowie

$$\begin{aligned} &2 \left(\int_{[0, \delta]} (\delta - s) \partial_y v(x, \phi_m(x) + s) d\lambda(s) \right)^2 \\ &\leq 2 \left(\left(\int_{[0, \delta]} (\delta - s)^2 d\lambda(s) \right)^{\frac{1}{2}} \left(\int_{[0, \delta]} (\partial_y v(x, \phi_m(x) + s))^2 d\lambda(s) \right)^{\frac{1}{2}} \right)^2 \\ &= 2 \int_{[0, \delta]} (\delta - s)^2 d\lambda(s) \int_{[0, \delta]} (\partial_y v(x, \phi_m(x) + s))^2 d\lambda(s) \\ &= 2 \frac{1}{3} \delta^3 \int_{[0, \delta]} (\partial_y v(x, \phi_m(x) + s))^2 d\lambda(s). \end{aligned}$$

e) Einsetzen von d) in c) liefert

$$\delta^2 v(x, \phi_m(x))^2 \leq 2\delta \int_{[0, \delta]} v(x, \phi_m(x) + t)^2 d\lambda(t) + \frac{2}{3}\delta^3 \int_{[0, \delta]} \partial_y v(x, \phi_m(x) + s)^2 d\lambda(s)$$

und mit Division durch δ^2 und dann Integration bzgl. x erhalten wir

$$\begin{aligned} &\int_{[x_1, x_2]} v(x, \phi_m(x))^2 d\lambda(x) \\ &\leq \frac{2}{\delta} \int_{\Omega_m^\delta} v(x, y)^2 d\lambda(x) d\lambda(y) + \frac{2}{3}\delta \int_{\Omega_m^\delta} (\partial_y v(x, y))^2 d\lambda(x) d\lambda(y) \\ &\leq \max \left\{ \frac{2}{\delta}, \frac{2}{3}\delta \right\} \int_{\Omega_m^\delta} v^2 + \partial_y v^2 d\lambda^2 \\ &\leq \max \left\{ \frac{2}{\delta}, \frac{2}{3}\delta \right\} \int_{\Omega_m^\delta} v^2 + \partial_y v^2 + \partial_x v^2 d\lambda^2 \\ &= \max \left\{ \frac{2}{\delta}, \frac{2}{3}\delta \right\} \|v\|_{H^1(\Omega)}^2 \\ &=: \xi \|v\|_{H^1(\Omega)}^2. \end{aligned}$$

f) Γ_m ist der Graph einer Funktion $\phi_m \in C^1([x_1, x_2])$, und daher

$$\begin{aligned} \int_{\Gamma_m} v^2 d\lambda &= \int_{[x_1, x_2]} v(x, \phi_m(x))^2 \sqrt{1 + \phi'_m(x)^2} d\lambda(x) \\ &\leq \|\sqrt{1 + \phi'_m(x)^2}\|_{L^\infty([x_1, x_2])} \int_{[x_1, x_2]} v(x, \phi_m(x))^2 d\lambda(x) \\ &=: C_m \int_{[x_1, x_2]} v(x, \phi_m(x))^2 d\lambda(x) \end{aligned}$$

und somit folgt dank e)

$$\int_{\Gamma_m} v^2 d\lambda \leq C_m \xi \|v\|_{H^1(\Omega)}^2.$$

g) Insgesamt erhalten wir

$$\|\gamma(v)\|_{L^2(\partial\Omega)}^2 \leq \sum_{m=1}^n \int_{\Gamma_m} v^2 d\lambda \leq \sum_{m=1}^n C_m \xi \|v\|_{H^1(\Omega)}^2$$

und d.h. mit $C_\gamma := (\sum_{m=1}^n C_m \xi)^{\frac{1}{2}}$ ist $\gamma: C^1(\bar{\Omega}) \rightarrow L^2(\partial\Omega), v \mapsto v|_{\partial\Omega}$ beschränkt durch

$$\|\gamma(v)\|_{L^2(\partial\Omega)} \leq C_\gamma \|v\|_{H^1(\Omega)}. \quad \square$$

Definition 3.10. (Sobolev-Dualräume)

Für $p \in [1, \infty)$ bezeichnen wir mit

$$H^{-m,p}(\Omega) := (H_0^{m,p}(\Omega))'$$

den Dualraum von $H_0^{m,p}(\Omega)$. Wir schreiben $H^{-m}(\Omega) := H^{-m,2}(\Omega)$.

Damit kann man DGL betrachten, deren rechte Seite Funktionale statt Funktionen sind (Details später).

Im Folgenden sehen wir, dass Räume $H_0^{m_1,p_1}(\Omega)$ stetig in $H_0^{m_2,p_2}(\Omega)$ eingebettet werden können, falls m_1, m_2, p_1, p_2 geeignet gewählt sind. Weiter lassen sich diese in klassische Hölderräume $C^{m,\alpha}(\bar{\Omega})$ einbetten, damit sind $H^{m,p}(\Omega)$ -Funktionen also „klassisch“ differenzierbar.

Satz 3.11. (1. Sobolev'scher Einbettungssatz)

Sei $\Omega \subseteq \mathbb{R}^d$ offen & beschränkt. Sei $m_1, m_2 \in \mathbb{N}_0$ mit $m_1 \geq m_2$, und $p_1, p_2 \in [1, \infty)$.

Falls $m_1 - \frac{d}{p_1} \geq m_2 - \frac{d}{p_2}$ gilt, so existiert eine stetige Einbettung

$$\mathcal{J}: H_0^{m_1,p_1}(\Omega) \hookrightarrow H_0^{m_2,p_2}(\Omega)$$

insb. gilt

$$\exists C \geq 0 \forall u \in H_0^{m_1,p_1}(\Omega): \|\mathcal{J}u\|_{H_0^{m_2,p_2}(\Omega)} \leq C \|u\|_{H_0^{m_1,p_1}(\Omega)}.$$

Falls Ω ein Lipschitz-Gebiet ist, dann lässt sich \mathcal{J} auf $H^{m_1,p_1}(\Omega)$ zu $H^{m_2,p_2}(\Omega)$ stetig erweitern.

Für den Beweis verweisen wir auf Alt. Satz 8.9.

Satz 3.12. (2. Sobolev'scher Einbettungssatz)

Sei $\Omega \subseteq \mathbb{R}^d$ offen & beschränkt. Sei $m, k \in \mathbb{N}_0$ mit $m \geq k$, und $p \in [1, \infty)$.

Falls es ein $\alpha \in (0, 1)$ existiert, s.d. $m - \frac{d}{p} \geq k + \alpha$ gilt, dann existiert eine stetige Einbettung

$$\mathcal{J}: H_0^{m,p}(\Omega) \hookrightarrow C^{m,\alpha}(\bar{\Omega})$$

insb. gilt

$$\exists C \geq 0 \forall u \in H_0^{m,p}(\Omega): \|\mathcal{J}u\|_{C^{m,\alpha}(\bar{\Omega})} \leq C \|u\|_{H^{m,p}(\Omega)}.$$

Falls Ω ein Lipschitz-Gebiet ist, dann lässt sich \mathcal{J} auf $H^{m,p}(\Omega)$ stetig erweitern.

Für den Beweis verweisen wir auf Alt. Satz 8.13.

Bemerkung.

Für einen Sobolev-Raum $H^{m,p}(\Omega)$ mit $\Omega \subseteq \mathbb{R}^d$ nennen wir die Größe $m - \frac{d}{p}$ den Sobolev-Index. Dies ist offenbar wichtig, denn dies charakterisiert die Regularität von Sobolev-Funktionen.

Folgerung 3.13. (Stetigkeit von $H_0^1(\Omega)$ -Funktion für $d = 1$)

Sei $d = 1$, d.h. $\Omega \subseteq \mathbb{R}$, dann hat jedes $u \in H_0^1(\Omega)$ einen stetigen Repräsentant.

Beweis.

Mit $\alpha \leq \frac{1}{2}$, $k = 0$, $p = 2$ und $m = 1$ gilt $m - \frac{d}{p} = 1 - \frac{1}{2} = \frac{1}{2} \geq \alpha = \alpha + k$, also existiert nach Satz 3.12 eine stetige Einbettung $\mathcal{J}: H_0^1(\Omega) \hookrightarrow C^{0,\alpha}(\bar{\Omega})$. Das bedeutet, für jedes $u \in H_0^1(\Omega)$ ist $\mathcal{J}u \in C^{0,\alpha}(\Omega)$ mit $u = \mathcal{J}u$ f.ü..

Wegen Satz 2.4 Schachtelung der Hölderräume ist also $\mathcal{J}u \in C^0(\Omega)$. □

Bemerkung.

- Für $d > 1$ ist dies falsch: $m - \frac{d}{p} = 1 - \frac{d}{2} \leq 0$ ist nie größer gleich $k + \alpha$ für $\alpha \in (0, 1)$, also Satz 3.12 nicht anwendbar.
- Für $d > 1$ enthält $H^1(\Omega)$ tatsächlich Funktionen mit Punktsingularitäten, z.B.:
 - im Fall $d = 2$: $u(x) = \log\left(\log\left(\frac{2}{\|x\|}\right)\right) \in H^1(B_1(0))$
 - im Fall $d \geq 3$: $u(x) = \|x\|^{-\beta} \in H^1(B_1(0))$ für $\beta \in (0, \frac{d-2}{2})$.

Satz 3.14. (Poincaré-Friedrich)

Sei $\Omega \subseteq \mathbb{R}^d$ offen und beschränkt, sowie $s := \text{diam}(\Omega)$.

Dann gilt

$$\forall v \in H_0^1(\Omega): \|v\|_{L^2(\Omega)} \leq s |v|_{H^1(\Omega)}.$$

Beweis. (Argument für $d=1$ und $v \in C^1$ schon in Bew. 1.41. gesehen)

Da $C_0^\infty(\Omega)$ dicht in $H_0^1(\Omega)$ ist (Satz von Meyers-Serrin), genügt es, Ungleichung für $v \in C_0^\infty(\Omega)$ zu zeigen.

OBdA sei das Koordinatensystem so verschoben, dass es gilt

$$\Omega \subseteq [0, s]^d.$$

also ist jedes $v \in C_0^\infty(\Omega)$ auf $R := [0, s]^d$ durch 0 fortsetzbar, und es gilt für jedes $v \in C_0^\infty(\Omega)$ und jedes $x = (x_1, \dots, x_d) \in R$

$$\begin{aligned} v(x_1, \dots, x_d) &= v(0, x_2, \dots, x_d) + \int_{[0, x_1]} \partial_{x_1} v(t, x_2, \dots, x_d) d\lambda(t) \\ &= 0 + \int_{[0, x_1]} \partial_{x_1} v(t, x_2, \dots, x_d) d\lambda(t). \end{aligned}$$

Damit erhalten wir

$$\begin{aligned} |v(x)|^2 &= \left| \int_{[0, x_1]} 1 \cdot \partial_{x_1} v(t, x_2, \dots, x_d) d\lambda(t) \right|^2 \\ &\leq \int_{[0, x_1]} 1^2 dt \cdot \int_{[0, x_1]} |\partial_{x_1} v(t, x_2, \dots, x_d)|^2 d\lambda(t) \\ &= s \cdot \int_{[0, x_1]} |\partial_{x_1} v(t, x_2, \dots, x_d)|^2 d\lambda(t) \\ &\leq s \cdot \int_{[0, s]} |\partial_{x_1} v(t, x_2, \dots, x_d)|^2 d\lambda(t) \end{aligned}$$

wobei die zweite Zeile aus Cauchy-Schwarz folgt, und damit folgt

$$\begin{aligned} \int_{[0, s]} |v(x_1, \dots, x_d)|^2 d\lambda(x_1) &\leq \int_{[0, s]} s \int_{[0, s]} |\partial_{x_1} v(t, x_2, \dots, x_d)|^2 d\lambda(t) d\lambda(x_1) \\ &= s^2 \int_{[0, s]} |\partial_{x_1} v(t, x_2, \dots, x_d)|^2 d\lambda(t) \end{aligned}$$

da das innere Integral auf der rechten Seite unabhängig von x_1 ist.

Integration über andere Koordinaten ergibt

$$\int_R |v(x)|^2 d\lambda^d(x) \leq s^2 \int_R (\partial_{x_1} v(x))^2 d\lambda^d(x) \leq s^2 \int_R \|\nabla v(x)\|^2 d\lambda^d(x) \quad (3.2)$$

wobei die Abschätzung aus der zweiten Zeile aus der Definition von ∇v folgt, und per Definition dann

$$\|v\|_{L^2}^2 = \int_R |v(x)|^2 d\lambda^d(x) \leq s^2 \int_R \|\nabla v(x)\|^2 d\lambda^d(x) = s^2 \|v\|_{H^1}^2. \quad \square$$

Folgerung 3.15. (Normäquivalenz auf $H_0^m(\Omega)$)

Sei $\Omega \subseteq \mathbb{R}^d$ offen & beschränkt mit $\text{diam}(\Omega) \leq s$.

Dann sind in $H_0^m(\Omega)$ die Norm $\|\bullet\|_{H^m(\Omega)}$ und Seminorm $|\bullet|_{H^m(\Omega)}$ äquivalent mit

$$\forall v \in H_0^m(\Omega): |v|_{H^m(\Omega)} \leq \|v\|_{H^m(\Omega)} \leq (1+s)^m |v|_{H^m(\Omega)}. \quad (3.3)$$

Beweis.

Die linke Seite von (3.3) ist klar, und die rechte Seite zeigen wir mit vollständiger Induktion.

Beim I.A. $m = 0$ ist $|\bullet|_{H^0} = \|\bullet\|_{L^2}$ und $(1+s)^0 = 1$, also (3.3) gilt mit Gleichheit.

Nun zum I.S. $m - 1 \rightarrow m$:

Gelte die Behauptung für $m - 1$, dann gilt für jedes $v \in H_0^m(\Omega)$:

$$\|v\|_{H^m(\Omega)}^2 = \|v\|_{H^{m-1}(\Omega)}^2 + |v|_{H^m(\Omega)}^2 \leq (1+s)^{2(m-1)} |v|_{H^{m-1}(\Omega)}^2 + |v|_{H^m(\Omega)}^2$$

wobei wir uns daran erinnern, dass

$$|v|_{H^{m-1}(\Omega)}^2 = \sum_{|\beta|=m-1} \|\partial^\beta v\|_{L^2(\Omega)}^2.$$

Für jedes $\beta \in \mathbb{N}_0^d$ mit $|\beta| \leq m - 1$ wenden wir die erste Ungleichheit in (3.2) auf $\partial^\beta v$ an und erhalten

$$\|\partial^\beta v\|_{L^2(\Omega)} \leq s \|\partial_{x_1} \partial^\beta v\|_{L^2(\Omega)}$$

und somit

$$\begin{aligned} \|v\|_{H^m(\Omega)}^2 &\leq (1+s)^{2(m-1)} s^2 \sum_{|\beta|=m-1} \|\partial_{x_1} \partial^\beta v\|_{L^2(\Omega)}^2 + |v|_{H^m(\Omega)}^2 \\ &\leq (1+s)^{2(m-1)} s^2 \sum_{|\beta|=m} \|\partial^\beta v\|_{L^2(\Omega)}^2 + |v|_{H^m(\Omega)}^2 \\ &= (1+s)^{2(m-1)} s^2 \sum_{|\beta|=m} \|\partial^\beta v\|_{L^2(\Omega)}^2 + \sum_{|\beta|=m} \|\partial_{x_1} \partial^\beta v\|_{L^2(\Omega)}^2 \\ &= ((1+s)^{2(m-1)} s^2 + 1) \sum_{|\beta|=m} \|\partial^\beta v\|_{L^2(\Omega)}^2. \end{aligned}$$

Mit

$$\begin{aligned} (1+s)^{2m} &= (1+s)^{2(m-1)}(1+s)^2 \\ &= s^2(1+s)^{2(m-1)} + 2s(1+s)^{2(m-1)} + (1+s)^{2(m-1)} \\ &\geq (1+s)^{2(m-1)}s^2 + 1 \end{aligned}$$

folgt daher

$$\|v\|_{H^m(\Omega)}^2 \leq (1+s)^{2m} \sum_{|\beta|=m} \|\partial^\beta v\|_{L^2(\Omega)}^2 = (1+s)^{2m} |v|_{H^m(\Omega)}^2. \quad \square$$

3.2. SCHWACHE LÖSUNGEN FÜR ELLIPTISCHE PDES

Wir illustrieren die **Idee** mittels eines Poisson-RWP mit Nullrandwerten, also

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

für $\Omega \subseteq \mathbb{R}^d$ offen & beschränkt und $f \in C(\Omega)$.

PDE in solcher Darstellung nennt man „starke Form der PDE“.

Sei $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$ eine klassische Lösung zum Poisson-RWP.

Durch Multiplizieren mit einer Testfunktion $v \in C_0^1(\bar{\Omega})$ und dann Integrieren erhalten wir

$$-\int_{\Omega} \Delta u \cdot v d\lambda^d = \int_{\Omega} f \cdot v d\lambda^d$$

und mit partieller Integration folgt

$$-\int_{\Omega} \Delta u \cdot v d\lambda^d = \int_{\Omega} \nabla u \bullet \nabla v d\lambda^d - \int_{\partial\Omega} (\nabla u \bullet n) v d\lambda^{d-1} = \int_{\Omega} \nabla u \bullet \nabla v d\lambda^d - 0$$

da $v \in C_0^1(\bar{\Omega})$.

