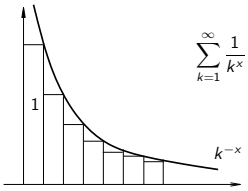
Verallgemeinerung der harmonischen Reihe: Riemannsche Zetafunktion

$$\zeta(x) = \sum_{k=1}^{\infty} \frac{1}{k^x} \quad \text{für} \quad x > 1$$

Konvergenzbeweis mittels Integralschranke



$$\sum_{k=1}^{\infty} \frac{1}{k^x} \leq 1 + \int_1^{\infty} \frac{dy}{y^x} = 1 + \frac{y^{-x+1}}{1-x}\Big|_1^{\infty}$$

$$= 1 - \frac{1}{1-x} = \frac{x}{x-1}$$

$$\begin{aligned}
\Delta\zeta_n(x) &= \zeta(x) - \zeta_n(x) = \sum_{k-1}^{\infty} k^{-x} - \sum_{k=1}^{n} k^{-x} \\
&= \sum_{k=n+1}^{\infty} k^{-x} \leq \int_n^{\infty} k^{-x} dk = \frac{k^{1-x}}{1-x}\Big|_{k=n}^{\infty} = \frac{1}{k^{1-x}(1-x)}\Big|_{k=n}^{\infty} \\
&= 0 - \frac{1}{n^{x-1}(1-x)} = \frac{1}{n^{x-1}(x-1)} \leq \texttt{tol} \\
\Rightarrow \quad & n \geq \sqrt[x-1]{\frac{1}{\texttt{tol}(x-1)}}
\end{aligned}$$

Partialsummen:

$\zeta_n(x) = \sum_{k=1}^{n} \frac{1}{k^x}$ wachsen monoton mit n und sind nach oben durch $\frac{x}{x-1}$ beschränkt, haben also einen eindeutigen Grenzwert $\zeta(x)$.

Praktische Notwendigkeit: Diskretisierung

Hier, wie häufig in numerischer Mathematik muss mathematisches Problem durch Ausführung endlich vieler Operationen auf endlich vielen Variablen annäherungsweise gelöst werden. Hier einfach Annäherung von $\zeta(x)$ durch $\zeta_n(x)$. Der entsprechende Abbruchfehler $|\zeta(x) - \zeta_n(x)|$ kann hier einfach mit Hilfe einer Integralschranke abgeschätzt werden. Unabhängig vom in der Numerischen Analysis betrachteten *Diskretisierungsfehler* ist der Rundungsfehler zu berücksichtigen.

Rundungsfehlerabschätzung bei Riemann

Für $b_i > 0$

$$\begin{aligned}
& fl\big(\big(\ldots\big(\big(\big(b_1 + b_2\big) + b_3\big) + b_4\big)\ldots + b_{n+1}\big) + b_n\big) \\
= \;& \big(\ldots\big(\big(\big(b_1 + b_2\big)\big(1+\varepsilon_1\big) + b_3\big)\big(1+\varepsilon_2\big) + b_4\big)\big(1+\varepsilon_3\big)\ldots + b_n\big)\big(1+\varepsilon_{n-1}\big) \\
= \;& b_1\big(1+\tilde{\varepsilon}_1\big)^{n-2} + b_2\big(1+\tilde{\varepsilon}_1\big)^{n-1} + b_3\big(1+\tilde{\varepsilon}_2\big)^{n-2} + \ldots + b_n\big(1+\tilde{\varepsilon}_{n-1}\big)^1 \\
\Longrightarrow \;& \big|fl\big(b_1 + \ldots + b_n\big) - \big(b_1 + b_2 + \ldots + b_n\big)\big| \\
\leq \;& b_1\Big[\big(1+\text{eps}\big)^{n-1} - 1\Big] + b_2\Big[\big(1+\text{eps}\big)^{n-1} - 1\Big] + \ldots + b_n\big(1+\text{eps}\big) \\
\approx \;& \big[\big(b_1 + b_2\big)(n-1) + (n-2)b_3 + (n-3)b_4 + \ldots + b_n\big]\,\text{eps}
\end{aligned}$$

Mit anderen Worten:

Der an der $j+1$-ten Stelle eingebrachte Summand wird $(n-j)$ -mal in den Operationen von einer Rundung betroffen und trägt entsprechend zur Gesamtfehlerschranke bei.

