Kapitel 2

Finite–Differenzen–Verfahren

2.1 Finite Differenzen

Bemerkung 2.1 Idee. Die grundlegende Idee von Finite–Differenzen–Verfahren besteht darin, dass man die Ableitungen in der Differentialgleichung durch geeignete finite Differenzen ersetzt. Dazu wird das Intervall $[0, 1]$ mittels eines äquidistanten Gitters zerlegt:

$$\begin{aligned} x_i &= ih, \quad i = 0, \ldots, N, \quad h = 1/N, \\ \omega_h &= \{x_i \ : \ i = 0, \ldots, N\} \ - \text{Gitter.} \end{aligned}$$

□

Definition 2.2 Gitterfunktion. Ein Vektor $\mathbf{u}_h = (u_0, \ldots, u_N)^T \in \mathbb{R}^{N+1}$, der jedem Gitterpunkt einen Funktionswert zuordnet, heißt Gitterfunktion. Die Restriktion einer Funktion $u \in C([0, 1])$ auf eine Gitterfunktion wird mit $R_h u$ bezeichnet, das heißt

$$R_h u := (u(x_0), u(x_1), \ldots, u(x_N))^T .$$

□

Beispiel 2.3 Sei ein Gitter mit den Punkten $\{0, 0.25, 0.5, 0.75, 1\}$ gegeben. Dann ist die Gitterfunktion zu $u(x) = x^2$

$$R_h u = \left(0, \frac{1}{16}, \frac{1}{4}, \frac{9}{16}, 1\right)^T .$$

Unterschiedliche Funktionen können für ein gegebenes Gitter die gleiche Gitterfunktion haben. Betrachte beispielsweise $u(x) = \sin(4\pi x)$ auf dem obigen Gitter. Die zugehörige Gitterfunktion ist

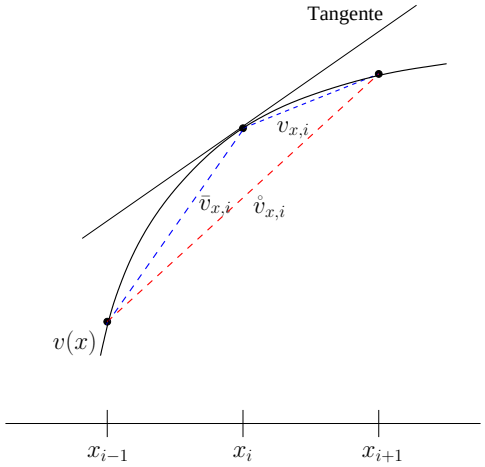
$$R_h u = (0, 0, 0, 0, 0)^T .$$

Dies ist offensichtlich auch die Gitterfunktion von $u(x) = 0$. Das obige Gitter ist zu grob, um die Funktion $u(x) = \sin(4\pi x)$ vernünftig auflösen zu können. □

Definition 2.4 Differenzenoperatoren. Sei $v(x)$ eine genügend glatte Funktion. Bezeichne $v_i = v(x_i)$, wobei x_i Knoten eines Gitters ist. Die folgenden Differenzen-

quotienten (finite Differenzen) nennt man

$$\begin{aligned}
D^+v(x_i) &= v_{x,i} &&= \frac{v_{i+1}-v_i}{h} && \text{– Vorwärtsdifferenz,}\\
D^-v(x_i) &= v_{\overline{x},i} &&= \frac{v_i-v_{i-1}}{h} && \text{– Rückwärtsdifferenz,}\\
D^0v(x_i) &= v_{\mathring{x},i} &&= \frac{v_{i+1}-v_{i-1}}{2h} && \text{– zentrale Differenz,}\\
D^+D^-(v)(x_i) &= v_{\overline{x}x,i} &&= \frac{v_{i+1}-2v_i+v_{i-1}}{h^2} && \text{– zweite Differenz.}
\end{aligned}$$



□

Bemerkung 2.5 Die Formel für $D^+D^-(v)(x_i)$ kontrolliert man durch direktes Nachrechnen. Weiter gilt

$$D^0v(x_i) = \frac{1}{2}(D^+v(x_i) + D^-v(x_i)).$$

□

Definition 2.6 Konsistenz eines Differenzenoperators, diskrete Maximumsnorm. Sei L ein Differentialoperator. Der Differenzenoperator $L_h : \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$ heißt mit L konsistent mit der Ordnung k, wenn

$$\max_{0\le i\le N} |(Lu)(x_i) - (L_hu_h)_i| =: \|(Lu)(x_i) - (L_hu_h)_i\|_{\infty,d} = \mathcal{O}(h^k)$$

gilt. Hierbei ist $\|\cdot\|_{\infty,d}$ die diskrete Maximumsnorm im Raum der Gitterfunktionen.

□

Die Konsistenz ist ein Maß für die Approximationsgüte von L_h.

Beispiel 2.7 Aus der Taylor[1]–Entwicklung für $v(x)$ an der Stelle x_i ergibt sich

$$\begin{aligned}
D^+v(x_i) &= v'(x_i) + \mathcal{O}(h),\\
D^-v(x_i) &= v'(x_i) + \mathcal{O}(h),\\
D^0v(x_i) &= v'(x_i) + \mathcal{O}(h^2),\\
D^+D^-(v)(x_i) &= v''(x_i) + \mathcal{O}(h^2).
\end{aligned}$$

Die Differenzenoperatoren $D^+v(x_i), D^-v(x_i), D^0v(x_i)$ sind damit konsistent zu $L = \frac{d}{dx}$ mit der Ordnung 1,1 beziehungsweise 2. Der Operator $D^+D^-(v)(x_i)$ ist von zweiter Ordnung konsistent mit $L = \frac{d^2}{dx^2}$. □

[1]Brook Taylor (1685 – 1731)

Beispiel 2.8 Betrachtet wird der Differentialoperator

$$Lu = \frac{d}{dx}\left(k(x)\frac{du}{dx}\right),$$

wobei $k(x)$ stetig differenzierbar ist. Wir definieren den Differenzenoperator L_h wie folgt

$$\begin{aligned}(L_h u_h)_i &= D^+(aD^-u(x_i)) = \frac{1}{h}\Big(a(x_{i+1})D^-u(x_{i+1}) - a(x_i)D^-u(x_i)\Big)\\ &= \frac{1}{h}\left(a_{i+1}\frac{u_{i+1}-u_i}{h} - a_i\frac{u_i-u_{i-1}}{h}\right),\end{aligned}$$

wobei a eine Gitterfunktion ist, die geeignet gewählt werden soll. Es folgt mit Produktregel beziehungsweise mit Taylor–Entwicklung

$$\begin{aligned}(Lu)_i &= k'(x_i)(u')_i + k(x_i)(u'')_i,\\ (L_h u_h)_i &= \frac{a_{i+1}-a_i}{h}(u')_i + \frac{a_{i+1}+a_i}{2}(u'')_i + \frac{h(a_{i+1}-a_i)}{6}(u''')_i + \mathcal{O}(h^2).\end{aligned}$$

Für die Differenz ergibt sich

$$\begin{aligned}(Lu)_i - (L_h u_h)_i &= \left(k'(x_i) - \frac{a_{i+1}-a_i}{h}\right)(u')_i + \left(k(x_i) - \frac{a_{i+1}+a_i}{2}\right)(u'')_i\\ &\quad -\frac{h(a_{i+1}-a_i)}{6}(u''')_i + \mathcal{O}(h^2).\end{aligned}$$

Damit L_h von zweiter Ordnung mit L konsistent ist, müssen somit gelten

$$\frac{a_{i+1}-a_i}{h} = k'(x_i) + \mathcal{O}(h^2), \quad \frac{a_{i+1}+a_i}{2} = k(x_i) + \mathcal{O}(h^2).$$

Aus der ersten Forderung folgt $a_{i+1} - a_i = \mathcal{O}(h)$, womit außerdem folgt, dass der dritte Summand in der Fehlergleichung von Ordnung $\mathcal{O}(h^2)$ wird. Mögliche Varianten sind (*Übungsaufgaben?*)

$$a_i = \frac{k_i + k_{i-1}}{2}, \quad a_i = k\left(x_i - \frac{h}{2}\right), \quad a_i = (k_i k_{i-1})^{1/2}.$$

Man beachte, die ,,natürliche" Wahl $a_i = k_i$ garantiert nur Konsistenz von erster Ordnung, siehe die Taylorentwicklung für $D^+v(x_i)$. □

2.2 Klassische Konvergenztheorie für zentrale Differenzen

Bemerkung 2.9 In diesem Abschnitt wird das 2–Punkt–Randwertproblem

$$Lu := -u'' + b(x)u' + c(x)u = f(x), \quad \text{für } x \in (0,1), \quad u(0) = u(1) = 0, \tag{2.1}$$

betrachtet, das heißt $\varepsilon = 1$, um die klassische Lösungstheorie darzustellen. Es wird angenommen, dass die Parameterfunktionen b, c, f hinreichend glatt sind und dass $c(x) \geq 0$ für alle $x \in [0,1]$ gilt. □

Definition 2.10 Zentrales Differenzenschema. Das zentrale Differenzenschema für (2.1) besitzt die Gestalt

$$\begin{aligned}-D^+D^-u_i + b_i D^0 u_i + c_i u_i &= f_i, \quad \text{für } i = 1, \ldots, N-1,\\ u_0 = u_N &= 0.\end{aligned} \tag{2.2}$$

□

Bemerkung 2.11

- Das zentrale Differenzenschema führt auf ein tridiagonales System linearer Gleichungen

$$r_i u_{i-1} + s_i u_i + t_i u_{i+1} = f_i, \quad i = 1, \ldots, N-1, \quad u_0 = u_N = 0,$$

mit

$$r_i = -\frac{1}{h^2} - \frac{1}{2h} b_i, \quad s_i = c_i + \frac{2}{h^2}, \quad t_i = -\frac{1}{h^2} + \frac{1}{2h} b_i.$$

- Die folgenden Fragen müssen beantwortet werden:
  - Welche Eigenschaften besitzt das diskrete Problem (2.2)?
  - Was kann man über den Fehler $|u(x_i) - u_i|$ aussagen?