D.h., die klassische Lösung u löst auch die „schwache Form der PDE“

$$\forall v \in C_0^1(\Omega): \quad \int_{\Omega} \nabla u \bullet \nabla v d\lambda^d = \int_{\Omega} f \cdot v d\lambda^d. \quad (3.4)$$

Hierzu bemerken wir:

- Terme in (3.4) machen Sinn für $u \in C_0^1(\Omega)$ oder $u \in H_0^1(\Omega)$, daher macht es Sinn, nach „schwachen Lösungen“, d.h. Lösung von (3.4) zu suchen.
- Eine klassische Lösung ist also eine schwache Lösung.
- Für allgemeinere f , z.B. unstetige, kann es vorkommen, dass keine klassische Lösung, aber schwache Lösung $u \in H_0^1(\Omega)$ existiert.
- Durch eine Bilinearform

$$a(u, v) := \int_{\Omega} \nabla u \bullet \nabla v d\lambda^d$$

und eine Linearform

$$L_f(v) := \int_{\Omega} f \cdot v d\lambda^d$$

kann (3.4) umgeschrieben werden als

$$\forall v \in H_0^1(\Omega): \quad a(u, v) = L_f(v).$$

Wegen Approximierbarkeit von $H_0^1(\Omega)$ bzw. $C_0^1(\Omega)$ durch C_0^∞ -Funktionen ist es egal, ob $\forall v \in C_0^1(\Omega)$ oder $\forall v \in H_0^1(\Omega)$ gefordert wird.

- Interessanten Fragen dabei gehören u.a. Existenz & Eindeutigkeit, Stabilität & Regularität solcher schwachen Lösungen.

Definition 3.16. (Stetigkeit)

Sei V ein reeller Hilbertraum mit induzierter Norm $\|\bullet\|$.

Eine Bilinearform $a: V \times V \rightarrow \mathbb{R}$ heißt stetig mit Stetigkeitskonstante $\gamma_a \in \mathbb{R}$, g.d.w. es gilt

$$\sup_{u, v \in V \setminus \{0\}} \frac{|a(u, v)|}{\|u\| \cdot \|v\|} =: \gamma_a < \infty.$$

Eine Linearform $L: V \rightarrow \mathbb{R}$ heißt stetig, g.d.w. es gilt

$$\|L\|_{V'} := \sup_{v \in V \setminus \{0\}} \frac{|L(v)|}{\|v\|} < \infty.$$

Bemerkung.

Also ist z.B. das Skalarprodukt stetig mit $\gamma_{(\bullet, \bullet)} = 1$, denn mit Cauchy-Schwarz gilt für jedes $u, v \in V$: $\langle u, v \rangle \leq \|u\| \cdot \|v\|$, und somit für $u, v \in V \setminus \{0\}$: $\frac{|\langle u, v \rangle|}{\|u\| \cdot \|v\|} \leq 1$ und Gleichheit nur bei $u = v$.

Wir definieren noch einen Begriff für „Beschränkung nach unten“:

Definition 3.17. (Koerzivität)

Sei V ein reeller Hilbertraum mit induzierter Norm $\|\bullet\|$.

Eine Bilinearform $a: V \times V \rightarrow \mathbb{R}$ heißt koerziv mit Koerzivitätskonstante $\alpha_a \in \mathbb{R}$, g.d.w. es gilt

$$\inf_{u \in V \setminus \{0\}} \frac{a(u, u)}{\|u\| \cdot \|u\|} =: \alpha_a > 0.$$

Bemerkung.

- Für eine Bilinearform $a: V \times V \rightarrow \mathbb{R}$ ist ihre Koerzivitätskonstante kleiner gleich ihrer Stetigkeitskonstante, da

$$\alpha_a = \inf_{u \in V \setminus \{0\}} \frac{a(u, u)}{\|u\| \cdot \|u\|} \leq \sup_{u \in V \setminus \{0\}} \frac{|a(u, u)|}{\|u\| \cdot \|u\|} \leq \sup_{u, v \in V \setminus \{0\}} \frac{|a(u, v)|}{\|u\| \cdot \|v\|} = \gamma_a.$$

- Man sieht leicht, dass ein Skalarprodukt $\langle \bullet, \bullet \rangle$ koerziv mit $\alpha_{(\bullet, \bullet)} = 1$ ist.
- Eine Bilinearform $a: V \times V \rightarrow \mathbb{R}$ ist koerziv g.d.w. ihr symmetrischer Anteil $a_s(u, v)$ koerziv ist, da mit $a_s(u, v) := \frac{1}{2}(a(u, v) + a(v, u))$ ist $a(u, u) = a_s(u, u)$.
- γ_a, α_a lassen sich durch geeignete EWP / SVD berechnen.

Definition 3.18. (Bilinearform / Linearform für elliptische PDEs)

Sei $\Omega \subseteq \mathbb{R}^d$ beschränkt und eine elliptische PDE mit Nullrandwerten gegeben durch

$$\begin{aligned} -\nabla \bullet (A \nabla u) + \nabla \bullet (bu) + cu &= f && \text{in } \Omega \\ u &= 0 && \text{auf } \partial\Omega \end{aligned}$$

mit $A(x) = (a_{ij}(x))_{i,j=1}^d \in (L^\infty(\Omega))^{d \times d}$, $b(x) = (b_i(x))_{i=1}^d \in (L^\infty(\Omega))^d$, $c \in L^\infty(\Omega)$ und $f \in L^2(\Omega)$.

Dann definieren wir für $u, v \in H^1(\Omega)$

$$\begin{aligned} a(u, v) &:= \int_{\Omega} (A \nabla u) \bullet \nabla v - (b \bullet \nabla v)u + cuv \, d\lambda^d \\ L_f(v) &:= \int_{\Omega} fv \, d\lambda^d. \end{aligned}$$

Satz 3.19. (Stetigkeit & Koerzivität für $b=0$ und $c=0$)

Sei A gleichmäßig elliptisch, d.h.

$$\exists \tilde{\alpha} \in \mathbb{R}_+ \forall x \in \Omega \forall z \in \mathbb{R}^d: \quad \langle z, A(x)z \rangle_2 \geq \tilde{\alpha} \|z\|_2^2$$

und gleichmäßig beschränkt d.h.

$$\exists C \in \mathbb{R}_+ \exists \text{Matrixnorm } \|\bullet\|_\sim \text{ auf } \mathbb{R}^{d \times d} \forall x \in \Omega: \quad \|A(x)\|_\sim \leq C.$$

Sei zudem $b=0$ und $c=0$.

Dann ist die Bilinearform $a(u, v)$ aus 3.18 stetig auf $H^1(\Omega)$ und koerziv auf $H_0^1(\Omega)$.

Beweis. Zur Stetigkeit auf $H^1(\Omega)$:

Für $u, v \in H^1(\Omega)$ gilt:

$$\begin{aligned} a(u, v) = \int_{\Omega} (A \nabla u) \bullet \nabla v \, d\lambda^d &\leq \int_{\Omega} \|A \nabla u\|_2 \cdot \|\nabla v\|_2 \, d\lambda^d \\ &\leq \int_{\Omega} \|A\|_2 \cdot \|\nabla u\|_2 \cdot \|\nabla v\|_2 \, d\lambda^d \\ &\leq \int_{\Omega} C' \cdot \|A\|_\sim \cdot \|\nabla u\|_2 \cdot \|\nabla v\|_2 \, d\lambda^d \\ &\leq \int_{\Omega} C' C \cdot \|\nabla u\|_2 \cdot \|\nabla v\|_2 \, d\lambda^d \\ &=: \int_{\Omega} \tilde{C} \cdot \|\nabla u\|_2 \cdot \|\nabla v\|_2 \, d\lambda^d \\ &\leq \tilde{C} \int_{\Omega} \|\nabla u\|_2^2 \, d\lambda^d \cdot \int_{\Omega} \|\nabla v\|_2^2 \, d\lambda^d \\ &= \tilde{C} \|u\|_{H^1(\Omega)} \cdot \|v\|_{H^1(\Omega)} \\ &\leq \tilde{C} \|u\|_{H^1(\Omega)} \cdot \|v\|_{H^1(\Omega)} \end{aligned}$$

wobei die erste Zeile wegen Cauchy-Schwarz auf Integranden folgt, dritte Zeile wegen Äquivalenz von Matrizenormen, 6. Zeile wegen C.S. in L^2 und letzte Zeile per Definition von $H^1(\Omega)$ -Norm.

Zur Koerzivität auf $H_0^1(\Omega)$:

Da Ω beschränkt, schreiben wir $s := \text{diam}(\Omega)$.

Damit gilt für jedes $u \in H_0^1(\Omega)$:

$$a(u, u) = \int_{\Omega} (A \nabla u) \bullet \nabla u \, d\lambda^d \geq \int_{\Omega} \tilde{\alpha} \|\nabla u\|^2 \, d\lambda^d = \tilde{\alpha} \|u\|_{H^1(\Omega)}^2 \geq \frac{\tilde{\alpha}}{(1+s)^2} \|u\|_{H^1(\Omega)}^2$$

wobei die erste Ungleichheit aus glm. Elliptizität folgt und die zweite wegen Normäquivalenz 3.15.

Somit sehen wir

$$\alpha_a := \inf_{u \in H_0^1 \setminus \{0\}} \frac{a(u, u)}{\|u\|_{H^1(\Omega)}^2} \geq \frac{\tilde{\alpha}}{(1+s)^2} > 0.$$

□

Bemerkung.

- $a(\bullet, \bullet)$ ist nicht koerziv auf $H^1(\Omega)$:
Wähle $u \in H^1(\Omega)$ konstant, also $u(x) = k$ für ein $k \in \mathbb{R}_{\neq 0}$, dann gilt
$$\frac{a(u, u)}{\|u\|_{H^1(\Omega)}^2} = \frac{0}{\|u\|_{H^1(\Omega)}^2} = 0.$$
- Da Stetigkeit sich auf Teilmengen vererbt, ist insb. $a(\bullet, \bullet)$ stetig auf $H_0^1(\Omega)$.
- Da Koerzivität sich auf Teilmengen vererbt, ist insb. $a(\bullet, \bullet)$ koerziv auf Teilmengen von $H_0^1(\Omega)$.
- Falls A symmetrisch, so ist $a(\bullet, \bullet)$ symmetrisch.
- Ähnliche Aussage wie 3.19 gilt falls $b \neq 0$ aber $c > 0$ „genügend groß“. In diesem Fall stellt sich kein Problem mit der Stetigkeit dar, aber für die Koerzivität wird man sehen, dass im Fall $b \neq 0$ gewisse zusätzliche Voraussetzung von c verlangt werden muss.

Bemerkung. (Stetigkeit der rechten Seite)

Die Linearform

$$L_f: H^1(\Omega) \rightarrow \mathbb{R}, v \mapsto \int_{\Omega} f v \, d\lambda^d$$

ist offensichtlich linear.

Wegen $H^1(\Omega) \subseteq L^2(\Omega)$ gilt mit $f \in L^2(\Omega)$:

$$|L_f(v)| = |\langle f, v \rangle_{L^2}| \leq \|f\|_{L^2} \cdot \|v\|_{L^2} \leq \|f\|_{L^2} \cdot \|v\|_{H^1}$$

also ist L_f stetig auf $H^1(\Omega)$ und $H_0^1(\Omega)$.

Falls $v \in H_0^1(\Omega)$ gilt die Eigenschaft sogar für L_f mit allgemeinerem f , z.B. mit einem $f \notin L^2(\Omega)$, solange noch $L_f \in (H^1(\Omega))$.

Definition 3.20. (Energie-Skalarprodukt)

Falls $a(\bullet, \bullet)$ bilinear und koerziv, dann ist der symmetrische Anteil ein Skalarprodukt, das sogenannte Energieskalarprodukt

$$\langle u, v \rangle_a := \frac{1}{2}(a(u, v) + a(v, u)).$$

Bemerkung.

- Dabei sind die Bilinearität und Symmetrie klar. Positive Definitheit resultiert genau aus Koerzivität.
- Damit ergibt sich die für die Fehleranalyse nützliche Energienorm

$$\|u\|_a := \sqrt{\langle u, u \rangle_a}.$$

- Es folgt Normäquivalenz auf $H_0^1(\Omega)$ von $\|\bullet\|_a$ zu $\|\bullet\|_{H^1}$ (oder $|\bullet|_{H^1}$):

$$\alpha \|u\|_{H^1}^2 \leq a_s(u, u) = \|u\|_a^2 = a(u, u) \leq \gamma_a \|u\|_{H^1}^2$$

wobei die erste Ungleichheit aus Koerzivität und die Zweite aus Stetigkeit von $a(\bullet, \bullet)$ folgt.

Definition 3.21. (Schwache Lösung)

Seien $a(\bullet, \bullet)$, $L_f(\bullet)$ einer elliptischen PDE wie in 3.18 gegeben.

Wir nennen $u \in H_0^1(\Omega)$ schwache Lösung der PDE, g.d.w. es gilt

$$\forall v \in H_0^1(\Omega): \quad a(u, v) = L_f(v).$$

Satz 3.22. (Klassische Lösung ist schwache Lösung)

Sei $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$ klassische Lösung der PDE und $f \in C^0(\Omega)$.

Dann ist u schwache Lösung der PDE.

Beweis. Wie zu Beginn von Abschnitt 3.2: Multiplikation mit $v \in H_0^1(\Omega)$, dann Integration & partielle Integration... \square

Wir behandeln im Folgenden die Existenz & Eindeutigkeit der schwachen Lösungen für elliptische PDEs, und zwar zunächst zum Poisson-Problem, dann allgemein.

Für die Existenz & Eindeutigkeit von schwachen Lösungen der Poisson-Gleichung benötigen wir zwei Hilfssätze:

Satz 3.23. (Orthogonale Projektion)

Sei V ein Hilbertraum und $W \subseteq V$ ein abgeschlossener Teilraum.

Dann existiert genau ein $P: V \rightarrow W$ mit der Eigenschaft

$$\forall v \in V \forall w \in W: \quad \langle v - Pv, w \rangle = 0.$$

P ist ein stetiger linearer Operator und zwar die sogenannte orthogonale Projektion auf W .

Beweis zu 3.23. wird als Übung überlassen. Für $V \cong \mathbb{C}^m$ siehe auch NUM I, 5.20 & 5.21.

Satz 3.24. (Riesz'scher Darstellungssatz)

Sei V ein Hilbertraum, so ist die Riesz-Abbildung \mathcal{J} definiert als

$$\mathcal{R}: V \rightarrow V', v \mapsto \langle v, \bullet \rangle$$

eine stetige, lineare, bijektive Isometrie.

Insbesondere existiert zu jedem $L \in V'$ ein eindeutiger Riesz-Repräsentant $v_L := \mathcal{R}^{-1}(L) \in V$ mit $L(\bullet) = \langle v_L, \bullet \rangle$.

Beweis.

i. Linearität ist klar.

ii. Isometrie folgt aus

$$\|\mathcal{R}(v)\| = \sup_{w \in V \setminus \{0\}} \frac{|\mathcal{R}(v)(w)|}{\|w\|} = \sup_{w \in V \setminus \{0\}} \frac{|\langle v, w \rangle|}{\|w\|} = \frac{|\langle v, v \rangle|}{\|v\|} = \frac{\|v\|^2}{\|v\|} = \|v\|$$

wobei die dritte Gleichheit wegen Cauchy-Schwarz gilt: C.S. liefert eine Ungleichung, bei der die Gleichheit g.d.w. beide Terme gleich gilt.

iii. Stetigkeit folgt direkt aus Isometrie, also

$$\|\mathcal{R}\| = \sup_{v \in V \setminus \{0\}} \frac{\|\mathcal{R}(v)\|}{\|v\|} = 1 < \infty.$$

iv. Zur Injektivität:

$\mathcal{R}(v) = 0$ bedeutet $\|\mathcal{R}(v)\| = 0$ und wegen Isometrie heißt es $\|v\| = 0$ also $v = 0$ und damit ist $\ker(\mathcal{R})$ trivial.

v. Zur Surjektivität:

Für jedes $L \in V' \setminus \{0\}$ ist $\ker(L)$ abgeschlossen, also gibt es nach 3.23 eine orthogonale Projektion $P: V \rightarrow \ker(L)$.

Sei $v_0 \in V$ mit $L(v_0) = 1$ und setze $v_1 := v_0 - Pv_0$.

Es ist $L(v_1) = L(v) = 1$, also $v_1 \neq 0$, und $v_1 \perp \ker(L)$.

