Kapitel 1

Einleitung

Partielle Differentialgleichungen (*partial differential equations - PDEs*) gehören zu den am häufigsten auftretenden mathematischen Modellen realer Prozesse. Verschiedenste Prozesse wie Wärmeleitung, Ausbreitung von Wasser- oder Schallwellen, Strömungen, Dynamik von biologischen Populationen werden heute mit PDEs modelliert, und immer neue Anwendung wie die Preisbestimmung von Finanzprodukten oder Glättung von Bildern und Computergraphiken kommen dazu.

In den wenigsten Fällen ist die analytische Lösung dieser Gleichungen möglich, und eine numerische Lösung wird nötig. Dazu ersetzt man die Gleichungen durch Gleichungssysteme in $\mathbb{R}^n$ (*Diskretisierung*), mit möglichst grossem n um die (unendlichdimensionale) Differentialgleichungen sinnvoll approximieren können. Nach der Diskretisierung verbleibt noch die Aufgabe, das endlichdimensionale Problem numerisch zu lösen, was meist eine weitere Herausforderung wegen der Grösse der Gleichungssysteme darstellt. In dieser Vorlesung werden wir uns sowohl mit verschiedenen Diskretisierungsverfahren als auch mit der effizienten Lösung der diskretisierten Probleme befassen. Wie auch bei der Theorie der partiellen Differentialgleichungen ist meist eine Unterscheidung nach Typ notwendig, da sich die unterschiedlichen Eigenschaften elliptischer, parabolischer, und hyperbolischer Gleichungen auch in der Numerik niederschlagen. Im folgenden werden wir für drei einfache Beispiele die Grundprobleme darstellen.

1.1 Beispiele der Numerik partieller Differentialgleichungen

1.1.1 Elliptische Probleme: Poisson Gleichung

Wir beginnen mit der einfachsten Form einer elliptischen Differentialgleichung, nämlich einem Randwertproblem für die eindimensionale Poisson-Gleichung.

$$-\frac{\partial^2 u}{\partial x^2}(x) = f(x), \qquad x \in (0,1), u(0) = 0, u(1) = 0. \tag{1.1}$$

Streng genommen ist (1.1) nicht einmal eine partielle, sondern nur eine gewöhnliche Differentialgleichung, aber dennoch (oder gerade deswegen) ist dieses Problem gut geeignet um die Grundzüge der Numerik elliptischer Randwertprobleme darzustellen.

Der erste Schritt in den meisten Diskretisierungsverfahren (und wir werden hier nur solche behandeln) ist die Auswahl eines geeigneten Gitters auf dem gegebenen Gebiet, d.h. auf dem Intervall $(0,1)$ im Fall von (1.1). Die einfachste Wahl ist ein reguläres Gitter mit den Punkten

$x_j = j/(n+1)$, $j = 0, \ldots, n+1$. Als Gitterfeinheit h bezeichnen wir den maximalen Abstand zwischen benachbarten Punkten, d.h. $h = \frac{1}{n+1}$.

Die erste Diskretisierungsart, die wir diskutieren werden, sind finite Differenzen, d.h., wir versuchen u durch eine Funktion u^h zu approximieren, die wir in den Gitterpunkten berechnen, d.h, wir suchen einen Vektor $(u_j^h)_{j=0,\ldots,n+1}$ wobei $u_j^h = u^h(x_j)$. Wir sehen sofort, dass wir durch die Randbedingungen zwei Werte sofort berechnen können, nämlich $u_0^h = u^h(0) = 0$ und $u_{n+1}^h = u^h(1) = 0$. Also eliminieren wir diese zwei Unbekannten und suchen nur nach dem Vektor $U_h = (u_j^h)_{j=1,\ldots,n}$.

Um nun die Werte von u_j^h an den inneren Gitterpunkten zu berechnen, konstruieren wir finite Differenzen mittels Taylorentwicklung. Für eine glatte Funktion φ gilt ja (beachte $h = x_{j+1} - x_j = x_j - x_{j-1}$)

$$\begin{aligned} \varphi(x_{j+1}) &= \varphi(x_j) + \varphi'(x_j)h + \frac{1}{2}\varphi''(x_j)h^2 + \frac{1}{6}\varphi'''(x_j)h^3 + \mathcal{O}(h^4) \\ \varphi(x_{j-1}) &= \varphi(x_j) - \varphi'(x_j)h + \frac{1}{2}\varphi''(x_j)h^2 - \frac{1}{6}\varphi'''(x_j)h^3 + \mathcal{O}(h^4). \end{aligned}$$

Addieren wir diese Gleichungen und dividieren durch h^2, so erhalten wir die Formel

$$\varphi''(x_j) = \frac{1}{h^2}\left(\varphi(x_{j+1}) - 2\varphi(x_j) + \varphi(x_{j-1})\right) + \mathcal{O}(h^2).$$

Daraus erhalten wir den klassischen Differenzenquotienten zur Approximation der zweiten Ableitung, d.h.,

$$\frac{\partial^2 u^h}{\partial x^2}(x_j) \approx \frac{1}{h^2}(u_{j+1}^h - 2u_j^h + u_{j-1}^h)$$

und wir ersetzen (1.1) durch die diskretisierte Version

$$-\frac{1}{h^2}(u_{j+1}^h - 2u_j^h + u_{j-1}^h) = f(x_j) := f_j, \qquad j = 1, \ldots, n. \tag{1.2}$$

Mit der Matrix $\mathbf{K}_h \in \mathbb{R}^{n\times n}$

$$\mathbf{K}_h = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & 0 & \ldots & 0 & 0 \\ -1 & 2 & -1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 2 & -1 & \ldots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & -1 & 0 \\ 0 & 0 & 0 & 0 & \ldots & 2 & -1 \\ 0 & 0 & 0 & 0 & \ldots & -1 & 2 \end{pmatrix} \tag{1.3}$$

und dem Vektor $F_h = (f(x_j))_{j=1,\ldots,n}$ können wir (1.2) in der Form

$$\mathbf{K}_h U_h = F_h. \tag{1.4}$$

schreiben.

