Bachelorarbeit
im Studiengang "Mathematik"

IMEX Verfahren für Reaktions-Konvektions-Diffusions-Gleichungen

Sophia Scholtka

Institut für Numerische und Angewandte Mathematik
Universität Göttingen

Betreut durch Prof. Dr. Gert Lube

14.8.2010

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Göttingen, den 14.8.2010

Inhaltsverzeichnis

1 Einleitung

1.1 Ziel der Arbeit

Die vorliegende Arbeit befasst sich mit der numerischen Behandlung des Anfangswertproblems

$$\partial_t u - \varepsilon \partial_{xx} u + \partial_x g(u) + h(u) = 0$$

mit der Anfangsbedingung

$$u(0, x) = u_0(x)$$

und der periodischen Randbedingung

$$u(t, 0) = u(t, 1),$$

wobei besonderes Augenmerk auf die explizit-implizite Behandlung der Zeitdiskretisierung gelegt wird.
Außerdem werden Anpassungen für den Fall $\varepsilon << 1$ behandelt.

1.2 Vorgehensweise

Wir beginnen mit Überlegungen zu den tatsächlichen Lösungen unserer Anfangswertprobleme.
Anschließend werden die Grundlagen der Ortsdiskretisierung besprochen, wobei wir bereits auf Upwind-Verfahren eingehen.
Es folgt die Theorie der Runge-Kutta-Verfahren, auf die wir die IMEX-Verfahren aufbauen.
Zuletzt werden einige Beispielrechnungen durchgeführt und die Ergebnisse interpretiert.

2 Grundlagen

2.1 Klassische Lösung

Definition 2.1. Eine Funktion $u \in C((0,T)) \cap C^2([0,1])$ mit Argumenten $(t,x) \in G = (0,T) \times [0,1]$ heißt *klassische Lösung* des Anfangswertproblems

$$\partial_t u - \varepsilon \partial_{xx} u + \partial_x g(u) + h(u) = 0$$

mit der Anfangsbedingung

$$u(0,x) = u_0$$

und der periodischen Randbedingung

$$u(t,0) = u(t,1),$$

falls die Differentialgleichung, Anfangsbedingung und Randbedingung punktweise erfüllt sind.

Wir beschäftigen uns zunächst mit der klassischen Lösung von einigen Spezialfällen der Differentialgleichung. Dabei heißt $\varepsilon \partial_{xx} u$ der Diffusionsterm, $\partial_x g(u)$ der Konvektionsterm und $h(u)$ der Reaktionsterm.
Das einfachste Reaktionsproblem ist

$$\partial_t u = \lambda u,\ u(0,x) = u_0$$

mit der Lösung

$$u(x,t) = u_0(x) \mathrm{e}^{\lambda t}.$$

Diese Differentialgleichung spielt bei der A- und L-Stabilität eine wichtige Rolle.
Das einfachste Diffusionsproblem ist die Wärmeleitungsgleichung

$$\partial_t u = \varepsilon \partial_{xx} u,\ u(0,x) = u_0,$$

deren Lösung

$$u(t,x) = \frac{1}{\sqrt{4\pi\varepsilon t}} \int_G u_0(y) \mathrm{e}^{\frac{-(x-y)^2}{4\varepsilon t}} dy$$

auf einem Gebiet G mit $x \in G$ bekannt ist [Lub07].
Das einfachste Konvektionsproblem ist das Transportproblem

$$\partial_t u + b\partial_x u = 0, \quad u(0,x) = u_0(x).$$

Es hat die Lösung

$$u(t,x) = u_0(x - bt).$$

Sei $u(t,x) = u_0(x-bt)$, dann gilt $\partial_t u(t,x) = -bu_0(x-bt)$ und $\partial_x u(t,x) = u_0(x-bt)$, also gilt $\partial_t u + b\partial_x u = 0$. □
Beim nichtlinearen Fall beschränken wir uns auf die Burgers-Gleichung

$$\partial_t u + bu\partial_x u = \varepsilon \partial_{xx} u, \quad u(0,x) = u_0(x).$$

Wir stellen um zu

$$\partial_t u + b\partial_x \left(\frac{u^2}{2} \right) = \varepsilon \partial_{xx} u.$$

Mit $u = \partial_x \varphi$ erhält man

$$\partial_{xt}\varphi + b\partial_x \left(\frac{(\partial_x \varphi)^2}{2} \right) = \varepsilon \partial_{xxx}\varphi.$$

Integration über x ergibt

$$\partial_t \varphi + b\frac{(\partial_x \varphi)^2}{2} = \varepsilon \partial_{xx}\varphi.$$

Nun substituieren wir $\varphi = \frac{-2\varepsilon}{b} \ln v$ und erhalten

$$\frac{-2\varepsilon}{bv}\partial_t v + \frac{b}{2}\frac{4\varepsilon^2}{b^2 v^2}(\partial_x v)^2 = \frac{-2\varepsilon^2}{bv}\partial_{xx} v + \frac{2\varepsilon^2}{bv^2}(\partial_x v)^2,$$

was wir zu

$$\partial_t v = \varepsilon \partial_{xx} v$$

kürzen können.

Dies ist die Wärmeleitungsgleichung, deren Lösung

$$v(t,x) = \frac{1}{\sqrt{4\pi\varepsilon t}} \int_G v_0(y) \mathrm{e}^{\frac{-(x-y)^2)}{4\varepsilon t}} dy$$

wir bereits kennen.
Wir substituieren zurück über $u = \partial_x \varphi = -2b\varepsilon \frac{\partial_x v}{v}$

$$v_0(y) = \mathrm{e}^{\frac{-1}{2b\varepsilon} \int_0^y u_0(s) ds}$$

und erhalten

$$u(t,x) = b \frac{\int_G \frac{x-y}{t} \mathrm{e}^{\frac{-1}{2b\varepsilon} \int_0^y u_0(s) ds} \mathrm{e}^{\frac{-(x-y)^2}{4\varepsilon t}} dy}{\int_G \mathrm{e}^{\frac{-1}{2b\varepsilon} \int_0^y u_0(s) ds} \mathrm{e}^{\frac{-(x-y)^2}{4\varepsilon t}} dy}.$$

Im Grenzfall $\varepsilon \to 0$ betrachten wir die Charakteristiken der Differentialgleichung

$$\partial_t u + bu\partial_x u = 0$$

beziehungsweise den etwas allgemeineren Fall

$$\partial_t u + \partial_x g(u) = r(x).$$

Die Charakteristiken sind die Lösungen der gewöhnlichen Differentialgleichung

$$\chi'(t) = g'(u(t, \chi(t))).$$

Man rechnet leicht nach, dass die Ableitung von u entlang der Charakteristiken gerade r ist:

$$\frac{\mathrm{d}}{dt} u(t,\chi(t)) = \partial_t u(t,\chi(t)) + \chi'(t)\partial_x u(t,\chi(t)) = \partial_t u + g'(u)\partial_x u = \partial_t u + \partial_x g(u) = r.$$

Wir setzen $g(u) = b\frac{u^2}{2}, b > 0$ und $r(x) = 0$ und sehen, dass nun u entlang der Charakteristiken konstant ist. Den Wert $u(t,x) = u_0(s)$ erhält man dann durch Auflösen der Gleichung $x = s + bu_0(s)t$ nach s.
Ist u_0 differenzierbar mit $u_0'(s) \geq 0$, so ist die Lösung s der Gleichung

$$x = s + bu_0(s)t$$

für jede Stelle (t, x) eindeutig bestimmt, da

$$s_1 < s_2 \Rightarrow u_0(s_1) \leq u_0(s_2) \Rightarrow s_1 + bu_0(s_1)t < s_2 + bu_0(s_2)t.$$

Ist aber $u_0'(s) < 0$, so kann der Fall

$$s_1 < s_2,\ u_0(s_1) > u_0(s_2),\ s_1 + bu_0(s_1)t = s_2 + bu_0(s_2)t$$

auftreten, das heißt für $u(x, t)$ sind zwei verschiedene Werte $u_0(s_1)$ und $u_0(s_2)$ möglich. Wir bezeichnen mit $\hat{t}$ den ersten Zeitpunkt, zu dem dieses Phänomen auftritt. Die Methode der Charakteristiken liefert dann nur für $t \in [0, \hat{t})$ eine eindeutige Lösung.

Beispiel 2.2. Für die Burgers-Gleichung mit periodischen Randbedingungen und einer Sinusschwingung als Startwert

$$\partial_t u + u\partial_x u = 0,\ u_0(x) = \sin(2\pi x),\ u(t, 0) = u(t, 1)$$

kreuzen sich die Charakteristiken für alle $t > 0$, wie in Abbildung 2.1 angedeutet. Die Methode der Charakteristiken liefert daraufhin das Ergebnis in Abbildung 2.2, bei dem man sehr gut sieht, dass die Zuordnung eines Punktes (t, x) zu einer Charakteristik nicht mehr eindeutig ist. Die Lösung ist für jedes feste $t > 0$ an der Stelle $x = 0.5$ nicht mehr stetig, da es immer zwei Charakteristiken gibt, die hier zusammenlaufen. Man nennt diese Unstetigkeitsstelle einen Schock.

In einem solchen Fall kann es keine klassische Lösung mehr geben. Daher führen wir jetzt den Begriff der schwachen Lösung ein, indem wir die Differentialgleichung mit einer Testfunktion ϕ multiplizieren

$$\phi\partial_t u + \phi u\partial_x u = 0$$

und partiell integrieren.

Definition 2.3. u ist eine *schwache Lösung* der Differentialgleichung

$$\partial_t u + \partial_x g(u) = 0, u_0(x) = \sin(2\pi x), u(t, 0) = u(t, 1),$$

falls für alle Testfunktionen $\phi \in C_0^1((0, \infty) \times [0, 1])$, also für alle stetig differenzierbaren

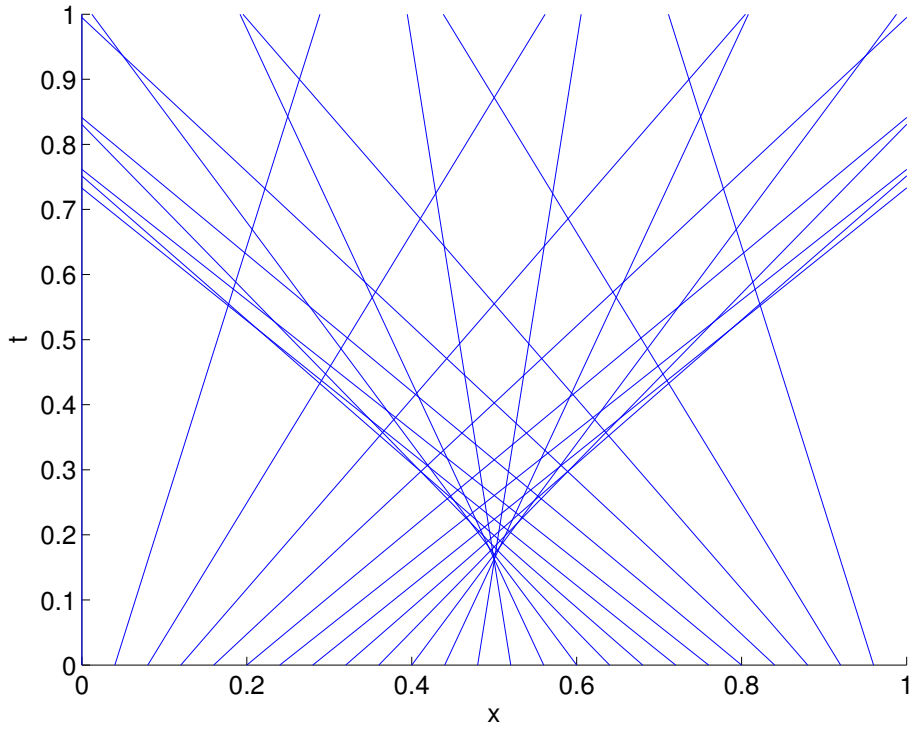

Abbildung 2.1: Charakteristiken für $\partial_t u + u\partial_x u = 0$, $u_0(x) = \sin(2\pi x)$ in der x, t-Ebene

Funktionen auf $(0, \infty) \times [0, 1]$ mit kompaktem Träger gilt:

$$\int_x \int_t (u\phi_t + g(u)\partial_x \phi) + \int_x u_0 \phi(0, x) = 0.$$

Bei der Burgers-Gleichung ist $g(u) = \frac{1}{2}u^2$ und

$$u(t, x) = \begin{cases} 2x & \forall\ 0 \leq x < 0.5 \\ 2x - 1 & \forall\ 0.5 < x \leq 1 \end{cases} \tag{2.4}$$

ist eine schwache Lösung der Burgers-Gleichung mit periodischen Randbedingungen und $u_0(x) = \sin(2\pi x)$.
Wir betrachten nun die Unstetigkeitsstelle genauer und greifen dabei auf die Rankine-Hugoniot-Bedingung und die Entropiebedingungen zurück, die bei Quarteroni und Valli [QV08] im 14. Kapitel ausführlich besprochen werden.
Eine Lösung erfüllt die Rankine-Hugoniot-Bedingung, falls für jeden Schock in der

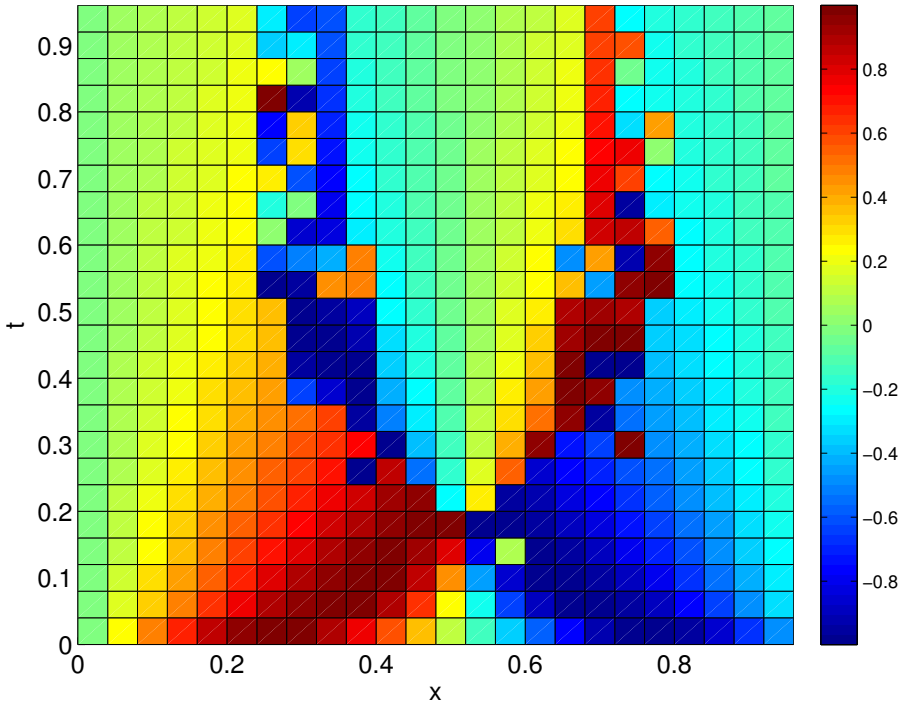

Abbildung 2.2: Mit der Methode der Charakteristiken berechnete Lösung für $\partial_t u + u\partial_x u = 0$, $u_0(x) = \sin(2\pi x)$ in der x, t-Ebene

Lösung gilt, dass der Schock entlang einer Kurve $(t, x(t))$ mit

$$\frac{dx}{dt} = s = \frac{g(u_L) - g(u_R)}{u_L - u_R}$$

liegt, wobei u_L und u_R die links- und rechsseitigen Grenzwerte der Lösung u an der Schockstelle sind.
In unserem Fall ist $g(u) = \frac{1}{2}u^2$, $u_L = 1, u_R = -1$ und $s = 0$. Das bedeutet, dass die Rankine-Hugoniot-Bedingung erfüllt und der Schock stationär ist. Die Charakteristiken enden jetzt jeweils im Schock, wie Abbildung 2.3 zeigt.
Wegen der Differenzierbarkeit von u_0 wird die x, t-Ebene für diese Burgers-Gleichung an jedem Punkt durch eine Charakteristik abgedeckt. Das Anfangswertproblem ist also nun mittels der Charakteristiken eindeutig lösbar.
Ist u_0 allerdings nicht differenzierbar, so kann es in der x, t-Ebene Bereiche geben, durch die zunächst keine Charakteristik verläuft.

