Vorlesung *Modellierung und Simulation I*

Prof. Dr. Gillian Queisser

1. Oktober 2013

Inhaltsverzeichnis

1 Gewöhnliche Differentialgleichungen

Dieses Kapitel beschäftigt sich mit den Grundlagen zur Lösung von gewöhnlichen Differentialgleichungen. Die behandelten numerischen Verfahren lassen sich später im Rahmen der zeitlichen Diskretisierung bei partiellen Differentialgleichungen einsetzen.

1.1 Motivation

Bespiel überlegen

1.2 Anfangswertproblem

Definition 1. ***Anfangswertproblem (AWP)*** *Sei $A : D \subset \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$ dann heißt*

$$\frac{\partial u}{\partial t} = A(t, u(t)) \qquad t \in [T_0, T] \tag{1.1}$$
$$u(T_0) = u_0 \tag{1.2}$$

1.3 Einschrittverfahren

Definition 2. ***Einschrittverfahren*** *Ein Einschrittverfahren berechnet iterativ eine diskrete Approximation $u_{\Delta t}(t + \Delta t)$ für $u(t + \Delta t)$ bei bekanntem Startzeitpunkt T_0 und bekannten Anfangswerten u_0 nach folgender Vorschrift:*

$$u_{\Delta t}(T_0) = u_0 \tag{1.3}$$
$$u_{\Delta t}(t + \Delta t) = u_{\Delta t}(t) + \Delta t \cdot \phi(t, u_{\Delta t}(t + \Delta t), u_{\Delta t}(t)) \tag{1.4}$$

hierbei heißt $\phi(t, u_{\Delta t}(t + \Delta t), u_{\Delta t}(t))$ Verfahrensfunktion.

1.3.1 Explizites Eulerverfahren

Wir betrachten eine gewöhnliche Differentialgleichung (ODE: ordinary differential equation) der Art

$$y'(x) = f(x, y(x)) \tag{1.5}$$

Gesucht ist die Lösung $y(x)$ mit vorgegebenem Startwert $y(x_0) = y_0$, d.h. die Steigung am Startpunkt ist gegeben.

Idee 1. *Lokale lineare Approximation der gesuchten Funktion. Gegeben sei x_k, y_k und suche y_{k+1} für x_{k+1}:*

$$y_{k+1} = y_k + hf(x_k, y_k) \tag{1.6}$$

Dieses Verfahren heißt *Eulerverfahren* und ist ein explizites Einschrittverfahren.

Definition 1. *Ein Einschrittverfahren heißt* explizit*, wenn f nur von (x_k, y_k) abhängt und nicht von y_{k+1}. Ein Verfahren heißt* implizit*, wenn f von (x_k, y_k, y_{k+1}) abhängt.*

1.3.2 Implizites Eulerverfahren

1.4 Taylorreihenverfahren

Idee 2. *Verwende mehr Summanden aus der Taylorreihe.*

$$y(x) = y(x_0) + \frac{(x-x_0)}{1!}y'(x_0) + \frac{(x-x_0)^2}{2!}y''(x_0) + \ldots + \frac{(x-x_0)^p}{p!}y^{(p)}(x_0) + R_{p+1}. \tag{1.7}$$

Mit $x_{k+1} - x_k =: h$ folgt

$$y_{k+1} = y_k + \frac{h}{1!}y'_k + \frac{h^2}{2!}y''_k + \frac{h^3}{3!}y_k^{(3)} + \ldots + \frac{h^p}{p!}y_k^{(p)} + R_{p+1}. \tag{1.8}$$

Der spätere Abbruch der Taylorreihe liefert eine bessere lokale Approximation der gesuchten Funktion, die Schwierigkeit besteht jedoch in der Berechnung zusätzlicher höherer Ableitung bis zur p-ten Ordnung.

Beispiel 1. *Betrachte die Funktion*

$$y' = -2xy^2 \tag{1.9}$$

mit $y(0) = 1$. Die Taylorentwicklung von y um x liefert

$$y(x) = y_k + c_1(x - x_k) + c_2(x - x_k)^2 + c_3(x - x_k)^3 + c_4(x - x_k)^4 + \ldots \tag{1.10}$$

mit unbekannten Koeffizienten c_i. Setze in $y' = -2xy^2$ ein, mit $x = x_k + h$:

$$\begin{aligned}
c_1 + 2c_2h + 3c_3h^2 + 4c_4h^3 + \ldots &= -2(x_k + h)\left(y_k + c_1h + c_2h^2 + c_3h^3 + c_4h^4 + \ldots\right)^2 = \\
&= -2(x_k + h)\left(y_k^2 + 2c_1y_kh + (c_1^2 + 2c_2y_k)h^2 + \right. \\
&\quad \left. + (2c_1c_2 + 2c_3y_k)h^3 + \ldots\right) \\
&= -2x_ky_k^2 + (-2y_k^2 - 4c_1x_ky_k)h + \\
&\quad \left(-4c_1y_k - 2x_k(c_1^2 + 2c_2y_k)\right)h^2 + \\
&\quad + \left(-2(c_1^2 + 2c_2y_k) - 4x_k(c_1c_2 + c_3y_k)\right)h^3 + \ldots
\end{aligned}$$

Ein Koeffizientenvergleich liefert

$$c_1 = -2x_k y_k^2 \tag{1.11}$$
$$c_2 = -(y_k + 2c_1 x_k)y_k \tag{1.12}$$
$$c_3 = \frac{-(4c_1 y_k + 2x_k(c_1^2 + 2c_2 y_k))}{3} \tag{1.13}$$
$$c_4 = -\left(\frac{1}{2}c_1^2 + c_2 y_k + x_k(c_1 c_2 + c_3 y_k)\right) \tag{1.14}$$

Bei Betrachtung des Approximationsfehlers $e_k := y(x_k) - y_k$ weist dieses Verfahren eine bessere Näherung als das explizite Eulerverfahren auf.

1.5 Diskretisierungsfehler und Fehlerordnung

Frage 1. *Wie gut sind unsere Approximationen und wie schnell wird die Approximation besser für $h \to 0$?*

Betrachte

$$y_{k+1} = y_k + h\Phi(x_k, y_k, y_{k+1}, h) \tag{1.15}$$

Die Funktion Φ wird *Verfahrensfunktion* genannt. Für unsere bisherigen Verfahren ist Φ gegeben durch

1. Explizites Eulerverfahren: $\Phi(x_k, y_k, h) = f(x_k, y_k)$
2. Implizites Eulerverfahren: $\Phi(x_k, y_k, y_{k+1}, h) = f(x_k, y_k)$
3. Tayloralgorithmus: $\Phi(x_k, y_k, y_{k+1}, h) = c_1 + c_2 h + c_3 h^2 + \ldots + c_p h^{(p-1)}$

1.5.1 Lokaler und globaler Diskretisierungsfehler

Seinen im Folgenden $x_k = x_0 + k \cdot h$ äquidistante Stützpunkte, $y(x_k)$ die exakte Lösung in x_k und y_k die approximierte Lösung in x_k.

Definition 2. *Unter dem* lokalen Diskretisierungsfehler *in x_{k+1} versteht man den Wert*

$$d_{k+1} := y(x_{k+1}) - y(x_k) - h \cdot \Phi(x_k, y(x_k), y(x_{k+1}), h) \tag{1.16}$$

Die Bezeichnung *lokal* wird deshalb gewählt, weil d_{k+1} der Fehler aus einem enzelnen Schritt des Einschrittverfahrens ist und gibt Auskunft über die Qualität der Verfahrensfunktion. In der Praxis ist die folgende Frage wichtig.

Frage 2. *Wie gut ist die berechnete Approximation nach mehreren Schritten?*

Definition 3. *Als* globalen Fehler *g_k in x_k bezeichnet man die Differenz*

$$g_k := y(x_k) - y_k \tag{1.17}$$

Idee 3. *Bringe d_k und g_k in Beziehung und schätze g_k mit Hilfe von d_k ab.*

Voraussetzung: $\Phi(x, y, z, h)$ erfülle die lokale Lipschitz-Bedingung

$$\begin{aligned} |\Phi(x, y, z, h) - \Phi(x, y^*, z, h)| &\leq L\,|y - y^*| && (1.18)\\ |\Phi(x, y, z, h) - \Phi(x, y, z^*, h)| &\leq L\,|z - z^*| && (1.19)\\ & && (1.20) \end{aligned}$$

für $x, y, y^*, z, z^*, h \in B$ und $0 > L < \infty$. Es folgt aus $d_{k+1} := y(x_{k+1}) - y(x_k) - h \cdot \Phi(x_k, y(x_k), y(x_{k+1}), h)$:

$$y(x_{k+1}) = y(x_k) + h\Phi(x_k, y(x_k), y(x_{k+1}), h) + d_{k+1} \tag{1.21}$$

Subtrahiere

$$y_{k+1} = y_k + h\Phi(x_k, y_k, y_{k+1}, h) \tag{1.22}$$

$$\begin{aligned} \Rightarrow g_{k+1} &= g_k + h\,(\Phi(x_k, y(x_k), y(x_{k+1}), h) - \Phi(x_k, y_k, y(x_{k+1}), h) && (1.23)\\ &\quad +\Phi(x_k, y_k, y(x_{k+1}), h) - \Phi(x_k, y_k, y_{k+1}, h) + d_{k+1}) && (1.24) \end{aligned}$$

Wegen der Lipschitzbedingung und $hL < 1$ folgt

$$|g_{k+1}| \leq |g_k| + h\,(L \cdot |y(x_k) - y_k| + L \cdot |y(x_{k+1} - y_k|) + d_{k+1} \tag{1.25}$$

$$\Rightarrow |g_{k+1}| \leq \frac{1 + hL}{1 - hL}|g_k| + \frac{|d_{k+1}|}{1 - hL} \tag{1.26}$$

Für den expliziten Fall gilt:

$$|g_{k+1}| \leq (1 + hL)|g_k| + |d_{k+1}| \tag{1.27}$$

Es existiert ein $K > 0$, sodass $\frac{1+hL}{1-hL} = 1 + hK$. Dann können wir schreiben

$$|g_{k+1}| \leq (1 + a)|g_k| + b \tag{1.28}$$

mit geeigneten Konstanten a, b.

1.5.2 Abschätzen des lokalen Diskretisierungsfehlers

Sei $D \geq 0$ so definiert, dass

$$\max_k |d_k| \leq D \tag{1.29}$$

Hilfssatz 1. *Erfülle g_k die Bedingung*

$$|g_{k+1}| \leq (1 + a)|g_k| + b. \tag{1.30}$$

Dann gilt:

$$|g_n| \leq \frac{(1 + a)^n - 1}{b} + (1 + a)^n|g_0| \leq \frac{b}{a}\,(e^{na} - 1) + e^{na}|g_0| \tag{1.31}$$

Beweis.

$$\begin{aligned} |g_n| &\leq (1+a)|g_{n-1}| + b \leq (1+a)^2|g_{n-2}| + ((1+a)+1)b \\ &\vdots \\ &\leq (1+a)^n|g_0| + \left((1+a)^{n-1} + \ldots + (1+a) + 1\right) b \\ &= \frac{(1+a)^n - 1}{a} \cdot b + (1+a)^n|g_0| \end{aligned}$$

Ferner gilt: e^t konvex $\Rightarrow (1+t) \leq e^t$, $\forall t$.

$$\Rightarrow (1+a)^n \leq e^{na}$$

□

Mit $g_0 = y(x_0) - y_0 = 0$ ergibt sich aus dem Hilfssatz folgender

Satz 1. *Für den globalen Fehler g_n an der Stelle $x_n = x_0 + nh$ gilt für eine explizite Methode*

$$|g_n| \leq \frac{D}{hL}\left(e^{nhL} - 1\right) \leq \frac{D}{hL} \cdot e^{nhL}. \tag{1.32}$$

Für den impliziten Fall gilt

$$|g_n| \leq \frac{D}{hK(1-hL)}\left(e^{nhK} - 1\right) \leq \frac{D}{hK(1-hL)} e^n hK \tag{1.33}$$

Der gobale Fehler hängt also stark von D ab, sowie von K und L.

Beispiel 2. *Für das explizite Eulerverfahren gilt*

$$d_{k+1} = y(x_{k+1}) - y(x_k) - hf(x_k, y(x_k))$$

Entwickle $y(x_{k+1})$ mit der Taylorexpansion

$$y(x_{k+1}) = y(x_k) + hy'(x_k) + \frac{1}{2}h^2 y''(x_k + \Theta h)$$

mit $0 < \Theta < 1$. Wegen $f(x_k, y(x_k)) = y'(x_k)$ folgt

$$d_{k+1} = y(x_k) + hy'(x_k) + \frac{1}{2}h^2 y''(x_k + \Theta h) - y(x_k) - hy'(x_k) = \frac{1}{2}h^2 y''(x_k + \Theta h)$$

Sei

$$M_2 := \max_{x_0 \leq \xi \leq x_n} \left|y''(\xi)\right|$$

$$\Rightarrow \max |d_{k+1}| \leq \frac{1}{2}h^2 M_2$$

Eingesetzt:

$$|g_n| \leq \frac{D}{hL} e^{nhL} = \frac{hM_2}{2L} e^{nhL}$$

D.h. g_n nimmt proportional zu h ab. Das Eulerverfahren besitzt die Fehlerordnung 1.

1.5.3 Fehlerordnung und Konsistenz

Definition 4. *Ein Einschrittverfahren besitzt die Fehlerordnung p, falls für seinen lokalen Diskretisierungsfehler d_k gilt:*

$$\max |d_k| \leq D = C \cdot h^{p+1} = \mathcal{O}(h^{p+1}) \tag{1.34}$$

Dabei ist C eine Konstante.

Folgerung. Der globale Fehler g_n einer expliziten und impliziten Methode ist beschränkt durch

$$|g_n| \leq \frac{C}{L} e^{nhL} h^p = \mathcal{O}(h^p) \tag{1.35}$$

Beispiel 3. *Für die Taylorreihenmethode bis Ordnung p gilt:*

$$\begin{aligned} d_{k+1} &= y(x_{k+1}) - y(x_k) - hy'(x_k) - \frac{h^2}{2!}y''(x_k) - \ldots - \frac{h^p}{p!}y^{(p)}(x_k) \\ &= \frac{h^{p+1}}{(p+1)!}y^{(p+1)}(x_k + \Theta h) \end{aligned} \tag{1.36}$$

für $0 < \Theta < 1$ Daraus ergibt sich für die Taylorreihenmethode die Fehlerordnung p.

Definition 5. *Ein Einschrittverfahren heißt mit der gewöhnlichen Differentialgleichung* konsistent*, falls ihre Fehlerordnung mindestens 1 ist.*

1.6 Verbesserte Methoden

Im Folgenden werden verbesserte Diskretisierungsverfahren für ODEs hergeleitet.

1.6.1 Verbesserte Polygonzugmethode

Idee 4. *Kombiniere zwei Schrittweiten.*

$$y_{k+1}^{(1)} = y_k + hf(x_k, y_k) \tag{1.37}$$

$$y_{k+\frac{1}{2}}^{(2)} = y_k + \frac{h}{2}f(x_k, y_k) \tag{1.38}$$

$$y_{k+1}^{(2)} = y_{k+\frac{1}{2}}^{(2)} + \frac{h}{2}f(x_k + \frac{h}{2}, y_{k+\frac{1}{2}}^{(2)}) \tag{1.39}$$

Durch Anwendung der *Richardson-Extrapolation* folgt:

$$\begin{aligned} y_{k+1} &= 2y_{k+1}^{(2)} - y_{k+1}^{(1)} = 2y_{k+\frac{1}{2}}^{(2)} + hf\left(x_k + \frac{h}{2}, y_{k+\frac{1}{2}}^{(2)}\right) - y_k - hf(x_k, y_k) \\ &= y_k + hf\left(x_k + \frac{h}{2}, y_k + \frac{h}{2}f(x_k, y_k)\right) \end{aligned}$$

Daraus folgt die

Verbesserte Polygonzugmethode

$$
\begin{aligned}
k_1 &:= f(x_k, y_k) && (1.40)\\
k_2 &:= f\left(x_k + \frac{h}{2}, y_k + \frac{1}{2}k_1\right) && (1.41)\\
y_{k+1} &= y_k + hk_2 && (1.42)
\end{aligned}
$$

1.6.2 Trapezmethode

Durch eine äquivalente Integraldarstellung des ursprünglichen ODE-Problems gelangen wir zu der *Trapezmethode*.

Äquivalente Integraldarstellung

Im Folgenden sei die Gleichung

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

zugrunde gelegt. Auf dem Intervall $[x_k, x_{k+1}]$ wird auf beiden Seiten integriert:

$$
\begin{aligned}
\int_{x_k}^{x_{k+1}} y'(x)dx &= \int_{x_k}^{x_{k+1}} f(x, y(x))dx && (1.43)\\
\Leftrightarrow y(x_{k+1}) - y(x_k) &= \int_{x_k}^{x_{k+1}} f(x, y(x))dx && (1.44)
\end{aligned}
$$

Da $y(x)$ unbekannt ist, wird das Integral auf der rechten Seite approximiert durch

$$
\begin{aligned}
\int_{x_k}^{x_{k+1}} f(x, y(x))dx &\approx f(x_k, y_k)(x_{k+1} - x_k) + \\
&\quad + \frac{1}{2}(x_{k+1} - x_k)\left(f(x_{k+1}, y_{k+1}) - f(x_k, y_k)\right)\\
&= \frac{h}{2}\left(f(x_k, y_k) + f(x_{k+1}, y_{k+1})\right)
\end{aligned}
$$

Definition 6. *Die so konstruierte Trapezmethode ist durch die Verfahrensfunktion*

$$y_{k+1} = y_k + \frac{h}{2}\left(f(x_k, y_k) + f(x_{k+1}, y_{k+1})\right) \tag{1.45}$$

definiert.

Da die Trapezmethode ein implizites Verfahren ist, muss eine Näherung für y_{k+1} gefunden werden. Dies kann durch eine Fixpunktiteration geschehen.

Fixpunktiteration.

$$y_{k+1}^{(0)} = y_k + hf(x_k, y_k) \tag{1.46}$$

$$y_{k+1}^{(n+1)} = y_k + \frac{h}{2}\left(f(x_k, y_k) + f(x_{k+1}, y_{k+1}^{(n)}\right) \tag{1.47}$$

Diese Iteration konvergiert gegen einen Fixpunkt y_{k+1}, falls die Lipschitz-Bedingung erfüllt ist und $\frac{hL}{2} < 1$, bzw. $h < \frac{2}{L}$.

Fehlerordnung der Trapezmethode

Die Trapezmethode besitzt die Verfahrensfunktion

$$\Phi(x_k, y_k, y_{k+1}, h) := \frac{1}{2}\left(f(x_k, y_k) + f(x_{k+1}, y_{k+1})\right).$$

Der lokale Diskretisierungsfehler lässt demnach berechnen durch

$$\begin{aligned} d_{k+1} &= y(x_{k+1}) - y(x_k) - \frac{h}{2}\left(f(x_k, y(x_k)) + f(x_{k+1}, y(x_{k+1}))\right) \\ &= y(x_{k+1}) - y(x_k) - \frac{h}{2}\left(y'(x_k) + y'(x_{k+1})\right) \\ &= hy'(x_k) + \frac{h^2}{2}y''(x_k) + \frac{h^3}{6}y'''(x_k) + \mathcal{O}(h^4) \\ &\quad -\frac{h}{2}\left(y'(x_k) + y'(x_k) + hy''(x_k) + \frac{h^2}{2}y'''(x_k) + \mathcal{O}(h^3)\right) \\ &= -\frac{1}{12}h^3 y'''(x_k) + \mathcal{O}(h^4) \end{aligned}$$

Daraus folgt, dass der Hauptteil des lokalen Diskretisierungsfehlers proportional zu h^3 ist und damit hat die Trapezmethode Fehlerordnung 2. Dies ist identisch mit der verbesserten Polygonzugmethode, wir werden aber später sehen, dass die Trapezmethode bessere *Stabilitätseigenschaften* hat.

1.6.3 Verfahren von Heun

In obigem Ansatz bleibt zu entscheiden wie viele Iterationsschritte in der Fixpunktiteration ausgeführt werden. In der Praxis wird oft nur ein Schritt ausgeführt, aufgrund dessen, dass durch das numerische Verfahren ohnehin Näherungen $y_{k+1} \approx y(x_{k+1})$ berechnet werden. Das Verfahren von Heun besteht aus einem Eulerschritt als *Prädiktor* und einem *Korrekturschritt* und wird deshalb auch als Prädiktor-Korrektor-Methode bezeichnet:

$$y_{k+1}^{(p)} = y_k + hf(x_k, y_k) \tag{1.48}$$

$$y_{k+1} = y_k + \frac{h}{2}(f(x_k, y_k) + f(x_{k+1}, y_{k+1}^{(p)}) \tag{1.49}$$

Das algorithmische Vorgehen ist dabei

$$\begin{aligned} k_1 &= f(x_k, y_k) && (1.50)\\ k_2 &= f(x_k + h, y_k + hk_1) && (1.51)\\ y_{k+1} &= y_k + \frac{1}{2}h(k_1 + k_2) && (1.52) \end{aligned}$$

Dies entspricht einer Mittelung der Steigungen k_1 und k_2 in den Punkten (x_k, y_k) und $(x_{k+1}, y_{k+1}^{(p)}$. Die Fehlerordnung den Heun-Verfahrens ist 2, der Beweis ist analog zur Polygonzugmethode zu führen.

1.7 Stabilität

Neben der Feststellung der Konsistenz eines Einschrittverfahrens, stellen wir uns die Frage nach der Konvergenz eines Verfahrens, speziell die

Frage 3. *Welche Anforderung muss an die Zeitschrittweite gestellt werden, um* Konvergenz *des Verfahrens zu garantieren?*

Wir betrachten folgendes

Beispiel 4.

$$u'(t) = \lambda \cdot u(t),\ \lambda < 0$$

Anwendung des expliziten Eulerverfahrens. *Die Verfahrensfunktion lautet*

$$\begin{aligned} \Phi(t, u_h(t), u_h(t+h), h) &= f(t, u_h(t), h) = \lambda \cdot u_h(t)\\ \Longrightarrow \qquad u(t+h) &= u(t) + \lambda \cdot h \cdot u(t) = (1 + \lambda h)u(t)\\ \Longrightarrow \qquad u((k+1) \cdot h) &= (1 + \lambda h)^k \cdot u_0 \end{aligned}$$

Dies bedeutet, dass das Verfahren nur dann konvergiert, wenn

$$\begin{aligned} |1 + \lambda h| &\leq 1\\ \Rightarrow 1 \geq 1 + \lambda h &> -1\\ \Rightarrow h &< -\frac{2}{\lambda} \end{aligned}$$

Diese Anforderung an die Zeitschrittweite heißt Courant-Friedrichs-Levy-Bedingung *und* $h \cdot \lambda$ *die* CFL-Zahl.

Anwendung des impliziten Eulerverfahrens. *Die Verfahrensfunktion lautet*

$$\begin{aligned} \Phi(t, u_h(t), u_h(t+h), h) &= f(t, u_h(t+h)) = \lambda u_h(t+h)\\ \Rightarrow u_h(t+h) &= u_h(t) + hf(t, u_h(t+h))\\ &= u_h(t) + h\lambda u_h(t+h)\\ \Rightarrow u_h(t+h) &= \frac{u_h(t)}{1 - h\lambda} = \frac{1}{1 - h\lambda)^k} u_0 \end{aligned}$$

Dies zeigt, dass implizite Eulerverfahren ist stabil für alle $\lambda < 0$.

