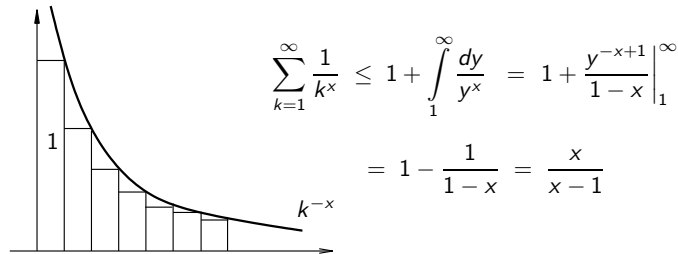


## Verallgemeinerung der harmonischen Reihe: Riemannsche Zetafunktion

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

Konvergenzbeweis mittels Integralschranke



- 33 -

## 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.

- 35 -

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

- 34 -

## Rundungsfehlerabschätzung bei Riemann

Für  $b_i > 0$

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

## 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.

- 36 -

### 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:)

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

### Bemerkung:

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

## Konvergenzbeschleunigung (1. Stufe nach Wijngaard)

Beobachtung bei Riemann:

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

### Idee:

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

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

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

$$a_j = \sum_{i=1}^{\infty} (b_{j2^i}) 2^i.$$

## 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:

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

### 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	$j \setminus k$	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.

## D-4 Lösung (nicht-)linearer Gleichungssysteme

Methoden zur Lösung des linearen Problems  $Ax = b$  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  $\approx PA = LU$  Faktorisierung  
 (  $P$  Permutation,  $L$  unterhalb und  $U$  oberhalb dreiecksförmig )
- ▶ Schmidt-Orthogonalisierung  $\approx A = QR$  Faktorisierung  
 (  $Q$  orthogonal,  $R$  oberhalb dreiecksförmig )
- ▶ Fixpunkt Iteration  $x \leftarrow x - MF(x)$  mit  $F(x) \equiv Ax - b$   
 (  $M \in \mathbb{R}^{n \times n}$  angenäherte Inverse so dass  $MA \approx I$  )

### 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.

## Schlussfolgerungen aus dem Summationsbeispiel

- ▶ Die Behandlung mathematischer und anderer Modellierungsprobleme bedingt das Auftreten von *Abbruchs-*  $\equiv$  *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 ( $\equiv$  Programme, Daten) laufen entweder im Sekunden- oder Stundenbereich. Beobachtung der Abarbeitung im Minutenbereich ist relativ selten.
- ▶ *Mathematisch endlich* ist nicht gleich *rechentechnisch endlich*.

## Linearisierung des 'Freistoss' Beispiels

Das nichtlineare System von 3 Gleichungen in 3 Unbekannten

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

hat die *Jacobimatrix*

$$\begin{aligned} F'(x) &\equiv \frac{\partial}{\partial x} F(x) \equiv \left[ \frac{\partial F_i}{\partial x_j} \right]_{i=1,2,3}^{j=1,2,3} \\ &\equiv \begin{bmatrix} x_2 - 9.8 * x_1 & x_1 & 0 \\ \frac{x_3}{1+0.1*x_1*x_3} & 0 & \frac{x_1}{1+0.1*x_1*x_3} \\ z(x) & \frac{1}{x_3} + 0.1 * x_1 & -\frac{x_2 - 9.8 * x_1}{x_3^2} \end{bmatrix} \end{aligned}$$

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

## Linearisierung durch Jacobimatrix

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

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

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

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

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

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

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

- 45 -

## 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).

- 47 -

## 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 *Newtonschrift*.

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

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

Hierbei zählt der hochgestellte Index ( $k$ ) die Iterationen.

- 46 -

## Lokale Konvergenz von Newton

### Satz D.9 (Satz von Kantorovich)

Sei die Vektorfunktion  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  einmal differenzierbar und besitze ihre Jacobimatrix  $F'(x) \in \mathbb{R}^{n \times n}$  die Lipschitzkonstante  $\gamma$ . 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

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

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

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

### Bemerkung:

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

- 48 -