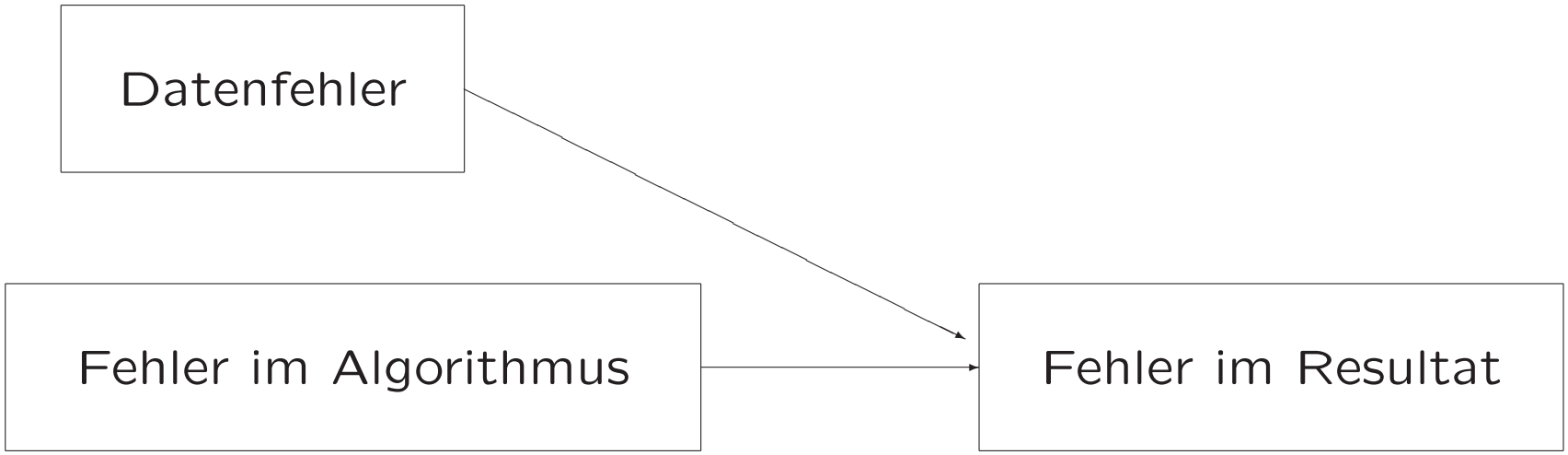

Datenfehler
Fehler im Algorithmus
Fehler im Resultat

Analyse der Fehlerverstärkung bei Datenfehlern:

Konzept der *Kondition* eines Problems.

Dies ist zunächst *unabhängig* von einem speziellen Lösungsweg (Algorithmus) und gibt nur an, welche Genauigkeit man bestenfalls (bei exakter Rechnung) bei gestörten Eingangsdaten erwarten kann.

Um dies etwas präziser beschreiben zu können, fassen wir den „mathematischen Prozeß“ oder das „Problem“ als Aufgabe auf, eine gegebene Funktion

$$f : X \to Y$$

an einer Stelle $x \in X$ *auszuwerten*.

Beispiel 2.1.
Die Berechnung der Multiplikation von x_1 und x_2:

$$f(x_1, x_2) = x_1 x_2,$$

und $X = \mathbb{R}^2$, $Y = \mathbb{R}$. △

Beispiel 2.2.
Die Berechnung der Summe von x_1 und x_2:

$$f(x_1, x_2) = x_1 + x_2,$$

und $X = \mathbb{R}^2, Y = \mathbb{R}$. △

Beispiel 2.3.
Man bestimme die kleinere der Nullstellen der Gleichung

$$y^2 - 2x_1 y + x_2 = 0,$$

mit $x_1^2 > x_2$. Die Lösung y^* ist

$$y^* = f(x_1, x_2) = x_1 - \sqrt{x_1^2 - x_2}$$

In diesem Fall gilt $X = \{\, (x_1, x_2) \in \mathbb{R}^2 \mid x_1^2 > x_2 \,\},\ Y = \mathbb{R}$. $\triangle$

Beispiel 2.4.
Bestimmung des Schnittpunktes zweier Geraden:

$$\begin{aligned} G_1 &= \{\,(y_1, y_2) \in \mathbb{R}^2 \mid a_{1,1}y_1 + a_{1,2}y_2 = x_1\,\} \\ G_2 &= \{\,(y_1, y_2) \in \mathbb{R}^2 \mid a_{2,1}y_1 + a_{2,2}y_2 = x_2\,\}, \end{aligned}$$

wobei $x = (x_1, x_2)^T \in \mathbb{R}^2$ und $a_{i,j}$ gegeben seien. Schreibt man kurz

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix},$$

so gilt

$$Ay = x.$$

Annahme: $\det A \neq 0$, Dann ist u gerade durch

$$y = A^{-1}x$$

gegeben. Also Auswertung der Funktion

$$f(x) = A^{-1}x,$$

d.h. $X = Y = \mathbb{R}^2$. △

Beispiel 2.5. Es soll das Integral

$$I_n = \int_0^1 \frac{t^n}{t+5}\, dt$$

für $n = 30$ berechnet werden. Für I_n $(n = 1, 2, \ldots)$ gilt die Rekursionsformel

$$I_n + 5I_{n-1} = \int_0^1 \frac{t^n + 5t^{n-1}}{t+5}\, dt = \int_0^1 t^{n-1}\, dt = \frac{1}{n}.$$

Durch die Rekursion

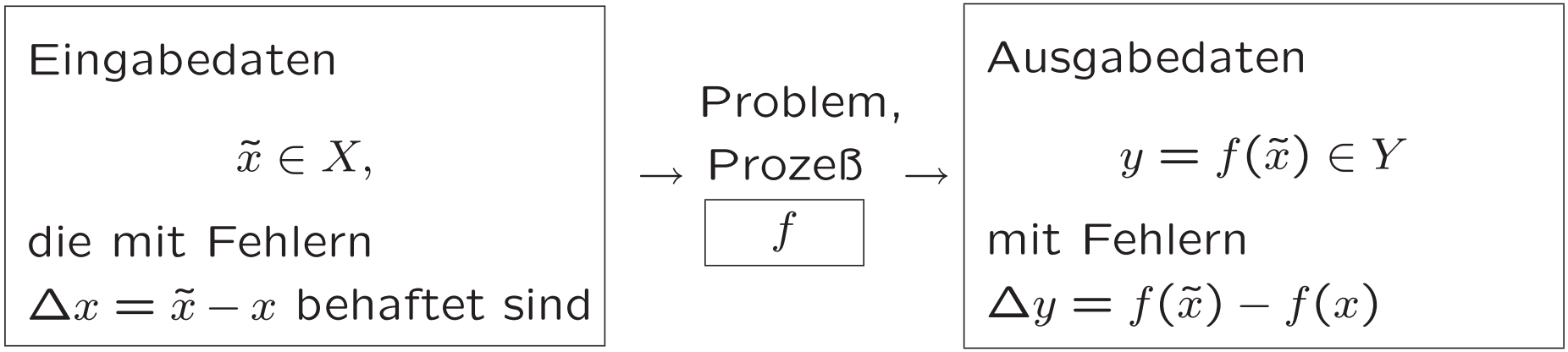
$$J_0 \in \mathbb{R}, \quad J_n = \frac{1}{n} - 5J_{n-1}, \quad n = 1, 2, \ldots, 30,$$

wird eine Funktion $f : \mathbb{R} \to \mathbb{R}$, $f(J_0) = J_{30}$ definiert.

