CoMa1 Formelsammlung

Zweierkomplement

Negative Binärzahlen im Zweierkomplement werden dargestellt als

$$n = -\left(1 + \sum_{i=0}^{N-2} (1 - d_i) \cdot 2^i\right)$$

Sprich n wird dargestellt als $\neg n + 1$.

Gleitkommazahl

$$x = (-1)^s \cdot a \cdot q^e$$

Für die *Mantisse* a gilt im *normalisierten Fall*:

$$\sum_{i=1}^{l} a_i \cdot q^{-i} \text{ sprich } a = 0,1\ldots$$

Als Bitmuster werden Gleitkommazahlen im PC abgespeichert als

$$s|e_1e_2e_3e_4|a_1a_2a_3a_4a_5a_6a_7a_8$$

a_1 wird oft nicht abgespeichert (man spricht von *Hidden-Bit Darstellung*), da a_1 normalisiert immer 1 ist. Es gibt zwei Sonderfälle: $e = 0$ ist reserviert für die Darstellung der 0, $e_i = 1\ \forall i$ kodiert NaN („Not a number“, ein Ausdruck z.B. für das Ergebnis von $\frac{n}{0}$) und ∞. Gleitkommazahlen stehen für intervalle in $\Re$, sind also Mengen, keine Zahlen. Insbesondere sind sie auch kein Körper!

Bei gegebenen q, a und e gilt (Mantissenlänge l, Exponentenlänge M):

$$\begin{aligned} max(x) &:= (l - q^1) \cdot q^{e_{max}} \\ min(x) &:= q^{e_{min}-1} \\ e_{max} &= q^{M-1} - 1 \\ e_{min} &= -e^{M-1} \end{aligned}$$

Oft wird der Exponent mit Bias b gespeichert, bei gespeichertem Exponenten E und echtem Exponenten e gilt dann $E = e + b$. Bei doppelt genauen Fließkommazahlen ist $b = 1023$, bei einfach genauen $b = 127$.

Runden

$$rd(x) := \sum_{i=1}^{l} a_i \cdot q^{-i} + \begin{cases} 0, & \text{falls } a_{l+1} < \frac{q}{2} \\ q^{-l}, & \text{falls } a_{l+1} \geq \frac{q}{2} \end{cases}$$

Fehler

Absoluter Fehler

$$|x - rd(x)|$$

Relativer Fehler

$$\frac{|x - rd(x)|}{|x|}$$

Den relativen Fehler kann man in Abhängigkeit von q und l abschätzen:

$$\frac{|x - rd(x)|}{|x|} \leq q^{-(l-1)} =: eps(q, l)$$

eps ist die Maschinengenauigkeit. *eps* ist $1,11 \cdot 10^{16}$ für doppelt genaue und $5,96 \cdot 10^{-8}$ für einfach genaue Fließkommazahlen.

Kondition

Es gilt:

$$\text{Fehlerverstärkung} = \frac{\text{Eingabefehler}}{\text{Ausgabefehler}}$$

$$\begin{aligned} \kappa_{abs}(\tilde{x}) &= \frac{|f(x_0) - f(\tilde{x})|}{|x_0 - \tilde{x}|} \\ &= |f'(x_0)| \\ \kappa_{rel}(\tilde{x}) &= \kappa_{abs}(\tilde{x}) \cdot \frac{|x_0|}{|f(x_0)|} \end{aligned}$$

κ_{abs} und κ_{rel} sind als die kleinsten Zahlen definiert, die

$$\begin{aligned} |f(x_0) - f(\tilde{x})| &\leq \kappa_{abs}(\tilde{x}) \cdot |x_0 - \tilde{x}| + o(|x_0 - \tilde{x}|) \\ \frac{|f(x_0) - f(\tilde{x})|}{f(x_0)} &\leq \kappa_{rel}(\tilde{x}) \cdot \frac{|x_0 - \tilde{x}|}{|x_0|} + o\left(\frac{|x_0 - \tilde{x}|}{|x_0|}\right) \end{aligned}$$

erfüllen.

Die Grundrechenarten sind konditioniert mit $\kappa_{rel} = 2$ für Multiplikation und Division und

$$\kappa_{rel} = \frac{|x| + |y|}{|x + y|}$$

für Addition (sprich 1 bei positiven Summanden, sonst beliebig schlecht).

Ist eine Funktion nicht stetig, so ist die absolute Kondition unendlich. Ist sie differenzierbar, ist sie sicher endlich.

Ist eine Funktion Lipschitz stetig mit

$$|f(x_1) - f(x_2) \leq L \cdot |x_1 - x_2| \, \forall x_1, x_2 \in \Re,$$

so ist sicher

$$\kappa_{abs} \leq L.$$

Eine Funktion hat genau dann einen Punkt mit unendlicher Kondition, wenn sie nicht Lipschitz stetig ist.

Die Kondition beschreibt den Ausgabefehler, sie ist Eigenschaft des Problems.

Geschachtelte Funktionsauswertungen erlauben die Abschätzung

$$f(x) = h \circ g(x) \Rightarrow \kappa_{rel}(f, x) \leq \kappa_{rel}(h, g(x)) \cdot \kappa_{rel}(f, x)$$

Algorithmen

Man kann jede Funktion f in elementare Funktionen zerlegen:

$$f(x) := g_n \circ g_{n-1} \cdots g_1 \circ g_0(x)$$

Dabei kann das Ergebnis jeder Elementarfunktion gestört sein, sprich $g_i(x) \tilde{=} rd \circ g_i(x) = g_i(x) \cdot (1 + \epsilon_i)$. Also sind auch die Funktionsauswertungen gestört.