  Dazu werden die Konzepte von Konsistenz und Stabilität verwendet.

□

Definition 2.12 Konsistenz eines Differenzenschemas und Konsistenzordnung. Betrachte ein Differenzenschema der Gestalt $L_h u_h = R_h(Lu)$. Dabei seien die Randbedingungen derart integriert, dass die erste und letzte Zeile von L_h identisch zur ersten und letzten Zeile der Einheitsmatrix sind und $R_h(Lu)_0 = u_0$ sowie $R_h(Lu)_N = u_N$ gelten. Das Schema wir konsistent von der Ordnung k in der diskreten Maximumsnorm genannt, falls

$$\|L_h R_h u - R_h(Lu)\|_{\infty,d} \leq c h^k$$

ist, wobei die positiven Konstanten c und k unabhängig von h sind. □

Lemma 2.13 Konsistenzordnung des zentralen Differenzenschemas. *Unter der Annahme, dass $u \in C^4([0,1])$ gilt, besitzt das zentrale Differenzenschema (2.2) die Konsistenzordnung* 2.

Beweis: Mit Taylor–Entwicklung, Übungsaufgabe. ■

Definition 2.14 Stabilität eines Differenzenschemas. Ein Differenzenschema $L_h u_h = f_h$ wird stabil in der diskreten Maximumsnorm genannt, wenn es eine Stabilitätskonstante c_S unabhängig von h gibt, so dass

$$\|u_h\|_{\infty,d} \leq c_S \|L_h u_h\|_{\infty,d}$$

für alle Gitterfunktionen u_h gilt. □

Definition 2.15 Konvergenz eines Differenzenschemas und Konvergenzordnung. Ein Differenzenschema für (2.1) ist konvergent von Ordnung k in der diskreten Maximumsnorm, falls es positive Konstanten c und k unabhängig von h gibt, so dass

$$\|u_h - R_h u\|_{\infty,d} \leq c h^k.$$

□

Satz 2.16 Konsistenz + Stabilität $\Longrightarrow$ Konvergenz. *Ein konsistentes und stabiles Differenzenschema ist konvergent. Konsistenz– und Konvergenzordnung sind gleich.*

Beweis: Es gilt

$$\begin{aligned}\|u_h - R_h u\|_{\infty,d} &\overset{\text{Stab.}}{\leq} c_S \|L_h (u_h - R_h u)\|_{\infty,d} \overset{\text{lin.}}{=} c_S \|L_h u_h - L_h R_h u\|_{\infty,d} \\ &= c_S \|f_h - L_h R_h u\|_{\infty,d} = c_S \|R_h f - L_h R_h u\|_{\infty,d} \\ &= c_S \|R_h L u - L_h R_h u\|_{\infty,d} \overset{\text{Kons.}}{\leq} K h^k,\end{aligned}$$

wobei die Konstante K das Produkt aus den Konstanten der Stabilitäts– und Konsistenzbedingung ist. ■

Bemerkung 2.17 Man muss also Konsistenz und Stabilität untersuchen.

- Konsistenzuntersuchungen basieren üblicherweise auf Taylor–Entwicklungen und sie laufen in der Regel nach dem gleichen Muster ab.
- Stabilitätsuntersuchungen werden nicht an Funktionen sondern mit Matrizen und Funktionen durchgeführt, siehe Definition 2.14. Sie sind nicht so einfach und es werden einige neue Begriffe benötigt, die im folgenden bereitgestellt werden.

□

Definition 2.18 Natürliche Ordnung von Vektoren und Matrizen, invers–monotone Matrix. Seien $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Dann schreibt man $\mathbf{x} \leq \mathbf{y}$ genau dann, wenn $x_i \leq y_i$ für alle $i = 1, \ldots, n$ gilt. Die Notation $\mathbf{x} \geq \mathbf{1}$ bedeutet $x_i \geq 1$ für alle $i = 1, \ldots, n$. Analog bedeutet für eine Matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ die Bezeichnung $A \geq 0$, dass $a_{ij} \geq 0$ für alle $i, j = 1, \ldots, n$ gilt.

Eine Matrix A, für welche A^{-1} existiert mit $A^{-1} \geq 0$, wird invers–monotone Matrix genannt. □

Das nächste Lemma gibt ein diskretes Analogon zum Vergleichsprinzip von Folgerung 1.27.

Lemma 2.19 Diskretes Vergleichsprinzip. *Sei $A \in \mathbb{R}^{n \times n}$ invers–monoton. Gilt $A\mathbf{v} \leq A\mathbf{w}$ für $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, dann folgt $\mathbf{v} \leq \mathbf{w}$.*

Beweis: Nach Voraussetzung gilt

$$A(\mathbf{v} - \mathbf{w}) := \mathbf{b} \leq \mathbf{0}.$$

Multiplikation mit A^{-1} ergibt

$$\mathbf{v} - \mathbf{w} = A^{-1}\mathbf{b} \leq \mathbf{0}.$$

Die letzte Ungleichung folgt daraus, dass A invers–monoton ist. Nichtnegative Matrixeinträge werden mit nichtpositiven Vektoreinträgen von $\mathbf{b}$ multipliziert. Man erhält als Ergebnis einen Vektor mit nichtpositiven Komponenten. ■

Eine wichtige Teilklasse der Klasse der invers–monotonen Matrizen ist die folgende.

Definition 2.20 M–Matrix. Eine Matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ wird M–Matrix genannt, falls:

1. $a_{ij} \leq 0$ für $i \neq j$,
2. A^{-1} existiert mit $A^{-1} \geq 0$.

□

Lemma 2.21 *Sei $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ eine M–Matrix. Dann gilt $a_{ii} > 0$, $i = 1, \ldots, n$.*

Beweis: Übungsaufgabe. ■

Die zweite Bedingung der Definition ist im allgemeinen schwer zu überprüfen. Es gibt aber auch handlichere Charakterisierungen von M–Matrizen.

Satz 2.22 M–Matrix–Kriterium. *Sei $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ mit $a_{ij} \leq 0$ für $i \neq j$. Dann ist A eine M–Matrix genau dann, wenn ein Vektor $\mathbf{e} \in \mathbb{R}^n$, $\mathbf{e} > 0$ existiert, so dass $A\mathbf{e} > \mathbf{0}$. Dann gilt für die Zeilensummennorm*

$$\left\|A^{-1}\right\|_\infty \leq \frac{\|\mathbf{e}\|_{\infty,d}}{\min_k (A\mathbf{e})_k}. \tag{2.3}$$

Der Vektor $\mathbf{e}$ wird majorisierendes Element genannt.

Beweis: Siehe Literatur [Boh81, AK90]. ■

Bemerkung 2.23 Zum M–Matrix–Kriterium.

- Das folgende Rezept ist oft erfolgreich, um ein majorisierendes Element zu konstruieren:
  - Finde eine Funktion $e(x) > 0$, so dass $(Le)(x) > 0$ für $x \in (0, 1)$. Das ist ein majorisierendes Element des Differentialoperators L.
  - Schränke $e(x)$ auf die Gitterfunktion e_h ein.

  Falls der erste Schritt in dieser Herangehensweise möglich ist, und die Diskretisierung konsistent ist, dann funktioniert die Herangehensweise im allgemeinen, zumindest für hinreichend kleine Gitterweite.
- Mit (2.3) kann man die Konstante c_S in der Stabilitätsdefinition abschätzen

  $$\|u_h\|_{\infty,d} = \left\|A^{-1}(f_h)\right\|_{\infty,d} \leq \left\|A^{-1}\right\|_\infty \|f_h\|_{\infty,d} = \left\|A^{-1}\right\|_\infty \|L_h u_h\|_{\infty,d},$$

  also gilt für diese Konstante

  $$c_S \leq \frac{\|\mathbf{e}\|_{\infty,d}}{\min_k (A\mathbf{e})_k}.$$

- Für Dirichletrandbedingungen eliminiert man die Variablen u_0 und u_N bevor man Satz 2.22 anwendet.

□

Beispiel 2.24 Zum M–Matrix–Kriterium. Betrachte (2.1) mit $b(x) \equiv 0$

$$Lu(x) = -u''(x) + c(x)u(x), \quad c(x) \geq 0 \text{ in } [0, 1].$$

Wähle $e(x) := \frac{1}{2}x(1 - x)$. Dann folgt

$$Le(x) = 1 + c(x)e(x) \geq 1.$$

Nun setzt man $e_h := R_h e$. Damit ergibt sich

$$(L_h e_h)_i = -D^+D^- e_{h,i} + c_i e_{h,i} = 1 + c_i e_{h,i} \geq 1,$$

weil der zweite Differenzenquotient die zweiten Ableitungen von quadratischen Funktionen in inneren Gitterpunkten exakt diskretisiert, siehe Beispiel 2.7. Das heißt

$$L_h e_h \geq (1, \ldots, 1)^T \quad \Longleftrightarrow \quad A\mathbf{e} \geq \mathbf{1}.$$

Für die Abschätzung der Stabilitätskonstanten erhält man

$$c_S \leq \frac{\|\mathbf{e}\|_{\infty,d}}{\min_k (A\mathbf{e})_k} \leq \frac{e_h(1/2)}{1} = \frac{1/8}{1} = \frac{1}{8}.$$

Dieses Beispiel zeigt, dass im Fall $b(x) \equiv 0$ die M–Matrix Eigenschaft ohne Einschränkungen an das Gitter gilt. □

Lemma 2.25 Stabilität des zentralen Differenzenschemas für hinreichend feines Gitter. *Für hinreichend kleine Gitterweite h ist das zentrale Differenzenschema (2.2) für das Randwertproblem (2.1) in der diskreten Maximumsnorm stabil. Die Koeffizientenmatrix ist eine M–Matrix.*