Das Problem besteht in der Auswertung dieser Funktion an der Stelle I_0:

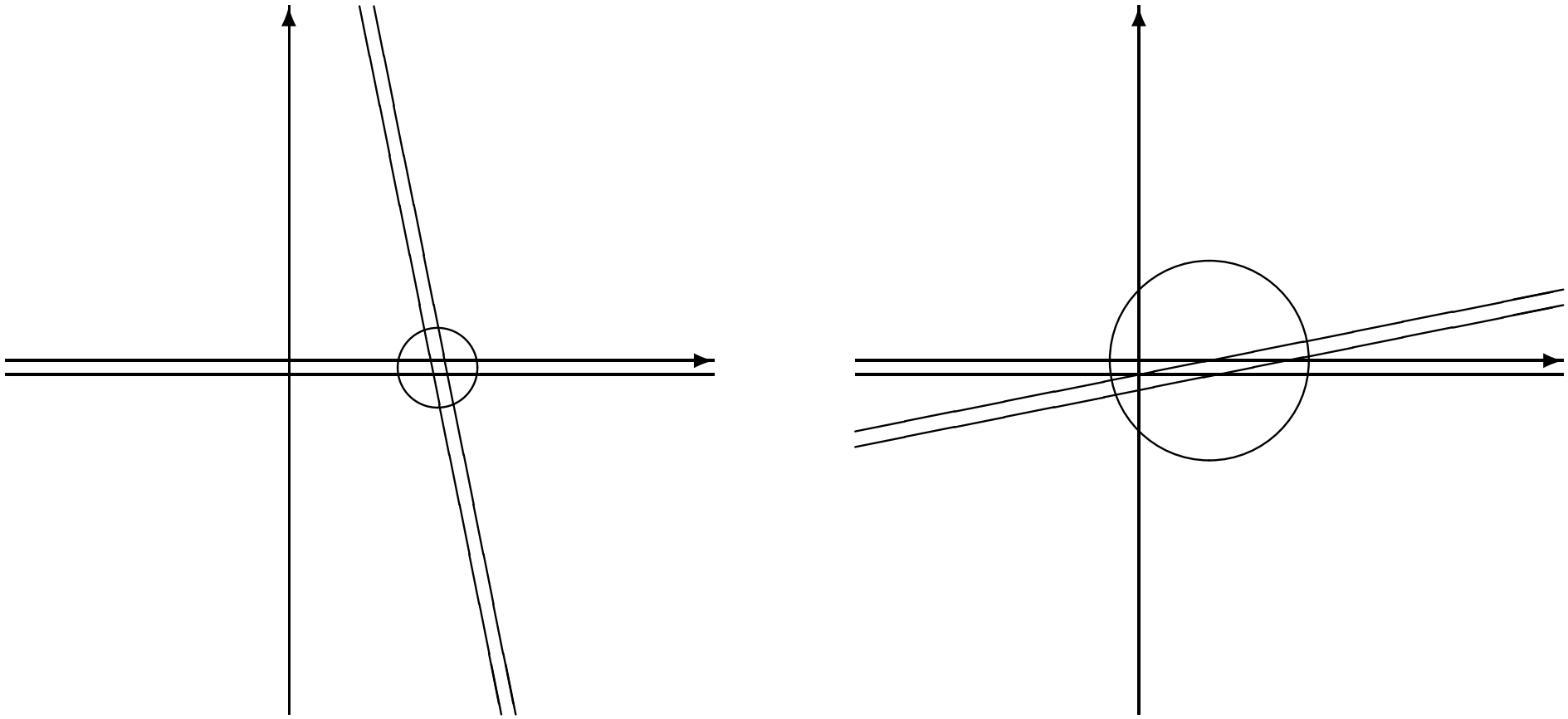
$$I_{30} = f(I_0) = f(\ln(\frac{6}{5})), \qquad X = Y = \mathbb{R}\ .$$

△



Es geht nun darum, den Ausgabefehler Δy ins Verhältnis zum Eingabefehler Δx zu setzen.

Kondition bei der Bestimmung des Schnittpunktes:



Eine Abbildung $\|\cdot\| : V \to \mathbb{R}$ heißt *Norm* auf V, falls

(N1) $\|v\| \geq 0$, $\forall\ v \in V$ und $\|v\| = 0$ impliziert $v = 0$;

(N2) Für alle $a \in \mathcal{K}, v \in V$ gilt $\|av\| = |a|\,\|v\|$;

(N3) Für alle $v, w \in V$ gilt die *Dreiecksungleichung*

$$\|v + w\| \leq \|v\| + \|w\|.$$

Wenn eine Norm auf V definiert ist, nennt man V oft einen *linearen normierten Raum*.

Die ∞- oder Sup-Norm

$$\|x\|_\infty := \max_{i=1,\dots,n} |x_i|, \quad x \in \mathbb{R}^n,$$
$$\|f\|_\infty := \|f\|_{L_\infty(I)} := \max_{t\in I} |f(t)|, \quad f \in C(I),$$

definiert eine Norm auf $\mathbb{R}^n$ bzw. $C(I)$.

Es kostet etwas mehr Mühe zu zeigen, daß für $1 \le p < \infty$ auch mit

$$\|x\|_p := \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}, \quad x \in \mathbb{R}^n,$$
$$\|f\|_p = \|f\|_{L_p(I)} := \left(\int_I |f(t)|^p\, dt\right)^{1/p}, \quad f \in C(I),$$

Normen auf $\mathbb{R}^n$ bzw. $C(I)$ gegeben sind. Für $p = 2$:

$$\|x\|_2 = (x,x)^{1/2}, \quad (x,y) := x^T y = \sum_{i=1}^n x_i y_i,$$

d.h, Norm wird durch ein *Skalarprodukt* induziert.

Beispiel 2.9. Man sollte beim Begriff „endlich-dimensionaler Vektorraum“ nicht nur an $\mathbb{R}^n$ denken:

$$\Pi_m := \Big\{ \sum_{i=0}^{m} a_i x^i \mid a_i \in \mathbb{R} \Big\}$$

ist ein $\mathbb{R}$-Vektorraum der Dimension $m+1$.

Die *Monome* $m_i(x) := x^i$, $i = 0, \ldots, m$, dienen hier als *Basis* (ein System von Elementen, deren Linearkombinationen den ganzen Raum ausfüllen und die *linear unabhängig* sind).

Π_m läßt sich z.B. folgendermaßen normieren. Man fixiere ein Intervall, z.b. $I = [0,1]$ und verwende die Sup-Norm für Funktionen

$$\|P\| := \|P\|_{L_\infty(I)} = \max_{x \in I} |P(x)|.$$

△

Auf einem **endlich-dimensionalen Vektorraum** V sind alle Normen äquivalent. Das heißt, zu je zwei Normen $\|\cdot\|_*$, $\|\cdot\|_{**}$ existieren beschränkte, positive Konstanten c, C, so daß

$$c\|v\|_* \leq \|v\|_{**} \leq C\|v\|_* \quad \text{für alle} \quad v \in V.$$

Beispiel:

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty, \quad \text{für alle} \quad x \in \mathbb{R}^n.$$

absolute Fehler: $\|\Delta x\|_X, \quad \|\Delta y\|_Y$,

relative Fehler: $\delta_x = \dfrac{\|\Delta x\|_X}{\|x\|_X}, \quad \delta_y = \dfrac{\|\Delta y\|_Y}{\|y\|_Y}$.

Mit der *relativen/absoluten Kondition* eines (durch f beschriebenen) Problems bezeichnet man nun das Verhältnis

$$\frac{\delta_y}{\delta_x} \quad \text{bzw.} \quad \frac{\|\Delta y\|_Y}{\|\Delta x\|_X}$$

des relativen/absoluten Ausgabefehlers zum relativen/absoluten Eingabefehler – also die **Sensitivität** des Problems unter Störung der Eingabedaten.

Wenn man über die Kondition eines Problems spricht, wird meistens die **relative** Kondition gemeint.
Ein Problem ist umso besser **konditioniert**, je kleinere Schranken für δ_y/δ_x (mit $\delta_x \longrightarrow 0$) existieren.

Seien $g, h: \mathbb{R}^n \to \mathbb{R}^m$ und $\|\cdot\|_{\mathbb{R}^n}$, $\|\cdot\|_{\mathbb{R}^m}$ Normen auf $\mathbb{R}^n$ bzw. $\mathbb{R}^m$. Sei $x_0 \in \mathbb{R}^n$.

