

5 Fehleranalyse (Stabilität)

Daten x_1, \dots, x_n $\xrightarrow{\text{Vorschrift}}$ Resultate z_1, \dots, z_m

$$F(x) = z, \quad F_j(x_1, \dots, x_n) = z_j, \quad j = 1, \dots, m.$$

Vorschrift/Funktion/Verfahren liefert bei Störungen eine Näherung \tilde{z} für z :

absoluter Fehler $\varepsilon := \tilde{z} - z$,

relativer Fehler $\delta := \frac{\tilde{z} - z}{z} = \frac{\varepsilon}{z}$ ($z \neq 0$) $\Rightarrow \tilde{z} = z(1 + \delta)$;

ein relativer Fehler von 0.01 bedeutet 1% Fehler.

Beispiele

- Berechnung eines Funktionswertes $\ln(x - \sqrt{x^2 - 1}) = z$
- Polynomauswertung $p(x) = z \rightarrow$ Horner-Schema
- LGS $Ax = b \rightarrow$ Gauß-Elimination

Zahldarstellung und Gleitkomma-Arithmetik

Gleitkommazahlen: $\xi = \pm a_{10^b}$ mit $0.1 \leq a < 1$ (normalisiert)

- Mantisse a : $a = 0.\alpha_1\alpha_2 \dots \alpha_t$ mit t Stellen
- Exponent b : ganze Zahl mit e Stellen

Maschinenzahl: Zahl, die in der Maschine mit t -stelliger Mantisse und e -stelligem Exponenten exakt dargestellt werden kann.

Menge der Maschinenzahlen (bezüglich Basis 10): $M_{10}(t, e)$

Definition: Unter einer Rundung versteht man die Approximation einer reellen Zahl x durch die nächstgelegene Maschinenzahl $rd(x)$.

Dargestellt sei x als normalisierte Gleitkommazahl

$$x = \alpha_{10^b}, \quad |\alpha| = 0.\alpha_1 \dots \alpha_t \alpha_{t+1} \dots, \quad \alpha_1 \neq 0, \quad x \notin M_{10}(t, e),$$

dann gilt

$$rd(x) = \text{sign}(x) \cdot a_{10^b}, \quad \text{wobei } a := \begin{cases} 0.\alpha_1 \dots \alpha_t, & 0 \leq \alpha_{t+1} \leq 4, \\ 0.\alpha_1 \dots \alpha_t + 10^{-t}, & 5 \leq \alpha_{t+1} \leq 9. \end{cases}$$

Satz 5.1 Bei t -stelliger normalisierter Gleitkomma-Arithmetik (mit $e = \infty$) gilt für den relativen Rundungsfehler die Abschätzung

$$rd(x) = x(1 + \delta), \quad |\delta| \leq 5 \cdot 10^{-t}$$

mit der Maschinengenauigkeit $\text{eps} := 5 \cdot 10^{-t}$.

Arithmetische Operationen $* \in \{ +, -, \cdot, : \}$

$$rd(x * y) = (x * y)(1 + \delta), \quad |\delta| \leq \text{eps}$$

Fehlerarten

1. *Eingangsfehler*: Fehlerhafte Daten $\tilde{x}_1, \dots, \tilde{x}_n$
 $\tilde{x}_j = x_j + \varepsilon_j$, ε_j absoluter Fehler,
 $\tilde{x}_j = x_j(1 + \delta_j)$, δ_j relativer Fehler.
2. *Formelfehler* (Verfahrensfehler) \hat{F} statt F :
 z.B. n -tes Glied einer Folge statt Grenzwert, Differenzenquotient
 statt Ableitung, Riemann-Summe/Quadraturformel statt Integral
3. *Rundungsfehler* (RF) entstehen während der laufenden Rechnung
4. *Menschlicher Irrtum* und *Maschinenfehler* \rightarrow Kontrollmaßnahmen

Stabilität/Instabilität

Auswirkung von Störungen \tilde{x} statt x auf die Resultate $F(\tilde{x}) = \tilde{z}$:

$$|\text{Auswirkung}| \leq K \cdot |\text{Störung}|$$

K Verstärkungsfaktor oder Konditionszahl:

gut konditioniert z. B. $K = 1 \dots 10 \dots 100$
 schlecht konditioniert z.B. $K = 10^3 \dots 10^6$

Natürliche Stabilität/Instabilität

Einfluss der Eingangsfehler auf das Endresultat \rightarrow Kondition!

Numerische Stabilität/Instabilität

Einfluss der Rundungsfehler auf das Endresultat!

Beispiele

1. *Funktionsauswertung*

$$z = F(x) := \ln(x - \sqrt{x^2 - 1}), \quad F(20) = -3.68 \dots$$

Algorithmus:

- (a) $\sqrt{x^2 - 1}$ Heron-Verfahren \Rightarrow Formelfehler
- (b) $u - v$ Auslöschung \Rightarrow Rundungsfehler
- (c) $\ln y$ Näherung \Rightarrow Formelfehler

$$\text{Formelfehler } \hat{F} \text{ statt } F \quad \hat{z} = \hat{F}(x), \quad |\hat{z} - z| \leq ?$$

$$\text{Eingangsfehler } \tilde{x} \text{ statt } x \quad \tilde{z} = \hat{F}(\tilde{x}), \quad |\tilde{z} - z| \leq ?$$

$$\text{Rundungsfehler } \tilde{\tilde{z}} \text{ statt } \tilde{z} \quad |\tilde{\tilde{z}} - z| \leq ?$$

2. *LGS* $Ax = b$: Gauß-Elimination $x = F(A, b)$

Eingangsfehler:

$$\left. \begin{array}{l} A + \Delta A \quad \text{statt } A \\ b + \Delta b \quad \text{statt } b \end{array} \right\} \Rightarrow (A + \Delta A)\tilde{x} = b + \Delta b$$

Fehlerfortpflanzung: Einfluss auf $\tilde{x} = x + \Delta x \rightarrow$ Verstärkungsfaktoren

Keine Formelfehler, da äquivalente Umformungen!

Rundungsfehler im 1. Schritt sind Eingangsfehler im 2. Schritt, usw.,

d.h. Rundungsfehleranalyse ist notwendig.