Zudem gilt für jedes $v \in V$: $v - L(v)v_1 \in \ker(L)$, und damit

$$\begin{aligned} \left\langle \frac{v_1}{\|v_1\|^2}, v \right\rangle &= \left\langle \frac{v_1}{\|v_1\|^2}, v - L(v)v_1 + L(v)v_1 \right\rangle \\ &= \left\langle \frac{v_1}{\|v_1\|^2}, v - L(v)v_1 \right\rangle + \left\langle \frac{v_1}{\|v_1\|^2}, L(v)v_1 \right\rangle \\ &= 0 + L(v) \frac{\langle v_1, v_1 \rangle}{\|v_1\|^2} \\ &= L(v) \end{aligned}$$

wobei die dritte Gleichheit aus $v_1 \perp \ker(L)$ folgt.

Daher ist $L = \mathcal{R}\left(\frac{v_1}{\|v_1\|^2}\right) \in \text{im}(\mathcal{R})$, also \mathcal{R} surjektiv. \square

Folgerung 3.25. (Ex. & Eind. schwacher Lösung von Poisson-Problem)

Betrachte $A(x) = I$, $c(x) = 0$ und $b(x) = 0$ in Def. 3.18, also

$$\forall v \in H_0^1(\Omega): \quad a(u, v) = \int_{\Omega} \nabla u \bullet \nabla v \, d\lambda^d \quad \wedge \quad L_f(v) = \int_{\Omega} f v \, d\lambda^d$$

für ein beliebiges $f \in L^2(\Omega)$.

Dann existiert eine eindeutige schwache Lösung $u \in H_0^1(\Omega)$ der Poisson-Gleichung mit Nullrandwerten, d.h.

$$\forall v \in H_0^1(\Omega): \quad a(u, v) = L_f(v)$$

und es gilt $|u|_{H^1} \leq C \|L_f\|_{H^{-1}(\Omega)}$ mit C unabhängig von L_f .

Beweis.

$A(x)$ ist glm. elliptisch & glm. beschränkt, also ist $a(\bullet, \bullet)$ nach 3.19 koerziv & stetig auf $H_0^1(\Omega)$.

Zudem ist $a(\bullet, \bullet)$ symmetrisch, also genau das Energie-Skalarprodukt aus Def. 3.20 und es gilt $a(u, u) = |u|_{H^1(\Omega)}^2$.

Da $(H_0^1(\Omega), \|\bullet\|_{H^1(\Omega)})$ ein Hilbertraum ist, folgt mit Normäquivalenz 3.15, dass $(H_0^1(\Omega), |\bullet|_{H^1(\Omega)})$ auch ein Hilbertraum ist.

Somit existiert nach Satz 3.24 ein stetiger isometrischer Isomorphismus

$$\mathcal{R}: H_0^1(\Omega) \xrightarrow{\sim} H_0^{-1}(\Omega)$$

und mit $f \in L^2(\Omega)$ ist $L_f \in H^{-1}(\Omega)$ also nach 3.24 ist

$$u := \mathcal{R}^{-1}(L)$$

das eindeutige Element mit

$$\forall v \in H_0^1(\Omega): \quad a(u, v) = \langle u, v \rangle_a = \langle \mathcal{R}^{-1}(L_f), v \rangle_a = L_f(v).$$

Wegen Isometrie von \mathcal{R} bzgl. $|\bullet|_{H^1(\Omega)}$ und zugehöriger induzierter Norm auf $H^{-1}(\Omega)$ gilt

$$|u|_{H^1(\Omega)} = \sup_{v \in H_0^1(\Omega) \setminus \{0\}} \frac{|L_f(v)|}{\|v\|_{H^1}} \leq \frac{1}{c} \sup_{v \in H_0^1(\Omega) \setminus \{0\}} \frac{|L_f(v)|}{\|v\|_{H^1}} =: C \|L\|_{H^{-1}}$$

wobei die Ungleichheit aus Normäquivalenz 3.15 für $c\|v\|_{H^1} \leq |v|_{H^1}$ folgt. \square

Bemerkung.

- "Schwache Form des Differentialoperators $-\Delta u$ " ist also

$$a: H_0^1 \rightarrow H^{-1}, u \mapsto a(u, \bullet).$$

- Riesz-Abbildung ist also deren Umkehrabbildung / Lösungsoperator für Poisson-Problem mit Nullrandbedingung:

$$a(u, v) = L_f(v) \quad : \Leftrightarrow \quad u = \mathcal{R}^{-1}(L).$$

Für Existenz & Eindeutigkeit für allgemeine elliptische PDEs benötigen wir folgenden Hilfssatz:

Satz 3.26. (Lax-Milgram)

Sei V ein Hilbertraum, $a: V \times V \rightarrow \mathbb{R}$ eine stetige, koerzive Bilinearform mit Koerzivitätskonstante $\alpha \in \mathbb{R}_+$.

Dann existiert genau eine lineare stetige Isomorphismus $\mathcal{A}: V \rightarrow V$ mit der Eigenschaft

$$\forall u, v \in V: \quad a(u, v) = \langle \mathcal{A}u, v \rangle$$

und bei der Inverse gilt

$$\|\mathcal{A}^{-1}\| = \sup_{v \in V \setminus \{0\}} \frac{\|\mathcal{A}^{-1}v\|}{\|v\|} \leq \frac{1}{\alpha}.$$

Beweis.

i. Wir definieren \mathcal{A} :

Sei $u \in V$ fest, dann ist $a(u, \bullet): V \rightarrow \mathbb{R}$ linear & stetig, also $a(u, \bullet) \in V'$ und daher existiert nach Satz 3.24 ein eindeutiges Riesz-Präsentant $v_u \in V$ so dass $a(u, \bullet) = \langle v_u, \bullet \rangle$ gilt.

Nun definieren wir

$$\mathcal{A}: V \rightarrow V, u \mapsto v_u$$

mit v_u so gewählt wie oben.

ii. Linearität von \mathcal{A} folgt aus Linearität von $\langle \bullet, \bullet \rangle$ und von $a(\bullet, \bullet)$.

iii. Zur Stetigkeit von \mathcal{A} :

Mit Stetigkeit von $a(\bullet, \bullet)$ gilt

$$\|\mathcal{A}u\|^2 = \langle \mathcal{A}u, \mathcal{A}u \rangle = a(u, \mathcal{A}u) \leq \gamma_a \|u\| \cdot \|\mathcal{A}u\|$$

also $\|\mathcal{A}u\| \leq \gamma_a \|u\|$, und somit

$$\sup_{u \in V \setminus \{0\}} \frac{\|\mathcal{A}u\|}{\|u\|} \leq \gamma_a < \infty.$$

iv. Zur stetigen Invertierbarkeit von \mathcal{A} auf $\text{im}(\mathcal{A})$:

Mit Koerzivität von $a(\bullet, \bullet)$ folgt

$$\alpha \|u\|^2 \leq a(u, u) = \langle \mathcal{A}u, u \rangle \leq \|\mathcal{A}u\| \cdot \|u\|$$

also

$$\forall u \in V: \quad \|\mathcal{A}u\| \geq \alpha \|u\| \tag{3.5}$$

und damit ist $\text{im}(\mathcal{A})$ abgeschlossen, \mathcal{A} injektiv und es existiert die Inverseabbildung $\mathcal{A}^{-1}: \text{im}(\mathcal{A}) \rightarrow V$, die auch stetig ist, denn mit $v = \mathcal{A}u$ liefert (3.5):

$$\|v\| = \|\mathcal{A}u\| \geq \alpha \|u\| = \alpha \|\mathcal{A}^{-1}v\|$$

also $\|\mathcal{A}^{-1}v\| \leq \frac{1}{\alpha} \|v\|$ und somit

$$\|\mathcal{A}^{-1}\| = \sup_{v \in \text{im}(\mathcal{A}) \setminus \{0\}} \frac{\|\mathcal{A}^{-1}v\|}{\|v\|} \leq \frac{1}{\alpha}.$$

v. Es bleibt noch die Surjektivität von \mathcal{A} :

Sei $v \in V$ beliebig und wir zeigen $v \in \text{im}(\mathcal{A})$.

Mit der orthogonalen Projektion $P: V \rightarrow \text{im}(\mathcal{A})$ und $w := v - Pv$ gilt

$$\forall u \in V: \quad \langle \mathcal{A}u, w \rangle = 0$$

also erhalten wir insbesondere für $u = w$:

$$0 = \langle \mathcal{A}w, w \rangle = a(w, w) \geq \alpha \|w\|^2$$

und d.h. $\|w\| = 0$, also $w = 0$, daher $v = Pv \in \text{im}(\mathcal{A})$.

Insgesamt ist \mathcal{A} beschränkt invertierbar auf dem ganzen V . \square

Folgerung 3.27. (Ex. & Eind. der schwachen Lsg. für all. ellip. PDEs)

Seien wie in Def. 3.18

$$\begin{aligned} a(u, v) &:= \int_{\Omega} (A \nabla u) \bullet \nabla v - (b \bullet \nabla v) u + c u v \, d\lambda^d \\ L_f(v) &:= \int_{\Omega} f v \, d\lambda^d \end{aligned}$$

für alle $u, v \in H_0^1(\Omega)$ mit A gleichmäßig elliptisch, A, b, c gleichmäßig beschränkt, $c \geq 0$ genügend groß, s.d. $a(u, v)$ koerziv auf $H_0^1(\Omega)$ mit Koerzivitätskonstante $\alpha > 0$ ist.

Dann existiert für alle $f \in L^2(\Omega)$ eine eindeutige schwache Lösung $u \in H_0^1(\Omega)$, d.h.

$$\forall v \in H_0^1(\Omega): \quad a(u, v) = L_f(v)$$

und diese ist insbesondere beschränkt durch

$$\|u\|_{H^1(\Omega)} \leq \frac{1}{\alpha} \|L_f\|_{H^{-1}(\Omega)}.$$

Beweis.

$L_f \in (H_0^1(\Omega))'$, also liefert der Satz von Riesz einen eindeutigen Riesz-Repräsentant $v_L := \mathcal{R}^{-1}(L_f) \in H_0^1(\Omega)$ mit $\forall v \in V: \langle v_L, v \rangle = L_f(v)$.

Da $a(\bullet, \bullet)$ stetig & koerziv ist, existiert nach Lax-Milgram ein eindeutiges stetig invertierbares $\mathcal{A}: H_0^1(\Omega) \rightarrow H_0^1(\Omega)$.

Setze $u := \mathcal{A}^{-1}v_L$. Dann gilt für alle $v \in H_0^1(\Omega)$:

$$a(u, v) = \langle \mathcal{A}u, v \rangle = \langle \mathcal{A}\mathcal{A}^{-1}v_L, v \rangle = \langle v_L, v \rangle = L_f(v)$$

also ist u eine schwache Lösung.

Eindeutigkeit von u ist garantiert durch Eindeutigkeit von \mathcal{A} und \mathcal{R} .

Beschränktheit ist erfüllt, da aus 3.26 folgt $\|\mathcal{A}^{-1}\| \leq \frac{1}{\alpha}$, und somit

$$\|u\| = \|\mathcal{A}^{-1}v_L\| \leq \|\mathcal{A}^{-1}\| \cdot \|v_L\| = \|\mathcal{A}^{-1}\| \cdot \|L\|_{H^{-1}(\Omega)} \leq \frac{1}{\alpha} \|L\|_{H^{-1}(\Omega)}$$

wobei $\|v_L\| = \|L\|_{H^{-1}(\Omega)}$ aus der Isometrie von \mathcal{R} folgt. \square

Ähnlich wie in Kapitel 2 impliziert Beschränktheit stetige Abhängigkeit:

Folgerung 3.28. (Stetige Abhängigkeit von rechter Seite)

Seien $u, \tilde{u} \in H_0^1(\Omega)$ Lösungen eines allgemeinen elliptischen RWP mit Nullrandwerten zu identischem $a(\bullet, \bullet)$ aber unterschiedlichen $L_f, \tilde{L}_f \in H^{-1}(\Omega)$.

Dann gilt

$$\|u - \tilde{u}\|_{H^1(\Omega)} \leq \frac{1}{\alpha} \|L_f - \tilde{L}_f\|_{H^{-1}(\Omega)}.$$

Beweis.

$w := u - \tilde{u}$ löst $a(w, v) = L_f(v) - \tilde{L}_f(v)$ und mit Beschränktheit aus 3.27 folgt
 $\|w\|_{H^1(\Omega)} \leq \frac{1}{\alpha} \|L_f - \tilde{L}_f\|_{H^{-1}(\Omega)}$ □

Satz 3.29. (Schwache Form als Minimierungsproblem)

Falls $a(\bullet, \bullet): V \times V \rightarrow \mathbb{R}$ stetig, koerziv & symmetrisch ist, dann ist die schwache Form äquivalent zu einem Minimierungsproblem, also

$$u \text{ löst } a(u, \bullet) = L_f(\bullet) \quad \Leftrightarrow \quad u = \arg \min_{v \in V} \frac{1}{2} a(v, v) - L_f(v).$$

Beweis.

Wir machen zunächst einige Vorüberlegungen:

Für $I: V \rightarrow \mathbb{R}, v \mapsto \frac{1}{2} a(v, v) - L_f(v)$ gilt

$$\begin{aligned} I(u + \varepsilon v) &= \frac{1}{2} a(u + \varepsilon v, u + \varepsilon v) - L_f(u + \varepsilon v) \\ &= \frac{1}{2} a(u, u) + \frac{\varepsilon}{2} a(u, v) + \frac{\varepsilon}{2} a(v, u) + \frac{\varepsilon^2}{2} a(v, v) - L_f(u) - \varepsilon L_f(v) \\ &= \frac{1}{2} a(u, u) + \frac{\varepsilon}{2} a(u, v) + \frac{\varepsilon}{2} a(u, v) + \frac{\varepsilon^2}{2} a(v, v) - L_f(u) - \varepsilon L_f(v) \\ &= I(u) + \varepsilon a(u, v) + \frac{\varepsilon^2}{2} a(v, v) - \varepsilon L_f(v) \end{aligned}$$

wobei bei der dritten Zeile genau Kommutativität von a genutzt wird, und in der letzter Zeile die Definition von I , also erhalten wir

$$I(u + \varepsilon v) = I(u) + \varepsilon a(u, v) - \varepsilon L_f(v) + \frac{\varepsilon^2}{2} a(v, v) \quad (\#)$$

und damit gilt

$$\left. \frac{d}{d\varepsilon} I(u + \varepsilon v) \right|_{\varepsilon=0} = [a(u, v) - L_f(v) + \varepsilon a(v, v)]|_{\varepsilon=0} = a(u, v) - L_f(v). \quad (\#\#)$$

„ \Leftarrow “

Sei u ein Maximierer von I .

Dann gilt für alle $v \in V$:

$$0 = \left. \frac{d}{d\varepsilon} I(u + \varepsilon v) \right|_{\varepsilon=0} \quad (\#\#)$$

denn andererfalls gäbe es ein v mit $0 \neq \frac{d}{d\varepsilon} I(u + \varepsilon v)$ Abstiegs/Aufstiegsrichtung und das wäre ein Widerspruch zu Minimalität von u .

„ \Rightarrow “

Aus (#) und Koerzivität $\forall v \neq 0: a(u, v) \geq \alpha \|v\|$ folgt, dass I (strikt) konvex ist.

Sei $a(u, v) = L_f(v)$, also mit (#) gilt (###) für alle $v \in V$.

Es bleibt noch zu zeigen: u ist Minimierer.

Falls nicht, d.h. $\exists \tilde{u} \in V: I(\tilde{u}) < I(u)$.

Setze $g(\varepsilon) := I(u + \varepsilon v)$ mit $v := \tilde{u} - u$, dann gilt $g(\tilde{u}) = g(1) < g(0) = I(u)$, also ist g (strikt) konvex und $g'(0) = 0$.

g' ist monoton wachsend wegen Konvexität, also $g'(\tau) \geq g'(0) = 0$ für $\tau \geq 0$, und damit

$$g(1) = g(0) + \int_0^1 g'(\tau) d\lambda(\tau) \geq 0$$

da $g'(\tau) \geq 0$ und das ist ein Widerspruch zu $g(1) < g(0)$. \square

Bemerkung.

- Für nicht-symmetrisches $a(\bullet, \bullet)$ existiert keine solche Interpretation.
- Erinnerung an NUM II: Dort haben wir die Äquivalenz zwischen quadratischen Problemen und LGS für symmetrisches A behandelt. 3.29 ist ∞ -dimensionales Analogon.
- Im Obigen wird noch einmal klar, dass Vollständigkeit von V essentiell für die Existenz eines Minimums ist: Über $C_0^1(\Omega)$ nimmt $I(v)$ i.A. kein Minimum an, auch wenn Minimalfolgen existieren.

Einige Bemerkung zur Erweiterungsmöglichkeiten:

Bemerkung. (Verallgemeinerte Randbedingungen)

i. Inhomogene Dirichlet-Randbedingung

Gegeben sei

$$-\Delta u = f \quad \text{in } \Omega, \quad u = g \quad \text{auf } \partial\Omega. \quad (1)$$

Sei g derart, dass $\tilde{g} \in C^2(\Omega) \cap C^0(\bar{\Omega})$ existiert mit $g = \tilde{g}$ auf $\partial\Omega$.