Fazit. *Senkenterme* in der Differentialgleichung sollten implizit behandelt werden, *Quellenterme* explizit.

Definition 7. *Folgende Aussagen definieren die* Steifigkeit *eines Problems:*

1. *Ein Problem heißt* steif*, wenn explizite Einschrittverfahren zu kleine Schrittweiten benötigen.*
2. *Ein Problem heißt* steif*, wenn für die Eigenwerte λ der Jacobi-Matrix $D_\lambda f$ der gewöhnlichen Differentialgleichung gilt:*

$$\min_{\lambda_i<0} \lambda_i \ll \max_{\lambda_i<0} \lambda_i$$

Fazit. Explizite Verfahren sind nicht geeignet für steife Probleme.

2 Partielle Differentialgleichungen

In diesem Kapitel werden Herleitungen für partielle Differentialgleichungen (PDEs) unterschiedlicher physikalischer Phänomene gezeigt. Die so motivierten PDEs werden in den späteren Kapiteln numerisch behandelt.

2.1 Gängige Operatoren in mehrdimensionaler Analysis

Im Folgenden werden einige grundlegenden Begriffe und Operatoren definiert, die später in den klassischen Anwendungsfällen, d.h. für klassische partielle Differentialgleichungen, benötigt werden. Dieser Abschnitt liefert also nur einen marginalen Einblick in die mehrdimensionale Analysis, die ohnehin eine Grundvoraussetzung für die Analyse höherdimensionaler Probleme ist.

Definition 8. *Sei $U \subset \mathbb{R}^n$ eine offene Menge und $f : U \to \mathbb{R}$ eine reelle Funktion. f heißt im Punkt $x \in U$* partiell differenzierbar *bzgl. der i-ten Koordinatenrichtung, falls*

$$D_i f(x) := \lim_{h \to 0} \frac{f(x + he_i) - f(x)}{h} \tag{2.1}$$

existiert. Dabei ist $e_i \in \mathbb{R}^n$ der i-te Einheitsvektor

$$e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$$

Definition 9. *Sei $U \subset \mathbb{R}^n$ offen und $f : U \to \mathbb{R}$ partiell differenzierbar. Dann heißt*

$$\text{grad } f(x) := \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x) \right) \tag{2.2}$$

der Gradient *von f im Punkt $x \in U$.*

Satz 2. *Seien $f, g : U \to \mathbb{R}$ zwei partiell differenzierbare Funktionen. Dann gilt*

$$\text{grad } (f \cdot g) = g \cdot \text{grad } f + f \cdot \text{grad } g \tag{2.3}$$

Bemerkung 1. *Statt* grad(f) *wird häufig auch ∇f geschrieben mit*

$$\nabla := \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n} \right)$$

als vektorwertiger Differentialoperator.

Definition 10. *Sei $U \subset \mathbb{R}^n$ eine offene Menge und*

$$V = (v_1, \ldots, v_n) : U \to \mathbb{R}^n$$

eine partiell differenzierbares Vektorfeld (d.h. alle v_i sind partiell differenzierbar). Dann heißt die Funktion

$$\text{div } v := \sum_{i=1}^{n} \frac{\partial v_i}{\partial x_i} \tag{2.4}$$

die Divergenz *des Vektorfeldes v.*

Definition 11. *Sei U eine offene Menge im $\mathbb{R}^3$. Für ein partiell differenzierbares Vektorfeld $v : U \to \mathbb{R}^3$ bezeichnet man*

$$\text{rot } v := \left(\frac{\partial v_3}{\partial x_2} - \frac{\partial v_2}{\partial x_3}, \frac{\partial v_1}{\partial x_3} - \frac{\partial v_3}{\partial x_1}, \frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2} \right) \tag{2.5}$$

als Rotation *von v.*

Bemerkung 2. *Die Rotation von v lässt sich auch als Vektorprodukt*

$$\text{rot } v = \nabla v$$

schreiben.

Definition 12. *Sei $U \to \mathbb{R}^n$ offen und $f : U \to \mathbb{R}$ zweimal stetig partiell differenzierbar. Dann ist der* Laplace-Operator *definiert als*

$$\Delta f := \text{div} \cdot \text{grad } f = \text{div} \cdot \nabla f = \frac{\partial^2 f}{\partial x_1^2} + \ldots + \frac{\partial^2 f}{\partial x_n^2} \tag{2.6}$$

$$\Delta := \frac{\partial^2}{\partial x_1^2} + \ldots + \frac{\partial^2}{\partial x_n^2} = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2} \tag{2.7}$$

2.2 Beispiele partieller Differentialgleichungen

Im vorigen Kapitel wurden gewöhnliche Differentialgleichungen vom Typ

$$\frac{\mathrm{d}u}{\mathrm{d}t} = f(t, u)$$

betrachtet. Die numerische Behandlung dieser Gleichung führte zu unterschiedlichen Zeitschrittverfahren, d.h. der Operator $\frac{\mathrm{d}}{\mathrm{d}t}$ wurde diskret behandelt, wobei $f(t, u)$ eine kontinuierliche Funktion war. Im jetzigen Kontext sei f jedoch keine explizite Funktion, sondern nur über ihre Ortsableitungen definiert. Dies verlangt nach einer diskreten Beschreibung im Ort und führt somit zu Gleichungen mit Differentialoperatoren in Zeit und Ort, also partiellen Differentialgleichungen.

2.2.1 Die Diffusionsgleichung

Sei $c(x,t)$ eine Funktion in Raum und Zeit, welche die Konzentrationsverteilung einer Spezies, z.B. Kalziumionen in einer Nervenzelle, oder die Wärmeverteilung in einem Wärmeleiter beschreibt. Dann ist der Fluss F von $c(x,t)$

$$F = -D \cdot \text{grad } c, \tag{2.8}$$

mit Diffusionskoeffizient D. Durch Forderung der *Massenerhaltung* und Anwendung des *Gauss'schen Integralsatzes* erhält man die Diffusionsgleichung.

Massenerhaltung. Sei V ein beliebiges Volumenelement und c die Konzentration in V. Die Änderung von c in V ist beschrieben durch

$$\int_V \frac{\partial c}{\partial t} d\vec{x}.$$

Unter der Voraussetzung, dass keine Quellen oder Senken in V enthalten sind gilt:

$$-\int_{\partial V} F \cdot \vec{n} dS = \int_V \frac{\partial u}{\partial t} d\vec{x}. \tag{2.9}$$

Satz 3. *Der* Gauss'sche Integralsatz *besagt*

$$\int_V \text{div } F(\vec{x}) d\vec{x} = \int_{\partial V} F(\vec{x}) \cdot \vec{n} dS \tag{2.10}$$

Daraus folgt

$$\begin{aligned} -\int_{\partial V} F \cdot \vec{n} dS &= -\int_V \text{div } F(\vec{x}) d\vec{x} = \int_V \frac{\partial c}{\partial t} d\vec{x} && (2.11)\\ \Longrightarrow -\text{div } F &= \frac{\partial c}{\partial t} && (2.12)\\ \Longrightarrow \frac{\partial c}{\partial t} &= -\text{div } (D \nabla c) \quad \text{(Diffusionsgleichung)} && (2.13) \end{aligned}$$

2.2.2 Die Wellengleichung

Wir betrachten für die Wellengleichung die Größen *Geschwindigkeit* v, *Dichte* ρ und *Druck* p.

1. Es gilt
$$\frac{\partial \rho}{\partial t} = -\rho_0 \text{div } v \tag{2.14}$$
wobei ρ_0 eine feste Dichte definiert. Die Herleitung obiger Gleichung geschieht analog zur Diffusionsgleichung also über die Annahme der Massenerhaltung, gefolgt von der Anwendung des Gauss'schen Integralsatzes.

2. **Newton'sches Gesetz:** Es gilt

$$\rho_0 \frac{\partial v}{\partial t} = -\text{grad } p \tag{2.15}$$

und bedeutet, dass eine räumliche Änderung des Druckfeldes eine Beschleunigung bewirkt.

3. **Zustandsgleichung:** Der Druck p ist bei konstanter Temperatur proportional zur Dichte

$$\Rightarrow p = c^2 \cdot \rho \tag{2.16}$$

$$\Rightarrow \frac{\partial^2}{\partial t^2}\rho \quad = -\rho_0 \text{div } \left(\tfrac{\partial v}{\partial t}\right) = \quad -\text{div } \left(\rho_0 \frac{\partial v}{\partial t}\right) \tag{2.17}$$

$$\Rightarrow \frac{1}{c^2}\frac{\partial^2}{\partial t^2} p \qquad\qquad = \qquad\qquad \text{div (grad } p) \tag{2.18}$$

$$\Leftrightarrow \frac{\partial^2}{\partial t^2} p = c^2 \cdot \text{div (grad } p) = c^2 \Delta p \tag{2.19}$$

Beispiel 5. *Wellengleichung in $\mathbb{R}^1$ und $\mathbb{R}^2$.*

1. *Im $\mathbb{R}^1$ beschreibt die Wellengleichung eine schwingende Saite:*

$$u_{tt} = u_{xx}$$

2. *In $\mathbb{R}^2$ beschreibt die Wellengleichung eine schwingende Membran:*

$$u_{tt} = c^2 \Delta u$$

2.2.3 Poisson- und Potential-Gleichung

Sei $\Omega \subset \mathbb{R}^d$, $d = 2, 3$ und $f : \Omega \to \mathbb{R}$ eine bekannte Ladungsdichteverteilung in Ω. Für die Spannung u gilt

$$-\Delta u = f(x) \quad \text{in } \Omega \tag{2.20}$$

Ein Spezialfall der Poissongleichung ist die *Potentialgleichung*

$$\Delta u = 0 \quad \text{in } \Omega \subset \mathbb{R}^2 \tag{2.21}$$

und beschreibt einen Ladungsfreien Raum.

Beispiel 6. *Lösung der Potentialgleichung auf einer Kreisscheibe mit Radius 1. Betrachte* $\Omega := \{(x, y) \in \mathbb{R}^2; x^2 + y^2 < 1\}$ *und transformiere* x *und* y *in Polarkoordinaten um:*

$$\begin{aligned} x &:= r \cdot \cos\phi \\ y &:= r \cdot \sin\phi \end{aligned}$$

Transformiere die Potentialgleichung in Polarkoordinaten um:

$$\Rightarrow \Delta u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \phi^2} \tag{2.22}$$

Dann erfüllen $r^k \cos k\phi$ *und* $r^k \sin k\phi$ *die Potentialgleichung. Zu wählen ist noch die* Randbedingung *für Radius* $r = 1$*:*

$$u|_{Rand} = u(\cos\phi, \sin\phi) = a_0 + \sum_{k=1}^{\infty}(a_k \cos k\phi + b_k \sin k\phi) \tag{2.23}$$

Dann lautet die Lösung im Innern ($r < 1$)*:*

$$u(x, y) = a_0 \sum r^k \cdot (a_k \cos k\phi + b_k \sin k\phi) \tag{2.24}$$

2.2.4 Poisson-Nernst-Planck Gleichungen

Als ein Beispiel für *gekoppelte* und *nichtlineare* PDEs können die Poisson-Nernst-Planck Gleichungen erwähnt werden. Diese Gleichungen beschreiben den Prozess der *Elektrodiffusion*, also die Diffusion von Teilchen gekoppelt mit einem Konvektionsterm, der durch das zu berechnende elektrische Feld bestimmt ist:

$$\frac{\partial c_i}{\partial t} = \nabla\left(D_i \nabla c_i + D_i \frac{z_i F}{RT} c_i \nabla\Phi\right) \tag{2.25}$$

$$\nabla(\epsilon\nabla\Phi) = -\left(\rho_f + \sum_i c_i z_i F\right) \tag{2.26}$$

Dabei sind die c_i verschiedene Ionenspezies, D_i die zugehörigen Diffusionstensoren, z_i die Teilchenladung, F die Faraday-Konstante, R die allg. Gaskonstante, T die Temperatur und Φ das Potential mit spezifischer Konstante ϵ und statischer Ladungsverteilung ρ_f.

3 Diskretisierung I: Differenzenverfahren für partielle Differentialgleichungen

In diesem Kapitel werden wir ein Approximationsverfahren für das kontinuierliche PDE-Problem herleiten, welches auf der Approximation der Ortsableitungen fundiert. Das als *Differenzenverfahren* bezeichnete Diskretisierungsverfahren wird für ein- und zweidimensionale Fälle hergeleitet. Weiter werden wir Eigenschaften der Systemmatrix des hergeleiteten Gleichungssystems analysieren, die uns am Ende des Kapitels zu einem Konvergenzbeweis für das Differenzenverfahren führen. Die sichtbar werdenden Vor- und Nachteile dieses Verfahrens leiten über in das folgende Kapitel alternativer und allgemeinerer Diskretisierungsverfahren.

Betrachte die Poissongleichung

$$-\Delta u = f \quad \text{auf } \Omega \tag{3.1}$$

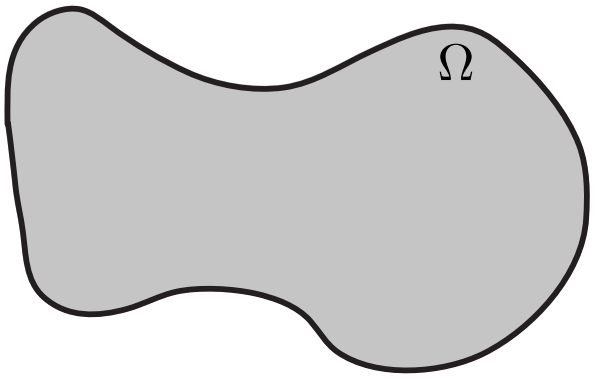

Abbildung 3.1: Kontinuierliches Rechengebiet Ω

Die Poissongleichung ist zunächst auf einem kontinuierlichen Gebiet Ω definiert, d.h. an unendlich vielen Punkten. Dies ist für ein numerisches Verfahren nicht zugänglich.

Idee 5. *Wähle endlich viele Punkte in Ω aus, in denen $-\Delta u = f$ erfüllt ist. Dies führt zu einer Approximation des kontinuierlichen Gebiets, sowie der kontinuierlichen Gleichung.*

3.1 Gebietsdiskretisierung

Betrachte beispielsweise das Einheitsquadrat als kontinuierliches Gebiet

$$\Omega = \{(x, y) : 0 < x < 1, 0 < y < 1\}. \tag{3.2}$$

Das Vorgehen zur Approximation des kontinuierlichen Gebiets, d.h. *Diskretisierung* von Ω ist das Folgende:

1. Überziehe Ω mit einem gleichmäßigen Gitter. Die Menge der Gitterknoten wird als Ω_h bezeichnet. Eine feste Schrittweite h liefert

$$\Omega_h = \left\{(x, y) \in \Omega : \frac{x}{h}, \frac{y}{h} \in \mathbb{Z}\right\}.$$

2. Erfülle die Differentialgleichung in jedem Punkt aus Ω_h.
   - ersetze $u(x)$ durch $u_h(x)$. Dabei ist $u(x)$ die kontinuierliche und $u_h(x)$ die diskrete Lösung.
   - Approximiere die Ableitungen:

$$\lim_{h \to 0} \frac{u(x+h) - u(x)}{h} \approx \frac{u(x+h) - u(x)}{h}$$

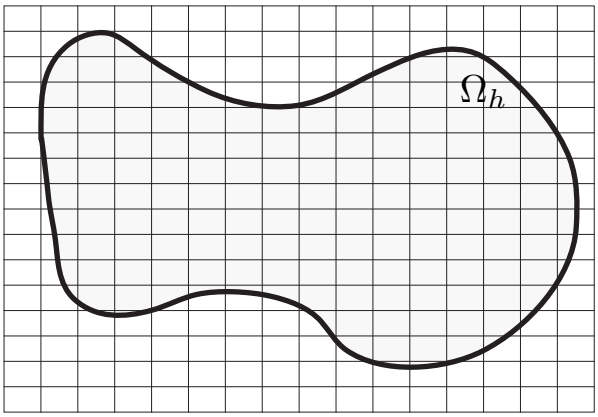

Abbildung 3.2: Diskretisiertes Gebiet Ω_h. Dieses entsteht durch Überziehen des kontinuierlichen Gebiets Ω mit einem Gitter. Die Gitterknoten definieren das endliche diskretisierte Gebiet

3.2 Approximationseigenschaften im $\mathbb{R}^1$

Betrachte das eindimensionale Problem

$$\begin{aligned} u''(x) &= f(x) \quad in\Omega = (0,1) \\ u(0) &= \varphi_0 \\ u(1) &= \varphi_1 \end{aligned}$$

Die Approximation der Ableitung kann auf verschiedene Arten geschehen:

1. **rechtsseitig:** $\delta^+ u(x) = \frac{u(x+h)-u(x)}{h}$
2. **linksseitig:** $\delta^- u(x) = \frac{u(x)-u(x-h)}{h}$
3. **symmetrisch:** $\delta^0 u(x) = \frac{u(x+h)-u(x-h)}{2h}$

Für die zweite Ableitung können links- und rechtsseitige Differenzen δ^+ und δ^- kombiniert werden:

$$u''(x) = \delta^+\delta^- u(x) = \frac{\frac{u(x+h)-u(x)}{h} - \frac{u(x)-u(x-h)}{h}}{h} = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} \tag{3.3}$$

Lemma 1. *Sei $[x-h, x+h] \subset \bar{\Omega}$. Es gilt*

1. $\delta^{\pm} u(x) = u'(x) + hR$ *mit* $|R| \leq \frac{1}{2}\|u''\|_\infty$
2. $\delta^0 u(x) = u'(x) + h^2 R$ *mit* $|R| \leq \frac{1}{6}\|u'''\|_\infty$
3. $\delta^+\delta^- u(x) = u''(x) + h^2 R$ *mit* $|R| \leq \frac{1}{12}\|u^{(4)}\|_\infty$

Beweis.
zu 1.:

$$\begin{aligned} u(x \pm h) &= u(x) \pm hu'(x) + \frac{h^2}{2}u''(x) + \dots \\ &= u(x) \pm hu'(x) + \frac{h^2}{2}u''(\xi), \text{ mit } x \leq \xi \leq x+h \\ \Leftrightarrow \frac{u(x+h)-u(x)}{h} &= u'(x) + \frac{h}{2}u''(\xi) \\ &\leq u'(x) + \frac{h}{2}\|u''\|_\infty \\ &\Rightarrow 1. \end{aligned}$$

zu 2.: Es gelten die Taylorentwicklungen um $x \pm h$:

$$\begin{aligned} u(x+h) &= u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u'''(\xi) \\ u(x-h) &= u(x) - hu'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u'''(\tilde{\xi}) \end{aligned}$$

Subtraktion ergibt:

$$\begin{aligned}
\delta^0 u(x) &= \frac{2hu'(x) + \frac{h^3}{6}(u'''(\xi) + u'''(\tilde{\xi}))}{2h} \\
&= u'(x) + \frac{h^2}{12}(u'''(\xi) + u'''(\tilde{\xi}) \\
&\leq u'(x) + \frac{h^2}{12} \cdot 2\|u'''\|_\infty = u'(x) + \frac{h^2}{6}\|u'''\|_\infty \\
&\quad \Rightarrow 2.
\end{aligned}$$

zu 3.: Betrachte die Entwicklung um $x \pm h$ bis zur Ordnung 4:

$$\begin{aligned}
u(x+h) &= u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u'''(x) + \frac{h^4}{24}u^{(4)}(\xi) \\
u(x-h) &= u(x) - hu'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u'''(x) + \frac{h^4}{24}u^{(4)}(\tilde{\xi})
\end{aligned}$$

Addition der obigen Gleichungen, Subtraktion von $2u(x)$ und Division durch h^2 liefert

$$\begin{aligned}
\delta^+\delta^- u(x) &= \frac{h^2 u''(x) + \frac{h^4}{24}\left(u^{(4)}(\xi) + u^{(4)}(\tilde{\xi})\right)}{h^2} \\
\Rightarrow \delta^+\delta^- u(x) &= u''(x) + \frac{h^2}{24}\left(u^{(4)}(\xi) + u^{(4)}(\tilde{\xi})\right) \\
&\leq u''(x) + \frac{h^2}{12}\|u^{(4)}\|_\infty \\
&\quad \Rightarrow 3.
\end{aligned}$$

□

3.3 Erstellen eines linearen Gleichungssystems

Durch die Approximation der Ableitungen auf einem diskreten Gebiet entsteht ein endliches lineares Gleichungssystem, dass es schlußendlich zu lösen gilt. Wir betrachten weiter die Poisson-Gleichung

$$u''(x) = \Delta u(x) = f(x),$$

und approximieren $\Delta \approx \Delta_h = \delta^+\delta^-$. Wir erhalten also eine diskretisierte Gleichung

$$\delta^+\delta^- u(x) = f(x) + \mathcal{O}(h^2)$$

In Matrix-Vektor-Schreibweise lässt sich obige Gleichung zu dem Beispiel darstellen als

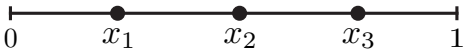

Abbildung 3.3: Beispieldiskretierung eines eindimensionalen Gebiets mit drei inneren Knoten

$$\delta^+\delta^- u(x) = \frac{1}{h^2}\begin{pmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{pmatrix} \cdot \begin{pmatrix} u_h(x_1) \\ u_h(x_2) \\ u_h(x_3) \end{pmatrix} = \begin{pmatrix} f(x_1) - \frac{1}{h^2}u(0) \\ f(x_2) \\ f(x_3) - \frac{1}{h^2}u(1) \end{pmatrix}$$

Im allgemeinen Fall erhalten wir

$$\frac{1}{h^2} \cdot \begin{pmatrix} -2 & 1 & 0 & \dots & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & \dots & 0 \\ & \ddots & & & \ddots & & \\ & & \ddots & & & \ddots & \\ 0 & 0 & \dots & 0 & 1 & -2 & 1 \\ 0 & 0 & \dots & 0 & 0 & 1 & -2 \end{pmatrix} \begin{pmatrix} u_1 \\ \\ \\ \vdots \\ \\ \\ u_n \end{pmatrix} = \begin{pmatrix} f_1 - \frac{u_0}{h^2} \\ f_2 \\ \\ \\ \vdots \\ \\ f_{n-1} \\ f_n - \frac{u_n}{h^2} \end{pmatrix}$$

Wir erhalten also ein Gleichungssystem von der Form

$$K_h \cdot u_h = f_h$$

Bemerkung 3. *K_h ist* dünnbesetzt*, d.h.*

$$\#\{K_{i,j} \neq 0; i, j = 1 \dots n\} = \mathcal{O}(n)$$

3.4 Finite Differenzen in $\mathbb{R}^2$

Wir betrachten nun ein zweidimensionales Gebiet Ω und zwar das kontinuierliche, wie auch diskretisierte Einheitsquadrat

$$\begin{aligned} \Omega &:= \{(x, y) : 0 < x < 1, 0 < y < 1\} \\ \Omega_h &:= \{(x, y) : (x, y) \in \Omega; x/h, y/h \in \mathbb{Z}\} \end{aligned}$$

mit den Gebietsrändern

$$\begin{aligned} \Gamma &:= \{(x, y) : x \in \{0, 1\}, y \in \{0, 1\}\} \\ \Gamma_h &:= \{(x, y) \in \Gamma : x/h, y/h \in \mathbb{Z}\} \end{aligned}$$

Wir betrachten nun das Randwertproblem

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega \\ u &= \varphi \text{ auf } \Gamma \end{aligned}$$

mit der diskreten Form

$$-\Delta_h u_h := (-\delta_x^- \delta_x^+ - \delta_y^- \delta_y^+) u_h(x) \tag{3.4}$$

Definition 13. *Mit u_h wird die* Gitterfunktion *von u bezeichnet, also die Reduktion von u auf das Gitter Ω_h.*

Wendet man den diskreten Laplace-Operator Δ_h auf die Gitterfunktion u_h an, so erhält man

$$\begin{aligned} -\Delta_h u_h &= (-\delta_x^- \delta_x^+ - \delta_y^- \delta_y^+) u_h(x) \\ &= -\frac{1}{h^2} \left(u(x-h,y) + u(x+h,y) + u(x,y-h) + u(x,y+h) - 4u(x,y) \right) \end{aligned}$$

Dies zeigt, dass u an 5 Gitterpunkten ausgewertet wird. Deshalb wird obige Darstellung auch als *Fünfpunktformel* bezeichnet.