Fehlerfortpflanzung und Stabilität

Daten x_1, \dots, x_n , Resultate z_1, \dots, z_m :

$$F(x) = z, F_j(x_1, \dots, x_n) = z_j, j = 1, \dots, m, F \text{ hinreichend glatt}$$

Gestörte Daten $\tilde{x}_j = x_j + \varepsilon_j = x_j(1 + \delta_j) \rightarrow F(\tilde{x}) =: \tilde{z}$

Fehler $\varepsilon := \tilde{x} - x, \delta := \tilde{z} - z, \tilde{z}_j := z_j + \tilde{\varepsilon}_j = z_j(1 + \tilde{\delta}_j)$

Taylor-Entwicklung von $F(\tilde{x})$ um x ergibt:

Satz 5.2 *Es gelten folgende Fehlerdarstellungen ("Zwischenstelle" ξ):*

$$\text{absoluter Fehler } \tilde{z}_j - z_j = \tilde{\varepsilon}_j = \underbrace{\sum_{k=1}^n \left(\frac{\partial F_j}{\partial x_k} \right)_{|\xi}}_{VF} \varepsilon_k,$$

$$\text{relativer Fehler } \frac{\tilde{z}_j - z_j}{z_j} = \tilde{\delta}_j = \underbrace{\sum_{k=1}^n \left(\frac{\partial F_j}{\partial x_k} \right)_{|\xi} \frac{x_k}{z_j}}_{VF} \delta_k.$$

Anmerkungen: Verstärkungsfaktoren (VF) heißen Konditionszahlen. Partielle Ableitungen sind groß, falls Lipschitzkonstante groß ist; $\frac{x_k}{z_j}$ ist groß, falls kleine Resultate aus großen Daten berechnet werden.

Arithmetische Operationen

Addition $z = F(x_1, x_2) := x_1 + x_2, \tilde{z} = \tilde{x}_1 + \tilde{x}_2$

$$\text{absoluter Fehler } \tilde{\varepsilon} := \tilde{z} - z = 1 \cdot \varepsilon_1 + 1 \cdot \varepsilon_2$$

$$\text{relativer Fehler } \tilde{\delta} := \frac{\tilde{z} - z}{z} = \underbrace{\frac{x_1}{x_1 + x_2}}_{VF} \delta_1 + \underbrace{\frac{x_2}{x_1 + x_2}}_{VF} \delta_2$$

Verstärkungsfaktoren sind groß, falls $x_1 \approx -x_2 \rightarrow$ **Auslöschung**

Multiplikation $z = F(x_1, x_2) := x_1 \cdot x_2, \tilde{z} = \tilde{x}_1 \cdot \tilde{x}_2$

$$\text{absoluter Fehler } \tilde{\varepsilon} := \tilde{z} - z = x_2 \cdot \varepsilon_1 + x_1 \cdot \varepsilon_2 + \varepsilon_1 \varepsilon_2 \doteq x_2 \cdot \varepsilon_1 + x_1 \cdot \varepsilon_2$$

(\doteq bedeutet "in 1. Näherung"; Vernachlässigung höherer Potenzen);

$$\text{relativer Fehler } \frac{\tilde{z} - z}{z} = \tilde{\delta} = \delta_1 + \delta_2 + \delta_1 \delta_2 \doteq 1 \cdot \delta_1 + 1 \cdot \delta_2.$$

Division $z = F(x_1, x_2) := \frac{x_1}{x_2}, \tilde{z} = \frac{\tilde{x}_1}{\tilde{x}_2};$

$$\text{absoluter Fehler } \tilde{\varepsilon} := \tilde{z} - z \doteq \frac{1}{x_2} \cdot \varepsilon_1 - \frac{x_1}{x_2^2} \cdot \varepsilon_2;$$

$$\text{relativer Fehler } \tilde{\delta} := \frac{\tilde{z} - z}{z} \doteq 1 \cdot \delta_1 - 1 \cdot \delta_2;$$

VF bei absolutem Fehler sind groß, falls Division mit kleiner Zahl.

Vorsicht: Bei Addition kann **Auslöschung** auftreten!

Bei Division mit kleiner Zahl große VF!

Stabilität numerischer Grundaufgaben

1. Rekursion / Iteration

$$x_0 = a, \quad x_{\nu+1} = \varphi(x_\nu), \quad \nu = 0, 1, \dots, n-1 \quad (\varphi : I \rightarrow \mathbb{R}, \|\varphi'\| \leq L).$$

$$\text{Eingangsfehler: } \tilde{x}_0 = a + \varepsilon = x_0(1 + \delta), \quad \tilde{x}_{\nu+1} = \varphi(\tilde{x}_\nu), \quad \nu = 0, 1, 2, \dots$$

Ergebnis 5.3 Es gelten die Abschätzungen

$$|\tilde{x}_n - x_n| \leq L^n \varepsilon, \quad \left| \frac{\tilde{x}_n - x_n}{x_n} \right| \leq L^n \left| \frac{a}{x_n} \right| \delta.$$

Verstärkungsfaktor groß bzw. klein, falls $L > 1$ und $n \geq n_0$ bzw. $L \leq 1$.
Iterationsverfahren erfüllen $L < 1$ und sind daher gut konditioniert.

2. Polynomauswertung $p(x) = a_n x^n + \dots + a_0$

Störungen (Eingangsfehler) in den Koeffizienten

$$\tilde{a}_j := a_j(1 + \delta_j), \quad |\delta_j| \leq \delta, \quad \tilde{p}(x) = \sum_{j=0}^n \tilde{a}_j x^j.$$

Ergebnis 5.4 Es gilt die Abschätzung

$$\left| \frac{\tilde{p}(x) - p(x)}{p(x)} \right| \leq \frac{\sum_{j=0}^n |a_j| \cdot |x|^j}{\left| \sum_{j=0}^n a_j x^j \right|} \delta.$$

Verstärkungsfaktor groß, falls die $a_j x^j$ alternierendes Vorzeichen haben.