Beweis: Sei $e(x)$ die Lösung des Randwertproblems

$$-w'' + b(x)w' = 1, \quad w(0) = w(1) = 0.$$

Nach dem Maximumprinzip, Lemma 1.25, gilt $e(x) \geq 0$ für $x \in (0,1)$. Da $c(x) \equiv 0$ ist, folgt nach Folgerung 1.32, dass das obige Problem eindeutig lösbar ist und insbesondere $e \in C([0,1])$ gilt. Damit ist $e(x)$ beschränkt. Für innere Gitterpunkte gilt wegen $c(x) \geq 0$

$$\begin{aligned}
(L_h e_h)_i &= (R_h L e)_i + (L_h e_h - R_h L e)_i \\
&= \left(R_h \left(1 + c(x)e(x)\right)\right)_i + \left(-D^+D^- e_h + b_i D^0 e_h + c_i e_h - 1 - c_i e_h\right)_i \\
&\geq 1 + \left(-D^+D^- e_h + b_i D^0 e_h - 1\right)_i \\
&= \left(-D^+D^- e_h + b_i D^0 e_h\right)_i .
\end{aligned}$$

Da e_h die zu $e(x)$ gehörende Gitterfunktion ist, approximiert der Ausdruck in der letzten Zeile für hinreichend feines h den Term $-e''(x_i) + b(x_i)e'(x_i)(= 1)$ hinreichend gut, siehe Beispiel 2.7. Insbesondere gibt es ein $H > 0$, so dass für alle $h \in (0, H]$ gilt

$$(L_h e_h)_i \geq \frac{1}{2}.$$

Das M–Matrix–Kriterium beweist nun die Aussage des Satzes. ■

Folgerung 2.26 Konvergenz zweiter Ordnung des zentralen Differenzenschemas. *Unter der Annahme, dass $u \in C^4([0,1])$ gilt, konvergiert das zentrale Differenzenschema (2.2) von zweiter Ordnung.*

Beweis: Das folgt aus Satz 2.16, indem man Lemmata 2.13 und 2.25 kombiniert. ■

Beispiel 2.27 Betrachte das 2–Punkt–Randwertproblem

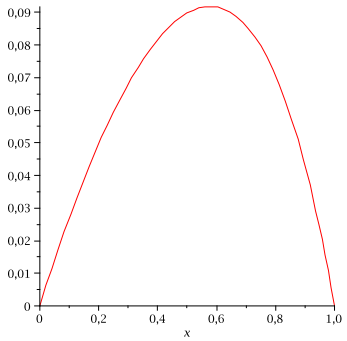
$$-u''(x) + 2u'(x) + 3u(x) = 1 \quad \text{in } (0,1), \quad u(0) = u(1) = 0.$$

Die Lösung dieses Problems lautet

$$u(x) = \frac{1}{3}\left(1 + \frac{1 - e^{-1}}{e^{-1} - e^3} e^{3x} + \frac{e^3 - 1}{e^{-1} - e^3} e^{-x}\right).$$

Man erhält die folgenden Fehler für unterschiedliche Gitterweiten:

Intervalle N	$\|u - u_h\|_{\infty,d}$
4	4.2388e-4
8	9.8811e-5
16	2.4529e-5
32	6.1537e-6
64	1.5368e-6
128	3.8440e-7
256	9.6093e-8
512	2.4023e-8
1024	6.0058e-9



Man erkennt, dass für hinreichend feine Gitter, sich der Fehler bei einer Halbierung der Gitterweite um den Faktor Vier verringert. Das ist zweite Ordnung Konvergenz. □

2.3 Upwind–Verfahren

Bemerkung 2.28 Von nun an werden Finite–Differenzen–Verfahren für das singulär gestörte Randwertproblem

$$Lu := -\varepsilon u'' + b(x)u' + c(x)u = f(x), \quad \text{für } x \in (0,1), \tag{2.4}$$

mit den Randbedingungen

$$u(0) = u(1) = 0, \tag{2.5}$$

unter den Voraussetzungen

$$\begin{aligned} \varepsilon &> 0, \\ b(x) &> 0 \quad \text{für alle } x \in [0,1], \\ c(x) &\geq 0 \quad \text{für alle } x \in [0,1], \end{aligned}$$

mit hinreichend glatten Funktion $b(x), c(x)$ und $f(x)$. Das Problem nennt man dann singulär gestört, wenn $\varepsilon \ll |b(x)|$ ist. Der Parameter ε wird singulärer Störungsparameter genannt. Für die Konvektion ist nur wichtig, dass $b(x) \neq 0$ für alle $x \in [0,1]$ gilt. Ist $b(x) < 0$ in $[0,1]$, gelangt man mit der Variablentransformation $x \mapsto 1 - x$ auf ein Problem mit den obigen Voraussetzungen.

Ist ε klein, so besitzt die Lösung von (2.4), (2.5) im allgemeinen eine Randgrenzschicht bei $x = 1$, vergleiche Beispiel 1.7. Diese Grenzschicht beeinflusst sowohl Stabilität als auch Konsistenz eines numerischen Verfahrens. Sind die Randbedingungen so gewählt, dass keine Grenzschicht auftritt, dann verbessert sich der Konsistenzfehler, aber die Stabilität des Verfahrens kann immer noch ein Problem sein. □

Beispiel 2.29 Zentrales Differenzenschema angewandt auf ein vereinfachtes singulär gestörtes Problem. Betrachte das Problem

$$-\varepsilon u'' + u' = 0 \text{ auf } (0,1), \quad u(0) = 0,\ u(1) = 1.$$

Die Lösung dieses Problems lautet

$$u(x) = \frac{e^{-(1-x)/\varepsilon} - e^{-1/\varepsilon}}{1 - e^{-1/\varepsilon}}.$$

Die Transformation $u(x) := x + v(x)$ würde auf ein Problem mit homogenen Randbedingungen führen. Man kann das Differenzenverfahren aber direkt auf das Problem mit inhomogenen Randbedingungen anwenden. Das wird in der Praxis auch so gemacht. Das diskrete Problem ist

$$-\varepsilon D^+ D^- u_i + D^0 u_i = 0, \quad u_0 = 0,\ u_N = 1.$$

Die Lösung dieses Gleichungssystems ist *Übungsaufgabe*

$$u_i = \frac{r^i - 1}{r^N - 1} \quad \text{mit} \quad r = \frac{2\varepsilon + h}{2\varepsilon - h}.$$

Es gilt insbesondere $|r| > 1$.

Ist $h \gg 2\varepsilon$, dann gilt $r \approx -1$ und es folgt

$$u_i \approx \frac{(-1)^i - 1}{(-1)^N - 1}.$$

Ist N gerade, so wird durch eine sehr kleine positive Zahl dividiert. Für gerade i, ist der Zähler ebenfalls klein und positiv, für ungerade i ist der Zähler negativ. Die diskrete Lösung oszilliert sehr stark, siehe Abbildung 2.1.

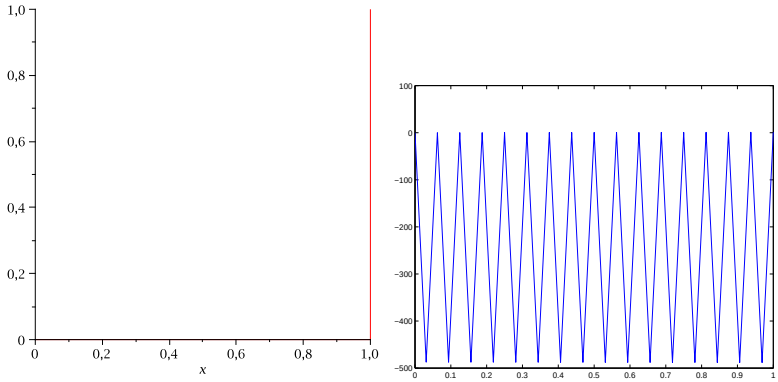


Abbildung 2.1: Lösung und diskrete Lösung mit dem zentralen Differenzenschema für $\varepsilon = 10^{-6}$ und $h = 1/32$.

Ist jedoch $h < 2\varepsilon$, dann erhält man mit dem zentralen Differenzenschema eine sinnvolle Approximation der Lösung. In Anwendungen ist jedoch oft $\varepsilon \leq 10^{-6}$, so dass man sehr feine Gitter braucht, um das zentrale Differenzenschema anwenden zu können. In einer Dimension ist das heute oft möglich, für Probleme in zwei oder drei Dimensionen jedoch nicht. □

Bemerkung 2.30 Zentrales Differenzenschema angewandt auf das allgemeine singulär gestörte Problem. Betrachte nun das singulär gestörte Problem (2.4), (2.5) und schreibe das Differenzenschema in der Form aus Bemerkung 2.11

$$r_i u_{i-1} + s_i u_i + t_i u_{i+1} = f_i, \quad i = 1, \ldots, N-1, \quad u_0 = u_N = 0,$$

mit

$$r_i = -\frac{\varepsilon}{h^2} - \frac{1}{2h} b_i, \quad s_i = c_i + \frac{2\varepsilon}{h^2}, \quad t_i = -\frac{\varepsilon}{h^2} + \frac{1}{2h} b_i, \quad b_i > 0.$$

Damit erhält man eine M–Matrix und somit Stabilität, falls man

$$t_i \leq 0 \quad \Longrightarrow \quad h \leq h_0(\varepsilon) = \frac{2\varepsilon}{\|b\|_\infty}$$

voraussetzt. Dies verallgemeinert die Beobachtung aus Beispiel 2.29. Man beachte, dass $h_0(\varepsilon) \to 0$ für $\varepsilon \to 0$ gilt. □

Bemerkung 2.31 Motivation für Upwind–Verfahren. Eine andere heuristische Erklärung für das Versagen des zentralen Differenzenverfahrens für $\varepsilon \ll h$ ist wie folgt. In diesem Fall besitzt das Verfahren für Beispiel 2.29 im wesentlichen die Gestalt

$$D^0 u_i = 0, \quad \Longleftrightarrow \quad \frac{u_{i+1} - u_{i-1}}{2h} = 0.$$

Daraus folgt insbesondere $u_{N-2} \approx u_N = 1$. Das ist eine schlechte Approximation des exakten Wertes $u(x_{N-2}) \approx 0$.