3.4.1 Matrix-Vektor-Schreibweise in $\mathbb{R}^2$

Im eindimensionalen Fall existierte eine natürliche Nummerierung der Knoten, welche die Matrixstruktur festlegte (die Ordnung der Knoten wurde stillschweigend verwendet). In $\mathbb{R}^2$ sind unterschiedliche Ordnungen der Knoten denkbar.

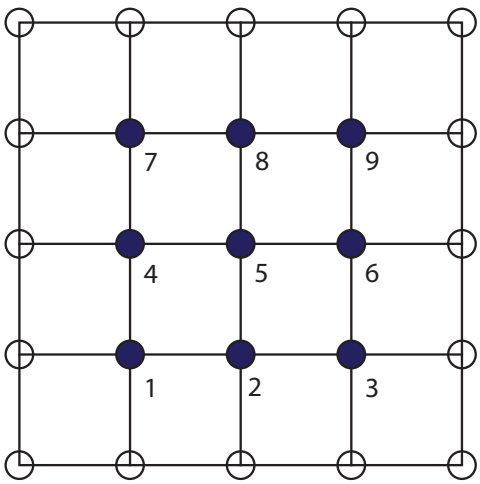

Abbildung 3.4: Lexikographische Nummerierung der inneren Knoten

Lexikographische Knotennummerierung

Die lexikographische Knotennummerierung ist eine zeilenweise Nummerierung der Knoten und definiert eine Matrix der Form

$$\frac{1}{h^2}\begin{pmatrix} 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 \end{pmatrix} \cdot u_h = \tilde{f}_h \tag{3.5}$$

wobei mit $\tilde{f}_h$ die rechte Seite f versehen mit den Einträgen die aus der Randbedingung entstehen bezeichnet wird.
Obige Matrix hat eine Blocktridiagonalstruktur der Form

$$K_h = \frac{1}{h^2}\begin{pmatrix} D & -I & 0 & 0 \\ -I & D & -I & 0 \\ & & \ddots & \\ 0 & 0 & -I & D \end{pmatrix} \tag{3.6}$$

mit

$$D = \begin{pmatrix} 4 & -1 & 0 & \cdots & 0 \\ -1 & 4 & -1 & 0 & \cdots \\ & & \ddots & & \\ & & & \ddots & \\ 0 & \cdots & 0 & 4 & -1 \end{pmatrix}$$

$$-I = \begin{pmatrix} -1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 0 & \cdots & 0 \\ & & \ddots & & \\ & & & \ddots & \\ 0 & \cdots & 0 & 0 & -1 \end{pmatrix}$$

Schachbrettnummerierung

Die Schachbrettnummerierung nummeriert die Knoten in der Schwarz/Weiß-Abfolge eines Schachbretts.
Daraus entsteht folgende Matrixstruktur

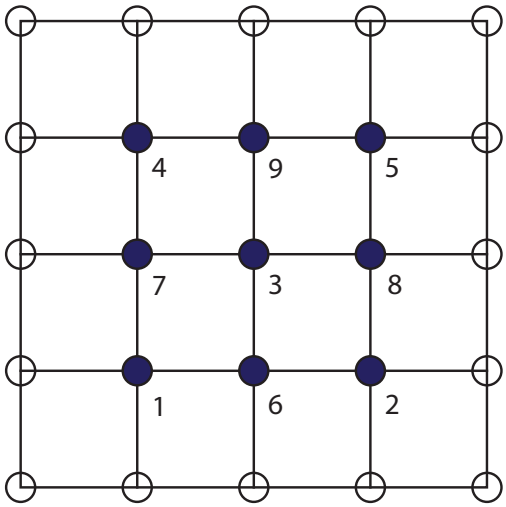

Abbildung 3.5: Schachbrettnummerierung der inneren Knoten

$$\frac{1}{h^2}\begin{pmatrix} 4 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 & -1 & 0 & -1 & 0 \\ 0 & 0 & 4 & 0 & 0 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 4 & 0 & 0 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 4 & 0 & 0 & -1 & -1 \\ -1 & -1 & -1 & 0 & 0 & 4 & 0 & 0 & 0 \\ -1 & 0 & -1 & -1 & 0 & 0 & 4 & 0 & 0 \\ 0 & -1 & -1 & 0 & -1 & 0 & 0 & 4 & 0 \\ 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 4 \end{pmatrix} \cdot u_h = \tilde{f}_h \tag{3.7}$$

Dies definiert eine Matrix mit der Struktur

$$K_h = \begin{pmatrix} D_1 & L \\ L^T & D_2 \end{pmatrix} \tag{3.8}$$

mit den aus K_h ersichtlichen Submatrizen D_i, L und L^T.

Bemerkung 4. *Die Kontennummerierung ändert nichts an den algebraischen Eigenschaften des Systems, kann aber technische Aspekte bei der Implementierung beeinflussen.*

3.4.2 Sternoperatoren

Die Fünfpunktformel definiert die Rechenvorschrift zur Approximation des Laplace-Operators in $\mathbb{R}^2$ über finite Differenzen. Eine vereinfachte Schreibweise hierfür ist die Definition eines Sternoperators.

Definition 14. *Die Fünfpunktformel wird über einen* Fünfpunktstern *folgendermaßen definiert*

$$
\begin{aligned}
-\Delta_h u_h &= \frac{1}{h^2}\left(-u(x-h,y)-u(x+h,y)-u(x,y-h)-u(x,y+h)+4u(x,y)\right) \\
&=: \frac{1}{h^2}\begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix} = -\Delta_h
\end{aligned}
$$

Bemerkung 5. *Der Fünfpunktstern ist keine Matrix, sondern lediglich eine Schablone die auf Ω_h aufgelegt die Rechenvorschrift der Fünfpunktformel vorgibt und damit am Ende auch die Matrixstruktur von K_h. Die Nummerierung der Knoten geht nicht in den Fünfpunktstern ein, da er direkt auf das Gitter aufgelegt wird.*

Weitere Sternoperatoren und Rechenvorschriften

1. in $\mathbb{R}^1$
   a) $\delta^+ = \frac{1}{h} \cdot \begin{bmatrix} 0 & -1 & 1 \end{bmatrix}$
   b) $\delta^- = \frac{1}{h} \cdot \begin{bmatrix} -1 & 1 & 0 \end{bmatrix}$
   c) $\delta^0 = \frac{1}{2h} \cdot \begin{bmatrix} -1 & 0 & -1 \end{bmatrix}$
2. **Allgemeiner Differenzenstern:**

$$
\begin{aligned}
&\frac{1}{h^k}\begin{bmatrix} & \vdots & & \\ \cdots & c_{-1,1}(x,y) & c_{0,1}(x,y) & c_{1,1}(x,y) & \\ & c_{-1,0}(x,y) & c_{0,0}(x,y) & c_{1,0}(x,y) & \cdots \\ & c_{-1,-1}(x,y) & c_{0,-1}(x,y) & c_{1,-1}(x,y) & \\ & & \vdots & & \end{bmatrix} \\
&\qquad = \frac{1}{h^k} \cdot \sum_{i,j} c_{ij} u_h(x+ih, y+jh)
\end{aligned}
$$

3. **Multiplikation von Sternen:** Am Beispiel von zwei eindimensionalen Sternen wird die Multiplikation von Sternen demonstriert.

$$
\begin{aligned}
\begin{bmatrix} a & b & c \end{bmatrix}\begin{bmatrix} d & e & f \end{bmatrix} u_h &= \begin{bmatrix} a & b & c \end{bmatrix} \cdot (d \cdot u_h(x-h) + e \cdot u_h(x) + f \cdot u_h(x+h)) \\
&= a(du_h(x-2h) + eu_h(x-h) + fu_h(x)) \\
&\quad + b(du_h(x-h) + eu_h(x) + fu_h(x+h)) \\
&\quad + c(du_h(x) + eu_h(x+h) + fu_h(x+2h)) \\
&= \begin{bmatrix} ad & ae+bd & af+be+cd & bf+ce & cf \end{bmatrix}
\end{aligned}
$$

3.4.3 Eigenschaften von Differenzensternen

Wir wollen einige typische (aber nicht allgemein gültige) Eigenschaften von Differenzensternen zusammenstellen. Basierend darauf definieren wir danach eine Klasse von Matrizen, die abschließend wichtige Eigenschaften der Systemmatrix der diskretisierten Modellgleichung liefern wird.
Folgende gängige Eigenschaften von Differenzensternen, bzw. der Systemmatrix werden betrachtet:

1. Die Zeilensumme $\sum_{j=1}^{n} c_{ij}$ ist Null.

$$\sum_{j=1}^{n} c_{ij} = 0, \quad \forall i = 1 \dots n$$

2. **Vorzeichenmuster:** Für Δ_h gilt:

$$c_{ii} > 0,\, c_{ij} \leq 0 \quad (i \neq j)$$

   Man beachte das Vorzeichenmuster gilt nicht immer, z.B. für Δ_h^2 ist das Muster nicht erfüllt.

3. **Diagonaldominanz:**
   a) *Schwache* Diagonaldominanz

$$a_{ii} \geq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_i j|$$

   .
   b) *Starke* Diagonaldominanz

$$a_{ii} > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_i j|$$

4. K_h ist *symmetrisch*

Aus diesen Eigenschaften lässt sich eine Klasse von speziellen Matrizen, den sogenannten M-Matrizen, aufbauen.

3.5 M-Matrizen

In diesem Abschnitt wird die Definition einer M-Matrix eingeführt. Anschließend zeigen wir, dass die Systemmatrix K_h M-Matrix Eigenschaften besitzt. Diese werden bei der Analyse von K_h nützlich sein, um abschließend wichtige Matrixeigenschaften zu spezifizieren.

3.5.1 Wiederholung von speziellen Matrixeigenschaften

Definition 15. *Eine Matrix $A \in M(n \times n, K)$ heißt* invertierbar*, wenn es eine Matrix $\tilde{A} \in M(n \times n, k)$ gibt, mit*

$$A \cdot \tilde{A} = \tilde{A} \cdot A = E_n$$

wobei E_n die Einheitsmatrix ist.

Definition 16. *Eine Matrix $A \in M(n \times n, K)$ heißt* regulär*, wenn*

$$\sum_{j=1}^{n} \lambda_i A_j = 0 \Leftrightarrow \lambda_i = 0, \quad \forall i = 1 \dots n$$

wobei A_j die Spaltenvektoren von A sind.

Lemma 2. *Eine* reguläre*, bzw.* nicht-singuläre *Matrix $A \in M(n \times n, K)$ ist invertierbar und lässt sich als endliches Produkt von Elementarmatrizen darstellen.*

Lemma 3. *Für $A \in M(n \times n, K)$ sind folgende Bedingungen äquivalent:*

1. *A ist invertierbar*
2. *A^T ist invertierbar*
3. *Spaltenrang von $A = n$, Zeilenrang von $A = n$*

Insbesondere gilt: $(A^T)^{-1} = (A^{-1})^T$.

3.5.2 Eigenschaften von M-Matrizen

Im Folgenden werden Matrizen $A, B \in M(n \times n, K)$ mit Einträgen $(a_{ij})_{i,j=1\dots n}$ bzw. $(b_{ij})_{i,j=1\dots n}$ betrachtet.

Definition 17. *$A \geq B$ wird komponentenweise definiert, d.h. es gilt $a_{ij} \geq b_{ij}, \quad \forall i, j = 1 \dots n$. Analog lässt sich $A \leq B$, $A < B$ und $A > B$ definieren.*

Definition 18. *Eine $n \times n$-Matrix heißt* M-Matrix*, wenn folgende Eigenschaften erfüllt sind:*

1. Vorzeichenbedingung: $a_{ii} > 0$, $a_{ij} \leq 0 \quad \forall i, j = 1 \dots n$, $i \neq j$
2. *A ist* regulär *und $A^{-1} \geq 0$.*

Frage 4. *Gelten für die Matrix K_h aus dem Modellproblem*

$$\begin{aligned} -\Delta u &= f \\ K_h u_h &= f_h \end{aligned} \tag{3.9}$$

die M-Matrix Eigenschaften?

Die Vorzeichenbedingung lässt sich direkt an K_h ablesen. Zu zeigen bleibt, dass A regulär ist und $A^{-1} \geq 0$.

Definition 19. *Eine Matrix A heißt* irreduzibel*, falls jeder Index $i \in \{1, \ldots, n\}$ mit jedem Index $j \in \{1, \ldots, n\}$* verbunden *ist.*

1. *i ist mit j* direkt verbunden*, wenn $a_{ij} \neq 0$.*
2. *i ist mit j* verbunden*, wenn eine Kette aus direkten Verbindungen i_k, $k = 1 \ldots p$ existiert, sodass*

$$i = i_1, i_2, \ldots, i_p = j$$

   mit $a_{i_{k-1} i_k} \neq 0$.

3. *Eine Alternative Definition ist*

$$A \text{ irreduzibel } \Leftrightarrow \text{ Es existiert keine Permutation } \Pi$$

   mit

$$\Pi^T A \Pi = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$$

3.5.3 Abschätzen der Eigenwertbereiche einer Matrix

Wir werden das Kriterium von Gerschgorin in zwei Fassungen beweisen. Dieses Kriterium wird uns Auskunft über die Eigenwertbereiche geben.

Kriterium von Gerschgorin

Gegeben sei eine Matrix $A = (a_{ij}) \in \mathbb{C}^{n \times n}$. Betrachte dazu

$$\bar{K}_i := \{z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|\},\ i = 1 \ldots n$$

Satz 4. *Die n Eigenwerte von A liegen in der Vereinigung*

$$\bigcup_{i=1}^{n} \bar{K}_i.$$

Beweis.
Sei λ Eigenwert von A und v Eigenvektor von A.
oBdA. kann $\|v\|_\infty := \max_{i=1}^{n} |v_i| = 1 = |v_r|$ angenommen werden. Es gilt

$$(A - \lambda E_n)v = 0$$

Für die r-te Zeile folgt

$$(a_{rr} - \lambda)v_r = - \sum_{i=1, i \neq r}^{n} a_{ri} v_i$$

Anwendung der Dreiecksungleichung ergibt

$$\begin{aligned}
|a_{rr} - \lambda| &= |(a_{rr} - \lambda)v_r| = \left| \sum_{i=1, i \neq j}^{n} a_{ri} v_i \right| \\
&\leq \sum_{i=1, i \neq j}^{n} |a_{ri}||v_i| \leq \sum_{i=1, i \neq j}^{n} |a_{ri}| \\
\Rightarrow \quad & |a_{rr} - \lambda| \leq \sum_{i=1, i \neq j}^{n} |a_{ri}| \\
\Leftrightarrow \quad & \lambda \in \bar{K}_r = \{z \in \mathbb{C} : |z - a_{rr}| \leq \sum_{i=1, i \neq r}^{n} |a_{ri}|\}
\end{aligned}$$

Damit ist gezeigt, dass jeder Eigenwert in der Vereinigung aller Kreisscheiben $\bar{K}_i$, $i = 1 \ldots n$ liegt. Wendet man das Kriterium von Gerschgorin auf die Systemmatrix K_h an, dann folgt:

1. $\sum_{i=1, i \neq j}^{n} a_{ij} = 4$, $\forall j \Rightarrow$ Radius für alle Kreise ist 4.
2. $a_{jj} = 4$

$$\Longrightarrow \lambda \in \left[0, 8h^{-2}\right]$$

Im nächsten Schritt wollen wir zeigen, dass die Eigenwerte von K_h echt größer Null sind, d.h. wir wollen zeigen

$$\lambda \in \left(0, 8h^{-2}\right)$$

Dazu beweisen wir die

Verschärfung des Kriteriums von Gerschgorin

Unter der Voraussetzung, dass A irreduzibel ist, gilt:

$$\lambda \in \left(\bigcup_{r=1}^{n} K_r \right) \cup \left(\bigcap_{r=1}^{n} \partial K_r \right)$$

mit

$$
\begin{aligned}
K_r &:= \{z \in \mathbb{C} : |z - a_{rr}| < \sum_{i=1, i\neq r}^{n} |a_{ri}|\} \\
\partial K_r &:= \{z \in \mathbb{C} : |z - a_{rr}| = \sum_{i=1, i\neq r}^{n} |a_{ri}|\}
\end{aligned}
\tag{3.10}
$$

Beweis. λ sei ein Eigenwert von A mit dem zugehörigen Eigenvektor v, mit $\|v\|_\infty = 1$. Falls $\lambda \in \bigcup_{r=1}^{n} K_r \Rightarrow$ Verschärfte Fassung von Gerschgorin bereits erfüllt. Sei also $\lambda \notin \bigcup_{r=1}^{n} K_r$ sondern in der Vereinigung der Ränder.

Hilfsbehauptung. Sei $a_{rj} \neq 0$ (also r direkt verbunden mit j). Dann folgt

$$
\begin{aligned}
|v_r| &= 1 \text{ und } |\lambda - a_{rr}| = \sum_{i=1, i\neq r}^{n} |a_{ri}| \\
\Rightarrow |v_j| &= 1 \text{ und } |\lambda - a_{jj}| = \sum_{i=1, i\neq j}^{n} |a_{ji}|
\end{aligned}
$$

Der Satz von Gerschgorin beweist die Existenz eines $r \in \{1, \ldots, n\}$ mit $|v_r| = 1$ und $|\lambda - a_{rr}| \leq \sum_{i=1, i\neq r}^{n} |a_{ri}|$. Da λ auf dem Rand liegt (siehe Annahme) lässt sich der Hilfssatz auf λ anwenden.
Da A irreduzibel $\Longrightarrow$ Für jedes $j \in \{1, \ldots, n\}$ existiert eine Verbindung

$$
r = i_0, i_1, \ldots, i_k = j \text{ mit } a_{i_{p-1} i_p} \neq 0
$$

Die Hilfsbehauptung liefert

$$
|v_{i_p}| = 1 \text{ und } |\lambda - a_{i_p i_p}| = \sum_{i=1, i\neq i_p}^{n} |a_{i_p i_p i}| \quad \forall p = 0, \ldots, k.
$$

Insbesondere gilt $|v_j| = 1$ und $|\lambda - a_{jj}| \in \partial K_j$. j wurde dabei beliebig gewählt.
$\Longrightarrow \lambda$ muss auf jedem Rand $\partial K_i \, \forall i = 1, \ldots, n$ liegen.
$\Longrightarrow \lambda \in \bigcap_{i=1}^{n} \partial K_i$.
$\Longrightarrow$ Behauptung

□

Beweis der Hilfsbehauptung. Aus $|\lambda - a_{rr}| \leq \sum_{i=1, i\neq r}^{n} |a_{ri}|$ folgt wegen der Annahme $|\lambda - a_{rr}| = \sum_{i=1, i\neq r}^{n} |a_{ri}|$ und $|v_r| = 1$. Aus der Dreiecksungleichung folgt

$$
\sum_{i=1, i\neq r}^{n} |a_{ri}||v_i| = \sum_{i=1, i\neq r}^{n} |a_{ri}|
$$

Dies gilt insbesondere für $i = j$ und da $|v_j| \leq 1$ muss summandenweise gelten:

$$\begin{aligned} |a_{rj}||v_j| &= |a_{rj}| \text{ und } |a_{rj}| \neq 0 \\ \Rightarrow |v_j| &= 1 \end{aligned}$$

□

Frage 5. *Was folgt mit den Kriterien von Gerschgorin für die Matrix K_h?*

Es treten die folgenden 3 Fälle auf:

$$\begin{aligned} a_{ii} &= \frac{4}{h} \text{ und} \\ r_i &= \sum_{j=1, j\neq i}^{n} |a_{ij}| = \begin{cases} 2/h^2 & \text{für die Eckpunkte} \\ 3/h^2 & \text{für die Seitenpunkte} \\ 4/h^2 & \text{für die inneren Punkte} \end{cases} \end{aligned}$$

Da K_h symmetrisch und reell ist, sind alle Eigenwerte von K_h reell und liegen auf dem reellen Zahlenstrahl. Die Kreisränder ∂K_j haben den gleichen Mittelpunkt mit unterschiedlichen Radien und sind deshalb disjunkt aus dem verschärften Kriterium von Gerschgorin folgt deshalb

$$\lambda \in \left(0, \frac{8}{h^2}\right)$$

3.5.4 Zusammenhang zwischen M-Matrix und Spektralradius

Zunächst definieren wir die Begriffe *Diagonaldominanz*, *Irreduzibilität* und den *Spektralradius*. Anschließend wird der Zusammenhang zwischen M-Matrix und Spektralradius analysiert.

Definition 20. *(Diagonaldominanz)*

1. *A heißt* diagonaldominant*, falls*

$$\sum_{j, j\neq i} |a_{ij}| < |a_{ii}| \quad \forall i = 1, \ldots, n \tag{3.11}$$

2. *A heißt* irreduzibel diagonaldominant*, falls A irreduzibel ist und*

$$\sum_{j, j\neq i} |a_{ij}| < |a_{ii}| \quad \text{für ein } i \tag{3.12}$$

$$\sum_{j, j\neq i} |a_{ij}| \leq |a_{ii}| \quad \forall i = 1, \ldots, n \tag{3.13}$$

Bemerkung 6. *Aus* irreduzibel *und* diagonaldominant *folgt* irreduzibel diagonaldominant*, die Rückrichtung gilt jedoch nicht.*

Definition 21. *(Spektralradius)*
Der Spektralradius $\varrho(A)$ *einer Matrix* A *ist definiert als*

$$\varrho(A) := \max\{|\lambda| : \lambda \text{ ist Eigenwert von } A\}. \tag{3.14}$$

Satz 5. *Folgende Aussagen gelten für die Matrix* $D^{-1}B$ *mit* $A = D - B$, $D := diag\{a_{ii} : i = 1, \ldots, n\}$ *und* $B := D - A$*:*

1. *A sei diagonaldominant oder irreduzibel diagonaldominant*

$$\Rightarrow \varrho(D^{-1}B) < 1.$$

2. *A erfülle die Vorzeichenbedingung. Dann gilt*

$$A \text{ ist M-Matrix } \Leftrightarrow \varrho(D^{-1}B) < 1.$$

Beweis von 1. Betrachte $C := D^{-1}B$ mit

$$\begin{aligned} c_{ij} &= -\frac{a_{ij}}{a_{ii}}, \quad (i \neq j) \\ c_{ii} &= 0 \end{aligned}$$

1. Sei Diagonaldominanz vorausgesetzt. Dann gilt:

$$r_i := \sum_{j=1, j\neq i}^{n} |c_{ij}| < 1 \quad \forall i = 1, \ldots, n.$$

Aus dem Kriterium von Gerschgorin folgt:

$$\begin{aligned} \lambda \in \bigcup_{i=1}^{n} \bar{K}_{r_i}(c_{ii}) &= \bigcup_{i=1}^{n} \bar{K}_{r_i}(0) \\ \Rightarrow |\lambda| &\leq \max_{i=1\ldots n} r_i < 1 \\ \Rightarrow \varrho(C) &= \varrho(D^{-1}B) < 1 \end{aligned}$$

□

2. Sei nun die irreduzible Diagonaldominanz vorausgesetzt:

$$A \text{ irreduzibel diagonaldominant} \Rightarrow \begin{array}{ll} r_j \leq 1 & \forall j = 1, \ldots, n \\ r_j < 1 & \text{für mindestens ein } i \end{array}$$

Aus der scharfen Version von Gerschgorin folgt

$$\lambda \in \bigcup_{j=1}^{n} K_{r_j}(0) \cup \left(\bigcap_{j=1}^{n} \partial K_{r_j}(0) \right)$$

Zu zeigen ist nun $\bigcup_{j=1}^{n} K_{r_j}(0) \cup \left(\bigcap_{j=1}^{n} \partial K_{r_j}(0)\right) \subset K_1(0)$.
Fall 1: Alle $r_j = r$ sind gleich. Da es mindestens ein i gibt mit $r_i < 1$, folgt

$$\bigcap_{j=1}^{n} \partial K_{r_j}(0) = \partial K_r(0) \subset K_1(0) \quad \surd$$

Fall 2: Falls nicht alle r_j gleich sind, dann ist $\bigcap_{j=1}^{n} \partial K_{r_j}(0) = \emptyset$ (da alle Kreise den Ursprung 0 haben).
$\Rightarrow$ Behauptung.