Dann gilt:

$$u \text{ löst (1)} \Leftrightarrow \tilde{u} := u - \tilde{g} \text{ löst (2), also}$$

$$\begin{aligned} -\Delta \tilde{u} &= -\Delta(u - \tilde{g}) = -\Delta u + \Delta \tilde{g} = f + \Delta \tilde{g} && \text{in } \Omega \\ \tilde{u} &= u - \tilde{g} = g - \tilde{g} = 0 && \text{auf } \partial\Omega. \end{aligned} \quad (2)$$

Lösungsansatz somit:

Löse RWP (2) für \tilde{u} in schwacher Form, dann löst $u := \tilde{u} + \tilde{g}$ ursprüngliches RWP.

Daraus folgt insbesondere Existenz & Eindeutigkeit schwacher Lösung für inhomogene Dirichlet-RWP.

ii. Gemischte Dirichlet/Neumann-Randbedingung

Sei $\partial\Omega = \Gamma_N \cup \Gamma_D$ mit Γ_N , Γ_D nicht verschwindendes $d - 1$ dimensionales Maß, und wir betrachten

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ auf } \Gamma_D, \quad \nabla u \bullet n = g_N \text{ auf } \Gamma_N \quad (3)$$

die starke Form der PDE.

Definiere dafür einen Lösung-/Testräume

$$V := H_{\Gamma_D}^1(\Omega) := \{v \in H^1(\Omega) \mid v|_{\Gamma_D} = 0\}$$

also $H_0^1(\Omega) \subsetneq H_{\Gamma_D}^1(\Omega) \subsetneq H^1(\Omega)$.

Multiplizieren (3) mit Testfunktion $v \in H_{\Gamma_D}^1(\Omega)$, Integration & partielle Integration liefert

$$\int_{\Omega} \nabla u \bullet \nabla v d\lambda^d - \int_{\partial\Omega} (\nabla u \bullet n) v d\lambda^{d-1} = \int_{\Omega} f v d\lambda^d$$

und wegen

$$\begin{aligned} \int_{\partial\Omega} (\nabla u \bullet n) v d\lambda^{d-1} &= \int_{\Gamma_N} (\nabla u \bullet n) v d\lambda^{d-1} + \int_{\Gamma_D} (\nabla u \bullet n) v d\lambda^{d-1} \\ &= \int_{\Gamma_N} g_N v d\lambda^{d-1} + \int_{\Gamma_D} 0 v d\lambda^{d-1} \end{aligned}$$

erhalten wir eine schwache Form der PDE, nämlich:

Finde $u \in H_{\Gamma_D}^1(\Omega)$ mit

$$\int_{\Omega} \nabla u \bullet \nabla v d\lambda^d = \int_{\Omega} f v d\lambda^d + \int_{\Gamma_N} g_N v d\lambda^{d-1}.$$

Beachte dabei:

- Dirichlet-Randbedingungen werden also in Konstruktion des Lösungs-/Testraums eingebaut („wesentliche/essentielle Randbedingungen“).
- Neumann-Randbedingungen werden über Zusatz-Terme in schwacher Form behandelt (natürliche Randbedingungen).

Gemäß 3.27 existiert genau eine schwache Lösung $u \in H_0^1(\Omega)$, falls $a(\bullet, \bullet)$ koerziv & stetig ist. Unter welchen Bedingungen ist $u \in H^m(\Omega)$ für ein $m \in \mathbb{N}$ oder sogar $u \in C^\infty(\Omega)$?

Definition 3.30. ($H^s(\Omega)$ -Regularität)

Sei $H_0^1(\Omega) \subseteq V \subseteq H^1(\Omega)$.

Eine PDE in schwacher Form

$$v \in V: \quad a(u, v) = \langle f, v \rangle_{L^2(\Omega)}$$

mit a koerziv auf V heißt $H^s(\Omega)$ -regulär, wenn es eine Konstante $C_R \geq 0$ gibt, s.d. zu jedem $f \in H^{s-2}(\Omega)$ eine Lösung $u \in H^s(\Omega)$ existiert mit

$$\|u\|_{H^s(\Omega)} \leq C_R \|f\|_{H^{s-2}(\Omega)}.$$

Beispiel.

Sei $\alpha \in (0, 2)$ und $\Omega := \left\{ x \in \begin{pmatrix} r \cos \varphi \\ r \sin \varphi \end{pmatrix} \mid r \in (0, 1), \varphi \in (0, 2\pi) \right\}$ mit Randsegmenten $\Gamma_1, \Gamma_2, \Gamma_3$ wobei Γ_2 der „Bogenteil“ und Γ_1 und Γ_3 die „Geradeteile“ wie z.B.



Wir sehen, mit $u(x) = \|x\|^{1/\alpha} \sin\left(\frac{\varphi(x)}{\alpha}\right)$ und $\varphi(x) = \arctan\left(\frac{x_2}{x_1}\right)$ ist klassische Lösung von

$$-\Delta u = 0 \quad \text{in } \Omega, \quad u(x) = \begin{cases} \sin\left(\frac{\varphi(x)}{\alpha}\right), & \text{auf } \Gamma_2 \\ 0, & \text{auf } \Gamma_1 \\ \|x\|^{1/\alpha}, & \text{auf } \Gamma_3 \end{cases}$$

also ist u schwache Lösung des inhomogenen RWP $\|x\|^{1/\alpha} \sin \pi = 0$ auf Γ_2 (inh. im Sinne von Nicht-Null auf Γ_2).

Man kann zeigen: Für $\alpha \leq 1$ ist $u \in H^2(\Omega)$; falls $\alpha > 1$ ist $u \notin H^2(\Omega)$.

Satz 3.31. (Satz von Friedrichs)

Sei $\Omega \subseteq \mathbb{R}^d$ offen, beschränkt mit glattem Rand (mindestens C^2) oder ein konkaves Lipschitz-Gebiet und $f \in L^2(\Omega)$.

Dann ist das Poisson-RWP mit Dirichlet-Randbedingungen $H^2(\Omega)$ -regulär.

Für den Beweis bzgl. glatter Gebiete verweisen wir auf Alt A10.3.

Bemerkung.

Verallgemeinerung:

$$f \in H^{s-2}(\Omega) \wedge \Omega \text{ hat } C^s \text{ Rand} \quad \Rightarrow \quad u \in H^s(\Omega).$$

Beispiel.

Für $d = 2$ und $\Omega := \{x \in \mathbb{R}^2 \mid 1 \leq \|x\| \leq 2\}$ Kreisring ist $u(x) = \ln(\|x\|)$ gemäß Def.2.21 eine klassische Lösung des inhomogenen RWP

$$-\Delta u = 0 \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \partial B_1(0), \quad u = \ln(2) \quad \text{auf } \partial B_2(0)$$

(inhomogen im Sinne von auf $\partial B_2(0)$ ist $u \neq 0$).

Ω hat offensichtlich C^s -Randbedingung für jedes $s \in \mathbb{N}$, dazu ist $f = 0 \in H^m(\Omega)$ für jedes $m \in \mathbb{N}$ und $u \in C^\infty(\Omega)$, also ist tatsächlich $u \in H^s(\Omega)$ für jedes $s \in \mathbb{N}$.

Damit schließen wir den Theorieabschnitt ab und kommen zu numerischen Verfahren.

3.3. GALERKIN-VERFAHREN

Wir behandeln zunächst das Galerkin-Verfahren, deren **Grundidee** lautet: Einschränkung der schwachen Form der PDE auf endlich dimensionalen Teilraum.

Definition 3.32. (Galerkin-Projektion)

Sei $a(\bullet, \bullet)$ eine stetige, koerzive Bilinearform auf Hilbertraum V , $L \in V'$ und

$$\forall v \in V: \quad a(u, v) = L(v) \tag{3.6}$$

eine schwache Form der PDE.

Sei $V_h \subseteq V$ ein endlich dimensionaler Teilraum.

Gesucht ist ein $u_h \in V_h$, die sogenannte diskrete Lösung von

$$\forall v \in V_h: \quad a(u_h, v) = L(v). \tag{3.7}$$

Folgerung 3.33. (Existenz, Eindeutigkeit & Beschränktheit von u_h)

Die diskrete Lösung $u_h \in V_h$ von Def. 3.32 existiert eindeutig und ist beschränkt durch

$$\|u_h\| \leq \frac{1}{\alpha} \|L\|_{V'}$$

mit α der Koerzivitätskonstante von $a(\bullet, \bullet)$ auf V .

Beweis.

Da Stetigkeit & Koerzivität von $a(\bullet, \bullet)$ vererbt sich auf Teilmengen, ist Lax-Milgram 3.26 auf V_h anwendbar und die Aussage folgt analog zu Kor. 3.27. \square

Bemerkung.

Das „ h “ in V_h indiziert, dass der Raum V_h durch einen Diskretisierungsparameter (z.B. Gitterweite) $h \in \mathbb{R}^+$ charakterisiert wird und es besteht die Hoffnung/das Ziel, dass $\lim_{h \rightarrow 0} u_h = u$ und die Konvergenz genügend schnell wird.

Wir werden uns mit den folgenden Fragen beschäftigen:

- Wie ist V_h geschickt zu wählen?
- Existieren a-priori-Fehlerschranken der Form $\|u - u_h\| \leq C_u h^p$?
- Existieren a-posteriori-Fehlerschranken der Gestalt $\|u - u_h\| \leq C_{u_h} h^p$?
- Wie stellt man das zugehörige LGS effizient auf?
- Wie kann man das LGS effizient lösen?

Lemma 3.34. (Galerkin-Orthogonalität)

Seien $u \in V$, $u_h \in V_h$ Lösungen von (3.6) bzw. (3.7).

Dann gilt

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

Beweis.

$$a(u - u_h, v) = a(u, v) - a(u_h, v) = L(v) - L(v) = 0 \quad \square$$

Bemerkung.

- Falls $a(\bullet, \bullet)$ symmetrisch ist, definiert $a(\bullet, \bullet)$ ein Skalarprodukt $\langle \bullet, \bullet \rangle_a$ und Lem. 3.34 entspricht gerade der orthogonalen Projektion von V auf V_h .
- Falls $a(\bullet, \bullet)$ nicht symmetrisch ist, ist die Galerkin-Projektion nicht die orthogonalen Projektion bzgl. $\langle \bullet, \bullet \rangle_a$, aber Lem. 3.34 gilt immer noch und man spricht immer noch von „Galerkin-Orthogonalität“, auch wenn kein (offensichtliches) Skalarprodukt der Orthogonalität zugrunde liegt.

Folgerung 3.35. (Reproduktion von Lösungen)

Sei $u \in V$ eine Lösung der schwachen Form der PDE und $u \in V_h$.

Dann gilt $u = u_h$.

Beweis.

Mit $u \in V_h$ ist $u - u_h \in V_h$ und daher gilt wegen 3.34

$$0 = a(u - u_h, u - u_h) \geq \alpha \|u - u_h\|^2 \geq 0$$

also $\|u - u_h\| = 0$ und somit $u = u_h$. \square

Bemerkung.

Also ist z.B. $V_h := \text{span}\{u\}$ optimaler 1-dimensionaler Lösungsraum. Dieser ist natürlich so „teuer“ zu bestimmen wie u , daher inpraktikabel.

Satz 3.36. (Diskretes Problem als LGS)

Sei $V_h \subseteq V$ endlich dimensionaler Teilraum mit Basis $(\varphi_j)_{j=1}^n$.

Definiere die „Steifigkeitsmatrix“ $A_h = (a_{ij})_{i,j=1}^n \in \mathbb{R}^{n \times n}$ sowie die „rechte Seite“ $b_h = (b_i)_{i=1}^n \in \mathbb{R}^n$ durch

$$\forall i, j \in \{1, \dots, n\}: \quad a_{ij} := a(\varphi_j, \varphi_i), \quad b_i := L(\varphi_i)$$

und schreibe die Lösung des LGS

$$A_h d = b_h \tag{3.8}$$

als $d = (d_j)_{j=1}^n \in \mathbb{R}^n$.

Dann ist die diskrete Lösung gegeben durch $u_h = \sum_{j=1}^n d_j \varphi_j$.

Beweis.

Wir sollen zeigen, $\sum_{j=1}^n d_j \varphi_j$ löst die schwache Form der PDE für alle Testfunktion $v \in V_h$. Wegen $(\varphi_j)_{j=1}^n$ Basis von V_h und Linearität von $a(\bullet, \bullet)$ sowie L reicht es, die Eigenschaft für φ_i zu prüfen:

$$a\left(\sum_{j=1}^n d_j \varphi_j, \varphi_i\right) = \sum_{j=1}^n a(d_j \varphi_j, \varphi_i) = (A_h d)_i = (b_h)_i = b_i = L(\varphi_i). \quad \square$$

Bemerkung. (Eigenschaften von A_h)

- Mit $a(\bullet, \bullet)$ koerziv folgt A_h positiv definit, denn für $\forall d \in \mathbb{R}^d \setminus \{0\}$:

$$\langle d, A_h d \rangle = \sum_{i,j=1}^n d_i d_j a_{ij} = a\left(\sum_{i=1}^n d_i \varphi_i, \sum_{j=1}^n d_j \varphi_j\right) \geq \alpha \left\| \sum_{i=1}^n d_i \varphi_i \right\|^2 > 0$$

wobei die letzte Stelle gilt, weil $d \neq 0$ und $(\varphi_j)_{j=1}^n$ Basis.

- Falls $a(\bullet, \bullet)$ symmetrisch ist, dann ist A_h symmetrisch. Dann kann man z.B. das LGS $A_h d = b_h$ durch CG-Verfahren lösen.

Beispiel. (Optimales A_h)

Falls V_h gespannt aus Eigenfunktionen des Differentialoperators, bilden diese Eigenfunktionen eine optimale Basis bei unbekannter/variabler rechter Seite.

Dazu sollen wir zunächst eine „schwache form“ des Eigenwertproblems zum Differentialoperator finden, also:

Finde Eigenfunktion $w \in V \subseteq L^2(\Omega)$ und Eigenwert λ , s.d.

$$\forall v \in L^2(\Omega): \quad a(w, v) = \lambda \langle w, v \rangle_{L^2}.$$

Unter gewissen Voraussetzungen (z.B. $a(\bullet, \bullet)$ symmetrisch & Ω zusammenhängend, sieh Evans K6.5) kann man zeigen:

- Es existieren nur reelle positive Eigenwerte
- Es existieren unendlich viele Eigenwerte und die Folge der Eigenwerte ist unbeschränkt, d.h. $\exists (\lambda_j)_{j \in \mathbb{N}} : 0 < \lambda_1 \leq \lambda_2 \leq \dots \wedge \lim_{j \rightarrow \infty} \lambda_j = \infty$.
- Es existiert eine orthonormale Menge $\{w_j\}_{j \in \mathbb{N}}$ in $L^2(\Omega)$ mit w_j als Eigenfunktion zum Eigenwert λ_j .

Aus diem Beispiel von oben können wir einen Ansatz formulieren:

- Wähle $k \in \mathbb{N}$, $h := \frac{1}{k}$ und $V_h := \text{span}\{\varphi_1, \dots, \varphi_k\} \subseteq V$ mit $\varphi_j := w_j$.
- Für Steifigkeitsmatrix folgt

$$a(\varphi_j, \varphi_i) = a(w_j, w_i) = \lambda_j \langle w_j, w_i \rangle_{L^2} = \lambda_j \delta_{ij}$$

also LGS (3.8)

$$\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_k \end{pmatrix} \begin{pmatrix} d_1 \\ \vdots \\ d_k \end{pmatrix} = \begin{pmatrix} L(w_1) \\ \vdots \\ L(w_k) \end{pmatrix}$$

und die Lösung ist dann gegeben durch

$$d_j := \frac{L(w_j)}{\lambda_j}.$$

Also ist eine optimale Struktur von A_h diagonal, aber das Verfahren ist in der Praxis inpraktikabel, da w_j , λ_j selten bekannt sind.

Beispiel. (V_h als polynomialer Ansatzraum)

Sei $d = 1$, $\Omega = (0, 1)$ sowie $a(\bullet, \bullet)$ definiert auf $H_0^1(\Omega)$ durch

$$a(u, v) := \int_{\Omega} u' v' \, d\lambda.$$

Polynom mit Nullrandwerten hat Gestalt $p(x) = x(1-x)q(x)$, daher kann man $k \in \mathbb{N}$, $h := \frac{1}{k}$ und $V_h := \text{span}\{\varphi_1, \dots, \varphi_k\} \subseteq V$ wählen wobei $\varphi_j(x) := x(1-x)x^{j-1}$.

Damit gilt es für die Steifigkeitsmatrix $A_h = (a_{ij})_{i,j=1}^n$:

$$a_{ij} = a(\varphi_j, \varphi_i) = \int_{\Omega} \varphi'_j(x) \varphi'_i(x) \, d\lambda$$

und i.A. ist $a_{ij} = a(\varphi_j, \varphi_i) \neq 0$.

D.h., die Steifigkeitsmatrix A_h ist i.A. nicht dünn besetzt.

Bei sehr großem k ist der Ansatz inpraktikabel wegen Speicherproblemen.

Bei „mäßigem“ k (so 10 oder 20) nennt man den Ansatz „Spektralverfahren“.

Wir versuchen nun eine Fehlerschranke für $u - u_h$ herzuleiten:

Lemma 3.37. (Céa)

Sei $a(\bullet, \bullet)$ koerzive, stetige Bilinearform auf V mit Stetigkeitskonstante γ und Koerzivitätskonstante α . Sei dazu L eine stetige Linearform.