Wir sehen sofort, dass n gross und damit h klein sein muss, damit die obige Taylorentwicklung und Approximation sinnvoll ist. Damit erhalten wir ein grosses lineares Gleichungssystem. Wir beobachten die folgenden Eigenschaften der Systemmatrix $\mathbf{K}_h$:

- Die Matrix $\mathbf{K}_h$ ist *dünnbesetzt*, d.h. nur ein kleiner Teil der Einträge ist von Null verschieden. Wie wir sehen werden, ist diese Eigenschaft kein Zufall, sondern tritt bei allen Verfahren die wir kennen lernen auf. Der Grund dafür ist die Lokalität der Differentialoperatoren, es ist intuitiv einleuchtend dass Funktionswerte in weiter entfernten Gitterpunkten für den Wert der Ableitung unbedeutend sind (und dies führt dann zu den Nulleinträgen in der Systemmatrix).

- Die Matrix $\mathbf{K}_h$ ist symmetrisch. Dies ist ein Resultat der Symmetrie des Differentialoperators (siehe unten).

- Die Matrix $\mathbf{K}_h$ ist positiv definit (siehe unten). Dies ist ebenfalls kein Zufall, sondern ein Resultat der Elliptizität der Gleichung.

- Die Matrix $\mathbf{K}_h$ ist monoton (siehe Kapitel 2), d.h. aus $F_h \geq 0$ folgt $U_h = \mathbf{K}_h^{-1} F_h \geq 0$. Dies ist eine Konsequenz aus dem Maximumprinzip (Monotonie) für elliptische Differentialgleichungen und wird in diesem Fall durch die Diskretisierung erhalten (was nicht für jede Diskretisierung der Fall ist).

Wir sehen also, dass das Gleichungssystem (1.4) viel Struktur aufweist, die wir auch bei der numerischen Lösung verwenden können. Die Dünnbesetztheit kann z.B. speichertechnisch ausgenutzt werden, man muss nur die Indizes und Werte der Nichtnullelemente speichern. Wir werden später sehen, dass auch andere strukturelle Eigenschaften für die effiziente Lösung von (1.4) wichtig sind.

Während die Lösung von (1.4) ein Problem der *numerischen linearen Algebra* ist, verbleiben noch die klassischen Probleme der *numerischen Analysis*:

- *Existenz und Eindeutigkeit:* Existiert die diskrete Lösung u^h (bzw. U_h) und ist sie eindeutig ?

- *Stabilität:* Bleibt die Lösung u^h für $h \to 0$ beschränkt (in einem noch zu klärenden Sinn) ?

- *Konsistenz:* Ergibt sich ein kleines Residuum, wenn man die Lösung u der Differentialgleichung in das diskretisierte Problem einsetzt, bzw. konvergiert dieses Residuum gegen Null für $h \to 0$?

- *Konvergenz:* Konvergiert für $h \to 0$ u^h gegen die kontinuierliche Lösung u (in einem noch zu klärenden Sinn) ?

- *Fehlerabschätzung:* Können wir die Fehler $u - u^h$ sinnvoll abschätzen als Funktion von h (in einer passenden Norm) ?

All diese Probleme werden wir im Laufe dieser Vorlesung behandeln. Für unser spezielles Beispiel ergibt sich natürlich Existenz und Eindeutigkeit sofort aus der positiven Definitheit der Systemmatrix. Weiter sehen wir aus dem obigen Argument für den Differenzenquotienten (angewandt auf $\varphi = u$) sofort die Konsistenz, falls u glatt genug ist. Dies ist im eindimensionalen Fall sofort durch die Eigenschaften von f nachprüfbar: $f \in C^k$ impliziert $u \in C^{k+2}$. Im mehrdimensionalen steckt hinter solchen Aussagen allerdings die komplizierte Regularitätstheorie für Lösungen partieller Differentialgleichungen.