Beweis von 2. A ist genau dann eine M-Matrix, wenn $\varrho(D^{-1}B) < 1$ und die Vorzeichenbedingung gilt. Zu zeigen ist also: A ist nicht-singulär und $A^{-1} \geq 0 \Leftrightarrow \varrho(D^{-1}B) < 1$. Wir zeigen zunächst die

Rückrichtung. Sei $\varrho(D^{-1}B) = \varrho(C) < 1$. Dann konvergiert die Neumann-Reihe

$$S := \sum_{\nu=0}^{\infty} C^{\nu} = (I - C)^{-1}$$

$$\begin{aligned} \Rightarrow S(I - C) &= I \\ \Leftrightarrow SD^{-1}(D - B) &= SD^{-1}A = I \\ \Rightarrow A^{-1} = SD^{-1} & \end{aligned}$$

Da $D^{-1} \geq 0$, $B \geq 0 \Rightarrow C \geq 0$, $C^{\nu} \geq 0$ und $S \geq 0$.
$\Rightarrow A^{-1} \geq 0 \Rightarrow$ 2. Bedingung für M-Matrix.
$\Rightarrow A$ ist eine M-Matrix.

Hinrichtung. Sei A eine M-Matrix. λ sei Eigenwert zu dem Eigenvektor $u \neq 0$ von $D^{-1}B$. Es gilt

$$|\lambda| \cdot |u| = |\lambda u| = |D^{-1}Bu| \leq D^{-1}B|u|$$

Ferner gilt: $A^{-1} \geq 0$ und $D \geq 0 \Rightarrow A^{-1}D \geq 0$.

$$\begin{aligned} \Rightarrow -A^{-1}DD^{-1}B|u| &\leq -A^{-1}D|\lambda||u| \\ \Rightarrow |u| = A^{-1}(D - B)|u| &= A^{-1}D(I - D^{-1}B)|u| \\ &\leq A^{-1}D|u| - A^{-1}D|\lambda||u| = (1 - |\lambda|)A^{-1}D|u| \end{aligned}$$

Wäre nun $|\lambda| \geq 1 \Rightarrow |u| - (1 - |\lambda|)A^{-1}D|u| \leq 0$. Da

$$1 - (1 - |\lambda|)A^{-1}D \geq 0$$

folgt $|u| \leq 0 \Rightarrow u = 0$. Dies ist ein Widerspruch zur Annahme.

□

Satz 6. *Eine irreduzible M-Matrix A hat eine positive Inverse*

$$A^{-1} > 0.$$

Beweis. Seien $\alpha, \beta \in \{1, \ldots, n\}$. Da A irreduzibel ist, existiert eine Verbindung

$$\alpha = \alpha_0, \alpha_1, \ldots, \alpha_k = \beta$$

$$\begin{aligned} C &:= D^{-1}B \text{ und } c_{\alpha_{i-1}\alpha_i} > 0 \\ \Rightarrow (C^k)_{\alpha\beta} &= \sum_{\gamma_1,\ldots,\gamma_k} c_{\alpha\gamma_1} c_{\gamma_1\gamma_2} \ldots c_{\gamma_{k-1}\beta} \geq c_{\alpha\alpha_1} c_{\alpha_1\alpha_2} \cdot \ldots \cdot c_{\alpha_{k-1}\beta} > 0 \end{aligned}$$

Wie oben bewiesen gilt $\varrho(C) < 1$.
$\Rightarrow S := \sum_{\nu=0}^{\infty} C^\nu$ konvergiert. Da $S_{\alpha\beta} \geq (C^k) > 0$ ist S durch $C^k > 0$ beschränkt.
$\Rightarrow S > 0$. Mit $A^{-1} = SD^{-1} > 0$ folgt $A^{-1} > 0$

□

3.6 Eigenschaften der Systemmatrix der Poisson-Gleichung

Nachdem verschiedene Matrixnormen eingeführt werden, wenden wir uns in diesem Abschnitt den Eigenschaften der Matrix K_h aus der Poisson-Gleichung zu.

Definition 22. *(Vektornorm)*
V sei ein Vektorraum über K (= $\mathbb{R}$ oder $\mathbb{C}$). $\|.\|$ heißt Norm *in V, falls gilt*

$$\begin{aligned} \|u\| &= 0 \Longleftrightarrow u = 0 && (3.15) \\ \|u+v\| &\leq \|u\| + \|v\| \quad \forall u, v \in V && (3.16) \\ \|\lambda u\| &= |\lambda| \|u\| \qquad \lambda \in K, u \in V && (3.17) \end{aligned}$$

Definition 23. *(Matrixnorm)*
V sei ein Vektorraum versehen mit einer Vektornorm $\|.\|$. Dann ist

$$\|A\|_M := \sup_{u \in V} \left\{ \frac{\|Au\|}{\|u\|} : 0 \neq u \in V \right\} \tag{3.18}$$

die der Vektornorm $\|.\|$ zugeordnete Matrixnorm.

Lemma 4. *Es gilt*

$$\|A\|_M \geq \varrho(A) \tag{3.19}$$

Beweis. Mit $\|A\|_M = \sup_{u \in V} \left\{ \frac{\|Au\|}{\|u\|} : 0 \neq u \in V \right\}$ und einem Eigenvektor v von A gilt:

$$\begin{aligned} \frac{\|Av\|}{\|v\|} &= \frac{\|\lambda v\|}{\|v\|} = |\lambda| \\ \Rightarrow \sup \left\{ \frac{\|Au\|}{\|u\|} \right\} &\geq \max_{v \in V} \frac{\|Av\|}{\|v\|} = |\lambda^*| = \varrho(A) \end{aligned}$$

□

3.6.1 Gebräuchliche Matrixnormen und positiv definite Matrizen

Zeilensummennorm

Die zur Maximumsnorm $\|.\|_\infty$ zugehörige Matrixnorm

$$\|A\|_\infty := \max_{\alpha\in\{1,\ldots,n\}} \left\{ \sum_{\beta\in\{1,\ldots,n\}} |a_{\alpha\beta}| \right\} \tag{3.20}$$

heißt *Zeilensummennorm.*

$$\Longrightarrow \qquad \|Au\|_\infty = \| \sum_j a_{ij} u_{ji} \|_\infty \tag{3.21}$$

für ein i so, dass $\sum_j |a_{ij}|$ maximal.

Bemerkung 7. *Aus $A \geq B$ folgt $\|A\|_\infty \geq \|B\|_\infty$, da $A \geq B$ komponentenweise definiert ist.*

Für einen späteren Beweis werden wir folgenden Satz benötigen:

Satz 7. *Sei A eine M-Matrix. Existiert ein Vektor w mit $Aw \geq \mathbb{1}$, dann gilt*

$$\|A^{-1}\|_\infty \leq \|w\|_\infty \tag{3.22}$$

Beweis. Es gilt

$$|u| \leq \|u\|_\infty \cdot \mathbb{1} \leq \|u\|_\infty \cdot Aw$$

für

$$|u| := \begin{pmatrix} |u_1| \\ |u_2| \\ \vdots \\ |u_i| \end{pmatrix}$$

Da A eine M-Matrix ist, gilt $A^{-1} \geq 0$.

$$\begin{aligned} |A^{-1}u| &\leq A^{-1}|u| \leq A^{-1}\|u\|_\infty Aw \\ &= \|u\|_\infty A^{-1}Aw = \|u\|_\infty \cdot w \\ \Rightarrow \frac{|A^{-1}u|}{\|u\|_\infty} &\leq w \text{ und speziell } \frac{\|A^{-1}u\|_\infty}{\|u\|_\infty} \leq \|w\|_\infty \\ \Rightarrow \|A^{-1}\|_\infty &\leq \|w\|_\infty \end{aligned}$$

□

Satz 8. *Seien A und B M-Matrizen mit $B \geq A$. Dann gilt:*

$$0 \leq B^{-1} \leq A^{-1} \text{ und } \|B^{-1}\|_\infty \leq \|A^{-1}\|_\infty \tag{3.23}$$

Beweis. Es gilt

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}.$$

Da A, B M-Matrizen sind, folgt $A^{-1} \geq 0$ und $B^{-1} \geq 0$ und mit $B \geq A$ folgt $B - A \geq 0$.

$$\Rightarrow A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1} \geq 0$$

$$\Leftrightarrow A^{-1} \geq B^{-1} \text{ und } \|B^{-1}\|_\infty \leq \|A\|_\infty$$

□

Spektralnorm

Zur euklidischen Vektornorm $\|u\|_2 = \sqrt{c \sum_{i=1}^n |u_i|^2}$, $c = const > 0$ lässt sich eine zugehörige Matrixnorm, die *Spektralnorm* definieren.

Lemma 5. *Die zur Vektornorm gehörige Matrixnorm* $\|.\|_2$ *heißt* Spektralnorm *und lässt sich schreiben als*

$$\|A\|_2 = \sqrt{\varrho(A^t A)} \tag{3.24}$$

Beweis.

$$\|A\|_2 = \sup_{u \in V} \left\{ \frac{\|Au\|_2}{\|u\|_2} : u \neq 0 \right\} = \max_{u \in V, \|u\|_2 = 1} \|Au\|_2$$

$$\Rightarrow \|A\|_2^2 = \max_{u \in V, \|u\|_2 = 1} \|Au\|_2^2 = \max_{\|u\|_2 = 1} \langle Au, Au \rangle = \max_{\|u\|_2 = 1} \langle A^t Au, u \rangle$$

Da $A^t A$ symmetrisch ist, liefert eine Hauptachsentransformation

$$P^t A^t A P = diag(\lambda_i) = D$$

mit $P = (v_1, \ldots, v_n)$, mit den Eigenvektoren v_i von $A^t A$. D enthält die zugehörigen Eigenwerte auf der Diagonalen. Für P gilt $PP^t = 1$.

$$\begin{aligned}
\Rightarrow \|A\|_2^2 &= \max_{\|u\|_2 = 1} \langle A^t Au, u \rangle = \max_{\|u\|_2 = 1} \langle A^t APP^t u, PP^t u \rangle \\
&\underbrace{=}_{\tilde{u} = P^t u} \max_{\|\tilde{u}\|_2 = 1} \langle A^t AP\tilde{u}, P\tilde{u} \rangle = \max_{\|\tilde{u}\|_2 = 1} \langle P^t A^t AP\tilde{u}, \tilde{u} \rangle = \max_{\|\tilde{u}\|_2 = 1} \langle D\tilde{u}, \tilde{u} \rangle \\
&= \max_{\|\tilde{u}\|_2 = 1} \langle \sum_{i=1}^n \lambda_i \tilde{u}_i, \tilde{u}_i \rangle = \max_{\|\tilde{u}\|_2 = 1} \left(\sum_{i=1}^n \lambda_i |\tilde{u}_i|^2 \right) = \lambda_{max} = \varrho(A^t A) \\
&\Rightarrow \|A\|_2 = \sqrt{\varrho(A^t A)}
\end{aligned}$$

□

Bemerkung 8. *Falls A symmetrisch ist, kann man obiges Vorgehen direkt auf A anwenden.*

$$\Rightarrow \|A\|_2 = \varrho(A), \text{ für } A \text{ symmetrisch}$$

Positiv definite Matrizen

Bevor wir im folgenden Abschnitt die wichtigsten Eigenschaften der Systemmatrix zur diskreten Poisson-Gleichung zusammenstellen, benötigen wir noch den Begriff der *positiven Definitheit.*

Definition 24. *Eine Matrix A heißt* positiv definit*, wenn A symmetrisch ist und*

$$\langle Au, u\rangle > 0 \quad \forall u \neq 0. \tag{3.25}$$

Lemma 6. *A ist positiv definit, genau dann wenn alle Eigenwerte von A positiv sind.*

Beweis. A symmetrisch $\Rightarrow \exists P : P^tAP = diag(\lambda_i)$. Dabei sind λ_i die Eigenwerte von A und P die Matrix zusammengesetzt aus den Eigenvektoren.

$$\begin{aligned} \Rightarrow \langle Au, u\rangle &= \langle APu, Pu\rangle = \langle P^tAPu, u\rangle \\ &= \langle \sum_{i=1}^{n} \lambda_i u, u\rangle = \sum_{i=1}^{n} \langle \lambda_i u, u\rangle > 0, \text{ gdw } \lambda_i > 0 \end{aligned}$$

□

Daraus folgt folgendes

Lemma 7. *Eine positiv definite Matrix ist regulär und besitzt eine positiv definite Inverse.*

Lemma 8. *Für A symmetrisch und diagonaldominant (oder irreduzibel diagonaldominant) mit $a_{ii} > 0$, so ist A positiv definit.*

Beweis.

$$\sum_{j=1, j\neq i}^{n} |a_{ij}| < a_{ii} \Rightarrow \text{ Gerschgorinkreise liegen in } (0, \infty)$$

$\Rightarrow \lambda_i$ positiv.
$\Rightarrow A$ ist positiv definit.

□

Lemma 9. *λ_{min} und λ_{max} seien minimaler bzw. maximaler Eigenwert von A (positiv definit). Dann gilt:*

$$\begin{aligned} \|A\|_2 &= \lambda_{max} \\ \|A^{-1}\|_2 &= \frac{1}{\lambda_{min}} \end{aligned}$$

Beweis. A ist positiv definit. $\Rightarrow A$ symmetrisch $\Rightarrow \|A\|_2 = \varrho(A)$

$$\begin{aligned} \Rightarrow \|A\|_2 &= \lambda_{max} \\ \|A^{-1}\|_2 &= \varrho(A^{-1}) = \frac{1}{\lambda_{min}} \end{aligned}$$

□

3.6.2 Matrixeigenschaften von K_h

Die Summe der in den vorigen Abschnitten zusammengefassten Sätze und Matrixeigenschaften ermöglichen den Beweis folgender Aussagen:

Satz 9. *Die Matrix K_h, definiert durch den Stern*

$$\begin{bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{bmatrix}$$

besitzt folgende Eigenschaften:

1. *K_h ist eine M-Matrix.*
2. *K_h ist positiv definit.*
3. *$\|K_h\|_\infty \leq \frac{8}{h^2}$ und $\|K_h^{-1}\|_\infty \leq \frac{1}{8}$.*
4. *$\|K_h\|_2 \leq \frac{8}{h^2}\cos^2\left(\frac{\pi h}{2}\right) < \frac{8}{h^2}$ und*
5. *$\|K_h^{-1}\|_2 \leq \frac{1}{8}h^2\sin^{-2}\left(\frac{\pi h}{2}\right) = \frac{1}{2\pi^2} + \mathcal{O}(h^2) < \frac{1}{16}$*

Beweis.

1. K_h erfüllt die Vorzeichenbedingung und ist irreduzibel diagonaldominant. In Satz 5 haben wir gezeigt, dass K_h dann auch eine M-Matrix ist.
2. K_h ist symmetrisch und irreduzibel diagonaldominant $\Rightarrow$ K_h ist positiv definit (siehe Lemma 8).
3. a) $\|K_h\|_\infty = \max_{i=1,\ldots,n}\left(\sum_{j=1}^n |K_{ij}|\right) = \frac{1}{h^2}\max 6,7,8 = \frac{8}{h^2}$

□

   b) Zu zeigen ist: $\|K_h^{-1}\|_\infty \leq \frac{1}{8}$. Nutze dazu Satz 7 mit

$$w(x,y) = \frac{x(1-x)}{2}$$

   und betrachte

$$K_h w_h = -\frac{(x-h)(1-(x-h))}{2h^2} - \frac{(x+h)(1-(x+h))}{2h^2} + \frac{2\cdot x(1-x)}{2h^2} = 1$$

   $K_h w \geq \mathbb{1}$. Für $\|w\|_\infty$ gilt:

$$\frac{d}{dx}w(x,y) = -x + \frac{1}{2} \Rightarrow \max_{x,y} w(x,y) = \frac{1}{2}$$

   Wegen Satz 7 gilt: $\|K_h^{-1}\|_\infty \leq \|w\|_\infty = \frac{1}{8}$.

□

4. Folgt aus folgendem

Lemma 10. *Die Eigenvektoren und Eigenwerte von K_h sind*

$$u^{\nu\mu}(x,y) = \sin(\nu\pi x)\sin(\mu\pi y), \quad (x,y) \in \Omega \tag{3.26}$$

mit

$$\lambda_{\nu\mu} = \frac{4}{h^2}\left(\sin^2\left(\frac{\nu\pi h}{2}\right) + \sin^2\left(\frac{\mu\pi h}{2}\right)\right), \quad 1 \leq \nu,\mu \leq n-1 \tag{3.27}$$

Beweis.

$$\begin{aligned} K_h u^{\nu\mu} &= \frac{1}{h^2}\left(-\sin(\nu\pi(x-h))\sin(\mu\pi y) + 4\sin(\nu\pi x)\sin(\mu\pi y)\right. \\ &\quad \left. -\sin(\nu\pi(x+h))\sin(\mu\pi y) - \sin(\nu\pi x)\sin(\mu\pi(y-h)) - sin(\nu\pi x)\sin(\mu\pi(y+h))\right) \end{aligned}$$

Wende die Regel

$$\sin(x \pm y) = \sin x \cos y \pm \sin y \cos x$$

mit $x := \nu\pi x$ und $y := \nu\pi h$ auf obigen Ausdruck an.

$$\Rightarrow K_h u^{\nu\mu} = \lambda_{\nu\mu} u^{\nu\mu}$$

□

$$\Rightarrow \|K_h\|_2 = \varrho(A) = \lambda_{max} \leq \tfrac{8}{h^2}\sin^2\left(\tfrac{\pi(n-1)h}{2}\right) = \tfrac{8}{h^2}\cos^2\left(\tfrac{\pi h}{2}\right) < \tfrac{8}{h^2}$$

□

5. Es gilt

$$\begin{aligned} \|K_h^{-1}\|_2 &= \varrho(K_h^{-1}) = \frac{1}{\lambda_{min}} \text{ mit} \\ \lambda_{min} &= \lambda_{11} = \frac{8}{h^2}\sin^2\left(\frac{\pi h}{2}\right) = \frac{8}{h^2}\left(\frac{\pi h}{2} + \mathcal{O}(h^2)\right)^2 \\ &= \frac{8}{h^2}\left(\left(\frac{\pi h}{2}\right)^2 + \mathcal{O}(h^4)\right) = 2\pi^2 + \mathcal{O}(h^2) \\ \Rightarrow \|K_h^{-1}\|_2 &\leq \frac{1}{2\pi^2} + \mathcal{O}(h^2) \end{aligned}$$

□

3.7 Konvergenzuntersuchung für das Finite Differenzen Verfahren

Dass ein numerisches Approximationsverfahren mit zunehmender Gitterfeinheit gegen die kontinuierliche Lösung konvergiert, ist Grundvoraussetzung für den effizienten Einsatz des Verfahrens. Wir werden im Folgenden, dass das behandelte Differenzenverfahren für die Poissongleichung *stabil* und *konvergent* ist.

3.7.1 Stetige Abhängigkeit von den Randdaten

Maximumsprizip

Lemma 11. *Sei $u_h \neq 0$ eine Lösung von $-\Delta_h u_h = f_h$ mit $f_h = 0$ und $u_h|_{\partial\Omega_h} = \varphi_h$. Dann nimmt u_h sein Maximum bzw. Minimum auf dem Rand $\partial\Omega_h$ an, oder u_h ist konstant.*

Beweis. Für $-\Delta_h u_h = 0$ gilt

$$u_h(x,y) = \frac{1}{4}(u_h(x-h,y) + u_h(x+h,y) + u_h(x,y-h) + u_h(x,y+h))$$

Angenommen u_h ist nicht konstant und $u_h(x,y)$ sei Maximum in Ω_h.
$\Rightarrow u_h(x \pm h, y)$ und $u_h(x, y \pm h)$ müssen maximal sein. Da K_h aber irreduzibel ist, setzt sich diese Bedingung auf ganz Ω_h fort.
$\Rightarrow u_h$ ist konstant. Dies ist ein Widerspruch zu obiger Annahme.

□

Vergleichsprinzip

Seien u_h, v_h Lösungen von

$$-\Delta_h u_h = f_h \text{ mit } u_h|_{\partial\Omega_h} = \varphi_h^u$$
$$-\Delta_h v_h = f_h \text{ mit } v_h|_{\partial\Omega_h} = \varphi_h^v$$

Dann gilt:

1. $\|u_h - v_h\|_\infty \leq \max_{x\in\partial\Omega_h} |\varphi^u(x) - \varphi^v(x)|$
2. $u_h \leq v_h$, falls $\varphi^u \leq \varphi^v$ auf $\partial\Omega_h$.

Beweis. Betrachte $w_h := v_h - u_h$. w_h ist Lösung der Poisson-Gleichung $-\Delta_h w_h = 0$ und $w_h \geq 0$ auf $\partial\Omega_h$, falls

$$\varphi^u \leq \varphi^v.$$

Aus dem Maximumsprinzip folgt w_h ist konstant oder $w_h > 0$. $\Rightarrow$ Behauptung 2.
Mit $u_h = \varphi^u$ und $v_h = \varphi^v$ auf $\partial\Omega_h$ gilt

$$-\max|\varphi^u - \varphi^v| \leq w_h \leq +\max|\varphi^u - \varphi^v| \text{ auf } \partial\Omega_h$$

Aufgrund des Maximumsprinzips gilt dies auf ganz $\Omega_h \Rightarrow$ Behauptung 1.