Diese Beobachtung sagt uns, dass es zur Approximation von $u'(x_{N-1})$ besser ist, den Wert an der Stelle u_N nicht zu verwenden. Der einfachste Kandidat, der diese Bedingung erfüllt, ist

$$u'(x_i) \approx \frac{u_i - u_{i-1}}{h}.$$

Verfolgt man das Ziel, die Matrixeinträge des diskreten Problems vom zentralen Differenzenschema so zu modifizieren, dass man eine M–Matrix erhält, so kann man ebenfalls diese Approximation motivieren. □

Definition 2.32 Einfaches Upwind–Verfahren. Das einfache Upwind–Verfahren für das singulär gestörte Randwertproblem (2.4), (2.5) besitzt die Gestalt

$$\begin{aligned} -\varepsilon D^+D^-u_i + b_i D^{\mathcal{N}} u_i + c_i u_i &= f_i, \quad \text{für } i = 1, \ldots, N-1, \\ u_0 = u_N &= 0 \end{aligned} \tag{2.6}$$

mit

$$D^{\mathcal{N}} := \begin{cases} D^+ & \text{für } b < 0, \\ D^- & \text{für } b > 0. \end{cases}$$

□

Bemerkung 2.33 Zum einfachen Upwind–Verfahren.

- Upwind, deutsch stromaufwärts, bedeutet, dass die finite Differenzenapproximation des konvektiven Termes mit Werten aus der Stromaufwärts–Richtung genommen. Bei konvektions–dominanten Problem erfolgt der Informationstransport in Richtung der Konvektion. Aus der Stromaufwärts–Richtung kommt daher die Information.
- Mit dem einfachen Upwind–Verfahren erhält man eine viel bessere Approximation der Lösung aus Beispiel 2.29, siehe Abbildung 2.2.

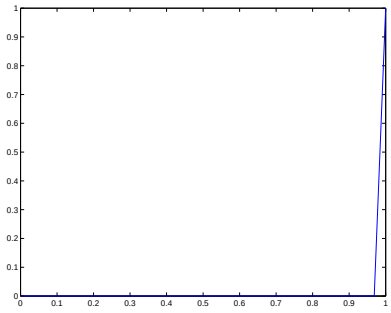


Abbildung 2.2: Diskrete Lösung mit dem einfachen Upwind–Verfahren für $\varepsilon = 10^{-6}$ und $h = 1/32$.

- Beim einfachen Upwind–Verfahren wird die Approximation zweiter Ordnung D^0 durch eine Approximation erster Ordnung, D^+ oder D^- ersetzt. Das wird sich natürlich in der Genauigkeit des Verfahrens bemerkbar machen.
- Sei L_h die Matrix des Upwind–Verfahrens, nachdem die Randwerte u_0 und u_N eliminiert wurden. In der Form von Bemerkung 2.11 besitzen die Matrixeinträge die Gestalt

  $$r_i = -\frac{\varepsilon}{h^2} - \frac{1}{h}\max\{0, b_i\}, \quad s_i = c_i + \frac{2\varepsilon}{h^2} + \frac{1}{h}\left|b_i\right|, \quad t_i = -\frac{\varepsilon}{h^2} + \frac{1}{h}\min\{0, b_i\}.$$

  Nun sind die Nichtdiagonaleinträge nichtpositiv, unabhängig von der Größe von ε und h.

□

Satz 2.34 Stabilität des einfachen Upwind–Verfahrens. *Unter den in Bemerkung 2.28 gemachten Voraussetzungen ist die Koeffizientenmatrix L_h des einfachen Upwind–Verfahrens (2.6) eine M–Matrix. Das einfache Upwind–Verfahren ist gleichmäßig stabil bezüglich des Parameters ε, das heißt es gilt*

$$\|u_h\|_{\infty,d} \leq c_S \|L_h u_h\|_{\infty,d}$$

mit einer von ε und h unabhängigen Stabilitätskonstanten $c_S > 0$.

Beweis: Betrachte nur den Fall $b(x) \geq \beta > 0$. Man konstruiert einen geeignetes majorisierendes Element. Wähle $e(x) = x$, dann gilt

$$Le(x) = -\varepsilon e''(x) + b(x)e'(x) + c(x)e(x) = b(x) + xc(x) \geq \beta.$$

Für das einfache Upwind–Verfahren und die Gitterfunktion e_h erhält man

$$\begin{aligned}
(L_h e_h)_i &= r_i x_{i-1} + s_i x_i + t_i x_{i+1} \\
&= \left(-\frac{\varepsilon}{h^2} - \frac{1}{h} b_i\right)(x_i - h) + \left(c_i + \frac{2\varepsilon}{h^2} + \frac{1}{h} b_i\right) x_i - \frac{\varepsilon}{h^2}(x_i + h) \\
&= \left(-\frac{\varepsilon}{h^2} - \frac{1}{h} b_i + c_i + \frac{2\varepsilon}{h^2} + \frac{1}{h} b_i - \frac{\varepsilon}{h^2}\right) x_i + \left(\frac{\varepsilon}{h^2} + \frac{1}{h} b_i - \frac{\varepsilon}{h^2}\right) h \\
&= c_i x_i + b_i \geq \beta.
\end{aligned}$$

Nach Satz 2.22 ist L_h eine M–Matrix. Mit der Abschätzung der Stabilitätskonstanten aus Bemerkung 2.23 erhält man schließlich

$$c_S \leq \frac{\|e_h\|_{\infty,d}}{\min_k (L_h e_h)_k} = \frac{1}{\beta}.$$

■

Zur Konsistenzuntersuchung benötigt man eine relativ genaue Abschätzung der Ableitungen der Lösung des stetigen Problems.

Lemma 2.35 *Seien $b(x) \geq \beta > 0$ und $b(x), c(x), f(x)$ hinreichend glatt. Dann erfüllt die Lösung $u(x)$ von (2.4), (2.5)*

$$\left|u^{(i)}(x)\right| \leq C\left[1 + \varepsilon^{-i} \exp\left(-\beta \frac{1-x}{\varepsilon}\right)\right], \quad i = 1, 2, \ldots, q,$$

für $x \in [0, 1]$. Die maximale Ordnung q hängt von der Glätte der Daten ab.

Beweis: Der Beweis erfolgte in [KT78], man findet ihn in [RST08, S. 21]. ■

Satz 2.36 Konsistenz des einfachen Upwind–Verfahrens. *Unter den in Bemerkung 2.28 gemachten Voraussetzungen mit $b(x) \geq \beta > 0$ existiert eine positive Konstante β^*, die nur von β abhängt, so dass der Fehler des einfachen Upwind–Verfahrens (2.6) in den inneren Gitterpunkten $\{x_i \ : \ i = 1, \ldots, N-1\}$*

$$|u(x_i) - u_i| \leq \begin{cases} Ch[1 + \varepsilon^{-1} \exp(-\beta^*(1 - x_i)/\varepsilon)] & \text{falls } h < \varepsilon, \\ Ch + C \exp(-\beta^*(1 - x_{i+1})/\varepsilon) & \text{falls } h \geq \varepsilon \end{cases}$$

erfüllt.

Beweis: Der Beweis folgt [KT78]. Hier wird nur der interessante Fall $h \geq \varepsilon$ betrachtet und auch für diesen wird der Beweis nicht ganz vollständig angegeben. Den vollständigen Beweis findet man in [RST08, S. 49f.].

Im Fall $h \geq \varepsilon$ zerlegt man die Lösung von (2.4), (2.5) in

$$u(x) = -u_0(1) \exp\left(-\frac{b(1)(1-x)}{\varepsilon}\right) + z(x) =: v(x) + z(x),$$

wobei $u_0(x)$ die reduzierte Lösung ist. Analog zu Lemma 2.35 findet man

$$\left|z^{(k)}(x)\right| \leq C\left[1 + \varepsilon^{1-k} \exp\left(-\frac{b(1)(1-x)}{\varepsilon}\right)\right], \quad k = 1, 2, 3.$$

Es gilt

$$L_h u_h = f_h = R_h(f) = R_h(Lu) = R_h\left(L(v+z)\right) = R_h(Lv) + R_h(Lz).$$

Damit hat man eine Zerlegung $u_h = v_h + z_h$, wobei die Gitterfunktionen durch

$$L_h v_h = R_h(Lv) \quad \text{und} \quad L_h z_h = R_h(Lz)$$

definiert sind, wobei v_h und z_h mit $v(x)$ beziehungsweise $z(x)$ in x_0 und x_N übereinstimmen. Mit Dreiecksungleichung gilt

$$|u(x_i) - u_i| = |v(x_i) + z(x_i) - (v_i + z_i)| \leq |v(x_i) - v_i| + |z(x_i) - z_i| \,.$$

Betrachte nun den Konsistenzfehler für $z(x)$. Dazu wird die Taylor–Entwicklung von $z(x)$ im Punkt x_i verwendet und man erhält im ersten Schritt *Übungsaufgabe*

$$|\tau_i| := |L_h z(x_i) - f(x_i) - v(x_i)| \leq C \int_{x_{i-1}}^{x_{i+1}} \left(\varepsilon \left| z^{(3)}(t) \right| + \left| z^{(2)}(t) \right| \right) \, dt.$$