Wenn es Konstanten $C > 0$, $\delta > 0$ gibt, so daß für alle x mit $\|x - x_0\|_{\mathbb{R}^n} < \delta$

$$\|g(x)\|_{\mathbb{R}^m} \leq C\|h(x)\|_{\mathbb{R}^m}$$

gilt, sagt man:

„g *ist von der Ordnung groß* $\mathcal{O}$ *von* h *für* x *gegen* x_0“.

Dafür wird oft die Notation

$$g(x) = \mathcal{O}(h(x)) \qquad (x \to x_0)$$

verwendet.

Beispiel 2.11. Für $n = m = 1$ gilt

$$\begin{aligned} \sin x &= \mathcal{O}(x) \quad (x \to a) \quad \text{für alle} \quad a \in \mathbb{R} \ , \\ x^2 + 3x &= \mathcal{O}(x) \quad (x \to 0) \ , \\ x^2 - x - 6 &= \mathcal{O}(x-3) \quad (x \to 3) \ . \end{aligned}$$

Für $n = 2, m = 1$, $g(x_1, x_2) = x_1^2(1 - x_2) + (x_2^3 + x_1)(1 - x_1^2)$ gilt

$$\begin{aligned} g(x_1, x_2) &= \mathcal{O}(x_1 + x_2^3) \quad ((x_1, x_2) \to (0,0)) \ , \\ g(x_1, x_2) &= \mathcal{O}(|1 - x_1| + |1 - x_2|) \quad ((x_1, x_2) \to (1,1)) \ . \end{aligned}$$

△

Für hinreichend oft differenzierbares $f : \mathbb{R} \to \mathbb{R}$ gilt

$$f(\tilde{x}) = f(x) + f'(x)(\tilde{x} - x) + \frac{f^{(2)}(x)}{2}(\tilde{x} - x)^2 + \ldots$$
$$+ \frac{f^{(k-1)}(x)}{(k-1)!}(\tilde{x} - x)^{k-1} + \frac{f^{(k)}(\xi)}{k!}(\tilde{x} - x)^k,$$

wobei ξ eine Zahl zwischen $\tilde{x}$ und x ist. Das Polynom

$$p_{k-1}(\tilde{x}) := f(x) + f'(x)(\tilde{x} - x) + \frac{f^{(2)}(x)}{2}(\tilde{x} - x)^2 + \ldots + \frac{f^{(k-1)}(x)}{(k-1)!}(\tilde{x} - x)^{k-1}$$

wird das **Taylorpolynom** vom Grad $k - 1$ in x genannt.
Für $k = 1$ erhält man als Spezialfall den *Mittelwertsatz*

$$\frac{f(\tilde{x}) - f(x)}{\tilde{x} - x} = f'(\xi),$$

wobei ξ eine Zahl zwischen $\tilde{x}$ und x ist. Oft wird die Darstellung

$$f(\tilde{x}) = p_{k-1}(\tilde{x}) + \mathcal{O}(|\tilde{x} - x|^k) \quad (\tilde{x} \to x)$$

verwendet.

Für hinreichend oft differenzierbares $f : \mathbb{R}^n \to \mathbb{R}$ gilt

$$f(\tilde{x}) = f(x) + \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(x)(\tilde{x}_j - x_j)$$
$$+ \sum_{i,j=1}^{n} \frac{1}{2} \frac{\partial^2 f(x)}{\partial x_i \partial x_j}(\tilde{x}_i - x_i)(\tilde{x}_j - x_j) + \mathcal{O}(\|\tilde{x} - x\|_2^3), \quad \tilde{x} \to x.$$

Setzt man kurz

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n} \right)^T \quad \textit{Gradient},$$
$$f''(x) = \left(\frac{\partial^2 f(x)}{\partial x_i \partial x_j} \right)_{i,j=1}^{n} \quad \textit{Hesse-Matrix},$$

läßt sich dies kompakt auch folgendermaßen schreiben:

$$f(\tilde{x}) = f(x) + \big(\nabla f(x)\big)^T (\tilde{x} - x) + \frac{1}{2}(\tilde{x} - x)^T f''(x)(\tilde{x} - x) + \mathcal{O}(\|\tilde{x} - x\|_2^3).$$

$$f(\tilde{x}) = f(x) + \big(\nabla f(x)\big)^T (\tilde{x} - x) + \mathcal{O}(\|\tilde{x} - x\|^2), \quad (\tilde{x} \to x).$$

Man schreibt

$$f(\tilde{x}) \doteq f(x) + \Big(\nabla f(x)\Big)^T (\tilde{x} - x),$$

um anzudeuten, daß beide Seiten nur in den Anteilen nullter und erster Ordnung übereinstimmen. Hieraus folgt dann

$$\frac{f(\tilde{x}) - f(x)}{f(x)} \doteq \sum_{j=1}^{n} \frac{\partial f(x)}{\partial x_j} \cdot \frac{x_j}{f(x)} \cdot \frac{\tilde{x}_j - x_j}{x_j}.$$

Definiert man also die *Verstärkungsfaktoren*

$$\phi_j(x) = \frac{\partial f(x)}{\partial x_j} \cdot \frac{x_j}{f(x)},$$

erhält man

$$\underbrace{\frac{f(\tilde{x}) - f(x)}{f(x)}}_{\substack{\text{rel. Fehler} \\ \text{der Ausgabe}}} \doteq \underbrace{\sum_{j=1}^{n} \phi_j(x)}_{\substack{\text{Fehler-} \\ \text{verstärkung}}} \cdot \underbrace{\frac{\tilde{x}_j - x_j}{x_j}}_{\substack{\text{rel. Fehler} \\ \text{der Eingabe}}}$$

$$\left|\frac{f(\tilde{x})-f(x)}{f(x)}\right| \stackrel{.}{\leq} \kappa_{\mathsf{rel}}(x) \sum_{j=1}^{n} \left|\frac{\tilde{x}_j - x_j}{x_j}\right| \ ,$$

$$\text{mit} \quad \kappa_{\mathsf{rel}}(x) = \kappa_{\mathsf{rel}}^{\infty}(x) = \max_j |\phi_j(x)| \ .$$

Ein besonders einfacher Fall ergibt sich noch, wenn $n = 1$ ist, die Funktion f also nur von einer Variablen abhängt, $X = Y = \mathbb{R}$.
Man erhält dann die Form

$$\left|\frac{f(\tilde{x})-f(x)}{f(x)}\right| \doteq \kappa_{\mathsf{rel}}(x) \left|\frac{\tilde{x}-x}{x}\right| \ ,$$

$$\text{mit} \quad \kappa_{\mathsf{rel}}(x) := \left|f'(x)\frac{x}{f(x)}\right| \ .$$

Beispiel 2.12

Gegeben sei die Funktion

$$f : \mathbb{R} \to \mathbb{R}, \quad f(x) = e^{3x^2}.$$

Für die relative Konditionszahl erhält man

$$\kappa_{\text{rel}}(x) = \left| f'(x) \frac{x}{f(x)} \right| = 6x^2.$$

Daraus folgt, daß diese Funktion für $|x|$ klein (groß) gut (schlecht) konditioniert ist. Zum Beispiel:

$x = 0.1, \ \tilde{x} = 0.10001$
$\left|\frac{x-\tilde{x}}{x}\right| = 10^{-4}$

$\xrightarrow{f}$

$\left|\frac{f(x)-f(\tilde{x})}{f(x)}\right| = 6.03 * 10^{-6}$

$x = 4, \quad \tilde{x} = 4.0004$
$\left|\frac{x-\tilde{x}}{x}\right| = 10^{-4}$

$\xrightarrow{f}$

$\left|\frac{f(x)-f(\tilde{x})}{f(x)}\right| = 9.65 * 10^{-3}$

Beispiel 2.13. (Multiplikation)

$$x = (x_1, x_2)^T, \quad f(x) = x_1 x_2, \quad \frac{\partial f(x)}{\partial x_1} = x_2, \quad \frac{\partial f(x)}{\partial x_2} = x_1,$$

$$\phi_j(x) = \frac{x_1 x_2}{f(x)} = 1, \qquad j = 1, 2.$$

Daraus folgt, daß $\kappa_{\text{rel}}(x) = 1$ (von x unabhängig!).
Die Multiplikation ist also für alle Eingangsdaten **gut konditioniert**.
Für die Multiplikation ergibt sich dann

$$\begin{aligned} \left|\frac{\tilde{x}_1\tilde{x}_2 - x_1x_2}{x_1x_2}\right| &= \left|\frac{f(\tilde{x}_1, \tilde{x}_2) - f(x_1, x_2)}{f(x_1, x_2)}\right| \leq \kappa_{\text{rel}} \left(\left|\frac{\tilde{x}_1 - x_1}{x_1}\right| + \left|\frac{\tilde{x}_2 - x_2}{x_2}\right|\right) \\ &\leq 1 \cdot (|\delta_{x_1}| + |\delta_{x_2}|) \leq 2\epsilon. \end{aligned}$$

△

Für die Division gilt ein ähnliches Resultat:
Eine Verstärkung des relativen Fehlers um einen beschränkten Faktor ($\kappa_{\text{rel}} \leq 1$).