3. Skalarprodukt $x^T y = \sum_{j=1}^n x_j y_j = z$

Störungen in den Koeffizienten

$$\tilde{x}_j := x_j(1 + \delta_j), \quad |\delta_j| \leq \delta, \quad \tilde{y}_j := y_j(1 + \gamma_j), \quad |\gamma_j| \leq \gamma \Rightarrow \tilde{z} := \tilde{x}^T \tilde{y}$$

Ergebnis 5.5 Es gilt die Abschätzung (in 1. Näherung)

$$\left| \frac{\tilde{z} - z}{z} \right| \leq (\delta + \gamma) \frac{\sum_{j=1}^n |x_j| |y_j|}{\left| \sum_{j=1}^n x_j y_j \right|}.$$

4. Lineares Gleichungssystem $Ax = b$

$\|\cdot\|$, $N(\cdot)$ seien verträgliche Normen;

Störung in rechter Seite $\tilde{b} = b + \Delta b$, $A\tilde{x} = \tilde{b} \Rightarrow \tilde{x} = x + \Delta x$.

Ergebnis 5.6 Betrachte Störung in b , dann gilt die Abschätzung

$$\frac{\|\Delta x\|}{\|x\|} \leq N(A)N(A^{-1}) \frac{\|\Delta b\|}{\|b\|}.$$

Definition: Für nichtsinguläre Matrix A wird die *Kondition* (*Konditionszahl*) bezüglich einer Matrixnorm $N(\cdot)$ definiert als

$$\text{cond}_N(A) := N(A)N(A^{-1}).$$

Rundungsfehler

Numerische Stabilität/Instabilität: Auswirkung der Rundungsfehler (RF)

Gleitkomma-Arithmetik

$$rd(a) = a(1 + \delta) \quad \text{bzw.} \quad \frac{rd(a)-a}{a} = \delta$$

Kurzschrift (Symbolische Schreibweise)

$$rd(a) = a\rho \quad \text{bzw.} \quad \frac{rd(a)-a}{a} = \rho - 1$$

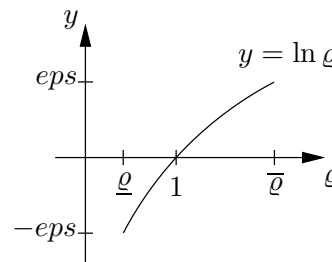
$|\ln \rho| \leq eps$ Maschinengenauigkeit.

$|\ln \rho| \doteq |\rho - 1|$ (\doteq Vernachlässigung höherer Potenzen), denn

$$\ln \rho = \ln(1 + (\rho - 1)) = \rho - 1 - \frac{(\rho-1)^2}{2} + \frac{(\rho-1)^3}{3} - + \dots$$

RF-Einheit $\pm(\rho - 1) \rightarrow$ maximaler Wert = eps

Kennzeichnung (Indizierung) weggelassen!



Relativer Fehler:

$a\rho$ bedeutet

$$a\rho \doteq a \pm (\rho - 1) \cdot |a| \quad \text{bzw.} \quad \frac{a\rho - a}{|a|} \doteq \pm(\rho - 1);$$

$a\rho^k$ bedeutet

$$a\rho^k \doteq a \pm k(\rho - 1)|a| \quad \text{bzw.} \quad \frac{a\rho^k - a}{|a|} \doteq \pm k(\rho - 1),$$

d.h. a ist mit k RF-Einheiten behaftet.

Rechenregeln: $\rho \cdot \rho = \rho^2$, $\frac{1}{\rho} = \rho^{-1}$, $\frac{\rho}{\rho} = \rho^0$.

Beispiel: Skalarprodukt $x^T y = z$

mit Kennzeichnung:

$$\begin{aligned} & ((x_1 y_1 (1 + \delta_1) + x_2 y_2 (1 + \delta_2))(1 + \delta_3) + x_3 y_3 (1 + \delta_4))(1 + \delta_5) \\ &= x_1 y_1 (1 + \delta_1)(1 + \delta_3)(1 + \delta_5) + x_2 y_2 (1 + \delta_2)(1 + \delta_3)(1 + \delta_5) + x_3 y_3 (1 + \delta_4)(1 + \delta_5) \\ &\doteq x_1 y_1 (1 + \delta_1 + \delta_3 + \delta_5) + x_2 y_2 (1 + \delta_2 + \delta_3 + \delta_5) + x_3 y_3 (1 + \delta_4 + \delta_5) =: \tilde{z} \end{aligned}$$

mit Kurzschrift:

$$\begin{aligned} & ((x_1 y_1 \rho + x_2 y_2 \rho) \rho + x_3 y_3 \rho) \rho = x_1 y_1 \rho^3 + x_2 y_2 \rho^3 + x_3 y_3 \rho^2 =: \tilde{z} \\ & \tilde{z} - z = x_1 y_1 \underbrace{(\rho^3 - 1)}_{\doteq \pm 3(\rho-1)} + x_2 y_2 (\rho^3 - 1) + x_3 y_3 (\rho^2 - 1) \end{aligned}$$

daraus folgt $\frac{|\tilde{z}-z|}{|z|} \leq 3 \cdot eps \cdot \frac{|x_1 y_1| + |x_2 y_2| + |x_3 y_3|}{|x_1 y_1 + x_2 y_2 + x_3 y_3|}$ (allgemein $\frac{|\tilde{z}-z|}{|z|} \leq n \cdot eps \cdot \frac{\sum |x_j y_j|}{|\sum x_j y_j|}$).

Idee: Rundungsfehler als Störungen in den Eingangsdaten auffassen
 → Rückwärtsanalyse!

$$\text{Vergleich mit Ergebnis 5.5: } \left| \frac{\tilde{z}-z}{z} \right| \leq \underbrace{(\delta + \gamma)}_{\text{Eingangsfehler}} \cdot \underbrace{\frac{\sum |x_j| |y_j|}{|\sum x_j y_j|}}_{\text{Konditionszahl}}$$

Algorithmus/Verfahren für $z = F(x)$: $F = g_m \circ \dots \circ g_1$

Faktorisierung der Abbildung, d.h. Zerlegung in Unteralgorithmen;
 Zwischenresultate $z^{(j)} := g_j(z^{(j-1)})$, $j = 1, \dots, m$, $z^{(0)} := x$;
 ein Rundungsfehler, der bei der Berechnung von $g_j(\cdot)$ auftritt,
 geht in die Restabbildung $h_j := g_m \circ \dots \circ g_{j+1}$ als Eingangsfehler ein
 und wird dann an das Endresultat weitergegeben.