Als alternative Methode zur Diskretisierung betrachten wir *finite Elemente* (FE). Die Grundidee einer FE Methode ist die Berechnung einer Näherungslösung als Linearkombination gegebener Basisfunktionen

$$u^h(x) = \sum_{j=1}^{n} u_j^h \phi_j^h(x)$$

wobei die Funktionen ϕ_j^h einen lokalen Träger haben. Die Basisfunktionen werden ebenfalls mit einem Gitter assoziert und erfüllen üblicherweise die Bedingung

$$\phi_j^h(x_k) = \delta_{jk}$$

mit dem Kronecker-Symbol δ_{jk}, das durch $\delta_{jj} = 1$ und $\delta_{jk} = 0$ für $k \neq j$ definiert ist. Man sieht leicht, dass unter dieser Bedingung die Werte u_j^h tatsächlich die Funktionswerte an den Gitterpunkten sind, d.h. $u^h(x_j) = u_j^h$. Um die Werte u_j^h durch direktes Einsetzen von u_h in die Differentialgleichung zu bestimmen, bräuchte man sehr glatte Funktionen ϕ_j^h und hätte ein sehr schlecht konditioniertes Gleichungssystem zu lösen. Deshalb geht man zur schwachen Formulierung der Differentialgleichung über, die man durch Multiplikation mit einer Testfunktion, Integration und anschliessender partieller Integraton erhält. Im Fall von (1.1) ist die schwache Formulierung gegeben durch die Variationsgleichung

$$\int_0^1 \frac{\partial u}{\partial x}(x)\frac{\partial \varphi}{\partial x}(x)\ dx = \int_0^1 f(x)\varphi(x) \tag{1.5}$$

für alle hinreichend glatten Testfunktionen φ. Zur Diskretisierung verwendet man nun einen Ansatz für u^h wie oben und wählt die Basisfunktionen φ_j^h als natürliche Testfunktionen. Damit erhält man die diskrete Variationsgleichung

$$\int_0^1 \frac{\partial u^h}{\partial x}(x)\frac{\partial \varphi_k^h}{\partial x}(x)\ dx = \int_0^1 f(x)\varphi_k^h(x)\ dx, \qquad k = 1, \ldots, n. \tag{1.6}$$

Durch Einsetzen der Linearkombination für u_j^h erhalten wir

$$\sum_j \int_0^1 \frac{\partial \varphi_j^h}{\partial x}(x)\frac{\partial \varphi_k^h}{\partial x}(x)\ dx u_j^h = \int_0^1 f(x)\varphi_k^h(x)\ dx, \qquad k = 1, \ldots, n.$$

Definieren wir wieder eine Systemmatrix $\mathbf{K}_h$ und einen Vektor F_h, in diesem Fall

$$\mathbf{K}_h = \left(\int_0^1 \frac{\partial \varphi_j^h}{\partial x}(x)\frac{\partial \varphi_k^h}{\partial x}(x)\ dx \right)_{j,k=1,\ldots,n}, \qquad F_h = \left(\int_0^1 f(x)\varphi_j^h(x)\ dx \right)_{j=1,\ldots,n},$$

dann lässt sich das diskrete Problem wieder in der Form (1.4) schreiben.

Wir sehen wieder, dass die Matrix $\mathbf{K}_h$ dünnbesetzt sein wird, denn für grosse Werte von $|j-k|$ werden sich die Träger von φ_j^h und φ_k^h nicht überschneiden und damit gilt entweder $\frac{\partial \varphi_j^h}{\partial x}(x) = 0$ oder $\frac{\partial \varphi_k^h}{\partial x}(x) = 0$. Dies wird noch deutlicher für die klassische Wahl stückweise linearer Ansatzfunktionen

$$\phi_j^h(x) = \begin{cases} \frac{x - x_{j-1}}{h} & \text{falls } x \in [x_{j-1}, x_j] \\ \frac{x_{j+1} - x}{h} & \text{falls } x \in [x_j, x_{j+1}] \\ 0 & \text{falls } |x - x_j| > h \end{cases} \tag{1.7}$$

Diese Ansatzfunktionen sind nicht C^1, wie wir aber noch sehen werden genügt es die Ableitungen stückweise in den Teilintervallen zu definieren. Wir erhalten dann

$$\frac{\partial \phi_j^h}{\partial x}(x) = \begin{cases} \frac{1}{h} & \text{falls } x \in [x_{j-1}, x_j] \\ -\frac{1}{h} & \text{falls } x \in [x_j, x_{j+1}] \\ 0 & \text{falls } |x - x_j| > h \end{cases}$$

Man berechnet leicht die Systemmatrix (Übung) als

$$\mathbf{K}_h = \frac{1}{h} \begin{pmatrix} 2 & -1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 2 & -1 \\ 0 & 0 & 0 & 0 & \dots & -1 & 2 \end{pmatrix} \tag{1.8}$$

d.h. $\mathbf{K}_h$ aus (1.3) und (1.8) unterscheiden sich nur um einen Faktor h. Diese unterschiedliche Skalierung ist eine Konsequenz der Integration, die bei der Definition der schwachen bzw. FE Lösung verwendet wurde. Der Unterschied zur Diskretisierung mit finiten Differenzen wird deutlicher an der rechten Seite, die dort aus Punktauswertungen bestand, im Fall der FE Diskretisierung aber aus lokalen (gewichteten) Mittelwerten. Dadurch kann die FE Methode auch leicht für unstetige (oder sogar distributionelle) rechte Seiten angewandt werden.