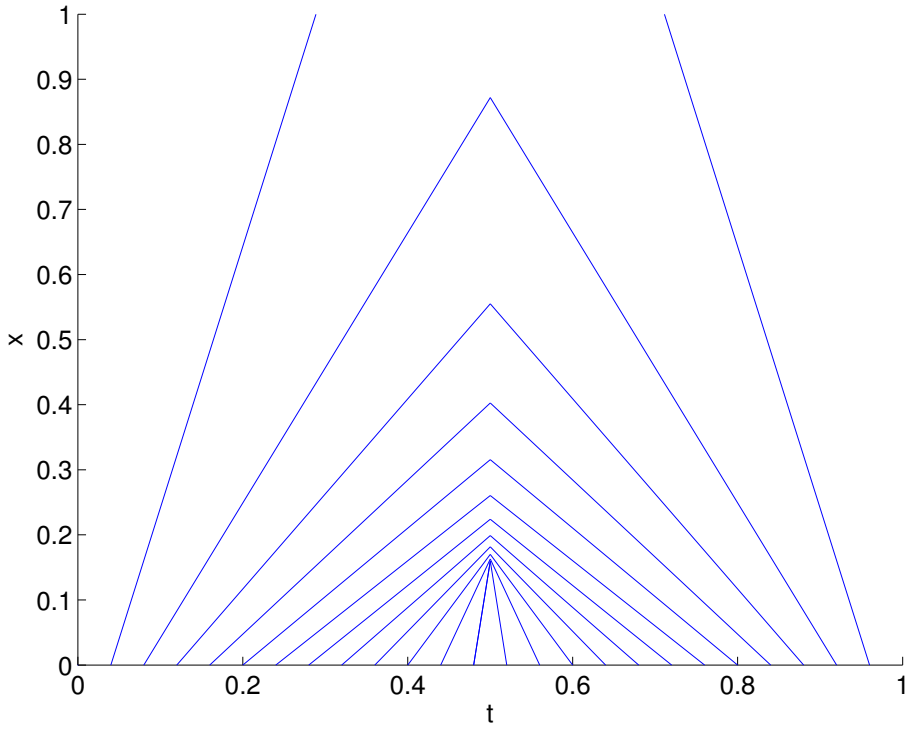

Abbildung 2.3: Charakteristiken für $\partial_t u + u\partial_x u = 0$, $u_0(x) = \sin(2\pi x)$ in der x, t-Ebene

Beispiel 2.5. Wir betrachten das Anfangswertproblem

$$\partial_t u + u\partial_x u = 0,\ u_0(x) = \begin{cases} 0 & x \leq 0 \\ 1 & x > 0 \end{cases}$$

für $(t, x) \in [0, \infty) \times \mathbb{R}$.
Die Charakteristiken sind in Abbildung 2.4 schwarz eingezeichnet und lassen einen Bereich frei. In diesem Bereich gibt es nun unendlich viele schwache Lösungen, zum Beispiel die beiden in Abbildung 2.5 gezeigten Lösungen, für die in Abbildung 2.4 passende Charakteristiken angegeben sind. Die rote Lösung erfüllt entlang der Schocklinien $x = \frac{1}{2}t$ und $x = t$ die Rankine-Hugoniot-Bedingung. Die grüne, stetige Lösung wird Verdünnungswelle genannt und erfüllt die Rankine-Hugoniot-Bedingung automatisch, da sie keine Unstetigkeitsstellen hat.
Um zwischen den beiden Lösungen zu unterscheiden, gibt es die Entropiebedingung. Ist der Fluss $g(u)$ konvex, wie bei der Burgers-Gleichung mit $g(u) = \frac{1}{2}u^2$, so gilt: Ein Schock, der sich mit der Geschwindigkeit $s = \frac{dx}{dt}$ fortbewegt, erfüllt die Entropiebe-

dingung, falls

$$g'(u_L) > s > g'(u_R).$$

Aus g konvex und $g'(u_L) > s > g'(u_R)$ folgt $u_L > u_R$. Das bedeutet, dass es nur Schocks geben darf, bei denen die Charakteristiken wie im Beispiel 2.2 ineinander laufen. Im umgekehrten Fall wird dagegen die stetige Lösung erzwungen. Die grüne Lösung und auch die angegebene schwache Lösung des Beispiels 2.2 erfüllen die Entropiebedingung, während die rote Lösung die Entropiebedingung nicht erfüllt.
Dass die Entropielösung für $\int_{\mathbb{R}} |u_0(x)|dx < \infty$ eindeutig bestimmt ist, wurde von Oleinik [Ole57] bewiesen.

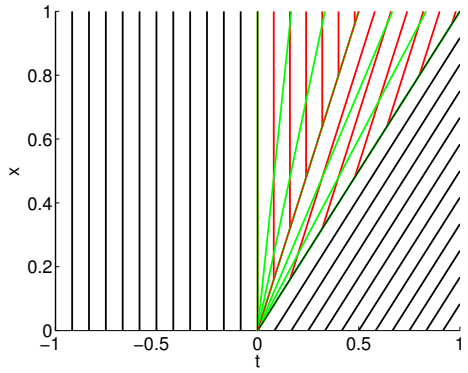

Abbildung 2.4: Charakteristiken

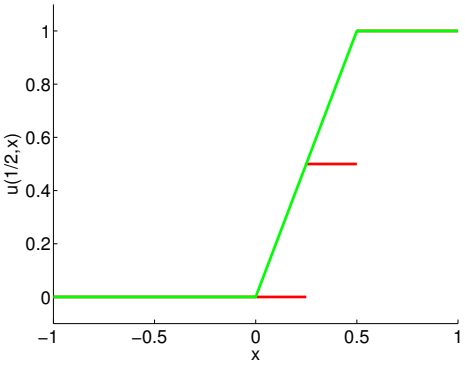

Abbildung 2.5: Lösungen

2.2 Ortsdiskretisierung mit Finiten Differenzen

Eine gängige Vorgehensweise zur numerischen Lösung partieller Differentialgleichungen ist es, zunächst die Ortsvariable x und die Ortsableitungen zu diskretisieren. Wir beschäftigen uns zunächst mit dem linearen Fall

$$\partial_t u - \varepsilon \partial_{xx} u + b(x)\partial_x u + c(x)u - r(x) = 0$$

und dem nichtlinearen Spezialfall

$$\partial_t u - \varepsilon \partial_{xx} u + b(x)u\partial_x u + c(x)u - r(x) = 0.$$

Für die Ortsdiskretisierung auf dem Intervall $[0, 1]$ mit Maschenweite $h = \frac{1}{n}, n \in \mathbb{N}$ wird ein äquidistantes Gitter

$$\{x_j = jh, j = 1..n\}$$

gewählt. Man betrachtet nun u an den Stellen

$$u_j(t) = u(x_j, t).$$

Die Ableitungen $\partial_{xx} u$ und $\partial_x u$ von u werden nun durch Differenzenquotienten dargestellt. Dabei kann das Bilden des Differenzenquotienten auch als Anwendung einer Matrix D auf den Vektor $u(t) = (u_1, u_2, \ldots, u_n)^t(t)$ betrachtet werden. Es gibt dabei für die erste Ableitung drei Möglichkeiten der Approximation erster und zweiter Ordnung:

a) den zentralen Differenzenquotienten

$$\partial_x u_j(t) \approx \frac{u_{j+1}(t) - u_{j-1}(t)}{2h}$$

mit der Matrix

$$D_h^0 = \frac{1}{2h} \begin{pmatrix} 0 & 1 & & & -1 \\ -1 & 0 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ 1 & & & -1 & 0 \end{pmatrix} \in \mathbb{R}^{n \times n}$$

b) den Vorwärts-Differenzenquotienten

$$\partial_x u_j(t) \approx \frac{u_{j+1}(t) - u_j(t)}{h}$$

mit der Matrix

$$D_h^+ = \frac{1}{h} \begin{pmatrix} -1 & 1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ 1 & & & & -1 \end{pmatrix} \in \mathbb{R}^{n \times n}$$

c) den Rückwärts-Differenzenquotienten

$$\partial_x u_j(t) \approx \frac{u_j(t) - u_{j-1}(t)}{h}$$

mit der Matrix

$$D_h^- = \frac{1}{h} \begin{pmatrix} 1 & & & & -1 \\ -1 & \ddots & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

Die Einträge an den Stellen $(1, n)$ beziehungsweise $(n, 1)$ werden durch die periodische Randbedingung nötig. Möchte man statt dessen mit der homogenen Dirichlet-Randbedingung $u(0, t) = u(1, t) = 0$ arbeiten, so müssen diese Einträge weggelassen

werden.
Für die zweite Ableitung wird immer der Differenzenquotient

$$\partial_{xx} u_j(t) \approx \frac{u_{j+1}(t) - 2u_j(t) + u_{j-1}(t)}{h^2}$$

mit der Matrix

$$D_h^2 = \frac{1}{h^2} \begin{pmatrix} -2 & 1 & & & 1 \\ 1 & -2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ 1 & & & 1 & -2 \end{pmatrix} \in \mathbb{R}^{n \times n}$$

verwendet, wobei für die Einträge $(1, n)$ und $(n, 1)$ das bereits Gesagte gilt.
Nun kann das semidiskrete Problem formuliert werden:

$$\partial_t u - \varepsilon \partial_{xx} u + b(x)\partial_x u + c(x)u - r(x) = 0 \text{ für } u : [0, T) \times \mathbb{R} \to \mathbb{R}$$

wird zu

$$\partial_t u - \varepsilon D_h^2 u + b(x) D_h u + c(x)u - r(x) = 0 \text{ für } u : [0, T) \to \mathbb{R}^n,$$

wobei wir noch $b(x)$,$c(x)$ und $r(x)$ diskretisieren müssen.
Mit

$$B_h = \begin{pmatrix} b(x_1) & & \\ & \ddots & \\ & & b(x_n) \end{pmatrix}, \quad C_h = \begin{pmatrix} c(x_1) & & \\ & \ddots & \\ & & c(x_n) \end{pmatrix}, \quad r_h = \begin{pmatrix} r(x_1) \\ \vdots \\ r(x_n) \end{pmatrix}$$

und

$$A_h = -\varepsilon D_h^2 + B_h D_h + C_h$$

ergibt sich folgendes System von gewöhnlichen Differentialgleichungen

$$\partial_t u + A_h u = r_h,$$

welches das semidiskrete Problem genannt wird. Für den nichtlinearen Spezialfall

$$\partial_t u - \varepsilon \partial_{xx} u + b(x) u \partial_x u + c(x)u - r(x) = 0$$

benötigen wir noch

$$U_h(t) = \begin{pmatrix} u(t,x_1) & & \\ & \ddots & \\ & & u(t,x_n) \end{pmatrix}$$

und setzen dann

$$A_h u(t) = [-\varepsilon D_h^2 + B_h U_h(t) D_h + C_h] u(t).$$

A_h ist nun allerdings keine lineare Abbildung mehr.
Das semidiskrete Problem kann nun zum Beispiel dadurch gelöst werden, dass $\partial_t u$ durch einen Vorwärts-Differenzenquotienten diskretisiert wird.

$$\begin{aligned} u_t &= r - A_h u \\ \frac{u(t+\tau) - u(t)}{\tau}(t) &\approx r_h - A_h u(t) \\ \Rightarrow u(t+\tau) &\approx L_h u(t) := u(t) - \tau(r_h - A_h u(t)) \end{aligned}$$

Die Approximation $u(t+\tau) = L_h u(t)$ nennt man Differenzenverfahren.
Wir müssen uns nun fragen, wie gut die Approximation durch das Differenzenverfahren ist. Zunächst betrachten wir den Fehler, der in jedem einzelnen Zeitschritt gemacht wird.

Definition 2.6. Wir nennen das Differenzenverfahren $u(t+\tau) = L_h u(t)$ *konsistent* mit Konsistenzordnung (p, q), falls für die hinreichend glatte exakte Lösung η gilt:

$$|||[\eta(t+\tau) - L_h \eta(t)]|||_\infty \leq C\tau(\tau^p + h^q) \quad \forall t \in G.$$

Beispiel 2.7. Wir betrachten noch einmal das Transportproblem

$$\partial_t u + b\partial_x u = 0, \quad u(0,x) = u_0(x).$$

Das semidiskrete Problem hat dann die Form

$$\partial_t u + b D_h u = 0.$$

Wir lösen es durch Diskretisierung von $\partial_t u$:

$$\frac{u(t+\tau)-u(t)}{\tau} + bD_h u(t) = 0$$
$$\Rightarrow u(t+\tau) = u(t) - \tau b D_h u(t) = L_h u(t).$$

Das Verfahren

$$u(t+\tau) = u(t) - \tau b D_h u(t)$$

ist für alle drei Diskretisierung D_h^+, D_h^-, D_h^0 konsistent.