Beispiel 2.14. (Addition)

$$x = (x_1, x_2)^T, \quad f(x) = x_1 + x_2, \quad \frac{\partial f(x)}{\partial x_1} = 1, \quad \frac{\partial f(x)}{\partial x_2} = 1,$$

$$\phi_j(x) = \frac{\partial f(x)}{\partial x_j} \cdot \frac{x_j}{f(x)} = \frac{x_j}{x_1 + x_2}, \quad j = 1, 2.$$

Daraus folgt

$$\kappa_{\text{rel}}(x) = \max \left\{ \left| \frac{x_1}{x_1 + x_2} \right|, \left| \frac{x_2}{x_1 + x_2} \right| \right\}.$$

Mit $f(x_1, x_2) = x_1 + x_2$ gilt dann

$$\begin{aligned} \left| \frac{(\tilde{x}_1 + \tilde{x}_2) - (x_1 + x_2)}{x_1 + x_2} \right| &= \left| \frac{f(\tilde{x}_1, \tilde{x}_2) - f(x_1, x_2)}{f(x_1, x_2)} \right| \\ &\leq \kappa_{\text{rel}} \left(\left| \frac{\tilde{x}_1 - x_1}{x_1} \right| + \left| \frac{\tilde{x}_2 - x_2}{x_2} \right| \right) \leq \kappa_{\text{rel}} 2\epsilon. \end{aligned}$$

Bei zwei Zahlen mit *gleichem* Vorzeichen: $\kappa_{\text{rel}} \leq 1$.
Aber: $\kappa_{\text{rel}}(x) \gg 1$ wenn $x_1 \approx -x_2$.

Beispiel 2.15. (Nullstelle)

Bestimmung der kleineren Nullstelle y^* von $y^2 - 2x_1 y + x_2 = 0$:

$$x = (x_1, x_2)^T, \quad f(x) = x_1 - \sqrt{x_1^2 - x_2} = y^*.$$

$$\frac{\partial f(x)}{\partial x_1} = \frac{\sqrt{x_1^2 - x_2} - x_1}{\sqrt{x_1^2 - x_2}} = \frac{-y^*}{\sqrt{x_1^2 - x_2}}, \quad \frac{\partial f(x)}{\partial x_2} = \frac{1}{2\sqrt{x_1^2 - x_2}}$$

$$\phi_1(x) = \frac{-y^*}{\sqrt{x_1^2 - x_2}} \frac{x_1}{y^*} = \frac{-x_1}{\sqrt{x_1^2 - x_2}}$$

$$\phi_2(x) = \frac{x_2}{2y^*\sqrt{x_1^2 - x_2}} = \frac{x_1 + \sqrt{x_1^2 - x_2}}{2\sqrt{x_1^2 - x_2}} = \frac{1}{2} - \frac{1}{2}\phi_1(x).$$

Die Kondition hängt stark von der Stelle (x_1, x_2) ab.

Wenn $x_2 < 0$: $|\phi_1(x)| \leq 1$ und $\kappa_{\text{rel}}(x) \leq 1$.
Wenn $x_2 \approx x_1^2$: $|\phi_1(x)| \gg 1$ und $\kappa_{\text{rel}} \gg 1$. $\triangle$

Eine Abbildung $\mathcal{L} : X \to Y$ heißt *linear*, falls für $x, z \in X$ und $\alpha, \beta \in \mathbb{R}$

$$\mathcal{L}(\alpha x + \beta z) = \alpha \mathcal{L}(x) + \beta \mathcal{L}(z).$$

Operatornorm von $\mathcal{L}$:

$$\|\mathcal{L}\|_{X \to Y} := \sup_{\|x\|_X = 1} \|\mathcal{L}(x)\|_Y.$$

Man sagt, $\mathcal{L}$ ist *beschränkt*, wenn $\|\mathcal{L}\|_{X \to Y}$ endlich ist.

Diese Definition ist äquivalent zu

$$\|\mathcal{L}\|_{X \to Y} := \sup_{x \neq 0} \frac{\|\mathcal{L}(x)\|_Y}{\|x\|_X}.$$

Daraus wiederum folgt sofort folgende wichtige Eigenschaft

$$\boxed{\|\mathcal{L}(x)\|_Y \leq \|\mathcal{L}\|_{X \to Y} \|x\|_X, \qquad \forall \quad x \in X.}$$

Sei $X = \mathbb{R}^n$, $Y \in \mathbb{R}^m$ und $B \in \mathbb{R}^{m \times n}$ eine $(m \times n)$-Matrix.

Stattet man X und Y mit der p-Norm für $1 \leq p \leq \infty$ aus, bezeichnet man die entsprechende Operatornorm kurz als $\|B\|_p := \|B\|_{X \to Y}$.

Es gilt:

$$\|B\|_\infty = \max_{i=1,\ldots,m} \sum_{k=1}^{n} |b_{i,k}|,$$

sowie

$$\|B\|_1 = \max_{i=1,\ldots,n} \sum_{k=1}^{m} |b_{k,i}|,$$

Ferner gilt für $A \in \mathbb{R}^{n \times n}$

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)},$$

wobei A^T die Transponierte von A ist, (d.h. $\left(A^T\right)_{i,j} = a_{j,i}$) und $\lambda_{\max}$ der größte Eigenwert ist. $\triangle$

Beispiel 2.21.

Für $A = \begin{pmatrix} 2 & -3 \\ 1 & 1 \end{pmatrix}$ ergibt sich:

$$\|A\|_\infty = 5, \quad \|A\|_1 = 4.$$

Die Eigenwerte der Matrix $A^T A = \begin{pmatrix} 5 & -5 \\ -5 & 10 \end{pmatrix}$ kann man über

$$\det \begin{pmatrix} 5-\lambda & -5 \\ -5 & 10-\lambda \end{pmatrix} = 0 \iff (5-\lambda)(10-\lambda) - 25 = 0$$

bestimmen. Also

$$\lambda_1 = \frac{1}{2}(15 - 5\sqrt{5}), \quad \lambda_2 = \frac{1}{2}(15 + 5\sqrt{5}),$$

und damit $\|A\|_2 = \sqrt{\frac{1}{2}(15 + 5\sqrt{5})}$. △

Satz 2.23.