Die Eigenschaften der Systemmatrizen resultierend aus finiten Differenzen bzw. finiten Elementen sind sehr ähnlich, manche Eigenschaften sind aber im FE Fall viel leichter nachzuprüfen. So sieht man sofort (auch ohne Berechnung) die Symmetrie von $\mathbf{K}_h$, da ja

$$(\mathbf{K}_h)_{jk} = \int_0^1 \frac{\partial \varphi_j^h}{\partial x}(x) \frac{\partial \varphi_k^h}{\partial x}(x)\, dx = \int_0^1 \frac{\partial \varphi_k^h}{\partial x}(x) \frac{\partial \varphi_j^h}{\partial x}(x)\, dx = (\mathbf{K}_h)_{kj}$$

gilt. Weiter sieht man sofort die positive Definitheit, da für einen Vektor $V \in \mathbb{R}^n \setminus \{0\}$ gilt

$$\begin{aligned} V \cdot (\mathbf{K}_h V) &= \sum_{j,k=1}^n v_j v_k \int_0^1 \frac{\partial \varphi_j^h}{\partial x}(x) \frac{\partial \varphi_k^h}{\partial x}(x)\, dx \\ &= \int_0^1 \left(\sum_{j=1}^n v_j \frac{\partial \varphi_j^h}{\partial x}(x) \right) \left(\sum_{j=1}^n v_k \frac{\partial \varphi_k^h}{\partial x}(x) \right) dx \\ &= \int_0^1 \left(\sum_{j=1}^n v_j \frac{\partial \varphi_j^h}{\partial x}(x) \right)^2 dx > 0. \end{aligned}$$

Auch die Stabilität der finiten Elemente Diskretisierung ist leicht nachprüfbar. Man be-

achte, dass

$$\begin{aligned}\int_0^1 \left(\frac{\partial u^h}{\partial x}\right)^2 \, dx &= \int_0^1 \left(\sum_{j=1}^n u_j^h \frac{\partial \varphi_j^h}{\partial x}(x)\right)^2 \, dx \\ &= U_h \cdot (\mathbf{K}_h U_h) = U_h \cdot F_h \\ &= \sum_{j=1}^n \int_0^1 u_j^h \varphi_j^h(x) f(x) \, dx \\ &= \int_0^1 u^h(x) f(x) \, dx \\ &\leq \sqrt{\int_0^1 |u^h(x)|^2 \, dx} \sqrt{\int_0^1 |f(x)|^2 \, dx},\end{aligned}$$

wobei wir die Cauchy-Schwarz Ungleichung in $L^2([0,1])$ für die letzte Zeile verwendet haben. Die Poincare-Ungleichung

$$\int_0^1 |\varphi(x)|^2 \, dx \leq \frac{1}{4} \int_0^1 \left(\frac{\partial \varphi}{\partial x}(x)\right)^2 \, dx$$

für Funktionen mit Randwerten $\varphi(0) = \varphi(1) = 0$ liefert dann die Stabilitätsabschätzung

$$\int_0^1 |u^h(x)|^2 \, dx \leq \frac{1}{4} \int_0^1 \left(\frac{\partial u^h}{\partial x}(x)\right)^2 \, dx \leq \frac{1}{16} \int_0^1 |f(x)|^2 \, dx.$$

Also erhalten wir eine von h unabhängige Schranke an die L^2-Norm sowohl von u^h als auch von $\frac{\partial u^h}{\partial x}$.

1.1.2 Parabolische Probleme: Die Wärmeleitungsgleichung

Als Beispiel für ein parabolisches Problem betrachten wir die eindimensionale Wärmeleitungsgleichung

$$\frac{\partial u}{\partial t}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) = f(x,t), \quad u(x,0) = u_0(x), u(0,t) = u(1,t) = 0, \quad x \in (0,1), t \in (0,T). \tag{1.9}$$

Nun haben wir ein Anfangs-Randwertproblem im Raum-Zeit Zylinder zu lösen und benötigen neben der Orts- auch noch eine Zeitdiskretisierung. Dabei stellt sich sofort die Frage nach der Reihenfolge der Diskretisierung: Soll zuerst im Ort und dann in der Zeit diskretisiert werden oder umgekehrt (man nennt diese beiden Zugänge *horizontale* bzw. *vertikale Linienmethode*). Man könnte auch direkt in der Raum-Zeit diskretisieren, z.B. durch geeignete mehrdimensionale finite Elemente. In den meisten Fällen führen alle Zugänge aber auf ähnliche Diskretisierungen, so dass wir uns hier auf die vertikale Linienmethode beschränken, d.h. wir diskretisieren zuerst im Ort.

Bei einer finiten Differenzen Diskretisierung im Ort suchen wir nun die Werte $u_j^h(t)$ an den Gitterpunkten x_j für jeden Zeitpunkt. Da wir den selben Differentialoperator diskretisieren, erhalten wir sofort (mit der obigen Notation) das semi-diskrete Problem

$$\frac{dU_h}{dt}(t) + \mathbf{K}_h U_h(t) = F^h(t), \quad U_h(0) = (u_0(x_j))_{j=1,\ldots n} \tag{1.10}$$

d.h. ein Anfangswertproblem für ein System gewöhnlicher Differentialgleichungen.