Beweis. a) Sei $\eta \in C^3([0,T]\times[0,1])$.
Taylorentwichlung ergibt

$$\eta(t,x+h) = \eta(t,x) + h\eta_x(t,x) + \frac{1}{2}h^2\eta_{xx}(t,x) + \frac{1}{6}h^3\eta_{xxx}(t,\xi_+)$$
$$\eta(t,x-h) = \eta(t,x) - h\eta_x(t,x) + \frac{1}{2}h^2\eta_{xx}(t,x) - \frac{1}{6}h^3\eta_{xxx}(t,\xi_-)$$
$$D_h^0\eta(t,x) = \frac{\eta(t,x+h)-\eta(t,x-h)}{2h} = \eta_x(t,x) + \frac{1}{12}h^2(\eta_{xxx}(t,\xi_+) + \eta_{xxx}(t,\xi_-))$$
$$\eta(t+\tau,x) = \eta(t,x) + \tau\eta_t(t,x) + \frac{1}{2}\tau^2\eta_{tt}(\theta,x)$$

Die einzelnen Taylorreihen werden jetzt zusammengesetzt und $\eta_t(t,x)$ kann durch $-b\eta_x(t,x)$ ersetzt werden, da η die exakte Lösung von $\eta_t = -b\eta_x$ ist.

$$\begin{aligned}
&||\eta(t+\tau,x) - (I-\tau bD_h^0)\eta(t,x)||_\infty \\
&\quad = ||\frac{1}{2}\tau^2\eta_{tt}(\theta,x) + b\tau\frac{1}{12}h^2(\eta_{xxx}(t,\xi_+)+\eta_{xxx}(t,\xi_-))||_\infty \\
&\quad \le C\tau(\tau+h^2)
\end{aligned}$$

Für b) und c) verwenden wir die oben angegebenen Taylorreihen, brechen die Ortsentwicklung aber schon mit der zweiten Ableitung ab.

$$\begin{aligned}
\text{b)} \qquad &||\eta(t+\tau,x) - (I-\tau bD_h^+)\eta(t,x)||_\infty \\
&= ||\frac{1}{2}\tau^2\eta_{tt}(\theta,x) + b\tau\frac{1}{2}h\eta_{xx}(t,\xi)||_\infty \\
&\le C\tau(\tau+h)
\end{aligned}$$

$$
\begin{aligned}
\text{c)} \qquad & ||\eta(t+\tau,x) - (I - \tau b D_h^-)\eta(t,x)||_\infty \\
& = ||\frac{1}{2}\tau^2 \eta_{tt}(\theta,x) - b\tau\frac{1}{2}h\eta_{xx}(t,\xi)||_\infty \\
& \leq C\tau(\tau+h)
\end{aligned}
$$

Diese Differenzenverfahren haben Konsistenzordnung $(1,2)$ beziehungsweise $(1,1)$. □

Um sicher zu stellen, dass Fehler in den Anfangsdaten u_0 nicht unendlich anwachsen können, betrachten wir die Stabilität des Differenzenverfahrens.

Definition 2.8. Wir nennen das Differenzenverfahren $u(t+\tau) = L_h u(t)$ *stabil*, falls für jede Zeit $T > 0$ eine von h und τ unabhängige Konstante $C_T > 0$ existiert mit

$$||L_h^k||_\infty \leq C_T \quad \forall k,\ \tau k < T.$$

Beispiel 2.9. Das Verfahren

$$u(t+\tau) = L_h u(t) := u(t) - \tau b D_h u(t)$$

ist mit D_h^0 nicht und mit D_h^+ oder D_h^- nur unter bestimmten Bedingungen stabil.

Beweis. Die Matrix $L_h^0 = I - \tau b D_h^0$ hat nach Hanke-Bourgeois [HB06] S.413 die Eigenwerte $\lambda_j = 1 - \frac{b\tau}{2h}\mathrm{e}^{-\mathrm{i}2\pi j/n} + \frac{b\tau}{2h}\mathrm{e}^{-\mathrm{i}2\pi j(n-1)/n} = 1 - \frac{b\tau}{h}\mathrm{i}\sin(4\pi j/n)$. Für $n > 4$ gibt es also mindestens einen Eigenwert λ_j mit $s_j = \sin(4\pi j/n) \neq 0$ also $|\lambda_j| = \sqrt{1 + \frac{b^2\tau^2}{h^2}s_j^2} > 1$. Daraus folgt

$$||\left(L_h^0\right)^k||_\infty \geq \left(\sqrt{1 + \frac{b^2\tau^2}{h^2}s_j^2}\right)^k,$$

weshalb es keine von h unabhängige Konstante geben kann, die $||\left(L_h^0\right)^k||_\infty$ beschränkt. Die Matrizen $L_h^+ = I - \tau b D_h^+$ und $L_h^- = I - \tau b D_h^-$ haben einen Eigenwert $1 + 2\frac{b\tau}{h}$ beziehungsweise $1 - 2\frac{b\tau}{h}$ passend zum Eigenvektor $(1,\cdots,1)^t$. Demnach ist L_h^+ für positive und L_h^- für negative b nicht stabil. Für das jeweils andere Vorzeichen von b lässt sich Stabilität folgendermaßen erreichen:
Für $-1 \leq \frac{b\tau}{h} \leq 0$ gilt

$$||L_h^+||_\infty = \max(1 + \frac{b\tau}{h}, \frac{-b\tau}{h}) \leq 1 \Rightarrow ||\left(L_h^+\right)^k||_\infty \leq 1.$$

Für $0 \leq \frac{b\tau}{h} \leq 1$ gilt

$$||L_h^-||_\infty = \max(1 - \frac{b\tau}{h}, \frac{b\tau}{h}) \leq 1 \Rightarrow || \left(L_h^-\right)^k ||_\infty \leq 1.$$

□

Beispiel 2.10. Wir fügen nun zum Transportproblem einen Diffusionsterm hinzu und wollen

$$\partial_t u - \varepsilon \partial_{xx} u + b\partial_x u = 0$$

diskretisieren. Dazu verwenden wir das Zweischichtenschema

$$u(t+\tau) = u(t) - (1-\theta)(bD_h^0 u(t) - \varepsilon\sigma D_h^2 u(t)) - \theta(bD_h^0 u(t+\tau) - \varepsilon\sigma D_h^2 u(t+\tau)).$$

Im Fall $\theta = 0, \sigma = 1$ erhält man das bereits vorgeschlagene Schema

$$u(t+\tau) = u(t) - bD_h^0 u(t) + \varepsilon D_h^2 u(t).$$

Für $\theta = 1$ erhält man das implizite Schema

$$u(t+\tau) = u(t) - bD_h^0 u(t+\tau) + \varepsilon D_h^2 u(t+\tau)$$

und für $0 < \theta < 1$ ein gemisches Schema, das man Crank-Nicolson-Verfahren nennt. Mit $\sigma > 1$ kann man zudem zusätzliche Diffusion einfügen. Dabei führt die Wahl $\sigma = \frac{h|b|}{2\varepsilon} \coth \frac{h|b|}{2\varepsilon}$ zum sogenannten Iljin-Schema. Analog zu Beispiel 2.9 kann man nun eine Stabilitätsbedingung herleiten. Das Zweischichtenschema mit $b > 0$ und $\theta \in [0, 1]$ ist stabil, falls

$$\sigma > \frac{hb}{2\varepsilon} \quad \text{und} \quad \frac{2\varepsilon\tau(1-\theta)\sigma}{h^2} \leq 1.$$

Insbesondere ist also das rein implizite Verfahren mit $\theta = 1$ immer stabil.

In Beispiel 2.9 haben wir gesehen, dass es von Vorteil ist, den Differenzenquotienten D_h entgegen der Flussrichtung $b = g'(u)$ zu wählen. Ein Differenzenschema, das dies berücksichtigt, nennt man Upwind-Schema.
Die Stabilitätsbedingung $0 \leq \frac{b\tau}{h} \leq 1$ für das Upwind-Schema L_h^- wurde zuerst von Courant, Friedrichs und Lewy [CFL28] beschrieben. Sie heißt daher CFL-Bedingung. Wir können die CFL-Bedingung anhand von Abbildung 2.6 veranschaulichen.

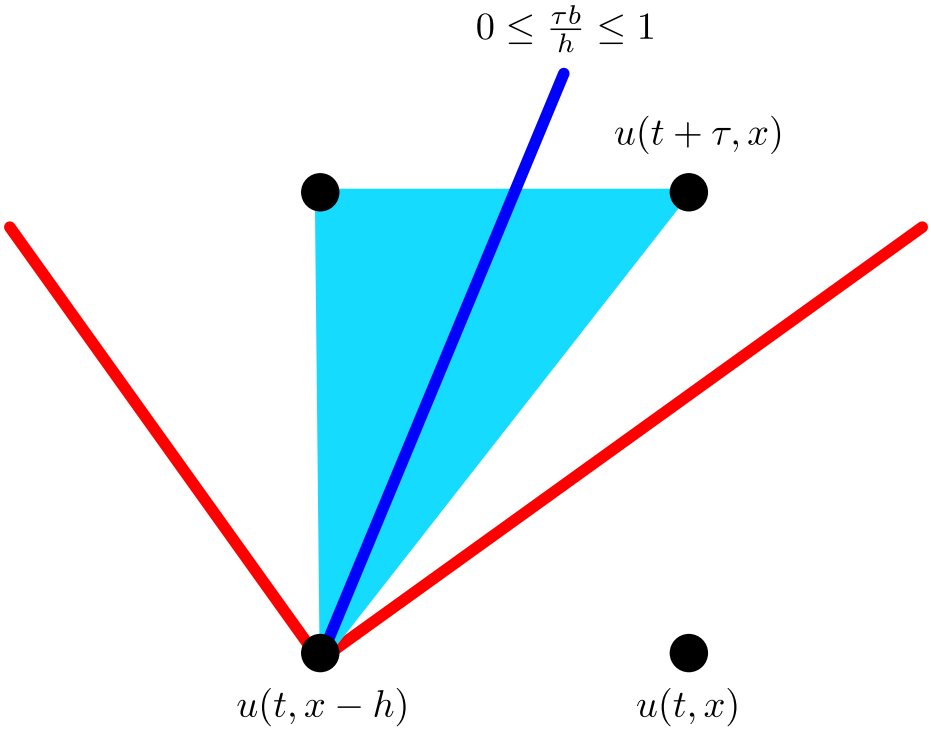

Abbildung 2.6: Diskretisierung mit D_h^- und mögliche Charakteristiken für $\partial_t u + b\partial_x u = 0$

Die aktuelle Information $u(t, x-h)$ beeinflusst bei Verwendung des Upwind-Schemas L_h^- die neuen Werte $u(t+\tau, x-h)$ und $u(t+\tau, x)$. Gilt nun die CFL-Bedingung, so wird tatsächlich die Information $u(t, x-h)$ entlang einer im blauen Bereich verlaufenden Charakteristik transportiert und beeinflusst daher die Werte $u(t+\tau, x-h)$ und $u(t+\tau, x)$. Ist die CFL-Bedingung dagegen nicht erfüllt, so sollte die Information $u(t, x-h)$ auch andere Werte zum Zeitpunkt $t+\tau$ beeinflussen, was durch das Schema L_h^- nicht geleistet werden kann.
Nachdem wir nun ein Upwind-Schema für den linearen Fall

$$\partial_t u + b\partial_x u = 0$$

kennen, müssen wir uns die Vorgehensweise im nichtlinearen Fall

$$\partial_t u + \partial_x g(u) = 0$$

überlegen. Hierbei können wir nicht davon ausgehen, dass die Flussrichtung $g'(u)$ immer das gleiche Vorzeichen hat.
Das Godunov-Verfahren ist ein Upwind-Schema, welches sich dem Vorzeichen der Flussrichtung $g'(u)$ anpasst und je nach Flussrichtung den Vorwärts- oder Rückwärts-Differenzenquotienten benutzt. Für die nichtlineare Gleichung

$$\partial_t u + \partial_x g(u) = 0$$

lautet es

$$u(t+\tau, x) = u(t,x) - \frac{\tau}{h}\left(h(u(t,x+h), u(t,x)) - h(u(t,x), u(t,x-h))\right)$$

mit

$$h(v,w) = \begin{cases} g(v) & \text{falls } v \geq w \text{ und } g(v) \geq g(w) \\ g(w) & \text{falls } v \geq w \text{ und } g(v) \leq g(w) \\ g(v) & \text{falls } v \leq w \text{ und } g'(v) \geq 0 \\ g(w) & \text{falls } v \leq w \text{ und } g'(w) \leq 0 \\ g((g')^{-1}(0)) & \text{sonst} \end{cases}$$

Weitere geeignete Schemata sind das Enquist-Osher-Schema

$$u(t+\tau, x) = u(t) - \frac{\tau}{h}\left(\int_{u(t,x-h)}^{u(t,x)} g'_+(s)ds + \int_{u(t,x)}^{u(t,x+h)} g'_-(s)ds\right)$$

mit $g'_+(s) = \max(g'(s), 0)$ und $g'_-(s) = \min(g'(s), 0)$, dessen Grundidee analog zum Godunov-Verfahren die Wahl eines zu $g'(u(t,x))$ passenden Differenzenquotienten ist, und das Lax-Friedrichs-Schema

$$\begin{aligned} u(t+\tau, x) = u(t) &- \frac{\tau}{2h}(g(u(t,x+h)) - g(u(t,x-h))) \\ &+ \frac{u(t,x+h) - 2u(t,x) + u(t,x-h)}{2}, \end{aligned}$$

bei dem ein künstlicher Diffusionsterm hinzugefügt wird, um das Verfahren zu stabilisieren.
Wir nennen ein Upwind-Schema $u(t+\tau,x) = H(u(t,x-h), u(t,x), u(t,x+h)) = u(t,x) + \lambda(h(u(t,x+h), u(t,x) - h(u(t,x), u(t,x-h)))$ monoton, falls H in allen drei Argumenten eine monoton wachsende Funktion ist und zitieren aus einem Artikel von

Harten et al. [HHLK76] den folgenden Satz:

Satz 2.11. *Sei H ein monotones Schema mit $h(w,w) = g(w)$. Dann konverigert die numerische Lösung von $\partial_t u + \partial_x g(u) = 0, u(0,x) = u_0(x)$ mit dem Schema H gegen die Entropielösung.*

Beispiel 2.12. Das Godunov-Verfahren, das Lax-Friedrichs-Schema und das Enquist-Osher-Schema erfüllen die Bedingungen an H, falls

$$\frac{\tau}{h} \max_u |g'(u)| \leq 1$$

gilt.

