5 Verfahren für parabolische Probleme

Wir diskutieren zunächst wieder die klassischen Differenzenapproximationen zur Lösung parabolischer Anfangs-Randwert-Aufgaben (ARWAn). Der Übersichtlichkeit halber beschränken wir uns dabei auf das Modellproblem der Wärmeleitungsgleichung in zwei Ortsdimensionen mit Dirichletschen Randbedingungen, d. h. auf die 1. ARWA:

$$\partial_t u + Lu = f \quad \text{in } \Omega \times (0,T), \quad u_{|\partial\Omega} = 0, \quad u_{|t=0} = u^0, \tag{5.0.1}$$

mit einem elliptischen Operator L, der hier exemplarisch als $L := -a\Delta$ mit einer Konstanten $a > 0$ gewählt wird.

Das Definitionsgebiet $\Omega \in \mathbb{R}^2$ wird wieder als glatt berandet oder als konvexes Polygongebiet vorausgesetzt. Die Problemdaten f, g, u^0 sind ebenfalls glatt und kompatibel, so dass die Lösung ebenfalls als glatt angenommen werden kann. Erweiterungen für Probleme mit weniger regulären Daten oder anderen Randbedingungen sowie auf den dreidimensionalen Fall werden gegebenenfalls in Bemerkungen berücksichtigt. Gelegentlich wird auch das eindimensionale Analogon von (5.0.1) betrachtet. Als Basis von Finite-ElemFente-Diskretisierungen dient wieder die variationelle Formulierung von (5.0.1):

$$(\partial_t u, \varphi) + a(u,\varphi) = (f,\varphi) \quad \forall \varphi \in V, \;\; t > 0, \quad u_{|t=0} = 0, \tag{5.0.2}$$

mit dem üblichen Sobolew-Raum $V := H_0^1(\Omega)$ und der symmetrischen und positiv definiten „Energie-Form“ $a(u,\varphi) := (a\nabla u, \nabla\varphi)_\Omega$.

Bei der Diskretisierung von instationären Problemen gibt es drei verschiedene Vorgehensweisen, die wir im Folgenden kurz beschreiben wollen.

i) „Linienmethode“: Zunächst wird eine Diskretisierung bzgl. der Ortsvariablen vorgenommen, d.h.: mit Hilfe eines Finite-Differenzen- oder Finite-Elemente-Ansatzes werden „diskrete“ Funktionen $u_h(t) = u_h(\cdot,t)$ bestimmt aus der Gleichung

$$u_h'(t) + \mathcal{A}_h u_h(t) = f_h(t), \;\; t \geq 0, \quad u_h(0) = u_h^0. \tag{5.0.3}$$

Im Falle eines Differenzenverfahrens auf einem Ortsgitter $\{x_i\}_{i=1,\dots,N}$ ist die diskrete Funktion $u_h(t) = (u_n(t))_{n=1}^N$ der Vektor der Knotenwerte $u_n(t) \approx u(x_n,t)$, $\mathcal{A}_h = A_h : \mathbb{R}^N \to \mathbb{R}^N$ die zum verwendeten Differenzenoperator korrespondierende Matrix und $f_h = b_h = (f(x_n))_{n=1}^N$. Beim Finite-Elemente-Ansatz ist $u_h(t) \in V_h$ eine Finite-Elemente-Funktion, $\mathcal{A}_h =: V_h \to V_h$ das durch

$$(\mathcal{A}_h v_h, \varphi_h) = a(v_h,\varphi_h), \quad v_h, \varphi_h \in V_h,$$

definierte diskrete Analogon zum Differentialoperators L und $f_h = P_h f \in V_h$ die L^2-Projektion der rechten Seite f auf V_h. Die Aufgabe (5.0.3) lautet demgemäß in variationeller Form wie folgt:

$$(u_h'(t),\varphi_h) + a(u_h(t),\varphi_h) = (f,\varphi_h) \quad \forall \varphi_h \in V_h, \;\; t \in I, \quad u_h(0) = P_h u^0. \tag{5.0.4}$$

Nach Einführung einer Knotenbasis $\{\varphi_h^{(n)},\, n=1,...,N=\dim(V_h)\}$ geht dieses Problem über in ein System für den Vektor $U_h(t)=(U_n(t))_{n=1}^N$ der zugehörigen Knotenwerte,

$$M_h U_h'(t)+A_h U_h(t)=b_h(t),\quad t\geq 0,\quad U_h(0)=U_h^0, \tag{5.0.5}$$

mit der „Steifigkeitsmatrix“ und „Massenmatrix“

$$A_h=(a(\varphi_h^{(n)},\varphi_h^{(m)}))_{n,m=1}^N,\quad M_h=((\varphi_h^{(n)},\varphi_h^{(m)}))_{n,m=1}^N$$

der Finite-Elemente-Basis. In beiden Fällen, (5.0.3) oder (5.0.5), handelt es sich um ein System von (linearen) gewöhnlichen Differentialgleichungen. Dieses wird nun mit einem der ublichen Schemata bzgl. der Zeit diskretisiert. Nach Wahl einer (zunächst konstanten) Zeitschrittweite k werden zu den „diskreten“ Zeitleveln $t_m=mk$ Approximationen $U_h^m=(U_n^m)_{n=1}^N$ zu $u(\cdot,t_m)$ bestimmt. Wir sprechen von einem „Einschritt-“ bzw. einem „Zweischrittverfahren“, wenn U_h^m aus den vorausgehenden Werten gemäß einer Formel der Form

$$U_h^m=F(U_h^m,U_h^{m-1})\quad\text{bzw.}\quad U_h^m=F(U_h^m,U_h^{m-1},U_h^{m-2})$$

berechnet wird. Im Falle

$$U_h^m=F(U_h^{m-1})\quad\text{bzw.}\quad U_h^m=F(U_h^{m-1},U_h^{m-2})$$

heißt die Methode „explizit“. Zur Durchführung einer nicht expliziten, d.h. „impliziten“, Methode müssen in jedem Zeitschritt Gleichungssysteme gelöst werden. Die hohe Dimension des Systems, $N=\#\{\text{Gitterpunkte } a_n\}$ bzw. $N=\dim(V_h)$, mit $N\sim 10^3-10^8$ impliziert im Hinblick auf die Lösungsökonomie Einschränkungen bei der Wahl der Verfahren. Es kommen in der Regel nur Schemata einfacher Struktur, d.h. mit wenigen Matrix-Vektor-Multiplikationen, und niedriger Ordnung $r=1-4$ in Frage. Eine weitere wesentliche Einschränkung besteht in der generischen Steifheit des Systems. Die Systemmatrix A_h hat in Abhängigkeit von der (gleichförmigen) Gitterfeinheit h die Kondition

$$\kappa_2(A_h)\approx h^{-2}.$$

Bei *expliziten* Zeitschrittschemata sind also einschneidende Schrittweitenrestriktionen einzuhalten, welche deren Verwendung in der Regel verbietet. Der formale Vorteil der expliziten Verfahren, dass in den einzelnen Zeitschritten keine impliziten Gleichungssysteme zu lösen sind, wird besonders in höheren Raumdimensionen (d $=2,3$) durch die hohe Zahl von durchzuführenden Zeitschritten (besonders bei Verwendung lokal verfeinerter Ortsgitter) schnell aufgehoben.

Beispiel numerischer Instabilität: Wir wollen dies anhand einer einfachen Modellsituation illustrieren. Die eindimensionale, homogene Version der ARWA (5.0.1)

$$\partial_t u-\partial_x^2 u=0\quad\text{in }\Omega=(0,1),\quad u_{|\partial\Omega}=0,\quad u_{|t=0}=u^0 \tag{5.0.6}$$

wird auf einem äquidistanten Gitter $0=x_0<...<x_n<...<x_{N+1}=1$ mit Hilfe

zentraler Differenzenquotienten 2. Ordnung,

$$\partial_x^2 u(x_n,t) \approx h^{-2}\Big\{U_{n-1}(t) - 2U_n(t) + U_{n+1}(t)\Big\}\,.$$

diskretisiert. Die Vektorfunktion $U_h(t) = (U_n(t))_{n=1}^N$ genügt dann dem System gewöhnlicher Differentialgleichungen

$$U_n'(t) - h^{-2}\{U_{n-1}(t) - 2U_n(t) + U_{n+1}(t)\} = 0\,,$$

wobei bei Berücksichtigung der Randbedingungen $U_0 \equiv U_{N+1} \equiv 0$ gesetzt ist. Die Anfangswerte sind naturgemäß $U_n(0) = u^0(x_n)$. Dies kann kompakt geschrieben werden als

$$U_h' + A_h U_h(t) = 0\,,\ t \geq 0, \quad U_h(0) = U^0\,, \tag{5.0.7}$$

mit der $(N \times N)$-Matrix

$$A_h \;=\; h^{-2}\begin{bmatrix} -2 & 1 & & & 0 \\ 1 & -2 & & & \\ & \ddots & \ddots & \ddots & \\ & & & -2 & 1 \\ 0 & & & 1 & -2 \end{bmatrix}.$$

Diese Matrix hat, wie wir bereits wissen, die Eigenwerte

$$0 < \lambda_1 \leq ... \leq \lambda_N = \frac{4}{h^2} + \mathcal{O}(h^2), \tag{5.0.8}$$

d. h.: Das nach Ortsdiskretisierung entstandene System (5.0.7) wird für kleines h zunehmend steif mit Steifigkeitsrate $\kappa = \mathcal{O}(h^{-2})$. Beim expliziten Euler[1]-Schema („Polygonzugmethode“) ist z. B. aus Stabilitätsgründen die Schrittweitenbedingung

$$-\lambda_n k \in [-2,0] \quad \Leftrightarrow \quad k \leq \tfrac{1}{2}h^2 \tag{5.0.9}$$

einzuhalten. Diese Schrittweitenbeschränkung für explizite Verfahren hat entscheidende praktische Bedeutung. Wir wollen das Phänomen der numerischen Instabilität illustrieren. Dazu betrachten wir als einfachstes explizites Zeitschrittschema das klassische Euler-Verfahren (Polygonzugmethode) mit äquidistanter Schrittweite k. Dies führt auf die folgenden Differenzengleichungen für die Approximationen $U_n^m \approx u(x_n,t_m)$:

$$U_n^{m+1} = U_n^m + \frac{k}{h^2}\Big(U_{n-1}^m - 2U_n^m + U_{n+1}^m\Big). \tag{5.0.10}$$

[1]Leonhard Euler (1707–1783), geb. in Basel: universeller Mathematiker und Physiker; bedeutendster und produktivster Mathematiker seiner Zeit; wirkte in Berlin und St. Petersburg; Arbeiten zu allen mathemischen Gebieten seiner Zeit.

Für $k = h^2$ ist dann

$$U_n^{m+1} = U_{n-1}^m - U_n^m + U_{n+1}^m.$$

Im Fall oszillierender Anfangsdaten $u_n^0 = (-1)^n$ ergibt sich

$$U_n^1 = (-1)^{n-1} - (-1)^n + (-1)^{n+1} = -3(-1)^n = -3U_n^0,$$

und bei Fortsetzung dieses Arguments:

$$U_n^m = (-3)^m U_n^0, \quad m \geq 1,\ n = 1, ..., N. \tag{5.0.11}$$

Dieses Verhalten bedeutet numerische „Instabilität“. Es mag unrealistisch erscheinen, eine oszillierende Anfangsbedingung der Art $U_n^0 = (-1)^n$ anzunehmen, doch bedingt durch Rundungsfehler könnte gelten:

$$U_n^0 = V_n^0 + \varepsilon(-1)^n \tag{5.0.12}$$

mit „glatten“ exakten Anfangsdaten V^0. Wegen der Linearität der betrachteten Differenzengleichungen folgt

$$U_n^m = V_n^m + \varepsilon(-3)^m(-1)^n, \tag{5.0.13}$$

so dass die anfänglich kleinen Anfangsstörungen schnell anwachsen; z. B. ist diese für $\varepsilon = 10^{-15} > 3^{-32}$ bereits nach nur 32 Zeitschritten auf Größe ≈ 1 angewachsen und zwar unabhängig von der Größe von h.

ii) „Rothe-Methode“: Bei der Rothe[2]-Methode wird die Differentialgleichung als gewöhnliche Differentialgleichung für eine Hilbertraum-wertige Funktion $U(t) \in V$ aufgefasst und zunächst mit einem A-stabilen Verfahren in der Zeit diskretisiert. Bei Verwendung z. B. des impliziten Euler-Schemas ergibt sich eine Folge von speziellen Randwertaufgaben

$$U^m + kLU^m = U^{m-1} + kf^m, \quad m \geq 1, \quad U^0(x) = u^0(x)\,.$$

Diese Probleme werden nun nacheinander auf möglicherweise wechselnden, dem Lösungsverlauf angepaßten Ortsgittern diskretisiert. Das Problem ist dabei der adäquate Transfer der jeweiligen Startlösung U^{m-1} vom alten auf das neue Ortsgitter. Hier zeigt sich wieder der systematische Vorteil einer Finite-Elemente-Galerkin-Methode, bei der sich ganz automatisch als *richtige* Wahl die L^2-Projektion von U^{m-1} auf das neue Gitter ergibt.

iii) Globale Orts-Zeit-Diskretisierung: Ähnlich wie bei den Transportproblemen in zwei Dimensionen könnte auch bei der Wärmeleitungsgleichung eine simultane Diskretisierung (etwa mit einem Finite-Elemente-Galerkin-Verfahren) auf einem unstrukturierten Gitter der ganzen (x,t)-Ebene erfolgen. Dieser theoretisch durchaus attraktive Ansatz wird aber bei höher dimensionalen Problemen wegen der globalen Kopplung aller Unbekannten sehr rechenaufwendig und spielt daher bei parabolischen Problemen in der Praxis keine wesentliche Rolle.

[2]Erich Rothe (1895–1988): Deutscher Mathematiker; Promotion und Habilitation in Berlin (1928), danach Assistent in Breslau, nach dem Krieg Prof. an der University of Michigan, USA.

Im Folgenden Abschnitt werden wir Differenzenapproximationen in Verbindung mit der Linienmethode betrachten. Die Rothe-Methode wird in Verbindung mit Finite-Elemente-Verfahren im Ort diskutiert. Dies mündet dann auch ohne Probleme in Galerkin-Diskretisierungen simultan in Ort und Zeit, den sog. „unstetigen“ oder „stetigen“ Galerkin-Verfahren (sog. „dG(r)-“ oder „cG(r)-Verfahren“).

5.1 Differenzenverfahren für parabolische Probleme

5.1.1 Zeitschrittverfahren

Wir beginnen mit der Diskussion der „Linienmethode“ zur Diskretisierung von parabolischen ARWAn der Art

$$\partial_t u + Lu = f \quad \text{in } Q_T := \Omega \times I, \quad u_{|\partial\Omega} = 0,\ u_{|t=0} = u^0, \tag{5.1.14}$$

mit $L := -a\Delta$ auf einem beschränkten (regulär berandeten) Gebiet $\Omega \subset \mathbb{R}^2$ und einem Zeitintervall $I = [0, T]$. Der Einfachheit halber wird die rechte Seite f gelegentlich als Null angenommen.

Ortsdiskretisierung von (5.1.14) mit einem der üblichen Differenzenverfahren (z. B. dem 5-Punkte-Operator mit geeigneter Randapproximation) führt auf ein System gewöhnlicher Differentialgleichungen

$$U_h'(t) + A_h U_h(t) = 0, \quad t > 0, \quad U_h(0) = U^0, \tag{5.1.15}$$

für den Vektor $U_h(t) \in \mathbb{R}^N$ der Knotenwerte. Da die Eigenwerte der Systemmatrix A_h alle reell sind (oder wenigstens nahe an der rellen Achse liegen), käme zur stabilen Integration des Systems (5.1.15) jede A(0)-stabile Formel in Frage. Dabei muss aber der hohe numerische Aufwand bei der Durchführung komplizierter impliziter Verfahren hoher Ordnung berücksichtigt werden. Durch Übertragung der klassischen Zeitschrittformeln für gewöhnliche Differentialgleichungen auf das System (5.1.15) erhalten wir unter Benutzung der oben eingeführten Bezeichnungen die folgenden einfachsten Einschrittverfahren:

1) *Explizites Euler-Verfahren (Polygonzugmethode):*

$$k^{-1}\{U_h^m - U_h^{m-1}\} + A_h U_h^{m-1} = f^{m-1}, \quad m \geq 1,$$

2) *Implizites Euler-Verfahren:*

$$k^{-1}\{U_h^m - U_h^{m-1}\} + A_h U_h^m = f^m, \quad m \geq 1,$$

3) *Crank[3]-Nicolson[4]-Verfahren (Trapezregel):*

$$k^{-1}\{U_h^m - U_h^{m-1}\} + \tfrac{1}{2}A_h(U_h^m + U_h^{m-1}) = \tfrac{1}{2}(f^m + f^{m-1}), \quad m \geq 1,$$

und die Zweischrittverfahren: 4) *BDF(2)-Verfahren (Rückwärtsdifferenzenformel):*

$$\tfrac{1}{2}k^{-1}\{3U_h^m - 4U_h^{m-1} + U_h^{m-2}\} + A_h U_h^m = f^m, \quad m \geq 2,$$

5) *Mittelpunkts-Verfahren:*

$$\tfrac{1}{2}k^{-1}\{U_h^m - U_h^{m-2}\} + A_h U_h^{m-1} = f^{m-1}, \quad m \geq 2,$$

6) *Simpson-Verfahren:*

$$\tfrac{1}{2}k^{-1}\{U_h^m - U_h^{m-2}\} + \tfrac{2}{3}A_h\{U_h^m + 4U_h^{m-1} + U_h^{m-2}\} = f^{m-1}, \quad m \geq 2.$$

Als Startwerte werden gewöhnlich (im Fall glatter Anfangsdaten) einfach die Restriktionen $U_n^0 = u^0(a_n)$ verwendet. Bei den Zweischrittverfahren wird der zweite erforderliche Startwert U_h^1 durch Anwendung einer Einschrittformel entsprechender Ordnung gewonnen. *Wegen ihrer inhärenten Instabilität (triviales Stabilitätsgebiet) kommen die Mittelpunktsformel und die Simpson-Formel für die praktische Anwendung nicht in Frage.*

Wie bei der Analyse von Differenzenverfahren üblich verwenden wir den „Abschneidefehler“ $\tau_{h,k}^m = (\tau_n^m)_{n=1}^N$ der Differenzenformeln. Diesen erhält man wieder durch formales Auswerten der Differenzenformeln auf der exakten Lösung:

$$k\,\tau_{h,k}^m := u^m - F(u^m, u^{m-1}, u^{m-2}).$$

Bei einer Ortsdiskretisierung der Ordnung p verhält sich der Abschneidefehler dann gemäß

$$\tau_{h,k}^m = \mathcal{O}(h^p + k^q),$$

wobei q die „Ordnung“ des Zeitschrittverfahrens ist. Von der Fehleranalyse der Zeitschrittverfahren für gewöhnliche Differentialgleichungen wissen wir bereits, dass die einfachen Euler-Verfahren die Ordnung $q = 1$ und das Crank-Nicolson- sowie das BDF(2)-Verfahren die Ordnung $q = 2$ haben. Später werden wir noch Verfahren der Ordnung $q = 3, 4$ kennenlernen. Bei der Analyse dieser Zeitschrittschemata für parabolische Probleme ist die genaue Abhängigkeit des Abschneidefehlers von der örtlichen und zeitlichen Regularität der Lösung interessant.