□

3.7.2 Konvergenz, Konsistenz und Stabilität

Wir wollen nun das kontinuierliche mit dem numerisch approximierten Problem vergleichen. Wir betrachten dazu

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega \\ u|_\Gamma &= \varphi \end{aligned}$$

und

$$\begin{aligned} -\Delta_h u_h &= f_h \text{ in } \Omega_h \\ u_h|_{\Gamma_h} &= \varphi_h \end{aligned}$$

Bemerkung 9. *Das kontinuierliche und diskrete Problem existieren in unterschiedlichen Gebieten. Um beide Probleme vergleichen zu können, müssen sie in einem Raum definiert sein.*

Definition 25. *(Restriktion)*

$$\begin{aligned} R_h : C(\bar{\Omega}) &\longrightarrow U_h \\ u &\mapsto R_h u \end{aligned}$$

mit $(R_h u)(\vec{x}) = u(\vec{x})$ *für alle* $\vec{x} \in \bar{\Omega}_h$ *ist die* Restriktion *der Lösung in* $\bar{\Omega}$ *auf* $\bar{\Omega}_h$. *Dabei sei* $h \in H \subset \mathbb{R}^+$ *und* H *eine Menge ohne Häufungspunkt (in unserem Fall ist* $H = \left\{\frac{1}{n} : n \in \mathbb{N}\right\}$*).*

Bemerkung 10. *Durch die Restriktion der kontinuierlichen Lösung auf das diskrete Gebiet können die Lösungen* u *und* u_h *nun verglichen werden.*

Konvergenz

Die diskrete Lösung $u_h \in U_h$ konvergiert bzgl. $\|.\|_h, h \in H$ gegen u, falls

$$\|u_h - R_h u_h\|_h \longrightarrow 0 \tag{3.28}$$

Definition 26. $u_h - R_h u$ *heißt* Diskretisierungsfehler *des Diskretisierungsverfahrens.*

Definition 27. *Die Diskretisierung* K_h *heißt* stabil *bzgl.* $\|.\|_\infty$ *für* $h \in H \subset \mathbb{R}^+$*, falls*

$$\sup_{h \in H} \|K_h^{-1}\|_\infty \leq C < \infty \tag{3.29}$$

Bemerkung 11. *Betrachte die zwei Systeme*

$$\begin{aligned} K_h(u_h) &= f_h \\ K_h(\tilde{u}_h) &= f_h + \epsilon \end{aligned}$$

Dann folgt

$$\begin{aligned} u_h &= K_h^{-1}(f_h) \\ \tilde{u}_h &= K_h^{-1}(f_h + \epsilon) \\ \Rightarrow \|\tilde{u}_h - u_h\| &\leq C \cdot \|\epsilon\| \end{aligned}$$

Stabilität bedeutet also, dass kleine Änderung auf der rechten Seite nur kleine Änderungen in der Lösung bewirken.

Bemerkung 12. *Der Fünfpunktstern und somit die Diskretisierung K_h zur Poissongleichung hat die Eigenschaft*

$$\|K_h^{-1}\|_\infty \leq \frac{1}{8},$$

ist also stabil.

Konsistenz

Sei $K_h u_h = f_h$ die Diskretisierung von $Ku = f$. K sei ein Differentialoperator der Ordnung m. Weiter seinen R_h und $\tilde{R}_h$ Restriktionsoperatoren für u und f. Die Diskretisierung $(K_h, R_h, \tilde{R}_h)$ des Differetialoperators K hat die *Konsistenzordnung* k bzgl. $\|.\|_\infty$, falls gilt

$$\|K_h R_h u - \tilde{R}_h K u\|_\infty \leq C \cdot h^k \cdot \|u\|_{C^{k+m}(\bar{\Omega})} \quad \forall u \in C^{k+m}(\bar{\Omega}) \tag{3.30}$$

Beispiel 7. *Sei $R_h = \tilde{R}_h$ gegeben durch*

$$(R_h u)(\vec{x}) = u(\vec{x}) \quad \forall \vec{x} \in \Omega_h.$$

Dann ist $(K_h, R_h, \tilde{R}_h) = (\Delta_h, R_h, R_h)$ konsistent mit Ordnung 2 bzgl. $\|.\|_\infty$.

Beweis. Wir erinnern uns an

$$(\partial^- \partial^+ u)(x) = u''(x) + h^2 R, \quad |R| \leq \frac{1}{12}\|u\|_{C^4(\bar{\Omega})}$$

Im $\mathbb{R}^2$ wenden wir diesen Ansatz in x und y-Richtung an. Taylorentwicklung liefert

$$\begin{aligned} -\Delta_h R u(x,y) &= -\Delta u(x,y) + h^2(R_x + R_y) \\ \text{mit } |R_x| &\leq \frac{1}{12}\|u^{(4)}\|_{C^0(\bar{\Omega})} \leq \frac{1}{12}\|u\|_{C^4(\bar{\Omega})} \end{aligned}$$

Analog dazu folgert man: $|R_y| \leq \frac{1}{12}\|u\|_{C^4(\bar{\Omega})}$.
$\Rightarrow C = \frac{1}{6}$ mit $\|K_h R_h u - R_h K u\| \leq C \cdot h^2 \|u\|_{C^4(\bar{\Omega})}$

□

Satz 10. *(Satz über die Konvergenz)*
Sei die Diskretisierung $(K_h, R_h, \tilde{R}_h)$ konsistent von der Ordnung k. Sei die zu K_h gehörende Matrix $\tilde{K}_h$ stabil bzgl. $\|.\|_\infty$. Dann ist das Verfahren konvergent *von der Ordnung k, falls $u \in C^{k+m}(\bar{\Omega})$. Dabei bezeichne m die Ordnung des Differentialoperators K_h.*

Beweis. Betrachte $w_h = u_h - R_h u$. Unser Ziel ist es zu zeigen, dass

$$w_h \to 0 \text{ für } h \to 0 \Rightarrow u_h \to u \text{ für } h \to 0.$$

Es gilt:

$$K_h w_h = K_h u_h - K_h R_h u = f_h - K_h R_h u = \tilde{R}_h f - K_h R_h u = \tilde{R}_h R u - K_h R_h u$$

Da $w_h|_{\Gamma_h} = 0$ gilt $K_h w_h = \tilde{K}_h w_h$.

$$\begin{aligned} \Rightarrow w_h &= K_h^{-1}(\tilde{R}_h K u - K_h R_h u) \\ \Rightarrow \|w_h\|_\infty &= \|u_h - R_h u\|_\infty \leq \|K_h^{-1}\|_\infty \cdot \|\tilde{R}_h K u - K_h R_h u\ \|_\infty \\ \Rightarrow \|u_h - R_h u\|_\infty &\leq C h^k \|u\|_{C^{k+m}(\bar{\Omega})} \end{aligned}$$

□

Frage 6. *Was ergibt sich dann für den 5-Punkt-Stern?*

Sei $u \in C^4(\bar{\Omega})$. Dann gilt:

$$\|K_h^{-1}\|_\infty < \frac{1}{8},\ c = \frac{1}{6} \Rightarrow \|u_h - R_h u\|_\infty \leq \frac{h^2}{48} \cdot \|u\|_{C^4(\bar{\Omega})}$$

3.8 Das Neumann-Problem

Bisher wurden die Funktionswerte der gesuchten Funktion auf dem Rand des Gebietes vorgegeben mit $u|_\Gamma = \varphi$. Stattdessen können auch die Ableitungen von u in Normalenrichtung auf dem Rand vorgegeben werden, d.h. wir betrachten das Problem

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega = (0,1) \times (0,1) \\ \frac{\partial u}{\partial n} &= \varphi \text{ auf } \Gamma \end{aligned}$$

Bemerkung 13. *Dadurch, dass $\frac{\partial u}{\partial n}$ vorgegeben wird, gibt es potentiell unendlich viele Lösungen. Denn falls u Lösung des Problems ist, so erfüllt auch $u + c$ das Problem.*

Eine eindeutige Lösung erfordert also eine zusätzliche Bedingung um die Variable c eindeutig festzulegen. Diese Zusatzbedingung nennt man *Kompatibilitätsbedingung.*

Lemma 12. *(Kompatibilitätsbedingung)*
Es muss gelten

$$\int_{\partial\Omega} \varphi ds = -\int_\Omega f dx \tag{3.31}$$

Beweis.

$$\begin{aligned} -\int_\Omega f dx &= \int_\Omega \Delta u dx = \int_\Omega div(\nabla u) dx \\ &= \int_{\partial\Omega} \nabla u \cdot \vec{n} ds = \int_{\partial\Omega} \frac{\partial u}{\partial \vec{n}} ds = \int_{\partial\Omega} \varphi ds \end{aligned}$$

□

3.8.1 Diskretisierung des Neumann-Problems

Wir betrachten das Neumann-Problem auf dem Gebiet $\bar{\Omega} = [0,1] \times [0,1]$. Dabei treten drei Typen von Knoten auf:

1. innere Knoten
2. Knoten auf Kanten
3. Knoten in den Eckpunkten von $\bar{\Omega}$

Die Diskretisierung des Laplace-Operators führt in diesen drei Fällen zu:

Innere Punkte Für Punkte im Innern, die keine Randknoten zum Nachbarn haben gilt:
$-\Delta u_h(x,y) = \frac{1}{h^2}\left(4u_h(x,y) - u_h(x-h,y) - u_h(x+h,y) - u_h(x,y-h) - u_h(x,y+h)\right)$

Randnahe Punkte Für innere Knoten am rechten Rand von $\bar{\Omega}$ ergibt sich:
$-\Delta u_h(x,y) = \frac{1}{h^2}\left(3u_h(x,y) - u_h(x-h,y) - u_h(x,y-h) - u_h(x,y+h)\right)$

Ecknahe Punkte Für den Knoten rechts unten ergibt sich:
$-\Delta u_h(x,y) = \frac{1}{h^2}\left(2u_h(x,y) - u_h(x-h,y) - u_h(x,y+h)\right)$

Behandlung der Neumann Randbedingung

Die Ableitung von u in Normalenrichtung $\vec{n}$ kann über einseitige Differenzenquotienten approximiert werden. Im eindimensionalen Fall lässt sich der rechte Rand folgendermaßen darstellen:

$$\frac{\partial u_h}{\partial \vec{n}}(x) \approx (\partial_n^- u_h)(x) = \frac{1}{h}(u_h(x) - u(x - h\vec{n})) = \varphi(x) \tag{3.32}$$

Für unser Modellbeispiel folgt damit:

$$\begin{aligned}
\frac{1}{h}\left(u_h(x,0) - u_h(x,h)\right) &= \varphi(x,0) \quad \text{(unten)} \\
\frac{1}{h}\left(u_h(x,1) - u_h(x,1-h)\right) &= \varphi(x,1) \quad \text{(oben)} \\
\frac{1}{h}\left(u_h(0,y) - u_h(h,y)\right) &= \varphi(0,y) \quad \text{(links)} \\
\frac{1}{h}\left(u_h(1,y) - u_h(1-h,y)\right) &= \varphi(1,y) \quad \text{(rechts)}
\end{aligned}$$

Das resultierende diskrete Problem

$$\begin{aligned}
K_h u_h &= \tilde{f}_h = f_h - \frac{1}{h}\varphi_h \\
\varphi_h &= \begin{cases} \varphi(x,y) & (x,y) \in \Gamma \\ 0 & \text{sonst} \end{cases}
\end{aligned}$$

ist nicht notwendigerweise regulär. Deshalb muss zusätzlich die *diskrete Kompatibilitätsbedingung* erfüllt sein.

Diskrete Kompatibilitätsbedingung

Die diskrete Kompatibilitätsbedingung lautet

$$-h^2 \sum_{x \in \Omega_h} f(x) = h \sum_{x \in \Gamma} \varphi(x). \tag{3.33}$$

Satz 11. *Ist neben $K_h u_h = \tilde{f}_h$ auch die diskrete Kompatibiltätsbedingung erfüllt, dann ist das Problem lösbar und zwei Lösungen unterscheiden sich nur in einer Konstante c.*

Beweis. Anwendung des Laplace-Operators auf eine konstante Funktion ergibt Null, d.h.

$$K_h \cdot c \cdot \mathbb{1} = 0 \Rightarrow c \cdot \mathbb{1} \in \ker(K_h)$$

und

$$\begin{aligned} \ker(K_h) &= span(\mathbb{1}) \\ \Rightarrow \dim(\ker(K_h)) &= 1 \end{aligned}$$

Ferner gilt: $K_h u_h = \tilde{f}_h$ lösbar $\Rightarrow \tilde{f}_h \in \text{Bild}(K_h)$ und

$$\begin{aligned} \text{Bild}(K_h) &= \ker(K_h)^\perp = span(\mathbb{1})^\perp \\ \Rightarrow \tilde{f}_h \in \perp \mathbb{1} &\Leftrightarrow \tilde{f}_h \cdot \mathbb{1} = 0 \Longleftrightarrow \sum_{x \in \Omega_h} \tilde{f}_h(x) = 0 \\ \Rightarrow \sum_{x \in \Omega_h} \tilde{f}_h(x) &= \sum_{\Omega_h} f_h - \frac{1}{h}\varphi_h \Longleftrightarrow \sum_{\Omega_h} f_h = \frac{1}{h} \sum_{\partial \Omega_h} \varphi_h \end{aligned}$$

□

3.8.2 Lösen des Neumann-Problems

Die Lösbarkeit des Neumann-Problems setzt, wie oben gezeigt, die Erfüllung der Kompatibilitätsbedingung voraus. Wir werden zur Vereinfachung der Notation im Folgenden mit f_h die rechte Seite mit den Beiträgen auf dem Rand bezeichnen, d.h. $f_h - \frac{1}{h} \rightsquigarrow f_h$.

Lemma 13. *Man wähle $x_0 \in \Omega_h$ beliebig und normiere u_h durch*

$$u_h(x_0) = 0.$$

Dann ist das um eine Zeile und eine Spalte reduzierte System

$$\hat{K}_h \hat{u}_h = \hat{f}_h$$

lösbar.

Beweis.

1. $\hat{K}_h$ ist symmetrisch (da K_h symmetrisch)

2. $\hat{K}_h$ ist eine M-Matrix, da $\hat{K}_h$ irreduzibel diagonaldominant ist

Daraus folgt, dass $\hat{K}_h$ regulär und somit das reduzierte System lösbar ist.

□

Zur Lösung des Problems wird also eine Korrektur vorgenommen. In obigem Fall durch Elimination eines Eintrags aus u_h und f_h.
$\Rightarrow$ Die Korrektur ist konzentriert in $f_h(x_0)$. Eine Alternative dazu ist, die Korrektur durch Erweiterung des Systems zu verteilen.

Verteilung der Korrektur

Betrachte

$$\bar{K}_h \bar{u}_h = \bar{f}_h$$

mit

$$\bar{K}_h = \begin{pmatrix} K_h & \mathbb{1} \\ \mathbb{1}^T & 0 \end{pmatrix}, \bar{u}_h = \begin{pmatrix} u_h \\ \lambda \end{pmatrix}, \bar{f}_h = \begin{pmatrix} f_h \\ \sigma \end{pmatrix}$$

mit beliebigem σ.

Lemma 14. *$\bar{K}_h \bar{u}_h = \bar{f}_h$ ist eindeutig lösbar.*

Beweis. $Rang(K_h, \mathbb{1}) = Rang(K_h) + 1 \Rightarrow \bar{K}_h$ ist regulär $\Rightarrow$ Lösbarkeit.

□

Lemma 15. *Ist $\bar{u}_h$ Lösung von $\bar{K}_h \bar{u}_h = \bar{f}_h$, so impliziert dies u_h ist Lösung von $K_h u_h = f_h$.*

Beweis.

1. $\lambda = 0$ impliziert eine Normierungsbedingung für u_h mit

$$\mathbb{1}^T \cdot u_h = \sum_{x \in \Omega_h} u_h(x) = \sigma.$$

2. Sei nun $\lambda \neq 0$. u_h ist Lösung von $K_h u_h = \hat{f}_h$ mit $\hat{f}_h := f_h - \lambda \cdot \mathbb{1}$.

□

Dadurch wird eine Verteilung der Korrektur auf alle Vektoreinträge erreicht.

3.9 Differenzenverfahren für allgemeine Probleme zweiter Ordnung

In diesem Abschnitt betrachten wir allgemeine Probleme zweiter Ordnung und die daraus resultierenden diskreten Gleichungen. Wir betrachten nun folgendes Problem

$$\begin{aligned} Ku &= f \quad \text{in } \Omega \\ K &= \sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial^2}{\partial x_i x_j} + \sum_{i=1}^{n} a_i(x)\frac{\partial}{\partial x_i} + a(x) \end{aligned} \tag{3.34}$$

Bemerkung 14. *Es gilt $a_{ij}(x) = a_{ji}(x)$, da $\frac{\partial^2}{\partial x_i x_j} = \frac{\partial^2}{\partial x_j x_i}$ für zweifach differenzierbare Funktionen.*
$\Rightarrow A(x) = (a_{ij}(x))_{i,j=1\ldots n}$ ist symmetrisch. Beachte, dass $a_{ij}(x)$ ortsabhängig sein können. Der oben definierte Differentialoperator zweiter Ordnung ist in der allgemeinsten Operatorform notiert.

Definition 28. *(Elliptizität)*
Ein Differentialoperator heißt elliptisch, *falls alle Eigenwerte des Hauptteils des Differentialoperators das gleiche Vorzeichen besitzen.*

Für den Operator K zweiter Ordnung (siehe oben) gilt $a_{ij} = a_{ji}$, d.h. $A = A^T \Rightarrow A$ ist positiv definit. Dies hat zur Folge, dass alle Eigenwerte positiv sind, d.h. K ist elliptisch.

Bemerkung 15. *$-\Delta u = F$ ist eine elliptische partielle Differentialgleichung.*

Bemerkung 16. *Ein allgemeiner Differentialoperator ist elliptisch, wenn*

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j > 0 \quad \forall x \in \Omega, 0 \neq \xi \in \mathbb{R}^n$$

Definition 29. *Ein Differentialoperator K heißt* gleichmäßig elliptisch *in Ω, falls gilt*

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq c(x)|\xi|^2,\ c(x) > 0 \quad \forall x \in \Omega,\ 0 \neq \xi \in \mathbb{R}^n$$

Im zweidimensionalen Fall auf dem Gebiet $\Omega = (0,1) \times (0,1)$ erhalten wir für randferne Punkte den diskreten Operator

$$\begin{aligned} & a_{11}(x,y)\partial_x^+\partial_x^- + 2a_{12}\partial_x^0\partial_y^0 + a_{22}(x,y)\partial_y^+\partial_y^- + a_1(x,y)\partial_x^0 + a_2(x,y)\partial_y^0 + a(x,y) \\ = \; & h^{-2}\begin{bmatrix} -\frac{1}{2}a_{12}(x,y) & a_{22}(x,y) & \frac{1}{2}a_{12}(x,y) \\ a_{11}(x,y) & -2(a_{11}(x,y)+a_{22}(x,y)) & a_{11}(x,y) \\ \frac{1}{2}a_{12}(x,y) & a_{22}(x,y) & -\frac{1}{2}a_{12}(x,y) \end{bmatrix} + \\ & + \frac{1}{2h}\begin{bmatrix} 0 & a_2(x,y) & 0 \\ -a_1(x,y) & 0 & a_1(x,y) \\ 0 & a_2(x,y) & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & a(x,y) & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{aligned}$$

Upwind-Methods missing (10.-11.12.2012)

4 Diskretisierung II: Finite Elemente Verfahren

Während das Diskretisierungsverfahren der Finiten Differenzen einen einfachen mathematischen Zugang über die direkte Approximation der Differentialoperatoren bietet, in der Implementierung verhältnismäßig leicht zu handhaben ist, so gibt es jedoch gewisse Einschränkungen aufgrund des fundamentalen Vorgehens in diesem Verfahren. Zum Beispiel erfordert die Behandlung komplexer Gebiete mit krummen Rändern besondere numerische Behandlung, was wiederum die Gesamtkomplexität beeinflusst. Ferner sahen wir in den Konvergenzbeweisen für das Finite Differenzen Verfahren sehr hohe Regularitätsannahmen für die Lösungsfunktion, die eine starke Einschränkung darstellen kann.
Um eine gebietsunabhängige Diskretisierungsherangehensweise aufzubauen, sowie ein Verfahren das substantiell schwächere Regularitätsvoraussetzungen stellt, werden wir einen funktionalanalytischen Weg finden. Anstelle der direkten Approximation von Differentialoperatoren tritt die Approximation der Lösungsräume in denen die Lösungsfunktion liegt durch endlich dimensionale, approximierende Funktionenräume. Ein daraus resultierendes numerisches Verfahren ist das Verfahren der *Finite Elemente*, welchem wir uns in diesem Kapitel widmen werden.

4.1 Funktionalanalytische Grundlagen

Wie oben schon erwähnt, werden wir über die Funktionalanalysis eine neue Kategorie der Diskretisierung herleiten. Dazu sind einige funktionalanalytische Grundlagen notwendig, die in diesem Abschnitt zusammengestellt werden.

4.1.1 Normierte Räume

Definition 30. *Sei X ein Vektorraum über $\mathbb{R}$ oder $\mathbb{C}$ und $\|.\| : X \to [0, \infty)$ eine Norm. Dann wird*

$$(X, \|.\|_X)$$

normierter Raum *genannt.*

Beispiel 8. *Die stetigen Funktionen auf $\bar{\Omega}$ bilden den normierten Raum $C^0(\bar{\Omega})$ mit der Supremumsnorm $\|.\|_\infty$.*

Definition 31. *Zwei Normen $\|.\|^{(1)}$ und $\|.\|^{(2)}$ auf X heißten* äquivalent*, wenn $0 < C < \infty$ existiert, mit*

$$\frac{1}{C}\|x\|^{(1)} \leq \|x\|^{(2)} \leq C\|x\|^{(1)} \quad \forall x \in X \tag{4.1}$$

Operatoren

Definition 32. *(Operatornorm)*
Seien X und Y normierte Räume mit Normen $\|.\|_X$ und $\|.\|_Y$. Die Operatornorm *eines Operators $T : X \to Y$ ist definiert als*

$$\|T\|_{Y \leftarrow X} := \sup_{x \in X} \left\{ \frac{\|Tx\|_Y}{\|x\|_X} : 0 \neq x \in X \right\} \tag{4.2}$$

Bemerkung 17. *Ist $\|T\|_{Y \leftarrow X}$ endlich, dann ist T beschränkt.*

Bemerkung 18. *Die beschränkten Operatoren bilden einen linearen Raum $L(X, Y)$ mit $(T_1 + T_2)x = T_1x + T_2x$.*

Offene Räume

Definition 33. *$(X, \|.\|)$ sein ein normierter Raum. $A \subset X$ heißt* offen*, falls für alle $x \in A$ ein $\epsilon > 0$ existiert, sodass*

$$K_\epsilon(x) := \{y \in X : \|x - y\| < \epsilon\}$$

in A enthalten ist.

4.1.2 Banach-Räume

Sei $(X, \|.\|)$ ein normierter Raum. X heißt *vollständig*, wenn jede Cauchy-Folge konvergiert. Ein normierter und vollständiger Raum heißt *Banach-Raum*

Definition 34. *(Cauchy-Folge)*
Eine Folge $\{x_h \in X : n \geq 1\}$ heißt Cauchy-Folge*, wenn gilt*

$$\sup\{\|x_n - x_m\|_X : n, m \geq k\} \to 0, \text{ für } k \to \infty \tag{4.3}$$

oder alternativ

$$\forall \epsilon > 0 \, \exists n_0 \in \mathbb{N} : \forall n, m \geq n_0 : \|x_n - x_m\| < \epsilon \tag{4.4}$$

Der Raum $\left(L^\infty(D), \|.\|_{L^\infty(D)}\right)$

$L^\infty(D)$ bezeichnet den Raum der auf D beschränkten lokal integrierbaren Funktionen. $L^\infty(D)$ besteht aus Äquivalenzklassen, wobei

$$\begin{aligned} f &= g, \text{ falls } f = g \textit{ fast überall} \\ \|u\|_{L^\infty(D)} &:= \inf_A \left\{ \sup_{x \in D \setminus A} \{|u(x)| : x \in D \setminus A : \mu(A) = 0\} \right\} \end{aligned}$$

Wir werden im Folgenden erklären, was mit obiger Terminologie gemeint ist.