Beweis. $h(w,w) = g(w)$ ist offensichtlich für alle drei Schemata erfüllt. Die Monotonie überprüfen wir, in dem wir H ableiten. Für das Godunov-Verfahren gilt

$$\begin{aligned}
\frac{dH}{du(t,x-h)} &= \begin{cases} \frac{\tau}{h} g'(u(t,x-h)) & \text{falls } g'(u(t,x-h)) > 0 \\ 0 & \text{sonst} \end{cases} \\
\frac{dH}{du(t,x+h)} &= \begin{cases} \frac{-\tau}{h} g'(u(t,x+h)) & \text{falls } g'(u(t,x+h)) < 0 \\ 0 & \text{sonst} \end{cases} \\
\frac{dH}{du(t,x)} &= 1 - \frac{\tau}{h} |g'(u(t,x))|.
\end{aligned}$$

Beim Enquist-Osher-Schema lauten die Ableitungen

$$\begin{aligned}
\frac{dH}{du(t,x-h)} &= \frac{\tau}{h} g'_+(u(t,x-h)) \\
\frac{dH}{du(t,x+h)} &= \frac{-\tau}{h} g'_-(u(t,x+h)) \\
\frac{dH}{du(t,x)} &= 1 - \frac{\tau}{h} |g'(u(t,x))|
\end{aligned}$$

und beim Lax-Friedrichs-Schema sind es

$$\begin{aligned}
\frac{dH}{du(t,x-h)} &= \frac{1}{2}\left(1 + \frac{\tau}{h} g'(u(t,x-h))\right) \\
\frac{dH}{du(t,x+h)} &= \frac{1}{2}\left(1 - \frac{\tau}{h} g'(u(t,x+h))\right) \\
\frac{dH}{du(t,x)} &= 0.
\end{aligned}$$

Demnach konvergieren die Lösungen aller drei Verfahren gegen die Entropielösung, falls $\frac{\tau}{h}|g'(u(t,x))| \leq 1$. □

Als Upwind-Schema zweiter Ordnung wird in dieser Arbeit das Lax-Wendroff-Schema verwendet.

$$\begin{aligned} u(t+\tau,x) = u(t,x) &- \frac{\tau}{2h}(g(u(t,x+h)) - g(u(t,x-h))) \\ &+ \frac{\tau^2}{h^2}g'(\frac{u(t,x+h)+u(t,x)}{2})(g(u(t,x+h) - g(u(t,x)) \\ &- \frac{\tau^2}{h^2}g'(\frac{u(t,x)+u(t,x-h)}{2})(g(u(t,x) - g(u(t,x-h)) \end{aligned}$$

2.3 Zeitdiskretisierung mit Runge-Kutta-Verfahren

2.3.1 Motivation der Runge-Kutta-Verfahren

Der einfachste Ansatz zur Lösung der gewöhnlichen Differentialgleichung

$$\partial_t u = f(t, u(t))$$

ist das explizite Eulerverfahren. Dabei ersetzt man $\partial_t u$ durch den vorwärts Differenzenquotienten und schreibt

$$\frac{u(t+\tau) - u(t)}{\tau} = f(t, u(t)),$$

was explizit durch

$$u(t+\tau) = u(t) + \tau f(t, u(t))$$

gelöst werden kann.
Eine andere Interpretation dieses Verfahrens ist die Integration der äquivalenten Integralgleichung

$$u(t+\tau) = u(t) + \int_t^{t+\tau} f(s, u(s))ds$$

mittels der "linker Eckpunkt"Rechteckregel

$$\int_t^{t+\tau} f(s, u(s))ds \approx \tau f(t, u(t)).$$

Nun kann die Rechteckregel auch anders definiert werden, etwa als "rechter Eckpunkt-Regel

$$\int_t^{t+\tau} f(s, u(s))ds \approx \tau f(t+\tau, u(t+\tau)),$$

was zum impliziten Eulerverfahren

$$u(t+\tau) = u(t) + \tau f(t+\tau, u(t+\tau))$$

führt. Ebenso ist

$$\int_t^{t+\tau} f(s, u(s))ds \approx \tau f(\frac{\tau}{2}, u(t+\frac{\tau}{2}))$$

möglich. Das daraus entwickelte Verfahren

$$u(t+\tau) = u(t) + \tau f(\frac{\tau}{2}, u(t+\frac{\tau}{2})),$$

für das $u(t+\frac{\tau}{2})$ mit einem Zwischenschritt der Art

$$u(t+\frac{\tau}{2}) = u(t) + \frac{\tau}{2} f(t, u(t))$$

berechnet werden muss, ist das Verfahren von Runge aus dem Jahr 1895 [Run95]. Auch Numerische Integrationsformeln höherer Ordnung sind als Grundlage solcher Zeitdiskretisierungen denkbar.
Insgesamt nennt man diese Klasse von Verfahren Runge-Kutta-Verfahren.

2.3.2 Butcher-Schema

Alle Runge-Kutta-Verfahren haben die Form

$$u(t+\tau) = u(t) + \tau \sum_{j=1}^{s} b_j k_j$$

mit

$$k_j = f(t + c_j\tau, u(t) + \sum_{i=1}^{s} a_{ji} k_i \tau)$$

und

$$c_j = \sum_{i=1}^{s} a_{ji}.$$

Dabei wird $\phi(u,t,\tau) = \sum_{j=1}^{s} b_j k_j$ die Verfahrensfunktion genannt.
Zur besseren Übersicht werden die Koeffizienten a_{ij}, b_j, c_j in einem Tableau, dem sogenannten Butcherschema, zusammengefasst.

$$\begin{array}{c|c} c & A \\ \hline & b \end{array} \quad = \quad \begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array}$$

Man nennt ein Runge-Kutta-Verfahren explizit, falls $A := (a_{ij})_{i,j=1\ldots s}$ eine strikte untere Dreiecksmatrix ist, da in diesem Fall alle k_j nacheinander berechnet werden

können. Um alle k_j zu bestimmen muss dagegen im impliziten Fall ein Gleichunssystem gelöst werden, welches durchaus nichtlinear sein kann.
Für die bisher besprochenen Verfahren ergeben sich folgende Tableaus:

0	0
	1

explizites Euler-Verfahren

$$k_1 = f(t, u(t))$$
$$u(t+\tau) = u(t) + \tau k_1$$

1	1
	1

implizites Euler-Verfahren

$$k_1 = f(t+\tau, u(t+\tau))$$
$$u(t+\tau) = u(t) + \tau k_1$$

0	0	0
1/2	1/2	0
	0	1

Verfahren von Runge

$$k_1 = f(t, u(t)), \quad k_2 = f(t + \tfrac{\tau}{2}, u(t) + \tfrac{\tau}{2} k_1), \quad u(t+\tau) = u(t) + \tau k_2$$

2.3.3 Konsistenz

Wir bezeichnen mit η die Lösung der Differentialgleichung $u_t = f(t, u(t))$.

Definition 2.13. Ein Einschrittverfahren hat Konsistenzordnung q, falls

$$||\eta(t+\tau) - u(t+\tau)||_\infty \leq C\tau^{q+1}$$

gilt, wobei $u(t+\tau)$ hier die nach dem Einschrittverfahren gebildete Näherungslösung ist.

Die Konsistenzordnung eines Verfahrens ermittelt man durch den Vergleich der Taylorreihen der Verfahrenslösung und der wahren Lösung. Voraussetzung dafür ist, dass die rechte Seite f der Differentialgleichung $u_t = f(t, u(t))$ hinreichend oft differenzierbar ist.

Beispiel 2.14. Zur Lösung der Differentialgleichung $u_t = f(t, u(t))$ wählen wir das Runge-Verfahren.

0	0	0
1/2	1/2	0
	0	1

Verfahren von Runge

$$
\begin{aligned}
k_1 &= f(t, u(t)) \\
k_2 &= f(t + \frac{\tau}{2}, u(t) + \frac{\tau}{2} k_1) \\
u(t + \tau) &= u(t) + \tau k_2
\end{aligned}
$$

Nun berechnen wir $\eta(t + \tau)$ mit der Taylorreihe

$$\eta(t + \tau) = \eta(t) + \tau \eta'(t) + \frac{\tau^2}{2} \eta''(t) + \mathcal{O}(\tau^3).$$

Für die wahre Lösung η gilt $\eta'(t) = f(t, \eta(t))$, also setzen wir ein

$$\eta(t + \tau) = \eta + \tau f + \frac{\tau^2}{2}(f_t + f_u f) + \mathcal{O}(\tau^3),$$

wobei auf der rechten Seite immer das Argument (t) für η b.z.w. $(t, \eta(t))$ für f weggelassen wurde.
Dies vergleichen wir mit der Taylorreihe der Verfahrenslösung

$$
\begin{aligned}
u(t + \tau) &= u(t) + \tau f(t + \frac{\tau}{2}, u(t) + \frac{\tau}{2} f(t, u(t))) \\
&= u(t) + \tau \left(f(t, u(t)) + \frac{\tau}{2} \left[f_t(t, u(t)) + f_u(t, u(t)) f(t, u(t)) \right] \right) + \mathcal{O}(\tau^3) \\
&= u + \tau f + \frac{\tau^2}{2}(f_t + f_u f) + \mathcal{O}(\tau^3).
\end{aligned}
$$

Wir stellen also fest, dass die Taylorreihen der wahren Lösung $\eta(t + \tau)$ und der Verfahrenslösung $u(t + \tau)$ sich maximal um einen Term der Größenordnung $\mathcal{O}(\tau^3)$ unterscheiden. Daher hat das Verfahren von Runge Konsistenzordnung 2.

Man kann die Berechnung der Taylorreihen auch allgemein für alle Runge-Kutta-Verfahren durchführen und damit allgemeine Bedinungen für jede Konsistenzordnung herleiten. Bis zur Ordnung q=3 wird dies bei Hanke-Bourgeois [HB06] S. 571 durchgeführt. Für höhere Ordnungen werden jedoch die Taylorreihen schnell unübersichtlich, weshalb man auf eine Repräsentation der Ableitungen durch Bäume zurückgreift. Dies kann man bei Hairer, Nørset, Waner [HNW93] im Kapitel II.2 nachlesen.

2.3.4 Konvergenz

Während die Konsistenz sich auf den Fehler in einem einzigen Schritt bezieht, geht es bei der Konvergenz um den Fehler bis zu einer festgelegten Zeit T.

Definition 2.15. Ein Einschrittverfahren zur Lösung des Anfangswertproblems $u_t = f(t, u(t)), u(t_0) = u_0$ auf dem Zeitintervall $[t_0, T]$ hat Konvergenzordnung q, falls es ein $\tau_0 > 0$ gibt, so dass

$$||\eta(t) - u(t)||_\infty \leq C\tau^q \qquad \forall t \in [t_0, T],\ \tau < \tau_0.$$

Wir zitieren nun zwei Sätze, die es erlauben, bei Runge-Kutta-Verfahren direkt von der Konsistenz auf die Konvergenz zu schließen.

Satz 2.16. *Die Verfahrensfunktion ϕ sei stetig bezüglich aller Variablen und genüge der Lipschitzbedingung*

$$||\phi(u, t, \tau) - \phi(v, t, \tau)||_\infty \leq M||u - v||_\infty \quad \forall (t, u), (t, v) \in G,\ 0 < \tau \leq \tau_0$$

bei gegebenenfalls hinreichend kleinem τ_0, dann ist das Einschrittverfahren genau dann konvergent, wenn es konsistent ist.
Ein Einschrittverfahren der Konsistenzordnung q hat unter diesen Voraussetzungen auch Konvergenzordnung q mit

$$||\eta(t_j) - u(t_j)||_\infty \leq \frac{C}{M}\left(\mathrm{e}^{M(t_j - t_0)}\right)\tau^q, \quad j = 0, 1, ..., n.$$

Beweis. Siehe [Lub99] S. 48

Eine Alternative ist folgender Satz.

Satz 2.17. *Die rechte Seite f der Differentialgleichung $u_t = f(t, u(t))$ sei q-mal stetig differenzierbar, f_u sei beschränkt und das Runge-Kutta-Verfahren habe Konsistenzordnung q. Dann existiert ein $\tau_0 > 0$, so dass bei Schrittweite $\tau \in (0, \tau_0)$ alle Näherungen $u(t_i)$ des Runge-Kutta-Verfahrens für $t_j = t_0 + j\tau \in [t_0, T]$ eindeutig definiert sind. Ferner gilt*

$$||\eta(t_j) - u(t_j)||_\infty \leq C\tau^q \qquad \tau < \tau_0$$

wobei die Konstante C von j und τ unabhängig ist, solange t_j in $[t_0, T]$ liegt.

Beweis. Siehe Hanke-Bourgeois [HB06] S. 575

Beide Sätze enthalten eine Schrittweitenbeschränkung der Form $0 < \tau \leq \tau_0$, die wesentlich von dem im folgenden besprochenen Stabilitätsgebiet des Verfahrens abhängt.

2.3.5 A- und L-Stabilität

Definition 2.18. Für ein Einschrittverfahren $u(t+\tau) = R(\tau\lambda)u(t)$ zur Lösung des Anfangswertproblems $u_t = \lambda u,\ u(0) = u_0$ heißt $R(\tau\lambda)$ Stabilitätsfunktion.

Falls $|R(\tau\lambda)| \leq 1$ ist, können Fehler in der Anfangsbedingung u_0 nicht unendlich anwachsen, da in diesem Fall $u(k\tau) = (R(\tau\lambda))^k u_0 \leq u_0$. Dies motiviert analog zu Definition 2.8 die folgende Definition.

Definition 2.19.

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

heißt Stabilitätsgebiet.

Beispiel 2.20. Das explizite Eulerverfahren $u(t+\tau) = u(t) + \tau\lambda u(t)$ hat die Stabilitätsfunktion

$$R(z) = 1 + z,$$

also muss man die Schrittweite τ mit $-1 \leq \tau\lambda \leq 0$ beschränken, um Stabilität zu erreichen. Wählen wir $\lambda = -1$ und $\tau = 0.99$, so ist das explizite Eulerverfahren stabil und konvergent. In Abbildung 2.7 sieht man aber, dass die Approximation sehr schlecht ist.
Das Ergebnis des impliziten Eulerverfahrens lässt sich dagegen kaum von der wahren Lösung unterscheiden, selbst wenn man die Schrittweite τ oder den Parameter λ vergrößert. Die Stabilitätsfunktion des impliziten Eulerverfahren ist $R(z) = \frac{1}{1-z}$, also ist es für $|1 - z| > 1$ stabil und damit auch für sehr große λ oder τ.
Die Lösung $u(t) = u_0 \mathrm{e}^{\lambda t}$ der Testgleichung $u_t = \lambda u,\ u(0) = u_0$ hat die Eigenschaften

$$\begin{aligned}
\mathfrak{Re}(\lambda) > 0 &\Rightarrow \lim_{t\to\infty} |u(t)| = \infty \\
\mathfrak{Re}(\lambda) < 0 &\Rightarrow \lim_{t\to\infty} |u(t)| = 0 \\
\mathfrak{Re}(\lambda) = 0 &\Rightarrow |u(t)| = 1 \quad \forall t > 0.
\end{aligned}$$

Wir verlangen nun von unseren Verfahren, dass sie diese Eigenschaften der Lösung beibehalten.