Schlussfolgerung:

Um Rundungsfehler zu minimieren sollten Summen möglichst vom kleinsten zum größten Summanden gebildet werden. Bei konvergenten (hoffentlich monoton fallenden) Reihen sollte von hinten, d.h. rückwärts summiert werden.

Beispiel D.7 ($\zeta(2)$ auf G's Laptop in einfacher Genauigkeit:)

$$\zeta(2) = \sum_{k=1}^{\infty} \frac{1}{k^2} \equiv \begin{cases} \pi^2/6 = 1.6449340\ldots & \text{exakt} \\ 1.6447253 & \text{vorwärts bis. liegen bleiben} \quad n = 4097 \\ 1.6446900 & \text{rückwärts vom gleichen} \quad n = 4097 \\ 1.6449339 & \text{rückwärts mit} \quad n = 2^{23} = 8388608 \end{cases}$$

Bemerkung:

Durch Rückwärtssummation können deutlich mehr Summanden der Form $1/k^{-}x$ mit $n > 4097$ ihren Beitrag zur Gesamtsumme leisten. Mehr Summanden zu benutzten bedeutet aber, den *Diskretisierungsfehler* zu verringern und damit den exakten Wert $\zeta(x)$ besser zu approximieren.

Abschätzung des Rundungsfehlers

Vorwärts:

$$\text{eps}\sum_{k=1}^{n}\frac{1}{k^2}(n-k) = \text{eps}\sum_{k=1}^{n}\left(\left(\frac{n}{k^2}\right) - \frac{1}{k}\right) \approx \text{eps}\left[n\frac{\pi^2}{6} - \ln(n)\right] \approx \text{eps}\cdot n\cdot\frac{\pi^2}{6}$$

Rückwärts:

$$\text{eps}\sum_{k=1}^{n}\frac{1}{k^2}k = \text{eps}\sum_{k=1}^{n}\frac{1}{k} \approx \text{eps}\cdot\ln(n)$$

Vergleich:

$$\text{eps}\cdot n\cdot\frac{\pi^2}{6} \gg \text{eps}\cdot\ln(n)$$

Konvergenzbeschleunigung (1. Stufe nach Wijngaard)

Beobachtung bei Riemann:

$$\zeta(x) = 1 + \frac{1}{2^x} + \cdots + \underbrace{\frac{1}{100^x} + \frac{1}{101^x} + \frac{1}{102^x} + \cdots}_{\text{spätere Terme ändern sich nur langsam}}$$

Idee:

Erste grobe Annäherung mit $b_k = \frac{1}{k^x}$

$$a_1 = b_1 + b_2\cdot 2 + b_4\cdot 4 + \cdots + (b_{2^i})\cdot 2^i \quad > \quad \zeta = b_1 + b_2 + b_3 + b_4\ldots$$

Reihe der $2^i b_{2^i}$ konvergiert viel schneller als $\sum b_k$. Die Korrektur erfolgt durch transformierte Terme

$$a_j = \sum_{i=1}^{\infty}(b_{j\,2^i})\,2^i.$$

Satz D.8

Satz: Für $b_k = k^{-x}$ oder andere monoton konvergierende Reihen gilt im Grenzwert

$$\sum_{k=1}^{\infty} b_k = \sum_{j=1}^{\infty}(-1)^{j-1}\,a_j \quad .$$

Bemerkung

Bemerkung: Die neue Reihe ist alternierend, wobei $a_j \geq b_j$, d.h. die einzelnen Terme gehen nicht schneller gegen Null als die der Ursprungsreihe.

Idee des Beweises:

Betrachte, wie oft b_k in a_j auftritt

Vorz	$j\backslash k$	1	2	3	4	5	6	7	8	9	10	11	12
+	1	1	2	–	4	–	–	–	8	–	–	–	–
–	2	–	1	–	2	–	–	–	4	–	–	–	–
+	3	–	–	1	–	–	2	–	–	–	–	–	4
–	4	–	–	–	1	–	–	–	2	–	–	–	–
+	5	–	–	–	–	1	–	–	–	–	2	–	–
–	6	–	–	–	–	–	1	–	–	–	–	–	2
+	7	–	–	–	–	–	–	1	–	–	2	–	–
$\sum$ mit Vorzeichen		1	1	1	1	1	1	1	.....				