Unter obigen Voraussetzungen gilt

$$\boxed{\frac{\|\mathcal{L}(\tilde{x}) - \mathcal{L}(x)\|_Y}{\|\mathcal{L}(x)\|_Y} \leq \kappa(\mathcal{L}) \frac{\|\tilde{x} - x\|_X}{\|x\|_X},}$$

wobei

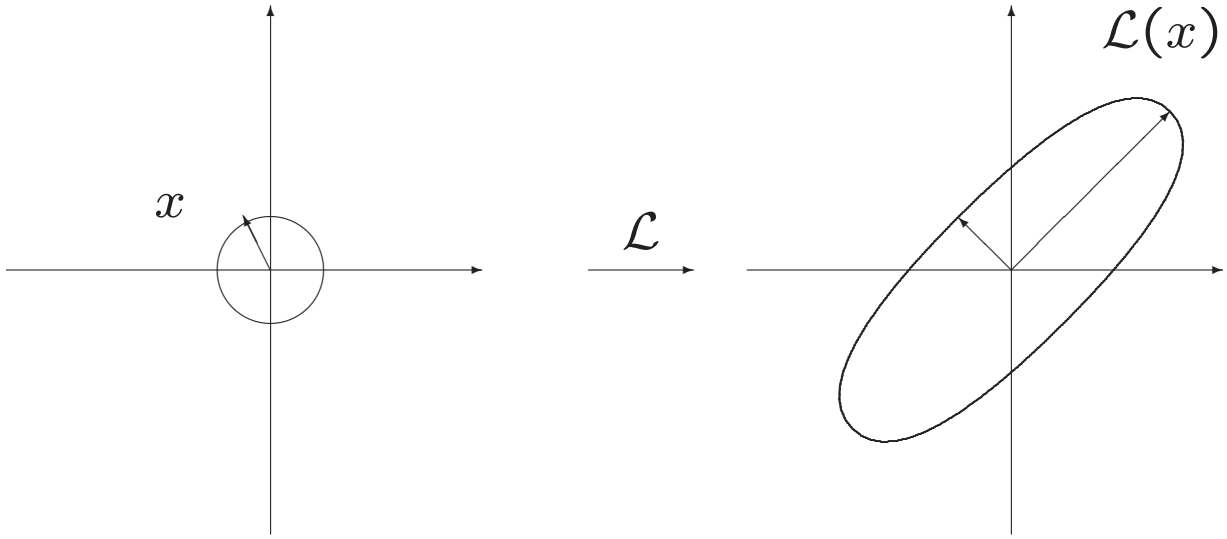
$$\boxed{\kappa(\mathcal{L}) = \frac{\sup_{\|x\|_X=1} \|\mathcal{L}(x)\|_Y}{\inf_{\|x\|_X=1} \|\mathcal{L}(x)\|_Y} = \frac{\|\mathcal{L}\|_{X\to Y}}{\inf_{\|x\|_X=1} \|\mathcal{L}(x)\|_Y}}$$

$\kappa(\mathcal{L})$ wird (relative) *Konditionszahl* von $\mathcal{L}$ (bezüglich der Normen $\|\cdot\|_X, \|\cdot\|_Y$). Offensichtlich gilt stets $\kappa(\mathcal{L}) \geq 1$.
Wenn $\mathcal{L} : X \to Y$ *bijektiv* ist, dann erhält man

$$\kappa(\mathcal{L}) = \|\mathcal{L}\|_{X\to Y} \|\mathcal{L}^{-1}\|_{Y\to X}.$$

Bemerkung 2.26

(i) Die Zahl $\kappa(\mathcal{L})$ ist eine *obere Schranke* für die relative Kondition des Problems der Auswertung der Funktion $\mathcal{L}(x)$. Sie ist *unabhängig* vom speziellen Auswertungspunkt x und wird oft *(relative) Konditionszahl* von $\mathcal{L}$ genannt.
(ii) Die Zahl $\kappa(\mathcal{L})$ hängt von den Normen $\|\cdot\|_V, \|\cdot\|_W$ ab.
(iv) Falls $\mathcal{L}$ bijektiv ist, haben $\mathcal{L}$ und $\mathcal{L}^{-1}$ *dieselbe Konditionszahl* !
(v) Geometrische Interpretation:



Beispiel 2.28

Die Bestimmung des Schnittpunktes der Geraden

$$\begin{aligned} 3u_1 + 1.001u_2 &= 1.999 \\ 6u_1 + 1.997u_2 &= 4.003, \end{aligned}$$

(fast parallel!) ergibt das Problem $u = A^{-1}b$ mit

$$A = \begin{pmatrix} 3 & 1.001 \\ 6 & 1.997 \end{pmatrix}, \quad b = \begin{pmatrix} 1.999 \\ 4.003 \end{pmatrix}.$$

Die Lösung ist $u = (1, -1)^T$.
Wir berechnen den Effekt einer Störung in b:

$$\tilde{b} = \begin{pmatrix} 2.002 \\ 4 \end{pmatrix}, \quad \tilde{u} := A^{-1}\tilde{b}.$$

Man rechnet einfach nach, daß

$$A^{-1} = \frac{-1}{0.015} \begin{pmatrix} 1.997 & -1.001 \\ -6 & 3 \end{pmatrix}, \quad \tilde{u} = \begin{pmatrix} 0.4004 \\ 0.8 \end{pmatrix}.$$

Als Norm wird die Maximumnorm genommen:

$$\|x\| = \|x\|_\infty := \max_i |x_i|.$$

Es gilt

$$\frac{\|\tilde{b} - b\|_\infty}{\|x\|_\infty} = \frac{3 * 10^{-3}}{4.003} \approx 7.5 * 10^{-4} \qquad \text{(Störung der Daten)}$$

und

$$\frac{\|\tilde{u} - u\|_\infty}{\|u\|_\infty} = \frac{1.8}{1} = 1.8 \qquad \text{(Änderung des Resultats).}$$

Schlechte Kondition wird quantifiziert durch

$$\|A\|_\infty \|A^{-1}\|_\infty = 4798.2 \quad . \qquad \triangle$$

Beispiel 2.29 (Integralberechnung über Rekursion)

Sei $\tilde{I}_0 \approx I_0$ ein gestörter Startwert. $f(\tilde{I}_0) = \tilde{I}_{30}$ folgt aus

$$\tilde{I}_n = \frac{1}{n} - 5\tilde{I}_{n-1}, \quad n = 1, 2, \ldots, 30.$$

Für das Resultat ohne Störung, $I_{30} = f(I_0)$, gilt

$$I_n = \frac{1}{n} - 5I_{n-1}, \quad n = 1, 2, \ldots, 30.$$

Daraus folgt, daß

$$\tilde{I}_{30} - I_{30} = -5(\tilde{I}_{29} - I_{29}) = 5^2(\tilde{I}_{28} - I_{28}) = \ldots = 5^{30}(\tilde{I}_0 - I_0)$$

und damit

$$\frac{|f(\tilde{I}_0) - f(I_0)|}{|f(I_0)|} = \frac{|\tilde{I}_{30} - I_{30}|}{|I_{30}|} = \frac{5^{30}|I_0|}{|I_{30}|} \cdot \frac{|\tilde{I}_0 - I_0|}{|I_0|}.$$

Für den gesuchten Wert I_{30} gilt $|I_{30}| = \ldots \leq \frac{1}{155}$. Also:

$$\kappa_{\text{rel}} := \frac{5^{30}\,|I_0|}{|I_{30}|} \geq 155 * 5^{30} \ln(\tfrac{6}{5}) = 2.6 * 10^{22}.$$

Tabelle 2.1. Integralberechnung über Rekursion

n	I_n	n	I_n
0	1.8232156e−01	16	9.8903245e−03
1	8.8392216e−02	17	9.3719069e−03
2	5.8038920e−02	18	8.6960213e−03
3	4.3138734e−02	19	9.1514726e−03
4	3.4306330e−02	20	4.2426370e−03
5	2.8468352e−02	21	2.6405862e−02
6	2.4324906e−02	22	−8.6574767e−02
7	2.1232615e−02	23	4.7635209e−01
8	1.8836924e−02	24	−2.3400938e+00
9	1.6926490e−02	25	1.1740469e+01
10	1.5367550e−02	26	−5.8663883e+01
11	1.4071341e−02	27	2.9335645e+02
12	1.2976630e−02	28	−1.4667466e+03
13	1.2039925e−02	29	7.3337673e+03
14	1.1228946e−02	30	−3.6668803e+04
15	1.0521935e−02		

Zahlendarstellungen Man kann zeigen, daß für jedes feste $b \in \mathbb{N}$, $b > 1$, jede beliebige reelle Zahl $x \neq 0$ sich in der Form

$$x = \pm\Big(\sum_{j=1}^{\infty} d_j b^{-j}\Big) * b^e$$

darstellen läßt, wobei der ganzzahlige Exponent e so gewählt werden kann, daß $d_1 \neq 0$ gilt.