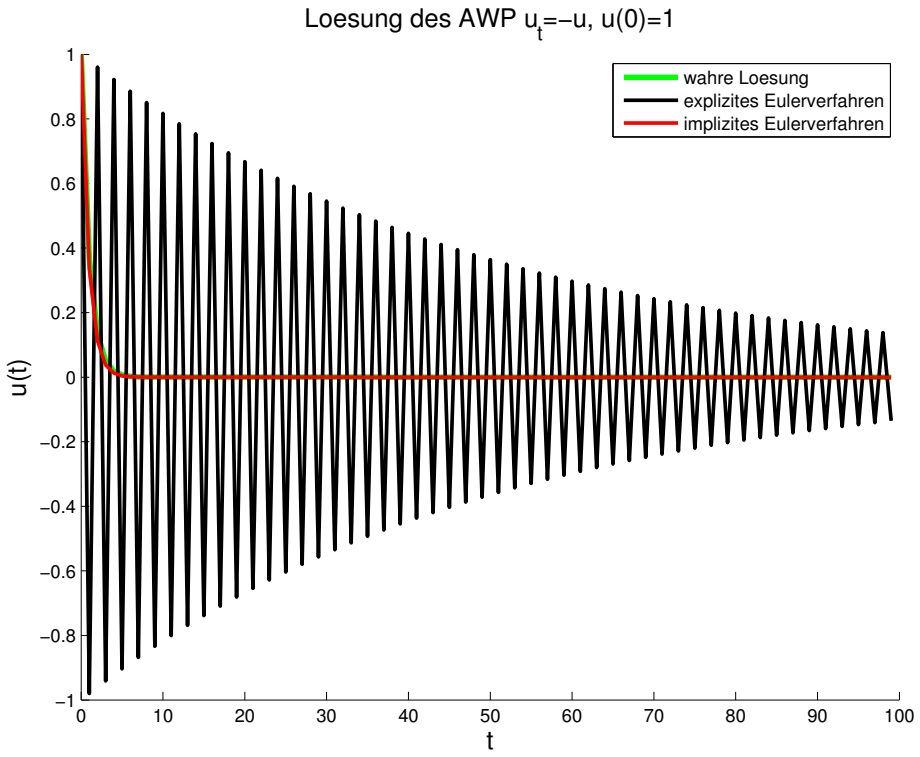

Abbildung 2.7: Lösung von $u_t = \lambda u$ mit explizitem und implizitem Eulerverfahren

Definition 2.21. Wir nennen ein Verfahren A-stabil, falls

$$|R(z)| \leq 1 \qquad \forall z \in \mathbb{C}, \mathfrak{Re}(z) \leq 0,$$

also falls das Verfahren für alle Schrittweiten τ und alle λ mit $\mathfrak{Re}(\lambda) < 0$ stabil ist.

Das explizite Eulerverfahren ist nicht A-stabil, da

$$|R(z)| = |1 + z| = (1 + \mathfrak{Re}(z))^2 + \mathfrak{Im}(z)^2 > 1 \text{ für } \mathfrak{Re}(z) < -2.$$

Das implizite Eulerverfahren ist A-stabil, da

$$|R(z)| = |\frac{1}{1-z}| = \frac{1}{(1 - \mathfrak{Re}(z))^2 + \mathfrak{Im}(z)^2} < 1 \text{ für } \mathfrak{Re}(z) < 0.$$

Definition 2.22. Wir nennen ein Verfahren L-stabil, falls

$$\lim_{\mathfrak{Re}(z) \to -\infty} R(z) = 0,$$

also falls es die Eigenschaft $\mathfrak{Re}(\lambda) < 0 \Rightarrow \lim_{t\to\infty} |u(t)| = 0$ erhält.

Das implizite Eulerverfahren ist L-stabil, da

$$\lim_{\mathfrak{Re}(z)->-\infty} |R(z)| = \lim_{\mathfrak{Re}(z)->-\infty} \frac{1}{(1-\mathfrak{Re}(z))^2 + \mathfrak{Im}(z)^2} = 0 \text{ für } \mathfrak{Im}(z) \text{ fest.}$$

Die Stabilitätsfunktion eines Runge-Kutta-Verfahrens hat allgemein die Form

$$R(z) = 1 + z\beta^*(I - z\alpha)^{-1}\mathbb{1},$$

wobei β und α der Vektor und die Matrix aus dem Butcherschema sind und $\mathbb{1} = (1, \ldots, 1)^t$.
Mit Hilfe der Stabilitätsfunktionen können wir nun die Stabilitätsgebiete einiger gängiger Verfahren graphisch darstellen.

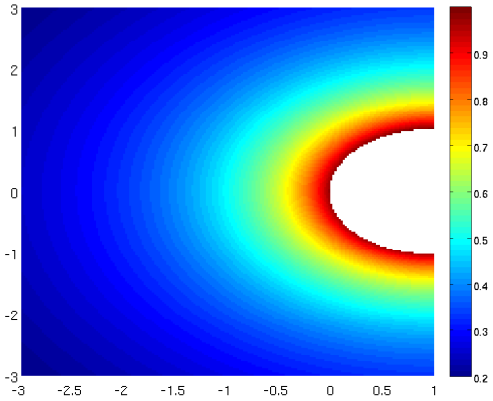

Abbildung 2.8: Stabilitätsgebiet des impliziten Euler-Verfahrens

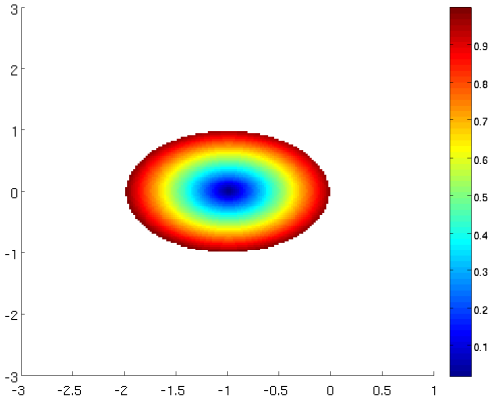

Abbildung 2.9: Stabilitätsgebiet des expliziten Euler-Verfahrens

Abbildung 2.8 zeigt das Stabilitätsgebiet des impliziten Eulerverfahrens. Man sieht, dass die gesamte negative Halbebene im Stabilitätsgebiet enthalten ist. Das Verfahren ist also A-stabil. Außerdem geht der Betrag der Stabilitätsfunktion für große negative Realteile gegen Null, da das Verfahren L-stabil ist.
Die Abbildungen 2.9, 2.10 und 2.11 zeigen, dass die Stabilitätsgebiete der vorgestellten expliziten Verfahren beschränkt sind. Daher können diese Verfahren weder A- noch L-stabil sein. Man erhält in jedem Fall eine Schrittweitenbeschränkung.

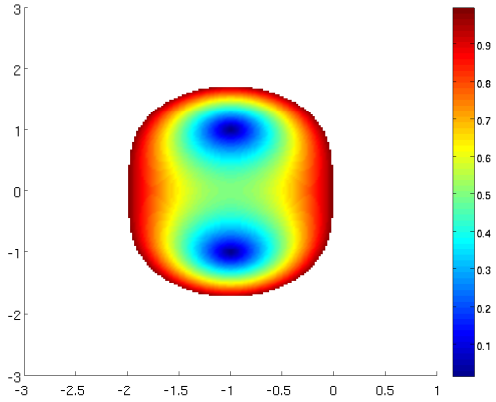

Abbildung 2.10: Stabilitätsgebiet des Verfahrens von Runge

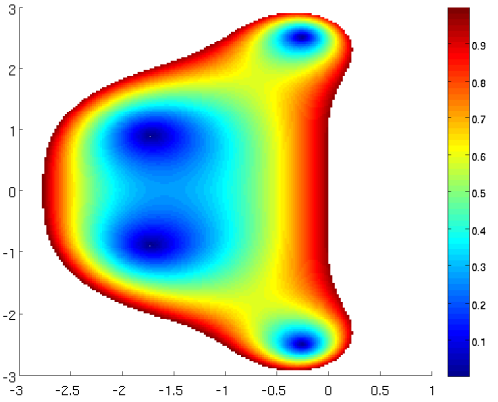

Abbildung 2.11: Stabilitätsgebiet des klassischen Runge-Kutta-Verfahrens

3 IMEX

3.1 IMEX-Idee

Wendet man implizite Runge-Kutta-Verfahren auf nichtlineare Probleme an, so muss man in jedem Schritt ein nichtlineares Gleichungssystem lösen, was einen sehr hohen Rechenaufwand zur Folge hat. Andererseits haben wir in den Grundlagen gesehen, dass explizite Runge-Kutta-Verfahren für steife Probleme ungeeignet sind, da sie ungünstige Bedingungen an die erlaubte Schrittweite stellen.
Bei der Burgers-Gleichung

$$\partial_t u + u\partial_x u = \varepsilon \partial_{xx} u$$

tritt aber sowohl der steife Term $\partial_{xx} u$ als auch der nichtlineare Term $u\partial_x u$ auf. Implizit-Explizite Verfahren, kurz IMEX-Verfahren, lösen dieses Problem, indem sie den linearen, steifen Teil implizit und den nichtlinearen, nicht steifen Teil explizit behandeln.
Wir schreiben die Differentialgleichung um,

$$\partial_t u = -u\partial_x u + \varepsilon \partial_{xx} u = f(u) + g(u)$$

wobei $f(u) = -u\partial_x u$ der nichtlineare und $g(u) = \varepsilon \partial_{xx} u$ der steife Teil ist.
Nun lösen wir das äquivalente System

$$\begin{aligned} u &= v + w \\ v_t &= \hat{f}(v, w) = f(v + w) \\ w_t &= \hat{g}(v, w) = g(v + w) \end{aligned}$$

wobei wir die zweite Differentialgleichung mit einem expliziten und die dritte mit einem impliziten Runge-Kutta-Verfahren lösen wollen.

Beispiel 3.1. Mit der Iterationsvorschrift

$$u(t+\tau,x) = u(t,x) + \tau \cdot (f(u(t,x)) + g(u(t+\tau,x)))$$

behandeln wir den nichtlinearen Teil explizit und den steifen Teil implizit. Dieses auf dem expliziten und dem impliziten Eulerverfahren basierende IMEX-Verfahren ist das einfachste der von Ascher et al. [ARS97] vorgestellten Verfahren. Für den unten vorgestellten Algorithmus ist es praktisch, das explizite Eulerverfahren

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

mit der Vorschrift $k_1 = f(t, u(t,x)), \quad u(t+\tau,x) = u(t,x) + \tau k_1$ aufzufüllen. Man erhält

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

mit der Vorschrift

$$\begin{aligned} k_1 &= f(t, u(t,x)) \\ k_2 &= f(t+\tau, u(t,x) + \tau k_1) \\ u(t+\tau,x) &= u(t,x) + \tau k_1. \end{aligned}$$

Im Algorithmus verwenden wir dann das IMEX-Paar

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

explizites Euler-Verfahren $u(t+\tau,x) = u(t,x) + \tau f(t, u(t,x))$

implizites Euler-Verfahren $u(t+\tau,x) = u(t,x) + \tau g(t+\tau, u(t+\tau,x)))$. (3.2)

3.1.1 Algorithmus

Algorithmus 1 beschreibt die Vorgehensweise für IMEX-Verfahren, bei denen ein s-stufiges diagonal implizites und ein σ-stufiges explizites Runge-Kutta-Verfahren kom-

biniert wird.

Eingabe : $u(t) \in \mathbb{R}^m$, Funktionen $f(u)$ und $g(u)$, Butcherschemata für explizites und implizites Verfahren
Ausgabe : $u(t+\tau) \in \mathbb{R}^m$

Berechne die Ableitung des explizit behandelten Teils $\hat{k}_1 = f(u(t))$
für $i = 1..s$
  Führe den Schritt zur Zwischenstelle u_i für f und g gemeinsam aus
  $u_i = u(t) + \tau \sum_{j=1}^{i} a_{i,j} k_j + \tau \sum_{j=1}^{i} \hat{a}_{i+1,j} \hat{k}_j$
  wobei $k_i = g(u_i)$ gelöst werden muss.
  Berechne die Ableitung des explizit behandelten Teils an der Zwischenstelle durch
  $\hat{k}_{i+1} = f(u_i)$
Berechne den neuen Wert $u(t+\tau)$ durch
$u(t+\tau) = u(t) + \tau \sum_{j=1}^{s} b_j k_j + \sum_{j=1}^{\sigma} \hat{b}_j \hat{k}_j$

Algorithmus 1 : Pseudocode des IMEX-Verfahrens

Für diese Arbeit wurde als erstes das oben vorgestellte IMEX-Paar der Eulerverfahren und als zweites die folgende Variante dieses Paares implementiert:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 0 & 1 \end{array} \qquad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \tag{3.3}$$

explizites Euler-Verfahren implizites Euler-Verfahren

Es ergibt sich die Vorschrift $u(t+\tau, x) = u(t,x) + \tau(f(u_1) + g(u_1))$ mit $u_1 = u(t,x) + \tau g(u_1) + \tau f(u(t,x))$. Beides sind Verfahren erster Ordnung.
Das einfachste IMEX-Verfahren zweiter Ordnung wird aus der expliziten und impliziten Mittelpunktsregel zusammengesetzt.