Definition 35. *(fast überall)*
Zwei Funktionen f und g heißen fast überall *gleich, falls*

$$\begin{aligned} f &= g \text{ bis auf Nullmenge } A \text{ gilt} \\ f &= g \text{ auf } \Omega \setminus A \end{aligned}$$

gilt.

Definition 36. *(Nullmenge)*
Eine Nullmenge *besitzt das Maß $\mu(A) = 0$.*

Definition 37. *(Maß)*
Sei $\mathcal{J}^n$ die Menge der beschränkten Intervalle des $\mathbb{R}^n$. Eine Intervallfunktion ist eine Abbildung

$$\varphi : \mathcal{J}^n \to \mathbb{R}$$

1. *φ ist* monoton*, wenn gilt*

$$I_1, I_2 \in \mathcal{J}^n \text{ und } I_1 \subset I_2 \Rightarrow \varphi(I_1) \leq \varphi(I_2).$$

2. *φ heißt* additiv*, wenn gilt*

$$I_1, I_2, I \in \mathcal{J}^n \text{ mit } I_1 \cap I_2 = \emptyset,\ I_1 \cup I_2 = I$$

$\Rightarrow \varphi(I) = \varphi(I_1) + \varphi(I_2)$
($\Rightarrow \varphi$ monoton und additiv $\Rightarrow \varphi(\emptyset) = 0$ und $\varphi(I) \geq 0, \forall I \in \mathcal{J}^n$).

3. *φ heißt* regulär*, falls gilt: Für jedes $\epsilon > 0$ gibt es zu jedem $I \in \mathcal{J}^n$ ein offenes Intervall I^* mit $I \subset I^*$, sodass gilt:*

$$\varphi(I) \leq \varphi(I^*) < \varphi(I) + \epsilon$$

Eine monotone, additive, reguläre Intervall-Funktion heißt Maß*.*

Beispiel 9. *Folgende gängige Maße können definiert werden:*

1. *Sei $I = [a_1, b_1] \times \ldots \times [a_n, b_n] \in \mathcal{J}^n$. Dann ist*

$$v : \mathcal{J}^n \to \mathbb{R} \text{ mit } v(I) = \Pi_{j=1}^n (b_j - a_j)$$

ein Maß (Volumenmaß).

2. *Für treppenfunktion-approximierbare Funktionen f ist*

$$\int f d\varphi = \int_{\mathbb{R}^n} f d\varphi = \sum_{i=1}^{m} \varphi(I_j) f(I_j)$$

auf den Intervallen $I_1, \ldots, I_m$ das Lebegue-Maß.

Definition 38. *(meßbar)*
$f : \mathbb{R}^n \to \mathbb{R}$ heißt meßbar*, wenn eine Folge von Treppenfunktionen $\{s_n\}_{n=1,2,\ldots}$ existiert, mit*

$$f = \lim s_n$$

.

4.1.3 Der Sobolev-Raum $L^2(\Omega)$

Definition 39. *(Skalarprodukt)*
Die Abbildung $(.,.) : X \times X \to \mathbb{K}$ *heißt* Skalarprodukt *auf* X*, falls gilt*

$$\begin{aligned} (x,x) &> 0 \quad \forall x \in X, x \neq 0 && (4.5)\\ (\lambda x + y, z) &= \bar{\lambda}(x,z) + (y,z) \quad \forall \lambda \in \mathbb{K}, x,y,z \in X && (4.6)\\ (x,y) &= \overline{(y,x)} && (4.7) \end{aligned}$$

Das Skalarprodukt definiert eine Norm durch

$$\|x\| := \sqrt{(x,x)}.$$

Definition 40. *(Hilbertraum)*
Ein Banach-Raum X *heißt* Hilbert-Raum*, wenn ein Skalarprodukt* $(.,.)_X$ *auf* X *existiert.*

Der Sobolev-Raum $L^2(\Omega)$

Sei Ω eine offene Teilmenge von $\mathbb{R}^n$. $L^2(\Omega)$ definiert den Raum der Äquivalenzklassen *meßbarer* und *quadrat-integrabler* Funktionen:

$$L^2(\Omega) := \{f : \Omega \to \mathbb{R} : f \text{ meßbar}, |f|^2 \text{ integrierbar}\} \tag{4.8}$$

Zwei Funktionen f und g sind gleich, wenn sie bis auf in A mit $\mu(A) = 0$ gleich sind. Mit dem Skalarprodukt

$$(u,v)_0 = (u,v)_{L^2(\Omega)} := \int_\Omega u(x)\overline{v(x)}dx \quad \forall u,v \in L^2(\Omega) \tag{4.9}$$

und Norm

$$\|u\|_0 = \|u\|_{L^2(\Omega)} := \sqrt{\int_\Omega |u(x)|^2} \tag{4.10}$$

ist $L^2(\Omega)$ ein Hilbertraum.

4.1.4 Schwache Differenzierbarkeit

In $L^2(\Omega)$ können keine Aussagen über die Differenzierbarkeit von $f \in L^2(\Omega)$ im klassischen Sinne gemacht werden. Aus der partiellen Integration folgernd gilt jedoch:

$$\int_\Omega f'(x)\varphi(x)dx = -\int_\Omega f(x)\varphi'(x)dx$$

für Funktionen φ mit $\varphi|_{\partial\Omega} = 0$.

Definition 41. *Falls für* $f \in L^2(\Omega)$ *eine Funktion* $g \in L^2(\Omega)$ *existiert, sodass gilt*

$$-\int_\Omega f'(x)\varphi(x)dx = -\int_\Omega g(x)\varphi(x)dx = \int_\Omega f(x)\varphi'(x)dx \tag{4.11}$$

dann heißt g *die* schwache Ableitung *von* f*.*

Bemerkung 19. *Folgende Zusammenhänge bestehen zwischen klassischer und schwacher Differenzierbarkeit:*

1. *Klassisch differenzierbare Funktionen sind auch schwach differenzierbar.*
2. *Die schwache Ableitung ist durch die integrale Definition nicht punktweise definiert.*
3. *Hinreichend oft schwach differenzierbar impliziert klassische Differenzierbarkeit (siehe Sobolevscher Einbettungssatz).*

Höhere schwache Differenzierbarkeit

Sei α ein Multiindex $\alpha = (\alpha_1, \ldots, \alpha_n)$ mit

$$\begin{aligned} |\alpha| &= \sum_{i=1}^{n} \alpha_i \\ D^\alpha &:= \frac{\partial^{|\alpha|}}{\partial^{\alpha_1} x_1 \ldots \partial^{\alpha_n} x_n} \end{aligned}$$

und $f \in L^2(\Omega)$. Dann heißt g α-fache schwache Ableitung von f, wenn gilt

$$\int_\Omega g(x)\varphi(x)dx = (-1)^{|\alpha|} \int_\Omega f(x) D^\alpha \varphi(x) dx \tag{4.12}$$

$$(g, \varphi)_0 = (-1)^{|\alpha|} (f, D^\alpha \varphi) \quad \forall \varphi \in C^\infty(\Omega), \varphi|_{\partial\Omega} = 0 \tag{4.13}$$

4.1.5 Die Hilbert-Räume $H^k(\Omega)$ und $H_0^k(\Omega)$

Die Menge aller Funktionen u aus $L^2(\Omega)$, die schwache Ableitungen $D^\alpha u \in L^2(\Omega)$ besitzen, bilden den *Sobolev-Raum*

$$H^k(\Omega) := \{u \in L^2(\Omega) : D^\alpha u \in L^2(\Omega) \text{ für } |\alpha| \leq k\} \tag{4.14}$$

mit $k \in \mathbb{N}_0$. $H^k(\Omega)$ wird in der Literatur auch als $W_2^k(\Omega)$ bzw. $W^{k,2}(\Omega)$ bezeichnet. $H^k(\Omega)$ ist ein Hilbert-Raum mit dem Skalarprodukt

$$(u, v)_k := (u, v)_{H^k(\Omega)} := \sqrt{\sum_{|\alpha| \leq k} \|D^\alpha u\|^2_{L^2(\Omega)}} \tag{4.15}$$

Satz 12. *(Sobolevscher Einbettungssatz)*
Es gilt

$$H^s(\mathbb{R}^n) \subset C^k(\mathbb{R}^n) \text{ falls } s > k + \frac{n}{2},\ k \in \mathbb{N}_0 \tag{4.16}$$

4.1.6 Dualräume

Sei X ein normierter, linearer Raum über $\mathbb{R}$. Mit X' wird der *Dualraum* bezeichnet, bestehend aus allen beschränkten, linearen Abbildungen von X nach $\mathbb{R}$:

$$X' = L(X, \mathbb{R})$$

Bemerkung 20. *Mit der Norm*

$$\|x'\| := \|x'\|_{\mathbb{R} \leftarrow X} := \sup \left\{ \frac{|x'(x)|}{\|x\|_X} : 0 \neq x \in X \right\}$$

ist X' ein Banachraum. $x' \in X'$ heißen lineare Funktionale *auf X:*

$$x'(x) = \langle x, x' \rangle_{X \times X'}$$

Lemma 16. *Seien X und Y normiert und $T \in L(X, Y)$. Für jedes $y' \in Y'$ definiert*

$$\langle Tx, y' \rangle_{Y \times Y'} = \langle x, x' \rangle_{X \times X'} \quad \forall x \in X \tag{4.17}$$

ein eindeutiges $x' \in X'$. Der zugehörige Dualoperator ist durch

$$\begin{aligned} T' : Y' &\longrightarrow X' \\ y' &\longmapsto x' \end{aligned}$$

definiert. D.h. $\langle Tx, y' \rangle_{Y \times Y'} = \langle x, T'y' \rangle_{X \times X'} = \langle x, x' \rangle_{X \times X'}$.

Lemma 17. *Es gilt*

$$\|T'\|_{X' \leftarrow Y'} = \|T\|_{Y \leftarrow X} \tag{4.18}$$

Beweis.

$$\begin{aligned} \|T'\|_{X' \leftarrow Y'} &= \sup_{y' \neq 0} \left\{ \frac{\|T'y'\|_{X'}}{\|y'\|_{Y'}} \right\} = \sup_{x, y \neq 0} \left\{ \frac{\langle x, T'y' \rangle_{X \times X'}}{\|x\|_X \|y'\|_{Y'}} \right\} \\ &= \sup_{x, y \neq 0} \left\{ \frac{\langle Tx, y' \rangle_{Y \times Y'}}{\|x\|_X \|y'\|_{Y'}} \right\} = \sup_{x' \neq 0} \left\{ \frac{\|Tx\|_Y}{\|x\|_X} \right\} = \|T\|_{Y \leftarrow X} \end{aligned}$$

Adjungierte Operatoren

Sei X ein Hilbert-Raum über $\mathbb{R}$. Jedes $y \in X$ definiert durch

$$f_y(x) := (x, y)_X \tag{4.19}$$

ein lineares Funktional $f_y \in X'$ mit $\|f_y\|_{X'} = \|y\|_X$. Umgekehrt definiert jedes Funktional f_y ein $y \in X$, wie folgender Satz zeigt:

Satz 13. *(Darstellungssatz von Riesz)*
X sei ein Hilbert-Raum und $f \in X'$ ein Funktional. Dann existiert genau ein $y_f \in X$, sodass

$$f(x) = (x, y)_X \quad \forall x \in X, \|f\|_{X'} = \|y_f\|_{X'} \tag{4.20}$$

Beweis. X ist abgeschlossen. Sei $N = \{x \in X : f(x) = 0\}$ der Kern von f.

1. $N = X$: Wähle $y = 0 \Rightarrow$ Behauptung.

2. Sei N ein abgeschlossener Teilraum von X. Urbilder abgeschlossener Mengen unter stetigen Funktionen sind abgeschlossen und auf abgeschlossenen Mengen existiert ein $z \in X$ mit
$$z \notin N \text{ und } \langle z, n\rangle = 0 \quad \forall n \in N.$$
Ferner gilt:
$$\begin{aligned} f\left(x - \frac{f(x)z}{f(z)}\right) &= f(x) - f\left(\frac{f(x)z}{f(z)}\right) \\ &= f(x) - \frac{f(x)}{f(z)} f(z) = 0 \\ \Rightarrow \quad x - \frac{f(x)z}{f(z)} &\in N \end{aligned}$$
Eingesetzt liefert
$$\begin{aligned} \langle x - \frac{f(x)}{f(z)} z, z\rangle &= 0 \Leftrightarrow \langle x, z\rangle - \frac{f(x)}{f(z)}\langle z, z\rangle = 0 \\ \Rightarrow f(x) &= \frac{\langle x, z\rangle}{\|z\|^2} f(z) = \langle x, \frac{f(z)z}{\|z\|^2}\rangle \end{aligned}$$
Mit $y := \frac{f(z)}{\|z\|^2} z$ folgt die Behauptung, bis auf die Eindeutigkeit.

Eindeutigkeit: Sei $\tilde{y}$ ein weiteres Element mit
$$\begin{aligned} f(x) &= \langle x, \tilde{y}\rangle = \langle x, y\rangle \\ &\Rightarrow \langle x, \tilde{y} - y\rangle = 0 \quad \forall x \in X \\ &\Rightarrow \tilde{y} - y = 0 \Rightarrow \text{ Eindeutigkeit} \end{aligned}$$

□

Bilinearformen

Definition 42. *Sei V ein Hilbertraum. Die Abbildung $a(.,.) : V \times V \longrightarrow \mathbb{R}$ heißt Bilinearform, falls*
$$a(x + \lambda y, z) = a(x, z) + \lambda a(y, z) \tag{4.21}$$
$$a(x, y + \lambda z) = a(x, y) + \lambda a(x, z) \quad \forall \lambda \in \mathbb{R},\ x, y, z \in V \tag{4.22}$$

Definition 43. $a(.,.)$ *heißt* stetig, *falls ein C_s existiert, sodass*
$$|a(x, y)| \leq C_s \|x\|_V \|y\|_V \quad \forall x, y \in V \tag{4.23}$$

Lemma 18. *Zu einer stetigen Bilinearform existiert ein eindeutiger Operator $A \in L(V, V')$ mit*

$$a(x,y) = \langle Ax, y\rangle_{V'\times V} \quad \forall x, y \in V \tag{4.24}$$

$$\|A\|_{V'\leftarrow V} \leq C_s \tag{4.25}$$

Beweis. Man halte $x \in V$ fest.

$$\varphi_x(y) := a(x,y)$$

ist ein Funktional $\varphi_x \in V'$ mit $\|\varphi_x\|_{V'} \leq C_s\|x\|_V$. Da φ_x über die Bilinearform definiert ist, ist φ_x linear.

$$Ax := \varphi_x$$

für $x \in V$ und es gilt mit

$$\|Ax\|_{V'} \leq C_s\|x\|_V \Leftrightarrow \frac{\|Ax\|_{V'}}{\|x\|_V} \leq C_s$$

auch

$$\|A\|_{V'\leftarrow V} := \sup\left\{\frac{\|Ax\|_{V'}}{\|x\|_V}\right\} \leq C_s$$

Analog dazu: Halte $y \in V$ fest.

□

Definition 44. *(V-Elliptizität)*
Eine Bilinearform heißt V-elliptisch*, falls sie auf $V \times V$ stetig ist und $C_E > 0$ existiert, sodass*

$$a(x,x) \geq C_E\|x\|_V^2 \quad \forall x \in V \tag{4.26}$$

4.2 Variationsformulierung

Das Ziel in diesem Abschnitt ist es, über die Funktionalanalysis einen neuen Zugang zum Modellproblem zu finden. Ein äquivalentes Problem zum klassischen Problem definieren wir über eine Variationsformulierung. Diese wird Lösungen des Modellproblems in Hilberträumen angeben, die bis auf Nullmengen auch klassische Lösungen sein werden.

Satz 14. *(Charakterisierungssatz)*
Sei V ein linearer Raum und

$$a : V \times V \longrightarrow \mathbb{R}$$

symmetrisch und positiv. $f : V \longrightarrow \mathbb{R}$ sei ein lineares Funktional. Die Abbildung

$$J(v) := \frac{1}{2}a(v,v) - f(v)$$

nimmt in V ihr Minimum genau dann bei u an, wenn gilt:

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

Die Lösung u ist eindeutig.

Beweis. Seien $u, v \in V$, $t \in \mathbb{R}$. Betrachte

$$\begin{aligned} J(u+tv) &= \frac{1}{2}a(u+tv, u+tv) - f(u+tv) = \\ &= \frac{1}{2}\left(a(u,u) + 2ta(u,v) + t^2 a(v,v)\right) - f(u) - tf(v) \\ &= J(u) + t(a(u,v) - f(v)) + \frac{1}{2}t^2 a(v,v) \end{aligned}$$

1. Gelte für $u \in V$: $a(u,v) = f(v)$

$$\begin{aligned} \overset{t=1}{\Rightarrow} J(u+v) &= J(u) + (f(v) - f(v)) + \frac{1}{2}a(v,v) \\ &= J(u) + \frac{1}{2}a(v,v) > J(u) \end{aligned}$$

   $\Rightarrow u$ ist Minimalpunkt.

2. Sei $u \in V$ Minimalpunkt.

$$\begin{aligned} J(u+tv) &= J(u) + t(a(u,v) - f(v)) + \frac{1}{2}t^2 a(v,v) \\ \Rightarrow 0 = \frac{dJ(u+tv)}{dt} &= a(u,v) - f(v) + ta(v,v) \\ \Rightarrow J'(u+tv)|_{t=0} &= a(u,v) - f(v) \\ \Leftrightarrow a(u,v) &= f(v) \end{aligned}$$

   $\Rightarrow a(u,v) = f(v) \Longleftrightarrow u$ ist Minimalpunkt.

□

4.2.1 Untersuchung des elliptischen Differentialoperators zweiter Ordnung

Wir betrachten den elliptischen Differentialoperator

$$Lu := -\sum_{i,k=1}^{n} \partial_i(a_{ik}\partial_k u) + a_0 u$$

mit $a_0(x) \geq 0\,(x \in \Omega)$. Das zugehörige Problem

$$\begin{aligned} Lu &= f \quad \text{in } \Omega \\ u &= g \quad \text{auf } \partial\Omega \end{aligned}$$

kann ohne Beschränkung der Allgemeinheit als homogenes Problem aufgefasst werden:

$$\begin{aligned} w &:= u - g \\ \Rightarrow Lw &= f_1 \quad \text{in } \Omega \\ w &= 0 \quad \text{auf } \partial\Omega \end{aligned}$$

Folgender Satz stellt den Zusammenhang zwischen Randwertaufgabe und Variationsproblem her:

Satz 15. *(Minimaleigenschaft)*
Jede klassische Lösung der Randwertaufgabe

$$-\sum_{i,k} \partial_i(a_{ik}\partial_k u) + a_0 u = f \text{ in } \Omega \tag{4.27}$$

$$u = 0 \text{ auf } \partial\Omega \tag{4.28}$$

ist Lösung des Minimierungsproblems

$$J(v) := \int_\Omega \left(\frac{1}{2} \sum_{i,k} a_{ik}\partial_i v \partial_k v + \frac{1}{2} a_0 v^2 - fv \right) dx \to \min \tag{4.29}$$

unter allen Funktionen aus $C^2(\Omega) \cap C^0(\bar{\Omega})$ mit Nullrandwerten.

Beweis. Die Greensche Formel besagt

$$\int_\Omega v\Delta u + \nabla v \nabla u dx = \int_{\partial\Omega} v \frac{\partial u}{\partial \vec{n}} ds$$

Angewandt auf $\int_\Omega v\partial_i w dx$ mit $w := a_{ik}\partial_k u$ liefert

$$\int_\Omega v\partial_i w + \partial_i v w dx = \int_{\partial\Omega} v \frac{\partial w}{\partial \vec{n}} ds$$

Mit $v|_{\partial\Omega} = 0$ folgt

$$\int_\Omega v\partial_i w = -\int_\Omega \partial_i v w$$

und mit $w = a_{ik}\partial_k u$

$$\int_\Omega v\partial_i(a_{ik}\partial_k u)dx = -\int_\Omega a_{ik}\partial_i v \partial_k u dx.$$

Setze

$$a(u,v) := \int_\Omega \sum_{i,k} a_{ik}\partial_i u \partial_k v + a_0 uv dx$$

$$f(v) := \int_\Omega fv dx$$

und summiere über i und j:

$$\Rightarrow \int_\Omega \sum_{i,j} v\partial_i(a_i k \partial_k u)dx = \underbrace{-\int_\Omega \sum_{i,j} a_{ik}\partial_i v \partial_k u dx}_{=a(u,v)-a_0 uv}$$

Eine Erweiterung um $\int_\Omega fv dx$ liefert schließlich

$$\begin{aligned} a(u,v) - f(v) &= \int_\Omega v\left(-\sum \partial_i(a_{ik}\partial_k u) + a_0 u - f\right) dx \\ &= \int_\Omega v(Lu - f)dx \underset{\text{falls } Lu=f}{=} 0 \end{aligned}$$

□

Der Charakterisierungssatz besagt also

$$u \text{ ist Lösung des Variationsproblems und}$$
$$u \in C^2(\Omega) \cap C^0(\bar{\Omega})$$
$$\Downarrow$$
$$u \text{ ist klassische Lösung}$$

Wir klären nun die Frage nach der Existenz einer Lösung.

4.2.2 Existenz und Eindeutigkeit für das Variationsproblem

Betrachte das Dirichlet-Integral (Energiefunktional)

$$J(u) := \int_\Omega |\nabla u|^2 dx = \int_\Omega \sum_{i=1}^{n} u_{x_i}^2(x) dx \tag{4.30}$$

Aus der Minimaleigenschaft folgt

$$\begin{aligned} J(u) \to \min \Longleftrightarrow \Delta u &= 0 \text{ in } \Omega \\ u &= \varphi \text{ auf } \Gamma \end{aligned}$$

Dirichlet-Prinzip

Dirichlet argumentierte, da $J(u) \geq 0$ muss für ein u das Minimum angenommen werden.
$\Rightarrow$ Es existiert eine Lösung zu

$$\begin{aligned} \Delta u &= 0 \text{ in } \Omega \\ u &= \varphi \text{ auf } \Gamma \end{aligned}$$

Einen Widerspruch zu dieser Aussage demonstrierte Weierstraß (1870):
Betrachte dazu

$$J(u) = \int_0^1 u^2(x) dx \longrightarrow \min \text{ für } u \in C^0([0,1])$$

und $u(0) = 1$, $u(1) = 0$.
Wir definieren nun die Funktionenfolge

$$u_n(x) = \begin{cases} 1 - nx & 0 \leq x \leq \frac{1}{n} \\ 0 & x > \frac{1}{n} \end{cases} \tag{4.31}$$

Es gilt $\lim_{n \to 0} u_n = 0$, d.h. das Infimum von $J(u)$ ist

$$\inf J(u) = 0,$$

wird aber für stetige Funktionen nicht angenommen. Dies ist ein Widerspruch zum Dirichlet-Prinzip. Dieses Beispiel zeigt, um eine eindeutige Lösung des Variationsproblems zu garantieren, muss der Funktionenraum richtig gewählt werden. Dies besagt der folgende Satz:

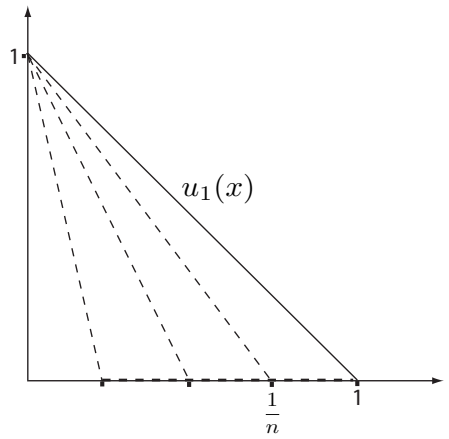

Abbildung 4.1: Folgenfunktionen $u_n(x)$

Existenz und Eindeutigkeit

Satz 16. *(Lax-Milgram)*
Sei V eine abgeschlossene, konvexe Menge in dem Hilbertraum H und $a : H \times H \longrightarrow \mathbb{R}$ eine elliptische Bilinearform. Für jedes l aus H' hat das Variationsproblem

$$J(v) := \frac{1}{2}a(v,v) - \langle l, v\rangle \longrightarrow \min \tag{4.32}$$

genau eine Lösung in V.