Normalisierte Gleitpunktdarstellung (floating point representation):

$$x = f * b^e, \quad \text{wobei}$$

- $b \in \mathbb{N} \setminus \{1\}$ die *Basis* (oder Grundzahl) ist,
- der *Exponent* e eine ganze Zahl ist:

$$r \leq e \leq R \ ,$$

- die *Mantisse* f eine feste Anzahl m (die *Mantissenlänge*) von Stellen hat:

$$f = \pm\ 0.d_1 \ldots d_m \ , \quad d_j \in \{0, 1, \ldots, b-1\} \ \text{ für alle } \ j \ .$$

Um die Eindeutigkeit der Darstellung zu erreichen, wird für $x \neq 0$ die Forderung $d_1 \neq 0$ gestellt (Normalisierung).

Mit dieser Darstellung erhält man somit: $x = \pm\left(\sum_{j=1}^{m} d_j b^{-j}\right) * b^e$.

Beispiel 2.31

Wir betrachten als Beispiel die Zahl

$$\begin{aligned} 123.75 &= 1*2^6 + 1*2^5 + 1*2^4 + 1*2^3 + 0*2^2 + 1*2^1 + 1*2^0 + 1*2^{-1} + 1*2^{-2} \\ &= 2^7(1*2^{-1} + 1*2^{-2} + 1*2^{-3} + 1*2^{-4} + 0*2^{-5} + \\ &\quad 1*2^{-6} + 1*2^{-7} + 1*2^{-8} + 1*2^{-9}). \end{aligned}$$

Diese Zahl wird in einem sechsstelligen dezimalen Gleitpunkt-Zahlensyste ($b = 10$, $m = 6$) als

$$0.123750 * 10^3$$

dargestellt. In einem 12-stelligen binären Gleitpunkt-Zahlensystem ($b = 2$, $m = 12$) wird sie als

$$0.111101111000 * 2^{111}$$

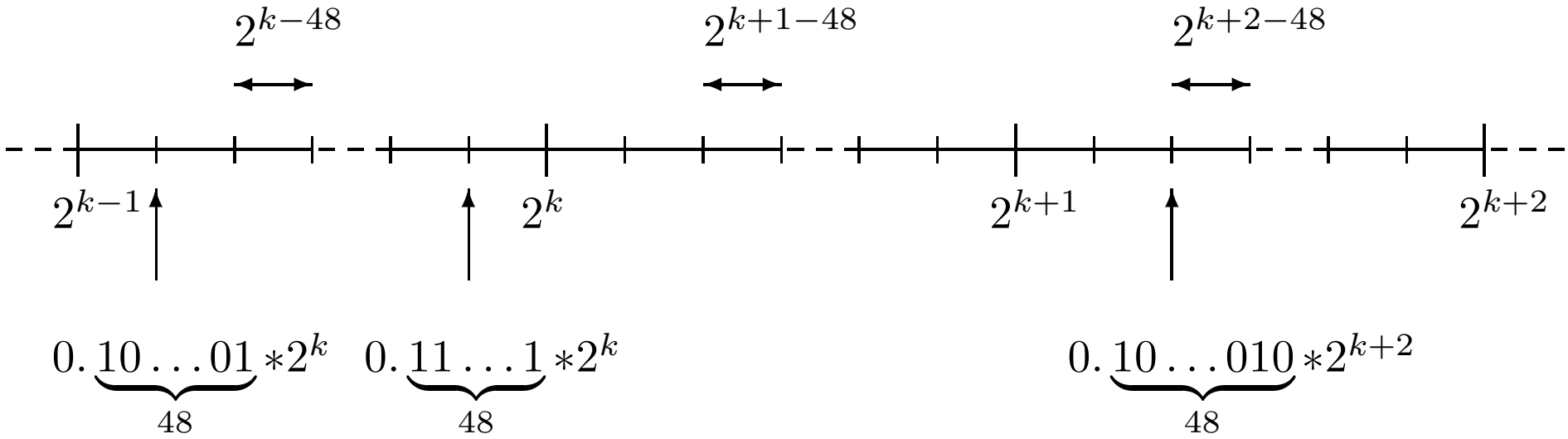
dargestellt. △

Beispiel 2.32

Menge $\mathbb{M}(2, 48, -1024, 1024)$ enthält $2^{47} * 2049$ positive Zahlen.

Anzahl der Zahlen in dieser Menge: $2 * 2^{47} * 2049 + 1 \approx 5.8 * 10^{17}$.

$$x_{\text{MIN}} = 2^{-1025} \approx 2.8 * 10^{-309},$$
$$x_{\text{MAX}} = (1 - 2^{-48}) * 2^{1024} \approx 1.8 * 10^{308}.$$



△

$x \in \pm[x_{\mathsf{MIN}}, x_{\mathsf{MAX}}]$ habe die Darstellung

$$x = \pm\Big(\sum_{j=1}^{\infty} d_j b^{-j}\Big) * b^e.$$

Die Reduktionsabbildung wird definiert durch

$$\mathsf{fl}(x) := \pm \begin{cases} \Big(\sum_{j=1}^{m} d_j b^{-j}\Big) * b^e & \text{falls } d_{m+1} < \frac{b}{2}, \\ \Big(\sum_{j=1}^{m} d_j b^{-j} + b^{-m}\Big) * b^e & \text{falls } d_{m+1} \geq \frac{b}{2}, \end{cases}$$

d.h., die letzte Stelle der Mantisse wird um eins erhöht bzw. beibehalten, falls die Ziffer in der nächsten Stelle $\geq \frac{b}{2}$ bzw. $< \frac{b}{2}$ ist.

Beispiel 2.33.

In einem Gleitpunkt-Zahlensystem mit Basis $b = 10$ und Mantissenlänge $m = 6$ erhält man folgende gerundete Resultate:

x	$\mathrm{fl}(x)$	$\left\lvert\frac{\mathrm{fl}(x)-x}{x}\right\rvert$
$\frac{1}{3} = 0.33333333\ldots$	$0.333333 * 10^{0}$	$1.0 * 10^{-6}$
$\sqrt{2} = 1.41421356\ldots$	$0.141421 * 10^{1}$	$2.5 * 10^{-6}$
$e^{-10} = 0.000045399927\ldots$	$0.453999 * 10^{-4}$	$6.6 * 10^{-7}$
$e^{10} = 22026.46579\ldots$	$0.220265 * 10^{5}$	$1.6 * 10^{-6}$
$\frac{1}{10} = 0.1$	$0.100000 * 10^{0}$	0.0

Im Fall $b = 2, m = 10$ erhält man:

x	$\mathrm{fl}(x)$	$\left\lvert\frac{\mathrm{fl}(x)-x}{x}\right\rvert$
$\frac{1}{3}$	$0.1010101011 * 2^{-1}$	$4.9 * 10^{-4}$
$\sqrt{2}$	$0.1011010100 * 2^{1}$	$1.1 * 10^{-4}$
e^{-10}	$0.1011111010 * 2^{-111}$	$3.3 * 10^{-4}$
e^{10}	$0.1010110000 * 2^{1111}$	$4.8 * 10^{-4}$
$\frac{1}{10}$	$0.1100110011 * 2^{-11}$	$2.4 * 10^{-4}$

$$\left|\frac{\mathrm{fl}(x)-x}{x}\right| \leq \frac{\frac{b^{-m}}{2}b^e}{b^{-1}b^e} = \frac{b^{1-m}}{2}.$$

Die Zahl

$$\mathrm{eps} := \frac{b^{1-m}}{2}$$

wird (relative) **Maschinengenauigkeit** genannt. Diese Zahl charakterisiert das *Auflösungsvermögen* des Rechners. Es gilt nämlich:

$$\mathrm{eps} = \inf\{\delta > 0 \mid \mathrm{fl}(1+\delta) > 1\}.$$

Diese Abschätzung besagt ferner, daß für eine Zahl ϵ mit $|\epsilon| \leq \mathrm{eps}$, nämlich $\epsilon = \frac{\mathrm{fl}(x)-x}{x}$,

$$\mathrm{fl}(x) = x(1+\epsilon)$$

gilt.