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array} \qquad \begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array} \tag{3.4}$$

explizite Mittelpunktsregel implizite Mittelpunktsregel

Ein L-stabiles Verfahren zweiter Ordnung ist

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \gamma & \gamma & 0 & 0 \\ 1 & \delta & 1-\delta & 0 \\ \hline & \delta & 1-\delta & 0 \end{array} \qquad \begin{array}{c|cc} \gamma & \gamma & 0 \\ 1 & 1-\gamma & \gamma \\ \hline & 0 & 1 \end{array} \tag{3.5}$$

explizit implizit

mit $\gamma = (2-\sqrt{2})/2$ und $\delta = 1 - 1/(2\gamma)$.
Als Beispiel für ein Verfahren dritter Ordnung wurde das folgende IMEX-Paar gewählt.

$$\begin{array}{c|ccccc} 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{2}{3} & \frac{11}{18} & \frac{1}{18} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{5}{6} & \frac{-5}{6} & \frac{1}{2} & 0 & 0 \\ 1 & \frac{1}{4} & \frac{7}{4} & \frac{3}{4} & \frac{-7}{4} & 0 \\ \hline & \frac{1}{4} & \frac{7}{4} & \frac{3}{4} & \frac{-7}{4} & 0 \end{array} \qquad \begin{array}{c|cccc} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{2}{3} & \frac{1}{6} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{-1}{2} & \frac{1}{2} & \frac{1}{2} & 0 \\ 1 & \frac{3}{2} & \frac{-3}{2} & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{3}{2} & \frac{-3}{2} & \frac{1}{2} & \frac{1}{2} \end{array} \tag{3.6}$$

explizit implizit

3.2 Analysis der IMEX-Verfahren

3.2.1 Konsistenz

Wir müssen nun wie im Abschnitt 2.3.3 die Taylorreihen der Verfahrenslösung und der wahren Lösung für unser System

$$\begin{aligned} u &= v + w \\ v_t &= \hat{f}(v, w) = f(v + w) \\ w_t &= \hat{g}(v, w) = g(v + w) \end{aligned}$$

miteinander vergleichen. Wir berechnen zunächst die Ableitungen von $v(t)$ und benuzten dabei die Abkürzungen $f = f(v + w)$ und $g = g(v + w)$.

$$\begin{aligned} v_t &= f \\ v_{tt} &= f_v f + f_w g \\ &= f'f + f'g \\ v_{ttt} &= f''ff + f''fg + f'f'f + f'f'g + f''fg + f''gg + f'g'f + f'g'g \end{aligned}$$

Für die Ableitungen von w müssen einfach f und g vertauscht werden.
Die Taylorreihe der wahren Lösung η hat also die Form

$$\begin{aligned} \eta(t+\tau) &= \eta(t) + \tau(f+g) + \frac{\tau^2}{2}(f'f + f'g + g'g + g'f) \\ &\quad + \frac{\tau^3}{6}(f''ff + f''fg + f'f'f + f'f'g + f''fg + f''gg + f'g'f + f'g'g \\ &\quad + g''gg + g''gf + g'g'g + g'g'f + g''gf + g''ff + g'f'g + g'f'f) + \mathcal{O}(\tau^4) \\ &= \eta(t) + \tau(f+g) + \frac{\tau^2}{2}(f'f + f'g + g'g + g'f) \\ &\quad + \frac{\tau^3}{6}(f''ff + 2f''fg + f''gg + g''gg + 2g''gf + g''ff \\ &\quad + f'f'f + f'f'g + 2f'g'f + 2f'g'g + g'g'g + g'g'f) + \mathcal{O}(\tau^4) \end{aligned}$$

Nun müssen wir die Taylorreihe der Verfahrenslösung berechnen. Dazu werfen wir einen Blick auf Zeile 9 des Algorithmus, stellen sie um und entwickeln zunächst $g(u_j)$ und $f(u_j)$ nach Taylor.

$$
\begin{aligned}
u(t+\tau) &= u(t) + \tau(\sum_{j=1}^{s} b_j k_j + \sum_{j=0}^{s} \hat{b}_{j+1}\hat{k}_{j+1}) \\
&= u(t) + \tau(\sum_{j=1}^{s} b_j g(u_j) + \sum_{j=0}^{s} \hat{b}_{j+1} f(u_j)) \\
&= u(t) \\
&+ \tau(\sum_{j=1}^{s} b_j [g + g' \cdot (u_j - u(t)) + \frac{1}{2} g'' \cdot (u_j - u(t))^2 + \frac{1}{6} g''' \cdot (u_j - u(t))^3 ...] \\
&+ \sum_{j=0}^{s} \hat{b}_{j+1} [f + f' \cdot (u_j - u(t)) + \frac{1}{2} f'' \cdot (u_j - u(t))^2 + \frac{1}{6} f''' \cdot (u_j - u(t))^3 ...])
\end{aligned}
$$

Nun müssen wir $(u_j - u(t))$ ersetzen, was mit Hilfe von Zeile 4 des Algorithmus geschieht.

$$
\begin{aligned}
u_j - u(t) &= \tau \left(\hat{a}_{j+1,1} f(u(t)) + \sum_{k=1}^{s} a_{jk} k_k + \hat{a}_{j+1,k+1} \hat{k}_{k+1} \right) \\
&= \tau \left(\hat{a}_{j+1,1} f(u(t)) + \sum_{k=1}^{s} a_{jk} g(u_k) + \hat{a}_{j+1,k+1} f(u_k) \right) \\
&= \tau \left(\sum_{k=1}^{s} a_{jk} [g + g' \cdot (u_k - u(t)) + \frac{1}{2} g'' \cdot (u_k - u(t))^2 + \frac{1}{6} g''' \cdot (u_k - u(t))^3 ...] \right. \\
&+ \hat{a}_{j+1,1} f(u(t)) \\
&\left. + \sum_{k=1}^{s} \hat{a}_{j+1,k+1} [f + f' \cdot (u_k - u(t)) + \frac{1}{2} f'' \cdot (u_k - u(t))^2 + \frac{1}{6} f''' \cdot (u_k - u(t))^3 ...] \right)
\end{aligned}
$$

Wiederholtes Einsetzen ergibt schließlich

$$\begin{aligned}
u(t+\tau) = u(t) &+ \tau(\sum_{j=0}^{s} \hat{b}_{j+1} f + \sum_{j=1}^{s} b_j g) \\
&+ \tau^2 \left(f'f \sum_{j=0}^{s} \hat{b}_{j+1}\hat{c}_{j+1} + f'g \sum_{j=1}^{s} \hat{b}_{j+1} c_j + g'f \sum_{j=1}^{s} b_j \hat{c}_{j+1} + g'g \sum_{j=1}^{s} b_j c_j \right) \\
&+ \tau^3 \left(f'f' \cdot \sum_{j=0}^{s} \hat{b}_{j+1} (\sum_{k=0}^{s} \hat{a}_{j+1,k+1}(\hat{c}_{k+1} f + c_k g) + \sum_{k=1}^{s} a_{jk}(\hat{c}_{k+1} f + c_k g)) \right. \\
&+ g'g' \cdot \sum_{j=1}^{s} b_j (\sum_{k=1}^{s} a_{jk}(c_k g + \hat{c}_{k+1} f) + \sum_{k=0}^{s} \hat{a}_{j+1,k+1}(c_k g + \hat{c}_{k+1} f)) \\
&+ \frac{1}{2} f'' \cdot \sum_{j=0}^{s} \hat{b}_{j+1} (\hat{c}_{j+1}^2 ff + \hat{c}_{j+1} c_j fg + c_j^2 gg) \\
&\left. + \frac{1}{2} g'' \cdot \sum_{j=1}^{s} b_j (c_j^2 gg + c_j \hat{c}_{j+1} gf + \hat{c}_{j+1}^2 ff) \right) + \mathcal{O}(\tau^4).
\end{aligned}$$

Der Vergleich mit der Taylorreihe der wahren Lösung führt zu den folgenden Ordnungsbedingungen.

Satz 3.7. *Ein IMEX-Runge-Kutta-Verfahren mit*

$$\sum_{j=0}^{s} \hat{b}_{j+1} = \sum_{j=1}^{s} b_j = 1$$

hat Konsistenzordnung 1. *Falls zusätzlich*

$$\sum_{j=0}^{s} \hat{b}_{j+1}\hat{c}_{j+1} = \sum_{j=1}^{s} \hat{b}_{j+1} c_j = \sum_{j=1}^{s} b_j \hat{c}_{j+1} = \sum_{j=1}^{s} b_j c_j = \frac{1}{2}$$

erfüllt ist, hat es Konsistenzordnung 2.
Falls zusätzlich die folgenden Bedingungen erfüllt werden, hat es Konsistenzordnung 3.

$$\begin{array}{lll}
\frac{1}{3} = \sum_{j=0}^{s} \hat{b}_{j+1}\hat{c}_{j+1}^2 & = \sum_{j=1}^{s} \hat{b}_{j+1}\hat{c}_{j+1} c_j & = \sum_{j=1}^{s} \hat{b}_{j+1} c_j^2 \\
= \sum_{j=1}^{s} b_j c_j^2 & = \sum_{j=1}^{s} b_j c_j \hat{c}_{j+1} & = \sum_{j=1}^{s} b_j \hat{c}_{j+1}^2
\end{array}$$

$$
\begin{array}{ll}
\frac{1}{6} & = \sum_{j=1}^{s} \hat{b}_{j+1} \sum_{k=1}^{s} \hat{a}_{j+1,k+1} \hat{c}_{k+1} \quad = \sum_{j=1}^{s} \hat{b}_{j+1} \sum_{k=1}^{s} \hat{a}_{j+1,k+1} c_k \\
& = \sum_{j=1}^{s} \hat{b}_{j+1} \sum_{k=0}^{s} a_{jk} \hat{c}_{k+1} \quad = \sum_{j=1}^{s} \hat{b}_{j+1} \sum_{k=0}^{s} a_{jk} c_k \\
& = \sum_{j=1}^{s} b_j \sum_{k=1}^{s} a_{jk} c_k \quad = \sum_{j=1}^{s} b_j \sum_{k=1}^{s} a_{jk} \hat{c}_{k+1} \\
& = \sum_{j=1}^{s} b_j \sum_{k=0}^{s} \hat{a}_{j+1,k+1} c_k \quad = \sum_{j=1}^{s} b_j \sum_{k=0}^{s} \hat{a}_{j+1,k+1} \hat{c}_{k+1}
\end{array}
$$

Beispiel 3.8. Die beiden Euler-IMEX-Verfahren haben Konsistenzordnung 1. Die IMEX-Verfahren 3.4 und 3.5 haben Konsistenzordnung 2 und das Verfahren 3.6 hat Konsistenzordnung 3.

3.2.2 A- und L-Stabilität

Wir haben die IMEX-Verfahren so konstruiert, dass wir die zweite Ableitung, $D_h^2 u$, implizit und die erste Ableitung, $D_h u$, explizit behandeln. D_h^2 hat reelle Eigenwerte, während D_h komplexe Eigenwerte hat. Dies motiviert uns, als Testgleichung für die IMEX-Verfahren

$$u_t = \mathrm{i}bu - au$$

mit $a, b > 0$ zu nehmen, wobei wir den reellen Teil implizit und den imaginären Teil explizit behandeln.
Wir müssen nun einen Schritt im IMEX-Verfahren in die Form

$$u(t + \tau) = R(z)u(t)$$

übersetzen. Zeile 4 des Algorithmus lautet

$$u_i = u(t) + \tau \sum_{j=1}^{i} \alpha_{i,j} k_j + \tau \sum_{j=1}^{i} \alpha_{i+1,j} \hat{k}_j$$

mit $k_i = g(u_i) = au_i$. Wir stellen um zu

$$u_i = \frac{(1 + \mathrm{i}\tau b)u(t) + \sum_{j=1}^{i-1} (\tau a \alpha_{ij} + \mathrm{i}\tau b \hat{\alpha}_{i+1,j}) u_j}{1 - \tau a \alpha_{ii}}$$

und schreiben den gesamten IMEX-Schritt mit $z = x + \mathrm{i}y = \tau a + \mathrm{i}\tau b$ als

$$u(t + \tau) = u(t) + z \sum_{j=1}^{s} b_j u_j = R(z)u(t).$$

Beispiel 3.9. Für das einfache Euler-Paar ergibt sich die Stabilitätsfunktion

$$R(z) = \frac{1 + \mathrm{i}y}{1 - x}$$

und für die Variante erhält man

$$R(z) = 1 + z\frac{1 + \mathrm{i}y}{1 - x}$$

.

Für die Mittelpunktsregel erhält man

$$R(z) = 1 + z\frac{1 + \mathrm{i}\frac{1}{2}y}{1 - \frac{1}{2}x}.$$

und für das L-stabile Verfahren zweiter Ordnung

$$R(z) = \frac{(1 + \mathrm{i}\delta y) + ((1 - \gamma)x + (1 - \delta)\mathrm{i}y)\frac{1+\mathrm{i}\gamma y}{1-\gamma x}}{1 - \gamma x}.$$

Für das Verfahren dritter Ordnung errechnet MuPAD die folgende Stabilitätsfunktion:

$$\begin{aligned} R(z) =& \frac{-7y^4 - 144y^2 + 288 + 48x^3 - 228x + 3x^2y^2 + 144xy^2}{18(x-2)^4} \\ &+ \mathrm{i}\frac{-48y^3 + 288y + 29x^3y + 57xy^3 - 288xy}{18(x-2)^4}. \end{aligned}$$

Zu den Stabilitätsfunktionen kann man nun die Stabilitätsgebiete nach Definition 2.19 zeichnen.

Wir betrachten nun die Stabilitätsgebiete und stellen als erstes fest, dass alle Stabilitätsgebiete nur jeweils einen kleinen Teil der imaginären Achse $x = 0$ abdecken. Das Stabilitätsgebiet des Euler-IMEX-Verfahrens in Abbildung 3.2 berührt gar nur für $z = 0$ die imaginäre Achse.
Da keines der Stabilitätsgebiete die gesamte negative Halbebene enthält, ist keines der IMEX-Verfahren A-stabil.
Für das reine Diffusionsproblem $u_t = \varepsilon u_{xx}$ dürfen wir dagegen für alle $\varepsilon > 0, \tau > 0$ ein Abklingen der Lösung erwarten, da alle fünf Stabilitätsgebiete die gesamte negative reelle Achse mit einschließen.

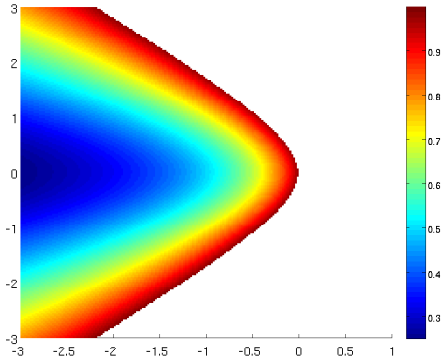

Abbildung 3.1: Stabilitätsgebiet des Euler-IMEX-Verfahrens 3.2

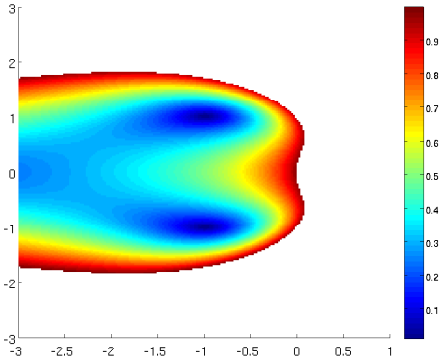

Abbildung 3.2: Stabilitätsgebiet der Variante des Euler-IMEX-Verfahrens 3.3

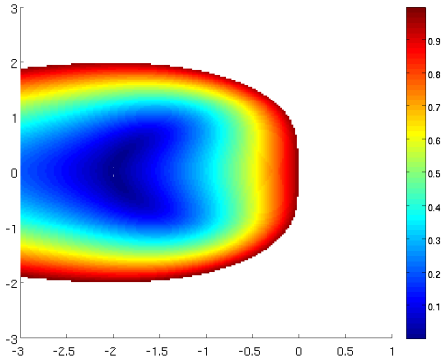

Abbildung 3.3: Stabilitätsgebiet des IMEX-Verfahrens mit Mittelpunktsregel 3.4

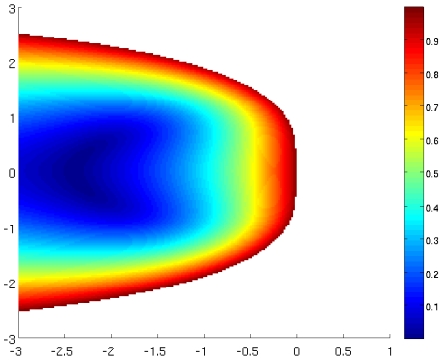

Abbildung 3.4: Stabilitätsgebiet des L-stabilen Verfahrens zweiter Ordnung 3.5

Allerdings sind nur die Verfahren 3.2, 3.5 und 3.6 L-stabil, das heißt, es gilt

$$\lim_{\mathfrak{Re}(z)\to\infty} |R(z)| = 0 \text{ für festes } \mathfrak{Im}(z).$$

Die Abbildungen 3.1, 3.4 und 3.5 veranschaulichen diese Tatsache, während man bei den Stabilitätsgebieten in Abbildung 3.2 und 3.3 erkennen kann, dass die Verfahren für größere Imaginärteile von z gar nicht stabil sein können.