Bei einer finiten Elemente Diskretisierung starten wir von der schwachen Form der Wärmeleitung

$$\int_0^1 \frac{\partial u}{\partial t}(x,t)\varphi(x) + \frac{\partial u}{\partial x}(x)\frac{\partial \varphi}{\partial x}(x)\ dx = \int_0^1 f(x,t)\varphi(x)\ dx$$

und machen für u^h wieder einen Ansatz als Linearkombination (mit zeitabhängigen Koeffizienten) der φ_j^h, die wir auch als Testfunktionen benützen. Damit erhalten wir ein diskretes System der Form

$$\mathbf{M}_h \frac{dU_h}{dt}(t) + \mathbf{K}_h U_h(t) = F^h(t), \quad U_h(0) = \left(\int_0^1 u_0(x)\varphi_j^h(x)\right)_{j=1,\dots n} \tag{1.11}$$

wobei wir nun zusätzlich eine Massenmatrix $\mathbf{M}_h$ erhalten, definiert durch

$$\mathbf{M}_h = \frac{h}{6}\begin{pmatrix} 4 & 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 4 & 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & 4 & 1 & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 4 & 1 \\ 0 & 0 & 0 & 0 & \dots & 1 & 4 \end{pmatrix}.$$

Man überprüft wieder leicht, dass $\mathbf{M}_h$ symmetrisch und positiv definit ist. Zumindest für theoretische Zwecke können wir deshalb $\mathbf{M}_h^{-1}$ anwenden und erhalten mit $\hat{\mathbf{K}}_h = \mathbf{M}_h^{-1}\mathbf{K}_h$ sowie $\hat{F}_h = \mathbf{M}_h^{-1}F_h$ wieder ein analoges System von gewöhnlichen Differentialgleichungen wie in (1.10) (durch Anwendung von $\mathbf{M}_h^{-1/2}$ von links und rechts kann auch die Symmetrie erhalten werden) . Deshalb beschränken wir uns im folgenden auf die Zeitdiskretisierung von (1.10).

Zur Zeitdiskretisierung führen wir ein Gitter auf dem Intervall $(0,T)$ ein, zur Vereinfachung wieder ein reguläres Gitter mit Punkten $t_k = k\tau, k = 0,\dots,m$, und Zeitschrittweite $\tau = \frac{T}{m}$. Nun approximieren wir U_h wieder durch Werte an den diskreten Zeitpunkten, d.h. wir suchen $U_{h,\tau}(t_k)$. Die Zeitableitung können wir wieder mit einem Differenzenquotienten berechnen. Dafür haben wir nun mehrere Möglichkeiten, wobei die einfachste ein Vorwärtsdifferenzenquotient ist, d.h.

$$\frac{dU_h}{dt}(t_k) \approx \frac{1}{\tau}\left(U_{h,\tau}(t_{k+1}) - U_{h,\tau}(t_k)\right). \tag{1.12}$$

Durch Einsetzen erhalten wir das *Vorwärts*-Euler Verfahren, eine *explizite Zeitdiskretisierung* der Form

$$U_{h,\tau}(t_{k+1}) = U_{h,\tau}(t_k) - \tau\left(\mathbf{K}_h U_{h,\tau}(t_k) - F_h(t_k)\right), \quad U_{h,\tau}(0) = U_h(0). \tag{1.13}$$

Bei der expliziten Diskretisierung ist keine Lösung eines Gleichungssystems nötig, in jedem Schritt können wir die diskrete Lösung direkt durch Anwenden der Matrix $\mathbf{K}_h$ aus dem vorherigen Zeitschritt bestimmen. Dadurch ist in diesem Fall die Existenz und Eindeutigkeit der diskreten Lösung klar. Nicht klar ist jedoch die Stabilität der diskreten Lösung. Zur Vereinfachung untersuchen wir dabei den Fall $F_h = 0$. Seien λ_j, $j = 1,\dots,n$ die Eigenwerte von $\mathbf{K}_h$ und sei Σ eine Orthogonalmatrix bestehend aus Eigenvektoren, sodass

$$\Sigma^T \mathbf{K}_h \Sigma = \ \text{diag}\ (\lambda_j).$$

Definieren wir nun $V^k = \Sigma^T U_{h,\tau}(t_k)$, dann erhalten wir die Differenzengleichung

$$V^{k+1} = V^k - \tau \text{ diag } (\lambda_j) V^k,$$

oder komponentenweise

$$V_j^{k+1} = (1 - \tau\lambda_j) V_j^k.$$

Daraus können wir die Lösung als

$$V_k^j = (1 - \tau\lambda_j)^k V_j^0$$

berechnen. Stabilität erhalten wir nur für $|1 - \tau\lambda_j| \leq 1$, da sonst V_k^j geometrisch anwächst. Da $\mathbf{K}_h$ positiv definit ist, sind alle λ_j positiv und damit $1 - \tau\lambda_j < 1$. Weiter muss aber gelten $1 - \tau\lambda_j \geq -1$, oder als Schranke für die Zeitschrittweite $\tau \leq \frac{2}{\lambda_j}$ (für alle j). Aus der Skalierung in (1.3) erwarten wir, dass der grösste Eigenwert von $\mathbf{K}_h$ von der Ordnung h^{-2} ist (dies lässt sich auch beweisen). Also erhalten wir eine Schranke der Form $\tau = \mathcal{O}(h^2)$, d.h. die Zeitschrittweite muss sehr klein im Vergleich zur örtlichen Gittergrösse sein.