Stabilität

Damit ein Algorithmus stabil ist, sollte man schlecht konditionierte Elementatfunktionen vermeiden oder zumindest zuerst ausführen. Für die Stabilität σ_{rel} gilt:

$$\frac{|f(x_0) - \tilde{f}(\epsilon, x_0)|}{|f(x_0)|} \leq \sigma_{rel} \cdot ||\epsilon|| + o(||\epsilon||)$$

Die Stabilität beschreibt den Auswertungsfehler, sie ist Eigenschaft des Algorithmus.

Auswertungsbäume / Stabilitätsabschätzung

Man zerlegt einen Algorithmus in Elementaroperationen und schreibt ihn als Baum auf. Für die Blätter (die Eingaben) gilt $\sigma_{rel} = 0$. Man bestimmt die Stabilitäten von den Blättern bis zur Wurzel per

$$\sigma_{rel}(g) \leq \kappa_{rel}(g) \cdot \sigma_{rel}(h) + 1$$

$$\sigma_{rel} \leq \sum_{j=1}^{n} \prod_{i=j+1}^{n} \kappa_i = 1 + \kappa_n(1 + \kappa_{n-1}(1 + \ldots \kappa_3(1 + \kappa_2)\ldots))$$

Gesamtfehler

$$\begin{aligned} \text{Gesamtfehler} &= \kappa \cdot \text{ Eingabefehler } + \sigma \cdot \text{ Ausgabefehler } + o \\ \frac{|f(x_0) - \tilde{f}(\epsilon, \tilde{x}_0)|}{|f(x_0)|} &\leq \kappa_{rel} \cdot \frac{|x_0 - \tilde{x}_0|}{|x_0|} + \sigma_{rel}(x_0) \cdot ||\epsilon|| + o(|x_0 - \tilde{x}_0| + ||\epsilon||) \end{aligned}$$

Landausymbole

$$
\begin{aligned}
f \in \mathcal{O}(g) &\Leftrightarrow 0 \leq \limsup_{x \to a} \left| \frac{f(x)}{g(x)} \right| < \infty \\
f \in \mathrm{o}(g) &\Leftrightarrow \lim_{x \to a} \left| \frac{f(x)}{g(x)} \right| = 0 \\
f \in \Theta(g) &\Leftrightarrow 0 < \liminf_{x \to a} \left| \frac{f(x)}{g(x)} \right| \leq \limsup_{x \to a} \left| \frac{f(x)}{g(x)} \right| < \infty
\end{aligned}
$$

$$
\mathcal{O}(1) < \mathcal{O}(\log x) < \mathcal{O}(x) < \mathcal{O}(x \cdot \log x) < \mathcal{O}(x^2) < \mathcal{O}(x^n) < \mathcal{O}(x^{\log x}) < \mathcal{O}(2^x)
$$

Effizienz

Betrachtet wird die Laufzeit eines Algorithmus. Man schätzt die Anzahl der Aufrufe der aufwändigsten Operation eines Algorithmus mit $\mathcal{O}$-Notation ab. Bis einschließlich einer Laufzeit von $\mathcal{O}(x^n)$ nennt man Algorithmen effizient. Ein Problem heißt komplex, wenn es keinen effizienten Algorithmus zu seiner Lösung gibt.

Vektornorm und Matrixnorm

Eine Vektornorm $|| \cdot || : V \to \Re_+$ weist jedem Vektor eine positive reelee Zahl zu. Für sie gilt:

1. $||a \cdot x|| = |a| \cdot ||x||$
2. $||x + y|| \leq ||x|| + ||y||$

Verschiedene Normen sind

1. Die euklidsche Norm $||x|| := \sqrt[2]{\sum_{i=1}^{n} x_i^2}$
2. Die Maximumsnorm $||x|| := \max |x_i|$
3. Die l^p Norm entspricht der euklidschen Norm für p als Exponent

Als Matrixnorm bezeichnet man eine Norm

$$
||A||_M = \sup_{x \in \Re^n \setminus 0} \frac{||Ax||}{||x||}
$$

Für sie gilt

1. $||Ax||_M \leq ||A||_M \cdot ||x||$
2. $||AB||_M \leq ||A||_M \cdot ||B||_M$
3. $||E||_M = 1$

Wir verwenden die Maximumsnorm, die definiert ist durch

$$
||A||_\infty = \max_i \sum_{j=1}^{n} |A_{i,j}|
$$

Lineare Gleichungssysteme

Eine Matrix heißt *regulär*, wenn ihre Zeilen linear unabhängig sind, sonst *singulär*. Man konditioniert die Lösung eines linearen Gleichungssystems mit

$$\frac{||x-\tilde{x}||}{||x||} \leq \underbrace{||A|| \cdot ||A^{-1}||}_{\kappa_{rel}\text{ der Matrix}} \cdot \frac{||A-\tilde{A}||}{||A||} + \frac{||b-\tilde{b}||}{||b||} + o(||A-\tilde{A}|| + ||b-\tilde{b}||)$$

Schlecht konditionierte Matritzen sind fast singulär. Beispiele für fast singuläre Matritzen sind der schleifende Schnitt

$$\begin{pmatrix} -1 & 1 \\ -1 & 1+\epsilon \end{pmatrix}$$

(Der schleifende Schnitt sagt aus: Je orthogonaler eine Gerade eine gestörte Gerade schneidet, desto kleiner ist der Fehler) und die Hilbertmatrix, die durch

$$h_{ij} = \frac{1}{i+j-1}$$

gegeben ist.

Zur Stabilität des Gaußalgorithmus

1. Ein kleines Pivotelement führt zu einem instabilen Algorithmus
2. Versuche stets durch Zeilentausch, das Pivotelement möglichst groß zu halten