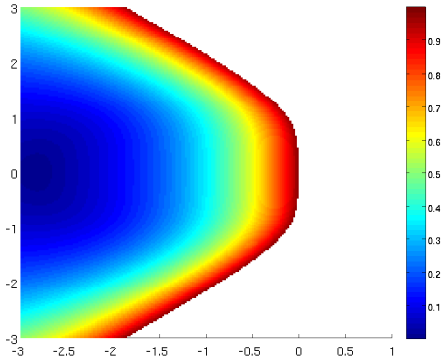

Abbildung 3.5: Stabilitätsgebiet des Verfahrens dritter Ordnung 3.6

3.2.3 Konvergenz

Für Differentialgleichungen der Form

$$\partial_t u = Au, u(0) = u_0$$

mit einer linearen Abbildung A, zum Beispiel das in Abschnitt 2.2 vorgestellte semidiskrete Problem

$$\partial_t u = A_h u = (-\varepsilon D_h^2 + B_h D_h + C_h)u,$$

sagt der Äquivalenzsatz von Lax [LR56] aus, dass ein konsistentes Verfahren genau dann konvergent ist, wenn es stabil ist.
In den Programmbeispielen in dieser Arbeit wird die Burgers-Gleichung

$$\partial_t u = -u\partial_x u + \partial_{xx} u, \quad u(0) = \frac{1}{2}\sin(2\pi x)$$

behandelt. Das zugehörige semidiskrete Problem

$$\partial_t u = -U_h D_h u + D_h^2 u, \quad u(0) = \frac{1}{2}\sin(2\pi x)$$

ist nicht linear, weshalb wir den Äquivalenzsatz von Lax nicht anwenden können. Wir zeigen für das Euler-IMEX-Verfahren die Konvergenz mit Hilfe von Satz 2.16. Dabei wollen wir $-U_h D_h u$ explizit und $D_h^2 u$ implizit behandeln. Die Verfahrensfunktion hat

also die Form

$$\phi(\tau, t, u) = -U_h(t)D_h u(t) + D_h^2 u(t+\tau).$$

Wir wissen bereits aus Abschnitt 2.1, wie die Lösung aussieht und können daher annehmen, dass $||U_h||_\infty \leq 1$. Mit dieser Annahme können wir die folgende Abschätzung durchführen:

$$\begin{aligned}
&||\phi(\tau, t, u) - \phi(\tau, t, v)||_\infty \\
&= || - U_h(t)D_h u(t) + D_h^2 u(t+\tau) + V_h(t)D_h v(t) - D_h^2 v(t+\tau)||_\infty \\
&\leq || - U_h(t)D_h u(t) + V_h(t)D_h v(t)||_\infty + ||D_h^2 u(t+\tau) - D_h^2 u(t+\tau)||_\infty \\
&\leq ||(-||U_h||_\infty D_h u(t) + ||V_h||_\infty D_h v(t))||_\infty + ||D_h^2 u(t+\tau) - D_h^2 v(t+\tau)||_\infty \\
&\leq || - D_h||_\infty ||u - v||_\infty + ||D_h^2||_\infty ||u - v||_\infty \\
&= (2n + 4n^2)||u - v||_\infty.
\end{aligned}$$

Die Verfahrensfunktion ist also lipschitzstetig, wenn auch mit der sehr großen Lipschitzkonstanten $(2n + 4n^2)$. Nach Satz 2.16 ist damit das Euler-IMEX-Verfahren konvergent. Für den Fehler gilt allerdings

$$E(\tau) = \max_j |\eta(t_j) - u(t_j)| \leq \frac{K}{2n + 4n^2}(\mathrm{e}^{(2n+4n^2)(t_n - t_0)} - 1)\tau,$$

das heißt, man muss die Zeitschrittweite unter Umständen sehr klein wählen, um den Fehler befriedigend klein zu halten.

4 Implementation und Ergebnisse

4.1 Anmerkungen zur Implementation

Alle in dieser Arbeit vorgestellten Upwind- und IMEX-Verfahren wurden in Matlab implementiert. Dabei musste Algorithmus 1 zunächst nur einmal allgemein implementiert werden, da man die IMEX-Verfahren einfach durch Austauschen der Butcherschemata wechseln kann.

```
%Butcherschemata in aex,bex bzw aim,bim gespeichert
%Schleife fuer numberofsteps Schritte
for m=1:numberofsteps-1
%IMEX-ALGORITHMUS
        kdach(:,1)=-a*(D_h*U(:,m));
        for l=1:s
                r=U(:,m);
                for g=1:l-1
                        r=r+tau*aim(l,g)*q(:,g)+tau*aex(l+1,g)*p(:,g);
                end
                r=r+tau*aex(l+1,l)*p(:,l);
                D=eye(n)-tau*epsilon*aim(l,l)*D_hh;
                u_i=D^-1*r;
                k(:,l)=epsilon*D_hh*u_i;
                kdach(:,l+1)=-a*(D_h*u_i);
        end
        U(:,m+1)=U(:,m)+tau*k*bim + tau*kdach*bex;
end
```

Für die Upwind-Verfahren war es allerdings nötig, die Differenzenquotienten direkt im Algorithmus anzupassen.

```
%Butcherschemata in aex,bex bzw aim,bim gespeichert
%Schleife fuer numberofsteps Schritte
for m=1:numberofsteps-1
%IMEX-ALGORITHMUS
        %ist au<0 so nimm den Rueckwaerts-Differenzenquotienten, andernfalls den Vorwaerts-
            Differenzenquotienten
        kdach(:,1)=c*U(:,m).*U(:,m).*U(:,m)+a*((Dminus_h*U(:,m)).*((a*U(:,m))<0)+(Dplus_h*U(:,m))
            .*((a*U(:,m))>=0)).*U(:,m);
        for l=1:s
                r=U(:,m);
                for g=1:l-1
                        r=r+tau*aim(l,g)*k(:,g)+tau*aex(l+1,g)*kdach(:,g);
                end
                r=r+tau*aex(l+1,l)*kdach(:,l);
                D=eye(n)-tau*epsilon*aim(l,l)*D_hh;
                u_i=D^-1*r;
                k(:,l)=D\(epsilon*D_hh*r);
                u_i=r+k*aim(l,l)*q(:,l);
                kdach(:,l+1)=c*u_i.*u_i.*u_i+a*((Dminus_h*u_i).*((a*u_i)<0)+(Dplus_h*u_i).*((a*u_i
                    )>=0)).*u_i;
```

```
        end
        U(:,m+1)=U(:,m)+tau*k*bim + tau*kdach*bex;
end
```

4.2 Numerische Resultate

4.2.1 Behandlung des linearen Problems mit IMEX-Verfahren

Wir beginnen mit dem linearen Problem

$$\partial_t u + a\partial_x u = \varepsilon \partial_{xx} u, \quad u_0(x) = u(0,x) = \sin(2\pi x)$$

für $x \in [0,1]$ mit der periodischen Randbedingung $u(t,0) = u(t,1)$.
Die exakte Lösung ist

$$u(x,t) = \sin(2\pi(x-at))\, \exp(-4\pi^2 \varepsilon t).$$

Die Ortsdiskretisierung wird mit dem zentralen Differenzenquotienten berechnet. Das semidiskrete Problem hat also die Form

$$\partial_t u = -aD_h^0 u + \varepsilon D_h^2 u,$$

für $u : [0,\infty) \to \mathbb{R}^n$ mit $h = \frac{1}{n}$.
Auf das semidiskrete Problem werden nun die IMEX-Verfahren angewandt.
Abbildung 4.1 zeigt links den größten absoluten und rechts den größten relativen Fehler im Intervall $t = [0,2]$ in Abhängigkeit von dem Diffusionsparameter ε für $a = 1$, $h = \frac{1}{63}$ und $\tau = \frac{1}{128}$.

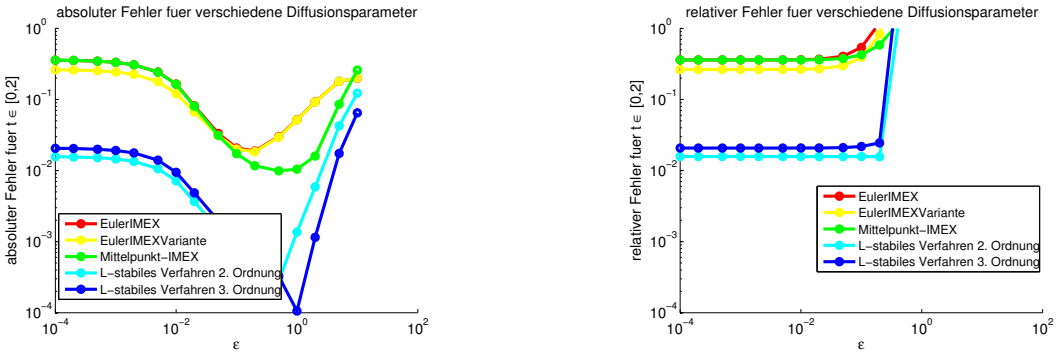

Abbildung 4.1: Absoluter und relativer Fehler für $h = \frac{1}{63}$

Abbildung 4.2 zeigt die mittels Mittelpunkt-IMEX-Verfahren berechnete Lösung für $\varepsilon = 0.02$, $a = 1$, $h = \frac{1}{63}$ und $\tau = \frac{1}{128}$ aus zwei verschiedenen Perspektiven.
Wir interessieren uns nun noch dafür, wie weit man die Ortsdiskretisierung verfeinern

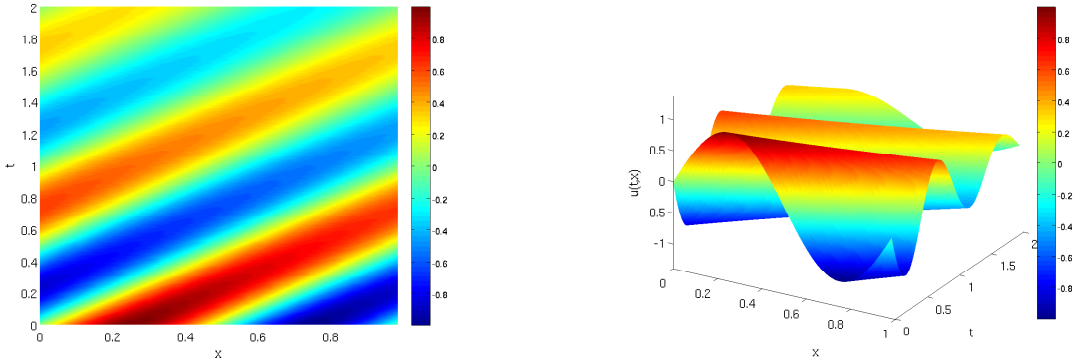

Abbildung 4.2: Lösung des Mittelpunkt-IMEX-Verfahrens

kann, ohne zu kleineren Zeitschritten gezwungen zu werden. Dazu wurde die Berechnung der absoluten und relativen Fehler für feinere Ortsdiskretisierungen wiederholt, während die Zeitschrittweite $\tau = \frac{1}{128}$ und der Parameter $a = 1$ festgehalten wurden. Abbildung 4.3 zeigt die Werte für $h = \frac{1}{126}$. Wir sehen, dass die feinere Ortsdiskretisierung bei den Verfahren 3.5 und 3.6 sogar eine bessere Lösung bewirkt, da der durch die Ortsdiskretisierung gemachte Fehler nun kleiner ist. Das Euler-IMEX-Verfahren und das Mittelpunkt-IMEX-Verfahren stoßen bei diesem Verhältnis von Orts- und Zeitschrittweite zumindest für kleine ε jedoch an ihre Grenzen.
In Abbildung 4.4 mit Werten für $h = \frac{1}{252}$ erkennt man, dass nun alle fünf Verfahren bei sehr kleinen oder großen ε keine guten Lösungen mehr liefern.