Eine Alternative zur expliziten Zeitdiskretisierung ist die Wahl eines Rückwärts-Differenzenquotienten

$$\frac{dU_h}{dt}(t_k) \approx \frac{1}{\tau} \left(U_{h,\tau}(t_k) - U_{h,\tau}(t_{k-1}) \right). \tag{1.14}$$

Mit dieser Wahl erhalten wir das *Rückwärts*-Euler Verfahren, eine *implizite Zeitdiskretisierung* der Form

$$U_{h,\tau}(t_k) + \tau \mathbf{K}_h U_{h,\tau}(t_k) = U_{h,\tau}(t_{k-1}) \tau F_h(t_k), \quad U_{h,\tau}(0) = U_h(0). \tag{1.15}$$

Im Gegensatz zu expliziten Verfahren erfordert die implizite Diskretisierung die Lösung eines linearen Gleichungssystems in jedem Zeitschritt. Die Systemmatrix $\mathbf{I} + \tau\mathbf{K}_h$ hat analoge Eigenschaften wie im elliptischen Fall. Würde man die obige Diskretisierung aus einer horizontalen Linienmethode herleiten, so sieht man, dass in jedem Zeitschritt eine Ortsdiskretisierung der elliptischen Gleichung

$$u^\tau(x, t_k) - \tau \frac{\partial^2 u^\tau}{\partial x^2}(x, t_k) = u^\tau(x, t_{k-1}) + \tau f(x, t_k)$$

gelöst wird. Damit ist natürlich der numerische Aufwand in jedem Schritt eines impliziten Verfahrens ungleich höher als in einem Schritt eines expliziten Verfahrens. Dies kann allerdings in den meisten Fällen durch eine grössere Zeitschrittweite kompensiert werden. Führen wir für $F_h = 0$ eine analoge Diagonalisierung wie im expliziten Fall durch, so erhalten wir die Rekursion

$$(1 + \tau\lambda_j) V_j^{k+1} = V_j^k$$

mit der Lösung

$$V_j^k = \frac{1}{(1 + \tau\lambda_j)^k} V_j^0.$$

Da nun $1 + \tau\lambda_j > 1$ gilt, erhalten wir sogar geometrischen Abfall der V_j^k (was die Wärmeleitung ohne Quelle natürlich besser approximiert) und damit insbesondere Stabilität unabhängig von der Zeitschrittweite.

1.1.3 Hyperbolische Probleme: Die Transportgleichung

Als einfaches Beispiel für hyperbolische Probleme (wie alle Differentialgleichungen erster Ordnung) betrachten wir die lineare eindimensionale Transportgleichung

$$\frac{\partial u}{\partial t}(x,t) + \frac{\partial u}{\partial x}(x,t) = 0, \quad u(x,0) = u_0(x), \quad u(0,t) = 0, \quad x \in (0,1), t \in (0,T). \tag{1.16}$$

Wegen der einfacheren Darstellung beschränken wir uns auf finite Differenzenverfahren zur Ortsdiskretisierung, diese sind auch die häufigst verwendeten für hyperbolische Probleme.

Wie schon zuvor bei der Zeitdiskretisierung haben wir verschiedene Möglichkeiten bei der Wahl der Differenzenquotienten für den Operator erster Ordnung $\frac{\partial u}{\partial x}(x,t)$. Bei der Wahl eines Vorwärtsdifferenzenquotienten erhalten wir das semidiskrete Verfahren

$$\frac{du_j^h}{dt}(t) = -\frac{1}{h}(u_{j+1}^h - u_j^h(t)) \tag{1.17}$$

bei einem Rückwärtsquotienten

$$\frac{du_j^h}{dt}(t) = -\frac{1}{h}(u_j^h - u_{j-1}^h(t)) \tag{1.18}$$

und bei einem zentralen Differenzenquotienten

$$\frac{du_j^h}{dt}(t) = -\frac{1}{2h}(u_{j+1}^h - u_{j-1}^h(t)). \tag{1.19}$$

In Matrixform erhalten wir in jedem Fall

$$\frac{dU_h}{dt}(t) = \mathbf{D}_h U_h(t)$$

mit den Matrizen

$$\mathbf{D}_h^+ = \frac{1}{h}\begin{pmatrix} 1 & -1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & -1 & \dots & 0 & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & -1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix},$$

$$\mathbf{D}_h^- = \frac{1}{h}\begin{pmatrix} -1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 1 & -1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & 0 & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 1 & -1 \end{pmatrix},$$

und

$$\mathbf{D}_h^c = \frac{1}{2h}\begin{pmatrix} 0 & -1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & -1 & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & -1 \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix}.$$

Man sieht sofort, dass beim Vorwärts- und beim zentralen Differenzenquotienten Probleme mit den Randbedingungen auftreten, da der Wert von u^h_{n+1}, d.h. von $u^h(1)$ benötigt würde. Beim Rückwärtsdifferenzenquotienten hingegen genügt es den Wert von $u^h_0 = u^h(0) = 0$ einzusetzen, was der gegebenen Randbedingung entspricht.

Im Fall des Rückwärtsdifferenzenquotienten können wir das semidiskrete Problem explizit lösen. Wegen $u^0_h = 0$ erhalten wir für u^1_h die Differentialgleichung

$$\frac{du^h_1}{dt}(t) = -\frac{1}{h}u^h_1, \qquad u^h_1(0) = u_0(x_1)$$

mit der Lösung $u^h_1(t) = e^{-\frac{t}{h}}u_0(x_1)$. Die weiteren Gleichungen können wir ebenfalls explizit lösen und erhalten induktiv

$$u^h_j(t) = e^{-\frac{t}{h}}\sum_{i=1}^{j} u_0(x_i)\frac{1}{(j-i)!}\left(\frac{t}{h}\right)^{j-i}.$$