Beweis. J ist nach unten beschränkt:

$$\begin{aligned} J(v) &\geq \frac{1}{2}C_E\|v\|^2 - \|l\|\|v\| \\ &= \frac{1}{2}C_E\|v\|^2 - \|l\|\|v\| + \|l\|^2 - \|l\|^2 \\ &= \frac{1}{2C_E}(C_E^2\|v\|^2 - 2C_E\|l\|\|v\| + \|l\|^2) - \frac{1}{2C_E}\|l\|^2 \\ &= \frac{1}{2C_E}(C_E\|v\| - \|l\|)^2 - \frac{1}{2C_E}\|l\|^2 \geq -\frac{\|l\|^2}{2C_E} \end{aligned}$$

Wir setzen nun $c_1 := \inf\{J(v) : v \in V\}$. Sei (v_n) eine Minimalfolge, d.h.

$$\lim_{n\to\infty} v_n = c_1.$$

Es gilt: $C_E\|v_n - v_m\|^2 \leq a(v_n - v_m, v_n - v_m)$, da $a(.,.)$ elliptisch ist. Mit Hilfe der Parallelogramgleichung

$$\|v_n + v_m\|^2 + \|v_n - v_m\|^2 = 2(\|v_n\|^2 + \|v_m\|^2) \tag{4.33}$$

folgt

$$\begin{aligned} a(v_n+v_m,v_n+v_m)+a(v_n-v_m,v_n-v_m) &= 2(a(v_n,v_n)+a(v_m,v_m)) \\ \Leftrightarrow a(v_n-v_m,v_n-v_m) &= 2a(v_n,v_n)+2a(v_m,v_m)-a(v_n+v_m,v_n+v_m) \end{aligned}$$

Angewandt auf

$$\begin{aligned} C_E\|v_n-v_m\|^2 &\leq a(v_n-v_m,v_n-v_m) \\ &= 2a(v_n,v_n)+2a(v_m,v_m)-a(v_n+v_m,v_n+v_m) \end{aligned}$$

Mit $J(v)=\frac{1}{2}a(v,v)-\langle l,v\rangle$ erhalten wir

$$\begin{aligned} 4J(v_n) &= 2a(v_n,v_n)-4\langle l,v_n\rangle \\ &= 4J(v_n)-4\langle l,v_n\rangle+4J(v_m)-4\langle l,v_m\rangle-\left(8\cdot J\left(\frac{v_n+v_m}{2}\right)-4\langle l,v_n+v_m\rangle\right) \\ &= 4J(v_n)+4J(v_m)-8J\left(\frac{v_n+v_m}{2}\right)\leq 4J(v_n)+4J(v_m)-8c_1 \end{aligned}$$

Die Forderung, dass V konvex ist, liefert

$$\frac{v_n+v_m}{2}\in V$$

(v_n) ist eine Minimalfolge $\Rightarrow \lim_{n\to\infty} 4J(v_{n,m})=4c_1$.

$$\Rightarrow C_E\|v_n-v_m\|^2\longrightarrow 0 \text{ für } n,m\to\infty,$$

also $\|v_n-v_m\|^2\longrightarrow 0$.
$\Rightarrow (v_n)$ ist eine Cauchy-Folge in H. H ist ein Hilbertraum, deshalb existiert ein $u\in H$ und mit V abgeschlossen auch $u\in V$ mit $\lim_{n\to\infty} v_n=u$ und $J(u)=\lim_{n\to\infty} J(v_n)=\inf_{v\in V} J(v)$.
$\Rightarrow J(v)=\frac{1}{2}a(v,v)-\langle l,v\rangle\longrightarrow$ min hat eine Lösung $u\in V$.

Eindeutigkeit: Seien u_1 und u_2 Lösungen, also Grenzwerte von Minimalfolgen. Dann ist auch $u_1,u_2,u_1,u_2,\ldots$ eine Minimalfolge.
$\Rightarrow u_1,u_2,u_1,u_2,\ldots$ ist Cauchy-Folge. Dies ist aber nur der Fall wenn $u_1=u_2\Rightarrow$ Eindeutigkeit.

□

Bemerkung 21. *Folgende Schlussfolgerung können gemacht werden:*

1. *Mit $V=H$ folgt: Zu jedem $l\in H'$ gibt es ein $u\in H$ mit*

$$a(u,v)=\langle l,v\rangle \quad \forall v\in H$$

2. *Sei $a(u,v):=(u,v)$. Der Rieszsche Darstellungssatz liefert: Zu jedem $l\in H'$ gibt es ein $u\in H$ mit*

$$(u,v)=\langle l,v\rangle \quad \forall v\in H$$

Damit erhält man eine Einbettung

$$\begin{aligned} H' &\longrightarrow H \\ l &\longmapsto v \end{aligned}$$

4.2.3 Schwache Lösung des Randwertproblems

Bei Finite Differenzenverfahren zeigte sich in den Konvergenzbeweisen die Forderung nach starker Regularität der Lösungsfunktionen. Differenzierbarkeit der Lösung wurde im klassischen Sinne aufgefasst. Um die funktionalanalytischen Eigenschaften von Hilbert-Räumen mit Integrabilitätsbedingungen (statt klassische Differenzierbarkeit) zu nutzen führen wir hier den Begriff der schwachen Differenzierbarkeit ein. Wir werden sehen, dass dieser Begriff in großen Teilen von Ω mit der klassischen Differenzierbarkeit übereinstimmen kann, Ausnahmemengen jedoch erlaubt sind, in denen Lösungsfunktionen nicht differenzierbar sein müssen. Dies ermöglich es uns später numerisch einfach handhabbare Funktionenräume zu definieren.

Definition 45. *Eine Funktion $u \in H_0^1(\Omega)$ heißt* schwache Lösung *der Randwertaufgabe 2. Ordnung*

$$\begin{aligned} Lu &= f \text{ in } \Omega \\ u &= 0 \text{ auf } \Gamma \end{aligned}$$

wenn gilt:

$$a(u,v) = (f,v)_0 \quad \forall v \in H_0^1(\Omega) \tag{4.34}$$

$$a(u,v) := \int_\Omega \sum_{i,k} a_{ik}\partial_i u \partial_k v + a_0 uv dx \tag{4.35}$$

Satz 17. *Sei L ein Differentialoperator 2. Ordnung. Dann hat das Randwertproblem*

$$\begin{aligned} Lu &= f \text{ in } \Omega \\ u &= 0 \text{ auf } \Gamma \end{aligned}$$

stets eine schwache Lösung in $H_0^1(\Omega)$ und ist Minimum des Problems

$$\frac{1}{2}a(v,v) - (f,v)_0 \longrightarrow \min \text{ in } H_0^1(\Omega)$$

Beispiel 10. *Für das Modellproblem*

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

ist die zugehörige Bilinearform

$$a(u,v) = \int_\Omega \nabla u \nabla v dx.$$

Das Variationsproblem lässt sich dann folgendermaßen stellen:
Finde $u \in H_0^1(\Omega)$, sodass

$$(\nabla u, \nabla v)_0 = (f,v)_0 \quad \forall v \in H_0^1(\Omega)$$

Dieses u findet man als Lösung des Minimierungsproblems

$$\frac{1}{2}\int_\Omega \nabla u \nabla v dx - (f,v)_0 \longrightarrow \min$$

4.2.4 Variationsproblem der Neumann-Randwertaufgabe

Wir betrachten das Problem

$$\begin{aligned} Lu &= f \text{ in } \Omega \\ \sum_{i,k} n_i a_{ik} \partial_k u &= g \text{ auf } \Gamma \end{aligned}$$

wobei n_i der i-te Anteil des lokalen Normalenvektors ist. Mit $f \in L_2(\Omega)$ und $g \in L_2(\Gamma)$ ist das lineare Funktional

$$\langle l, v \rangle := \int_\Omega f v dx + \int_\Gamma g v dx$$

definiert. Das Variationsproblem hat dann die Gestalt:
Finde $u \in H^1(\Omega)$, sodass

$$a(u, v) = (f, v)_{0,\Omega} + (g, v)_{0,\Gamma} \quad \forall v \in H^1(\Omega)$$

gilt.

Satz 18. *Die Variationsaufgabe auf Ω mit stückweise glattem Rand und erfüllter Kegelbedingung*

$$J(v) := \frac{1}{2} a(v, v) - (f, v)_{0,\Omega} - (g, v)_{0,\Gamma} \longrightarrow \min$$

ist äquivalent zu

$$\begin{aligned} Lu &= f \text{ in } \Omega \\ \sum_{i,k} n_i a_{ik} \partial_k u &= g \text{ auf } \Gamma \end{aligned}$$

falls $u \in C^2(\Omega) \cap C^1(\bar{\Omega})$.

Beachte: Die Variationsaufgabe hat in $H^1(\Omega)$ eine eindeutige Lösung (nicht notwendigerweise in $C^2(\Omega) \cap C^1(\bar{\Omega})$).

4.3 Galerkin-Verfahren

Um die schwache Lösung der Randwertaufgabe zu finden muss ein Minimierungsproblem gelöst werden, und zwar über einen unendlich dimensionalen Hilbertraum. Um einen numerischen Ansatz für dieses Problem zu finden, ist die Idee der Galerkin-Verfahren, den Funktionenraum endlich-dimensional zu approximieren. Wir werden also z.B. $H^1(\Omega)$ durch einen approximierten Raum V_h ersetzen.
Wir betrachten dazu die Variationsaufgabe
Finde $u \in H^1(\Omega)$ mit

$$a(u, v) = (f, v).$$

Beispiel 11. *Sei die Randwertaufgabe*

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega \\ u &= u_0 \text{ auf } \Gamma \end{aligned}$$

gegeben. Die schwache Formulierung lautet:
Suche $u \in H_0^1(\Omega)$ *mit*

$$\int_\Omega \nabla u \nabla v dx = \int_\Omega f v dx \quad \forall v \in H_0^1(\Omega)$$

mit

$$\begin{aligned} a(u,v) &:= \int_\Omega \nabla u \nabla v dx \\ (f,v) &:= \int_\Omega f v dx \end{aligned}$$

Problem. Hier muss ein Minimierungsproblem in einem unendlich-dimensionalen Raum $H^m(\Omega)$ bzw. $H_0^m(\Omega)$ gelöst werden. Für die numerische Behandlung des Minimierungsproblems benötigen wir einen endlich-dimensionalen Raum.

Idee 6. *Ersetze den Lösungsraum* V *durch* V_h*, wobei* V_h *endlich-dimensional ist, also durch eine endliche Basis darstellbar ist.*

Damit wird das obige Minimierungsproblem zu

$$J(v) := \frac{1}{2} a(v,v) - \langle l, v \rangle \longrightarrow \min_{V_h} \tag{4.36}$$

$u_h \in V_h$ ist Lösung des Minimierungsproblems, wenn gilt

$$a(u_h, v) = \langle l, v \rangle \quad \forall v \in V_h.$$

Das endliche Problem

Sei $\{\psi_1, \psi_2, \ldots, \psi_n\}$ eine Basis von V_h. Dann ist

$$a(u_h, v) = \langle l, v \rangle \quad \forall v \in V_h$$

äquivalent zu

$$a(u_h, \psi_i) = \langle l, \psi_i \rangle \quad i = 1, 2, \ldots, N.$$

$u_h \in V_h$ lässt sich als Linearkombination der ψ_i darstellen:

$$u_h = \sum_{k=1}^{N} z_k \psi_k \tag{4.37}$$

mit zu berechnenden Koeffizienten z_k. Dies führt zu einem Gleichungssystem

$$\sum_{k=1}^{N} a(\psi_k, \psi_i) z_k = \langle l, \psi_i \rangle \quad i = 1, 2, \ldots N$$

durch Einsetzen von $u_h = \sum_{k=1}^{N} z_k \psi_k$ in $a(u_h, \psi_i) = \langle l, \psi_i \rangle \, \forall i = 1, \ldots, N$. Mit $A_{ik} := a(\psi_k, \psi_i)$ und $b_i := \langle , \psi_i \rangle$ lässt sich das Gleichungssystem schreiben als

$$Az = b$$

Bemerkung 22. *Falls $a(.,.)$ V-elliptisch ist, so ist die Matrix A positiv definit:*

$$\begin{aligned} z^t A z &= \sum_{i,k} z_i A_{ik} z_k = a\left(\sum_k z_k \psi_k, \sum_i z_i \psi_i \right) \\ &= a(u_h, u_h) \geq C_E \|u_h\|_V^2 \end{aligned}$$

Frage 7. *Wie gut approximiert $u_h \in V_h$ die Lösung $u \in V$?*

Lemma 19. *(Céa-Lemma)*
Die Bilinearform a sei V-elliptisch und $H_0^m(\Omega) \subset V \subset H^m(\Omega)$. Dann gilt:

$$\|u - u_h\|_m \leq \frac{C_S}{C_E} \inf_{v_h \in V_h} \|u - v_h\|_m \tag{4.38}$$

Beweis. Es gilt:

$$\begin{aligned} a(u, v) &= \langle l, v \rangle \quad \forall v \in V \\ a(u_h, v) &= \langle l, v \rangle \quad \forall v \in V_h \end{aligned}$$

Da $V_h \subset V$ können wir schreiben

$$a(u, v) - a(u_h, v) = a(u - u_h, v) = 0 \text{ für } v \in V_h.$$

Mit $v_h \in V_h$ können wir $v \in V_h$ schreiben als

$$v = v_h - u_h \in V_h.$$

$\Rightarrow a(u - u_h, v_h - u_h) = 0$. Ferner ist $a(.,.)$ elliptisch, d.h.

$$\begin{aligned} C_E \|u - u_h\|_m^2 &\leq a(u - u_h, u - u_h) = a(u - u_h, u - u_h) + a(u - u_h, v_h - u_h) \\ &\leq C_S \|u - u_h\|_m \|u - v_h\|_m \end{aligned}$$

Daraus schließen wir

$$\begin{aligned} \|u - u_h\|_m &\leq \frac{C_S}{C_E} \|u - v_h\|_m \quad \forall v_h \in V_h \\ \Rightarrow \|u - u_h\|_m &\leq \frac{C_S}{C_E} \inf_{v_h \in V_h} \|u - v_h\|_m \end{aligned}$$

□

Bemerkung 23. *Aufgrund von $a(u - u_h, v) = 0$ ist der Approximationsfehler $u - u_h$ orthogonal zu V.*

Folgerung aus dem Céa-Lemma

Je besser V_h den Raum V approximiert, desto kleiner wird der Abstand zwischen u und u_h

$$\|u - u_h\|_m.$$

Frage 8. *Wie wählen wir V_h, bzw. die Basis von V_h?*

4.4 Finite Elemente Verfahren

Die Frage nach einer geeigneten Basis für V_h beantwortet das Finite Elemente Verfahren durch

1. Basisfunktionen niedriger Ordnung
2. lokale Funktionen, d.h. Funktionen mit kompaktem Träger

Die Approximationsgüte von Funktionen mit niedriger Ordnung ist zwar nur in kleinen Bereichen vertretbar, durch die Einfachheit der Funktionen kann eine bessere Approximation jedoch durch Gitterverfeinerung, auch in Kombination mit Erhöhung der Polynomordnung, erreicht werden. Durch Funktionen mit kompaktem Träger lassen sich die Einträge der Steifigkeitsmatrix und der rechten Seite durch lokale Integralapproximation einfach lösen und so das finale lineare Gleichungssystem aufstellen.
Dieses kann mitunter sehr groß werden, da die Gitterfeinheit aufgrund der obigen Grundannahmen für die Basisfunktionen hoch werden kann. Courant demonstrierte 1943 folgendes grundlegende Beispiel für das Vorgehen bei der Finite Elemente Methode.

4.4.1 Beispiel von Courant

Betrachte das Poisson-Problem

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega = (0,1) \times (0,1) \\ u &= 0 \text{ auf } \Gamma \end{aligned}$$

Ω wird dabei in 8 Dreiecke I-VIII zerlegt, mit den Knoten L, R, O, U, LO, RU und Z. Wir wählen V_h folgendermaßen

$$V_h = \{v \in C(\bar{\Omega}) : v \text{ linear und } v|_\Gamma = 0\}.$$

v lässt sich also in jedem Gitterdreieck darstellen als

$$v(x, y) = a + bx + cy.$$

Da wir die Werte in den Gitterknoten kennen, lassen sich a, b und c eindeutig bestimmen. Bei N inneren Gitterknoten ist $\dim V_h = N$. Wir benötigen also N Basisfunktionen für V_h. Die Basisfunktionen $\{\psi_i\}_{i=1}^N$ seien definiert durch

$$\psi_i(K_j) = \delta_{ij} \text{ mit } K_j = Knoten(j),\ j = 1, \ldots, N.$$

Die Basisfunktion über dem inneren Knoten Z ist damit

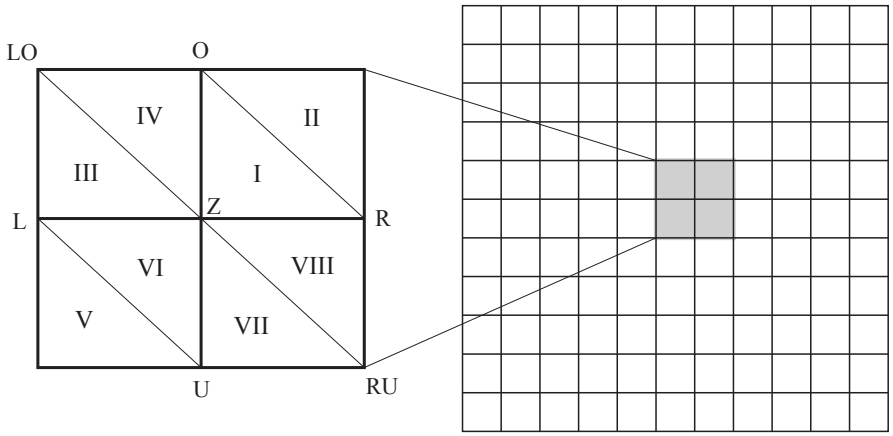

Abbildung 4.2: Unterteilung des Gebiets in Dreiecke

	I	II	III	IV	V	VI	VII	VIII
$\partial_1 \psi_Z$	$\frac{-1}{h}$	0	$\frac{1}{h}$	0	0	$\frac{1}{h}$	0	$\frac{-1}{h}$
$\partial_2 \psi_Z$	$\frac{-1}{h}$	0	0	$\frac{-1}{h}$	0	$\frac{1}{h}$	$\frac{1}{h}$	0

Tabelle 4.1: *Ableitungen von ψ_Z in die erste und zweite Raumrichtung*

1. linear in jedem Dreieck
2. Null auf den umliegenden Knoten

erfüllt also die Finite Elemente Richtlinien *niedrige Ordnung* und *lokal*. Wir können nun die Ableitungen der Basisfunktion ψ_Z in den Dreiecken I-VIII berechnen (siehe Tabelle 4.4.1)
Zu lösen ist das Gleichungssystem

$$Au = b$$

mit

$$A_{ij} = a(\psi_i, \psi_j).$$

In unserem Beispiel benötigen wir also $a(\psi_Z, \psi_Z)$, $a(\psi_Z, \psi_O)$, $a(\psi_Z, \psi_U)$, $a(\psi_Z, \psi_L)$, $a(\psi_Z, \psi_R)$, $a(\psi_Z, \psi_{LO})$ und $a(\psi_Z, \psi_{RU})$.

Für $a(\psi_Z, \psi_Z)$ gilt:

$$\begin{aligned} a(\psi_Z, \psi_Z) &= \int_\Omega (\nabla \psi_Z)^2 dxdy = \int_{I-VIII} (\nabla \psi_Z)^2 dxdy \\ &= 2 \cdot \int_{I+III+IV} \left((\partial_1 \psi_Z)^2 + (\partial_2 \psi_Z)^2 \right) \\ &= 2 \cdot \int_{I+III} (\partial_1 \psi_Z)^2 dxdy + 2 \cdot \int_{I+IV} (\partial_2 \psi_Z)^2 dxdy \\ &= \frac{2}{h^2} \int_{I+III} dxdy + \frac{2}{h^2} \int_{I+IV} dxdy \\ \Rightarrow a(\psi_Z, \psi_Z) &= 4 \end{aligned}$$

Für $a(\psi_Z, \psi_O)$ gilt:

$$\begin{aligned} a(\psi_Z, \psi_O) &= \int_{I-VIII} \nabla \psi_Z \nabla \psi_O dxdy \\ &= \int_{I+IV} \nabla \psi_Z \nabla \psi_O dxdy = \int_{I+IV} \partial_1 \psi_Z \partial_1 \psi_O + \partial_2 \psi_Z \partial_2 \psi_O dxdy \\ &= \int_{I+IV} \partial_2 \psi_Z \partial_2 \psi_O dxdy = \int_{I+IV} -\frac{1}{h} \cdot \frac{1}{h} dxdy \\ &= -\frac{1}{h} \cdot \int_{I+IV} dxdy = -1 \end{aligned}$$

Aus Symmetriegründen folgt:

$$a(\psi_Z, \psi_O) = a(\psi_Z, \psi_U) = a(\psi_Z, \psi_L) = a(\psi_Z, \psi_R) = -1$$

Durch Nachrechnen erhält man zudem:

$$a(\psi_Z, \psi_{RU}) = a(\psi_Z, \psi_{LO}) = 0$$

Damit erhalten wir einen Matrix-Stern der Form

$$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}$$

Bemerkung 24. *Die Identität der Sterne und damit der Systemmatrizen zwischen Finite Differenzen und Finite Elemente gilt im Allgemeinen* nicht.

Das Beispiel von Courant zeigt uns:

1. Wir benötigen eine Triangulierung mit bestimmten Eigenschaften
2. Wir benötigen Ansatzfunktionen über dem diskreten Gittergebiet

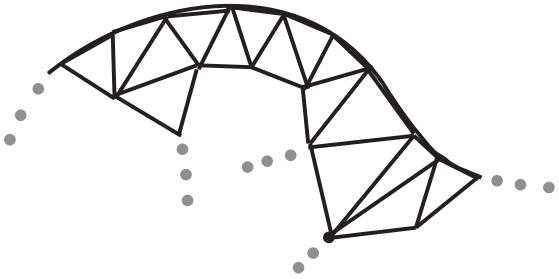

Abbildung 4.3: Triangulierung eines Gebiets

4.4.2 Triangulierung

Ein Gebiet mit gekrümmten Rändern kann lokal linear approximiert werden.