Beispiel 2.34.

Für die Zahlensysteme in Beispiel 2.33 ergibt sich:

$$b = 10, m = 6 \quad \rightarrow \quad \text{eps} = \frac{1}{2} * 10^{-5}$$
$$b = 2, m = 10 \quad \rightarrow \quad \text{eps} = \frac{1}{2} * 2^{-9} = 9.8 * 10^{-4}.$$

Die Werte für den relativen Rundungsfehler $|\epsilon|$, mit ϵ wie in (2.48), findet man in der dritten Spalte der Tabellen in Beispiel 2.33. $\triangle$

Die Verknüpfung von Maschinenzahlen durch eine *exakte* elementare arithmetische Operation liefert **nicht** notwendig eine Maschinenzahl.

Beispiel 2.35.

$b = 10, m = 3$:

$$0.346 * 10^2 + 0.785 * 10^2 = 0.1131 * 10^3 \neq 0.113 * 10^3 \qquad \triangle$$

Ähnliches passiert bei Multiplikation und Division.

Die üblichen arithmetischen Operationen müssen also durch geeignete Gleitpunktoperationen $\textcircled{\nabla}$, $\nabla \in \{+, -, \times, \div\}$, ersetzt werden (Pseudoarithmetik).

Forderung:
Für $\nabla \in \{+, -, \times, \div\}$ gelte

$$x \mathbin{\textcircled{\nabla}} y = \mathrm{fl}(x \nabla y) \quad \text{für } x, y \in \mathbb{M}(b, m, r, R).$$

Wegen (2.48) werden wir also stets annehmen, daß für $\nabla \in \{+, -, \times, \div\}$

$$x \mathbin{\textcircled{\nabla}} y = (x \nabla y)(1 + \epsilon) \quad \text{für } x, y \in \mathbb{M}(b, m, r, R)$$

und ein ϵ mit $|\epsilon| \leq \mathrm{eps}$ gilt.

Nichtsdestoweniger hat die Realisierung einer solchen Pseudoarithmetik eine Reihe unliebsamer Konsequenzen: Zum Beispiel geht die *Assoziativität* der Addition verloren, d.h., im Gegensatz zur exakten Arithmetik spielt es eine Rolle, welche Zahlen zuerst verknüpft werden.

Beispiel 2.36. Zahlensystem mit $b = 10,\ m = 3$. Maschinenzahlen

$$\begin{aligned} x &= 6590 = 0.659 * 10^4 \\ y &= \ \ 1 \ \ = 0.100 * 10^1 \\ z &= \ \ 4 \ \ = 0.400 * 10^1 \end{aligned}$$

Exakte Rechnung: $(x+y)+z = (y+z)+x = 6595$. Pseudoarithmetik:

$$x \oplus y = 0.659 * 10^4 \quad \text{und} \quad (x \oplus y) \oplus z = 0.659 * 10^4,$$

aber

$$y \oplus z = 0.500 * 10^1 \quad \text{und} \quad (y \oplus z) \oplus x = 0.660 * 10^4. \qquad \triangle$$

Entsprechend gilt auch das Distributivgesetz nicht mehr:

Beispiel 2.37 Für $b = 10, m = 3,\ x = 0.156 * 10^2$ und $y = 0.157 * 10^2$

$$\begin{aligned} (x - y) * (x - y) &= 0.01 \\ (x \ominus y) \otimes (x \ominus y) &= 0.100 * 10^{-1}, \end{aligned}$$

aber

$$(x \otimes x) \ominus (x \otimes y) \ominus (y \otimes x) \oplus (y \otimes y) = -0.100 * 10^1. \qquad \triangle$$

Bezeichne wieder δ_x, δ_y die relativen Fehler der Größen $\tilde{x}, \tilde{y}$ gegenüber den exakten Werten x, y, d.h.

$$\tilde{x} = x(1 + \delta_x), \quad \tilde{y} = y(1 + \delta_y).$$

Ferner nehmen wir an, daß $|\delta_x|, |\delta_y| \leq \epsilon < 1$.

In Beispiel 2.13 hatten wir bereits gesehen, κ_{rel} für die Multiplikation $f(x, y) = xy$ den Wert $\kappa_{\text{rel}} = 1$ hat.
Falls insbesondere $|\delta_x|, |\delta_y| \leq \epsilon \leq$ eps, bleibt bei der Multiplikation der relative Fehler im Rahmen der Maschinengenauigkeit, denn aus Beispiel 2.13 folgt

$$\left| \frac{\tilde{x}\tilde{y} - xy}{xy} \right| \leq 2 \text{ eps}.$$

Für die Division gilt ein ähnliches Resultat.

Beispiel 2.38. (Auslöschung)

Betrachte

$$x = 0.73563, \quad y = 0.73441, \quad x - y = 0.00122.$$

Bei 3-stelliger Rechnung ($b = 10, m = 3, \text{eps} = \frac{1}{2} * 10^{-2}$) ergibt sich

$$\begin{aligned} \tilde{x} = \text{fl}(x) = 0.736, \quad |\delta_x| = 0.50 * 10^{-3} \\ \tilde{y} = \text{fl}(y) = 0.734, \quad |\delta_y| = 0.56 * 10^{-3}. \end{aligned}$$

Die relative Störung im Resultat der Subtraktion ist hier

$$\left| \frac{(\tilde{x} - \tilde{y}) - (x - y)}{x - y} \right| = \left| \frac{0.002 - 0.00122}{0.00122} \right| = 0.64,$$

also sehr groß im Vergleich zu δ_x, δ_y.

Zusammenfassend:

$$\left|\frac{(x \,\overline{\nabla}\, y) - (x \nabla y)}{(x \nabla y)}\right| \leq \text{eps} \qquad x, y \in \mathbb{M}, \quad \nabla \in \{+, -, \times, \div\}$$

d.h., die relativen Rundungsfehler bei den elementaren Gleitpunktoperationen sind betragsmäßig kleiner als die Maschinengenauigkeit, wenn die Eingangsdaten x, y **Maschinenzahlen** *sind.*

Sei $f(x, y) = x \nabla y, \;\; x, y \in \mathbb{R}, \;\; \nabla \in \{+, -, \times, \div\}$ *und* κ_{rel} *die relative Konditionszahl von* f*. Es gilt*

$$\begin{aligned} \nabla \in \{\times, \div\}: &\quad \kappa_{\text{rel}} \leq 1 \quad \textit{für alle} \;\; x, y \;, \\ \nabla \in \{+, -\}: &\quad \kappa_{\text{rel}} \gg 1 \quad \textit{wenn} \;\; |x \nabla y| \ll \max\{|x|, |y|\}. \end{aligned}$$

Also: möglich sehr große Fehlerverstärkung bei $+, -$ (**Auslöschung**).

Ein Algorithmus heißt *gutartig* oder *stabil*, wenn die durch ihn im Laufe der Rechnung erzeugten Fehler in der Größenordnung des durch die Kondition des Problems bedingten unvermeidbaren Fehlers bleiben.

Beispiel 2.39

Bestimmung von $u^* = f(a_1, a_2) = a_1 - \sqrt{a_1^2 - a_2}$. Algorithmus I:

$$\begin{aligned} y_1 &= a_1 a_1 \\ y_2 &= y_1 - a_2 \\ y_3 &= \sqrt{y_2} \\ u^* &= y_4 = a_1 - y_3. \end{aligned}$$

Für $a_1 = 6.000227$, $a_2 = 0.01$ in einem Gleitpunkt-Zahlensystem mit $b = 10, m = 5$ bekommt man das Ergebnis

$$\tilde{u}^* = 0.90000 * 10^{-3}.$$

Exakte Lösung:

$$u^* = 0.83336 * 10^{-3}.$$

Da das Problem für diese Eingangsdaten a_1, a_2, *gut konditioniert* ist, ist der durch den Algorithmus erzeugte Fehler sehr viel größer als der unvermeidbare Fehler.
Algorithmus I ist also **nicht stabil**.