Die rechte Seite kommt durch die Restglieder. Nun verwendet man die obige Abschätzung für die Ableitungen von $z(x)$

$$\begin{aligned}
|\tau_i| &\leq C \int_{x_{i-1}}^{x_{i+1}} \left(\varepsilon + \varepsilon^{-1} \exp\left(-b(1) \frac{1-t}{\varepsilon} \right) + 1 + \varepsilon^{-1} \exp\left(-b(1) \frac{1-t}{\varepsilon} \right) \right) \, dt \\
&\leq \int_{x_{i-1}}^{x_{i+1}} (\varepsilon + 1) \, dt + \varepsilon^{-1} \int_{x_{i-1}}^{x_{i+1}} \left(\exp\left(-\beta \frac{1-t}{\varepsilon} \right) + \exp\left(-\beta \frac{1-t}{\varepsilon} \right) \right) \, dt \\
&\leq Ch + C\varepsilon^{-1} \int_{x_{i-1}}^{x_{i+1}} \exp\left(-\beta \frac{1-t}{\varepsilon} \right) \, dt \\
&= Ch + C\varepsilon^{-1} \left(\frac{\varepsilon}{\beta} \exp\left(-\beta \frac{1-t}{\varepsilon} \right) \Big|_{x_i - h}^{x_i + h} \right) \\
&= Ch + C \left[\exp\left(-\beta \frac{1 - x_i - h}{\varepsilon} \right) - \exp\left(-\beta \frac{1 - x_i + h}{\varepsilon} \right) \right] \\
&= Ch + C \exp\left(-\beta \frac{1-x_i}{\varepsilon} \right) \left[\exp\left(\frac{\beta h}{\varepsilon} \right) - \exp\left(-\frac{\beta h}{\varepsilon} \right) \right] \\
&= Ch + C \sinh\left(\frac{\beta h}{\varepsilon} \right) \exp\left(-\beta \frac{1-x_i}{\varepsilon} \right).
\end{aligned}$$

Es gilt

$$\sinh(t) = \frac{e^t - e^{-t}}{2} \leq \frac{e^t}{2} = Ce^t.$$

Damit folgt

$$|\tau_i| \leq Ch + C \exp\left(-\beta \frac{1-x_i}{\varepsilon} + \frac{\beta h}{\varepsilon} \right) = Ch + C \exp\left(-\beta \frac{1 - x_{i+1}}{\varepsilon} \right).$$

Auch für den Konsistenzfehler des zweiten Anteiles $v(x)$ findet man

$$|v(x_i) - v_i| \leq Ch + C \exp\left(-\beta \frac{1 - x_{i+1}}{\varepsilon} \right).$$

Die Kombination beider Anteile ergibt die Gesamtabschätzung. ■

Folgerung 2.37 Konvergenz des einfachen Upwind–Verfahrens außerhalb von Grenzschichten. *Unter den Voraussetzungen von Satz 2.34 und Satz 2.36 konvergiert das einfache Upwind–Verfahren in einem Intervall $[0, 1-\delta]$ für festes $\delta > 0$ von erster Ordnung mit einer Konvergenzkonstante unabhängig von ε.*

Bemerkung 2.38 Verhalten innerhalb der Grenzschicht basierend auf der Abschätzung. Sei $\varepsilon < h$, dann erhält man im Punkt x_{N-2} die Abschätzung

$$\begin{aligned}
|u(x_{N-2}) - u_{N-2}| &\leq Ch + C \exp\left(-\beta^* \frac{1 - x_{N-1}}{\varepsilon} \right) = Ch + C \exp\left(-\beta^* \frac{h}{\varepsilon} \right) \\
&\leq Ch + Ch = \mathcal{O}(h),
\end{aligned}$$

da die Exponentialfunktion mit einem betragsmäßig großen negativen Argument gegen die lineare Funktion abgeschätzt werden kann. Für x_{N-1} erhält man jedoch

$$|u(x_{N-1}) - u_{N-1}| \leq Ch + C \exp\left(-\beta^* \frac{1 - x_N}{\varepsilon}\right) = Ch + C = \mathcal{O}(1),$$

da $x_N = 1$. □

Beispiel 2.39 Die obige Beobachtung ist kein Problem der erzielten Abschätzung. Betrachte

$$-\varepsilon u''(x) - u'(x) = 0, \quad u(0) = 0,\ u(1) = 1.$$

Die Lösung dieses Problems besitzt eine Grenzschicht bei $x = 0$. Mit dem einfachen Upwind–Verfahren erhält man

$$u_i = \frac{1 - r^i}{1 - r^N}, \quad \text{mit} \quad r = \frac{\varepsilon}{\varepsilon + h}.$$

Für $h = \varepsilon$ erhält man

$$u_1 = \frac{1 - r}{1 - r^N} = \frac{1 - 1/2}{1 - (1/2)^N} = \frac{1/2}{1 - (1/2)^N} \approx \frac{1}{2}.$$

Für die Lösung gilt jedoch

$$u(x_1) = \frac{1 - e^{-1}}{1 - e^{-1/\varepsilon}} \approx 0.63$$

für kleine ε. Damit ist der Fehler $\mathcal{O}(1)$. Somit kann man nicht erwarten, die Abschätzung aus Satz 2.36 wesentlich zu verbessern. □

Bemerkung 2.40 Typisches Verhalten innerhalb der Grenzschicht in numerischen Simulationen. Betrachte konstantes ε und variables h. Ist h groß genug, dann liegen alle Gitterpunkte außerhalb der Grenzschicht. Wird h verkleinert, erhöht sich der Fehler, weil dann der erste Gitterpunkt von außerhalb sich in die Grenzschicht hineinbewegt, siehe Abbildung 2.3. Wenn h dann hinreichend klein wird, dann fällt der Fehler wieder. In diesem Falle greift die erste Abschätzung von Satz 2.36.

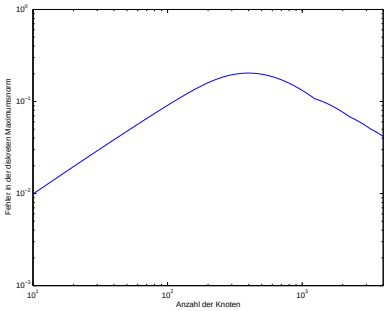


Abbildung 2.3: Fehler des einfachen Upwind–Verfahrens für Beispiel 1.7, $\varepsilon = 1e-3$ und unterschiedlicher Anzahl von Gitterpunkten.

□

Bemerkung 2.41 Interpretation des Upwind–Verfahrens als künstliche Diffusion. Die Schwierigkeiten der numerischen Lösung eines singulär gestörten Problems liegen in den unterschiedlichen Größenordnungen von Diffusion und Konvektion. Es ist klar, dass die numerische Lösung einfacher wird, je größer die Diffusion im Vergleich zur Konvektion ist.

Betrachte $b > 0$. Dann gilt

$$\begin{aligned} b_i D^{\mathcal{N}} u_i &= b_i D^+ u_i = \frac{u_i - u_{i-1}}{h} = b_i \frac{u_{i+1} - u_{i-1}}{2h} + b_i \frac{-u_{i+1} + 2u_i - u_{i-1}}{2h} \\ &= b_i D^0 u_i - \frac{b_i h}{2} D^+ D^- u_i. \end{aligned}$$

Damit kann das einfache Upwind–Verfahren (2.6) in der Form

$$\begin{aligned} -\left(\varepsilon + \frac{b_i h}{2}\right) D^+ D^- u_i + b_i D^0 u_i + c_i u_i &= f_i, \quad \text{für } i = 1, \ldots, N-1, \\ u_0 = u_N &= 0, \end{aligned}$$

geschrieben werden.

Der Diffusionskoeffizient wird also künstlich erhöht und er besitzt die Größenordnung $\mathcal{O}(h)$. Das einfache Upwind–Verfahren ist also nichts anderes als das zentrale Differenzenverfahren angewandt auf ein Problem mit hinreichend großer, $\mathcal{O}(h)$, Diffusions. Man hat bereits in Beispiel 2.29 gesehen, dass das zentrale Differenzenverfahren für einen Diffusionskoeffizienten der Größenordnung $\mathcal{O}(h)$ vernünftige Ergebnisse liefert. □

Man kann Verfahren mit künstlicher Diffusion auch direkt definieren.

Definition 2.42 Verfahren mit künstlicher Diffusion, angepasstes Upwind–Verfahren. Ein Finite–Differenzen–Verfahren mit künstlicher Diffusion ist durch

$$\begin{aligned} -\varepsilon \sigma(q(x_i)) D^+ D^- u_i + b_i D^0 u_i + c_i u_i &= f_i, \quad \text{für } i = 1, \ldots, N-1, \\ u_0 = u_N &= 0, \\ q(x) &:= \frac{b(x)h}{2\varepsilon}, \end{aligned} \tag{2.7}$$

gegeben. Man nennt das Verfahren auch angepasstes Upwind–Verfahren. □

Bemerkung 2.43

- Das einfache Upwind–Verfahren (2.6) erhält man für $\sigma(q) = 1 + q$.
- Die Einführung künstlicher Diffusion verfälscht die ursprüngliche Aufgabe erheblich. Betrachte beispielsweise

  $$-\varepsilon u'' + u' = 1 \quad \text{auf } (0,1), \quad u(0) = u(1) = 0$$

  mit der Lösung

  $$u(x) = x - \frac{\exp\left(-\frac{1-x}{\varepsilon}\right) - \exp\left(-\frac{1}{\varepsilon}\right)}{1 - \exp\left(-\frac{1}{\varepsilon}\right)}, \tag{2.8}$$

  siehe Beispiel 1.7. Der zweite Term ist für das Erfüllen der Randbedingung bei $x = 1$ verantwortlich. Er ist nur wesentlich von Null verschieden im Intervall $[1 - \varepsilon, 1]$, siehe Abbildung 2.4. Führt man künstliche Diffusion ein, dann erhält man eine gestörte Lösung und der Term, der für das Erfülltsein der Randbedingung verantwortlich ist, ist im Intervall $[1 - \varepsilon\sigma(q(x_{N-1})), 1]$ wesentlich von Null verschieden. Das bedeutet, die Grenzschicht ist (weit) weniger steil. Man sagt, die Grenzschicht wird verschmiert.