Dann gilt für Lösungen $u \in V$ und $u_h \in V_h$ für (3.6) bzw. (3.7)

$$\|u - u_h\| \leq \frac{\gamma}{\alpha} \inf_{v \in V_h} \|u - v\|.$$

Beweis.

Sei $v \in V_h$ beliebig, dann gilt

$$\begin{aligned} \alpha \|u - u_h\|^2 &\leq a(u - u_h, u - u_h) \\ &= a(u - u_h, u - v) + a(u - u_h, v - u_h) \\ &= a(u - u_h, u - v) + 0 \\ &\leq \gamma \|u - u_h\| \cdot \|u - v\| \end{aligned}$$

wobei die 3. Zeile aus Galerkin-Orthogonalität folgt, und die letzte Zeile wegen Stetigkeit von $a(\bullet, \bullet)$.

Insb. gilt die Abschätzung für $v \in V_h$ s.d. $\|u - v\|$ das Infimum annimmt. \square

Bemerkung. (Galerkin-Projektion & Bestapproximation)

- $\inf_{v \in V_h} \|u - v\|$ auf der rechten Seite von 3.37 ist der „Bestapproximationsfehler“ gegeben durch orthogonale Projektion bzgl. $\langle \bullet, \bullet \rangle_V$ auf V_h , und zwar unabhängig von $a(\bullet, \bullet)$.
- $\|u - u_h\|$ ist Galerkin-Projektionsfehler, und dies ist nach Lemma von Céa höchstens um einen konstanten, von h -unabhängigen Faktor $\frac{\gamma}{\alpha}$ schlechter als der Bestapproximationsfehler. Man nennt diesen Effekt auch „Quasi-Optimalität“ der Galerkin-Projektion.
- Offensichtlich ist ein gutes Konstruktionsprinzip für V_h gegeben durch:
Wähle V_h so, dass dieser gute Approximation ermöglicht, dann wird auch $\|u - u_h\|$ klein sein. Formal

$$\lim_{n \rightarrow \infty} \inf_{v \in V_h} \|u - v\| = 0 \quad \Rightarrow \quad \lim_{n \rightarrow 0} \|u - u_h\| = 0.$$

Bemerkung. (Verbesserung der Schranke für symmetrische Probleme)

Falls $a(\bullet, \bullet)$ zusätzlich symmetrisch ist, so gilt Aussage 3.37 mit besserem Faktor $\sqrt{\frac{\gamma}{\alpha}}$, denn:

Mit Stetigkeit von $a(\bullet, \bullet)$ ist $\|v\|_a^2 = a(v, v) \leq \gamma \|v\|^2$, und mit Koerzivität gilt $\alpha \|v\|^2 \leq a(v, v) = \|v\|_a^2$.

Beides kombiniert zusammen ergibt Normäquivalenz

$$\sqrt{\alpha} \|v\| \leq \|v\|_a \leq \sqrt{\gamma} \|v\|.$$

Dann gilt

$$\begin{aligned}\|u - u_h\|_a^2 &= a(u - u_h, u - u_h) \\ &= a(u - u_h, u - v) \\ &= \langle u - u_h, u - v \rangle_a \\ &\leq \|u - u_h\|_a \cdot \|u - v\|_a\end{aligned}$$

wobei die zweite Zeile analog wie im Beweis bei Céa aus Galerkin-Orthogonalität folgt, und die letzte Zeile wegen Cauchy-Schwarz, also gilt

$$\forall v \in V_h: \|u - u_h\|_a \leq \|u - v\|_a.$$

Mit Normäquivalenz folgt daher

$$\|u - u_h\| \leq \frac{1}{\sqrt{\alpha}} \|u - u_h\|_a \leq \frac{1}{\sqrt{\alpha}} \|u - v\|_a \leq \sqrt{\frac{\gamma}{\alpha}} \|u - v\|$$

also

$$\|u - u_h\| \leq \sqrt{\frac{\gamma}{\alpha}} \inf_{v \in V_h} \|u - v\|.$$

Bemerkung. (Koerzivität wesentlich bei Galerkin-Projektion)

Falls $a(\bullet, \bullet)$ nicht koerziv ist, kann die Galerkin-Projektion aus einem regulären System in V ein singuläres System in V_h erzeugen.

Als Beispiel betrachte $V := \mathbb{R}^2$, $a(u, v) := u^T \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} v$ und $L(v) := \begin{pmatrix} 1 & 1 \end{pmatrix} v$.

Wir sehen, $a(\bullet, \bullet)$ ist nicht koerziv, denn

$$\forall \alpha \in \mathbb{R}_+: a(e_2, e_2) = -1 < 0 \leq \alpha \|e_2\|^2 = \alpha$$

aber System in V ist wohlgestellt, da

$$\begin{aligned}\forall v \in V: a(u, v) = L(v) &\Leftrightarrow \forall v \in V: u^T \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} v = \begin{pmatrix} 1 & 1 \end{pmatrix} v \\ &\Leftrightarrow u^T \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \end{pmatrix} \\ &\Leftrightarrow u = \begin{pmatrix} 1 \\ -1 \end{pmatrix}\end{aligned}$$

(Mehr zu Wohlgestelltheit siehe NUMPDE).

Mit $V_h := \text{span} \left\{ \varphi_1 := \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}$ entsteht jedoch ein singuläres System, denn gemäß Satz 3.36 erhalten wir $u_h = d \cdot \varphi_1$ mit $d \in \mathbb{R}$ Lösung des LGS $A_h d = b_h$ wobei

$$A_h = (a_{11}) = a(\varphi_1, \varphi_1) = 0, \quad b_h = L(\varphi_1) = 2,$$

aber das System $0 \cdot d = 2$ ist nicht lösbar!

Ausweg dafür wäre Verwendung getrennter Ansatz- & Testräume, also die sogenannte „Petrov-Galerkin“-Projektion.

Dabei wählt man $V_h, \tilde{V}_h \subseteq V$ und suche nach $u_h \in V_h$ mit der Eigenschaft

$$\forall v \in \tilde{V}_h: \quad a(u_h, v) = L(v).$$

Man wählt V_h, \tilde{V}_h derart, so dass das resultierende projezierte System regulär ist, z.B. $\varphi_1 \in \mathbb{R}^2 \setminus \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}$ beliebig, $\tilde{\varphi}_1 := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \varphi_1$, $V_h := \text{span}\{\varphi_1\}$ und $\tilde{V}_h := \text{span}\{\tilde{\varphi}_1\}$.

Dann ist $A_h = (a(\varphi_1, \tilde{\varphi}_1)) = \varphi_1^T \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tilde{\varphi}_1 = \varphi_1 \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \varphi_1 = \|\varphi_1\|^2 > 0$ da $\varphi_1 \neq 0$ und $b_h := L(\tilde{\varphi}_1)$, also ist $A_h d = b_h$ eindeutig lösbar und $u_h = d\varphi_1$ die diskrete Lösung.

Mit dazu wird in Vorlesung NUMPDE behandelt.

3.4. FINITE ELEMENTE METHODE

Die **Idee der Finite-Elemente-Methode (FEM)** ist: Wähle Galerkin-Verfahren mit stückweise polynomiellen, global stetigen Ansatzfunktionen, welche lokalen Träger haben.

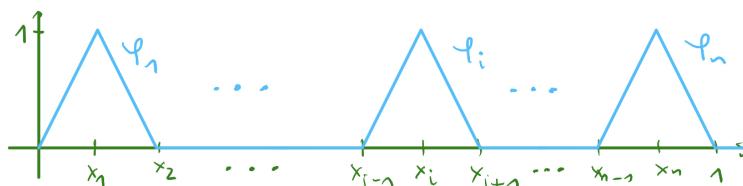
Wir betrachten ein einfaches Beispiel:

Für $d = 1$, $\bar{\Omega} = [0, 1]$ und $n \in \mathbb{N}$ wählen wir äquidistantes Gitter zu Gitterweite $h := \frac{1}{n+1}$ also $\{x_i := ih\}_{i=0}^{n+1}$.

Als Ansatzfunktion definieren wir die „Hütchenfunktion“

$$\varphi_j(x) := \begin{cases} (x - x_{j-1}) \frac{1}{h}, & x \in [x_{j-1}, x_j] \\ 1 - (x - x_j) \frac{1}{h}, & x \in [x_j, x_{j+1}] \\ 0, & \text{sonst} \end{cases}$$

also hat jedes φ_j nur $[x_{j-1}, x_{j+1}]$ als Träger und „es sieht so aus wie ein Hut“ wie das Bild darstellt:



Außerdem sieht man leicht, dass $\varphi_j(x_i) = \delta_{ij}$ gilt, also bildet $\{\varphi_j\}_{j=1}^n$ eine nodale Basis, und der Ansatzraum $V_h := \text{span}\{\varphi_1, \dots, \varphi_n\} \subseteq H_0^1(\Omega)$ ist ein Teilraum der linearen Splines.

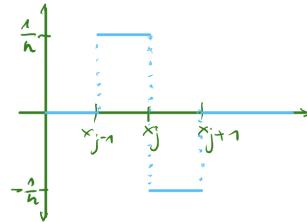
Für das 1d-Poisson-Problem $-u'' = f \wedge u(0) = u(1) = 0$ lautet zunächst

$$a(u, v) := \int_{[0,1]} u'v'd\lambda, \quad L(v) := \int_{[0,1]} fv d\lambda.$$

Die Ableitung der Ansatzfunktion φ_j lautet

$$\varphi'_j(x) = \begin{cases} \frac{1}{h}, & x \in (x_{j-1}, x_j) \\ -\frac{1}{h}, & x \in (x_j, x_{j+1}) \\ 0, & \text{sonst} \end{cases}$$

oder graphisch dargestellt



und damit erhalten wir für die Steifigkeitsmatrix $A = (a_{ij})_{i,j=1}^n$:

- Falls $i = j$:

$$a_{jj} = a(\varphi_j, \varphi_j) = \int_{[0,1]} (\varphi'_j)^2 d\lambda = \int_{[x_{j-1}, x_{j+1}]} \left(\frac{1}{h}\right)^2 d\lambda = 2h \left(\frac{1}{h}\right)^2 = \frac{2}{h}.$$

- Falls $i = j + 1$:

$$a_{j+1,j} = \int_{[0,1]} \varphi'_j \varphi'_{j+1} d\lambda = \int_{[x_j, x_{j+1}]} \frac{1}{h} \left(-\frac{1}{h}\right) d\lambda = -h \frac{1}{h^2} = \frac{-1}{h}.$$

- Falls $i = j - 1$:

Analog zu dem Fall $i = j + 1$, also $a_{j,j-1} = \frac{-1}{h}$.

- Falls $|i - j| \geq 2$:

Dann ist $a_{ij} = 0$ da dann $\varphi'_i \varphi'_j = 0$ in allen Teilintervallen.

Also erhalten wir

$$A_h = \frac{1}{h} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & \ddots & & \\ & \ddots & \ddots & -1 & \\ & & -1 & 2 & \end{pmatrix}, \quad b_h = \begin{pmatrix} \int_{[0,1]} \varphi_1 f d\lambda \\ \vdots \\ \int_{[0,1]} \varphi_n f d\lambda \end{pmatrix}$$

Insb. ist A_h eine Tridiagonalmatrix, sehr dünn besetzt, also kein Speicherproblem für große n falls nur Nicht-Null-Elemente gespeichert werden.

Dank guter Approximationsfähigkeit von Splines und des Lemma von Céa ergibt gute Approximation der diskrete Lösung der Galerkin-Projektion (später genauer).

Definition 3.38. (Simplex)

Seien $a_0, a_1, \dots, a_s \in \mathbb{R}^d$ so gewählt, dass $\dim \langle a_0, \dots, a_s \rangle_{\mathbb{R}} = s$ ist, also insbes. $s \leq d$.

i. Dann nennen wir die konvexe Hülle

$$T := \text{conv}(a_0, \dots, a_s) = \left\{ x \in \mathbb{R}^d \mid x = \sum_{j=0}^s \lambda_j a_j, \lambda_j \geq 0, \sum_{j=0}^s \lambda_j = 1 \right\}$$

ein (nicht-degeneriertes) s -dimensionales Simplex in \mathbb{R}^d . Die Punkte $\{a_j\}_{j=0}^s$ heißen Ecken des Simplex.

ii. Für $r \in \{0, \dots, s-1\}$ sei $\{a'_0, \dots, a'_r\} \subseteq \{a_0, \dots, a_s\}$. Dann ist

$$S := \text{conv}(a'_0, \dots, a'_r)$$

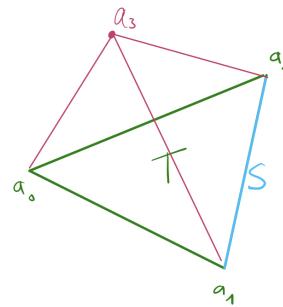
ein r -dimensionales Seitensimplex von T mit Eckenmenge

$$\mathcal{E}(s) := \{a'_0, \dots, a'_r\}.$$

iii. Falls $d=s$, $a_0=e_0=0$ und $a_j=e_j$ die Einheitsvektoren für $j \in \{1, \dots, d\}$, dann heißt $\hat{T} := T$ das Einheitssimplex oder Referenzelement in \mathbb{R}^d .

Bemerkung.

- Für $s=2$ und $d \geq 2$ ist T ein Dreieck.
- Für $s=3$ und $d \geq 3$ ist T ein Tetraeder.
- Für $s=1$ und $d \geq 1$ ist T eine Strecke.
- Für $r=1$ ist S eine Kante und für $r=0$ ist S eine Ecke von T .



Lemma 3.39. (Baryzentrische Koordinaten)

Sei $T = \text{conv}(a_0, \dots, a_s)$ ein s -dimensionales Simplex in \mathbb{R}^d und $x \in T$.

Dann sind die Zahlen $\{\lambda_j\}_{j=0}^s$ mit $x = \sum_{j=0}^s \lambda_j a_j$, $\sum_{j=0}^s \lambda_j = 1$ und $\lambda_j \geq 0$ eindeutig bestimmt, und sie sind die sogenannte Baryzentrischen Koordinaten (oder „lokale Koordinaten“).

Also ist die Abbildung $\lambda: T \rightarrow \mathbb{R}^{s+1}$, $x \mapsto \{\lambda_j\}_{j=0}^s$ wohldefiniert.

Beweis.

Die gesuchte $\{\lambda_j\}_{j=0}^s$ erfüllt LGS

$$\underbrace{\left(\begin{array}{c|c|c} a_0 & \cdots & a_s \\ \hline 1 & \cdots & 1 \end{array} \right)}_{=:M \in \mathbb{R}^{(d+1) \times (s+1)}} \left(\begin{array}{c} \lambda_0 \\ \vdots \\ \lambda_s \end{array} \right) = \left(\begin{array}{c} x \\ \hline 1 \end{array} \right).$$

Existenz einer Lösung $\lambda := (\lambda_0 \dots \lambda_s)^T$ mit $\lambda_i \geq 0$ folgt aus Definition des Simplex und $x \in T$.

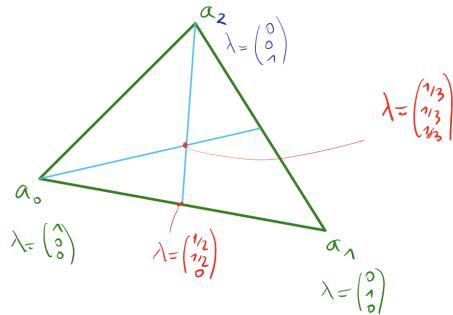
Eindeutigkeit folgt daraus, dass die Matrix M vollen Spaltenrang besitzt, denn Subtraktion der ersten Spalte von M von den Übrigen ergibt eine neue Matrix

$$\bar{M} := \left(\begin{array}{c|c|c|c} a_0 & a_1 - a_0 & \cdots & a_s - a_0 \\ \hline 1 & 0 & \cdots & 0 \end{array} \right)$$

wobei die letzten s Spalten von \bar{M} l.u. sind nach Definition des Simplex und die erste Spalte offensichtlich l.u. zu den Anderen sind. \square

Beispiel.

Betrachte das 2-dimensionale Einheitssimplex in \mathbb{R}^3 , also



Dabei ist der Punkt $\lambda = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right)$ der Schwerpunkt des Dreiecks und auch der Inkreismittelpunkt, denn er ist der Schnitt der Seitenhalbierend, z.B. via $w = \frac{2}{3}$ gilt $\lambda = (0, 0, 1) + w((\frac{1}{2}, \frac{1}{2}, 0) - (0, 0, 1))$.

Bemerkung.

Lemma 3.39 ist sehr praktisch, um zu testen, ob ein Punkt in einem Simplex, auf Seitensimplizes, etc., liegt:

Berechnen von $\lambda(x)$, dann Überprüfen ob $\lambda_i \geq 0$. Falls ja, ist $x \in T$, sonst $x \notin T$. Anzahl $|\{\lambda_i \mid \lambda_i \neq 0\}| - 1$ ist Dimension des Seitensimplex, auf dem x liegt.