[3]John Crank (1916–2006): Englischer Mathematiker; Prof. an der Brunel University, Uxbridge, England; Arbeiten zur Numerik partieller Differentialgleichungen.

[4]Phyllis L. Nicolson (1917–1968): Englische Physikerin; Lecturer in Leeds und Manchester.

Hilfssatz 5.1 (Konsistenz): *Für die ARWA (5.1.14) genügen die Abschneidefehler der betrachteten Differenzenverfahren den folgenden (scharfen) Abschätzungen:*

i) Explizites und implizites Euler-Verfahren:

$$\max_{\bar{Q}_T} |\tau_{h,k}^m| \le \max_{\bar{Q}_T} |\tau_h^m| + \tfrac{1}{2} k \max_{\bar{Q}_T} |\partial_t^2 u| \,; \tag{5.1.16}$$

ii) Crank-Nicolson-Verfahren:

$$\max_{\bar{Q}_T} |\tau_{h,k}^m| \le \max_{\bar{Q}_T} |\tau_h^m| + \tfrac{1}{12} k^2 \max_{\bar{Q}_T} |\partial_t^3 u| \,; \tag{5.1.17}$$

iii) BDF(2)-Verfahren:

$$\max_{\bar{Q}_T} |\tau_{h,k}^m| \le \max_{\bar{Q}_T} |\tau_h^m| + \tfrac{2}{3} k^2 \max_{\bar{Q}_T} |\partial_t^3 u| \,. \tag{5.1.18}$$

Dabei ist $\tau_h^m = \mathcal{O}(h^2)$ der Abschneidefehler der Ortsdiskretisierung.

Beweis: Der Abschneidefehler der Ortsdiskretisierung genügt i. Allg. der Abschätzung

$$|\tau_h^m| = |Lu^m - L_h u^m| \le c h^2 M_4^m(u),$$

wobei L_h der Ortsdifferenzenoperator ist und $M_4(u) := \max_{\bar{\Omega}} |\nabla^4 u^m|$. Speziell in einer Raumdimension mit $\Omega = (0,1)$ gilt

$$|\tau_h^m| = |\partial_x^2 u^m - L_h u^m| \le \tfrac{1}{12} h^2 \max_{[0,1]} |\partial_x^4 u^m|.$$

i) Für die explizite Euler-Formel gilt

$$\begin{aligned}
|k^{-1}(u^m - u^{m-1}) + L_h u^{m-1}| &= \Big| k^{-1} \int_{t_{m-1}}^{t_m} \partial_t u \, dt + L_h u^{m-1} \Big| \\
&= \Big| k^{-1} \int_{t_{m-1}}^{t_m} \partial_t u \, dt - \partial_t u^{m-1} - L u^{m-1} + L_h u^{m-1} \Big| \\
&\le k^{-1} \Big| \int_{t_{m-1}}^{t_m} \left\{ \partial_t u - \partial_t u^{m-1} \right\} ds \Big| + |L u^{m-1} - L_h u^{m-1}| \\
&\le k^{-1} \int_{t_{m-1}}^{t_m} (t - t_{m-1}) \, dt \max_{[t_{m-1}, t_m]} |\partial_t^2 u| + |\tau_h^{m-1}|
\end{aligned}$$

Es folgt

$$\max_{\bar{Q}_T} |\tau_{h,k}^m| \le \tfrac{1}{2} k \max_{\bar{Q}_T} |\partial_t^2 u| + \max_{\bar{Q}_T} |\tau_h^{m-1}|.$$

Dieselbe Abschätzung gilt auch für die implizite Euler-Formel.

ii) Für die Crank-Nicolson-Formel gilt

$$\begin{aligned}|k^{-1}(u^m-u^{m-1})+\tfrac{1}{2}L_h(u^m+u^{m-1})| &= \Big|k^{-1}\int_{t_{m-1}}^{t_m}\partial_t u\,dt-\tfrac{1}{2}(\partial_t u^m+\partial_t u^{m-1})\\ &\quad+\tfrac{1}{2}(Lu^m-L_hu^m)+\tfrac{1}{2}(Lu^{m-1}-L_hu^{m-1})\Big|\\ &\le k^{-1}\Big|\int_{t_{m-1}}^{t_m}\tfrac{1}{2}(t-t_m)(t-t_{m-1})\,dt\Big|\max_{[t_{m-1},t_m]}|\partial_t^3u|\\ &\quad+\tfrac{1}{2}(|\tau_h^m|+|\tau_h^{m-1}|)\end{aligned}$$

Wir erhalten damit

$$\max_{\bar{Q}_T}|\tau_{h,k}^m|\le\tfrac{1}{12}k^2\max_{\bar{Q}_T}|\partial_t^3u|+\max_{\bar{Q}_T}|\tau_h^m|.$$

iii) Für die BDF(2)-Formel gilt

$$\begin{aligned}\tfrac{1}{2}k^{-1}\{3u^m-4u^{m-1}+u^{m-2}+L_hu^m\} &= \tfrac{1}{2}k^{-1}\{3u^m-4u^{m-1}+u^{m-2}-2k\partial_t u^m\}\\ &\quad+L_hu^m-Lu^m\,.\end{aligned}$$

Taylor-Entwicklung um t_m liefert

$$3u^m-4u^{m-1}+u^{m-2}-2k\partial_t u^m=\tfrac{4}{3}k^3\partial_t^3u(\cdot,\eta^m)$$

mit gewissen Zwischenstellen $\eta^m\in[t_{m-2},t_m]$. Damit erhalten wir

$$\max_{\bar{Q}_T}|\tau_{h,k}^m|\le\tfrac{2}{3}k^2\max_{\bar{Q}_T}|\partial_t^3u|+\max_{\bar{Q}_T}|\tau_h^m|.$$

Dies vervollständigt den Beweis. Q.E.D.

Die Lösung der ARWA (5.1.14) besitzt die explizite Darstellung

$$u(x,t)=\sum_{n=1}^{\infty}u_n^0v^{(n)}(x)e^{-\lambda_nt},\quad(x,t)\in Q_T,\tag{5.1.19}$$

mit den Eigenwerten und (orthonormierten) Eigenfunktionen des regulären „elliptischen“ Operators $L=-a\Delta:V\subset L^2(\Omega)\to L^2(\Omega)$,

$$0<\lambda_1<...\le\lambda_n\le...\ \ (n\in\mathbb{N}),\quad v^{(n)}(x)\in V:\quad Lv^{(n)}=\lambda_nv^{(n)},$$

und den Entwicklungskoeffizienten der Startwerte

$$u^0(x)=\sum_{n=0}^{\infty}u_n^0v^{(n)}(x)\,,\quad u_n^0=(u^0,v^{(n)})_\Omega\,.$$

Diese Darstellung lässt sich wegen der gleichmäßigen Konvergenz der Reihen wie folgt

umformen:

$$\begin{aligned} u(x,t) &= \sum_{n=1}^{\infty} u_n^0 v^{(n)}(x)\Big(\sum_{i=0}^{\infty}(-1)^i\frac{\lambda_n^i t^i}{i!}\Big) = \sum_{i=0}^{\infty}(-1)^i\frac{t^i}{i!}\Big(\sum_{n=1}^{\infty} u_n^0\lambda_n^i v^{(n)}(x)\Big) \\ &= \sum_{i=0}^{\infty}(-1)^i\frac{t^i}{i!}\Big(\sum_{n=1}^{\infty} u_n^0 L^i v^{(n)}(x)\Big) = \sum_{i=0}^{\infty}(-1)^i\frac{t^i}{i!}L^i\Big(\sum_{n=1}^{\infty} u_n^0 v^{(n)}(x)\Big) \\ &= \Big(\sum_{i=0}^{\infty}(-1)^i\frac{1}{i!}L^i\Big)u^0(x) =: e^{-tL}u^0(x). \end{aligned}$$

Die Definition der Operatorfunktion e^{-tL} über eine konvergente Taylor-Reihe lässt sich auf beliebige analytische Funktionen übertragen. Wir betonen, dass eine solche kompakte Lösungsdarstellung nur im Fall zeitlich konstanter Koeffizienten a möglich ist. Auf dem diskreten Zeitgitter gilt dann

$$u(\cdot,t_m) = e^{-kL}u(\cdot,t_{m-1}), \quad m\in\mathbb{N}. \tag{5.1.20}$$

Dies legt es nahe, den Zeitschritt $t_{m-1}\to t_m$ mit Hilfe einer rationalen Approximation $R(z)\approx e^z$ der Exponentialfunktion der „Ordnung“ $q+1$ anzusetzen,

$$R(z) = \frac{P(z)}{Q(z)} = e^z + \mathcal{O}(|z|^{q+1}), \quad z\le 0, \tag{5.1.21}$$

mit geeigneten Polynomen $P\in P_r$ und $Q\in P_s$, wobei natürlich Q auf $z\in\mathbb{R}_-$ keine Nullstellen haben darf. Das Diskretisierungsschema lautet dann

$$U_h^m = R(-kA_h)U_h^{m-1} \quad \text{bzw.} \quad Q(-kA_h)U_h^m = P(-kA_h)U_h^{m-1}. \tag{5.1.22}$$

Die oben betrachteten Einschrittverfahren lassen sich in diesen Rahmen einordnen gemäß:

$$\begin{aligned} &\text{„Expliziter Euler“ :} && R(z) = 1+z\,, \\ &\text{„Impliziter Euler“ :} && R(z) = (1-z)^{-1}\,, \\ &\text{„Crank-Nicolson“ :} && R(z) = (1+\tfrac{1}{2}z)(1-\tfrac{1}{2}z)^{-1}\,. \end{aligned}$$

Durch die Ordnungsbedingung

$$e^z Q_{rs}(z) - P_{rs}(z) = \mathcal{O}(|z|^{r+s+1}), \quad z\le 0, \tag{5.1.23}$$

für den Ansatz $P_{rs}\in P_r$, $Q_{rs}\in P_s$ wird man auf die sog. „Padé[5]-Schemata“ geführt. Diese sind eindeutig bestimmt und werden gewöhnlich in der sog. „Padé-Tafel“ dargestellt:

[5]Henri Eugène Padé (1785–1836): Französischer Mathematiker; Prof. in Poitiers und Bordeaux; entwickelte die sog. Padé-Approximation.

$$\left|\begin{array}{ccccc}
\frac{1}{1} & \frac{1+z}{1} & \frac{1+z+\frac{1}{2}z^2}{1} & \frac{1+x+\frac{1}{2}z^2+\frac{1}{6}z^3}{1} & \cdots \\
\frac{1}{1-z} & \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z} & \frac{1+\frac{2}{3}z+\frac{1}{6}z^2}{1-\frac{1}{3}z} & \frac{1+\frac{3}{4}z+\frac{1}{4}z^2+\frac{1}{24}z^3}{1-\frac{1}{4}z} & \cdots \\
\frac{1}{1-z+\frac{1}{2}z^2} & \frac{1+\frac{1}{3}z}{1-\frac{2}{3}z+\frac{1}{6}z^2} & \frac{1+\frac{1}{2}z+\frac{1}{12}z^2}{1-\frac{1}{2}z+\frac{1}{12}z^2} & \cdots & \cdots \\
\cdots & \cdots & \cdots & \frac{1+\frac{1}{2}z+\frac{1}{10}z^2+\frac{1}{120}z^3}{1-\frac{1}{2}z+\frac{1}{10}z^2-\frac{1}{120}z^3} & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots
\end{array}\right| .$$

Abbildung 5.1: Padé-Tafel.

Offensichtlich sind alle bisher betrachteten Einschrittschemata Padé-Formeln und damit in diesem Sinne ordnungsoptimal. Aus der Padé-Tafel erhalten wir nun weitere Zeitschrittverfahren höherer Ordnung. Dabei kommen aus Ökonomiegründen nur die „diagonalen“ oder „subdiagonalen“ Padé-Schemata in Frage; z. B. die folgenden impliziten Verfahren 3. bzw. 4. Ordnung:

$$\begin{aligned}
(I + \tfrac{1}{3}kA_h)U_h^m &= (I - \tfrac{2}{3}kA_h + \tfrac{1}{6}k^2A_h^2)U_h^{m-1} \quad (\mathrm{q} = 3), \\
(I + \tfrac{1}{2}kA_h + \tfrac{1}{12}k^2A_h^2)U_h^m &= (I - \tfrac{1}{2}kA_h + \tfrac{1}{12}k^2A_h^2)U_h^{m-1} \quad (\mathrm{q} = 4).
\end{aligned}$$

Wir bemerken für die weitere Analyse, dass eine rationale Approximation $R(z)$ der Exponentialfunktion (der Ordnung $r \geq 1$) die folgende Eigenschaft hat:

$$|R(z)| \leq e^{\delta z}, \quad -1 \leq z \leq 0, \tag{5.1.24}$$

mit einem geeigneten $\delta > 0$. Die Wirkung der Zeitschrittschemata des Typs (5.1.22) lässt sich mit Hilfe der Spektralzerlegung der Matrix A_h wieder beschreiben durch:

$$U_h^m = \sum_{n=1}^{N} U_n^0 R(-k\lambda_n)^m v^{(n)}, \quad m \geq 1,$$

bzw. (mit der Euklidischen Vektornorm $|\cdot|$)

$$|U_h^m|^2 = \sum_{n=1}^{N} |U_n^0|^2 \, |R(-k\lambda_n)|^{2m}.$$

Ihr qualitatives Verhalten lässt sich also weitgehend durch die Eigenschaften der verwendeten rationalen Funktion $R(z)$ charakterisieren. Wir stellen einige wichtige Bedingungen für die folgende Analyse zusammen.

i) Die *A-Stabilität*

$$|R(z)| \leq 1, \quad z \leq 0.$$

sichert die „Stabilität“ der Zeititeration $\sup_{m\geq 0}|U_h^m| < \infty$.

ii) Die *strenge A-Stabilität*

$$|R(z)| \leq 1 - ck\,, \quad z \leq -1,$$

sichert die Beschränktheit der diskreten Lösung auch im Fall inhomogener rechter Seiten, $\sup_{m\geq 0}|U_h^m| < c\sup_{m\geq 0}|f^m|$.

iii) Die *starke A-Stabilität*

$$|R(z)| \leq \kappa < 1\,, \quad z \leq -1\,,$$

sichert die (exponentielle) Dämpfung „hochfrequenter“ Lösungsanteile und macht das Verfahren robust gegenüber lokalen Störungen der Daten („Glättungseigenschaft“).

iv) Zur korrekten Wiedergabe von Schwingungsprozessen (im Ort oszillierenden Lösungen) sollte

$$R(\pm i) \sim 1$$

sein, um diese Schwingungen möglichst wenig zu dämpfen („numerischen Dissipativität“).

Offensichtlich können nur *implizite* Verfahren die gelisteten Eigenschaften haben. Das implizite Euler-Schema (und genauso alle sub-diagonalen Padé-Schemata) ist stark A-stabil (mit Limes $\kappa = 0$), neigt aber zur Überdämpfung: $|(1+i)^{-1}| = 1/\sqrt{2}$. Dagegen ist das Crank-Nicolson-Schema (und genauso alle diagonalen Padé-Schemata) nur einfach A-stabil,

$$\lim_{z\to-\infty}\frac{1+\frac{1}{2}z}{1-\frac{1}{2}z} = -1,$$

besitzt aber praktisch auch keine numerische Dissipation: $|(1-i/2)(1+i/2)^{-1}| = 1$. Die fehlende *starke* A-Stabilität hat nachteilige Konsequenzen im Fall von irregulären Anfangswerten u^0 (z.B.: lokalen Temperaturspitzen). Die durch diese Anfangsdaten induzierten hochfrequenten Fehleranteile werden durch das Crank-Nicolson-Schema nur unzureichend ausgedämpft, so dass sich ein unphysikalisches Lösungsverhalten zeigen kann. Es sei daran erinnert, dass der kontinuierliche Differentialoperator stark dämpfend ist:

$$\|u(t)\| \leq e^{-\lambda_{min}t}\|u^0\|\,, \quad t \geq 0\,,$$

mit dem kleinsten Eigenwert des Ortsoperators, $\lambda_{min} > 0$.