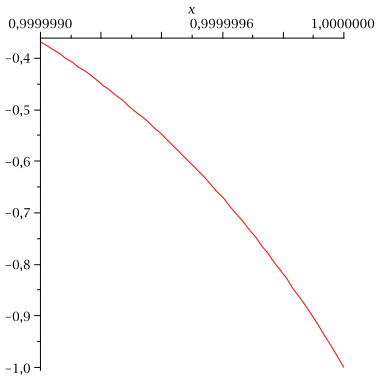


Abbildung 2.4: Zweiter Term der Lösung (2.8) für $\varepsilon = 10^{-6}$.

Beim einfachen Upwind–Verfahren ist

$$\varepsilon\sigma(q(x_{N-1})) = \varepsilon + \varepsilon\frac{b_{N-1}h}{2\varepsilon} = \varepsilon + \frac{b_{N-1}h}{2}.$$

Dieser Ausdruck ist in realistischen Situationen, das heißt für $\varepsilon \ll b_{N-1}$ und $\varepsilon \ll h$, um Ordnungen größer als ε.

□

Die Frage ist, ob man stabile Verfahren mit weniger Verschmierungen konstruieren kann.

Satz 2.44 Stabilität von Verfahren mit künstlicher Diffusion. *Seien $b(x) > \beta > 0$, $c(x) \geq 0$ und $\sigma(q) > q$. Dann ist die Koeffizientenmatrix des Verfahrens mit künstlicher Diffusion (2.7) eine M–Matrix und das Verfahren ist stabil in der diskrete Maximumsnorm. Die Stabilitätskonstante hängt nicht von ε ab.*

Beweis: Der Beweis geht im Prinzip wie der von Satz 2.34, Übungsaufgabe. ■

Satz 2.45 Konsistenz von Verfahren mit künstlicher Diffusion. *Seien die Voraussetzungen von Satz 2.44 gegeben, sei $u \in C^4([0,1])$ und sei*

$$|\sigma(q) - 1| \leq \min\{q, Mq^2\},$$

mit einer Konstanten $M > 0$. Dann ist für festes ε der Konsistenzfehler des Verfahrens mit künstlicher Diffusion (2.7) von zweiter Ordnung.

Beweis: Der Konsistenzfehler im Punkt x_i ist

$$\begin{aligned} |\tau_i| &= \big|\big[-\varepsilon\sigma(q_i)\, D^+D^-u(x_i) + b_i D^0 u(x_i) + c_i u(x_i)\big] \\ &\quad -\big[-\varepsilon u''(x_i) + b_i u'(x_i) + c_i u(x_i)\big]\big| \\ &= \big|\varepsilon\sigma(q_i)\left(u''(x_i) - D^+D^-u_i\right) + \varepsilon\left(1-\sigma(q_i)\right)u''(x_i) + b_i\left(D^0u(x_i) - u'(x_i)\right)\big|. \end{aligned}$$

Aus den Konsistenzfehlerabschätzungen aus Beispiel 2.7 folgt

$$|\tau_i| \leq C\left[\varepsilon\,|\sigma(q_i)|\,h^2\left\|u^{(4)}\right\|_\infty + \varepsilon\,|1-\sigma(q_i)|\left\|u^{(2)}\right\|_\infty + h^2\left\|u^{(3)}\right\|_\infty\right].$$

Mit der Voraussetzung des Satzes und der Definition von $q(x)$ ergeben sich

$$\begin{aligned} |\sigma(q_i)| &\leq |\sigma(q_i) - 1| + 1 \leq \min\left\{q_i, Mq_i^2\right\} + 1 \leq q_i + 1 \leq C\frac{h}{\varepsilon} + 1, \\ |1-\sigma(q_i)| &\leq Mq_i^2 \leq C\frac{h^2}{\varepsilon^2}. \end{aligned}$$

Durch Einsetzen folgt

$$\begin{aligned} |\tau_i| &\leq C\left[\left(\frac{h}{\varepsilon}+1\right)\varepsilon h^2\left\|u^{(4)}\right\|_\infty + \varepsilon C\frac{h^2}{\varepsilon^2}\left\|u^{(2)}\right\|_\infty + h^2\left\|u^{(3)}\right\|_\infty\right] \\ &\leq C(\varepsilon)h^2. \end{aligned} \tag{2.9}$$

■

Bemerkung 2.46

- Beispiel für Funktionen $\sigma(q)$, welche die Voraussetzungen von Satz 2.45 erfüllen sind *Übungsaufgabe*

$$\sigma(q) = \max\{1, q\}, \quad \sigma(q) = \sqrt{1+q^2}, \quad \sigma(q) = 1 + \frac{q^2}{1+q}.$$

  Die letzte Variante wird Samarskii[2]–Upwind–Verfahren genannt.
- Die Konsistenz ist nur für konstantes ε von zweiter Ordnung. Der Vorfaktor $C(\varepsilon)$ divergiert gegen Unendlich für $\varepsilon \to 0$, siehe mittlerer Summand in (2.9). Damit werden die Verfahren für $\varepsilon \to 0$ immer schlechter. Man kann zeigen, dass die Konsistenz unabhängig von ε außerhalb der Grenzschicht nur von erster Ordnung ist. Das typische Verhalten in der Grenzschicht ist wie beim einfachen Upwind–Verfahren, siehe Bemerkung 2.40.

□

Bemerkung 2.47 Fazit.

- Das zentrale Differenzenverfahren ist für singulär gestörte Probleme nicht geeignet.
- Das einfache Upwind–Verfahren ist stabil, aber zu ungenau (von erster Ordnung konsistent). Es verschmiert die Grenzschichten.
- Upwind–Verfahren lassen sich als Verfahren mit künstlicher Diffusion interpretieren.
- Angepasste Upwind–Verfahren können für festes ε von zweiter Ordnung konsistent sein. Diese Eigenschaft ist aber nicht gleichmäßig in ε.

Die bisher vorgestellten Upwind–Verfahren sind nicht befriedigend, da sie für kleine ε zu ungenau sind und die Konvergenz innerhalb der Grenzschicht von ε abhängt.

□

2.4 Gleichmäßig konvergente Verfahren

Bemerkung 2.48 Motivation. Ziel ist es, Verfahren zu entwickeln, die im gesamten Intervall $[0, 1]$ gleichmäßig konvergieren, also insbesondere auch innerhalb der Grenzschicht. Dazu werden zwei Wege vorgestellt:

- ein Verfahren, welches man durch eine geeignete Wahl der künstlichen Diffusion $\sigma(q)$ in (2.7) erhält,
- Verfahren, welche man durch die Wahl geeigneter Gitter definiert.

In der Praxis hat man oft sehr kleine Diffusionen. Deshalb ist es wichtig, dass numerische Verfahren auch für diese Fälle gute Ergebnisse liefern. Die Konstruktion solcher Verfahren ist nicht trivial. Das wird schon dadurch klar, dass der Grenzübergang $\varepsilon \to 0$ in gewisser Weise unstetig ist, weil sich dadurch die Ordnung der Differentialgleichung ändert. Damit ändert sich zum Beispiel die Anzahl der benötigten Randbedingungen, aber auch die Eigenschaften von Lösungen der Differentialgleichungen unterschiedlicher Ordnung sind unterschiedlich, zum Beispiel die Glätte.

[2] Alexander Andreewitsch Samarskii (1919 – 2008)

Ein gleichmäßig konvergentes numerisches Verfahren muss diesen Grenzübergang ohne Qualitätsverlust bewerkstelligen können.

Dieser Abschnitt folgt teilweise [GR05]. □

Definition 2.49 Gleichmäßige Konvergenz. Man nennt ein Verfahren zur Lösung von (2.4), (2.5) gleichmäßig konvergent von der Ordnung p bezüglich des singulären Störungsparameters ε in der diskreten Maximumsnorm, wenn eine Abschätzung der Form

$$\|u - u_h\|_{\infty,d} \leq Ch^p, \quad p > 0,$$

mit einer von ε unabhängigen Konstanten C gilt. □

2.4.1 Geeignete künstliche Diffusion

Die Wahl einer geeigneten künstlichen Diffusion $\sigma(q)$ lässt sich motivieren, indem man die Lösung von (2.4), (2.5) für $\varepsilon \to 0$ betrachtet.