Bemerkung

Bei Riemann können die $a_i = a_i(x)$ sogar explizit berechnet werden.

Schlussfolgerungen aus dem Summationsbeispiel

- Die Behandlung mathematischer und anderer Modellierungsprobleme bedingt das Auftreten von *Abbruchs-* ≡ *Diskretisierungsfehlern* sowie Rundungsfehlern. Beide sollten abgeschätzt und möglichst minimiert werden.
- Gleitpunktarithmetik ist weder kommutativ noch assoziativ, distributiv usw.
  Spezielle Konsequenz: Betragsmäßig fallende Reihen von hinten summieren!
- Es ist erstaunlich einfach, an die Grenzen der Gleitpunkt- und Ganzzahlarithmetik zu stoßen.
- Viele Jobs (≡ Programme, Daten) laufen entweder im Sekunden- oder Stundenbereich. Beobachtung der Abarbeitung im Minutenbereich ist relativ selten.
- *Mathematisch endlich* ist nicht gleich *rechentechnisch endlich*.

D - 4 Lösung (nicht-)linearer Gleichungssysteme

Methoden zur Lösung des linearen Problemes $Ax = b$ mit $\dim(x) = \dim(b) = n$

- Cramersche Regel $x_i = (-1)^i det(A_i)/det(A)$ für $i = 1..n$
  (In A_i wird die $i-$te Spalte von A durch b ersetzt)
- Gauss-Elimination $\approx$ P A = LU Faktorisierung
  (P Permutation, L unterhalb und U oberhalb dreiecksförmig)
- Schmidt-Ortogonalisierung $\approx$ A = QR Faktorisierung
  (Q orthogonal, R oberhalb dreiecksförmig)
- Fixpunkt Iteration $x \leftarrow x - M\,F(x)$ mit $F(x) \equiv Ax - b$
  ($M \in \mathbb{R}^{n\times n}$ angenäherte Inverse so dass $M\,A \approx I$)

Hinweise:

- Für (eindeutige) Lösbarkeit ist überall $det(A) \neq 0$ vorrauszusetzen.
- Löse $LUx = b$ bzw $QRx = b$ durch Substitution/Transponierung.
- Die letzte Methode lässt sich auch auf nichtlineares $F(x)$ anwenden.

Linearisierung des 'Freistoss' Beispieles

Das nichtlineare System von 3 Gleichungen in 3 Unbekannten

$$\begin{array}{lcll} F_1(x_1,x_2,x_3) & = & x_1 * x_2 - 4.9 * x_1^2 - 2 & = 0 \\ F_2(x_1,x_2,x_3) & = & 10 * ln(1 + 0.1 * x_3 * x_1) - 25 & = 0 \\ F_3(x_1,x_2,x_3) & = & (x_2 - 9.8 * x_1) * (\frac{1}{x_3} + 0.1 * x_1) + \frac{1}{\sqrt{3}} & = 0 \end{array}$$

hat die *Jacobimatrix*

$$F'(x) \equiv \frac{\partial}{\partial x}F(x) \equiv \left[\frac{\partial F_i}{\partial x_j}\right]_{j=1,2,3}^{i=1,2,3}$$

$$\equiv \begin{bmatrix} x_2 - 9.8 * x_1 & x_1 & 0 \\ \frac{x_3}{1+0.1*x_1*x_3} & 0 & \frac{x_1}{1+0.1*x_1*x_3} \\ z(x) & \frac{1}{x_3} + 0.1 * x_1 & -\frac{x_2 - 9.8*x_1}{x_3^2} \end{bmatrix}$$

mit $z(x) \equiv -9.8 * (\frac{1}{x_3} + \frac{x_1}{10}) + \frac{1}{10}(x_2 - 9.8 * x_1) = \frac{x_2}{10} - 9.8\left(\frac{1}{x_3} + \frac{1}{5}x_1\right)$