Bei Verwendung des Crank-Nicolson-Schemas für Rechnungen über lange Zeiträume sollte es stabilisiert werden, um wenigstens *strenge* A-Stabilität zu sichern. Dies kann ohne Reduktion der Konsistenzordnung durch einen leichten k-abhängigen Schift erfolgen:

$$\big(I + \tfrac{1}{2}(1+ck)kA_h\big)\,U_h^m = \big(I - \tfrac{1}{2}(1-ck)kA_h\big)\,U_h^{m-1}. \tag{5.1.25}$$

Verfahren höherer Ordnung erfordern die Invertierung der Operatorfunktion $Q(-kA_h)$. Dies ist in der Regel zu teuer. Einerseits ist die Besetzungstruktur von $Q(-kA_h)$ selbst bei Polynomgrad $j = 2$ bereits deutlich dichter als die von A_h, andererseits würde das Arbeiten mit der Linearfaktorzerlegung $Q(z) = (z-\mu)(z-\bar{\mu})$ die Verwendung (kostspie-

liger) komplexer Arithmetik erfordern. Geeignet wären dagegen Schemata, bei denen das Nennerpolynom in reelle Linearfaktoren zerfällt: $Q(z) = \prod_{j=1}^{s}(z-\mu_j)$, $\mu_j \in \mathbb{R}$. Durch diesen Ansatz sollten sich systematisch Verfahren mit günstigeren Eigenschaften als die der einfachen Basisschemata gewinnen lassen.

Ein Beispiel für einen solchen Ansatz ist die parameter-abhängige rationale Funktion

$$R_\theta(z) = \frac{(1+\alpha\theta' z)(1+\beta\theta z)^2}{(1-\alpha\theta z)^2(1-\beta\theta' z)} = e^z + O(|z|^3), \quad z \le 0\,,$$

mit $\theta = 1 - \frac{1}{2}\sqrt{2} = 0,292893...$, $\theta' = 1-2\theta$ und beliebigen Werten $\alpha \in (\frac{1}{2}, 1]$, $\beta = 1-\alpha$. Das auf dieser rationalen Funktion basierende Schema ist wegen

$$|R_\theta(z)| < 1\,,\ z < 0\,, \quad \lim_{z\to-\infty} |R_\theta(z)| = \frac{\beta}{\alpha} < 1\,.$$

stark A-stabil. Die Entwicklung

$$R_\theta(z) = 1 + z + \tfrac{1}{2}z^2\{1-(\alpha-\beta)(2\theta^2-4\theta+1)\} + \tfrac{1}{6}r(\theta,\alpha)z^3 + \mathcal{O}(|z|^4)$$

zeigt, dass für die obige Parameterwahl von der Ordnung $\mathcal{O}(k^2)$ ist. Für die Güte dieser Approximation im Vergleich zu der des Crank-Nicolson-Schemas ist die Größe der führenden Fehlerkonstante $r(\alpha)$ bestimmend. Eine Taylor-Entwicklung ergibt

$$\begin{aligned} r(\theta,\alpha) = {} & (18\theta' + 24\theta^3)\alpha^3 + (42\theta^2\theta' + 12\theta\theta'^2 + 30\theta'^3)\alpha^2\beta \\ & + (12\theta^3 + 30\theta^2\theta' + 24\theta\theta'^2 + 6\theta'^3)\alpha\beta^2 + (6\theta^2\theta' + 12\theta\theta'^2 + 6\theta'^3). \end{aligned}$$

Im betrachteten Bereich $\{0,5 < \alpha \le 1\}$ ist $|r(\theta,\alpha)| \le 0,5$. Damit ist die Fehlerkonstante dieser Approximation nur in akzeptablem Maß größer als die entsprechende Fehlerkonstante $\frac{1}{12}$ der Trapezregel. Das zugehörige Verfahren lässt sich in Form eines Teilschrittschemas schreiben (hier für den inhomogenen Fall),

Teilschritt-θ-Verfahren (Fractional-Step-θ-Method):

$$(I + \alpha\theta k A_h)U^{m-1+\theta} = (I - \beta\theta k A_h)U^{m-1} + \theta k f_h^{m-1}, \tag{5.1.26}$$

$$(I + \beta\theta' k A_h)U^{m-\theta} = (I - \alpha\theta' k A_h)U^{m-1+\theta} + \theta' k f_h^{m-\theta}, \tag{5.1.27}$$

$$(I + \alpha\theta k A_h)U^{m} = (I - \beta\theta k A_h)U^{m-\theta} + \theta k f_h^{m-\theta}. \tag{5.1.28}$$

Jeder der Teilschritte hat die Form eines geschifteten Crank-Nicolson-Schritts, so dass der Gesamtaufwand pro Zeitschritt dem von drei Crank-Nicolson-Schritten entspricht. Für den speziellen Wert

$$\alpha = (1-2\theta)(1-\theta)^{-1} = 0,585786...$$

ist $\alpha\theta = \beta\theta'$, so dass die zu invertierenden Matrizen in den drei Teilschritten übereinstimmen, was z.B. bei der direkten Lösung der Gleichungssysteme ausgenutzt werden kann. Eine genaue Analyse des Abschneidefehlers des FS-Schemas zeigt, dass seine führende

Fehlerkonstante nur wenig größer als die von drei kombinierten Crank-Nicolson-Schritten ist:

$$\tau_k^m = \hat{c}k^2 + \mathcal{O}(k^3), \quad \hat{c}_{FS} \sim \hat{c}_{3\times CN}.$$

Dies bedeutet, dass das FS-Schema gegenüber dem CN-Schema bzgl. Genauigkeit und Aufwand gleichwertig ist, aber über eine höhere Robustheit verfügt. Das FS-Schema hat sich in der Praxis als besonders geeignet zur Behandlung von parabolischen Problemen mit nicht notwendig regulären Daten und geringer natürlicher Eigendissipation erwiesen.

5.1.2 Stabilität und Konvergenz

Wir wollen nun die Stabilität und Konvergenz von Diskretisierungen der Wärmeleitungsgleichung untersuchen. Dabei bedienen wir uns exemplarisch verschiedener Techniken, die alle diskrete Analoga von Analysemethoden beim kontinuierlichen Problem sind.

i) „Maximumprinzipmethode“
Eine einfache, direkte Variante der Maximumprinzipmethode kann bei gewissen expliziten Differenzenschemata angewendet werden. Die Ortsdiskretisierung führe auf eine M-Matrix A_h. Für die explizite Euler-Formel

$$U_h^m = U_h^{m-1} - kA_hU_h^{m-1}$$

gilt dann wegen der Diagonaldominanz von A_h:

$$\begin{aligned} |U_n^m| &= |1 - ka_{nn}||U_n^{m-1}| + k\sum_{\nu\neq n} |a_{n\nu}||U_\nu^{m-1}| \\ &\leq |1 - ka_{nn}||U_n^{m-1}| + ka_{nn}\max_\nu |U_\nu^{m-1}|. \end{aligned}$$

Unter der Schrittweitenbedingung

$$k \leq \max_n\{a_{nn}^{-1}\} \sim ch^2 \tag{5.1.29}$$

folgt daher die L^∞-Stabilität des Verfahrens

$$\max_n |U_n^m| \leq \max_n |U_n^{m-1}| \leq ... \leq \max_n |U_n^0|, \quad m \geq 1. \tag{5.1.30}$$

Im Fall des 5-Punkte-Schemas ist $a_{nn} = 4h^{-2}$, so dass die Stabilitätsbedingung (5.1.29) die Form

$$k \leq \tfrac{1}{4}h^2 \tag{5.1.31}$$

erhält.

Satz 5.1 (Explizites Euler-Verfahren): *Unter der Schrittweitenbedingung (5.1.29) gilt für das explizite Euler-Verfahren die Konvergenzabschätzung*

$$\max_{\bar{Q}_T} |U_h^m - u(\cdot, t_m)| \leq T\Big\{\tfrac{1}{2}k \max_{Q_T} |\partial_t^2 u| + \max_{\bar{Q}_T} |\tau_h^m|\Big\}. \tag{5.1.32}$$

mit dem örtlichen Abschneidefehler $\tau_h^m = \mathcal{O}(h^q)$.

Beweis: Der Fehler $e^m := u(\cdot, t_m) - U^m$ genügt der Gleichung

$$k^{-1}(e^m - e^{m-1}) + A_h e^m = \tau_{h,k}^m$$

mit dem Abschneidefehler

$$\max_{\bar{Q}_T} |\tau_{h,k}^m| \leq \tfrac{1}{2}k \max_{\bar{Q}_T} |\partial_t^2 u| + \max_{\bar{Q}_T} |\tau_h^m| \,.$$

Das bei der Herleitung der Stabilitätsbedingung (5.1.29) verwendete Argument liefert

$$\max_{\bar{\Omega}} |e^m| \leq \max_{\bar{\Omega}} |e^{m-1}| + k \max_{\bar{Q}_T} |\tau_{h,k}^m|$$

Durch Iteration dieser Abschätzung folgt weiter wegen $e^0 = 0$:

$$\begin{aligned} \max_{\bar{\Omega}} |e^m| &\leq k \sum_{\mu=1}^{m} |\tau_{h,k}^{\mu}| \\ &\leq \tfrac{1}{2} t_m k \max_{\bar{Q}_T} |\partial_t^2 u| + t_m \max_{\bar{Q}_T} |\tau_h^m|. \end{aligned}$$

Dies impliziert die behauptete Fehlerabschätzung. Q.E.D.

Eine wichtige Eigenschaft des kontinuierlichen Wärmeleitungsoperators ist seine „inverse Monotonie“, d. h.: Lösungen zu nicht-negativen Anfangsdaten und rechter Seite bleiben nicht-negativ. Diese Eigenschaft überträgt sich auf die diskretisierten Probleme, wenn die Systemmatrix A_h M-Matrix ist.

i) Für das explizite Euler-Verfahren folgt unter der Schrittweitenbedingung (5.1.29) aus $U_n^{m-1} \geq 0$ und $f_n^m \geq 0$ notwendig auch

$$U_n^m = (1 - k a_{nn}) U_n^{m-1} + k \sum_{\nu \neq n} |a_{n\nu}| U_\nu^{m-1} + k f_n^m \geq 0.$$

ii) Für das implizite Euler-Verfahren ist im Falle $U_n^{m-1} \geq 0$ und $f_n^m \geq 0$

$$(I_h + k A_h) U_h^m = U_h^{m-1} + k f_h^m \geq 0.$$

Da mit A_h natürlich auch $I_h + k A_h$ M-Matrix ist, gilt $(I_h + k A_h)^{-1} \geq 0$. Es folgt $U_h^m \geq 0$.

In beiden Fällen ist also auch das diskrete Schema „invers-monoton“. Dies ist i.a. für das Crank-Nicolson-Schema nicht der Fall.

ii) „Von Neumannsche Methode“ (Fourier-Methode)

Wir beschränken uns bei der Beschreibung der auf von Neumann[6] zurückgehenden Analysemethode auf den örtlich eindimensionalen Fall mit $\Omega = (-\pi, \pi)$,

$$\partial_t u - \partial_x^2 u = 0 \quad \text{in } Q_T,$$

mit „periodischen“ Dirichlet-Randbedingungen

$$u(-\pi, t) = u(\pi, t), \quad t \geq 0.$$

In diesem Fall kann die Lösung der ARWA nach trigonometrischen Funktionen entwickelt werden (Fourier-Entwicklung). In komplexer Schreibweise lautet dies

$$u(x,t) = \sum_{\nu=0}^{\infty} a_\nu^0 e^{i\nu x} e^{\nu^2 t}, \tag{5.1.33}$$

mit den Entwicklungskoeffizienten a_ν^0 der Anfangsbedingung. Auf einem äquidistanten Punktgitter $\{x_n = -\pi + nh,\ n = 0, ..., N = 2\pi/h\}$ machen wir für die diskrete Lösung $U_h^m = \{U_n^m,\ n = 0, ..., N\},\ m \geq 0,$ den analogen Entwicklungsansatz

$$U_n^m = \sum_{\nu=0}^{N} a_\nu^m e^{i\nu nh} =: \sum_{\nu=0}^{N} a_\nu^0 \omega_\nu^m e^{i\beta_\nu n} \tag{5.1.34}$$

mit $\beta_\nu := \nu h$ und zu bestimmenden Parametern $\omega_\nu \in \mathbb{C}$. Wir fragen nach der Stabilität für $m \to \infty$ der Differenzendiskretisierung bzgl. der diskreten Spektralnorm

$$\|U_h^m\|_h := \Big(\sum_{n=1}^{N} |a_n^m|^2 \Big)^{1/2}.$$

Die Wirkung des (linearen) Differenzenschemas kann für jede einzelne Fourier-Komponente separat untersucht werden. Gesucht sind Bedingungen an k und h, unter denen $|\omega_\nu| \leq 1$ ist für alle möglichen β_ν. Dann liegt Stabilität vor in dem Sinne, dass

$$\|U_h^m\|_h^2 = \sum_{n=1}^{N} |a_n^0|^2 |\omega_\nu|^{2m} \leq \sum_{n=1}^{N} |a_n^0|^2 = \|U_h^0\|_h^2. \tag{5.1.35}$$

Wir führen diese Analyse wieder exemplarisch für das explizite Euler-Schema durch. Mit $r := kh^{-2}$ gilt

$$U_n^{m+1} = rU_{n-1}^m + (1 - 2r)U_n^m + rU_{n+1}^m.$$

[6] John von Neumann (1903–1957): US-Amerikanischer Mathematiker österreichisch-ungarischer Herkunft; Studium in Budapest, Berlin und Zürich; 1927 Privatdozent in Berlin; arbeitete dann mit Hilbert in Göttingen; ab 1933 Prof. in Princeton (USA); bedeutende Beitrg zur mathematischen Logik, Funktionalanalysis, Quantenmechanik und Spieltheorie; gilt als einer der Väter der Informatik.

Einsetzen von $U_n^m := \omega^m e^{i\beta n}$ ergibt

$$\omega^{m+1} e^{i\beta n} = r\omega^m e^{i\beta(n-1)} + (1-2r)\omega^m e^{i\beta n} + r\omega^m e^{i\beta(n+1)},$$

und nach Vereinfachung

$$\omega = re^{-i\beta} + (1-2r) + re^{i\beta}.$$

Es liegt Stabilität vor, wenn $|\omega| \le 1$ für beliebiges β. Unter Ausnutzung der Beziehungen

$$e^{i\beta} = \cos(\beta) + i\sin(\beta), \quad \cos(\beta) = 1 - 2\sin^2(\tfrac{1}{2}\beta),$$

folgt

$$\begin{aligned}\omega &= r\left(e^{i\beta} + e^{-i\beta}\right) + (1-2r) = r\left(\cos(\beta) + i\sin(\beta) + \cos(\beta) - i\sin(\beta)\right) + (1-2r)\\ &= r\left(2 - 4\sin^2(\tfrac{1}{2}\beta)\right) + (1-2r) = 1 - 4r\sin^2(\tfrac{1}{2}\beta).\end{aligned}$$

Stabilität liegt vor für

$$-1 \le 1 - 4r\sin^2(\tfrac{1}{2}\beta) \le 1 \quad \forall\beta,$$

was äquivalent ist zu

$$r\sin^2(\tfrac{1}{2}\beta) \le \tfrac{1}{2}.$$

Dies führt auf die schon bekannte Stabilitätsbedingung

$$k \le \tfrac{1}{2}h^2. \tag{5.1.36}$$

Die Fourier-Methode kann auch für „exotischere“ Differenzenformeln angewendet werden. Wir demonstrieren dies anhand des klassischen „Du Fort[7]-Frankel[8]-Verfahren“ (1953):

$$\frac{1}{2k}\Big(U_n^{m+1} - U_n^{m-1}\Big) - \frac{1}{h^2}\Big(U_{n-1}^m - (U_n^{m+1} + U_n^{m-1}) + U_{n+1}^m\Big) = 0. \tag{5.1.37}$$

Sein Abschneidefehler verhält sich wie

$$\max_{n,m} |\tau_n^m| = O(k^2/h + k^2 + h^2). \tag{5.1.38}$$

Die von Neumannsche Stabilitätsanalyse liefert für die Verstärkungsfaktoren die Darstellung ($r := k/h^2$)

[7]E.C. Du Fort (????–????): US-Amerikanischer Physiker; Publ. mit S. P. Frankel: Stability conditions in the numerical tratment of parabolic differential equations, Math. Tables and other Aids to Comput. (jetzt Math. Comput.) 7, 135-152 (1953).