Definition 3.40. (Geometrische Maße)

Für Simplex $T = \text{conv}(a_0, \dots, a_s)$ ist der Durchmesser definiert als

$$h_T := \text{diam}(T) = \max_{x,y \in T} \|x - y\|$$

und zudem ist der Inkugeldurchmesser definiert als

$$\rho_T := 2 \sup\{R \in \mathbb{R}_+ \mid \exists x_0 \in T : B_R(x_0) \subseteq T\}.$$

Wir definieren noch eine Größe

$$\sigma_T := \frac{h_T}{\rho_T}.$$

Bemerkung.

$\sigma_T \geq 1$ ist ein Maß für die Degeneriertheit eines Simplex:

- T hat sehr spitze Winkel $\Rightarrow \sigma_T$ groß
- T hat sehr ähnliche Winkel $\Rightarrow \sigma_T$ klein



Wir werden sehen: σ_T ist invariant unter euklidischen Bewegungen & Skalierungen.

Bei Fehleranalyse & Implementierung der FEM werden Operationen häufig auf dem Referenzelement durchgeführt und das Ergebnis auf das gewünschte Simplex transformiert.

Satz 3.41. (Referenzabbildung)

Sei $T = \text{conv}(a_0, \dots, a_s)$ ein s -dimensionales Simplex in \mathbb{R}^d mit Ecken $\{a_i\}_{i=0}^s$ und $\hat{T} \subseteq \mathbb{R}^s$ das Referenzelement, d.h. $\hat{T} = \text{conv}(e_0, e_1, \dots, e_s)$ wobei e_0 der Nullvektor und e_i die kanonischen Einheitsvektoren sind.

- i. Dann gibt es genau eine Matrix $B \in \mathbb{R}^{d \times s}$ mit vollem Spaltenrang und genau ein $t \in \mathbb{R}^d$, sodass es eine eindeutige affine Abbildung

$$F_T: \hat{T} \rightarrow T, \hat{x} \mapsto B\hat{x} + t$$

mit der Eigenschaft

$$\forall i \in \{0, \dots, s\}: \quad F_T(e_j) = a_j$$

existiert.

ii. Für die Norm von B gilt die Abschätzung

$$\|B\|_2 = \sup_{x \in \mathbb{R}^s \setminus \{0\}} \frac{\|B\hat{x}\|_2}{\|\hat{x}\|_2} \leq \frac{h_T}{\rho_{\hat{T}}}.$$

Falls $s=d$ gilt, also B regulär, dann gilt noch:

a) eine Abschätzung für die Norm von B^{-1} , also

$$\|B^{-1}\|_2 \leq \frac{h_{\hat{T}}}{\rho_T}$$

b) eine Formel für den Betrag der Determinante von B , also

$$|\det(B)| = \frac{\lambda^{s-1}(T)}{\lambda^{s-1}(\hat{T})}$$

c) sowie eine Abschätzung für $\det(B)$, also $\exists c, C \in \mathbb{R}_+$ unabhängig von T , s.d.

$$c\rho_T^d \leq |\det(B)| \leq Ch_T^d.$$

Den Beweis dazu überlassen wir als Übung.

Nun führen wir die numerische Aspekte ein:

Definition 3.42. (Triangulierung)

Sei $\Omega \subseteq \mathbb{R}^d$ offen und beschränkt mit polygonalem Rand.

Sei I eine endliche Indexmenge.

Wir nennen $\mathcal{T}_h := \{T_i \subseteq \Omega \mid i \in I\}$ eine zulässige Triangulierung von Ω , g.d.w.

- Für jedes $i \in I$ ist T_i ein d -dimensionales Simplex.
- Überdeckung $\bar{\Omega} = \bigcup_{i \in I} T_i$.
- Nicht-Überlappung $\forall i, j \in I$ mit $i \neq j$: $T_i^\circ \cap T_j^\circ = \emptyset$.
- Konformität $\forall i, j \in I$ mit $i \neq j$: die Menge $S := T_i \cap T_j$ ist entweder eine leere Menge oder ein k -dimensionaler Seitensimplex von T_i und von T_j .

Wir definieren noch die globale Gitterweite oder Feinheit von \mathcal{T}_h

$$h := \max_{i \in I} h_{T_i}$$

den minimalen Inkugelradius

$$\rho := \min_{i \in I} \rho_{T_i}$$

die Ecken- oder Knotenmenge

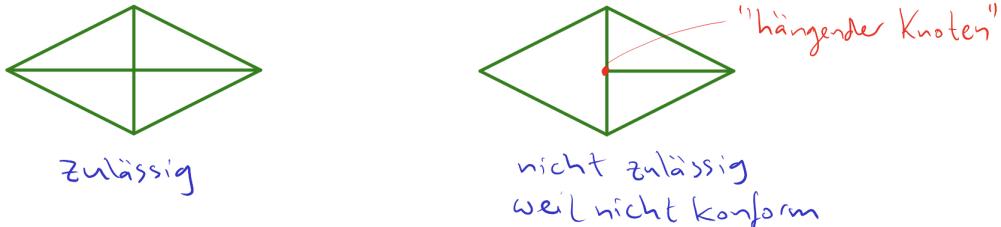
$$\mathcal{E}(\mathcal{T}_h) := \bigcup_{i \in I} \mathcal{E}(T_i)$$

und ein Maß für die Degeneriertheit

$$\sigma := \max_{i \in I} \sigma_{T_i}.$$

Bemerkung.

- Eine zulässige Triangulierung hat also keine „hängenden Knoten“, z.B.



Man kann FEM zwar auch für nichtkonforme (also nicht zulässige) Gitter definieren, dies ist aber technisch aufwändiger.

- Falls Ω nicht polygonalen Rand hat, kann man mittels sog. „isoparametrischer Elemente“ den Rand approximieren, d.h. man erlaubt nicht affine Referenzabbildung, wie z.B.



T ist kein Simplex.

Definition 3.43. (Lokale Polynomielle Räume)

Sei $S \subseteq \mathbb{R}^d$ ein s -dimensionales Simplex, also $s \leq d$, und wir definieren den Raum der polynomiauen Funktionen bis Grad $k \in \mathbb{N}$ auf S durch

$$\mathbb{P}_k(S) := \left\{ p: S \rightarrow \mathbb{R} \mid p(x) = \sum_{\tilde{\beta} \in \mathbb{N}_0^d, |\tilde{\beta}| \leq k} \alpha_{\tilde{\beta}} x^{\tilde{\beta}}, \alpha_{\tilde{\beta}} \in \mathbb{R} \right\}$$

wobei für $x = (x_1, \dots, x_d)^T \in \mathbb{R}^d$ und $\tilde{\beta} \in \mathbb{N}_0^d$ die Größe $x^{\tilde{\beta}} := x_1^{\tilde{\beta}_1} \cdots x_d^{\tilde{\beta}_d}$ bedeutet.

Damit definieren wir die global stetigen, stückweise polynomielien Räumen für \mathcal{T}_h durch

$$\mathbb{P}_k(\mathcal{T}_h) := \{ p \in C^0(\Omega) \mid \forall i \in I: p|_{T_i} \in \mathbb{P}_k(T_i) \}$$

sowie Teilräume mit Nullrandwerten

$$\mathbb{P}_{k,0}(\mathcal{T}_h) := \mathbb{P}_k(\mathcal{T}_h) \cap H_0^1(\Omega).$$

Die Räume $\mathbb{P}_k(\mathcal{T}_h)$ werden nachher als Ansatzräume verwendet.

Lemma 3.44. (Polynom in baryzentrischen Koordinaten)

Sei $T = \text{conv}(a_0, \dots, a_d)$ ein d -dimensionales Simplex in \mathbb{R}^d .

- i. Zu $p \in \mathbb{P}_k(T)$ existiert ein $\bar{p} \in \mathbb{P}_k(\mathbb{R}^{d+1})$ der Gestalt

$$\bar{p}(\lambda) = \sum_{\beta \in \mathbb{N}_0^{d+1}, |\beta| \leq k} d_\beta \lambda^\beta$$

s.d. $p(x) = \bar{p}(\lambda(x))$ für $x \in T$ mit $\lambda: T \rightarrow \mathbb{R}^{s+1}$, $x \mapsto \{\lambda_j\}_{j=0}^s$ den baryzentrischen Koordinaten aus 3.39.

- ii. Zu $\bar{p} \in \mathbb{P}_k(\mathbb{R}^{d+1})$ ist $\bar{p} \circ \lambda = p \in \mathbb{P}_k(T)$.

Beweis.

- i. Sei $p(x) = \sum_{\tilde{\beta} \in \mathbb{N}_0^d, |\tilde{\beta}| \leq k} \alpha_{\tilde{\beta}} x^{\tilde{\beta}} = \sum_{\tilde{\beta} \in \mathbb{N}_0^d, |\tilde{\beta}| \leq k} \alpha_{\tilde{\beta}} \prod_{i=0}^d x_i^{\tilde{\beta}_i}$.
Mit $x = \sum_{j=0}^s \lambda_j a_j$ und $a_j = (a_{ji})_{i=1}^d$ folgt

$$p(x) = \sum_{\tilde{\beta} \in \mathbb{N}_0^d, |\tilde{\beta}| \leq k} \alpha_{\tilde{\beta}} \prod_{i=0}^d \left(\sum_{j=0}^s \lambda_j a_{ji} \right)^{\tilde{\beta}_i} =: \tilde{p}(\lambda)$$

wobei \tilde{p} nur Monome in λ_j der Ordnung höchstens k enthält, daher gilt $p \in \mathbb{P}_k(\mathbb{R}^{d+1})$ und somit $\tilde{p}(\lambda) = \sum_{\beta \in \mathbb{N}_0^{d+1}, |\beta| \leq k} \tilde{d}_\beta \lambda^\beta$ mit geeigneten $\tilde{d}_\beta \in \mathbb{R}$.

Wir sollen noch den konstanten Anteil via $\sum_{j=0}^s \lambda_j = 1$ für jedes $x \in T$ eliminieren, d.h. wir nutzen $\tilde{d}_0 = \tilde{d}_0 \cdot 1 = \tilde{d}_0 \sum_{j=0}^s \lambda_j$ und definiere

$$\bar{p}(\lambda) := \sum_{\tilde{\beta} \in \mathbb{N}_0^d, |\tilde{\beta}| \leq k} d_\beta \lambda^\beta$$

mit

$$d_\beta := \begin{cases} \tilde{d}_\beta, & |\beta| > 1 \\ \tilde{d}_\beta + \tilde{d}_0, & |\beta| = 1. \end{cases}$$

Dann ist nach Konstruktion $p(x) = \tilde{p}(\lambda(x)) = \bar{p}(\lambda(x))$ für jedes $x \in T$.

- ii. Sei $\bar{p}(\lambda) = \sum_{\beta \in \mathbb{N}_0^{d+1}, |\beta| \leq k} d_\beta \lambda^\beta$.

Gemäß Lemma 3.39 ist λ Lösung von

$$M\lambda = \begin{pmatrix} x \\ 1 \end{pmatrix}$$

für $M \in \mathbb{R}^{(d+1) \times (s+1)}$, also

$$\lambda = M^{-1} \begin{pmatrix} x \\ 1 \end{pmatrix} =: (M_1 | m_1) \begin{pmatrix} x \\ 1 \end{pmatrix} =: M_1 x + m_1$$

wobei $M_1 \in \mathbb{R}^{(d+1) \times d}$ und $m_1 \in \mathbb{R}^{d+1}$, also eine affin-lineare Abbildung von x .

Damit gilt aber $p(x) := \bar{p}(\lambda(x)) = \sum_{\beta \in \mathbb{N}_0^{d+1}, |\beta| \leq k} d_\beta (M_1 x + m_1)^\beta \in \mathbb{P}_k(T)$. \square

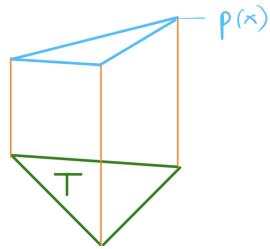
Satz 3.45. (Lineares Finites Element, Courant-Element)

Sei $T = \text{conv}(a_0, \dots, a_d)$ ein d -dimensionales Simplex in \mathbb{R}^d und $p_0, \dots, p_d \in \mathbb{R}$.

Dann existiert genau ein $p \in \mathbb{P}_1(T)$ mit $\forall j \in \{0, \dots, d\}: p(a_j) = p_j$.

Bemerkung.

- Für $d=1$ ist es klar wegen eindeutiger Polynominterpolation (vgl. NUM I).
- Für $d=2$ kann man die Situation schön visualisieren



Anschaulich ist es klar, da eine Ebene sich durch 3 Punkte in allgemeiner Lage bestimmt.

- Der allgemeiner Fall für $d \in \mathbb{N}_{\geq 3}$ ist es nun zu zeigen:

Beweis.

Sei $d \in \mathbb{N}_{\geq 3}$ und $p(x) = c_0 + \sum_{j=1}^d c_j x_j$ also $d+1$ Unbekannte c_0, \dots, c_d und $d+1$ Interpolationsbedingungen $\forall j \in \{0, \dots, d\}: p(a_j) = p_j$.

Wir nutzen die Beobachtung

$$\forall j \in \{0, \dots, d\}: p(a_i) = c_0 + \sum_{j=1}^d c_j x_j = (a_i^T \ 1)^T \begin{pmatrix} c_1 \\ \vdots \\ c_d \\ c_0 \end{pmatrix} = p_i$$

und erhalten damit ein LGS

$$\underbrace{\begin{pmatrix} a_{0,0} & \cdots & a_{0,d} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ a_{d,1} & \cdots & a_{d,d} & 1 \end{pmatrix}}_{=M^T} \begin{pmatrix} c_1 \\ \vdots \\ c_d \\ c_0 \end{pmatrix} = \begin{pmatrix} p_0 \\ p_1 \\ \vdots \\ p_d \end{pmatrix}$$

wobei M der Matrix aus Lemma 3.39 für $s=d$ entspricht.

Dort haben wir gezeigt, dass M vollen Spaltenrang hat.

Also ist M^T regulär, somit existiert $(c_1 \cdots c_d \ c_0)^T$ eindeutig. \square

Bemerkung.

Für Numerik wählt man konkrete lokale Basis $\Phi = \{\varphi_0, \dots, \varphi_d\}$ von $\mathbb{P}_1(T)$, z.B. die nodale Basis zu den Ecken und schreibt $p \in \mathbb{P}_1(T)$ also Linearkombination dieser Basis. Die Funktionen $\{\varphi_j\}_{i=0}^d$ nennt man Formfaktoren (auf Englisch „shape functions.“).

Satz 3.46. (Ex. & Eind. $\mathbb{P}_1(\mathcal{T}_h)$ Interpolationsfunktion)

Sei $\Omega \subseteq \mathbb{R}^d$ offen und beschränkt mit polygonalem Rand, \mathcal{T}_h eine zulässige Triangulierung von Ω mit $m_1 := |\mathcal{E}(\mathcal{T}_h)|$ und Knotenmenge $\mathcal{E}(\mathcal{T}_h) = \{v_i\}_{i=1}^{m_1}$. Seien $p_1, \dots, p_{m_1} \in \mathbb{R}$.

Dann gibt es genau ein $p \in \mathbb{P}_1(\mathcal{T}_h)$ s.d. $\forall i \in \{1, \dots, m_1\}: p(v_i) = p_i$.

Beweis.

Lokale eindeutige Existenz ist klar nach 3.45, d.h. $\forall T \in \mathcal{T}_h$ existiert eindeutiges $p_T \in \mathbb{P}_1(T)$.

Setze $p(x): \Omega \rightarrow \mathbb{R}$ mit $p|_T := p_T$ für $\forall T \in \mathcal{T}_h$.

Damit p wohl-definiert ist, müssen wir zeigen, dass die überlappenden Teile, also am Rand von T , keine Mehrdeutigkeit haben, also ist die globale Stetigkeit von p noch zu zeigen.

Seien T, T' zwei Simplizes mit $S := T \cap T'$ und S einem r -dimensionales Seiten-Simplex von T und T' mit $r \in \mathbb{N}$ (Für $r = 0$ ist nur ein Punkt gemeinsam, und dann ist Stetigkeit klar wegen Interpolationsbedingung in S).

Als ein r -dimensionales Simplex hat S die Eckenmenge $\mathcal{E}(S) \subseteq \{v_i\}_{i=1}^{m_1}$ mit $|\mathcal{E}(S)| = r + 1$.

Betrachte das Referenzsimplex $\hat{S} \subseteq \mathbb{R}^r$.

Mit 3.45 ist ein Interpolationspolynom $\hat{p} \in \mathbb{P}_1(\hat{S})$ durch $r + 1$ Werte eindeutig festgelegt.

Dank Bijektivität $\mathbb{P}_1(\hat{S}) \cong \mathbb{P}_1(S)$ ist also Interpolationspolynom $p|_S \in \mathbb{P}_1(S)$ eindeutig, also $p_T|_S = p|_S = p_{T'}|_S$, daher ist p stetig über Seitensimplex S hinweg. \square

Mit dem Satz folgt sofort Existenz über einer nodalen Basis

Definition 3.47. (Lagrange-Basis, $k = 1$)

Sei $\Omega \subseteq \mathbb{R}^d$ offen und beschränkt mit polygonalem Rand, \mathcal{T}_h eine zulässige Triangulierung von Ω mit $m_1 := |\mathcal{E}(\mathcal{T}_h)|$ und Knotenmenge $\mathcal{E}(\mathcal{T}_h) = \{v_i\}_{i=1}^{m_1}$.