Definition 46. *(Zulässige Triangulierung)*
Eine Zerlegung $\mathcal{T} = \{T_1, T_2, \ldots, T_M\}$ *von* Ω *in Dreiecks- bzw. Viereckselemente heißt* zulässig*, wenn folgende Eigenschaften erfüllt sind:*

1. $\bar{\Omega} = \bigcup_{i=1}^{M} T_i$
2. $T_i \cap T_j$ *besteht aus genau einem Punkt, dann ist dieser ein Eckpunkt von* T_i *und* T_j
3. $T_i \cap T_j$ *mehr als ein Punkt, dann ist* $T_i \cap T_j$ *eine Kante von* T_i *und* T_j

Die Punkte 2 und 3 definieren *konforme Gitter.*

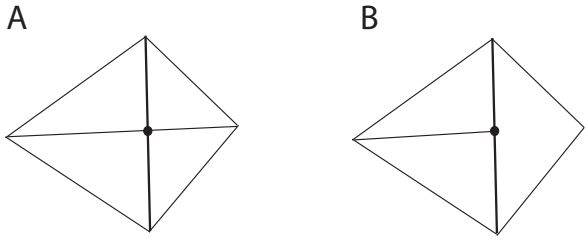

Abbildung 4.4: A: konformes Gitter, B: nicht konformes Gitter

In drei Raumdimensionen kommen verschiedene Gitterelemente zum Einsatz (siehe Abb. 4.4.2).

Definition 47. *(Finite Elemente Raum)*
Für ein Gitter Ω_h *über* $\Omega \subset \mathbb{R}^d$ *ist*

$$V_h^p(\mathcal{T}) = \{u \in H^1 : \text{ für alle } T \in \mathcal{T} : u|_T \in \mathbb{P}_p\}$$

der konforme Finite Elemente Raum der Ordnung p.

Aus dem Beispiel von Courant können wir ein allgemeines Verfahren im $\mathbb{R}^d$, $d = 1, 2, 3$, ableiten.

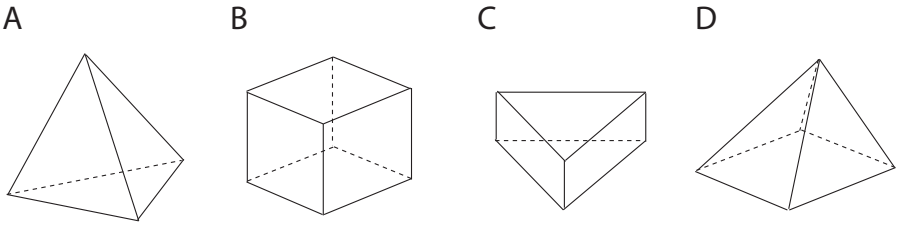

Abbildung 4.5: A: Tetraeder, B: Quader, C: Prisma, D: Pyramide

4.4.3 Finite Elemente im $\mathbb{R}^1$

Wir betrachten als Modellproblem die Helmholtz-Gleichung

$$-\Delta u + u = f \text{ mit } a(u,v) = \int_\Omega (\nabla u \nabla v + uv) dx \tag{4.39}$$

Sei $\mathcal{N} = \{a = x_0, x_1, x_2, \ldots, x_{N+1} = b\}$ die Diskretisierung des Gebiets $[a,b]$ mit Gitterweiten $h_i = x_{i+1} - x_i$ und lokalen Ansatzfunktionen φ_i für die gilt

$$\varphi_i(x_j) = \delta_{ij}$$

Damit hat die Lösung u_h die Gestalt

$$u_h = \sum_{i=1}^{N} a_i \varphi_i.$$

Berechnung von u_h auf Referenzintervallen

Die Ansatzfunktionen φ_i können durch Formfunktionen Φ_i dargestellt werden.
Durch eine bijektive affine Transformation wird das Intervall $[x_i, x_{i+1}]$ auf ein Referenzintervall $[0,1]$ abgebildet. Die Lösung u_i kann auf diese Weise immer auf dem Intervall $[0,1]$ über die Formfunktionen gelöst und dann zurücktransformiert werden. Dies hat technische Vorteile gegenüber der Berechnung auf dem Originalelement.

Transformationsabbildung

Sei $I_i = [x_i, x_{i+1}]$ und $\xi \in [0,1]$. Dann definieren

$$\begin{aligned} x_{I_i} : [0,1] &\longrightarrow I_i \\ \xi &\longmapsto x_i + h_i \xi \\ \xi_{I_i} : I_i &\longrightarrow [0,1] \\ x &\longmapsto \frac{(x - x_i)}{h_i} \end{aligned}$$

eine Bijektion zwischen $[0,1]$ und $[x_i, x_{i+1}]$. Auf dem Referenzintervall können wir unsere Lösung darstellen als

$$u_h(\xi) = \alpha_1 + \alpha_2 \xi.$$

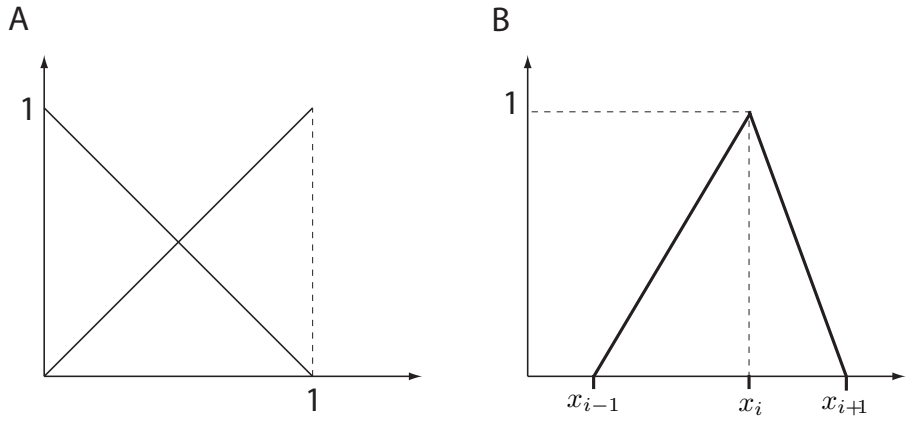

Abbildung 4.6: A: Formfunktionen über Referenzintervall, B: Ansatzfunktionen über Gitter

Dabei gilt $u_i = u_h(0) = \alpha_1$ und $u_{i+1} = u_h(1) = \alpha_1 + \alpha_2$.

$$\begin{aligned} \Rightarrow u_h(\xi) &= \alpha_1 + \alpha_2\xi = u_i + (u_{i+1} - u_i)\xi \\ &= (1-\xi)u_i + \xi u_{i+1} = u_i\Phi_1(\xi) + u_{i+1}\Phi_2(\xi) \end{aligned}$$

Bemerkung 25. *Für die Formfunktionen gilt:*

$$\forall \xi \in [0,1] : \Phi_1(\xi) + \Phi_2(\xi) = 1$$

Für die Ansatzfunktionen gilt

$$\varphi_i(x) = \begin{cases} \Phi_2(\xi(x)) & x \in I_{i-1} \\ \Phi_1(\xi(x)) & x \in I_i \\ 0 & sonst \end{cases}$$

Wir können nun auf dem Referenzintervall und abhängig von den Formfunktionen die Matrixeinträge der Systemmatrix berechnen.

Berechnung der Systemmatrix-Einträge

Zu berechnen sind die Einträge $a(\varphi_i, \varphi_j)$ der Systemmatrix A (in unserem Fall für das Helmholtz-Problem)

$$a(\varphi_i, \varphi_j) = \sum_{k=1}^{N} \int_{I_k} \nabla\varphi_i(x)\nabla\varphi_j(x) + \varphi_i(x)\varphi_j(x)\, dx.$$

Die φ_i, φ_j können durch Φ_n und Φ_m $(n, m \in \{1, 2\})$ ersetzt werden.

$$\begin{aligned}
\Rightarrow (A_{I_k})_{nm} &= \int_{I_k} \nabla_x \Phi_n(\xi(x)) \nabla_x \Phi_m(\xi(x)) + \Phi_n(\xi(x)) \Phi_m(\xi(x)) dx \\
&= (x_{k+1} - x_k) \int_0^1 \nabla_\xi \Phi_n(\xi) \xi'(x(\xi)) \xi'(x(\xi)) \nabla_\xi \Phi_m(\xi) + \Phi_n \Phi_m d\xi \\
&= h_k \int_0^1 \frac{1}{h_k^2} \nabla \Phi_n \nabla \Phi_m + \Phi_n \Phi_m d\xi
\end{aligned}$$

Man sieht, die Systemmatrix ist zusammengesetzt aus 2×2 Submatrizen A_{I_k}.
Die Matrixeinträge von A_{I_k} lauten:

$$\begin{aligned}
(A_{I_k})_{11} &= \int_0^1 \frac{1}{h_k} \nabla \Phi_1 \nabla \Phi_1 + \Phi_1 \Phi_1 \cdot h_k d\xi \\
&= \int_0^1 \frac{1}{h_k} + h_k (1 - \xi)^2 d\xi = \frac{1}{h_k} + \frac{1}{3} h_k \\
(A_{I_k})_{12} &= \int_0^1 \frac{1}{h_k} \nabla \Phi_1 \nabla \Phi_2 + \Phi_1 \Phi_2 \cdot h_k d\xi \\
&= \int_0^1 -\frac{1}{h_k} + \xi(1 - \xi) \cdot h_k d\xi \\
&= -\frac{1}{h_k} + \frac{1}{6} h_k \\
(A_{I_k})_{21} &= -\frac{1}{h_k} + \frac{1}{6} h_k \\
(A_{I_k})_{22} &= \frac{1}{h_k} + \frac{1}{3} h_k
\end{aligned}$$

$$\Rightarrow A_{I_k} = \tfrac{1}{h_k} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + h_k \begin{bmatrix} 1/3 & 1/6 \\ 1/6 & 1/3 \end{bmatrix}$$

Quadratische Elemente

Statt einer linearen Approximation können wir auch einen quadratischen Ansatz wählen. Dies erhöht die lokale Approximationsgüte. Wir stellen die Lösung $u_h(\xi)$ deshalb dar als

$$u_h(\xi) = \alpha_1 + \alpha_2 \xi + \alpha_3 \xi^2 \text{ auf } I = [0, 1].$$

Um dieses Polynom eindeutig zu bestimmen, benötigen wir 3 Stützstellen, zusätzlich zu u_i und u_{i+1} können wir z.B.

$$u_{i+\frac{1}{2}} = u_h \left(\frac{x_i + x_{i+1}}{2} \right).$$

An den Stützstellen gilt:

$$\begin{aligned} u_i &= u_h(0) = \alpha_1 \\ u_{i+1} &= u_h(1) = \alpha_1 + \alpha_2 + \alpha_3 \\ u_{i+\frac{1}{2}} &= u_h\left(\frac{1}{2}\right) = \alpha_1 + \frac{1}{2}\alpha_2 + \frac{1}{4}\alpha_3 \end{aligned}$$

Berechnung der α_i, $i = 1, \ldots, 3$ liefert

$$u_h(\xi) = u_i\Phi_1(\xi) + u_{i+1}\Phi_2(\xi) + u_{i+\frac{1}{2}}\Phi_3(\xi)$$

mit

$$\begin{aligned} \Phi_i(\xi) &= 2\left(\xi - \frac{1}{2}\right)(\xi - 1) \\ \Phi_2(\xi) &= 2\xi\left(\xi - \frac{1}{2}\right) \\ \Phi_3(\xi) &= 4\xi\,(1 - \xi\,) \end{aligned}$$

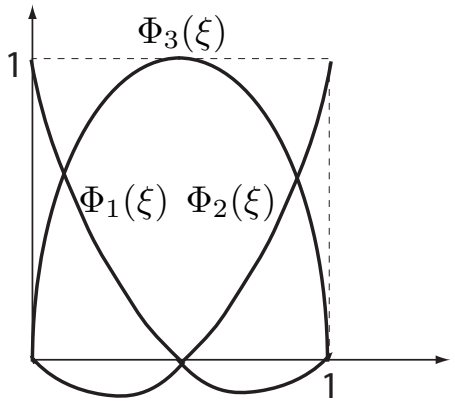

Abbildung 4.7: Ansatzfunktionen für quadratische Elemente

Die Submatrizen A_{I_k} sind jetzt 3×3-Matrizen und lassen sich analog zum linearen Fall ausrechnen.

Bemerkung 26. *Die Gradienten $\nabla\Phi_i(\xi)$ sind nun keine Konstanten über dem Referenzintervall. Man benötigt also noch ein numerischen Verfahren zur Approximation der Integrale.*

4.4.4 Finite Elemente im $\mathbb{R}^2$

Wir übertragen nun die Vorgehensweise der Finiten Elemente im $\mathbb{R}^1$ auf den zweidimensionalen Fall. Gitterelemente, die im Eindimensionalen Intervalle waren, sind nun Dreiecke oder Vierecke.

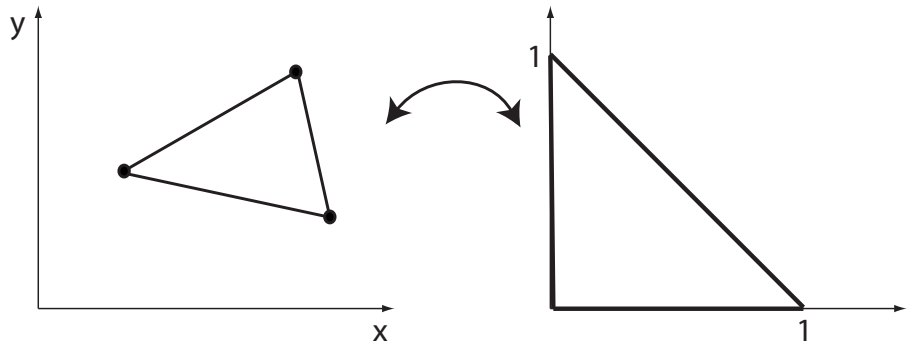

Abbildung 4.8: Affine Transformation von Dreiecken zwischen diskretisiertem Raum und Referenzelement

Die affine Transformation auf ein Referenzdreieck (bzw. Referenzviereck) ist deshalb eine bilinear, bijektive Abbildung mit der Eigenschaft (im Fall des Dreieckselements):

$$\begin{aligned} x &= x_1 + (x_2 - x_1)\xi + (x_3 - x_1)\eta \\ y &= y_1 + (y_2 - y_1)\xi + (y_3 - y_1)\eta \\ \Leftrightarrow \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix} \end{aligned}$$

Die Rücktransformation hat damit die Gestalt:

$$\begin{aligned} \begin{pmatrix} \xi \\ \eta \end{pmatrix} &= \underbrace{\begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix}^{-1}}_{A^{-1}} \begin{pmatrix} x - x_1 \\ y - y_1 \end{pmatrix} \\ &= \frac{1}{\det A} \begin{pmatrix} y_3 - y_1 & x_1 - x_3 \\ y_1 - y_2 & x_2 - x_1 \end{pmatrix} \begin{pmatrix} x - x_1 \\ y - y_1 \end{pmatrix} \end{aligned}$$

mit $\det A = (x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1)$. Für die partiellen Ableitungen gilt:

$$\begin{aligned} u_x &= u_\xi \xi_x + u_\eta \eta_x \\ u_y &= u_\xi \eta_y + u_\eta \eta_y \end{aligned}$$

und

$$\begin{aligned} \xi_x &= \frac{y_3 - y_1}{\det A} \\ \eta_x &= \frac{y_2 - y_1}{\det A} \\ \xi_y &= \frac{x_3 - x_1}{\det A} \\ \eta_y &= \frac{x_2 - x_1}{\det A} \end{aligned}$$

Bemerkung 27. $\det A$ *beschreibt dabei die Volumenänderung des Dreiecks unter der vorgegebenen Transformation:*

$$dxdy = \det A d\xi d\eta$$

Mit $\nabla\Phi_i$, $i = 1, 2, 3$ und $\xi_x, \xi_y, \eta_x, \eta_y$ sind die Komponenten der Matrix A_{I_k} vollständig bestimmbar. Dazu ist ein numerisches Integrationsverfahren notwendig.

Wahl der Φ_i: Lineare Elemente

Wir stellen die Lösung dar als

$$\begin{aligned} u_h(\xi, \eta) &= \alpha_1 + \alpha_2\xi + \alpha_3\eta \\ u_j &:= u_h(\overline{P_j}), \quad j = 1, 2, 3 \end{aligned}$$

$\overline{P_j}$ sind die Eckpunkte des Referenzdreiecks. Es gilt

$$\begin{aligned} u_1 &= u_h(0, 0) = \alpha_1 \\ u_2 &= u_h(1, 0) = \alpha_1 + \alpha_2 \\ u_3 &= u_h(0, 1) = \alpha_1 + \alpha_3 \end{aligned}$$

Eingesetzt liefert dies

$$u_h(\xi, \eta) = u_1 + (u_2 - u_1)\xi + (u_3 - u_1)\eta = (1 - \xi - \eta)u_1 + \xi u_2 + \eta u_3.$$

Mit $\Phi_1 = 1 - \xi - \eta$, $\Phi_2 = \xi$, $\Phi_3 = \eta$ folgt

$$u_h(\xi, \eta) = u_1\Phi_1 + u_2\Phi_2 + u_3\Phi_3.$$

mit

$$\begin{aligned} \Phi_i(\overline{P_j}) &= \delta_{ij} \quad i, j = 1, 2, 3 \\ \sum_{i=1}^{3} \Phi_i(\xi, \eta) &= 1 \quad \xi, \eta \in \bar{T} \end{aligned}$$

Wahl der Φ_i: Quadratische Elemente

Wir benötigen für den Fall quadratischer Ansatzfunktionen 3 weitere Stützstellen, da wir die Lösung u_h nun schreiben als

$$u_h(\xi, \eta) = \alpha_1 + \alpha_2\xi + \alpha_3\eta + \alpha_4\xi^2 + \alpha_5\xi\eta + \alpha_6\eta^2$$

Die Formfunktionen dazu lauten

$$\begin{aligned} \Phi_1 &= (1 - \xi - \eta)(1 - 2\xi + 2\eta) \\ \Phi_2 &= \xi(2\xi - 1) \\ \Phi_3 &= \eta(2\eta - 1) \\ \Phi_4 &= 4\xi(1 - \xi + \eta) \\ \Phi_5 &= 4\xi\eta \\ \Phi_6 &= 4\eta(1 - \xi - \eta) \end{aligned}$$

Bemerkung 28. *Im $\mathbb{R}^3$ geht man analog vor. Der lineare Fall definiert*

$$u_h(\xi, \eta, \zeta) = \alpha_1 + \alpha_2\xi + \alpha_3\eta + \alpha_4\zeta.$$

Das Tetraederelement im dreidimensionalen Fall liefert die nötigen 4 Knoten für die Bestimmung von $\alpha_1, \ldots, \alpha_4$.

4.4.5 Konvergenzaussagen zu Finite Elemente Verfahren

Abschätzung des Energiefehlers

Die Energienorm war definiert als

$$\|u\|_a := \sqrt{a(u,u)}$$

Wir wissen aus der Minimaleigenschaft

$$\begin{aligned} a(u-u_h, u-u_h) &= \min_{v_h \in V_h} a(u-v_h, u-v_h) \\ \Rightarrow \|u-u_h\|_a &= \min_{v_h \in V_h} \|u-v_h\|_a \end{aligned}$$

Satz 19. *Sei $\Omega \subset \mathbb{R}^d$, $d \leq 3$ konvex und $u \in H^2(\Omega)$ die schwache Lösung der Poisson-Gleichung. Sei $u_h \in V_h$ die Finite Elemente Lösung von*

$$a(u_h, v_h) = \langle f, v_h \rangle, \quad v_h \in V_h \subset H_0^1(\Omega)$$

mit linearen, konformen finiten Elementen. Dann gilt für den Diskretisierungsfehler

$$\|u-u_h\|_a \leq c \cdot h \cdot \|f\|_{L^2(\Omega)}, \quad f \in L^2(\Omega) \tag{4.40}$$

Bemerkung 29. *Falls für alle $f \in L^2(\Omega)$ gilt*

$$\|u\|_{H^2(\Omega)} \leq c\|f\|_{L^2(\Omega)}$$

*dann heißt das Problem H^2-*regulär*.*

Abschätzung des L^2-Fehlers

Satz 20. *Sei $\Omega \subset \mathbb{R}^d$, $d \leq 3$ und $u \in H^2(\Omega)$ schwache H^2-reguläre Lösung der Poisson-Gleichung. Dann gilt*

$$\|u-u_h\|_{L^2(\Omega)} \leq c \cdot h\|u-u_h\|_a \tag{4.41}$$

$$\|u-u_h\|_{L^2(\Omega)} \leq c \cdot h^2\|f\|_{L^2(\Omega)} \tag{4.42}$$

5 Ausblick

5.1 Finite Volumen Verfahren

Die Idee der Finite Volumen Diskretisierung ist sehr ähnlich zu den Finite Elemente Verfahren. Das Finite Elemente Konzept wird jedoch nicht auf das bisherige Gitter Ω_h angewandt, sondern auf ein verschobenenes, *duales* Gitter. Dies hat zur Folge, dass das Verfahren lokal Flusserhaltend ist. Wir erinnern uns an Konzentrationsänderungen in einem beliebigen Volumenelement die, ohne Quellen oder Senken im Volumen, gleich dem Fluss über den Elementrand sind:

$$\int_B \frac{\partial u}{\partial t} d\vec{x} = \int_B -\Delta u d\vec{x} = \int_{\partial B} F \cdot \vec{n} ds$$

Wählen wir ein solches Volumenelement über jedem Knoten in Ω_h, z.B. durch Verbinden von Kantenmittelpunkten und Flächenschwerpunkte, so erhält man sogenannte *Voronoie-Elemente* die zusammengesetzt ein konformes duales Gitter über dem Gebiet Ω erzeugen.

$$\Rightarrow -\int_\Omega div(\nabla u) dx = -\sum_{i=1}^{M} \int_{B_i} div(\nabla u) dx = \sum_{i=1}^{M} \int_{B_i} f dx$$

Für festes i gilt:

$$\begin{aligned} -\int_{B_i} div(\nabla u) dx &= \int_{B_i} f dx \\ \Leftrightarrow -\int_{\partial B_i} \nabla u \cdot \vec{n} dx &= \int_{B_i} f dx \\ \Leftrightarrow -\int_{\partial B_i} \frac{\partial u}{\partial \vec{n}} dx &= \int_{B_i} f dx \end{aligned}$$

Dies führt zu einem Gleichungssystem

$$K_h^{FV} u_h^{FV} = f_h^{FV}$$

mit

$$\left(K_h^{FV}\right)_{B_i} = -\int_{\partial B_i} \frac{\partial \varphi_{B_i}^{FV}}{\partial \vec{n}} ds$$

5.2 Lösen von linearen Gleichungssystemen

Aus allen besprochenen Diskretisierungsverfahren gewinnen wir ein, mitunter sehr großes, lineares Gleichungssystem

$$K \cdot u_h = b,$$

wobei wir am Ende u_h als Lösung des Problems berechnen wollen. D.h. wir benötigen die Darstellung

$$u_h = K^{-1} b,$$

müssen also K invertieren. Das Problem in realen Fällen ist jedoch, dass eine exakte Invertierung vom Rechen- und somit Zeitaufwand nicht vertretbar ist. Ein Ansatz zur Lösung des Gleichungssystems ist, die Inverse von A zu approximieren und ein iteratives Verfahren anzusetzen. Welche Vor- und Nachteile solche Verfahren haben und wie man effizient große lineare Systeme löst, ist Thema einer Folgevorlesung.