Definition: Ein Algorithmus heißt *numerisch instabil*, wenn die natürliche Instabilität der Restabb. h_1, \dots, h_{m-1} diejenige von F deutlich übertrifft, d.h. wenn eine der Konditionszahlen von g_1, \dots, g_m um einen Faktor κ (z.B. $\kappa = 50 \times$ Dimension des Problems) größer als die Konditionszahl von F ist; sonst heißt der Algorithmus *numerisch stabil*.

Beispiel

Berechne $z = F(x) := \ln(x - \sqrt{x^2 - 1})$ für $x = 20$; Ergebnis $z = -3.68 \dots$
 Relativer Fehler: Verstärkungsfaktor $\left| \frac{x}{F(x)} F'(x) \right| = 0.2713 \dots$,
 d.h. gut konditioniertes Problem (natürlich stabil, gedämpft).

Algorithmus: $F = g_3 \circ g_2 \circ g_1$:

$$\begin{aligned} g_1(x) &:= \sqrt{x^2 - 1} \\ g_2(u, v) &:= u - v \quad \rightarrow \text{Auslöschung} \\ g_3(y) &:= \ln y \end{aligned}$$

$$h_1 := g_3 \circ g_2, \quad h_2 := g_3, \quad h_1(u, v) = \ln(u - v)$$

Verstärkungsfaktor für h_1 (Satz 5.2): $\frac{1}{u-v} \cdot \frac{u}{\ln(u-v)} \approx \frac{1}{0.025} \cdot \frac{20}{-3.5} \approx -215$

⇒ Algorithmus ist numerisch instabil !

Besser: $z = F(x) := -\ln(x + \sqrt{x^2 - 1})$, $x = 20$, $z = -3.68 \dots$

Algorithmus $F = g_3 \circ g_2 \circ g_1$:

$$\begin{aligned} g_1(x) &:= \sqrt{x^2 - 1}, \quad g_2(u, v) := u + v, \quad g_3(y) := -\ln y; \\ h_1 &:= g_3 \circ g_2, \quad h_2 := g_3, \quad h_1(u, v) = -\ln(u + v) \end{aligned}$$

Konditionszahl für

- h_1 : $-\frac{1}{u+v} \cdot \frac{u}{-\ln(u+v)} \approx -\frac{1}{40} \cdot \frac{20}{-3.69} \approx \frac{1}{7.4} \approx 0.13$
- h_2 : $-\frac{1}{y} \cdot \frac{y}{-\ln y} \approx -\frac{1}{40} \cdot \frac{40}{-3.69} \approx \frac{1}{3.69} \approx 0.27$

⇒ Algorithmus ist numerisch stabil !

Rückwärtsanalyse der Rundungsfehler

Algorithmus rückwärts durchgehen und den Einfluss der RF zurückspielen auf Störungen in den Eingangsdaten $F(\tilde{x}) = \tilde{z}$:

$$\text{Fehler } \left| \frac{\tilde{x}-x}{x} \right| \leq N \cdot \text{eps} \Rightarrow \left| \frac{\tilde{z}-z}{z} \right| \leq N \cdot \text{eps} \cdot \text{cond}(F)$$

Beispiele

1. *Skalarprodukt* $x^T y = z$ (n Multipl., $n - 1$ Add.) (Ergebnis 5.5)

$$\text{Störungen } \tilde{x}_j = x_j(1 + \delta_j), \tilde{y}_j = y_j(1 + \gamma_j)$$

$$\Rightarrow \frac{\tilde{z}-z}{z} \doteq \sum' \frac{x_j y_j}{z} (\delta_j + \gamma_j) \Rightarrow \frac{|\tilde{z}-z|}{|z|} \leq n \cdot \varepsilon \sum \frac{|x_i y_i|}{|z|},$$

wobei $\delta_j + \gamma_j = \pm(j+1)(\rho-1)$ (Aufteilung der RF-Einheiten).

2. *Polynomauswertung* (nach Horner: n Multipl., n Add.)

$$\begin{aligned} p(x) &= a_n x^n + a_{n-1} x^{n-1} + \dots + a_0 \\ &= \left(\dots \left((a_n x + a_{n-1}) x + a_{n-2} \right) x + \dots + a_1 \right) x + a_0. \end{aligned}$$

Rückwärtsanalyse der Rundungsfehler:

$$\begin{aligned} &\left(\left(\dots \left((a_n x \rho + a_{n-1}) \rho x \rho + a_{n-2} \right) \rho x \rho + \dots + a_1 \right) \rho x \rho + a_0 \right) \rho \\ &= \left(\left(\dots \left((a_n x^2 \rho^3 + a_{n-1} x \rho^2 + a_{n-2}) \rho x \rho + \dots + a_1 \right) \rho x \rho + a_0 \right) \rho \right. \\ &\quad \vdots \\ &= a_n x^n \rho^{2n} + a_{n-1} x^{n-1} \rho^{2n-1} + a_{n-2} x^{n-2} \rho^{2n-3} + \dots + a_1 x \rho^3 + a_0 \rho \\ &= \sum_{j=0}^n \underbrace{(a_j \rho^{2j+1})}_{=\tilde{a}_j} x^j \quad (j = n: \text{Potenz } \rho^{2n}) \end{aligned}$$

Statt $p(x) = \sum_j a_j x^j$ wird das Polynom $\tilde{p}(x) = \sum_j \tilde{a}_j x^j$ mit gestörten Koeffizienten $\tilde{a}_j = a_j \rho^{2j+1} = a_j \pm (2j+1)(\rho-1)|a_j|$ ausgewertet.