[8]Stanley Phillips Frankel (1919–1978): US-Amerikanischer Informatiker; Mitglied der theoretischen Abteilung des „Manhattan Project“ in Los Alamos 1943 (Bau der ersten Atombombe); arbeitete mit dem ENIAC-Computer und in verschiedenen Instituten an der Nutzung mehrerer früher Computer-Systeme; Gruppenleiter am California Institute of Technology (CalTech) in Passadena, USA; Entwicklung der sog. „Monte-Carlo-Methode“ in der statistischen Physik.

$$\omega = \frac{2r\cos(\beta) \pm \sqrt{1 - 4r^2\sin^2(\beta)}}{1 + 2r}. \tag{5.1.39}$$

Dies impliziert , dass $|\omega| \leq 1$ für alle β, d.h.: Das DuFord-Frankel-Schema ist unbedingt stabil. Analog zeigt man, dass das sog. „Richardson-Verfahren“

$$\frac{1}{2k}(U_n^{m+1} - U_n^{m-1}) - \frac{1}{h^2}(U_{n-1}^m - 2U_n^m + U_{n-1}^m) = 0 \tag{5.1.40}$$

unbedingt *instabil* ist. Obwohl es die „optimale“ Konsistenzordnung $O(h^2 + k^2)$ besitzt, ist es also praktisch unbrauchbar. Dies ist nicht verwunderlich, da dieses Schema ein Derivat der Mittelpunktsregel mit dem Stabilitätspolynom $\pi(z, h\lambda)$ und den Wurzeln $z_{1,2} = h\lambda \pm (h^2\lambda^2 + 1)^{1/2}$ ist.

Die von Neumann'sche Fourier-Methode zur Stabilitätsanalyse von Differenzenschemata ist auf den Fall periodischer Dirichlet-Randbedingungen bzw. den Grenzfall von „Ganzraum-Problemen“ ($\Omega = \mathbb{R}^1$) beschränkt und erfordert äquidistante Ortsgitter. Für allgemeinere Ortsdiskretisierungen anwendbar ist die im folgenden präsentierte „Spektralmethode“.

iii) Spektral-Methode:
Die symmetrische, positiv definite Matrix A_h habe die Eigenwerte und zugehörigen (l_2-orthonormierten) Eigenvektoren

$$0 < \lambda_1 \leq ... \leq \lambda_N, \quad \{w^{(n)},\, n = 1, ..., N\}.$$

Jede Gitterfunktion besitzt dann eine Entwicklung der Form

$$U_h^m = \sum_{n=1}^{N} a_n w^{(n)}, \quad a_n = \langle U_h^m, w^{(n)} \rangle.$$

Dabei ist das Skalarprodukt $\langle \cdot, \cdot \rangle$ für eine FD-Diskretisierung im Ort wieder ein diskretes Analogon der kontinuierlichen L^2-Norm,

$$\langle v, w \rangle := \sum_{n=1}^{N} h_n^2 v_n w_n,$$

und für eine FE-Diskretisierung gerade diese: $\langle v, w \rangle := (v, w)_\Omega$. Entsprechend sind die zugehörigen Normen $\|u\| := \langle v, v \rangle^{1/2}$ definiert. Wir analysieren im folgenden isoliert den Zeitschrittfehler im Rahmen der Linienmethode.

Satz 5.2 (Glättungseigenschaft): *Jedes stark A-stabile Einschrittschema vom Typ (5.1.22) der Ordnung r besitzt die Glättungseigenschaft:*

$$\|U_h^m - u_h^m\| \leq c\,\frac{k^r}{t_m^r}\,\|u_h^0\|, \quad m > 0. \tag{5.1.41}$$

Beweis: Nach Voraussetzung ist $\sup_{z\geq 0}|R(-z)|\leq 1$, $\lim_{z\to\infty}|R(-z)|\leq\omega<1$ und

$$|R(-z)-e^{-z}|\leq c|z|^{r+1},\quad 0\leq z\leq 1.$$

O.B.d.A. nehmen wir an, dass $|R(-z)|\leq\omega<1$ für $z\geq 1$. Wir verwenden wieder das Spektralargument von oben. Mit den Eigenwerten $0<\lambda_1\leq ...\leq\lambda_N$ von A_h und einem zugehörigen Orthonormalsystem $\{w^{(n)},\, n=1,...,N\}$ von Eigenvektoren gilt wieder für den Anfangswert

$$u_h^0=\sum_{n=1}^{N}\alpha_n w^{(n)}$$

die Abschätzung ($\tau_n:=k\lambda_n$)

$$\begin{aligned}|U_h^m-u_h^m|^2&=\sum_{n=1}^{N}\alpha_n^2\Big|R(-k\lambda_n)^m-e^{-mk\lambda_n}\Big|^2\\&=\sum_{\tau_n\leq 1}...\;+\;\sum_{\tau_n>1}...\,.\end{aligned}$$

Für die erste Summe rechts gilt mit einem geeigneten $\delta>0$:

$$\begin{aligned}\sum_{\tau_n\leq 1}...&=\sum_{\tau_n\leq 1}\Big|R(-\tau_n)-e^{-\tau_n}\Big|^2\Big|\sum_{\mu=0}^{m-1}R(-\tau_n)^{m-1-\mu}e^{-\mu\tau_n}\Big|^2\alpha_n^2\\&\leq c\sum_{\tau_n\leq 1}\tau_n^{2r+2}m^2e^{-2\delta(m-1)\tau_n}\alpha_n^2\leq cm^{-2r}|u_h^0|^2.\end{aligned}$$

Für die zweite Summe rechts gilt entsprechend mit einem $\delta>0$:

$$\begin{aligned}\sum_{\tau_n>1}...&\leq 2\sum_{\tau_n>1}\alpha_n^2\Big\{|R(-\tau_n)|^{2m}+e^{-2m\tau_n}\Big\}\\&\leq ce^{-\delta m}\sum_{\tau_n>1}\alpha_n^2\leq cm^{-2r}|u_h^0|^2.\end{aligned}$$

Kombination dieser beiden Abschätzungen liefert wegen $m=t_m/k$:

$$|U_h^m-u_h^m|^2\leq c\frac{k^{2r}}{t_m^{2r}}|u_h^0|^2.\tag{5.1.42}$$

Dies vervollständigt den Beweis. Q.E.D.

Das populäre Crank-Nicolson-Schema

$$U_h^m=(I_h+\tfrac{1}{2}kA_h)^{-1}(I_h-\tfrac{1}{2}kA_h)U_h^{m-1}$$

besitzt als nicht *stark* A-stabiles Schema nicht die volle Glättungseigenschaft. Wir wollen diesen Defekt anhand einer Modellbetrachtung erläutern.

Sei

$$U_h^0 = \sum_{n=1}^{N} \alpha_n^0 w^{(n)} \quad \Rightarrow \quad U_h^m = \sum_{n=1}^{N} \alpha_n^0 \Big(\frac{1 - k\lambda_n/2}{1 + k\lambda_n/2}\Big)^m w^{(n)}.$$

Die Lösungskomponente zur höchsten Frequenz $\Lambda = \lambda_N$ verhält sich wie

$$\omega^m = \Big(\frac{1 - k\Lambda/2}{1 + k\Lambda/2}\Big)^m \sim e^{-t_m \Lambda},$$

was dem Abfall der „exakten“ Lösung entspricht.

i) Für $k\Lambda < 2$ ($\Leftrightarrow$ $k \sim h^2$) ist

$$|\omega| \le e^{-\delta}, \quad \delta > 0,$$

was den korrekten exponentiellen Abfall $e^{-\delta m}$ impliziert.

ii) Im Fall $k\Lambda \sim k/h^2 \sim 4/h$ ($\Leftrightarrow$ $k \sim h$) ist

$$\omega \sim -\frac{1 - h/2}{1 + h/2},$$

was oszillierendes Verhalten $(-1)^m e^{-hm}$ impliziert.

Zur Dämpfung dieser Oszillationen in den „hochfrequenten“ Komponenten können folgende Strategien verwendet werden:

a) Mittelbildung:

$$\tilde{U}_h^1 = \tfrac{1}{4}\{U_h^0 + 2U_h^1 + U_h^2\} = \sum_{n=1}^{N} \alpha_n^0 \Big\{\frac{1}{4} + \frac{1}{2}\frac{1 - \frac{1}{2}k\lambda_n}{1 + \frac{1}{2}k\lambda_n} + \frac{1}{4}\Big(\frac{1 - \frac{1}{2}k\lambda_n}{1 + \frac{1}{2}k\lambda_n}\Big)^2\Big\} w^{(n)}.$$

Auswertung des Ausdrucks in der Klammer ergibt

$$\frac{1 + k\lambda_n + \frac{1}{2}k^2\lambda_n^2 + 2 - \frac{1}{2}k^2\lambda_n^2 + 1 - k\lambda_n + \frac{1}{4}k^2\lambda_n^2}{4(1 + \frac{1}{2}k\lambda_n)^2} = \frac{1}{(1 + \frac{1}{2}k\lambda_n)^2}$$

und somit

$$\tilde{U}_h^1 = \sum_{n=1}^{N} \frac{\alpha_n^0}{(1 + \frac{1}{2}k\lambda_n)^2}\, w^{(n)}.$$

b) Euler-Dämpfung: Der Zeitschrittprozess wird mit zwei impliziten Euler-Schritten mit halber Schrittlänge gestartet. Dies ergibt

$$\tilde{U}_h^1 = \sum_{n=1}^{N} \alpha_n^0 \Big(\frac{1}{1 + \frac{1}{2}k\lambda_n}\Big)^2 w^{(n)}.$$

Satz 5.3 (Gedämpftes Crank-Nicolson-Verfahren): *Das durch zwei Euler-Schritte gedämpfte Crank-Nicolson-Verfahren besitzt die Glättungseigenschaft:*

$$\|U_h^m - u_h^m\| \le c\frac{k^2}{t_m^2}\|u_h^0\|. \tag{5.1.43}$$

Beweis: Für $z \ge 0$ gilt

$$\left|e^{-z} - \frac{1-\frac{1}{2}z}{1+\frac{1}{2}z}\right| \le c\frac{z^3}{1+\frac{1}{2}z}, \quad \left|e^{-z} - \frac{1}{1+z}\right| \le c\frac{z^2}{1+z}.$$

Wir verwenden dies in der folgenden Abschätzung:

$$\|U_h^m - u_h^m\|^2 = \sum_{n=1}^{N} \alpha_n^2 \Big(\Big(\frac{1-\frac{1}{2}k\lambda_n}{1+\frac{1}{2}k\lambda_n}\Big)^{m-2} \Big(\frac{1}{1+\frac{1}{2}k\lambda_n}\Big)^2 - e^{-mk\lambda_n}\Big)^2$$

Wir bezeichnen den Inhalt der äußeren Klammer mit σ_n^m und setzen $\tau_n := k\lambda_n$. Es gilt

$$\begin{aligned}
\sigma_n^m &= e^{-(m-2)\tau_n}\Big\{e^{-2\tau_n} - \Big(\frac{1}{1+\frac{1}{2}\tau_n}\Big)^2\Big\} + \Big\{e^{-(m-2)\tau_n} - \Big(\frac{1-\frac{1}{2}\tau_n}{1+\frac{1}{2}\tau_n}\Big)^{m-2}\Big\}\Big(\frac{1}{1+\frac{1}{2}\tau_n}\Big)^2 \\
&= e^{-(m-2)\tau_n}\Big(e^{-\tau_n} - \frac{1}{1+\frac{1}{2}\tau_n}\Big)\Big(e^{-\tau_n} + \frac{1}{1+\frac{1}{2}\tau_n}\Big) \\
&\quad + \Big(e^{-\tau_n} - \frac{1-\frac{1}{2}\tau_n}{1+\frac{1}{2}\tau_n}\Big)\Big(\frac{1}{1+\frac{1}{2}\tau_n}\Big)^2 \sum_{\mu=0}^{m-3} e^{-\mu\tau_n}\Big(\frac{1-\frac{1}{2}\tau_n}{1+\frac{1}{2}\tau_n}\Big)^{n-3-\mu}.
\end{aligned}$$

i) Fall $\tau_n \le 2$:

$$\frac{1-\frac{1}{2}\tau_n}{1+\frac{1}{2}\tau_n} \le e^{-\frac{1}{2}\tau_n}.$$

Dies sieht man wie folgt: Wegen $e^z \ge z$ gilt $-1 + ze^{-z} \le 0$. Die Funktion $f(z) := 1 - z - (1+z)e^{-z}$ hat die Eigenschaften $f(0) = 0$ und $f'(z) = -1 + ze^{-z} \le 0$ und folglich $f(z) \le 0$. Damit erschließen wir

$$\begin{aligned}
|\sigma_n^m| &\le c\Big\{e^{-m\tau_n}\tau_n^2 + \tau_i^3 e^{-m\tau_n/2}\sum_{\mu=0}^{m-3} e^{-\mu\tau_n/2}\Big\} \\
&\le c\Big\{e^{-m\tau_n}\tau_n^2 + \tau_n^3 e^{-m\tau_n/2}\frac{1-e^{-(m-2)\tau_n/2}}{1-e^{-\tau_n/2}}\Big\} \le \frac{c}{m^2} = c\frac{k^2}{t_m^2}.
\end{aligned}$$

ii) Fall $\tau_n > 2$:

$$\Big|\frac{1-\tau_n/2}{1+\tau_n/2}\Big| \le e^{-2/\tau_n}.$$

Damit erschließen wir:

$$|\sigma_n^m| \le c\Big\{ e^{-m\tau_n} + e^{-2(m-2)/\tau_n} \frac{1}{\tau_n^2} \Big\}$$
$$\le c\Big\{ (m\tau_n)^2 e^{-m\tau_n} \frac{1}{m^2} + e^{-2m/\tau_n} \Big(\frac{m}{\tau_n}\Big)^2 \frac{1}{m^2} \Big\} \le c\frac{1}{m^2} = c\frac{k^2}{t_m^2}.$$

Zusammenfassung der Resultate (i) und (ii) liefert nun:

$$\|U_h^m - u_h^m\|^2 \le c\frac{k^4}{t_m^4} \sum_{n=1}^{N} \alpha_n^2. \tag{5.1.44}$$

Dies vervollständigt den Beweis. Q.E.D.

Auch die Spektralmethode ist auf den Fall parabolischer Probleme mit zeitunabhängigen, selbstadjungierten Operatoren wie dem Laplace-Operator Δ beschränkt. Die weitreichendste Analysetechnik ist die sog. „Energie-Methode“ (Hilbertraum-Methode), welche auch für Probleme mit unsymmetrischen Operatoren mit zeitabhängigen Koeffizienten anwendbar ist. Wir demonstrieren diese Technik hier aber nur für die vorliegende Modellsituation.

iv) Energie-Methode:
Wir betrachten das populäre Crank-Nicolson-Schema. Für Funktionen $(v_n)_{n=1}^N$ auf einem äquidistanten Quadratgitter sind

$$(v,w)_h := h^d \sum_{n=1}^{N} v_n w_n\,, \quad \|v\|_h := (v,v)_h^{1/2},$$

diskrete Analoga des L^2-Skalarprodukts und der zugehörigen L^2-Norm.

Satz 5.4 (Crank-Nicolson-Verfahren): *Das Crank-Nicolson-Verfahren hat für hinreichend glatte Lösung u den globalen Diskretisierungsfehler*

$$\max_{\bar{Q}_T} \|u - U_h\|_h \le c(u)\, T\{h^2 + k^2\}\,, \tag{5.1.45}$$

mit einer Konstante $c(u) \approx \max_{\bar{Q}_T}\{|\partial_t^3 u| + a|\nabla^4 u|\}$.

Beweis: Für den Fehler $e^m := u^m - U^m$ gilt

$$k^{-1}(e^m - e^{m-1}) + \tfrac{1}{2}A_h(e^m + e^{m-1}) = \tau_{h,k}^m\,.$$

Multiplikation dieser Identität mit $e^m + e^{m-1}$ und Summation über m ergibt

$$k^{-1}\{\|e^m\|_h^2 - \|e^{m-1}\|_h^2\} + \tfrac{1}{2}(A_h(e^m + e^{m-1}), e^m + e^{m-1})_h = (\tau_{h,k}^m, e^m + e^{m-1})_h\,.$$

Der kleinste Eigenwert von A_h ist $\lambda > 0$. Damit erschließen wir

$$k^{-1}\{\|e^m\|_h^2 - \|e^{m-1}\|_h^2\} + \tfrac{1}{2}\lambda\|e^m + e^{m-1}\|_h^2 \leq \tfrac{1}{2}\lambda\|e^m + e^{m-1}\|_h^2 + \tfrac{1}{2}\lambda^{-1}\|\tau_{h,k}^m\|_h^2\,,$$

bzw.

$$\|e^m\|_h^2 \leq \|e^{m-1}\|_h^2 + \tfrac{1}{2}\lambda^{-1}k\|\tau_{h,k}^m\|_h^2\,.$$

Wir summieren nun über $\mu = m, ..., 1$ und erhalten

$$\|e^m\|_h^2 \leq \|e^0\|_h^2 + \tfrac{1}{2}\lambda^{-1}k\sum_{\mu=1}^{m}\|\tau_{h,k}^\mu\|_h^2\,.$$

Mit $e^0 = 0$ und der obigen Abschätzung für den Abschneidefehler folgt schließlich die Behauptung. Q.E.D.