Man sieht sofort die Stabilität des Verfahrens in der Supremum-Norm, es gilt

$$|u^h_j(t)| \le e^{-\frac{t}{h}}\sum_{i=1}^{j}|u_0(x_i)|\frac{1}{(j-i)!}\left(\frac{t}{h}\right)^{j-i} \le \max_i |u_0(x_i)|e^{-\frac{t}{h}}\sum_{i=0}^{j-1}\frac{1}{i!}\left(\frac{t}{h}\right)^{i} \le \max_i |u_0(x_i)| \le \|u_0\|_\infty.$$

Bei einem Vorwärtsdifferenzenquotienten erhalten wir hingegen

$$u^h_j(t) = \frac{1}{h}e^{\frac{t}{h}}\sum_{i=j}^{n} u_0(x_i)\frac{1}{i!}\left(-\frac{t}{h}\right)^{i-j} - \int_0^t (s-t)^{n-j}e^{\frac{t-s}{h}}u^h_{n+1}(s)\,ds$$

und sehen sofort dass wir durch den Faktor $e^{\frac{t}{h}}$ einen in der Zeit wachsenden Anteil erhalten, der für Instabilität des Verfahrens sorgt. Beim zentralen Differenzenquotienten erhält man in ähnlicher Weise Instabilität.

Der Grund für die Stabilität des Vorwärtsdifferenzenquotienten liegt in der Ausbreitung der Charakteristiken der hyperbolischen Gleichung (1.16). Die charakteristischen Gleichungen sind gegeben durch

$$\frac{dt}{d\tau} = 1, \qquad \frac{dx}{d\tau} = 1,$$

und somit erhält man Charakteristiken als Geraden der Form $x = t + c$. Entlang der Charakteristiken gilt in diesem Fall sogar $\frac{d}{d\tau}u(x(\tau), t(\tau)) = 0$, d.h. die Lösung ist konstant. Gehen wir vorwärts in der Zeit, dann wird die Information entlang der Charakteristiken also von links nach rechts ausgebreitet. Dieser Sichtweise entspricht der Rückwärtsdifferenzenquotient, da er zur Berechnung des Wertes am Gitterpunkt x_j nur Werte links dieses Gitterpunkts verwendet. Der Vorwärts- und zentrale Differenzenquotient verletzen hingegen die Kausalität, da sie zur Berechnung des Wertes in x_j auch auf den rechten Gitterpunkt x_{j+1} zugreifen. Man sieht in diesem Beispiel, dass die Charakteristiken bei hyperbolischen Problemen von grosser Bedeutung für die Konstruktion stabiler Verfahren sind. Wollen wir die Analysis also auf eine allgemeinere Gleichung, etwa

$$\frac{\partial u}{\partial t}(x,t) + v(x,t)\frac{\partial u}{\partial x}(x,t) = 0, \quad u(x,0) = u_0(x), \tag{1.20}$$

verallgemeinern, so berechnen wir zuerst die Charakteristiken

$$\frac{dt}{d\tau} = 1, \frac{dx}{d\tau} = v(x,t).$$

Hier können wir wieder $t = \tau$ setzen und erhalten also $\frac{dx}{dt} = v(x,t)$. Für die Ausbreitungsrichtung der Charakteristiken ist dann nur das Vorzeichen von v entscheidend. Ist v positiv verwenden wir wie oben den Rückwärtsdifferenzenquotienten, andernfalls den Vorwärtsquotienten. In Kurzform erhalten wir so das *Upwind-Verfahren*

$$\frac{du_j^h}{dt} + \max\{v(x_j,t),0\}\frac{u_j^h - u_{j-1}^h}{h} + \min\{v(x_j,t),0\}\frac{u_{j+1}^h - u_j^h}{h} = 0.$$

Im allgemeinen ist es nicht möglich, die semidiskreten Probleme wie oben zu lösen, weshalb auch eine Zeitdiskretisierung notwendig ist. Hierzu wählen wir wieder Gitterpunkte $t_k = k\tau$ auf der Zeitskala, mit kleinem Zeitschritt τ. Wir bezeichnen die Werte der diskreten Lösung am Ort x_j und zum Zeitpunkt t_k mit $u_{j,k}^h$. Nun haben wir wieder mehrere Möglichkeiten für die Wahl des Differenzenquotienten in der Zeit. Der Einfachheit halber wählen wir den Vorwärtsdifferenzenquotienten und erhalten wir das Vorwärts-Euler Verfahren

$$\frac{u_{j,k+1}^h - u_{j,k}^h}{\tau} + \frac{u_{j,k}^h - u_{j-1,k}^h}{h} = 0,$$

das die explizite Berechnung der Werte im nächsten Zeitschritt als

$$u_{j,k+1}^h = (1 - \frac{\tau}{h})u_{j,k}^h + \frac{\tau}{h}u_{j-1,k}^h$$

erlaubt. Aus dieser Formel sehen wir auch die Stabilität des Verfahrens abhängig von $\lambda = \frac{\tau}{h}$. Für $\lambda \leq 1$ ist die Lösung im nächsten Zeitschritt Konvexkombination von Werten im letzten Zeitschritt und Maximum und Minimum können deshalb im Verlauf der Zeit nicht grösser werden. Durch Rückeinsetzen der Zeitschritte erhalten wir

$$u_{j,k}^h = \sum_{i=0}^{k} \left(\begin{array}{c} n \\ i \end{array} \right) (1-\lambda)^{k-i}\lambda^i u_0(x_{j-i})$$

mit $u_0(x_\ell) = 0$ für $\ell < 0$. Um die Stabilität abzuschätzen verwenden wir für $\lambda \leq 1$

$$|u_{j,k}^h| \leq \sum_{i=0}^{k} \left(\begin{array}{c} n \\ i \end{array} \right) (1-\lambda)^{k-i}\lambda^i \max_j |u_0(x_{j-i})| = \|u_0\|_\infty.$$