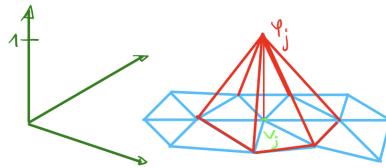
Dann existieren linear unabhängige Funktionen $\varphi_1, \dots, \varphi_{m_1} \in \mathbb{P}_1(\mathcal{T}_h)$ mit

$$\forall i, j \in \{1, \dots, m_1\}: \quad \varphi_i(v_j) = \delta_{ij}$$

und $\{\varphi_i\}_{i=1}^{m_1}$ spannen $\mathbb{P}_1(\mathcal{T}_h)$ auf. D.h. $\dim(\mathbb{P}_1(\mathcal{T}_h)) = m_1$ und $\{\varphi_i\}_{i=1}^{m_1}$ ist eine nodale oder Lagrange-Basis von $\mathbb{P}_1(\mathcal{T}_h)$.

Bemerkung.

- Illustration im Fall $d=2$:



Wir sehen, es ist $\varphi_j(v_i) = \delta_{ij}$ und die „Hütchenfunktion“ ist stückweise linear, global stetig und es gilt $\text{supp } \varphi_j = \bigcup_{i \in \{j \in I \mid v_j \in \mathcal{E}(T_i)\}} T_i$.

- Geringerer Polynomgrad ($\mathbb{P}_0(\mathcal{T}_h)$ anstatt $\mathbb{P}_1(\mathcal{T}_h)$) macht bei Forderung nach Stetigkeit keinen Sinn, denn $\mathbb{P}_0(\mathcal{T}_h) = \{p: \Omega \rightarrow \mathbb{R} \mid \exists c \in \mathbb{R}: p \equiv c\}$ besteht aus global konstanten Funktionen (im Fall Ω zusammenhängend).

Satz 3.48.

Sei $\Omega \subseteq \mathbb{R}^d$ offen und beschränkt mit polygonalem Rand, \mathcal{T}_h eine zulässige Triangulierung von Ω , und $k \in \mathbb{N}$.

Sei $v: \Omega \rightarrow \mathbb{R}$ mit $v|_{T^\circ} \in C^k(T)$ für jedes $T \in \mathcal{T}_h$.

Dann gilt

$$v \in C^{k-1}(\bar{\Omega}) \quad \Rightarrow \quad v \in H^k(\Omega).$$

Bemerkung.

Der Satz impliziert insbesondere $\mathbb{P}_1(\mathcal{T}_h) \subseteq H^1(\Omega)$.

Damit ist dann

$$\mathbb{P}_{1,0}(\mathcal{T}_h) := \{p \in \mathbb{P}_1(\mathcal{T}_h) \mid \forall v_i \in \mathcal{E}(\mathcal{T}_h) \cap \partial\Omega: p(v_i) = 0\} \subseteq H_0^1(\Omega)$$

also kann $V_h := \mathbb{P}_{1,0}(\mathcal{T}_h)$ im Galerkin-Verfahren verwendet werden.

Dies sind genau die sog. „lineare Finite-Elemente“.

Freiheitsgrade sind also nur die Funktionswerte in den „inneren“ Knoten, und wir verwenden häufig die Notationen

$$m_1 := \dim(\mathbb{P}_1(\mathcal{T}_h)) \quad \text{sowie} \quad m_{1,0} := \dim(\mathbb{P}_{1,0}(\mathcal{T}_h))$$

wobei wir noch anmerken, dass $m_{1,0}$ später die Größe der FM-Matrix wird.

Beweis. (von 3.48)

Es reicht, den Fall $k=1$ zu zeigen (Für $k \in \mathbb{N}_{\geq 2}$ folgt die Behauptung durch Betrachtung der Ableitung der Ordnung $k-1$).

Sei $v \in C^\circ(\Omega)$ und für $\forall i \in \{1, \dots, d\}$ definiere $w_i: \Omega \rightarrow \mathbb{R}$ durch

$$w_i|_{T^\circ} := \partial_{x_i} v$$

(und beliebigen Werten auf ∂T).

Für $\phi \in C_0^\infty(\Omega)$ und $n = (n_i)_{i=1}^d$ der äußere Normalenvektor folgt

$$\int_\Omega w_i \phi d\lambda^d = \sum_{T \in \mathcal{T}_h} \int_T \partial_{x_i} v \phi d\lambda^d = \sum_{T \in \mathcal{T}_h} \left(- \int_T v \partial_{x_i} \phi d\lambda^d + \int_{\partial T} v \phi n_i d\lambda^{d-1} \right).$$

Da v global stetig ist, verschwinden die Integrale über innere Seitensimplizies, und das Randintegral verschwindet wegen $\phi|_{\partial\Omega} = 0$.

Somit gilt

$$\int_\Omega w_i \phi d\lambda^d = - \int_\Omega v \partial_{x_i} \phi d\lambda^d$$

also ist w_i die schwache Ableitung von v , und $w_i \in L^2(\Omega)$ wegen Beschränktheit. \square

Definition 3.49. (Lagrange-FEM-Approximation für $k = 1$)

Sei $a(\bullet, \bullet)$ stetig und koerziv, $L(\bullet)$ stetig auf $H_0^1(\Omega)$.

Sei \mathcal{T}_h eine zulässige Triangulierung von Ω mit nodalen Basisfunktionen $\{\varphi_i\}_{i=1}^{m_{1,0}}$ bzgl. inneren Knoten $\mathcal{E}(\mathcal{T}_h) \setminus \partial\Omega =: \{v_i\}_{i=1}^{m_{1,0}}$.

Setze $V_h := \text{span}\{\varphi_i\}_{i=1}^{m_{1,0}} = \mathbb{P}_{1,0}(\mathcal{T}_h)$.

Dann ist $u_h \in V_h$ die FEM-Approximation der PDE g.d.w. gilt

$$\forall v \in V_h: \quad a(u_h, v) = L(v).$$

Wohlgestelltheit folgt aus Lax-Milgram zusammen mit Galerkin-Projektion eines koerziven Problems.

Im Folgenden behandeln wir das Thema **Assemblierung des Systems mittels Quadratur**.

Dazu überlegen uns zunächst

- Für Aufstellen des LGS, d.h. Berechnung von A_h, b_h („Assemblierung“) werden Integrale

$$\int_\Omega (A(x) \nabla \varphi_i) \bullet \nabla \varphi_j d\lambda^d, \quad \int_\Omega f \nabla \varphi_i d\lambda^d, \quad \int_{\Gamma_N} g_N \varphi_i d\lambda^d$$

durch Quadraturen berechnet / approximiert.

- Es reicht, Quadraturen auf Referenzelement zu definieren, welche dann auf beliebige Simplizies transformiert und zu einer zusammengesetzten Quadratur für Gebietsintegral kombiniert werden.

Definition 3.50. (Quadratur auf Referenzelement)

Sei $\hat{T} \subseteq \mathbb{R}^d$ Einheitssimplex, $l \in \mathbb{N}$ Anzahl der Stützstellen für Quadratur, $\hat{x}_1, \dots, \hat{x}_l \in \hat{T}$ sowie $w_1, \dots, w_l \in \mathbb{R}$. Dann heißt

$$\tilde{I}(g) := \sum_{i=1}^l w_i g(\hat{x}_i)$$

Quadratur für $g \in C^0(\hat{T})$.

Schreibe $I(g)$ für das exakte Integral von g auf \hat{T} .

Dann heißt \tilde{I} exakt auf $\mathbb{P}_k(\hat{T})$ oder exakt von Ordnung mindestens k , g.d.w. gilt

$$\forall g \in \mathbb{P}_k(\hat{T}): \quad I(g) = \tilde{I}(g).$$

Bemerkung.

- (**Integraltransformation**) Für ein d -dimensionales Simplex $T \subseteq \mathbb{R}^d$ mit Referenzabbildung

$$F_T: \hat{T} \rightarrow T, \hat{x} \mapsto B\hat{x} + t$$

(vgl. 3.41) gilt

$$\int_T g(x) d\lambda^d(x) = |\det(B)| \int_{\hat{T}} (g \circ F_T)(\hat{x}) d\lambda^d(\hat{x}) \approx |\det(B)| \tilde{I}(g \circ F_T) =: \tilde{I}_T(g)$$

und dabei sieht man leicht

$$\tilde{I}_T \text{ ist exakt auf } \mathbb{P}_k(T) \quad \Leftrightarrow \quad \tilde{I} \text{ ist exakt auf } \mathbb{P}_k(\hat{T}).$$

- (**Integration über Differentialausdrücke**) Wir wissen

$$\varphi \in C^1(T) \quad \Leftrightarrow \quad \hat{\varphi} := \varphi \circ F_T \in C^1(\hat{T})$$

und somit erhalten wir dank der Transformation der Ableitungen

$$\nabla_{\hat{x}} \hat{\varphi}(\hat{x}) = (\mathbf{D}_{\hat{x}} \hat{\varphi})(\hat{x})^T = (\mathbf{D}_x \varphi \cdot \mathbf{D}_{\hat{x}} F_T)^T = (\mathbf{D}_x \varphi \cdot B)^T = B^T \nabla_x \varphi$$

also die sog. „Jacobian inverse transposed“

$$\nabla_x \varphi = (B^T)^{-1} \nabla_{\hat{x}} \hat{\varphi}(\hat{x}).$$

Damit kann man z.B. die Einträge der Steifigkeitsmatrix umschreiben, d.h. für $\forall i, j \in \{1, \dots, m_{1,0}\}$:

$$\begin{aligned} & \int_T (\nabla_x \varphi_i(x))^T A(x) \nabla_x \varphi_j(x) d\lambda^d(x) \\ &= \int_{\hat{T}} (\nabla_{\hat{x}} \hat{\varphi}_i(\hat{x}))^T B^{-1} A(\hat{x}) (B^T)^{-1} \nabla_{\hat{x}} \hat{\varphi}_j(\hat{x}) |\det(B)| d\lambda^d(\hat{x}) \end{aligned}$$

für geeignete $\hat{i}, \hat{j} \in \{1, \dots, n_1\}$ und nodale basis $\{\hat{\varphi}_i\}_{i=1}^{n_1}$ von $\mathbb{P}_1(\hat{T})$ (Wir kümmern uns gleich darum, wie wir die „geeignete $\hat{i}, \hat{j} \in \{1, \dots, n_1\}$ “ finden können).

Beispiel. Sei \hat{T} ein d -dimensionales Referenzelement.

- Zunächst $d \in \mathbb{N}$ also für beliebige Dimension.
Setze $l := 1$ und $\hat{x}_s := \frac{1}{d+1}(1 \ 1 \cdots 1) \in \mathbb{R}^d$ also der Schwerpunkt von \hat{T} .
Dann ist

$$\tilde{I}(g) := \lambda^d(\hat{T}) g(\hat{x}_s)$$

exakt auf $\mathbb{P}_1(\hat{T})$.

Dies ist die Mittelpunktsintegration.

- Für besseres Verhalten sei nun $d = 2$.
Für $i, j \in \{0, 1, 2\}$ setze $\hat{m}_{ij} := \frac{1}{2}(e_i + e_j)$ also Kantenmittelpunkte von \hat{T} .
Dann ist

$$\tilde{I}(g) := \frac{1}{3}\lambda^2(\hat{T}) \sum_{i,j \in \{0,1,2\}, i < j} g(\hat{m}_{ij})$$

exakt auf $\mathbb{P}_2(\hat{T})$.

- Sei weiterhin $d = 2$.

Für \hat{x}_s, \hat{m}_{ij} wie vorhin setzen wir nun

$$\tilde{I}(g) := \frac{1}{60}\lambda^2(\hat{T}) \left(3 \sum_{i=0}^2 g(e_i) + 8 \sum_{i,j \in \{0,1,2\}, i < j} g(\hat{m}_{ij}) + 27g(\hat{x}_s) \right)$$

und dies ist exakt auf $\mathbb{P}_3(\hat{T})$.

Bemerkung. (Quadraturen)

- Gebietsintegrale werden dann einfach approximiert durch zusammengesetzte Quadraturen:

$$\int_{\Omega} g d\lambda^d = \sum_{T \in \mathcal{T}_h} \int_T g d\lambda^d \approx \sum_{T \in \mathcal{T}_h} \tilde{I}_T(g) = \sum_{T \in \mathcal{T}_h} |\det(B_T)| \sum_{i=1}^l w_i g(F_T(\hat{x}_i)).$$

- Ordnung der Quadratur sollte der (noch zu diskutierenden) Konvergenzordnung der FEM angepasst sein: Quadraturordnung sollte hoch genug sein, damit Quadraturfehler nicht Konvergenz der FEM für $h \rightarrow 0$ beeinträchtigt; Quadraturordnung sollte nicht zu hoch sein, um nicht zu viel Rechenzeit in Quadraturen zu verwenden.

Bemerkung. (Assemblierung über lokale Beiträge)

- Direkte Berechnung der FEM Systemmatrix ist sehr treuer.
Beispiel mit Poisson-Problem:

$$\forall i, j \in \{1, \dots, m_{1,0}\}: \quad a_{ij} = \int_{\Omega} (\nabla \varphi_i)^T \nabla \varphi_j d\lambda^d = \sum_{T \in \mathcal{T}_h} \int_T (\nabla \varphi_i)^T \nabla \varphi_j d\lambda^d$$

Mit $|\mathcal{E}(\mathcal{T}_h)| = \mathcal{O}(m_{1,0})$ beträgt der Gesamtaufwand $\mathcal{O}(m_{1,0}^3)$.

Stattdessen nutzt man „Lokalität der Basisfunktion“ & „globale Indexabbildung“.

- (**Globale Indexabbildung**) Den Zusammenhang zwischen lokalen & globalen Indizes stellt eine entsprechende Abbildung sicher:

Sei $n_1 := \dim(\mathbb{P}_1(\hat{T}))$ mit Basis $\{\hat{\varphi}_i\}_{i=1}^{n_1}$.

Sei $\hat{g}: \mathcal{T}_h \times \{1, \dots, n_1\} \rightarrow \{1, \dots, m_1\}$ derart, dass für globale Basis $\{\varphi_i\}_{i=1}^{m_1}$ von $\mathbb{P}_1(\mathcal{T}_h)$ gilt

$$\hat{\varphi}_{\hat{i}}(\hat{x}) = \varphi_i(F_T(\hat{x}))$$

für $i = \hat{g}(T, \hat{i})$, $\hat{i} \in \{1, \dots, n_1\}$ und $T \in \mathcal{T}_h$.

Hier ist \hat{i} der lokale Index und i der globale Index.

Damit ist Kenntnis von $\{\varphi_i\}_{i=1}^{m_1}$ niemals explizit notwendig. Es reicht, eine Basis auf dem Referenzelement zu definieren.

- Umformulieren der Matrixeinträge

$$\begin{aligned} a_{ij} &= \int_{\text{supp}(\varphi_i) \cap \text{supp}(\varphi_j)} (\nabla \varphi_i)^T \nabla \varphi_j \, d\lambda^d \\ &= \sum_{\substack{T \in \mathcal{T}_h, \\ T \subseteq \text{supp}(\varphi_i) \cap \text{supp}(\varphi_j)}} \int_T (\nabla \varphi_i(x))^T \nabla \varphi_j(x) \, d\lambda^d(x) \\ &= \sum_{\substack{T \in \mathcal{T}_h, \\ T \subseteq \text{supp}(\varphi_i) \cap \text{supp}(\varphi_j)}} \int_{\hat{T}} |\det(B)| (\nabla \hat{\varphi}_{\hat{i}}(\hat{x}))^T B^{-1} (B^T)^{-1} \nabla \hat{\varphi}_{\hat{j}}(\hat{x}) \, d\lambda^d(\hat{x}) \\ &=: \sum_{\substack{T \in \mathcal{T}_h, \\ T \subseteq \text{supp}(\varphi_i) \cap \text{supp}(\varphi_j)}} a_{\hat{i}\hat{j}, T} \end{aligned}$$

mit $\hat{i}, \hat{j} \in \{1, \dots, n_1\}$ derart, dass $\hat{g}(T, \hat{i}) = i$ und $\hat{g}(T, \hat{j}) = j$ gilt.

- Statt Schleife über i, j , führen wir Schleife über $T \in \mathcal{T}_h$ aus, um mit Schleife über \hat{i}, \hat{j} lokale Steifigkeitsmatrix $A_{h,T} \in \mathbb{R}^{n_1 \times n_1}$ zu berechnen

$$A_{h,T} := (a_{\hat{i}\hat{j}, T})_{\hat{i}, \hat{j}=1}^{n_1}.$$

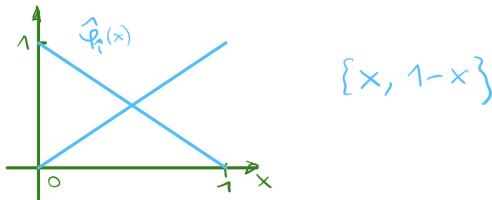
- „Assembliere“ globale Steifigkeitsmatrix A_h durch Addition der lokalen Beiträge in entsprechende Einträge von A_h :

$A_h = 0$
 for $T \in \mathcal{T}_h$:
 calculate $A_{h,T}$
 for $\hat{i}, \hat{j} \in \{1, \dots, n_1\}$:
 $(A_h)_{\hat{g}(T, \hat{i}), \hat{g}(T, \hat{j})} += (A_{h,T})_{\hat{i}, \hat{j}}$.