5.2 FE-Galerkin-Verfahren für parabolische Probleme

Wir diskutieren nun die Rothe-Methode zur Lösung des Problems

$$\partial_t u - \Delta u = f \quad \text{in } Q_T = \Omega \times [0,T], \tag{5.2.46}$$

mit den Nebenbedingungen $u_{|t=0} = u^0$ und $u_{|\partial\Omega} = 0$. Da die folgende Analyse exemplarischen Charakter hat, betrachten wir nur das implizite Euler-Schema. Dieses lautet angewendet auf das kontinuierliche Problem (5.2.46)

$$k_m^{-1}(U^m - U^{m-1}) - \Delta U^m = \bar{f}^m, \quad U^0 := u^0, \tag{5.2.47}$$

wobei die rechte Seite im zeitlichen Mittel ausgewertet wird gemäß

$$\bar{f}^m := k_m^{-1}\int_{t_{m-1}}^{t_m} f(t)\,dt = f^m + \mathcal{O}(k_m).$$

Die Zeitschrittweite $k_m := t_m - t_{m-1}$ darf hier variieren, um eine möglichst gute Anpassung an die Lösungseigenschaften zu erreichen. Mechanismen zur adaptiven Wahl der Zeitschrittweiten auf der Basis von a posteriori Fehlerabschätzungen werden weiter unten diskutiert.

Die einzelnen Zeitschritte seien mit Hilfe eines FE-Verfahrens mit Ansatzräumen $V_h^m \subset V$ auf möglicherweise von Zeit zu Zeit wechselnden Gittern $\mathbf{T}_h^m$ diskretisiert:

$$(U_h^m, \varphi) + k_m(\nabla U_h^m, \nabla\varphi) = (U_h^{m-1}, \varphi) + k_m(\bar{f}^m, \varphi) \quad \forall\varphi \in V_h^m. \tag{5.2.48}$$

Die Varianz der Ortsdiskretisierungen im Verlaufe der Zeititeration ermöglicht die dynamische adaptive Anpassung der Ortsgitter an die momentane Lösungsstruktur. In Ope-

ratorschreibweise lautet das Schema (5.2.47)

$$(I_h^m + k_m \mathcal{A}_h^m) U_h^m = P_h^m U_h^{m-1} + k_m P_h^m \bar{f}^m, \quad U_h^0 = P_h^0 u^0, \tag{5.2.49}$$

mit der L^2-Projektion P_h^m auf V_h^m. Bezüglich der üblichen Knotenbasen $\{\varphi_h^{m,n}, n = 1, ..., N_m = \dim V_h^m\}$ der Räume V_h^m lässt sich dies als lineares Gleichungssystem zur Bestimmung der zugehörigen Knotenwertevektoren $x_h^m \in \mathbb{R}^{N_m}$ schreiben. Dazu führen wir zusätzlich zu Massematrizen, Steifigkeitsmatrizen und Lastvektoren

$$M_h^m := \big((\varphi_h^{m,i}, \varphi_h^{m,j})\big)_{i,j=1}^{N_m}, \quad A_h^m := \big((\nabla\varphi_h^{m,i}, \nabla\varphi_h^{m,j})\big)_{i,j=1}^{N_m}, \quad b_h^m := \big((\bar{f}^m, \varphi_h^{m,j})\big)_{j=1}^{N_m}$$

auf dem Gitter $\mathbf{T}_h^m$ noch Transfermatrizen zwischen den Räumen V_h^{m-1} und V_h^m ein:

$$M_h^{m-1,m} := \big((\varphi_h^{m-1,j}, \varphi_h^{m,n})\big)_{j,n=1}^{N_{m-1},N_m}.$$

Damit schreibt sich

$$(U_h^{m-1}, \varphi_h^{m,n}) = \sum_{j=1}^{N_{m-1}} x_j^{m-1} (\varphi_h^{m-1,j}, \varphi_h^{m,n}) = M_h^{m-1,m} x_h^{m-1}$$

und folglich

$$(M_h^m + k_m A_h^m) x_h^m = M_h^{m-1,m} x_h^{m-1} + k_m M_h^m b_h^m. \tag{5.2.50}$$

Wir wollen dieses Verfahren im folgenden im Hinblick auf Stabilität, Konvergenz sowie a priori und a posteriori Fehlerabschätzung untersuchen.

5.2.1 A priori Konvergenzabschätzungen

Der natürliche Ansatz zur Analyse von FE-Diskretisierungen ist die „Energie-Methode“. Wir geben zunächst einen einfachen Beweis für das implizite Euler-Verfahren unter realistischen Annahmen an die Regularität der Lösung. Wir setzen (Die Zellen der Zerlegung $\mathbb{T}_h^m$ werden ab jetzt mit K bezeichnet.)

$$h_m := \max_{K \in \mathbf{T}_h^m} \operatorname{diam}(K), \quad k = \max_{1 \le m \le M} k_m$$

und $e_h^m := U_h^m - u^m$ mit $u^m := u(\cdot, t_m)$.

Satz 5.5 (Implizites Euler-Verfahren): *Für das implizite Euler-Schema in Verbindung mit einer FE-Diskretisierung 2. Ordnung gelten die folgenden Fehlerabschätzungen:*

i) Für beliebig variierende Ortsdiskretisierung:

$$\max_{1 \le m \le M} \|e_h^m\| \le c\, T^{1/2} \max_{0 \le m \le M} \Big\{ \frac{h_m^2}{k_m^{1/2}} \|\nabla^2 u^m\| \Big\} + c \Big(\sum_{m=1}^{M} k_m^2 \int_{t_{m-1}}^{t_m} \|\nabla \partial_t u\|^2 \, dt \Big)^{1/2}; \tag{5.2.51}$$

ii) Im Spezialfall $V_h^m = V_h$ gleich für alle m :

$$\max_{1\le m\le M}\|e_h^m\| \le c\,T^{1/2}\max_{0\le m\le M}\{h_m^2\|\nabla^2 u^m\|\} + c\Big(\sum_{m=1}^{M} k_m^2\int_{t_{m-1}}^{t_m}\|\nabla\partial_t u\|^2\,dt\Big)^{1/2}. \tag{5.2.52}$$

Beweis: Wir bezeichnen mit $R_h^m u \in V_h^m$ die „elliptische“ Ritz-Projektion der Lösung u^m zum Zeitlevel t_m auf den Finite-Elemente-Raum V_h^m, definiert durch

$$(\nabla R_h^m v, \nabla\varphi_h) = (\nabla v, \nabla\varphi_h) \quad \forall\varphi_h \in V_h^m. \tag{5.2.53}$$

Für deren Fehler gilt

$$\|v - R_h^m v\| + h_m\|\nabla(v - R_h^m v)\| \le c h_m^2\|\nabla^2 v\|. \tag{5.2.54}$$

Wir betrachten nun zunächst die Differenz $\eta_h^m := U_h^m - R_h^m u^m$. Für beliebiges $\varphi_h \in V_h^m$ ist dann unter Ausnutzung der Identität

$$(U_h^m - U_h^{m-1}, \varphi_h) + k_m(\nabla U_h^m, \nabla\varphi_h) = k_m(\bar f^m, \varphi_h)$$

und der Projektionseigenschaft von R_h^m:

$$\begin{aligned}(\eta_h^m - \eta_h^{m-1}, \varphi_h) &+ k_m(\nabla\eta_h^m, \nabla\varphi_h)\\ &= k_m(\bar f^m, \varphi_h) - (R_h^m u^m - R_h^{m-1}u^{m-1}, \varphi_h) - k_m(\nabla R_h^m u^m, \nabla\varphi_h)\\ &= k_m(\bar f^m, \varphi_h) - (R_h^m u^m - R_h^{m-1}u^{m-1}, \varphi_h) - k_m(\nabla u^m, \nabla\varphi_h).\end{aligned}$$

Wir setzen nun $\varphi_h := \eta_h^m$ und erhalten mit Hilfe der Identität $(a-b)a = \frac12 a^2 - \frac12 b^2 + \frac12(a-b)^2$ die Beziehung

$$\begin{aligned}\tfrac12\|\eta_h^m\|^2 - \tfrac12\|\eta_h^{m-1}\|^2 &+ \tfrac12\|\eta_h^m - \eta_h^{m-1}\|^2 + k_m\|\nabla\eta_h^m\|^2\\ &= k_m(\bar f^m, \eta_h^m) - (R_h^m u^m - R_h^{m-1}u^{m-1}, \eta_h^m) - k_m(\nabla u^m, \nabla\eta_h^m)\\ &= k_m(\bar f^m, \eta_h^m) - (u^m - u^{m-1}, \eta_h^m) - k_m(\nabla u^m, \nabla\eta_h^m)\\ &\quad + (u^m - R_h^m u^m, \eta_h^m) - (u^{m-1} - R_h^{m-1}u^{m-1}, \eta_h^{m-1})\\ &\quad + (u^{m-1} - R_h^{m-1}u^{m-1}, \eta_h^{m-1} - \eta_h^m).\end{aligned} \tag{5.2.55}$$

Weiter haben wir

$$\begin{aligned}k_m(\bar f^m, \eta_h^m) - (u^m - u^{m-1}, \eta_h^m) &- k_m(\nabla u^m, \nabla\eta_h^m)\\ &= \int_{t_{m-1}}^{t_m}(f - \partial_t u, \eta_h^m)\,dt - k_m(\nabla u^m, \nabla\eta_h^m)\\ &= \int_{t_{m-1}}^{t_m}(\nabla u, \nabla\eta_h^m)\,dt - k_m(\nabla u^m, \nabla\eta_h^m)\\ &= \int_{t_{m-1}}^{t_m}(t - t_{m-1})(\nabla\partial_t u, \nabla\eta_h^m)\,dt\\ &\le k_m\|\nabla\eta_h^m\|^2 + \tfrac14 k_m^2\int_{t_{m-1}}^{t_m}\|\partial_t\nabla u\|^2\,dt\end{aligned} \tag{5.2.56}$$

sowie

$$(u^{m-1} - R_h^{m-1}u^{m-1}, \eta_h^{m-1} - \eta_h^m) \le \tfrac{1}{2}\|\eta_h^{m-1} - \eta_h^m\|^2 + ch_m^4\|\nabla^2 u^{m-1}\|^2 \tag{5.2.57}$$

oder

$$(u^{m-1} - R_h^{m-1}u^{m-1}, \eta_h^{m-1} - \eta_h^m) \le k_m^{-1}\|\eta_h^{m-1} - \eta_h^m\|^2 + ck_m h_m^4\|\nabla^2 u^{m-1}\|^2 \tag{5.2.58}$$

i) Wir betrachten zunächst den Fall allgemein variierender Ortsdiskretisierung. Kombination der Beziehungen (5.2.55), (5.2.56) und (5.2.57) und Absorption von Termen in die linke Seite ergibt

$$\begin{aligned}\tfrac{1}{2}\|\eta_h^m\|^2 - \tfrac{1}{2}\|\eta_h^{m-1}\|^2 \le{}& (u^m - R_h^m u^m, \eta_h^m) - (u^{m-1} - R_h^{m-1}u^{m-1}, \eta_h^{m-1}) \\ &+ \tfrac{1}{4}k_m^2 \int_{t_{m-1}}^{t_m} \|\partial_t \nabla u\|^2\, dt + ch_m^4\|\nabla^2 u^{m-1}\|^2.\end{aligned}$$

Wir wenden diese Abschätzung rekursiv für $m, m-1, ..., 1$ an und finden

$$\begin{aligned}\|\eta_h^m\|^2 \le{}& \|\eta_h^0\|^2 + 2(u^m - R_h^m u^m, \eta_h^m) - 2(u^0 - R_h^0 u^0, \eta_h^0) \\ &+ c\sum_{\mu=1}^{m}\Big\{k_\mu^2 \int_{t_{\mu-1}}^{t_\mu} \|\partial_t \nabla u\|^2\, dt + ch_\mu^4\|\nabla^2 u^{\mu-1}\|^2\Big\}\end{aligned}$$

bzw.

$$\begin{aligned}\|\eta_h^m\|^2 \le{}& \|\eta_h^0\|^2 + \tfrac{1}{2}\|\eta_h^m\|^2 + \tfrac{1}{2}\|u^m - R_h^m u^m\|^2 - (u^0 - R_h^0 u^0, \eta_h^0) \\ &+ c\sum_{\mu=1}^{m} k_\mu^2 \int_{t_{\mu-1}}^{t_\mu} \|\partial_t \nabla u\|^2\, dt + ct_m \max_{0\le\mu\le m}\{k_m^{-1}h_\mu^4\|\nabla^2 u^\mu\|^2\}.\end{aligned}$$

Mit Hilfe der Abschätzung

$$\|u^\mu - R_h^\mu u^\mu\| + \|u^\mu - P_h^\mu u^\mu\| \le ch_\mu^2\|\nabla^2 u^\mu\|, \quad \mu = 1, ..., m, \tag{5.2.59}$$

folgt

$$\|e_h^m\| \le \|\eta_h^m\| + \|u^m - R_h^m u^m\| \le \|\eta_h^m\| + ch_m^2\|\nabla^2 u^m\|, \tag{5.2.60}$$

$$\|\eta_h^0\| \le \|u^0 - R_h^0 u^0\| + \|u^0 - P_h^0 u^0\| \le ch_0^2\|\nabla^2 u^0\| \tag{5.2.61}$$

und damit schließlich die Fehlerabschätzung (5.2.51):

$$\|e_h^m\|^2 \le ct_m \max_{0\le\mu\le m}\{k_\mu^{-1}h_\mu^4\|\nabla^2 u^\mu\|^2\} + c\sum_{\mu=1}^{m} k_\mu^2 \int_{t_{\mu-1}}^{t_\mu} \|\nabla\partial_t u\|^2\, dt. \tag{5.2.62}$$

ii) Wir nehmen nun an, dass $V_h^m = V_h$ bzw. $R_h^m = R_h$ für $m = 1, ..., M$. Kombination der Beziehungen (5.2.55), (5.2.56) und (5.2.58) und Absorption von Termen in die linke Seite ergibt:

$$\tfrac{1}{2}\|\eta_h^m\|^2 - \tfrac{1}{2}\|\eta_h^{m-1}\|^2 \le (u^m - R_h^m u^m, \eta_h^m) - (u^{m-1} - R_h^{m-1}u^{m-1}, \eta_h^{m-1}) + \tfrac{1}{4}k_m^2 \int_{t_{m-1}}^{t_m} \|\partial_t \nabla u\|^2 \, dt + k_m^{-1}\|\eta_h^{m-1} - \eta_h^m\|^2 + ck_m h_m^4 \|\nabla^2 u^{m-1}\|^2.$$

Wir wenden diese Abschätzung rekursiv für $m, m-1, ..., 1$ an und finden:

$$\|\eta_h^m\|^2 \le \|\eta_h^0\|^2 + 2(u^m - R_h^m u^m, \eta_h^m) - 2(u^0 - R_h^0 u^0, \eta_h^0) + c\sum_{\mu=1}^{m} \left\{ k_\mu^2 \int_{t_{\mu-1}}^{t_\mu} \|\partial_t \nabla u\|^2 \, dt + k_m h_\mu^4 \|\nabla^2 u^{\mu-1}\|^2 + k_\mu^{-1}\|\eta_h^{\mu-1} - \eta_h^\mu\|^2 \right\}$$

bzw. mit den Abschätzungen (5.2.59), (5.2.60) und (5.2.61),

$$\begin{aligned} \|e_h^m\|^2 \le{}& ct_m \max_{0\le\mu\le m} \{h_\mu^2 \|\nabla^2 u^\mu\|^2\} + \sum_{\mu=1}^{m} k_\mu^{-1}\|\eta_h^{\mu-1} - \eta_h^\mu\|^2 \\ &+ c\sum_{\mu=1}^{m} k_\mu^2 \int_{t_{\mu-1}}^{t_\mu} \|\partial_t \nabla u\|^2 \, dt. \end{aligned} \tag{5.2.63}$$

Im letzten Schritt schätzen wir die mittlere Summe rechts ab. Jetzt kann $\varphi_h = \varphi_h^m := \eta_h^m - \eta_h^{m-1} \in V_h$ als Testfunktion verwendet werden, und wir erhalten wie oben

$$\begin{aligned} &\|\varphi_h^m\|^2 + \tfrac{1}{2}k_m\|\nabla\eta_h^m\|^2 - \tfrac{1}{2}k_m\|\nabla\eta_h^{m-1}\|^2 + \tfrac{1}{2}k_m\|\nabla\varphi_h^m\|^2 \\ &= k_m(\bar{f}^m, \varphi_h^m) - (R_h^m u^m - R_h^{m-1}u^{m-1}, \varphi_h^m) - k_m(\nabla R_h^m u^m, \nabla\varphi_h^m) \\ &= k_m(\bar{f}^m, \varphi_h^m) - (u^m - u^{m-1}, \varphi_h^m) - k_m(\nabla u^m, \nabla\varphi_h^m) \\ &+ (u^m - u^{m-1} - R_h(u^m + u^{m-1}), \varphi_h^m). \end{aligned}$$

Weiter haben wir

$$\begin{aligned} k_m(\bar{f}^m, \varphi_h^m) - (u^m - u^{m-1}, \varphi_h^m) - k_m(\nabla u^m, \nabla\varphi_h^m) &= \int_{t_{m-1}}^{t_m} (f - \partial_t u, \varphi_h^m) \, dt - k_m(\nabla u^m, \nabla\varphi_h^m) \\ &= \int_{t_{m-1}}^{t_m} (\nabla u, \nabla\varphi_h^m) \, dt - k_m(\nabla u^m, \nabla\varphi_h^m) \\ &= \int_{t_{m-1}}^{t_m} (t - t_{m-1})(\nabla\partial_t u, \nabla\varphi_h^m) \, dt \\ &\le \tfrac{1}{2}k_m\|\nabla\varphi_h^m\|^2 + \tfrac{1}{2}k_m^2 \int_{t_{m-1}}^{t_m} \|\partial_t\nabla u\|^2 \, dt \end{aligned}$$

sowie

$$\begin{aligned} (u^m - u^{m-1} - R_h(u^m - u^{m-1}), \varphi_h^m) &\le \tfrac{1}{4}\|\varphi_h^m\|^2 + ch_m^2\|\nabla(u^m - u^{m-1})\|^2 \\ &\le \tfrac{1}{4}\|\varphi_h^m\|^2 + ch_m^2 k_m \int_{t_{m-1}}^{t_m} \|\nabla\partial_t u\|^2 \, dt. \end{aligned}$$

Kombination dieser Abschätzungen und Absorption von Termen in die linke Seite ergibt

$$\|\varphi_h^m\|^2 + k_m\|\nabla\eta_h^m\|^2 - k_m\|\nabla\eta_h^{m-1}\|^2 \le \{k_m^2 + ch_m^2 k_m\}\int_{t_{m-1}}^{t_m} \|\nabla\partial_t u\|^2\,dt.$$

Wir wenden diese Abschätzung wieder rekursiv für $m, m-1, ..., 1$ an und finden

$$\sum_{\mu=1}^{m} k_\mu^{-1}\|\varphi_h^m\|^2 + \|\nabla\eta_h^m\|^2 \le \qquad \|\nabla\eta_h^0\|^2 + c\sum_{\mu=1}^{m}\{k_\mu + h_\mu^2\}\int_{t_{\mu-1}}^{t_\mu} \|\partial_t\nabla u\|^2\,dt.$$

Mit $\|\nabla\eta_h^0\|^2 \le ch_0^4\|\nabla^2 u^0\|^2$ folgt schließlich

$$\sum_{\mu=1}^{m} k_\mu^{-1}\|\eta_h^\mu - \eta_h^{\mu-1}\|^2 \le ch_0^4\|\nabla^2 u^0\|^2 + c\sum_{\mu=1}^{m}\{k_\mu + h_\mu^2\}\int_{t_{\mu-1}}^{t_\mu} \|\partial_t\nabla u\|^2\,dt. \tag{5.2.64}$$

Wir setzen dies in (5.2.63) ein und erhalten die behauptete Abschätzung (5.2.52):

$$\|e_h^m\|^2 \le ct_m \max_{0\le\mu\le m}\{h_\mu^2\|\nabla^2 u^\mu\|^2\} + \sum_{\mu=1}^{m} k_\mu^{-1}\|\eta_h^{\mu-1} - \eta_h^\mu\|^2 + c\sum_{\mu=1}^{m} k_\mu^2\int_{t_{\mu-1}}^{t_\mu} \|\partial_t\nabla u\|^2\,dt,$$

was den Beweis vervollständigt. Q.E.D.