Lemma 2.50 *Sei $u(x,\varepsilon)$ die Lösung von (2.4), (2.5) mit $b(x) \geq \beta > 0$, $c(x) \geq 0$ und sei $u_0(x)$ die Lösung des reduzierten Problems. Dann gilt für alle $x \in [0, x_0)$ mit $x_0 < 1$*

$$\lim_{\varepsilon \to 0} u(x,\varepsilon) = u_0(x).$$

Beweis: Der Beweis beruht auf dem Vergleichsprinzip, Folgerung 1.27. Setze

$$v_1(x) := \gamma \exp(\beta x), \quad \gamma > 0,$$

dann folgt

$$(Lv_1)(x) = \gamma(-\varepsilon\beta^2 + b(x)\beta + c(x)) \exp(\beta x) \geq \gamma\beta^2(1-\varepsilon)\exp(\beta x) \geq 1$$

für hinreichend großes γ. Weiter setzt man

$$v_2(x) := \exp\left(-\beta\frac{1-x}{\varepsilon}\right).$$

Dann gilt

$$\begin{aligned}(Lv_2)(x) &= \left(-\varepsilon\frac{\beta^2}{\varepsilon^2} + b(x)\frac{\beta}{\varepsilon} + c(x)\right)\exp\left(-\beta\frac{1-x}{\varepsilon}\right)\\ &\geq \frac{\beta}{\varepsilon}\left(-\beta + b(x)\right)\exp\left(-\beta\frac{1-x}{\varepsilon}\right) \geq 0.\end{aligned}$$

Betrachte nun

$$v(x) := M_1\varepsilon v_1(x) + M_2 v_2(x).$$

Dann sind

$$\begin{aligned}(Lv)(x) &= M_1\varepsilon(Lv_1)(x) + M_2(Lv_2)(x) \geq M_1\varepsilon(Lv_1)(x) \geq M_1\varepsilon \geq \varepsilon\left|u_0''(x)\right|\\ &= \left|L(u-u_0)(x)\right|,\\ v(0) &= M_1\varepsilon v_1(0) + M_2 v_2(0) = M_1\varepsilon\gamma + M_2\exp(-\beta/\varepsilon) \geq 0 = \left|(u-u_0)(0)\right|,\\ v(1) &= M_1\varepsilon v_1(1) + M_2 v_2(1) = M_1\varepsilon\gamma\exp(\beta) + M_2 \geq M_2 \geq \left|u_0(1)\right|,\end{aligned}$$

für geeignet gewählte, von ε unabhängige, Konstanten M_1 und M_2. Die Konstanten müssen hinreichend groß sein und sie hängen nur von $u_0(x)$ ab. Nach dem Vergleichsprinzip folgt

$$\left|(u-u_0)(x)\right| \leq v(x) = M_1\varepsilon\gamma\exp(\beta x) + M_2\exp\left(-\beta\frac{1-x}{\varepsilon}\right).$$

Damit erhält man für $x < 1$

$$\lim_{\varepsilon\to 0}\left|(u-u_0)(x)\right| = 0.$$

■

Lemma 2.51 *Unter den Voraussetzungen von Lemma 2.50 existiert eine von x und ε unabhängige Konstante C, so dass für die Lösung von (2.4), (2.5) gilt*

$$\left| u(x,\varepsilon) - \left[u_0(x) - u_0(1)\exp\left(-b(1)\frac{1-x}{\varepsilon}\right)\right]\right| \leq C\varepsilon, \quad x \in [0,1].$$

Beweis: Der Beweis ist ähnlich wie der von Lemma 2.50. ∎

Bemerkung 2.52 Notwendige Bedingung für eine geeigneten Funktion $\sigma(q)$. Seien $\rho^* := h/\varepsilon$ fest und i fest. Das heißt, für $h \to 0$ gilt auch $\varepsilon \to 0$. Ziel ist es, in diesem Falle eine Bedingung für eine geeignete Funktion $\sigma(q)$ zu finden. Wegen $\varepsilon \to 0$ für $h \to 0$ folgt nach Lemma 2.51

$$\begin{aligned}\lim_{h\to 0} u(1-ih) &= \lim_{h\to 0} u((N-i)h) = \lim_{h\to 0}\left(u_0(1) - u_0(1)\exp\left(-b(1)\frac{ih}{\varepsilon}\right)\right)\\ &= u_0(1) - u_0(1)\exp\left(-ib(1)\rho^*\right)\\ &= u_0(1)\left(1-\exp(-2iq(1)\right). \end{aligned} \tag{2.10}$$

Das angepasste Upwind–Verfahren besitzt die Gestalt

$$-\varepsilon\sigma\left(q(b_i)\right)\frac{u_{i+1}-2u_i+u_{i-1}}{h^2} + b_i\frac{u_{i+1}-u_{i-1}}{2h} = f_i - c_iu_i$$

oder nach Erweiterung mit h^2/ε

$$\begin{aligned}-\sigma\left(q(b_i)\right)\left(u_{i+1}-2u_i+u_{i-1}\right) + q(b_i)\left(u_{i+1}-u_{i-1}\right) &= \frac{h^2}{\varepsilon}\left(f_i - c_iu_i\right)\\ &= h\rho^*\left(f_i - c_iu_i\right).\end{aligned}$$

Für den rechten Rand, das heißt $i = N-1$ gilt insbesondere

$$\begin{aligned}&\lim_{h\to 0}\left(-\sigma\left(q_{N-1}\right)\left(u_N - 2u_{N-1}+u_{N-2}\right) + q_{N-1}\left(u_N - u_{N-2}\right)\right)\\ &\quad= \lim_{h\to 0} h\rho^*\left(f_{N-1} - c_{N-1}u_{N-1}\right) \quad \Longrightarrow\\ 0 &= \lim_{h\to 0}\left(-\sigma\left(q_{N-1}\right)\left(u_N - 2u_{N-1}+u_{N-2}\right) + q_{N-1}\left(u_N - u_{N-2}\right)\right).\end{aligned}$$

Einsetzen von (2.10) liefert, wobei ohne Beschränkung der Allgemeinheit $u_0(1) \neq 0$ angenommen wird,

$$\begin{aligned}0 &= -\sigma\left(q(1)\right)\Big(-\exp(-2Nq(1)) + 2\exp(-2(N-1)q(1))\\ &\quad -\exp(-2(N-2)q(1))\Big) + q(1)\Big(-\exp(-2Nq(1)) + \exp(-2(N-2)q(1))\Big)\end{aligned}$$

und nach Division durch $-\exp(-2Nq(1)) \neq 0$

$$0 = -\sigma\left(q(1)\right)\Big(1 - 2\exp(2q(1)) + \exp(4q(1))\Big) + q(1)\Big(1-\exp(4q(1))\Big).$$

Nun gilt

$$\frac{1-e^{4x}}{1-2e^{2x}+e^{4x}} = \frac{e^{-2x}-e^{2x}}{e^{-2x}-2+e^{2x}} = \frac{\left(e^x - e^{-x}\right)\left(e^x+e^{-x}\right)}{\left(e^x-e^{-x}\right)^2} = \frac{e^x+e^{-x}}{e^x-e^{-x}} = \coth(x).$$

Damit folgt

$$\sigma\left(q(1)\right) = q(1)\frac{1-\exp(4q(1))}{1-2\exp(2q(1))+\exp(4q(1))} = q(1)\coth\left(q(1)\right).$$

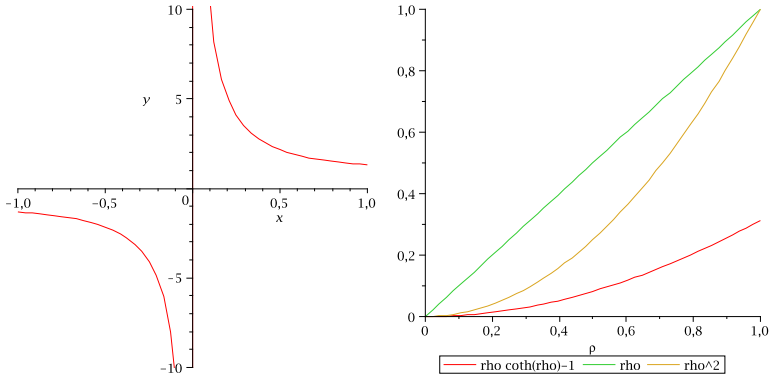


Abbildung 2.5: $\coth(x)$ und Vergleich zu den Bedingungen aus Satz 2.45.

Eine Wahl, die diesem Grenzwert genügt ist

$$\sigma(q) = q \coth(q).$$

Diese Funktion erfüllt auch die Bedingungen für die Konsistenz von Verfahren mit künstlicher Diffusion, Satz 2.45, siehe Abbildung 2.5.

□

Definition 2.53 Iljin[3]–Verfahren, Iljin–Allen[4]–Southwell[5]–Verfahren. Das Verfahren

$$\begin{aligned} -\frac{h}{2} b_i \coth\left(\frac{h}{2\varepsilon} b_i\right) D^+ D^- u_i + b_i D^0 u_i + c_i u_i &= f_i, \quad \text{für } i = 1, \ldots, N-1, \\ u_0 = u_N &= 0, \end{aligned}$$

wird Iljin–Verfahren oder Iljin–Allen–Southwell–Verfahren genannt. □

Satz 2.54 Gleichmäßige Konvergenz des Iljin–Allen–Southwell–Verfahrens. *Das Iljin–Allen–Southwell–Verfahren konvergiert auf* $[0,1]$ *gleichmäßig von erster Ordnung in der diskreten Maximumsnorm, das heißt*

$$\|u(x_i) - u_i\|_{\infty,d} \leq Ch$$

mit einer von ε *und* h *unabhängigen Konstanten* C*.*

Beweis: Der Beweis ist relativ rechenaufwändig, deshalb wird auf die Literatur, [RST08], verwiesen. ■

Beispiel 2.55 Betrachte

$$-\varepsilon u'' + u' = 1 \quad \text{auf } (0,1), \quad u(0) = u(1) = 0,$$

mit der Lösung

$$u(x) = x - \frac{\exp\left(-\frac{1-x}{\varepsilon}\right) - \exp\left(-\frac{1}{\varepsilon}\right)}{1 - \exp\left(-\frac{1}{\varepsilon}\right)}.$$

Für den Fehler in der diskreten Maximumsnorm im Fall $\varepsilon = 10^{-3}$ erhält man

[3] A.M. Iljin
[4] D.N. de G. Allen
[5] Richard V. Southwell

Intervalle	zentr. Diff.	einf. Upwind	IAS–Verfahren
2	124.5	0.00199	0
4	31.004	0.00398	0
8	7.715	0.00793	0
16	2.0235	0.01574	0
32	0.91132	0.03100	2.2204e-16
64	0.77305	0.06015	1.5543e-15
128	0.59276	0.11307	8.3598e-15
256	0.34287	0.18371	1.2388e-14
512	0.12997	0.19679	1.0976e-14
1024	0.03277	0.12933	5.8457e-14
2048	0.00750	0.07486	1.5675e-13
4096	0.00183	0.04076	2.8882e-13

Man sieht, dass das Iljin–Allen–Southwell–Verfahren immer die genauesten Ergebnisse liefert. Sind die Knoten hinreichend entfernt von der Grenzschicht, dann sind die Ergebnisse in den Knoten sogar exakt. □

2.4.2 Grenzschichtangepasste Gitter

Bemerkung 2.56 Motivation. Es wurde bereits gezeigt, dass die Lösung von singulär gestörten Problemen aus zwei Bestandteilen besteht:

- der Lösung des reduzierten Problems, diese ist im allgemeinen glatt und einfach zu approximieren,
- einem Korrekturterm, der das Erfülltsein der Randbedingung am Ausflussrand erzwingt. Dieser ist dafür verantwortlich, dass die Grenzschicht auftritt, dass sich die Lösung in einem sehr kleinen Intervall dramatisch verändert.

