Kapitel 2

Eigenschaften numerischer Verfahren

2.1 Fehlertypen

Die Vereinfachung mathematischer Modelle und die Konstruktion praktischer Verfahren führt zu zwei Typen von Verfahrensfehlern:

- Abbruchfehler: Fehler beim Übergang von infiniten zu finiten Partialsummen,
- Diskretisierungsfehler: Entsteht durch Ersetzen einer Funktion f durch eine endliche Anzahl von Zahlenwerten. $f(x_0), \ldots, f(x_n)$ oder Koeffizienten a_0, a_1, a_2.

Dem gegenüber steht der Modellfehler:

- Mathematisches Modell (Gleichungen) gibt die Wirklichkeit nur fehlerbehaftet wieder.

Bemerkung 2.1

- *Modelle und Verfahren können als Transformationen (Abbildungen) aufgefasst werden, die Eingangsdaten auf Ausgangsdaten (Lösungen) abbilden.*
- *Störungstheorie untersucht das Fehlerverhalten.*
- *Es gilt folgender Zusammenhang:*
  - *„gut“ gestelltes Problem: kleiner Eingangsfehler → kleiner Ausgangsfehler.*

  - *„schlecht“ gestelltes Problem: kleiner Eingangsfehler → großer Ausgangsfehler.*

- *Regularisierung: Ersetzen eines „schlecht“ gestellten Problems durch ein „gut“ gestelltes Problem.*

Fehler der numerischen Methode:

- Eingabefehler
  → unvermeidbarer Fehler

- Rundungsfehler

Definition 2.1 (Absoluter/relativer Fehler) *Ist $\tilde{x} \in \mathbb{R}$ eine Näherung für $x \in \mathbb{R}$, so ist $|\tilde{x} - x|$ der absolute Fehler und $\frac{|\tilde{x}-x|}{|x|}$ der relative Fehler für $x \neq 0$.*

Beispiel 2.1 *Abbruchfehler der Exponentialfunktion:*

$$e^x = \sum_{j=0}^{\infty} \frac{x^j}{j!}; \quad x \in \mathbb{R} \quad \rightarrow \quad P_n(x) = \sum_{j=0}^{n} \frac{x^j}{j!}$$
$$x \leq 0 : |e^x - P_n(x)| \leq |x|^{n+1} \cdot \frac{1}{(n+1)!}.$$
$$x > 0 : |e^x - P_n(x)| \leq e^x \cdot |x|^{n+1} \cdot \frac{1}{(n+1)!}.$$

Rechenformat für Gleitkommazahlen laut IEEE754 bei 64 Bit Gleitkomma-Arithmetik: Basis 2, 52 Stellen für Mantisse, 1 Vorzeichen-Bit, 11 Stellen für Charakteristik.
Es gilt das Rundungsgesetz:

$$|x - rd(x)| \leq |x| \epsilon \quad \forall x \in \mathbb{R}$$

mit Maschinengenauigkeit $\epsilon = 2^{-51} \approx 4{,}4409 \cdot 10^{-16} \Rightarrow$, d.h. die 16. Dezimalstelle ist bis auf 5 Einheiten genau.

Bemerkung 2.2

- *IEEE 754 sichert zu, dass der maximale Fehler pro Operation $+, -, \cdot$, $sin(x), sqrt(x), exp(x)$ ϵ nicht überschreitet.*

- *Für eine Sequenz von Operationen der Länge n ist der maximal mögliche Rundungsfehler auf $n \cdot \epsilon$ beschränkt.*

Definition 2.2 (Rundungsfehler) *Der Rundungsfehler eines numerischen Verfahrens ist der durch Gleitkommarechnung enstandene Fehler des Endergebnisses bei exakten reellen Eingabedaten.*

Definition 2.3 (Gleitkommazahl) *Eine B-adische, m-stellige normalisierte Gleitkommazahl hat die Form:*

$$x = \pm B^e \sum_{k=-m}^{-1} x_k B_k; \quad x_{-1} \neq 0 \text{ (normalisiert)}, \quad x_k \in \{0, 1, \ldots, B-1\}$$

$e \in \mathbb{Z}$ *Exponent,* B *Basis,* x_k *Ziffern der Mantisse*

Eine zentrale Frage ist die Fehlerfortpflanzung von Störungen $\Delta x = (\Delta x_1, \ldots, \Delta x_n)^T \in \mathbb{R}^n$ einer Abbildung $F : \mathbb{R}^n \to \mathbb{R}$ auf das Ergebnis $F(x + \Delta x)$.

Bei Stetigkeit folgt aus dem Mittelwertsatz der Differentialrechnung die Fehlerabschätzung:

$$\begin{aligned} |F(x + \Delta x) - F(x)| &= \left| \sum_{j=1}^{n} \frac{\partial F(x + \tau \Delta x)}{\partial x_j} \cdot \Delta x_j \right| \\ &\leq \max_{1 \leq j \leq n} |\Delta x_j| \cdot \max_{0 \leq \tau \leq 1} \left\{ \sum_{j=1}^{n} \left| \frac{\partial F(x + \tau \cdot \Delta x)}{\partial x_j} \right| \right\} \end{aligned}$$

Somit ist der Vergrößerungsfaktor des absoluten Fehlers durch die Ableitung wesentlich bestimmt.

Definition 2.4 (Kondition) *Die Kondition eines Problems ist der maximale Vergrößerungsfaktor für den Einfluss von relativen Eingangsfehlern auf relative Resultatfehler. Ist die Kondition groß, heißt das Problem schlecht konditioniert (i.A. mehrere Zehnerpotenzen).*

Bemerkung 2.3 *Der Verlust an relativer Genauigkeit bei Differenzbildung fast gleicher Zahlen ist die schwerwiegendste Fehlerquelle im numerischen Rechnen.*

Beispiel 2.2 *Wir nehmen eine Arithmetik mit 2-stelliger Genauigkeit an.*
Es sind $x = 0{,}344152$ *und* $y = 0{,}344135$ *gegeben.*
Die Differenz $x - y = 0{,}000017 = 0{,}17 \cdot 10^{-4}$.
Das bedeutet einen Verlust von 2 Stellen Genauigkeit.
Ein Fehler von $0{,}01\%$ *in* x *bewirkt einen relativen Fehler von* 200% *in* $x - y$.

2.2 Landausymbole

Die Landausymbole sind ein wichtiges Hilfsmittel zur quantitativen Beschreibung von Grenzprozessen. Man schreibt:

$$f = O(g)$$

für $x \to x_0$, falls $\exists C > 0$ und ϵ mit $|x - x_0| < \epsilon$, sodass $|f(x)| \leq C\,|g(x)|$. Entsprechend schreibt man

$$f = o(g)$$

für $x \to x_0$, falls $\forall C > 0$ ein $\epsilon(C)$ existiert, sodass $|f(x)| \leq C \cdot |g(x)|$ für alle x mit $|x - x_0| \leq \epsilon$.