Die Konvergenzordnung in (5.2.51) ist nicht optimal. Unter der Bedingung $h_m^{4/3} \le ck_m$ ergibt sich aber die zeit-optimale Konvergenzordnung $\mathcal{O}(k_m)$. Das optimale Resultat (5.2.52) lässt sich auch unter den weniger einschränkenden Bedingungen $V_h^{m-1} \subset V_h^m$ oder $h_m^2 k_m^{-1} \le \kappa$ hinreichend klein beweisen.

Hilfssatz 5.2 (A-priori Schranke): *Für die Lösung der ARWA (5.2.46) gilt die a priori Abschätzung*

$$\max_{[0,T]}\|\nabla^2 u\| + \Big(T^{-1}\int_0^T \|\nabla\partial_t u\|^2\,dt\Big)^{1/2} \le c\,\|\nabla^2 u^0\| + c\max_{[0,T]}\{\|f\| + \|\partial_t f\|\}. \tag{5.2.65}$$

Beweis: Der Beweis verwendet die „Energie-Technik“, wird hier aber nicht ausgeführt. Q.E.D.

Wir wollen noch die Frage nach der „inversen Monotonie“ der Orts-Zeit-Diskretisierung diskutieren. Unter bestimmten Bedingungen an das Ortsgitter (z.B. alle Innenwinkel einer Triangulierung $\omega \le \pi/2$) ist die Steifigkeitsmatrix A_h eine M-Matrix. Die Systemmatrix $M_h + kA_h$ muss aber nicht automatisch M-Matrix sein. Um dies dennoch sicherzustellen, werden die Elemente von M_h,

$$m_{ij} = (\varphi_h^{m,i}, \varphi_h^{m,j}),$$

mit Hilfe der Trapezregel ausgewertet. Bei stückweise linearen Ansätzen liefert dies eine Diagonalmatrix $\tilde{M}_h = M_h + \mathcal{O}(h^2)$ mit positiven Diagonalelementen, so dass $\tilde{M}_h + kA_h$ M-Matrix wird. Dieser „Mass-Lumping“ genannte Prozess erhält die Konvergenzordnung des Gesamtschemas und stellt seine „inverse Monotonie“ sicher (Übungsaufgabe).

5.2.2 Fehlerkontrolle und Schrittweitensteuerung

Zur Herleitung von *a posteriori* Fehlerabschätzungen erweist sich eine globale Betrachtung simultan in Ort und Zeit (ohne die bisherige Aufspaltung in Orts- und Zeitdiskretisierung) als angemessen. Wir führen dazu das Konzept der „unstetigen Galerkin-Verfahren“ für parabolische ARWAn ein, als deren einfachster Spezialfall das implizite Euler-Verfahren (5.2.48) erscheinen wird. Die Vorgehensweise entspricht der bereits von den gewöhnlichen AWAn her bekannten, ergänzt um die Aspekte der Ortsdiskretisierung.

Ausgehend von oben formulierten Diskretisierungen $\mathbf{T}_h^m = \{K_n^m\}$ des Ortsgebiets $\bar{\Omega}$ und $0 = t_0 < ... < t_\mu < ... < t_M = T$ des Zeitintervalls $I = [0,T]$ führen wir die folgenden Bezeichnungen ein (Man beachte, dass die Zeitschrittweite nicht bzgl. des Orts variiert.):

$$
\begin{aligned}
&I_m := (t_{m-1}, t_m], \quad k_m = t_m - t_{m-1},\\
&k := \max_{m=1,...,M} k_m, \quad h_m := \max_{K\in\mathbf{T}_h^m} h_K, \quad h := \max_{m=1,...,M} h_m,\\
&v^{m\pm} := \lim_{s\downarrow 0} v(\cdot, t_m \pm s), \quad [v]^m := v^{m+} - v^{m-},\\
&Q_n^m := K_n^m \times I_m, \quad \partial Q_n^m := \partial K_n^m \times I_m, \quad Q^m := \Omega \times I_m.
\end{aligned}
$$

Die ARWA (5.2.46) lässt sich äquivalent schreiben in der Form

$$
\sum_{m=1}^{M} \left\{ \int_{I_m} \{(\partial_t u, \varphi) + (\nabla u \nabla \varphi)\}\, dt + ([u]^{m-1}, \varphi^{(m-1)+}) \right\} = \int_I (f, \varphi)\, dt \tag{5.2.66}
$$

für beliebige in der Zeit stetige Testfunktion $\varphi(\cdot,t) \in V$, wobei die Anfangsbedingung durch die Setzung $u^{0-} := u^0$ berücksichtigt ist. Jede (glatte) Lösung von (5.2.46) erfüllt offenbar die Beziehung (5.2.66), und umgekehrt muss jede Lösung von (5.2.66) in den Teilintervallen I_m der Wärmeleitungsgleichung genügen und bei t_m auch stetig sein. Damit folgt dann wieder, dass es sich um eine Lösung der ARWA handeln muss. Dieses Problem wird nun mit einem Galerkin-Ansatz auf dem ganzen Orts-Zeit-Zylinder Q_T diskretisiert. Dazu führen wir die folgenden Finite–Elemente–Räume ein:

$$
V_h = \{v : \bar{Q}_T \to \mathbb{R} |\ v_{t\in I_m}(\cdot,t) \in V_h^m, v_{t\in I_m}(x,\cdot) \in P_r(I_m),\ (x,t) \in Q_T\}.
$$

Die Funktionen in V_h sind also bzgl. des Ortes stückweise in V_h^m (d.h. linear oder bilinear und stetig) und bzgl. der Zeit stückweise polynomial vom Grad r (und unstetig). Der Galerkin-Ansatz sucht dann Approximationen $U_h \in V_h$ zu bestimmen durch die Vorschrift

$$
\sum_{m=1}^{M} \left\{ \int_{I_m} \{(\partial_t U_h, \varphi_h) + (\nabla U_h, \nabla \varphi_h)\}\, dt + ([U_h]^{m-1}, \varphi_h^{(m-1)+}) \right\} = \int_I (f, \varphi_h)\, dt \tag{5.2.67}
$$

für beliebige Testfunktion $\varphi_h \in V_h$ mit dem Anfangswert $U_h^{0-} = P_h^0 u^0$ (P_h^0 die L^2-Projektion auf V_h^0). Da die Testfunktionen unstetig in der Zeit sein dürfen, zerfällt dieses formal zeitlich global gekoppelte System in lokale Teilprobleme auf jedem Zeitstreifen

$Q^m = \Omega \times I_m$. Dieses Schema wird „unstetiges Galerkin-Verfahren“ (abgekürzt „$dG(r)$-Verfahren“) genannt. Wir wollen im folgenden nur den einfachsten Fall $r = 0$, d.h. das $dG(0)$-Verfahren, betrachten. In diesem Fall reduziert sich das globale Schema (5.2.67) auf die folgende Sequenz von lokalen Gleichungen auf den Zeitintervallen I_m, $m = 1, ..., M$:

$$\int_{I_m} \{(\partial_t U_h, \varphi_h) + (\nabla U_h, \nabla \varphi_h)\}\, dt + ([U_h]^{m-1}, \varphi_h) = \int_{I_m} (f, \varphi_h)\, dt \tag{5.2.68}$$

für beliebiges $\varphi_h \in V_h^m$. Mit der Setzung $U_h^m := U_h^{m,-}$ ergibt sich wegen $\partial_t U_h \equiv 0$ auf I_m:

$$k_m(\nabla U_h^m, \nabla \varphi_h) + (U_h^m - U_h^{m-1}, \varphi_h) = k_m(\bar{f}^m, \varphi_h) \tag{5.2.69}$$

für beliebiges $\varphi_h \in V_h^m$. Dies ist gerade das implizite Euler-Verfahren (5.2.48), welches sich also in diesem Rahmen als $dG(0)$-Verfahren interpretieren lässt. Diese Sicht ändert zwar nichts am Verfahren selbst, bietet jedoch einen systematischen Zugang zu seiner Fehleranalyse.

Bemerkung 5.1: Die $dG(r)$-Verfahren höheren Grades $r \geq 1$ entsprechen keinem der oben diskutierten Zeitschritt-Schemata, sie sind vielmehr Varianten gewisser impliziter Runge[9]-Kutta[10] -Verfahren. Wir bemerken aber, dass sich auch das populäre Crank-Nicolson-Verfahren in den Rahmen der Galerkin-Verfahren einordnen lässt. Dazu macht man einen modifizierten Ansatz mit bzgl. der Zeit stückweise *linearen*, aber diesmal *global stetigen* Funktionen. Der Raum der Testfunktionen ist dagegen derselbe wie beim $dG(0)$-Verfahren. Dies wird dann ein „Petrow-Galerkin-Verfahren“ (sog. $cG(1)$-Verfahren), welches sich ähnlich analysieren lässt wie das $dG(0)$-Verfahren:

$$(U_h^m, \varphi_h) + \tfrac{1}{2} k_m(\nabla(U_h^m + U_h^{m-1}), \nabla \varphi_h) = (U_h^{m-1}, \varphi_h) + k_m(\bar{f}^m, \varphi_h), \tag{5.2.70}$$

für beliebige Testfunktion $\varphi_h \in V_h^m$.

Wir wollen jetzt eine a posteriori Fehlerabschätzung für das $dG(0)$-Verfahren ableiten, welche als Basis für eine simultane Anpassung des Ortsgitters und der Zeitschrittweite an den Lösungsverlauf dienen kann. Dazu setzen wir $e_h := U_h - u$. Ein typisches Beispiel ist die Kontrolle des L^2-Fehlers zum Endzeitpunkt $J(e_h) := \|e_h^{N+}\|$.

Satz 5.6 (A posteriori Fehlerschranke): *Für das $dG(0)$-Verfahren gilt bei Kontrolle des örtlichen L^2-Fehlers zum Endzeitpunkt, $J(e_h) := \|e_h^{N+}\|$, die a posteriori Fehlerabschätzung*

$$J(e_h) \leq \eta(U_h) := c_i \sum_{\mu=1}^{m} \sum_{K_n^\mu \in \mathbf{T}_h^\mu} \rho_n^\mu(U_h)\, \omega_n^\mu(z), \tag{5.2.71}$$

[9] *Carle David Tolmé Runge (1856–1927): Deutscher Mathematiker; Prof. in Hannover und Göttingen; Beiträge zur Spektraltheorie mit Anwendungen in der Atom-Physik.*

[10] *Martin Wilhelm Kutta (1867–1944): Deutscher Mathematiker; Prof. in Stuttgart; Beiträge zur Aerodynamik und zur Numerik von Differentialgleichungen.*

mit den Residuentermen

$$\rho_n^\mu(U_h) := \|R(U_h)\|_{Q_n^\mu} + h_n^{-1/2}\|[\partial_n U_h]\|_{\partial K\times I^\mu} + k_\mu^{-1/2}\|[U_h]^{\mu-1}\|_{K_n^\mu},$$

$R(U_h) := f - \partial_t U_h + \Delta U_h$ *, und einer „Interpolationskonstante“* c_i *. Mit der Lösung* z *des „dualen Problems“*

$$-\partial_t z - \Delta z = 0 \quad \textit{in } \Omega\times[0,t_m], \quad z_{|\partial\Omega} = 0, \;\; z_{|t=t_m} = \|e^{m+}\|^{-1}e^{m+}. \tag{5.2.72}$$

haben die Gewichte $\omega_n^\mu(z)$ *die Gestalt:*

$$\omega_n^\mu(z) := k_\mu\|\partial_t z\|_{Q_n^\mu} + k_\mu^2\|\partial_t^2 z\|_{Q_n^\mu} + h_n^2\|\nabla^2 z\|_{Q_n^\mu}.$$

Beweis: i) Das erste „Standbein“ des Beweises ist ein (parabolisches) Dualitätsargument. Wir führen zunächst für ein festes $t_m \in I$ das (kontinuierliche) „duale Problem“ ein:

$$-\partial_t z - \Delta z = \psi \quad \text{in } \; Q_m := \Omega\times[0,t_m], \quad z_{|\partial\Omega}, \;\; z_{t=t_m} = \chi, \tag{5.2.73}$$

bzw. in semi-variationeller Formulierung

$$-(\partial_t z,\varphi) + (\nabla z,\nabla\varphi) = (\psi,\varphi) \quad \forall\varphi\in V. \tag{5.2.74}$$

Umschreiben in die Form (5.2.66) ergibt

$$\sum_{\mu=1}^m \Big\{ \int_{I_\mu} \{-(\varphi,\partial_t z) + (\nabla\varphi,\nabla z)\}\,dt + (\varphi^{\mu+},[z]^\mu)\Big\} = \int_{[0,t_m]} (\varphi,\psi)\,dt$$

für beliebige in der Zeit stetige Testfunktion $\varphi(\cdot,t)\in V$, wobei die Anfangsbedingung wieder durch die Vorschrift $z^{m+} := \chi$ berücksichtigt ist. Durch partielle Integration in der Zeit wird dies umgeformt zu

$$\begin{aligned}\sum_{\mu=1}^m \Big\{ \int_{I_\mu} \{(\partial_t\varphi, z) + (\nabla\varphi,\nabla z)\}\,dt + ([\varphi]^{\mu-1}, z^{(\mu-1)+})\Big\} \\ = (\varphi^{m+},\chi) + \int_{[0,t_m]} (\varphi,\psi)\,dt\end{aligned} \tag{5.2.75}$$

für beliebige in der Zeit (stückweise) differenzierbare Testfunktion $\varphi(\cdot,t)\in V$.

ii) Das zweite „Standbein“ des Beweises ist die Galerkin-Orthogonalität des Fehlers des $dG(0)$-Verfahrens. Durch Vergleich der beiden Gleichungen (5.2.66) für u und (5.2.67) für U_h erhalten wir für den Fehler $e_h := U_h - u$ die „Galerkin-Orthogonalität“:

$$\sum_{m=1}^M \Big\{ \int_{I_m} \{(\partial_t e_h,\varphi_h) + (\nabla e_h\nabla\varphi_h)\},dt + ([e_h]^{m-1},\varphi_h^{(m-1)+})\Big\} = 0 \tag{5.2.76}$$

für beliebige stetige Testfunktion $\varphi_h(\cdot,t)\in V_h$. Wir wählen nun in (5.2.75) die spezielle

Testfunktion $\varphi := e_h$ und erhalten:

$$\sum_{\mu=1}^{m} \Big\{ \int_{I_\mu} \{(\partial_t e_h, z) + (\nabla e_h, \nabla z)\}\, dt + ([e_h]^{\mu-1}, z^{(\mu-1)+}) \Big\} = (e_h^{m+}, \chi) + \int_{[0,t_m]} (e_h, \psi)\, dt.$$

Unter Ausnutzung von (5.2.76) erhalten wir damit die allgemeine Fehleridentität:

$$\begin{aligned}(e_h^{m+}, \chi) + \int_{[0,t_m]} (e_h, \psi)\, dt = \sum_{\mu=1}^{m} \Big\{ \int_{I_\mu} \{(\partial_t e_h, z - z_h) + (\nabla e_h, \nabla(z - z_h))\}\, dt \\ + ([e_h]^{\mu-1}, (z - z_h)^{(\mu-1)+}) \Big\}\end{aligned} \tag{5.2.77}$$

mit einer beliebigen Approximation $z_h \in V_h$ zur dualen Lösung z. Wir setzen nun im dualen Problem $t_m = T$, $\psi := 0$ und $\chi := \|e^{N+}\|^{-1} e^{N+}$. Aus der allgemeinen Fehleridentität (5.2.77) folgt

$$\|e_h^{m+}\| = \sum_{\mu=1}^{m} \Big\{ \int_{I_\mu} \{(\partial_t e_h, z - z_h) + (\nabla e_h, \nabla(z - z_h))\}\, dt + ([e_h]^{\mu-1}, (z - z_h)^{(\mu-1)+}) \Big\} \tag{5.2.78}$$

mit einer beliebigen Approximation $z_h \in V_h$ zur dualen Lösung z. Wir nutzen nun die Lösungseigenschaften von u und integrieren auf jeder Zelle K_n^m partiell bzgl. des Orts:

$$\begin{aligned}\|e_h^{m+}\| = \sum_{\mu=1}^{m} \sum_{K_n^\mu \in \mathbf{T}_h^\mu} \Big\{ \int_{I_\mu} \{(\partial_t u - \partial_t U_h, z - z_h)_{K_n^\mu} - (\Delta u - \Delta U_h, z - z_h)_{K_n^\mu} \\ - (\partial_n (U_h - u), z - z_h)_{\partial K_n^\mu}\}\, dt + \ ([U_h - u]^{\mu-1}, (z - z_h)^{(\mu-1)+})_{K_n^\mu} \Big\}.\end{aligned}$$

Dies ergibt

$$\begin{aligned}\|e_h^{m+}\| = \sum_{\mu=1}^{m} \sum_{K_n^\mu \in \mathbf{T}_h^\mu} \Big\{ \int_{I_\mu} \{(R(U_h), z - z_h)_{K_n^\mu} - \tfrac{1}{2}([\partial_n U_h], z - z_h)_{\partial K_n^\mu}\}\, dt \\ + \ ([U_h]^{\mu-1}, (z - z_h)^{(\mu-1)+})_{K_n^\mu} \Big\},\end{aligned}$$

mit dem Residuum $R(U_h) := f - \partial_t U_h + \Delta U_h$ und dem Sprung $[\partial_n U_h]$ von $\partial_n U_h$ über die Zellkanten. Durch Anwendung der Hölder'schen Ungleichung erhalten wir

$$\begin{aligned}\|e_h^{m+}\| \le \sum_{\mu=1}^{m} \sum_{K_n^\mu \in \mathbf{T}_h^\mu} \Big\{ \|R(U_h)\|_{Q_n^\mu} \|z - z_h\|_{Q_n^\mu} \\ + \ \tfrac{1}{2} \|[\partial_n U_h]\|_{\partial Q_n^\mu} \|z - z_h\|_{\partial Q_n^\mu}\}\, dt + \|[U_h]^{\mu-1}\|_{K_n^\mu} \|(z - z_h)^{(\mu-1)+}\|_{K_n^\mu} \Big\},\end{aligned} \tag{5.2.79}$$

Wir wählen nun für z_h die natürliche Interpolierende $I_{h,k} z \in V_h$, welche zellweise definiert

ist durch die Vorschrift

$$I_{h,k} z_{|I_\mu} = k_\mu^{-1} \int_{I_\mu} z \, dt, \quad I_{h,k} z_{|K_n^\mu} = \text{konstante Interpolation von } z. \tag{5.2.80}$$

Dann gelten die Interpolationsabschätzungen (Beweis mit Hilfe einer Variante des Bramble-Hilbert-Lemmas)

$$\|z - I_{h,k} z\|_{Q_n^\mu} \le c_i \, \omega_n^\mu(z), \tag{5.2.81}$$

$$h_n^{1/2} \|z - I_{h,k} z\|_{\partial Q_n^\mu} \le c_i \, \omega_n^\mu(z), \tag{5.2.82}$$

$$k_\mu^{1/2} \|(z - I_{h,k} z)^{(\mu-1)+}\|_{K_n^\mu} \le c_i \, \omega_n^\mu(z), \tag{5.2.83}$$

wobei

$$\omega_n^\mu(z) := k_\mu \|\partial_t z\|_{Q_n^\mu} + k_\mu^2 \|\partial_t^2 z\|_{Q_n^\mu} + h_n^2 \|\nabla^2 z\|_{Q_n^\mu}.$$

Die „Interpolationskonstante“ c_i, welche in die a posteriori Fehlerabschätzung (5.2.71) eingeht, hat dabei in der Regel die Größe $c_i \sim 0,1-1$. Einsetzen dieser Abschätzungen in (5.2.79) und Anwendung der Schwarz'schen Ungleichung ergeben

$$\|e_h^{m+}\| \le c_i \sum_{\mu=1}^{m} \sum_{K_n^\mu \in \mathbf{T}_h^\mu} \left\{ \|R(U_h)\|_{Q_n^\mu} + h_n^{-1/2} \|[\partial_n U_h]\|_{\partial Q_n^\mu} + k_\mu^{-1/2} \|[U_h]^{\mu-1}\|_{K_n^\mu} \right\} \omega_n^\mu(z),$$

Dies impliziert die Behauptung. Q.E.D.

Zur konkreten Auswertung der *a posteriori* Fehlerabschätzung (5.2.71) müssen die Gewichte $\omega_n^m(z)$ berechnet werden. Dazu wird analog zum stationären, elliptischen Fall das duale Problem (5.2.72) numerisch auf dem aktuellen Gitter gelöst. Mit der resultierenden diskreten dualen Lösung z_h wird dann approximiert gemäß:

$$\omega_n^\mu(z) \approx \tilde{\omega}_n^\mu := k_\mu \|k_\mu^{-1} [z_h]^{\mu-1}\|_{K_n^\mu} + h_n^2 \|\nabla_h^2 z_h\|_{Q_n^\mu}, \tag{5.2.84}$$

wobei $\nabla_h^2 z_h \sim \nabla^2 z$ ein geeigneter Differenzenqotient ist. Auf der Basis der approximativen *a posteriori* Fehlerabschätzung

$$\|e_h^{m+}\| \approx \tilde{\eta}(U_h) := c_i \sum_{\mu=1}^{m} \sum_{K_n^\mu \in \mathbf{T}_h^\mu} \rho_n^\mu(U_h) \, \tilde{\omega}_n^\mu(z) \tag{5.2.85}$$

können nun simultan das Ortsgitter und die Zeitschrittweite adaptiert werden. Dabei werden in einem iterativen Prozess die Ortsresiduen

$$\rho_n^{m,1}(U_h) := \|R(U_h)\|_{Q_n^\mu} + h_n^{-1/2} \|[\partial_n U_h]\|_{\partial Q_n^\mu},$$

und Zeitresiduen

$$\rho_n^{m,2}(U_h) := k_\mu^{-1/2} \|[U_h]^{\mu-1}\|_{K_n^\mu},$$

durch Anpassen von h_n und k_m balanciert, bis $\tilde{\eta}(U_h) \le$ TOL für eine vorgegebene Fehlertoleranz TOL.

Analoge Resultate gelten für andere Fehlermaße, z. B. den globalen L^2-Fehler auf Q_T:

$$J(e_h) := \|e_h\|_{Q_T} = \Big(\int_I \|e_h\|^2 \, dt \Big)^{1/2}. \tag{5.2.86}$$

In diesem Fall gilt die *a posteriori* Fehlerabschätzung (5.2.85) mit dem dualen Problem

$$-\partial_t z - \Delta z = \|e_h\|_{Q_T}^{-1/2} e_h \quad \text{in } \Omega \times [0, t_m], \quad z_{|\partial\Omega} = 0, \;\; z_{|t=t_m} = 0. \tag{5.2.87}$$

Die Wirksamkeit der adaptiven Steuerung der Ortsgitterweite auf der Basis dieser *a posteriori* Fehlerabschätzungen wird anhand eines einfachen Beispiels demonstriert.

Beispiel 5.1: Wir lösen die inhomogene Wärmeleitungsgleichung

$$\partial_t u - \Delta u = f \quad \text{in } Q_T, \quad u_{|\partial\Omega} = 0, \;\; u_{|t=0} = u^0, \tag{5.2.88}$$

auf dem Einheitsquadrat $\Omega = (0,1)^2$. Als „exakte“ Lösung wird angesetzt:

$$u(x,t) := \frac{1}{1 + \alpha |x - x^0|^2}, \quad x^0 := \big(\tfrac{1}{2} + \tfrac{1}{4}\cos(2\pi t), \tfrac{1}{2} + \tfrac{1}{4}\sin(2\pi t) \big)^T,$$

woraus sich Anfangswert und rechte Seite ergeben zu

$$u^0(x) := u(x,0), \quad f(x,t) := \partial_t u(x,t) - \Delta u(x,t).$$

Diese Lösungsfunktion beschreibt einen „Hügel“ im Gebiet $(0,1)^2$, der während des Zeitintervalls $I = [0,1]$ einmal im Kreis um den Punkt $(\frac{1}{2}, \frac{1}{2})$ herumläuft. Die Größe bzw. Steigung des Hügels lässt sich durch Wahl des Parameters α steuern. Für den Test wird $\alpha = 50$ gesetzt. Die sich damit ergebenden Gitter sind in Abb. 5.2 und 5.3 gezeigt.

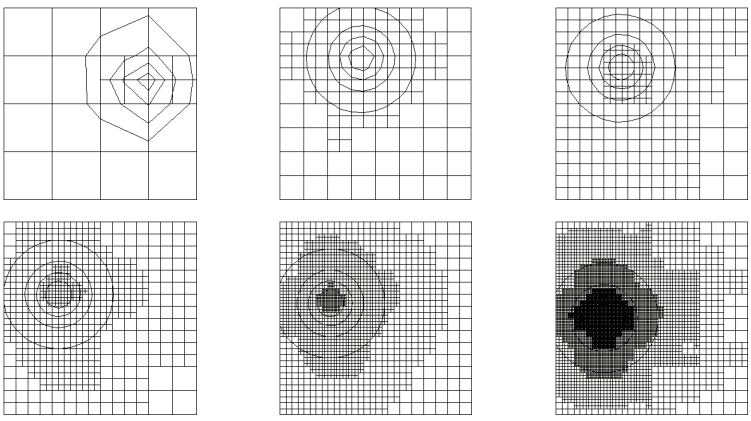

Abbildung 5.2: Gittersequenzen bei Kontrolle des Endzeit-L^2-Fehlers $\|e_M\|_\Omega$; Quelle: R. Hartmann, „A posteriori Fehlerschätzung ... für die Wärmeleitungsgleichung“, Diplomarbeit, Univ. Heidelberg, 1998.

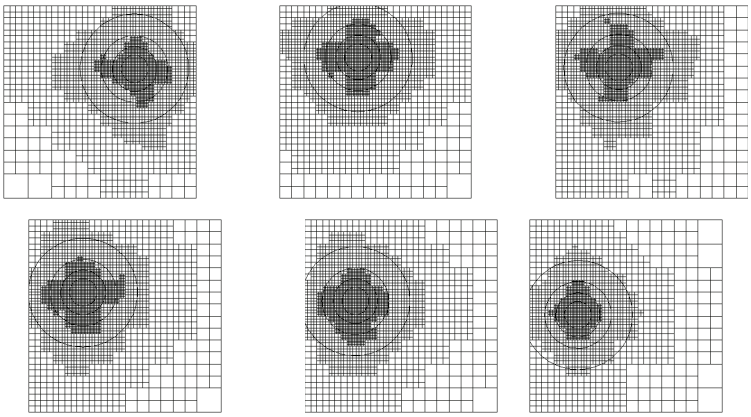

Abbildung 5.3: Gittersequenzen bei Kontrolle des globalen L^2-Fehlers $\|e\|_{\Omega\times I}$ (unten); Quelle: R. Hartmann, „A posteriori Fehlerschätzung ... bei Galerkin-Verfahren für die Wärmeleitungsgleichung“, Diplomarbeit, Univ. Heidelberg, 1998.

5.3 Verallgemeinerungen und Lösungsaspekte

Wir haben als Modellfall die Wärmeleitungsgleichung

$$\partial_t u - a\Delta u = f \quad \text{in } Q_T = \Omega \times [0,T], \tag{5.3.89}$$

mit homogenen Dirichlet-Randbedingungen $u_{|\partial\Omega} = 0$ und einem konstanten „Diffusionskoeffizienten“ $a > 0$ betrachtet.

i) Verallgemeinerungen:
Wir lassen nun zu, dass der Koeffizient $a = a(x,t)$ vom Ort und von der Zeit abhängt. Das Problem schreibt sich dann in der Form

$$\partial_t u - \nabla \cdot \{a\nabla u\} = f \quad \text{in } Q_T = \Omega \times [0,T]. \tag{5.3.90}$$

Die oben betrachteten Zeitschrittverfahren lassen sich in der Regel leicht auf diesen allgemeineren Fall übertragen. Wir wollen dies anhand des Crank-Nicolson-Verfahren diskutieren:

$$k^{-1}\left(U_h^m - U_h^{m-1}\right) + \tfrac{1}{2}\left(A_h^m U_h^m + A_h^{m-1}U_h^{m-1}\right) = \tfrac{1}{2}\left(f_h^m + f_h^{m-1}\right), \tag{5.3.91}$$

mit der Ortsdiskretisierung $A_h(t)$ des Operators $-\nabla\cdot(a(\cdot,t)\nabla)$. Dabei werden für Funktionen $w(t)$ die abkürzenden Bezeichnungen $w^m := w(t_m)$, $w^{m-1/2} := w(t_{m-1/2})$ und $t_{m-1/2} := \frac{1}{2}(t_m + t_{m-1})$ verwendet. Dies entspricht der sog. „Sehnentrapezregel“; Anwen-

dung der „Tangententrapezregel“ führt auf das Schema:

$$k^{-1}\left(U_h^m - U_h^{m-1}\right) + \tfrac{1}{2}A_h^{m-1/2}\left(U_h^m + U_h^{m-1}\right) = f_h^{m-1/2}. \tag{5.3.92}$$

Beide Verfahrensvarianten sind *unbedingt stabil* und von der Konvergenzordnung $\mathcal{O}(k^2 + h^2)$. Zu Ihrer Analyse ist das elegante Spektralargument aus dem vorigen Abschnitt leider nicht mehr geeignet, da der Operator $A_h(t)$ nun mit der Zeit variiert. Statt dessen verwendet man die flexiblere „Energietechnik“ und behält im wesentlichen dieselben Aussagen wie im autonomen Fall, allerdings mit wesentlich mehr Aufwand. Im (praktisch wichtigen) nichtlinearen Fall $a = a(u(t))$ wird zweckmäßigerweise die Tangententrapezform des Crank-Nicolson-Schemas verwendet:

$$k^{-1}\left(U_h^m - U_h^{m-1}\right) + A_h\left(\tfrac{1}{2}(U_h^m + U_h^{m-1})\right) = f_h^{m-1/2}. \tag{5.3.93}$$

Hier ist auch die BDF(2)-Formel gut anwendbar:

$$2k^{-1}\left(3U_h^m - 4U_h^{m-1} + U_h^{m-2}\right) + A_h(U_h^m) = f_h^m. \tag{5.3.94}$$

Eine Stabilitäts- und Konvergenzanalyse steht aber außerhalb des Rahmens dieses Textes.

ii) Berechnung der Startwerte:
Bei der Durchführung jedes Zeitschritts ist die rechte Seite aufzubauen, welche die Information vom vorausgehenden Zeitlevel beinhaltet. Dabei ist unter Umständen eine L^2-Projektion auf das aktuelle Gitter vorzunehmen.

a) Anfangswert: Die Auswertung des Anfangswerts U_h^0 kann meist durch einfache, lokale Interpolation (oder Restriktion) des kontinuierlichen Anfangswerts u^0 auf das Gitter erfolgen. Im Fall eines irregulären Anfangswerts, etwa $u^0 \notin C(\bar{\Omega})$, ist jedoch Vorsicht geboten. Zur Gewährleistung der vollen „Glättungseigenschaft“ der Diskretisierung (im Ort sowie in der Zeit) sollte U_h^0 als L^2-Projektion ausgewertet werde gemäß:

$$(U_h^0, \varphi_h) = (u^0, \varphi_h) \quad \varphi_h \in V_h. \tag{5.3.95}$$

b) Ortsgitterwechsel: Verändert sich die Ortsdiskretisierung vom Zeitlevel t_{m-1} zum Zeitlevel t_m, so muss die vorausgehende Näherung $U_h^{m-1} \in V_h^{m-1}$ auf das neue Gitter transferiert werden. Im FE-Kontext geschieht dies zwangsläufig gemäß

$$(U_h^{m-1}, \varphi_h) \quad \forall \varphi_h \in V_h^m, \tag{5.3.96}$$

was gleichbedeutend mit der Auswertung der L^2-Projektion $P_h^m U_h^{m-1} \in V_h^m$ ist. Dieser unscheinbare Schritt kann unter Umständen die „teure“ Komponente des ganzen Lösungsprozesses sein. Dies ist dann der Fall, wenn die beiden Gitter $\mathbf{T}_h^{m-1}$ und $\mathbf{T}_h^m$ völlig unabhängig voneinander erzeugt werden. Zur Auswertung von (5.3.96) müssen Zellintegrale über Produkte von Knotenbasisfunktionen berechnet werden:

$$\int_{K_n^m} \varphi_h^{m-1,j} \varphi_h^{m,n}\, dx.$$

Normalerweise geschieht dies mit Hilfe von Quadraturformeln. Da die Funktion $\varphi_h^{m-1,j}$ auf der Zelle K_n^m aber in der Regel nur stückweise glatt ist,wäre das zu ungenau. Der dadurch in jedem Zeitschritt eingeschleppte Fehler würde im Verlaufe der Rechnung akkumulieren und das Ergebnis stark verfälschen. Die Verwendung von Quadraturformeln besonders hoher Ordnung (z.B. 3×3-Gauß-Formeln) behebt diese Schwierigkeit nicht, da letztere zur Erreichung ihrer hohen Genauigkeit natürlich auch eine entsprechend hohe Regularität des Integranden benötigen. Gerade diese ist aber im betrachteten Fall nicht gegeben. Es gibt im wesentlichen drei Wege zur Lösung dieses technischen Problems:

- Es werden „summierte“ Quadraturformeln auf den einzelnen Zellen K_n^m verwendet; etwa durch Unterteilung in $4-16$ Unterzellen. Dies erhöht zwar nicht die Ordnung der Integration, vermindert aber die relevante Fehlerkonstante.

- Die Integration wird für den stückweise polynomialen Integranden „exakt“ durchgeführt. Dazu ist die Bestimmung aller Teilstücke von K_n^m erforderlich, auf denen $\varphi_h^{m-1,j}$ glatt ist. Dies wird bei, unstrukturierten Gittern in 3-D allerdings sehr aufwendig.

- Gehören die Gitter $\mathbf{T}_h^{m-1}$ und $\mathbf{T}_h^m$ zu einer Familie von hierarchisch verfeinerten Gittern, kann diese Strukturinformation zur effizienten Berechnung der Integrale verwendet werden, da die Lage der kritischen Knicklinien von $\varphi_h^{m-1,j}$ durch die reguläre Verfeinerung bestimmt ist.

iii) Lösungskomplexität:
Wir betrachten wieder den Modellfall der homogenen Wärmeleitungsgleichung auf dem Einheitsquadrat $\Omega=(0,1)^2\subset\mathbb{R}^2$,

$$\partial_t u-\Delta u=0 \quad \text{in } Q_T, \tag{5.3.97}$$

welche auf einem äquidistanten Gitter mit dem 5-Punkte-Differenzenoperator diskretisiert sei.

– *Explizite* Verfahren (etwa das explizite Euler-Schema) erfordern in jedem Zeitschritt eine Matrix-Vektor-Multiplikation mit einem arithmetischen Aufwand $\mathcal{O}(N)$. Die Stabilitätsbedingung $k\le ch^2$ erzwingt etwa $M\sim h^{-2}\sim N$ Zeitschritte pro Zeiteinheit. Dies bedeutet einen Gesamtaufwand von $\mathcal{O}(N^2)$ OP.

– *Implizite* Verfahren erfordern in jedem Zeitschritt die Lösung eines linearen Gleichungssystems, erlauben aber größere Zeitschritte. Zur Überbrückung einer Zeiteinheit sind aus Genauigkeitsgründen in der Regel $M\sim h^{-1}$ Zeitschritte erforderlich. Wir diskutieren exemplarisch das Crank-Nicolson-Verfahren. Bei zeilenweiser Numerierung der Gitterpunkte haben die resultierenden Gleichungssysteme

$$(I_h+\tfrac{1}{2}akA_h)U^m=(I_h-\tfrac{1}{2}akA_h)U^{m-1} \tag{5.3.98}$$

die Koeffizientenmatrix $L_h := I_h + \frac{1}{2}akA_h$, wobei wieder

$$A_h = \left.\begin{bmatrix} B_m & -I_m & & \\ -I_m & B_m & -I_m & \\ & -I_m & B_m & \ddots \\ & & \ddots & \ddots \end{bmatrix}\right\} N \qquad B_m = \left.\begin{bmatrix} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & \ddots \\ & & \ddots & \ddots \end{bmatrix}\right\} m$$

mit der $m \times m$-Einheitsmatrix I_m. Die Eigenwerte dieser Matrix sind gegeben durch:

$$\lambda_{kl}(L_h) = 1 + \tfrac{1}{2}akh^{-2}\{4 - 2(\cos(kh\pi) + \cos(lh\pi))\}, \quad k, l = 1, ..., m.$$

Damit ergibt sich ihre Spektralkondition zu

$$\mathrm{cond}_2(L_h) = \frac{\lambda_{max}(L_h)}{\lambda_{min}(L_h)} \sim \frac{1 + 4akh^{-2}}{1 + ak\pi^2}. \tag{5.3.99}$$

Wir haben also unterschiedlich kritische Konditionierung abhängig von der Relation zwischen k und h. Im Hinblick auf eine Balancierung von Orts- und Zeitfehler ist die Wahl $k \sim h$ sinnvoll. In diesem Fall ist dann

$$\kappa_2(L_h) = \mathcal{O}(h^{-1}). \tag{5.3.100}$$

Im parabolischen Fall ist in der Regel die Konditionierung der zu lösenden impliziten Gleichungssysteme also weniger kritisch als bei elliptischen Problemen. Bei Einhaltung der (natürlich unrealistischen) Schrittweitenrelation $k \sim h^2$ wird sogar $\kappa_2(L_h) = \mathcal{O}(1)$, und der Lösungsaufwand der impliziten Verfahren nähert sich dem der expliziten.

Zur Lösung des Systems (5.3.98) können alle oben diskutierten Methoden verwendet werden. Da sich im autonomen Fall die Matrix L_h von Zeitschritt zu Zeitschritt nicht ändert, bietet sich die direkte Lösung mit Hilfe einer einmaligen Cholesky-Zerlegung zu Beginn der Rechnung an (wenn dies speichertechnisch möglich ist). Dieser Ansatz erfordert es aber, den Zeitschritt k konstant zu halten, was in den meisten praktischen Fällen nicht ökonomisch ist. I. Allg. muss in jedem Zeitschritt eine „neue“ Matrix invertiert werden, was die Verwendung iterativer Lösungsverfahren impliziert. Dabei steht mit U_h^{m-1} ein meist recht guter Startwert zur Verfügung. Bei Verwendung eines Mehrgitterverfahrens bietet sich daher die Organisation im F-Zyklus an. Häufig ist wegen der vergleichsweise moderaten Konditionierung der Matrizen L_h (bei kleiner Zeitschrittweite) zu ihrer Invertierung ein normales CG-Verfahren ausreichend schnell, so dass sich der Einsatz der komplizierten Mehrgitteriteration erübrigt. Dies hängt aber sehr von der jeweiligen konkreten Situation ab.

iv) Splitting-Methoden (ADI-Verfahren):
In höheren Raumdimensionen ist die Lösung der Gleichungssysteme in jedem Zeitschritt eines *impliziten* Verfahrens kostspielig und kann Rechnungen über sehr lange Zeitintervalle $T \gg 1$ unmöglich machen. Der Übergang zu *expliziten* Schemata ist in der Regel wegen der damit verbundenen Zeitschrittrestriktion auch nicht sinnvoll. In dieser Situa-

tion stellen die sog. „Splitting-Methoden“ eine attraktive Alternative dar. Diese zerlegen die Lösung der vollen d-dimensionalen Gleichungssysteme in eine Folge von tridiagonalen Systemen (wie im 1-dimensionalen Fall), welche mit optimaler $\mathcal{O}(N)$-Komplexität gelöst werden können. Ein Vertreter dieses Verfahrenstyps ist das „ADI-Verfahren“ (Alternating Direction Implicit Iteration) nach Peaceman-Rachford, welches wir bereits im Zusammenhang mit iterativen Lösungsverfahren für spezielle „separable“ Gleichungssysteme kennengelernt haben. Bei mehrdimensionalen, parabolischen Problemen wird es nun als *Diskretisierungsverfahren* eingesetzt.

Wir betrachten wieder die Wärmeleitungsgleichung auf dem Einheitsquadrat $\Omega = (0,1)^2$,

$$\partial_t u - a\Delta u = 0 \quad \text{in } Q_T, \tag{5.3.101}$$

welche auf einem äquidistanten Gitter mit dem 5-Punkte-Differenzenoperator diskretisiert sei. Bei zeilenweiser Numerierung der Gitterpunkte haben die aus dem Crank-Nicolson-Verfahren resultierenden Gleichungssysteme die Gestalt

$$(I_h + \tfrac{1}{2}akA_h)U^m = (I_h - \tfrac{1}{2}akA_h)U^{m-1}. \tag{5.3.102}$$

Der Differenzenoperator A_h wird auf dem kartesischen Tensorprodukt-Gitter in seine Bestandteile bzgl. der einzelnen Koordinatenrichtungen zerlegt gemäß

$$A_h = A_{h,1} + A_{h,2}.$$

Entsprechend erhält das Gleichungssystem (5.3.102) des Crank-Nicolson-Schemas die Form

$$(I_h + \tfrac{1}{2}ak(A_{h,1} + A_{h,2}))U^m = (I_h - \tfrac{1}{2}ak(A_{h,1} + A_{h,2}))U^{m-1}$$

Die $A_{h,i}$ sind Tridiagonalmatrizen. Es wird dann unter Einführung von Zwischenwerten $U_h^{m-1/2}$ wie folgt iteriert:

$$\begin{aligned}(I_h + \tfrac{1}{2}akA_{h,1})U_h^{m-1/2} &= (I_h - \tfrac{1}{2}akA_{h,2})U_h^{m-1}\\ (I_h + \tfrac{1}{2}akA_{h,2})U_h^{m} &= (I_h - \tfrac{1}{2}akA_{h,1})U_h^{m-1/2}.\end{aligned}$$

Die ADI-Methode kann als ein Mehrschritt-Differenzenschema interpretiert werden, wobei allerdings die Zwischenwerte keine physikalische Relevanz haben. In jedem Teilschritt müssen Gleichungssysteme mit Tridiagonalgestalt gelöst werden. Wir wissen bereits von der Diskussion der iterativen Lösungsverfahren, dass der ADI-Algorithmus für jeden Wert des Parameters $ak > 0$ gegen die Lösung U_h^∞ des Gleichungssystems $A_h U_h^\infty = 0$ konvergiert. Dies ist auch „physikalisch“ sinnvoll, da ja im homogenen Fall ($f \equiv 0$) auch für die exakte Lösung gilt $u(t) \to 0\,(t \to \infty)$. Damit erweist sich das ADI-Verfahren automatisch als *unbedingt numerisch stabil*. Durch Elimination des Zwischenwertes $U_h^{m-1/2}$ erhalten wir

$$(I + \tfrac{1}{2}kA_{h,1})(I + \tfrac{1}{2}kA_{h,2})U_h^m = (I - \tfrac{1}{2}kA_{h,1})(I - \tfrac{1}{2}kA_{h,2})U_h^{m-1}. \tag{5.3.103}$$

Der Abschneidefehler dieser Differenzenformel erlaubt die Abschätzung

$$|\tau_{h,k}^m| \le c\Big\{k^2 \max_{\bar{Q}_T} |\partial_t^3 u| + h^2 \max_{\bar{Q}_T} |\nabla^4 u|\Big\}. \tag{5.3.104}$$

Eine Konvergenzanalyse ist leicht möglich, wenn wir wieder annehmen, dass die Zerlegungsmatrizen $A_{h,1}$ und $A_{h,2}$ kommutieren, d.h.: $A_{h,1}A_{h,2} = A_{h,2}A_{h,1}$. In diesem Fall besitzen sie ein gemeinsames ONS von Eigenvektoren $\{v^{(n)},\, n = 1, ..., N\}$ zu Eigenwerten $\lambda_n = \lambda_n(A_{h,1})$ und $\mu_n = \mu_n(A_{h,2})$. Die Koeffizienten in der Entwicklung

$$U_h^m = \sum_{\nu=1}^{N} \alpha_\nu^m v^{(n)}$$

werden dann durch das ADI-Schema wie folgt fortgepflanzt

$$\alpha_n^m = \frac{(1 - \frac{1}{2}k\lambda_n)(1 - \frac{1}{2}k\mu_n)}{(1 + \frac{1}{2}k\lambda_n)(1 + \frac{1}{2}k\mu_n)}\, \alpha_n^{m-1}. \tag{5.3.105}$$

Hieraus folgt wieder, analog zum Crank-Nicolson-Schema, direkt die unbedingte Stabilität des ADI-Schemas. Für die Fourier-Koeffizienten $\alpha_n(t)$ der örtlich semi-diskreten Approximation $u_h(t)$ gilt

$$\alpha_n(t_m) = e^{-k(\lambda_n+\mu_n)}\alpha_n(t_{m-1}). \tag{5.3.106}$$

Aus der Beziehung für $z = z_1 + z_2,\ z_i \le 0$,

$$\frac{1 + \frac{1}{2}z_1}{1 - \frac{1}{2}z_1} \cdot \frac{1 + \frac{1}{2}z_2}{1 - \frac{1}{2}z_2} = \big\{e^{z_1} + \mathcal{O}(|z_1|^3)\big\}\,\big\{e^{z_2} + \mathcal{O}(|z_2|^3)\big\} = e^z + \mathcal{O}(|z|^3) \tag{5.3.107}$$

erhält man durch Adaption des Arguments, welches bei der Konvergenzanalyse der Padé-Verfahren verwendet wurde, den folgenden Satz.

Satz 5.7 (ADI-Verfahren): *Angewendet auf die 5-Punkte-Diskretisierung der Wärmeleitungsgleichung auf dem Einheitsquadrat ist das ADI-Verfahren für jede Zeitschrittweite k stabil und mit 2. Ordnung konvergent:*

$$\|U_h^m - u^m\|_h \le c\Big\{k^2 \max_{\bar{Q}_T} |\partial_t^3 u| + h^2 \max_{\bar{Q}_T} |\nabla^4 u|\Big\}. \tag{5.3.108}$$

Beweis: Der Beweis bedient sich wieder der Spektralmethode und wird ausgelassen. Q.E.D.

Der ADI-Ansatz ist generell auf (kartesischen) Tensorproduktgittern in beliebigen Raumdimensionen möglich, wenn der zugrunde liegende Differentialoperator „separabel“ ist, d.h. additiv in eindimensionale Operatoren zerfällt, wie z.B. der Operator

$$Lu = -\partial_1^2 u - \partial_2^2 u + \partial_1 u + \partial_2 u + u.$$

In allgemeineren Situationen (z. B. bei Auftreten gemischter Ableitungen $\partial_1\partial_2 u$ oder auf unstrukturierten Gittern) sind additive Zerlegungen von A_h in tridiagonale Teilmatrizen nicht mehr möglich, und der Wert des ADI-Ansatzes wird zweifelhaft.

5.4 Übungen

Übung 5.1: Das implizite Euler-Verfahren angewendet auf eine FE-Diskretisierung des (homogenen) Wärmeleitungsproblems

$$\partial_t u - \Delta u = 0, \quad t \geq 0, \quad u_{|t=0} = u^0, \quad u_{|\partial\Omega} = 0,$$

mit linearen oder bilinearen Ansatzfunktionen führt auf eine Folge von linearen Systemen

$$M_h U^m + kA_h U^m = M_h U^{m-1}, \quad m \geq 1, \quad U^0 = P_h u^0,$$

mit der zugehörigen Massematrix M_h und Steifigkeitsmatrix A_h. Man zeige:

a) Die Auswertung der Elemente der Massematrix mit der Trapezregel ergibt eine Diagonalmatrix (sog. „Masse-Lumping“).

b) Auf Triangulierungen ohne stumpfe Innenwinkel ist die durch Masse-Lumping entstehende Systemmatrix $\tilde{M}_h + kA_h$ diagonal-dominant und vom nicht-negativen Typ, d. h. eine M-Matrix.

c) Anspruchsvolle Zusatzaufgabe: Der Lumping-Prozess bewirkt einen Zusatzfehler der Größe $|\tilde{U}^m - U^m| = \mathcal{O}(h^2)$.

Übung 5.2: Man leite eine Bedingung für die L^2–Stabilität sowie die L^∞-Stabilität des Wärmeleitungsproblems aus Aufgabe 13.1 auf dem Einheitswürfel im R^3 her. Die Ortsdiskretisierung erfolge mit dem 7-Punkte-Differenzenoperator auf einem (äquidistanten) kartesischen Gitter mit Gitterweite h. (Hinweis: Man übertrage die Argumentation aus der Vorlesung von zwei auf drei Raumdimensionen.)

Übung 5.3: Die Wärmeleitungsgleichung

$$\partial_t u - a\Delta u = f, \quad t \geq 0, \quad u_{t=0} = u^0, \quad u_{|\partial\Omega} = 0,$$

auf einem Polygongebiet $\Omega \subset \mathbb{R}^2$ mit Wärmeleitkoeffizient $a > 0$ werde im Ort mit einem linearen FE-Ansatz auf einer quasi-gleichförmigen Folge von Triangulierungen der Gitterweite h und in der Zeit mit dem impliziten Euler-Schema mit Schrittweite k diskretisiert:

$$M_h U^m + kA_h U^m = M_h U^{m-1} + kb^m, \quad m \geq 1, \quad U^0 = P_h u^0.$$

Man untersuche die Abhängigkeit der Konditionierung der zugehörigen Systemmatrix $M_h + kA_h$ von den Diskretisierungsparametern h und k. (Hinweis: Man betrachte zunächst den Spezialfall, dass Ω das Einheitsquadrat mit einem gleichförmigen Rechteckgitter ist und die Massematrix M_h unter Anwendung der Trapezregel („Masse-Lumping“) nur näherungsweise berechnet wird.)