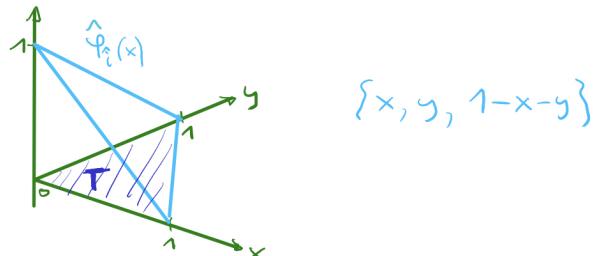
- Gesamtkomplexität beträgt dann $\mathcal{O}(|\mathcal{E}(\mathcal{T}_h)| \cdot n_1^2) = \mathcal{O}(m_{1,0} \cdot n_1^2)$ und dies ist wesentlich schneller als $\mathcal{O}(m_{1,0}^3)$.
- Ähnlich ist Assemblierung der rechten Seite möglich, sogar ohne zusätzliche Schleife über T , d.h. simultan mit A_h wird b_h auch berechnet.
Anders ausgedrückt: „Einzelner Gitterlauf“ reicht zur Assemblierung des Systems A_h, b_h .

Beispiel. Basen auf Referenzelemente für

- $d=1, \hat{T}=[0, 1], n_1=2$



- $d=2, \hat{T}$ Einheitsdreieck, $n_1=3$



Bemerkung. (Struktur von A_h)

Dünnbesetztheit, also sobald $\text{supp}(\varphi_i) \cap \text{supp}(\varphi_j)$ eine Nullmenge in \mathbb{R}^d ist, gilt $(A_h)_{i,j} = 0$. Konkreter:

Sei r maximale Kantenzahl aus einem Knoten, z.B. für $r=5$ sieht es so aus



Dann existieren in jeder Zeile von A_h höchstens $r+1$ Nicht-Null-Einträge.

Dies muss bei Implementierung durch Sparse-Matrizen für Speichereffizienz berücksichtigt werden, insbesondere muss bei Verfeinerung darauf geachtet werden, dass r nicht unbegrenzt wächst.

3.5. FEM-KONVERGENZ/FEHLERANALYSE

Wir schildern zunächst den **groben Plan der Fehleranalyse**:

Gegeben ist $I_h: V \rightarrow V_h$ ein linearer Interpolationoperator mit Approximationsgüte $\|u - I_h u\| \leq C \cdot h^p$ mit möglichst großen p . Dann folgt mit Céa für Galerkin-Projektion

$$\|u - u_h\| \leq \frac{\gamma}{\alpha} \inf_{v \in V_h} \|u - v\| \leq \frac{\gamma}{\alpha} \|u - I_h u\| \leq \frac{\gamma}{\alpha} C \cdot h^p$$

also Konvergenz für $h \rightarrow 0$, und zudem sieht man auch die Konvergenzordnung & Fehlerschranke.

Wir sollen uns einige Überelegung zum linearen Interpolationsoperator machen:

Sei $\Omega \subseteq \mathbb{R}^d$ polygonal berandet, \mathcal{T}_h Triangulierung von Ω und I_h ein Lagrange-Interpolationsoperator.

Damit $I_h: H^m(\Omega) \rightarrow \mathbb{P}_1(\mathcal{T}_h)$ sinnvoll definiert ist, muss $H^m(\Omega)$ Punktauswertungen erlauben, d.h. $v \in H^m(\Omega)$ muss einen stetigen Repräsentanten $\tilde{v} \in C^0(\Omega)$ besitzen, d.h. $\|v - \tilde{v}\|_{H^m(\Omega)} = 0$, damit Punktauswertung von v als Punktauswertung von \tilde{v} definiert werden kann.

Erinnerung:

- Für $d = m = 1$ hat $v \in H^1(\Omega)$ stetigen Repräsentanten.
- Allgemeiner gilt der 2. Sobolev'sche Einbettungssatz:
Für Sobolev-Index $m - \frac{d}{p} > 0$ hat $v \in H^{m,p}(\Omega)$ stetigen Repräsentanten.

Sei im Folgenden d, m derart, dass I_h wohldefiniert ist.

Definition 3.51. (Gebrochene Normen)

Sei \mathcal{T}_h eine Triangulierung von Ω .

Für $m \geq 0$ definieren wir gitterabhängige Normen & Räume

$$H^m(\mathcal{T}_h) := \{v: \Omega \rightarrow \mathbb{R} \mid \forall T \in \mathcal{T}_h: v|_T \in H^m(T)\}$$

mit Seminorm

$$|v|_{H^m(\mathcal{T}_h)} := \sqrt{\sum_{T \in \mathcal{T}_h} |v|_{H^m(T)}^2}$$

und Norm

$$\|v\|_{H^m(\mathcal{T}_h)} := \sqrt{\sum_{T \in \mathcal{T}_h} \|v\|_{H^m(T)}^2}.$$

Bemerkung. Man sieht leicht, dass es gilt

- $H^m(\Omega) \subseteq H^m(\mathcal{T}_h)$.

- $v \in H^m(\Omega) \Rightarrow \|v\|_{H^m(\mathcal{T}_h)} = \|v\|_{H^m(\Omega)}$.
- $H^0(\mathcal{T}_h) = L^2(\mathcal{T}_h) := L^2(\Omega)$.

Satz 3.52. (Interpolationsabschätzung)

Sei \mathcal{T}_h zulässige Triangulierung von $\Omega \subseteq \mathbb{R}^d$ und $m \in \{0, 1, 2\}$.

Seien dazu $\sigma_{\max}, h_{\max} \in \mathbb{R}_+$ s.d. für jedes $T \in \mathcal{T}_h$ gilt $\sigma_T \leq \sigma_{\max}$ und $h_T \leq h_{\max}$.

Sei $I_h: H^2(\Omega) \rightarrow \mathbb{P}_1(\mathcal{T}_h)$ der Lagrange-Interpolationsoperator.

Dann existiert ein $C = C(\Omega, \sigma_{\max}, h_{\max}, m, d)$ s.d. es gilt

$$\forall u \in H^2(\Omega): \|u - I_h u\|_{H^m(\mathcal{T}_h)} \leq C \cdot h^{2-m} |u|_{H^2(\Omega)}. \quad (3.9)$$

Für den Beweis dieses Satzes verweisen wir auf Satz 6.4 in Braess: FEM.

Bemerkung.

Beispielsweise sind die Eigenschaften für $d = 2$ und $m = 2$ schön:

- Mit Sobolev-Index $m - \frac{d}{p} = 2 - \frac{2}{2} = 1 > 0$ ist $I_h: H^2(\Omega) \rightarrow \mathbb{P}_1(\mathcal{T}_h)$ wohldefiniert wegen sinnvoller Punktauswertung.
- Für die Abschätzung gilt dann $\|u - I_h u\|_{H^2(\mathcal{T}_h)} \leq C \cdot h^{2-2} |u|_{H^2(\Omega)} = C |u|_{H^2(\Omega)}$.

Der Satz 3.52 ist die wesentliche Aussage, welche wir auf Gittersequenz anwenden, um Konvergenz von FEM sicherzustellen. Nun definieren wir Gittersequenzen:

Definition 3.53. (Gittersequenz-Charakterisierung)

Eine Folge $\{\mathcal{T}_i\}_{i \in \mathbb{N}}$ von Triangulierungen mit $h_i := h(\mathcal{T}_i)$ und $\lim_{i \rightarrow \infty} h_i = 0$ heißt

- nicht-entartet, g.d.w. es gilt

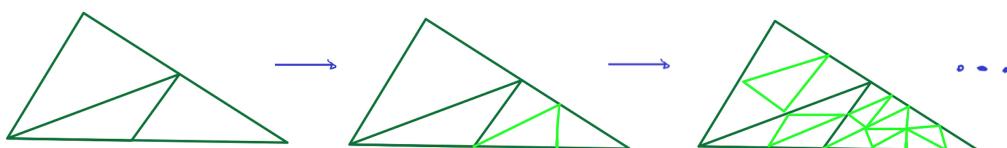
$$\exists \sigma_{\max} \in \mathbb{R}_+ \forall i \in \mathbb{N} \forall T \in \mathcal{T}_i: \sigma_T = \frac{h_T}{\rho_T} \leq \sigma_{\max}$$

- quasi-uniform, g.d.w. es gilt

$$\exists \sigma_{\max} \in \mathbb{R}_+ \forall i \in \mathbb{N} \forall T \in \mathcal{T}_i: \sigma_T = \frac{h_i}{\rho_T} \leq \sigma_{\max}.$$

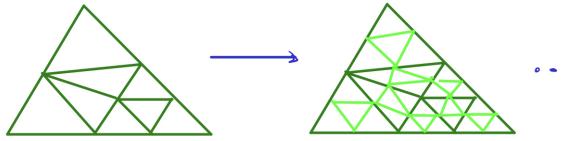
Bemerkung. (Illustration von Def. 3.53.)

- Eine Gittersequenz wie



also zunächst nur rechtes Dreieck verfeinernd, dann alle Dreiecke vierteilend, dann nur rechts, dann alle... ist nicht-entartet, aber nicht quasi-uniform, denn ρ_T fällt schneller ab als h_i .

- Eine Gittersequenz wie



also immer alle Dreiecke vierteilend ist quasi-uniform und auch nicht-entartet, d.h. $\exists C: h_T \leq h_i \leq C \cdot h_T$.

Bemerkung.

- "Quasi-uniform" impliziert „nicht-entartet“.
- Nicht-entartete gitter lassen auch lokal verfeinerte Gitter zu.
- In nicht-entarteten Gittern sind innere Winkel durch $\phi_{\min} > 0$ nach unten beschränkt.

Folgerung 3.54. (Konvergenz der Interpolation)

Sei $\{\mathcal{T}_i\}_{i \in \mathbb{N}}$ eine Folge nicht-entarteter Triangulierungen, $h_{\max} := \max_{i \in \mathbb{N}} h_i$ und $\lim_{i \rightarrow \infty} h_i = 0$.

Für jedes $i \in \mathbb{N}$ sei $I_{h_i}: H^2(\Omega) \rightarrow \mathbb{P}_1(\mathcal{T}_i)$ der Lagrange-Interpolationsoperator.

Dann gilt

$$\forall m \in \{0, 1\} \quad \forall u \in H^m(\Omega): \quad \lim_{i \rightarrow \infty} \|u - I_{h_i} u\|_{H^m(\mathcal{T}_i)} = 0.$$

Beweis.

Es ist $h_i^{2-m} \leq h_{\max}^{1-m} h_i$ und daher $0 \leq \lim_{i \rightarrow \infty} h_i^{2-m} < h_{\max}^{1-m} \lim_{i \rightarrow \infty} h_i \leq h_{\max}^{1-m} \cdot 0 = 0$.

In (3.9) ist C unabhängig vom Index i , daher erhalten wir mit 3.52

$$\lim_{i \rightarrow \infty} \|u - I_{h_i} u\|_{H^m(\mathcal{T}_i)} \leq \lim_{i \rightarrow \infty} C \cdot h_i^{2-m} |u|_{H^2(\Omega)} = C |u|_{H^2(\Omega)} \lim_{i \rightarrow \infty} h_i^{2-m} = 0. \quad \square$$

Satz 3.55. (FEM a-priori Fehlerschranke in H^1)

Sei $\Omega \subseteq \mathbb{R}^d$ beschränkt und polygonal berandet.

Sei $a(\bullet, \bullet)$ koerativ und stetig, $L(\bullet)$ stetig auf $H_0^1(\Omega)$, und $u \in H_0^1(\Omega)$ die eindeutige schwache Lösung der PDE schwacher Form bzgl. $a(\bullet, \bullet)$ und $L(\bullet)$.

Sei \mathcal{T}_h eine zulässige Triangulierung von Ω .

Seien dazu $\sigma_{\max}, h_{\max} \in \mathbb{R}_+$ s.d. für jedes $T \in \mathcal{T}_h$ gilt $\sigma_T \leq \sigma_{\max}$ und $h_T \leq h_{\max}$.

Sei $u_h \in V_h := \mathbb{P}_{1,0}(\mathcal{T}_h)$ die Lagrange-FEM-Lösung.

Falls $u \in H^2(\Omega)$ ist, so existiert ein $\tilde{C} = \tilde{C}(\Omega, \sigma_{\max}, h_{\max}, m, d)$ mit

$$\|u - u_h\|_{H^1(\Omega)} \leq \tilde{C} \cdot h \|u\|_{H^2(\Omega)}.$$

Beweis. Wir erhalten eine Abschätzungskette

$$\|u - u_h\|_{H^1(\Omega)} \leq \frac{\gamma}{\alpha} \inf_{v \in V_h} \|u - v\|_{H^1(\Omega)} \leq \frac{\gamma}{\alpha} \|u - I_h u\|_{H^1(\Omega)} \leq \frac{\gamma}{\alpha} C \cdot h |u|_{H^2(\Omega)} \leq \frac{\gamma}{\alpha} C \cdot h \|u\|_{H^2(\Omega)}$$

wobei die erste Ungleichheit wegen Lemma von Céa, und die dritte Ungleichheit wegen $\|\bullet\|_{H^1(\Omega)} = \|\bullet\|_{H^1(\mathcal{T}_h)}$ und 3.52 mit $m = 1$.

Deswegen gilt die Behauptung mit $\tilde{C} := \frac{\gamma}{\alpha} C$. \square

Folgerung 3.56. (Konvergenz von FEM)

Sei $\{\mathcal{T}_i\}_{i \in \mathbb{N}}$ eine Folge von nicht-entarteter zulässigen Triangulierungen mit $h_i := h(\mathcal{T}_i)$, $\sigma_i := \sigma(\mathcal{T}_i)$, $h_{\max} := \sup_{i \in \mathbb{N}} h_i$, $\sigma_{\max} := \sup_{i \in \mathbb{N}} \sigma_i$ und $\lim_{i \rightarrow 0} h_i = 0$.

Seien die weiteren Voraussetzungen von 3.55 erfüllt.

Dann gilt in $H^1(\Omega)$

$$\lim_{i \rightarrow \infty} u_{h_i} = u.$$

Bemerkung.

Für lineare FEM braucht man die $H^2(\Omega)$ -Regularität, dann gilt Konvergenz erster Ordnung in h bzgl. H^1 -Norm, also

$$\|u - u_h\|_{H^1(\Omega)} \leq C \cdot h \|u\|_{H^2(\Omega)} \leq C \cdot C_R h \|f\|_{L^2(\Omega)}$$

wobei erste Ungleichheit aus 3.55 folgt, die Zweite aus 3.31 Satz von Friedrichs.

Bemerkung. (Hinsicht aus einem numerischen Beispiel)

Bei einem numerischen Beispiel (siehe Vorlesung 28 gegen Ende) kann man tatsächlich Konvergenzordnung 1 bzgl. $\|\bullet\|_{H^1}$ -Fehler beobachten, zudem interessanterweise aber noch Konvergenzordnung 2 bzgl. $\|\bullet\|_{L^2}$ -Fehler. Dies wird in Vorlesung ENUMPDE in WS22/23 behandelt.

Bemerkung. (Method of Lines)

Für zeitabhängige PDEs lassen sich Kapitel 1 & Kapitel 2/3 verbunden.

Bspw. kann man ein Anfangsrandwertproblem

$$\begin{aligned} \partial_t u - \Delta_x u &= f && \text{in } (0, T) \times \Omega \\ u(0, x) &= u_0(x) && x \in \Omega \\ u(t, x) &= g(t, x) && x \in \partial\Omega \end{aligned}$$

„semi-diskretisieren“, also

- Führe eine Ortsdiskretisierung gemäß Kapitel 2/3 mit inneren Knoten $\{x_i\}_{i=1}^n \subseteq \Omega$ durch.

- Finde $U(t) = (U_i(t))_{i=1}^n$ mit $U_i(t) \approx u(t, x_i)$ und $U_i(0) = u_0(x_i)$, d.h. Lösung entlang „Linien“ im Orts-Zeit-Raum

$$\frac{d}{dt}U(t) - A_h U(t) = f_h(t) \quad (3.10)$$

mit geeignetem A_h , z.B. FD-Matrix und f_h Quellterm plus Randwerte.

- (3.10) ist ein ODE-System, das mit Kapitel 1 numerisch gelöst werden kann.
- Stabilitätsbedingungen bzw. Konvergenzordnung ergeben häufig Zusammenhang von Zeit- & Ortsdiskretisierungsparameter τ bzw. h .

Details dazu werden in Vorlesung ENUMPDE in WS22/23 behandelt.

Ende der Vorlesung