Alternative:

$$u^* = \frac{a_2}{a_1 + \sqrt{a_1^2 - a_2}}$$

Algorithmus II:

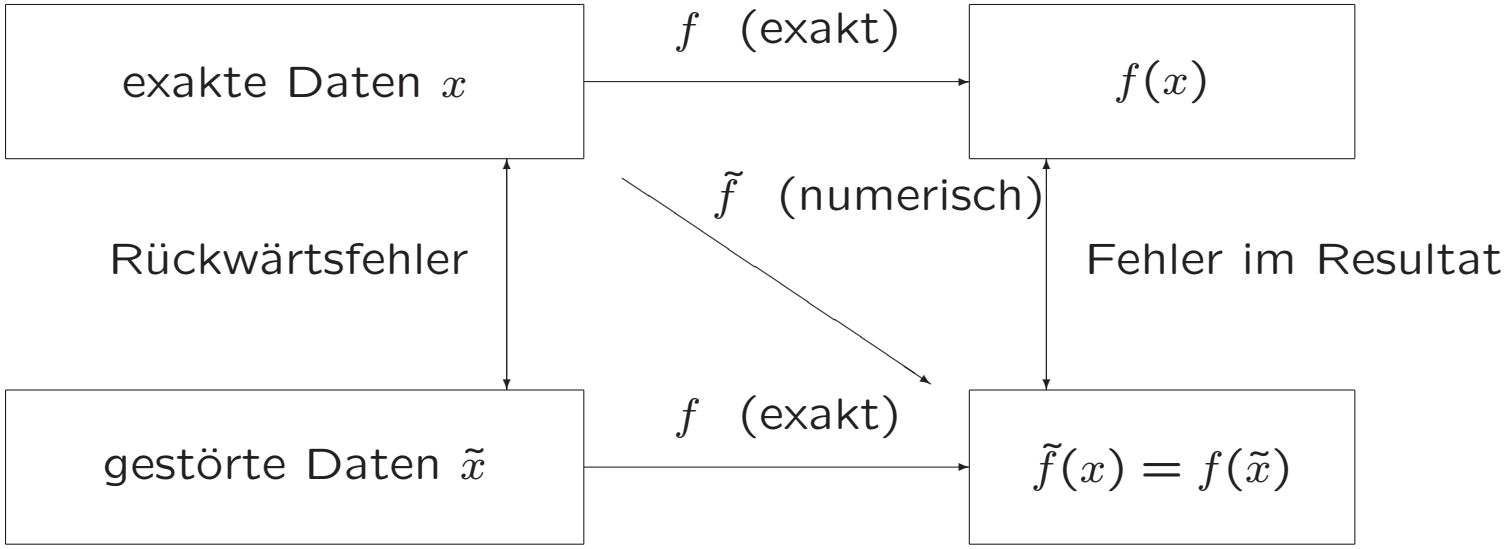
$$\begin{aligned} y_1 &= a_1 * a_1 \\ y_2 &= y_1 - a_2 \\ y_3 &= \sqrt{y_2} \\ y_4 &= a_1 + y_3 \\ u^* &= y_5 = \frac{a_2}{y_4}. \end{aligned}$$

Hiermit ergibt sich mit $b = 10$, $m = 5$

$$\tilde{u}^* = 0.83333 * 10^{-3}.$$

Hier tritt keine Auslöschung auf. Der Gesamtfehler bleibt im Rahmen der Maschinengenauigkeit.
Algorithmus II ist somit **stabil**. △

- Interpretiere sämtliche im Laufe der Rechnung auftretenden Fehler als Ergebnis *exakter* Rechnung zu geeignet gestörten Daten.
- Abschätzungen für diese Störung der Daten, verbunden mit Abschätzungen für die Kondition des Problems, ergeben dann Abschätzungen für den Gesamtfehler.



Beispiel 2.40.

x_1, x_2, x_3 seien Maschinenzahlen. Maschinengenauigkeit $=$ eps.
Aufgabe: Berechne die Summe $S = (x_1 + x_2) + x_3$.
Man erhält

$$\tilde{S} = ((x_1 + x_2)(1 + \epsilon_2) + x_3)(1 + \epsilon_3),$$

mit $|\epsilon_i| \leq \text{eps}, \quad i = 2, 3$. Daraus folgt

$$\begin{aligned}
\tilde{S} &= x_1(1+\epsilon_2)(1+\epsilon_3) + x_2(1+\epsilon_2)(1+\epsilon_3) + x_3(1+\epsilon_3) \\
&\doteq x_1(1+\epsilon_2+\epsilon_3) + x_2(1+\epsilon_2+\epsilon_3) + x_3(1+\epsilon_3) \\
&= x_1(1+\delta_1) + x_2(1+\delta_2) + x_3(1+\delta_3) \\
&=: \hat{x}_1 + \hat{x}_2 + \hat{x}_3,
\end{aligned}$$

wobei

$$|\delta_1| = |\delta_2| = |\epsilon_2 + \epsilon_3| \leq 2 \text{ eps}, \quad |\delta_3| = |\epsilon_3| \leq \text{eps}\,.$$

Also: fehlerbehaftetes Resultat $\tilde{S}$ als *exaktes* Ergebnis zu *gestörten* Eingabedaten $\hat{x}_i = x_i(1 + \delta_i)$.

Sei $f(x) = f(x_1, x_2, x_3) = x_1 + x_2 + x_3$ mit rel. Konditionszahl κ_{rel}, dann gilt für den *unvermeidbaren* Fehler

$$F_{\text{Daten}}(x) = \left| \frac{f(\tilde{x}) - f(x)}{f(x)} \right| \leq \kappa_{\text{rel}}(x) \sum_{j=1}^{3} \left| \frac{\tilde{x}_j - x_j}{x_j} \right| \leq \kappa_{\text{rel}}(x)\, 3 \text{ eps},$$

wobei angenommen wird, daß die Daten mit höchstens Maschinengenauigkeit gestört werden ($\tilde{x}_i = x_i(1+\epsilon)$, $|\epsilon| \leq \text{eps}$).

Der durch *Rechnung* bedingte Fehler ist höchstens

$$\begin{aligned} F_{\text{Rechnung}}(x) &= \left| \frac{f(\widehat{x}) - f(x)}{f(x)} \right| \leq \kappa_{\text{rel}}(x) \sum_{j=1}^{3} \left| \frac{\widehat{x}_j - x_j}{x_j} \right| \\ &\leq \kappa_{\text{rel}}(x) \sum_{j=1}^{3} \left| \delta_j \right| \leq \kappa_{\text{rel}}(x)\, 5 \text{ eps}\,. \end{aligned}$$

Man sieht: $F_{\text{Rechnung}}(x)$ ist höchstens in der Größenordnung $F_{\text{Daten}}(x)$.

Deshalb ist die Berechnung von S ein *stabiler* Algorithmus. $\triangle$

Zusammenfassend:

- Kenntnisse über die Kondition eines Problems sind oft für die Interpretation oder Bewertung der Ergebnisse von entscheidender Bedeutung.

- $\times$ und $\div$ sind Operationen die für alle Eingangsdaten gut konditioniert sind. $\pm$ kann schlecht konditioniert sein. Dadurch können bei einer Subtraktion Rundungsfehler enorm verstärkt werden: *Auslöschung*.

- In einem Algorithmus sollen (wegen Stabilität) Auslöschungseffekte vermieden werden.

- Bei einem stabilen Lösungsverfahren bleiben die im Laufe der Rechnung erzeugten Rundungsfehler in der Größenordnung der durch die Kondition des Problems bedingten unvermeidbaren Fehler. $\triangle$