Betrachte als typisches Beispiel das 2–Punkt–Randwertproblem aus Beispiel 2.55. Im Intervall $[0, 1-\varepsilon]$ hat die Lösung praktisch die Gestalt $u(x) = x$, ist also sehr einfach auf einem groben Gitter zu approximieren. Der interessante Bestandteil der Lösung ist im Intervall $[1-\varepsilon, 1]$. Wählt man ein äquidistantes Gitter der Schrittweite h, dann gilt im allgemeinen $h > \varepsilon$ und das Intervall $[1-\varepsilon, 1]$ ist in $[x_{N-1}, x_N] = [1-h, 1]$ enthalten. Man kann nicht erwarten, damit das Verhalten der Lösung in $[1-\varepsilon, 1]$ zu auflösen zu können.

Die Idee von grenzschichtangepassten Gittern besteht darin, in der Grenzschicht ein (wesentlich) feineres Gitter zu wählen als außerhalb derselben. Damit besteht die Möglichkeit, die Lösung in der Grenzschicht gut zu approximieren. □

Bemerkung 2.57 Shishkin[6]–Gitter. Betrachte der einfacheren Notation halber ein Problem, bei welchem die Grenzschicht sich bei $x = 0$ befindet. Desweiteren sei $b = -\beta \in \mathbb{R}^+$ eine Konstante. Nun werden die Gitterpunkte gemäß

$$x_i = \phi(i/N),$$

verteilt, wobei die Funktion $\phi(\xi)$ so gewählt werden muss, dass bei $x = 0$ ein hinreichend feines Gitter entsteht. Die Anzahl N der Intervalle ist vorgegeben. Ein Gitter von Shishkin–Typ ist gegeben durch

$$\phi(\xi) = \begin{cases} \dfrac{\sigma\varepsilon}{\beta}\hat{\phi}(\xi) \text{ mit } \hat{\phi}(1/2) = \ln(N) & \text{für } \xi \in [0, 1/2], \\ 1 - 2\left(1 - \dfrac{\sigma\varepsilon}{\beta}\ln(N)\right)(1-\xi) & \text{für } \xi \in [1/2, 1], \end{cases}$$

[6] Grigory I. Shishkin

wobei $\sigma > 0$ ein Parameter ist. Das Shishkin–Gitter (1988) erhält man für

$$\hat{\phi}(\xi) = 2\ln(N)\xi.$$

Damit hat man für die Gitterpunkte $x_0, \ldots, x_{N/2}$

$$x_i - x_{i-1} = \phi\left(\frac{i}{N}\right) - \phi\left(\frac{i-1}{N}\right) = \frac{\sigma\varepsilon}{\beta} 2\ln(N)\left(\frac{i}{N} - \frac{i-1}{N}\right) = 2\frac{\sigma\varepsilon}{\beta}\frac{\ln(N)}{N}$$

unabhängig von i. Für die Gitterpunkte $x_{N/2+1}, \ldots, x_N$ gilt

$$\begin{aligned} x_i - x_{i-1} &= \phi\left(\frac{i}{N}\right) - \phi\left(\frac{i-1}{N}\right) \\ &= 1 - 2\left(1 - \frac{\sigma\varepsilon}{\beta}\ln(N)\right)\left(1 - \frac{i}{N}\right) - 1 + 2\left(1 - \frac{\sigma\varepsilon}{\beta}\ln(N)\right)\left(1 - \frac{i-1}{N}\right) \\ &= \frac{2}{N} - 2\frac{\sigma\varepsilon}{\beta}\frac{\ln(N)}{N}, \end{aligned}$$

unabhängig von i. Dies ist ein stückweise äquidistantes Gitter. Der Übergangspunkt vom sehr feinen auf das grobe Gitter ist bei

$$\tau = x_{N/2} = \frac{\sigma\varepsilon}{\beta}\ln(N).$$

□

Die Wahl des Shishkin–Gitters wird mit dem folgenden Satz gerechtfertigt.

Satz 2.58 Konvergenz des einfachen Upwind–Verfahrens auf einem Shishkin–Gitter. *Betrachte das einfache Upwind–Verfahren auf einem Shishkin–Gitter mit dem Übergangspunkt*

$$\tau = \min\left\{\frac{1}{2}, \frac{\varepsilon}{\beta}\ln(N)\right\},$$

also mit $\sigma = 1$. Dann gilt die Fehlerabschätzung

$$\|u(x_i) - u_i\|_{\infty,d} \leq CN^{-1}\ln(N),$$

mit einer von ε und N unabhängigen Konstanten C.

Beweis: Der Beweis basiert auf der Zerlegung der Lösung in den Anteil vom reduzierten Problem (glatter Anteil) und den Korrekturterm. Er ist relativ aufwändig, siehe [RST08]. ■

Bemerkung 2.59

- Die Konvergenz ist wegen des Faktors $\ln(N)$ leicht suboptimal. Man sieht aber in numerischen Beispielen, dass die obige Abschätzung scharf ist, dass dieser Faktor also nicht entfallen kann.
- Die Idee der Verwendung grenzschichtangepasster Gitter geht bereits auf Bachvalov[7] (1969) zurück. Bei Bachvalov–Gittern gibt es einen glatten Übergang vom feinem zum groben Gittern. Numerische Verfahren sind auf Bachvalov–Gittern schwieriger zu analysieren als auf Shishkin–Gittern.
- Die a priori (vor der numerischen Lösung) Konstruktion geeigneter grenzschichtangepasster Gitter erfordert im wesentlichen die Kenntnis der Lösung. Dies ist in der Praxis vollkommen unrealistisch, insbesondere bei Problemen in zwei oder drei Dimensionen. Man benötigt vielmehr eine a posteriori (während der numerischen Lösung) Konstruktion von angepassten Gittern. Auch dazu gibt es Wege, siehe Kapitel **??**.

[7] Nikolai Sergejewitsch Bachvalov (1934 – 2005)

- Die wesentliche Erkenntnis der Analysis von Verfahren auf a priori grenzschichtangepassten Gittern besteht darin, dass gezeigt wird, dass man auf einem geeigneten Gitter ein einfaches Verfahren verwenden kann und damit vernünftige Fehlerabschätzungen erhält.
- Bei der Nutzung von Shishkin–Gittern muss man jetzt Differenzenquotienten im Knoten $x_{N/2}$ erklären, für welchen die anliegenden Intervalle nicht gleich lang sind.
  Sei x_i ein Knoten und haben die Intervalle $[x_{i-1}, x_i]$ und $[x_i, x_{i+1}]$ die Längen h_i und h_{i+1}. Für den Rückwärts– und Vorwärtsquotienten ändert sich nichts zur Definition 2.4, da man dort immer nur eines der anliegenden Intervalle braucht. Ansonsten definiert man

  $$\tilde{h_i} := \frac{h_i + h_{i+1}}{2}.$$

  Der zentrale Differenzenquotient ist das gewichtete Mittel

  $$D^0 v(x_i) = \frac{1}{2\tilde{h_i}} \left(h_i D^+ v(x_i) + h_{i+1} D^- v(x_i) \right)$$

  Diese Approximation ist von zweiter Ordnung konsistent. Die zweite Ableitung wird wie folgt approximiert

  $$v''(x_i) \approx \delta^2 v_i := \frac{1}{\tilde{h_i}} \left(D^+ v(x_i) - D^- v(x_i) \right) = \frac{1}{\tilde{h_i}} \left(\frac{v_{i+1} - v_i}{h_{i+1}} - \frac{v_i - v_{i-1}}{h_i} \right).$$

  Diese Approximation ist nicht mehr von zweiter Ordnung konsistent. *Übungsaufgabe*
- Die Matrizen, die man bei der Nutzung von grenzschichtangepassten Gittern erhält, sind sehr schlecht konditioniert.

□

Beispiel 2.60 Betrachte wieder

$$-\varepsilon u'' + u' = 1 \quad \text{auf } (0,1), \quad u(0) = u(1) = 0,$$

mit der Lösung

$$u(x) = x - \frac{\exp\left(-\frac{1-x}{\varepsilon}\right) - \exp\left(-\frac{1}{\varepsilon}\right)}{1 - \exp\left(-\frac{1}{\varepsilon}\right)}.$$

Die Grenzschicht ist in diesem Beispiel bei $x = 1$. Deshalb wählt man den Übergangspunkt hier

$$\tau = x_{N/2} = 1 - \frac{\sigma\varepsilon}{\beta} \ln(N) = 1 - \sigma\varepsilon \ln(N).$$

Die Fehler in der diskreten Maximumsnorm für $\varepsilon = 10^{-6}$ und $\sigma = 2$ sind

Intervalle	$\|u - u_h\|_{\infty,d}$
4	0.25584
8	0.16455
16	0.10833
32	0.069125
64	0.043656
128	0.026335
256	0.015402
512	0.0087902
1024	0.0049257
2048	0.0027225
4096	0.0014891

Man stellt fest, das die Ergebnisse stark von der Wahl von σ abhängen, *Übungsaufgabe*. □

Bemerkung 2.61 Fazit. Gleichmäßig konvergente Verfahren kann man auf zwei Arten bekommen:

- Verwendung eines geeignet modifizierten Verfahrens auf einem einfachen Gitter,
- Verwendung eines einfachen Verfahrens auf einem geeignet gewählten Gitter.

□