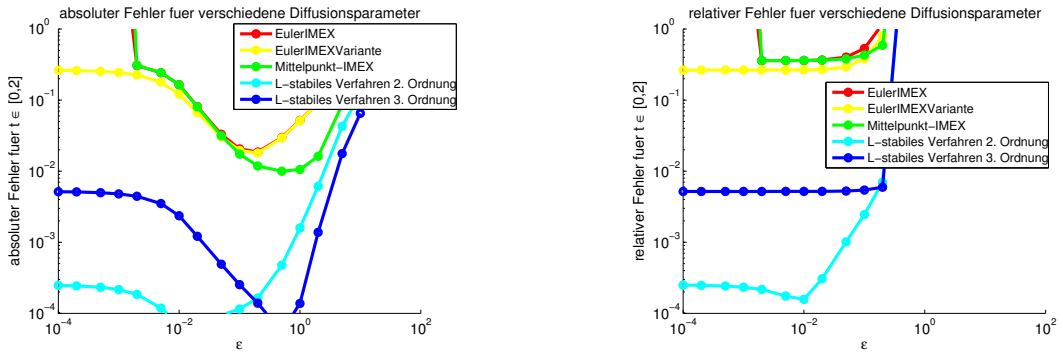

Abbildung 4.3: Absoluter und relativer Fehler für $h = \frac{1}{126}$

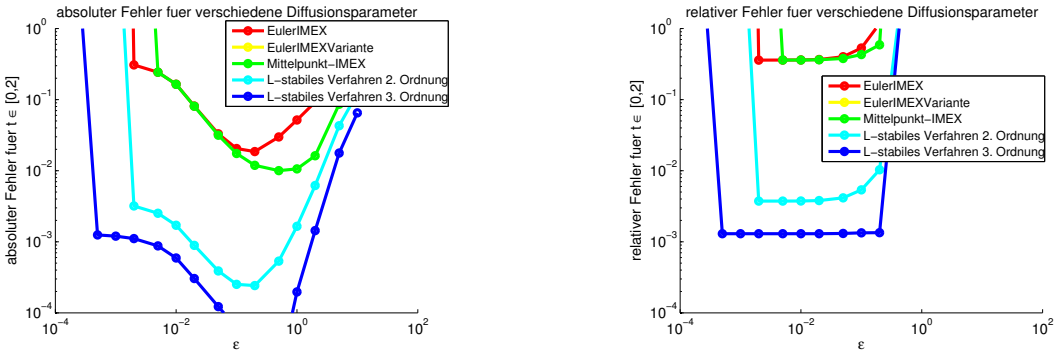

Abbildung 4.4: Absoluter und relativer Fehler für $h = \frac{1}{252}$

4.2.2 Behandlung des nichtlinearen Problems mit IMEX- und Upwind-Verfahren

Für das nichtlineare Problem

$$\partial_t u + u\partial_x u = \varepsilon \partial_{xx}$$

sind die Ergebnisse für $\varepsilon \geq 0.01$ ähnlich gut. Für kleinere Diffusionsparameter werden die Approximationen dagegen beliebig schlecht.
An dieser Stelle greifen wir auf die in Abschnitt 2.2 diskutierten Upwind-Verfahren zurück. Wir testen sie zunächst am Beispiel

$$\partial_t u + u\partial_x u = 0$$

und notieren die maximalen Zeitschrittweiten, für die wir hierbei gute Lösungen erhalten.

Verfahren	$h = \frac{1}{126}$	$h = \frac{1}{252}$
Lax-Friedrichs	$\tau \leq \frac{1}{100}$	$\tau \leq \frac{1}{170}$
Godunov	$\tau \leq \frac{1}{70}$	$\tau \leq \frac{1}{185}$
Enquist-Osher	$\tau \leq \frac{1}{40}$	$\tau \leq \frac{1}{95}$
Lax-Wendroff	$\tau \leq \frac{1}{80}$	$\tau \leq \frac{1}{165}$

Halbieren wir die Ortsschrittweite, so halbiert sich auch die maximal mögliche Zeitschrittweite. Dies entspricht der in Abschnitt 2.2 besprochenen CFL-Bedingung.
Nun kombinieren wir die IMEX-Verfahren mit dem Godunov-Verfahren, indem wir $\partial_x u$ mit einem jeweils passend entgegen der Flussrichtung gewählten einseitigen Dif-

ferenzenquotienten diskretisieren. Mit den so konstruierten IMEX-Upwind-Verfahren können wir

$$\partial_t u + u\partial_x u = \varepsilon \partial_{xx}$$

auch für $\varepsilon < 0.01$ lösen. Abbildung 4.5 zeigt die Lösungen für $h = \frac{1}{126}$, $\tau = \frac{1}{40}$ und $\varepsilon = 0.001$ beziehungsweise $\varepsilon = 0$, die mit dem Mittelpunkt-IMEX-Verfahren mit Upwind berechnet wurden. Man erkennt deutlich die schwache Lösung 2.4 des Problems mit $\varepsilon = 0$.

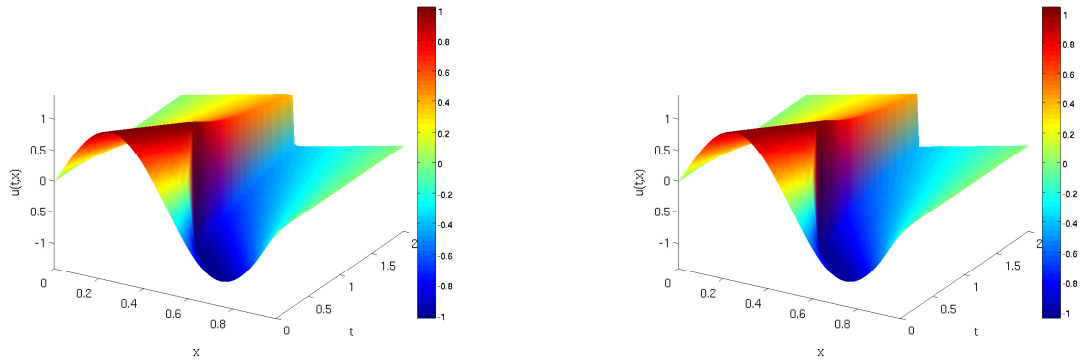

Abbildung 4.5: IMEX-Upwind-Lösungen für $\varepsilon = 0.001$ beziehungsweise $\varepsilon = 0$

Bisher haben wir hauptsächlich Konvektions-Diffusions-Gleichungen behandelt. Mit unserem kombinierten IMEX-Upwind-Verfahren können wir aber auch Konvektions-Reaktions-Diffusions-Gleichungen behandeln. Wir testen dies an dem Beispiel

$$\partial_t u + \partial_x u = 0.02\partial_{xx} u + u^3, \quad u_0(x) = u(0,x) = \sin(2\pi x)$$

für $x \in [0,1]$ mit der periodischen Randbedingung $u(t,0) = u(t,1)$. Der Reaktionsterm $+u^3$ verstärkt die Amplitude der Lösung. Mit $\tau = h = \frac{1}{126}$ erhalten wir die Lösung in Abbildung 4.6.
Ein weiterer interessanter Ansatz ist es, auch $u\partial_x u$ teilweise implizit zu behandeln. Das Verfahren

$$u(t+\tau) = u(t) + \tau \left[-u(t)D_h^0 + \varepsilon D_h^2\right] u(t+\tau)$$

nennen wir semiimplizites Verfahren. Es kann, genau wie die IMEX-Verfahren, auf

$$\partial_t u + u\partial_x u = \varepsilon \partial_{xx}$$

angewandt werden, falls $\varepsilon \geq 0.01$. Dabei erlaubt es recht große Zeitschrittweiten.

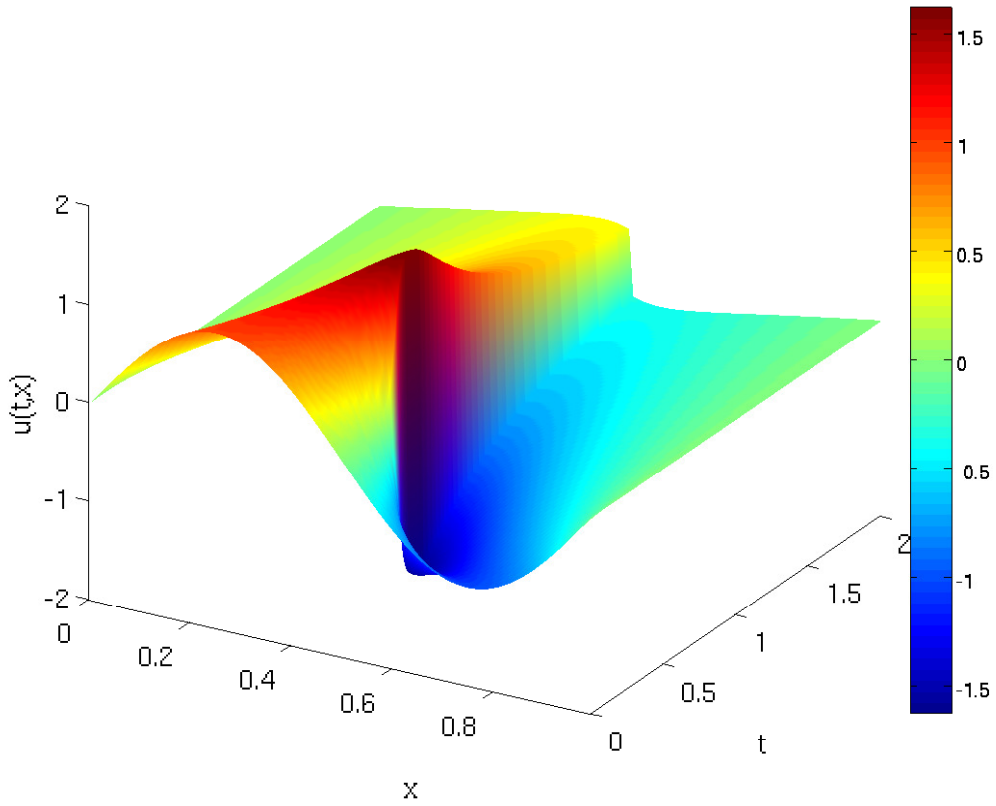

Abbildung 4.6: Lösung einer Konvektions-Reaktions-Diffusions-Gleichung

Abbildung 4.7 zeigt die mit $\varepsilon = 0.01$, $h = \frac{1}{126}$ und $\tau = \frac{1}{40}$ errechnete Lösung.

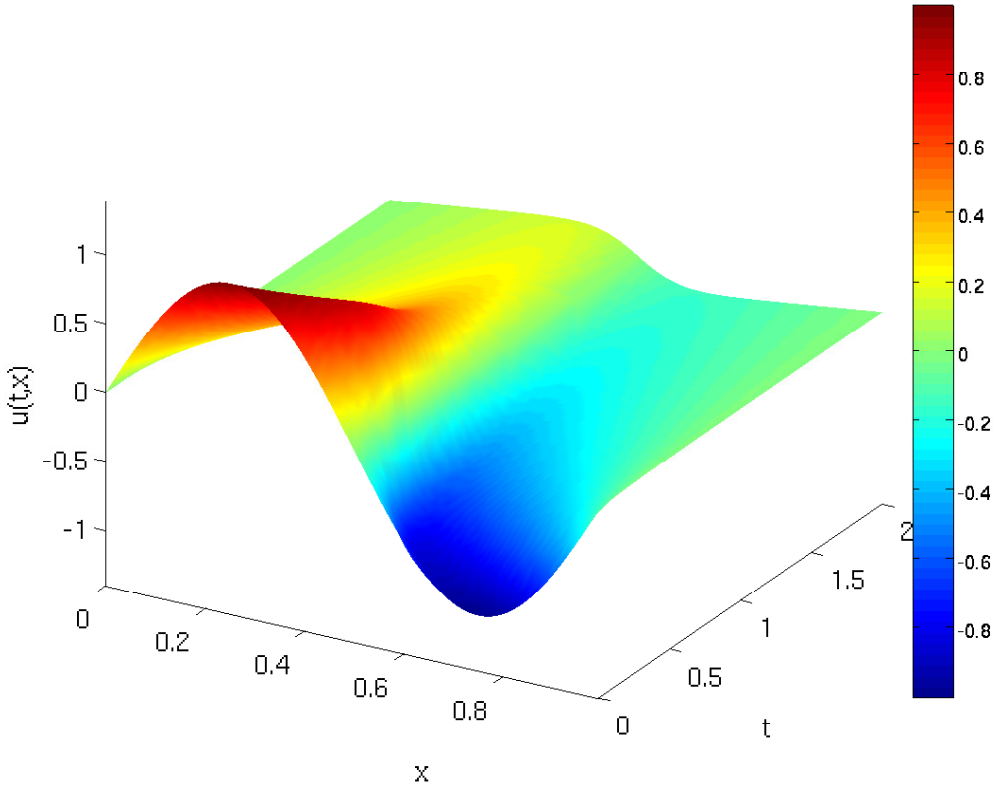

Abbildung 4.7: Lösung des semiimpliziten Verfahrens

5 Zusammenfassung und Ausblick

Diese Arbeit gibt eine Übersicht über die numerische Behandlung von Konvektions-Reaktions-Diffusions-Problemen, wobei der Schwerpunkt auf der Transportgleichung und der Burgers-Gleichung liegt. Dafür wurden zunächst die grundlegenden Methoden besprochen.
Aufbauend auf den Grundlagen konnten IMEX-Verfahren konstruiert werden, die in Kapitel 3 ausführlich analysiert wurden.
Die in hier vorgestellten IMEX-Verfahren eignen sich besonders für die Transportgleichung

$$\partial_t u + \partial_x u = \varepsilon \partial_{xx} u$$

mit mäßiger Diffusion $0.01 < \varepsilon < 1$, da in diesem FAll relativ große Zeitschrittweiten möglich werden.
Mit Hilfe von Upwind-Ortsdiskretisierungen können die IMEX-Verfahren auch auf allgemeinere Konvektions-Reaktions-Diffusions-Gleichungen angewandt werden. Da hierbei künstliche Diffusion eingesetzt wird, entfällt die Bedingung $0.01 < \varepsilon$.
Die Anwendung der IMEX-Verfahren auf weitere partielle Differentialgleichungen in höheren Dimensionen bis hin zur Navier-Stokes-Gleichung ist eine mögliche Fortsetzung dieser Arbeit.

Literaturverzeichnis

[ARS97] U.M. Ascher, S.J. Ruuth, and R.J. Spiteri. Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations. *Applied Numerical Mathematics*, 25(2-3):151–167, 1997. 3.1

[CFL28] R. Courant, K. Friedrichs, and H. Lewy. Über die partiellen Differenzengleichungen der mathematischen Physik. *Mathematische Annalen*, 100(1):32–74, 1928. 2.2

[HB06] M. Hanke-Bourgeois. *Grundlagen der Numerischen Mathematik und des wissenschaftlichen Rechnens*. Vieweg+ Teubner Verlag, 2006. 2.9, 2.3.3, 2.3.4

[HHLK76] A. Harten, JM Hyman, P.D. Lax, and B. Keyfitz. On finite-difference approximations and entropy conditions for shocks. *Communications on Pure and Applied Mathematics*, 29(3):297–322, 1976. 2.2

[HNW93] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving ordinary differential equations: Nonstiff problems*. Springer, 1993. 2.3.3

[LR56] PD Lax and RD Richtmyer. Survey of the stability of linear finite difference equations. *Communications on Pure and Applied Mathematics*, 9(2):267–293, 1956. 3.2.3

[Lub99] G. Lube. Skript zur Vorlesung Theorie und Numerik gewöhnlicher Differentialgleichungen, 1999. 2.3.4

[Lub07] G. Lube. Skript zur Vorlesung Numerik instationärer partieller Differentialgleichungen, 2007. 2.1

[Ole57] O.A. Oleinik. Discontinuous solutions of non-linear differential equations. *Uspekhi Matematicheskikh Nauk*, 12(3):3–73, 1957. 2.5

[QV08] A.M. Quarteroni and A. Valli. *Numerical approximation of partial differential equations*. Springer Verlag, 2008. 2.1

[Run95] C. Runge. Ueber die numerische Auflösung von Differentialgleichungen. *Mathematische Annalen*, 46(2):167–178, 1895. 2.3.1