Linearisierung durch Jacobimatrix

Falls für $F : \mathbb{R}^n \to \mathbb{R}^n$ die n^2 Komponenten der Jacobimatrix

$$F'(x) \equiv \frac{\partial}{\partial x} F(x) \equiv \left[\frac{\partial F_i}{\partial x_j} \right]_{j=1,\ldots,n}^{i=1,\ldots,n}$$

bezüglich jeder der Variablen $x_1, \ldots, x_n$ Lipschitz-stetig sind, so lässt sich aus dem Hauptsatz der Differential- und Integralrechnung herleiten, dass für jeden *Schritt* $s \in \mathbb{R}^n$ gilt

$$\| F(x+s) \; - \; [\, F(x) \; + \; F'(x)\, s \,] \, \| \quad \leq \quad \gamma \|s\|^2$$

Hierbei ist $F'(x)s$ ein Matrix-Vektor Produkt und $\| \cdot \|$ ist eine Vektor- bzw. Matrixnorm (siehe Abschnitt B-3) mit

$$\|F'(x) - F'(y)\| \quad \leq \quad \gamma \|x - y\|$$

$F_x(s) \equiv F(x) + F'(x)\, s$ ist als Funktion des variablen Vektors s die Linearisierung (verallgemeinerte Tangente) von F an der Stelle x.

Newton's Methode im Vektorfall

Setzt man die Linearisierung $F_x(s) = F(x) + F'(x)s$ zu null so erhält man das lineare Gleichungssystem

$$As = b \quad \text{mit} \quad A = F'(x) \quad \text{und} \quad b = -F(x)$$

Die Lösung lässt sich ausdrücken als

$$s = A^{-1} b = -F'(x)^{-1} F(x)$$

und heisst *Newtonschritt*.
Wiederholte Berechnung von s und anschliessende Inkrementierung $x \leftarrow x + s$ ergibt Newton's Methode

$$x^{(k+1)} \equiv x^{(k)} + s^{(k)} \quad \text{mit} \quad F'(x^{(k)})\, s^{(k)} = -F(x^{(k)}) \quad \text{für} \quad k = 0, 1, \ldots$$

Hierbei zählt der hochgestellte Index (k) die Iterationen.

Warnung:

- Das Verfahren muss abgebrochen werden wenn $det(F'(x^{(k)}))$ null oder sehr klein ist.
- Im letzteren Falle werden die Schritte $s^{(k)}$ typischerweise sehr gross und führen häufig zu Argumenten $x^{(k+1)}$ wo F garnicht mehr ausgewertet werden kann.
- Zur Vermeidung dieses Problems wird $s^{(k)}$ manchmal mit einem Dämpfungsfaktor $\alpha^{(k)} < 1$ multipliziert, der dann *Schrittweite* genannt wird. Wir iterieren also effektiv

  $$x^{(k+1)} = x^{(k)} - \alpha^{(k)} F'(x^{(k)})^{-1} F(x^{(k)})$$

  Die Bestimmung eines geeigneten $\alpha^{(k)}$ heisst auch Strahlsuche (engl: Line Search).

Lokale Konvergenz von Newton

Satz D.9 (Satz von Kantorovich)

Sei die Vektorfunktion $F : \mathbb{R}^n \to \mathbb{R}^n$ einmal differenzierbar und besitze ihre Jacobimatrix $F'(x) \in \mathbb{R}^{n \times n}$ die Lipschitzkonstante γ.
Weiterhin sei $x^{(0)}$ ein Punkt an dem $F'(x^{(0)})$ regulär ist und somit eine Inverse $F'(x^{(0)})^{-1}$ existiert. Mit $\| \cdot \|$ als induzierte Matrix-Norm folgt dann aus

$$\left\| F'(x^{(0)})^{-1} \right\|^2 \left\| F(x^{(0)}) \right\| \leq \frac{1}{2\gamma}$$

dass Newton's Methode zu einer Lösung $x^{(*)}$ mit $F(x^{(*)}) = 0$ konvergiert. Die Konvergenzgeschwindigkeit ist quadratisch in dem Sinne dass für eine Konstante c und alle k gilt

$$\left\| x^{(k+1)} - x^{(*)} \right\| \leq c \left\| x^{(k)} - x^{(*)} \right\|^2$$

Bemerkung:

Je nichtlinearer ein Problem umso grösser ist γ und desto stärker ist damit die Bedingung an $x^{(0)}$. Wird praktisch nie überprüft !!!!