Fehlerfortpflanzung (Ergebnis 5.4): Störung in den Koeffizienten

$$\tilde{a}_j = a_j(1 + \delta_j), \quad |\delta_j| \leq \delta = 2n \cdot \text{eps}$$

Ergebnis 5.7 Das Horner-Schema liefert eine Näherung $\tilde{p}(x)$, für deren relativen Fehler gilt

$$\left| \frac{\tilde{p}(x) - p(x)}{p(x)} \right| \leq \underbrace{2n}_{\text{numer. Kond.}} \cdot \text{eps} \cdot \underbrace{\frac{\sum |a_j| |x^j|}{|\sum a_j x^j|}}_{\text{natürl. Kond.}}.$$

Numerischer Verstärkungsfaktor ist $2 \times$ Polynomgrad.

Polynomauswertung ist numerisch stabil !

Beispielblatt: Fehleranalyse

Aufgabe: Berechne $z = e^x$ für $x = 1$ mittels Taylorentwicklung
mit einer absoluten bzw. relativen Genauigkeit von 10^{-6}

Konditionsabschätzungen: Mit $\tilde{x} = x \pm \varepsilon$ ($\varepsilon > 0$) folgt

$$|\tilde{z} - z| \leq e^{x+\varepsilon} |\tilde{x} - x| \quad \text{bzw.} \quad \left| \frac{\tilde{z} - z}{z} \right| \leq e^\varepsilon \left| \frac{\tilde{x} - x}{x} \right|$$

(Verstärkungsfaktoren in der Größenordnung e bzw. 1).

Näherungsverfahren: $v_n(x) = \sum_{\nu=0}^n \frac{1}{\nu!} x^\nu$,

es gelten dieselben Konditionsabschätzungen.

Formelfehler: $z - v_n(1) = \frac{1}{(n+1)!} e^\xi$ ($0 < \xi < 1$)

Wähle $n = 9$: $|z - v_9(1)| < e \cdot 0.27 \cdot 10^{-6}$, $\left| \frac{z - v_9(1)}{z} \right| < 0.27 \cdot 10^{-6}$

Rundungsfehler: Auswertung von

$$v_9(1) = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \dots + \frac{1}{9!} = (\dots ((\frac{1}{9} + 1)\frac{1}{8} + 1)\frac{1}{7} + \dots + 1)\frac{1}{2} + 1 + 1$$

Gleitkommazahlen (Mantissenlänge t): $\left| \frac{rd(x) - x}{x} \right| \leq eps := 5 \cdot 10^{-t}$

Rückwärtsanalyse (in "Kurzschrift"):

$$\hat{v}_9 = ((\dots (\frac{1}{9} + 1) \rho \frac{1}{8} \rho \dots + 1) \rho \frac{1}{2} \rho + 1) \rho \frac{1}{1} \rho + 1 = \sum_{\nu=0}^9 \frac{\rho^{2\nu}}{\nu!}$$

also \hat{v}_9 auffassen als $\hat{v}_9 = v_9(1 + (\rho^2 - 1))$ mit Störung $|\rho^2 - 1| \doteq 2 \cdot eps$

Fehlerfortpflanzung:

$$|\hat{v}_9 - v_9(1)| \leq e^{1+2 \cdot eps} \cdot 2 \cdot eps \quad \text{bzw.} \quad \left| \frac{\hat{v}_9 - v_9(1)}{v_9(1)} \right| \leq e^{2 \cdot eps} \cdot 2 \cdot eps$$

Wähle $t = 8$: $eps = 5 \cdot 10^{-8}$, $e^{2 \cdot eps} = 1.0000001$, $e^{1+2 \cdot eps} = e \cdot e^{2 \cdot eps}$

Absoluter Gesamtfehler:

$$|z - \hat{v}_9| \leq |z - v_9(1)| + |\hat{v}_9 - v_9(1)| < e(0.27 \cdot 10^{-6} + 10^{-7}) < \underline{1.006 \cdot 10^{-6}}$$

Relativer Gesamtfehler:

$$\left| \frac{z - \hat{v}_9}{z} \right| \leq \left| \frac{z - v_9(1)}{z} \right| + \left| \frac{\hat{v}_9 - v_9(1)}{v_9(1)} \right| < 0.27 \cdot 10^{-6} + 1.01 \cdot 10^{-7} < \underline{0.371 \cdot 10^{-6}}$$

Ergebnis: Mit $n = 9$ und $t = 8$ ergibt sich

Näherung $\hat{v}_9 = \underline{2.718\ 281\ 53}$

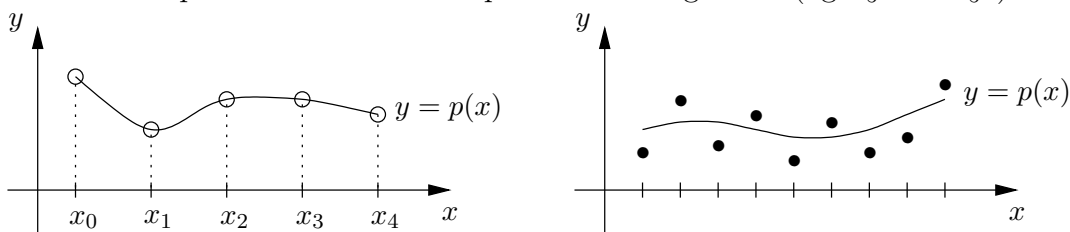
wahrer Wert: $e = 2.718\ 281\ 828 \dots$

absoluter Fehler $|e - \hat{v}_9| < 0.3 \cdot 10^{-6}$

relativer Fehler $\left| \frac{e - \hat{v}_9}{e} \right| < 0.11 \cdot 10^{-6}$

6 Algebraische Interpolation

Motivation: Gegeben sind Mess- oder Funktionswerte zu gewissen Zeitpunkten, durch die eine Kurve gelegt werden soll. Gesucht wird eine einfache Funktion $p(x)$, welche zu den gegebenen Zeitpunkten die Mess- bzw. Funktionswerte annimmt (linkes Bild). Gefordert wird, dass $p(x)$ einfach und dem Problem angepasst ist, z.B. algebraisches oder trigonometrisches Polynom, Spline, rationale Funktion. Bei sehr vielen Messpunkten (rechtes Bild) wird meistens eine Ausgleichsrechnung nach dem Prinzip der kleinsten Fehlerquadrate durchgeführt (vgl. §3 und §8).