Für $\lambda > 1$ können hingegen Instabilitäten auftreten, da man ein geometrisches Wachstum mit Faktor > 1 erhalten kann. Sei z.B. der Anfangswert so, dass $u_0(x_0) = 1$ und $u_0(x_j) = 0$ für $j > 0$ gilt. Dann ist $u_k^h = \lambda^k$, dieser Wert wächst also in der Zeit stark an und das Verfahren ist deshalb instabil. Man nennt die Beschränkung $\lambda \leq 1$ die CFL (Courant-Friedrichs-Levy) Bedingung. Für die allgemeinere Gleichung (1.20) wird Stabilität analog durch die CFL-Bedingung $\lambda \leq \|v\|_\infty$ erreicht.

Die CFL-Bedingung kann auch bezüglich der Charakteristiken interpretiert werden, da ja auch das diskrete Verfahren eine analoge Eigenschaft hat. Im diskreten Fall wird ja die Information entlang der Geraden mit Steigung λ fortgepflanzt, im stetigen Fall entlang der Charakteristiken mit Steigung 1. Die CFL-Bedingung impliziert also, dass die "diskreten Charakteristiken" nicht steiler sind als die kontinuierlichen, d.h. die Information wird im numerischen Verfahren nicht schneller fortgepflanzt als in der Differentialgleichung.

Bezüglich des zentralen Differenzenquotienten kann das Upwind-Verfahren auch als Verfahren mit künstlicher Diffusion dargestellt werden

$$\frac{u_{j,k+1}^h - u_{j,k}^h}{\tau} + \frac{u_{j+1,k}^h - u_{j-1,k}^h}{2h} = \frac{h}{2}\frac{u_{j+1,k}^h - 2u_{j,k}^h + u_{j-1,k}^h}{h^2}.$$

Der Differenzenquotient auf der rechten Seite ist eine Diskretisierung der zweiten Ableitung und damit approximieren wir die Differentialgleichung

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = \frac{h}{2}\frac{\partial^2 u}{\partial x^2},$$

d.h. wir erhalten künstliche Diffusion mit Koeffizienten $\frac{h}{2}$. Da der Diffusionskoeffizient von h abhängt, sollte dieser zusätzliche Effekt mit kleiner werdender Gittergrösse verschwinden.

Abschliessend können wir noch den Fehler bei der numerischen Approximation untersuchen und nehmen dazu an, dass die Lösung der Transportgleichung $u \in C^2$ erfüllt. Definieren wir den punktweisen Fehler als $e^h_{j,k} = u^h_{j,k} - u(x_j, t_k)$, so gilt

$$e^h_{j,k+1} = (1-\lambda)e^h_{j,k} + \lambda e^h_{j-1,k} + (1-\lambda)(u(x_j,t_k) - u(x_j,t_{k+1})) + \lambda(u(x_{j-1},t_k) - u(x_j,t_{k+1})).$$

Nun erhalten wir durch Taylor-Entwicklung

$$\begin{aligned} u(x_j,t_k) - u(x_j,t_{k+1}) &= -\frac{\partial u}{\partial t}(x_j,t_k)\tau + \mathcal{O}(\tau^2) \\ u(x_{j-1},t_k) - u(x_j,t_{k+1}) &= -\frac{\partial u}{\partial x}(x_j,t_k)h - \frac{\partial u}{\partial t}(x_j,t_k)\tau + \mathcal{O}(\tau^2+h^2) \end{aligned}$$

und damit

$$\begin{aligned} & (1-\lambda)(u(x_j,t_k) - u(x_j,t_{k+1})) + \lambda(u(x_{j-1},t_k) - u(x_j,t_{k+1})) \\ = \; & -(1-\lambda)\tau\frac{\partial u}{\partial t}(x_j,t_k) - \lambda\tau\frac{\partial u}{\partial t}(x_j,t_k) - \lambda h\frac{\partial u}{\partial x}(x_j,t_k) + \mathcal{O}(\tau^2+\lambda h^2) \\ = \; & -\tau\underbrace{\frac{\partial u}{\partial t}(x_j,t_k) + \frac{\partial u}{\partial x}(x_j,t_k)}_{=0} + \mathcal{O}(\tau^2+\lambda h^2). \end{aligned}$$

Für den maximalen Fehler in jedem Zeitschritt $e^h_k = \max_j |e^h_{j,k}|$ erhalten wir dann die Abschätzung

$$e^h_{k+1} \leq e^h_k + C(\lambda h^2 + \tau^2)$$

und nach Rückeinsetzen

$$e^h_k \leq e^h_0 + Ck\tau(h+\tau).$$

Da in jedem Fall $k = \mathcal{O}(\tau^{-1})$ gilt, folgt eine Fehlerabschätzung erster Ordnung

$$e^h_k \leq e^h_0 + C(h+\tau).$$