ZUSATZMATERIAL UND KORREKTUREN ZUR 2. AUFLAGE VON *NUMERIK 3X9*

S. BARTELS, 16.12.2024

Kap. 1: Auslöschung.

Zu Auslöschungseffekten kann es kommen, wenn betragsmäßig große Zwischenergebnisse oder Summanden in einer Rechnung auftreten.

Satz 1. *Sei $\phi(x,y) = x - y \neq 0$ und seien $\widetilde{x}, \widetilde{y}$ Störungen von x, y. Dann werden die relativen Fehler $\varepsilon_x = |x - \widetilde{x}|/|x|$ und $\varepsilon_y = |y - \widetilde{y}|/|y|$ mit dem Inversen der exakten Differenz $\delta = x - y$ und der Summe der Beträge von x und y verstärkt, das heißt es gilt*

$$\varepsilon_\phi = \frac{|(x - \widetilde{x}) + (y - \widetilde{y})|}{|\delta|} \leq |\delta|^{-1}(|x| + |y|) \max\{\varepsilon_x, \varepsilon_y\}.$$

(ii) Ist $s = y_1 + y_2 + \cdots + y_n$ und $\widetilde{s} = \widetilde{y}_1 + \widetilde{y}_2 + \cdots + \widetilde{y}_n$, mit relativen Fehlern $\varepsilon_i = |\widetilde{y}_i - y_i|/|y_i|$, dann gilt für den relativen Fehler $\varepsilon_s = |\widetilde{s} - s|/|s|$ der gestörten Summe, dass

$$\varepsilon_s \leq \Big(\frac{1}{|s|} \sum_{i=1}^{n} |y_i|\Big) \max_{j=1,\dots,n} \varepsilon_j,$$

eine starke Fehlerverstärkung tritt auf, falls $|y_1| + |y_2| + \cdots + |y_n| \gg |s|$ gilt.

Beweis: (i) Die erste Aussage folgt mit der Dreiecksungleichung.
(ii) Mit Faktoren $\sigma_i \in \{\pm 1\}$, sodass $\sigma_i \varepsilon_i = (\widetilde{y}_i - y_i)/y_i$ gilt, ist die gestörte Summe gegeben durch

$$\widetilde{s} = \sum_{i=1}^{n} (1 + \sigma_i \varepsilon_i) y_i = s + \sum_{i=1}^{n} \sigma_i \varepsilon_i y_i.$$

Die Dreiecksungleichung impliziert die behauptete Abschätzung, die optimal ist, falls beispielsweise $\sigma_i \varepsilon_i y_i \geq 0$ für $i = 1, 2, \dots, n$ gilt. □

Kap 5: Ausgleichsprobleme.

Die theoretische Konstruktion der Lösung des Ausgleichsproblems und die Verwendung der QR-Zerlegung sind in Abbildung 1 illustriert.

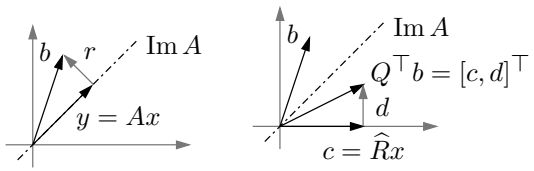

ABBILDUNG 1. Abstrakte Lösung des Ausgleichsproblems über die Zerlegung $b = Ax + r$ mit $r \in \ker A^\top$ (links), Transformation des Augleichsproblems mit der QR-Zerlegung von A (rechts).

Kap. 11: Baryzentrische Lagrange-Basis.

Ein alternativer Zugang zur Auswertung eines Interpolationspolynoms p folgt aus der *baryzentrischen Darstellung* der Lagrange-Basispolynome. Mit

$$\gamma_i = \prod_{j=0,\dots,n,\, j\neq i} (x_i - x_j)^{-1}, \quad w(x) = \prod_{j=0}^{n} (x - x_j),$$

gilt $L_i(x) = \gamma_i w(x)(x - x_i)^{-1}$ für $x \neq x_i$, und folglich

$$p(x) = w(x) \sum_{i=0}^{n} (x - x_i)^{-1} \gamma_i y_i$$

sowie $p(x_i) = y_i$ für $i = 0, 1, \dots, n$. Im Fall $y_0 = y_1 = \cdots = y_n = 1$ gilt $1 = w(x) \sum_{i=0}^{n} \gamma_i (x - x_i)^{-1}$ und somit

$$p(x) = \frac{\sum_{i=0}^{n} (x - x_i)^{-1} \gamma_i y_i}{\sum_{i=0}^{n} (x - x_i)^{-1} \gamma_i}.$$

Diese Darstellungen von p lassen sich in $\mathcal{O}(n)$ Operationen auswerten, sofern die Gewichte $(\gamma_i)_{i=0,\dots,n}$ zuvor berechnet wurden.

Kap. 12: Interpolation mit Splines.

Korollar 12.1: Sei s_3 die ... Die Funktion f erfülle (i) $f''(x) = 0$ für $x \in \{a, b\}$, (ii) $s_3'(x) = f'(x)$ für $x \in \{a, b\}$ oder (iii) $f^{(i)}(a) = f^{(i)}(b)$ für $i = 0, 1, 2$, im Fall natürlicher, vollständiger beziehungsweise periodischer Randbedingungen. Dann gilt ...

Kap 14: Summierte Trapezregel.

Bemerkung: Oft wird ein deutlich schnelleres Konvergenzverhalten der summierten Trapezregel beobachtet. Für periodische Funktionen $f \in C^k([0, 2\pi])$, deren Ableitungen bis zur Ordnung k periodisch sind, lässt sich für den Quadraturfehler $\delta_N = |Q^N(f) - I(f)|$ die Ordnung $\delta_N = \mathcal{O}(h^k)$ nachweisen. Wird das Intervall $[0, 2\pi]$ mit dem Einheitskreis in $\mathbb{C}$ identifiziert, und besitzt f eine holomorphe Fortsetzung in einer Umgebung des Kreises, so lässt sich die exponentielle Konvergenzeigenschaft $\delta_N = \mathcal{O}(r^{-N})$ mit einer Zahl $r > 1$ beweisen.

Aufgabe. Wir identifizieren periodische Funktionen $f \in C([0,2\pi])$ mit Funktionen auf der Einheitskreislinie $\partial B_1(0) \subset \mathbb{C}$ mittels $f(e^{\mathrm{i}\theta}) \equiv f(\theta)$ und betrachten

$$I(f) = \int_0^{2\pi} f(e^{\mathrm{i}\theta})\,\mathrm{d}\theta.$$

Die Funktion f sei fortsetzbar zu einer holomorphen Funktion in $B_{\widetilde{r}}(0)$, $\widetilde{r} > 1$. Integrale über Kreislinien $\partial B_r(0)$ werden mit der Abbildung $\gamma(\theta) = re^{\mathrm{i}\theta}$ als Linienintegrale verstanden, das heißt

$$\int_{\partial B_r(0)} f(w)\,\mathrm{d}w = \int_0^{2\pi} f(\gamma(\theta))\gamma'(\theta)\,\mathrm{d}\theta.$$

(i) Zeigen Sie, dass die summierte Trapezformel gegeben ist durch

$$Q^N(f) = \frac{2\pi}{N}\sum_{k=1}^{N} f(e^{\mathrm{i}k2\pi/N}),$$

und dass sich diese auch als summierte Mittelpunktregel interpretieren lässt.
(ii) Aus der Cauchy-Integralformel ergibt sich für $z \in B_{\widetilde{r}}(0)$ und $0 < r < \widetilde{r}$, dass

$$f(z) = \sum_{n=0}^{\infty} c_n z^n, \quad c_n = \frac{f^{(n)}(0)}{n!} = \frac{1}{2\pi\mathrm{i}}\int_{\partial B_r(0)} \frac{f(w)}{(w-0)^{n+1}}\,\mathrm{d}w.$$

Beweisen Sie, dass $I(f) = 2\pi c_0$ und $|c_n| \le M_r r^{-n}$ mit $M_r = \max_{w\in\partial B_r(0)} |f(w)|$.
(iii) Zeigen Sie, dass

$$Q^N(f) = 2\pi\sum_{\ell=0}^{\infty} c_{\ell N}$$

und folgern Sie, dass $|Q^N(f) - I(f)| \le 2\pi M_r(r^N-1)^{-1} = \mathcal{O}(r^{-N})$ für alle $1 < r < \widetilde{r}$ gilt.

Kap. 14: Aitkinscher Delta-Quadrat-Prozess.

Ist die Folge $(x_k)_{k\ge 0}$ linear konvergent gegen x^* mit Faktor $0 < q < 1$, so gilt approximativ $(x^* - x_k) \approx q(x^* - x_{k-1})$ sowie $(x^* - x_{k+1}) \approx q(x^* - x_k)$, sodass sich q eliminieren und x^* approximativ bestimmen lässt, das heißt

$$x^* \approx y_{k+1} = \frac{x_{k-1}x_{k+1} - x_k^2}{x_{k+1} - 2x_k + x_{k-1}} = x_{k+1} - \frac{(\delta x_k)^2}{\delta^2 x_k},$$

wobei $\delta x_k = x_{k+1} - x_k$ und $\delta^2 x_k = x_{k+1} - 2x_k + x_{k-1}$ seien. Unter bestimmten Voraussetzungen an die Differenzen $x_k - x^*$ kann quadratische Konvergenz gegen x^* für die Folge $(y_k)_{k\ge 2}$ bewiesen werden. Das Vorgehen wird als *Aitkinscher Delta-Quadrat-Prozess* bezeichnet.

Kap. 15: Fixpunktiterationen.

Beispiel 15.1: (ii) Ist $\Phi \in C^1(\mathbb{R})$, so ist die Fixpunktiteration $x_{k+1} = \Phi(x_k)$ zur Bestimmung eines Fixpunkts x^* von Φ lokal und linear konvergent, falls $|\Phi'(x^*)| < 1$ gilt, was unmittelbar aus dem Mittelwertsatz folgt. Im Fall $\Phi'(x_*) = 0$ gilt lokale, quadratische Konvergenz, wie eine Taylor-Approximation zeigt, das heißt

$$x^{k+1} - x^* = \Phi(x_k) - \Phi(x^*) = \frac{1}{2}\Phi''(\xi)(x_k - x^*)^2.$$

(iii) Konvergente Fixpunktiterationen sind robust bezüglich Rundungsfehlern, da sie selbststabilisierend sind in dem Sinne, dass jede Iterierte als neuer Startwert der Iteration betrachtet werden kann.

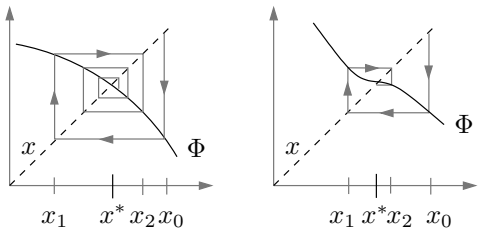

ABBILDUNG 2. Lineare (links) und quadratische (rechts) Konvergenz von Fixpunktiterationen.

Kap. 21: Konvergenz von Einschrittverfahren.

In wichtigen Spezialfällen lassen sich präzisere Fehlerabschätzungen beweisen, die zudem zu wichtigen allgemeinen Erkenntnissen führen.

Beispiel 21.3: Wir betrachten die Gleichung $y' = \lambda y$ mit einer Zahl $\lambda < 0$. Für das explizite Euler-Verfahren erhalten wir mit $e_k = y(t_k) - y_k$ die Fehlergleichung

$$e_{k+1} = (1 + \tau\lambda)e_k + \tau\widetilde{\mathcal{C}}_{expl}\big(t_k, y(t_k), \tau\big).$$

Sofern nicht $|1 + \tau\lambda| \leq 1$ oder $\tau \leq 1/(2|\lambda|)$ gilt, ist mit einer Fehlerverstärkung in den Zeitschritten zu rechnen. Dies erklärt die Schrittweitenbedingung des vorigen Resultats. Für das implizite Euler-Verfahren ergibt sich die Fehlergleichung

$$(1 - \tau\lambda)e_{k+1} = e_k + \tau\widetilde{\mathcal{C}}_{impl}\big(t_k, y(t_k), \tau\big).$$

Da $1 - \tau\lambda \geq 1$ gilt, tritt eine Fehlerdämpfung auf. Aus $|1 - \tau\lambda|^{-1} \leq 1$ folgt für die absoluten Fehler $u_k = |e_k|$, dass

$$u_{k+1} \leq u_k + c\tau^2.$$

Eine Summation über $k = 0, 1, \ldots, \ell - 1$ mit $0 \leq \ell \leq K$ zeigt

$$\max_{k=0,\ldots,K} |y(t_k) - y_k| \leq cT\tau.$$

Diese Abschätzung gilt ohne Schrittweitenbedingung und ohne exponentielle Abhängigkeit von T und $|\lambda|$.

Kap. 24: Initialisierung von Mehrschrittverfahren.

Bemerkung: Eine Zweischrittmethode mit quadratischer Konsistenzordnung kann mit einem Schritt des impliziten oder expliziten Euler-Verfahrens initialisiert werden. Ist im Fall des expliziten Euler-Verfahrens beispielsweise $y_1 = y_0 + \tau f(0, y_0)$, so ergibt sich der Konsistenzfehler

$$y(\tau) - y_0 - \tau f(0, y_0) = y(\tau) - y(0) - \tau y'(0) = \frac{\tau^2}{2} y''(\xi)$$

mit $\xi \in [0, \tau]$, und es folgt $y_1 - y(t_1) = \mathcal{O}(\tau^2)$. In der Konvergenzanalyse des Euler-Verfahrens wird ein Faktor τ beñotigt, um die Akkumulation der Fehler mehrerer Zeitschritte $t_k = k\tau$, $k = 1, 2, \ldots, K$, zu kontrollieren.

Wichtige Korrekturen.

S. 5	In Satz 1.1 wird die Summe der komponentenweisen relativen Fehler des Arguments betrachtet.
S. 36	In der Produktdarstellung von Q sind Einbettungen $\widehat{v_i} = [0, v_i] \in \mathbb{R}^m$ der Vektoren v_i zu verwenden.
S. 133	Die Bedingung an ε sollte die Konstante c_3 enthalten.
S. 295/296	In Aufgaben 29.7.2 und 29.7.3 sind λ_1^{-1} und λ_{min}^{-1} durch λ_1 beziehungsweise λ_{min} zu ersetzen.
S. 291	In Aufgabe 29.5.10 ist der Ausdruck $\mathcal{O}(h^{3\gamma})$ durch $\mathcal{O}(h^{\gamma+2})$ zu ersetzen.