Ziel: Approximation einer komplizierten Funktion durch eine einfache Funktion

Interpolation mit algebraischen Polynomen

Gegeben: Knoten $x_0, x_1, \dots, x_n \in \mathbb{R}$ (paarweise verschieden),
Daten $y_0, y_1, \dots, y_n \in \mathbb{R}$ (beliebig bzw. Funktionswerte $y_j := f(x_j)$).

Gesucht: Polynom $p \in \mathcal{P}_n$ mit $p(x_j) = y_j$, $j = 0, 1, \dots, n$
→ $n + 1$ Bedingungsgleichungen für $n + 1$ Koeffizienten

Vandermonde-Matrix

$$V_n = \begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{pmatrix}$$

$\det V_n = \prod_{\nu > \mu} (x_\nu - x_\mu) \neq 0$, falls $x_\nu \neq x_\mu$ für $\nu \neq \mu$.

Satz 6.1 Die Interpolationsaufgabe besitzt im Raum \mathcal{P}_n genau eine Lösung.

Beweis: Sei $p(x) = \sum_{\nu=0}^n a_\nu x^\nu$, $y = (y_0, \dots, y_n)^T$, $a = (a_0, \dots, a_n)^T$,

Werte einsetzen $p(x_j) = y_j$, $j = 0, 1, \dots, n$ äquivalent zu $V_n a = y$,
beachte $\det V_n \neq 0$, also eindeutige Lösung.

Variante (allgemein anwendbar): Inhomogenes Problem

$$p(x_j) = y_j, \quad j = 0, 1, \dots, n$$

ist genau dann eindeutig lösbar, wenn homogenes Problem $p(x_j) = 0$,
 $j = 0, 1, \dots, n$, nur die triviale Lösung besitzt. Da $p \in \mathcal{P}_n$ höchstens
 n Nullstellen besitzt oder identisch Null ist, muss $p = 0$ sein. \square

Darstellung des Interpolationspolynoms

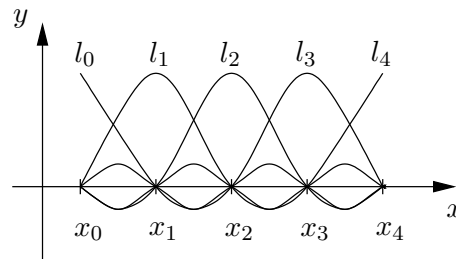
1. **Normalgestalt** $p(x) = \sum_{\nu=0}^n a_\nu x^\nu$
 Koeffizienten a_ν durch LGS $V_n a = y \rightarrow$ schlecht konditioniert
 z.B. äquidistante Knoten in $[0, 1]$:
 $\text{cond}_Z(V_5) \approx 10^3$, $\text{cond}_Z(V_{10}) \approx 10^7$, $\text{cond}_Z(V_{20}) \approx 10^{16}$

2. **Lagrange-Darstellung** $p(x) = \sum_{j=0}^n y_j \ell_j(x)$

Lagrange-Grundfunktionen

$$\ell_j(x) = \prod_{\substack{k=0 \\ k \neq j}}^n \frac{x-x_k}{x_j-x_k}, \quad \ell_j \in \mathcal{P}_n, \quad \ell_j(x_k) = \delta_{jk} \quad (\delta_{jk} \text{ Kronecker})$$

Knotenpolynom $w(x) = \prod_{\nu=0}^n (x - x_\nu)$, also $\ell_j(x) = \frac{w(x)}{(x-x_j)w'(x_j)}$



3. **Newton-Gestalt** $p(x) = \sum_{j=0}^n \alpha_j w_j(x)$

Grundfunktionen $w_0(x) = 1$, $w_j(x) = (x - x_0) \cdots (x - x_{j-1})$,
 $w_{n+1}(x)$ ist gerade das Knotenpolynom $w(x)$.

Vorteile dieser Darstellung:

- Koeffizienten α_i über dividierte Differenzen rekursiv definiert
- Einfache numerische Auswertung: Einschachtelung (Horner-ähnlich)

$$p(x) = (\cdots (\alpha_n(x - x_{n-1}) + \alpha_{n-1})(x - x_{n-2}) + \cdots + \alpha_1)(x - x_0) + \alpha_0$$
- Hinzunahme eines weiteren Knotens x_{n+1} :
 die bereits berechneten Koeffizienten $\alpha_0, \dots, \alpha_n$ bleiben unverändert:

$$p_{n+1} = p_n + \alpha_{n+1} w_{n+1} \rightarrow \text{ableitungsfreie Fehlerdarstellung (Satz 6.4)}$$

Dividierte Differenzen

Rekursive Berechnung der Koeffizienten α_j der Newton-Darstellung

$$\begin{aligned} p(x_0) &= \alpha_0 & &= y_0 \\ p(x_1) &= \alpha_0 + \alpha_1(x_1 - x_0) & &= y_1 \\ p(x_2) &= \alpha_0 + \alpha_1(x_2 - x_0) + \alpha_2(x_2 - x_0)(x_2 - x_1) & &= y_2 \\ &\vdots & &= \vdots \\ p(x_n) &= \alpha_0 + \cdots + \alpha_n(x_n - x_0) \cdots (x_n - x_{n-1}) & &= y_n \end{aligned}$$

Schema:

$$\begin{array}{c|ccc}
 0 & 1 & & \\
 & & \rangle & -2 \\
 1 & -1 & & \rangle \frac{7}{6} \\
 & & \rangle & \frac{3}{2} \\
 3 & 2 & & \rangle 0 \\
 & & \rangle & \frac{3}{2} \\
 2 & \frac{1}{2} & &
 \end{array}$$

Die IP-Polynome $p_2(x)$ und $p_3(x)$ für die ersten 3 bzw. 4 Knoten lauten:

$$p_2(x) = 1 - 2x + \frac{7}{6}x(x-1) \quad \text{bzw.} \quad p_3(x) = p_2(x) - \frac{7}{12}x(x-1)(x-3).$$

Eigenschaften dividierter Differenzen

1. Invarianz gegenüber Index-Vertauschungen; d.h. Wert der dividierten Differenz hängt nicht von der Reihenfolge der Knoten ab.
2. Zusammenhang mit Ableitungen:

Daten $y_j := f(x_j) \rightarrow$ Differenzen $f[x_0, x_1, \dots, x_k]$

$$f[x_0, x_1] := \frac{f(x_1) - f(x_0)}{x_1 - x_0} = f'(\eta) \quad (\text{Mittelwertsatz})$$

$$f[x_0, x_1, x_2] := \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = \frac{f''(\xi)}{2!} \quad (\text{Taylor}), \text{ usw.}$$

Approximation einer Funktion durch Interpolation

$f \in C[a, b]$:

Knoten $a \leq x_0 < \dots < x_n \leq b$ und Daten $y_j := f(x_j), j = 0, 1, \dots, n$

Interpolations-Operator

$$L_n : C[a, b] \rightarrow \mathcal{P}_n, \quad f \mapsto L_n f := p_f = \sum_{j=0}^n f(x_j) \ell_j$$

L_n linear, beschränkt und $L_n f = f$ für alle $f \in \mathcal{P}_n$

Satz 6.3 (Restglied von Cauchy) Sei $f \in C^{n+1}[a, b]$,
 $a \leq x_0 < \dots < x_n \leq b$. Dann gilt für jedes $x \in [a, b]$

$$f(x) - p_f(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi) w(x)$$

mit $\xi = \xi(x) \in (a, b)$ und Knotenpolynom $w(x) = (x - x_0) \cdots (x - x_n)$.

Beweis: *Satz von Rolle:*

$g \in C^1[a, b]$ mit $g(a) = g(b) = 0$, dann existiert $\xi \in (a, b)$ mit $g'(\xi) = 0$.
 $x \notin \{x_0, \dots, x_n\}$:

Hilfsfunktion $\phi(t) := f(t) - p_f(t) - \frac{w(t)}{w(x)} (f(x) - p_f(x))$, $\phi \in C^{n+1}[a, b]$.

$$\phi^{(n+1)}(t) = f^{(n+1)}(t) - \frac{(n+1)!}{w(x)} (f(x) - p_f(x))$$

$\phi(x) = 0$ und $\phi(x_j) = 0, j = 0, 1, \dots, n$, d.h.

ϕ besitzt mindestens $n + 2$ (paarweise verschiedene) Nullstellen in $[a, b]$.

Satz von Rolle:

$\phi'(t)$ besitzt mindestens $n + 1$ Nullstellen in (a, b)

$\phi''(t)$ besitzt mindestens n Nullstellen in (a, b)

\vdots

$\phi^{(n+1)}(t)$ besitzt mindestens 1 Nullstelle ξ in (a, b)

$\phi^{(n+1)}(\xi) = 0$ auflösen, dann folgt die Behauptung. \square

Vergleich mit Taylor–Rest bei Entwicklung um α :

$$f(x) - p(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi)(x - \alpha)^{n+1}$$

Satz 6.4 (Ableitungsfreies Restglied) Für den

Interpolationsfehler gilt mit $x \notin \{x_0, \dots, x_n\}$

$$f(x) - p_f(x) = f[x_0, \dots, x_n, x]w(x).$$

Beweis: Knoten $x_0, \dots, x_n \rightarrow$ IP–Polynom p_f

Interpolation nach Newton: Knoten x hinzufügen, dann gilt

$$p_{n+1} = p_f + \alpha_{n+1}w_{n+1} \quad \text{mit} \quad p_{n+1}(x) = f(x),$$

also $f(x) - p_f(x) = f[x_0, \dots, x_n, x]w(x)$. \square

Aus dieser Darstellung folgt sofort das Cauchy–Restglied (Satz 6.3):

Zusammenhang von dividierter Differenz zu Ableitung ausnützen.

Kubische Spline–Interpolation

Raum der kubischen Splines $S_3(\Delta)$ mit Dimension = $n + 3$

Gitter $\Delta : a = x_0 < x_1 < \dots < x_n = b$

Daten y_j (bzw. Funktionswerte $y_j = f(x_j)$), $j = 0, 1, \dots, n$

Gesucht: Interpolierender kubischer Spline $s \in S_3(\Delta)$ mit

$$s(x_j) = y_j \quad \text{für} \quad j = 0, 1, \dots, n$$

(beachte $n + 1$ Bedingungen für $n + 3$ Koeffizienten)

Randvorgaben (RVG): (a) $s'(x_0) = y_0^I, s'(x_n) = y_n^I$

($y_0^I, y_n^I \in \mathbb{R}$ zusätzlich gegeben) oder

(b) $s''(x_0) = s''(x_n) = 0$

Satz 6.5 Die Interpolationsaufgabe mit RVG (a) oder (b) ist eindeutig lösbar.

Darstellung des interpolierenden kubischen Splines $s(x)$

Ergebnis 6.6 Der interpolierende kubische Spline mit RVG (a) oder (b) lässt sich in jedem Intervall $[x_{j-1}, x_j]$ der Länge $h_j = x_j - x_{j-1}$ ($j = 1, \dots, n$) darstellen durch das Polynom

$$s(x) = \frac{1}{6h_j} \left(M_j(x - x_{j-1})^3 + M_{j-1}(x_j - x)^3 \right) + b_j \left(x - \frac{x_j + x_{j-1}}{2} \right) + a_j,$$

wobei

$$b_j = \frac{1}{h_j}(y_j - y_{j-1}) - \frac{1}{6}h_j(M_j - M_{j-1}),$$

$$a_j = \frac{1}{2}(y_j + y_{j-1}) - \frac{1}{12}h_j^2(M_j + M_{j-1}),$$

und die Momente M_j durch LGS mit symmetrischer Tridiagonalmatrix bestimmt sind:

RVG (a)

$$\left(\begin{array}{c|cccc|c} \frac{h_1}{3} & & & & & \\ \hline \frac{h_1}{6} & \frac{h_1}{6} & & & & \\ \frac{h_1}{6} & \frac{h_2+h_1}{3} & \frac{h_2}{6} & & & \\ & \frac{h_2}{6} & \frac{h_3+h_2}{3} & \frac{h_3}{6} & & \\ & & \ddots & \ddots & \ddots & \\ & & & \frac{h_{n-1}}{6} & \frac{h_n+h_{n-1}}{3} & \frac{h_n}{6} \\ \hline & & & \frac{h_n}{6} & \frac{h_n}{3} & \end{array} \right) \begin{pmatrix} M_0 \\ M_1 \\ \vdots \\ \vdots \\ M_{n-1} \\ M_n \end{pmatrix} = \begin{pmatrix} m_0 \\ m_1 \\ \vdots \\ \vdots \\ m_{n-1} \\ m_n \end{pmatrix},$$

$$m_j = \frac{1}{h_{j+1}}(y_{j+1} - y_j) - \frac{1}{h_j}(y_j - y_{j-1}), \quad j = 1, \dots, n-1,$$

$$m_0 = \frac{1}{h_1}(y_1 - y_0) - y_0^I, \quad m_n = -\frac{1}{h_n}(y_n - y_{n-1}) + y_n^I;$$

RVG (b) (d.h. $M_0 = M_n = 0$):

im LGS sind jeweils die ersten und letzten Zeilen und Spalten zu streichen.

Anmerkungen: Bei äquidistanten Knoten ($h_j = h$) lauten die Matrizen

$$RVG (a) : \frac{1}{6} \begin{pmatrix} 2 & 1 & & & & \\ 1 & 4 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & & 1 & 4 & 1 \\ & & & & 1 & 2 \end{pmatrix} \quad \text{mit } \text{cond} \leq 6,$$

$$RVG(b) : \frac{1}{6} \begin{pmatrix} 4 & 1 & & & & \\ 1 & 4 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & & 1 & 4 & 1 \\ & & & & 1 & 4 \end{pmatrix} \quad \text{mit } \text{cond} \leq 3.$$

Das LGS im Ergebnis 6.6 ist gut konditioniert, also ist das Verfahren der kubischen Spline-Interpolation numerisch stabil.

Beweisidee: Herleitung

s'' Polygonzug, Momente $M_j := s''(x_j)$, $j = 0, 1, \dots, n$ (unbekannt)

↓ Stammfunktion, Integrationskonstanten b_j

s' Stetigkeit

↓ Stammfunktion, Integrationskonstanten a_j

s Stetigkeit und Interpolationsbedingungen

Intervall $[x_{j-1}, x_j]$, $h_j = x_j - x_{j-1}$:

$$s''(x) = \frac{1}{h_j} (M_j(x - x_{j-1}) + M_{j-1}(x_j - x))$$

$$s'(x) = \frac{1}{2h_j} (M_j(x - x_{j-1})^2 - M_{j-1}(x_j - x)^2) + b_j$$

$$s(x) = \frac{1}{6h_j} (M_j(x - x_{j-1})^3 + M_{j-1}(x_j - x)^3) + b_j \left(x - \frac{x_j + x_{j-1}}{2}\right) + a_j$$

Unbekannte: $M_j, b_j, a_j \rightarrow$ Anzahl = $(n + 1) + n + n = 3n + 1$

Anzahl der Bedingungen: Stetigkeit für s' $\rightarrow n - 1$

Stetigkeit für s $\rightarrow n - 1$

IP-Bedingungen $\rightarrow n + 1$

Randvorgaben $\rightarrow \frac{2}{3n + 1}$

b_j und a_j durch die M_j ausdrücken, dann ergibt sich LGS für die M_j . \square

Approximation einer Funktion $f(x)$ mit Spline-Interpolation

Satz 6.7 Für den interpolierenden kubischen Spline s_f zu $f \in C^4[a, b]$ bezüglich äquidistanter Knoten mit RVG (a) oder (b) (falls $f \notin \mathcal{P}_3$) gilt

$$\|f - s_f\|_\infty \leq 2h^4 \|f^{(4)}\|_\infty.$$

Anmerkungen: Es gilt gleichmäßige Konvergenz mit $O(h^4)$ für $h \rightarrow 0$.

Vergleich mit Cauchy-Rest (Satz 6.3) für das algebraische IP-Polynom p_f

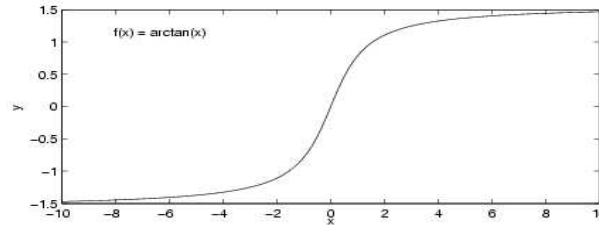
$$\|f - p_f\|_\infty \leq \frac{1}{(n+1)!} \underbrace{\|w\|_\infty}_{\approx \frac{1}{2^n}} \underbrace{\|f^{(n+1)}\|_\infty}_{\rightarrow ? \text{ für } n \rightarrow \infty}$$

d.h. i.A. ist gleichmäßige Konvergenz für Polynom-IP nicht gesichert!

Beispielblatt: Polynom- und kubische Spline-Interpolation

1. Interpolation von $f(x) = \arctan x$ im Intervall $[-10, 10]$ mit 21 Stützstellen:

$$-10, -9, \dots, -1, 0, +1, \dots, +9, +10$$



Fehler $e_p(x) := |\arctan x - p_{20}(x)|$ bei Polynom-IP,

Fehler $e_s(x) := |\arctan x - s_f(x)|$ bei Spline-IP mit $M_0 = M_n = 0$.

Qualitatives Verhalten der Fehlerfunktionen in Zwischenpunkten:

x	± 0.5	± 1.5	± 2.5	± 3.5	± 4.5	± 5.5	± 6.5	± 7.5	± 8.5	± 9.5
$e_p(x)$	0.02	0.02	0.01	0.02	0.01	0.02	0.06	0.3	1.5	16.1
$e_s(x)$	0.03	0.01	10^{-3}	10^{-3}	10^{-4}	10^{-4}	10^{-5}	10^{-6}	10^{-5}	10^{-4}

2. Interpolation der Daten:

x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
y	7	6	4	4	5	4	2	3	5	7	6	4	4	5	7

Funktionsverlauf des IP-Polynoms $p_{14}(x)$ und des interpolierenden kubischen Splines $s(x)$ mit $M_0 = M_n = 0$:

