Technische Universität München
Institut für Informatik
Prof. Dr. Thomas Huckle
Dipl.-Math. Jürgen Bräckle
Nikola Tchipev, M.Sc.

SoSe 2014

Numerisches Programmieren, Übungen

Musterlösung 8. Übungsblatt: Extrapolation, Diskrete Fourier-Transformation

1) Extrapolation: Numerische Quadratur mit hoher Ordnung

Vorbemerkung: Von Aitken-Neville zur Romberg-Quadratur – Herleitung des Algorithmus

Im Kapitel zur Polynominterpolation wurde erklärt, wie man zu n gegebenen Stützpunkten ein Polynom $n-1$-ten Grades aufstellt, das sämtliche Stützpunkte interpoliert. Der Algorithmus von Aitken-Neville wurde verwendet, um zu den gegebenen Stützpunkten direkt den Wert des zugehörigen Interpolationspolynoms an einer gegebenen Stelle zu berechnen.

Die Grundidee des Romberg-Verfahrens ist, zu verschiedenen h's (h ist die Breite eines einzelnen Teilintervalls bei der Trapezsumme) die Trapezsumme zu berechnen, und dann die h's als Stützstellen und die zugehörigen Trapezsummen als Stützwerte eines zu interpolierenden Polynoms zu interpretieren. Wenn man dieses Polynom an der Stelle $h = 0$ auswertet, so erhält man einen Näherungswert für die Trapezsumme mit $h \to 0$.

Da die Indizes des Aitken-Neville-Algorithmus und des Romberg-Algorithmus nicht übereinstimmen, müssen die Indizes zunächst umgerechnet werden. Betrachtet man die Visualisierung des Aitken-Neville-Dreiechsschemas (dieses findet sich auf Übungsblatt 5 in der ersten Aufgabe), so stehen in der ersten Zeile alle Werte, für die i_{ait} Null ist. Diese Werte sind aber zugleich jeweils das Ende einer Gegendiagonalen im Romberg-Algorithmus. Das Ende der Gegendiagonalen bedeutet im Romberg-Verfahren aber immer, daß k_{rom} Null ist. In der n-ten Zeile der Abbildung ist $i_{ait} = k_{rom} = n - 1$. Um den Aitken-Neville-Algorithmus also in den Romberg-Algorithmus zu überführen, wird i_{ait} durch k_{rom} ersetzt:

$$i_{ait} = k_{rom} \tag{1}$$

Betrachtet man nun eine Spalte in der Abbildung, so ist k_{ait} konstant, i_{rom} und k_{rom} steigen aber linear an (in jeder Zeile um 1). In der ersten Spalte ($k_{ait} = 0$) gilt $i_{rom} = k_{rom}$, in der zweiten Spalte ($k_{ait} = 1$) gilt $i_{rom} = k_{rom} + 1$, und in der n-ten Spalte ($k_{ait} = n - 1$) gilt $i_{rom} = k_{rom} + n - 1$. Es folgt also:

$$k_{ait} = i_{rom} - k_{rom} \tag{2}$$

Damit kann man nun aus der Gleichung des Aitken-Neville-Verfahrens die entsprechende Gleichung des Romberg-Verfahrens herleiten:

$$\begin{aligned}
p[i] &= \frac{(x_{i+k}-x)}{(x_{i+k}-x_i)}\cdot p[i] + \frac{(x-x_i)}{(x_{i+k}-x_i)}\cdot p[i+1] \\
&= \frac{(x_{i+k}-x)\cdot p[i] + (x-x_i)\cdot p[i+1]}{(x_{i+k}-x_i)} \\
&= \frac{(x_{i+k}-x)\cdot p[i] + (x-x_i+x_{i+k}-x_{i+k})\cdot p[i+1]}{(x_{i+k}-x_i)} \\
&= \frac{(x_{i+k}-x)\cdot p[i] + (x_{i+k}-x_i)\cdot p[i+1] - (x_{i+k}-x)\cdot p[i+1]}{(x_{i+k}-x_i)} \\
&= p[i+1] + \frac{(x_{i+k}-x)}{(x_{i+k}-x_i)}\cdot (p[i]-p[i+1]) \\
&= p[i+1] + \frac{(x-x_{i+k})}{(x_{i+k}-x_i)}\cdot (p[i+1]-p[i])
\end{aligned}$$

Bis zu diesem Punkt sind die Indizes noch diejenigen aus dem Aitken-Neville-Verfahren. Wenn diese mit Hilfe der Gleichungen (1) und (2) durch die Romberg-Indizes ersetzt werden erhält man:

$$\begin{aligned}
p[k] &= p[k+1] + \frac{(x-x_{k+i-k})}{(x_{k+i-k}-x_k)}\cdot (p[k+1]-p[k]) \\
&= p[k+1] + \frac{(x-x_i)}{(x_i-x_k)}\cdot (p[k+1]-p[k])
\end{aligned}$$

Nun werden die Stützstellen x_j durch h_j^2 ersetzt und $p[j]$ durch $T[j]$ ersetzt. Außerdem wird x auf Null gesetzt:

$$\begin{aligned}
T[k] &= T[k+1] + \frac{(0-h_i^2)}{(h_i^2-h_k^2)}\cdot (T[k+1]-T[k]) \\
&= T[k+1] + \frac{-h_i^2\cdot \frac{1}{h_i^2}}{(h_i^2-h_k^2)\cdot \frac{1}{h_i^2}}\cdot (T[k+1]-T[k]) \\
&= T[k+1] + \frac{1}{\frac{h_k^2}{h_i^2}-1}\cdot (T[k+1]-T[k])
\end{aligned}$$

Diese Gleichung entspricht genau der Gleichung aus dem Romberg-Verfahren. Zuletzt müssen nur noch die Schleifen des Algorithmus angepasst werden. Beim Aitken-Neville-Algorithmus aus der Vorlesung hat die erste Stützstelle den Index 0, beim Romberg-Algorithmus aber den Index 1. Passt man die Schleifen ensprechend an, so erhält man schließlich den Romberg-Algorithmus aus der Vorlesung:

```
for i=1:n
    waehle n[i];
    h[i]:=(b-a)/n[i];
    T[i]:=Trapezsumme zur Schrittweite h[i]
    for k:=i-1:-1:1
        T[k] := T[k+1] + 1/(h[k]^2/h[i]^2-1)*(T[k+1]-T[k])
```

```
    end
end
```

a) Wir leiten die Formel für die Romberg-Quadratur nun über die Reihenentwicklung her. Dazu multiplizieren wir $T(h_1)$ und $T(h_2)$ mit h_2^2 bzw.h_1^2 und ziehen die beiden Gleichungen voneinander ab:

$$\begin{array}{rrcll} & T(h_1) & = & I(f) + \tau_1 h_1^2 + \tau_2 h_1^4 + \ldots & \mid \cdot h_2^2 \\ - & T(h_2) & = & I(f) + \tau_1 h_2^2 + \tau_2 h_2^4 + \ldots & \mid \cdot h_1^2 \\ \hline \end{array}$$
$$T(h_1) \cdot h_2^2 - T(h_2) \cdot h_1^2 = (h_2^2 - h_1^2) I(f) + h_1^2 h_2^2 (h_1^2 - h_2^2)\tau_2 + \mathcal{O}(h^6)$$

$$\begin{aligned} I(f) &= \frac{T(h_1) \cdot h_2^2 - T(h_2) \cdot h_1^2}{h_2^2 - h_1^2} - h_1^2 h_2^2 \tau_2 + \mathcal{O}(h^4) \\ &= \frac{T(h_1) \cdot h_2^2 - T(h_2) \cdot h_1^2}{h_2^2 - h_1^2} + \mathcal{O}(h^4) \end{aligned}$$

Der Fehler beträgt $O(h_1^2 \cdot h_2^2)$.

Bemerkung: Um die im Algorithmus verwendete Formel zu erhalten, formen wir einfach um:

$$\begin{aligned} I(f) &\approx \frac{T(h_1) \cdot h_2^2 - T(h_2) \cdot h_1^2}{h_2^2 - h_1^2} = \frac{T(h_2) \cdot h_1^2 - T(h_1) \cdot h_2^2}{h_1^2 - h_2^2} \\ &= \frac{T(h_2) \cdot [(h_1^2 - h_2^2) + h_2^2] - T(h_1) \cdot h_2^2}{h_1^2 - h_2^2} = T(h_2) + \frac{(T(h_2) - T(h_1)) \cdot h_2^2}{h_1^2 - h_2^2} \\ &= T(h_2) + \frac{T(h_2) - T(h_1)}{(h_1^2/h_2^2) - 1} \end{aligned}$$

b) Es soll die Funktion $f(x) = -x^2 + 4$ mit Hilfe des Romberg-Verfahrens zwischen $a = -2$ und $b = 2$ integriert werden. Die Teilintervallbreite halbiert sich in jeden Schritt. Der analytisch korrekte Wert lautet:

$$\int_{-2}^{2} f(x)\, dx = 10\frac{2}{3}$$

Zunächst werden die Trapezsummen für $n = 1$, $n = 2$ und $n = 4$ berechnet:

- $h_1 = \frac{b-a}{2^0} = 4$:

$$\begin{aligned} Q_{TS}(f, h_1) &= h_1 \cdot \left(\frac{f_0}{2} + \frac{f_1}{2} \right) \\ &= 4 \cdot \left(\frac{0}{2} + \frac{0}{2} \right) = 0 \end{aligned}$$

- $h_2 = \frac{b-a}{2^1} = 2$:

$$\begin{aligned} Q_{TS}(f, h_2) &= h_2 \cdot \left(\frac{f_0}{2} + f_1 + \frac{f_2}{2} \right) \\ &= 2 \cdot \left(\frac{f(-2)}{2} + f(0) + \frac{f(2)}{2} \right) \\ &= 2 \cdot \left(\frac{0}{2} + 4 + \frac{0}{2} \right) = 8 \end{aligned}$$

- $h_3 = \frac{b-a}{2^2} = 1$:

$$\begin{aligned} Q_{TS}(f, h_3) &= h_3 \cdot \left(\frac{f_0}{2} + f_1 + f_2 + f_3 + \frac{f_4}{2} \right) \\ &= 1 \cdot \left(\frac{f(-2)}{2} + f(-1) + f(0) + f(1) + \frac{f(2)}{2} \right) \\ &= \left(\frac{0}{2} + 3 + 4 + 3 + \frac{0}{2} \right) = 10 \end{aligned}$$

Nun wird das Romberg-Verfahren durchgeführt:

- $i = 1$:

$$T[1] = Q_{TS}(f, h_1) = 0$$

- $i = 2$:

$$T[2] = Q_{TS}(f, h_2) = 8$$

  - $k = 1$

$$T[1] = T[2] + \frac{1}{\frac{h_1^2}{h_2^2} - 1} \cdot (T[2] - T[1]) = 8 + \frac{1}{\frac{16}{4} - 1} \cdot 8 = 10\frac{2}{3}$$

- $i = 3$:

$$T[3] = Q_{TS}(f, h_3) = 10$$

  - $k = 2$

$$T[2] = T[3] + \frac{1}{\frac{h_2^2}{h_3^2} - 1} \cdot (T[3] - T[2]) = 10 + \frac{1}{\frac{4}{1} - 1} \cdot 2 = 10\frac{2}{3}$$

  - $k = 1$

$$T[1] = T[2] + \frac{1}{\frac{h_1^2}{h_3^2} - 1} \cdot (T[2] - T[1]) = 10\frac{2}{3} + \frac{1}{\frac{16}{4} - 1} \cdot 0 = 10\frac{2}{3}$$

Dargestellt als Tableau (analog zum Schema von Aitken und Neville) sieht das Ganze wie folgt aus:

h_i	i	$O(h^2)$		$O(h^4)$		$O(h^6)$
4	1	0				
			$\searrow$			
2	2	8	$\rightarrow$	$10\frac{2}{3}$		
			$\searrow$		$\searrow$	
1	3	10	$\rightarrow$	$10\frac{2}{3}$	$\rightarrow$	$10\frac{2}{3}$

Das mit Hilfe des Romberg-Verfahrens berechnete Integral ist in diesem Fall exakt. Die Funktion $f(x)$ musste insgesamt nur fünf Mal ausgewertet werden. Bei der Trapezsumme mit $n = 8$ von Aufgabenblatt 6 musste die Funktion 9 mal ausgewertet werden, das Ergebnis war aber dennoch schlechter (10.5).

Hinweis:

Bei der Extrapolation mit drei Schrittweiten ist der Fehler zwischen echtem Integral und Wert der Extrapolation $O(h_1^2 \cdot h_2^2 \cdot h_3^2) = O(h^6)$.
Für Polynome vom Grad ≤ 5 liefert die Extrapolation hier das exakte Ergebnis.
Bei höheren Graden (z.B. x^8) nicht. Wenn beispielsweise die Funktion $g(x) = x^8$ von 0 bis 1 integriert werden soll, so ist das exakte Ergebnis $I(f) = \frac{1}{9} = 0,\overline{1}$

Es ergeben sich folgende Trapezsummen:

- $TS(h = 1) = 0.5$
- $TS(h = \frac{1}{2}) = 0.25195$
- $TS(h = \frac{1}{4}) = 0.15100$
- $TS(h = \frac{1}{8}) = 0.12141$

Berechnet man zu den ersten drei Trapezsummen wieder die Integration mit dem Romberg-Verfahren, so erhält man:

$$Q_{romberg} = 0.11389 \tag{3}$$

Dieses Ergebnis ist zwar nicht exakt, aber deutlich besser als die Trapezsumme mit $n = 8$, und das mit nur fünf statt neun Funktionsauswertungen.

2) Frequenzanalyse

Die Signale s_1, s_2, s_3 sind folgendermassen mit den Frequenzspektren verknüpft:

$$\begin{aligned} s_1 &= e^{3it} &&\leftrightarrow f_3 \\ s_2 &= 0.2\ i + 0.8\ e^{it} &&\leftrightarrow f_1 \\ s_3 &= e^{it} + 0.2\ e^{12it} &&\leftrightarrow f_2, \end{aligned}$$

wobei $t \in [0; 2\pi]$.
Zusätzliche Begründungen:

s_1 einzelne gleichmässige mittelfrequentierte Schwingung $\rightarrow f_3$

s_2 einzelne langsame Schwingung + leichte Verschiebung nach oben $\rightarrow f_1$

s_3 dominante langsame Schwingung + kleine hochfrequentierte Schwingung $\rightarrow f_2$

3) Eigenschaften der diskreten Fourier-Transformation (DFT)

Wiederholung einiger Regeln beim Rechnen mit komplexen Zahlen ($z \in \mathbb{C}$):

- $z = x + iy, \quad x = Re(z),\ y = Im(z)$
- $\overline{z} = x - iy$ (konjugiert Komplexes), $\quad \overline{z \cdot w} = \overline{z} \cdot \overline{w}$
- $|z| = \sqrt{Re(z)^2 + Im(z)^2} \ = \ \sqrt{x^2 + y^2}$
- e^{it} durchläuft den Einheitskreis gegen den Uhrzeigersinn (beginnend bei (1,0)). Außerdem ist die Funktion 2π-periodisch. Daher gilt:

$$\begin{aligned} e^{i\cdot 0} \ = \ e^{i\cdot 2k\pi} \ &= \ 1 \in \mathbb{R},\ k \in \mathbb{Z} \\ e^{-i\cdot\pi} \ &= \ -1 \in \mathbb{R}. \end{aligned}$$

a) Wir betrachten den Ausdruck auf der rechten Seite:

$$\begin{aligned} \left(\frac{1}{n}\,\overline{IDFT(\overline{v})}\right)_l \ &= \ \frac{1}{n}\overline{\sum_{k=0}^{n-1} \overline{v_k}\ \omega^{kl}} \ = \ \frac{1}{n}\sum_{k=0}^{n-1} \overline{\overline{v_k}\ \omega^{kl}} \ = \ \frac{1}{n}\sum_{k=0}^{n-1} \overline{\overline{v_k}}\ \overline{\omega}^{kl} \\ &= \ \frac{1}{n}\sum_{k=0}^{n-1} v_k\ \overline{\omega}^{kl} \ = \ DFT(v)_l, \quad l = 0, \ldots, n-1. \end{aligned}$$

b) Es gilt für alle $k = 0, \ldots, n-1$:

$$DFT(v+u)_k \ = \ \frac{1}{n}\sum_{j=0}^{n-1}(v+u)_j\ \overline{\omega}^{jk} \ = \ \frac{1}{n}\sum_{j=0}^{n-1} v_j\ \overline{\omega}^{jk} + \frac{1}{n}\sum_{k=0}^{n-1} u_j\ \overline{\omega}^{kl} \ = \ DFT(v)_k + DFT(u)_k.$$

c) Mit Hilfe der geometrischen Summenformel

$$\sum_{j=0}^{n} z^j \ = \ \frac{1 - z^{n+1}}{1 - z}, \quad \text{für } z \neq 1$$

berechnen wir mit $z = \overline{\omega}^l$:

$$\begin{aligned} \text{für } l = 0: \quad & \sum_{k=0}^{n-1} \overline{\omega}^{0k} \;=\; \sum_{k=0}^{n-1} e^0 \;=\; n \\ \text{für } l = 1, \ldots, n-1: \quad & \sum_{k=0}^{n-1} \overline{\omega}^{kl} \;=\; \frac{1 - (\overline{\omega}^l)^n}{1 - \overline{\omega}^l} \;=\; \frac{1 - e^{-i\frac{2\pi}{n}nl}}{1 - e^{-i\frac{2\pi}{n}l}} \;=\; 0. \end{aligned}$$

Außerdem erhalten wir mit dem gleichen Trick:

$$\sum_{k=0}^{n-1} \omega^{kl}\, \overline{\omega}^{kj} \;=\; \sum_{k=0}^{n-1} \omega^{k(l-j)} \;=\; \sum_{k=0}^{n-1} e^{i\frac{2\pi}{n}k(l-j)} \;=\; \begin{cases} n, & \text{für } l = j \\ 0, & \text{für } l \neq j \end{cases}.$$

d) Wir berechnen mit den Ergebnissen aus c):

$$\begin{aligned} IDFT(DFT(v))_l \;&=\; \sum_{k=0}^{n-1} c_k \omega^{kl} \;=\; \sum_{k=0}^{n-1} \left(\frac{1}{n} \sum_{j=0}^{n-1} v_j \overline{\omega}^{jk} \right) \omega^{kl} \\ &=\; \frac{1}{n} \sum_{j=0}^{n-1} v_j \left(\sum_{k=0}^{n-1} \overline{\omega}^{jk} \omega^{kl} \right) \;\overset{c)}{=}\; \frac{1}{n} \sum_{j=0}^{n-1} v_j \left(\delta_{jl} \cdot n \right) \;=\; v_l. \end{aligned}$$

e) Für die drei Beispielvektoren erhalten wir folgende Ergebnisse:

- $(DFT(a))_k = \dfrac{1}{n} \displaystyle\sum_{j=0}^{n-1} a_j \overline{\omega}^{jk} = \frac{1}{n}\omega^0 = \frac{1}{n}, \quad \forall k = 0, \ldots, n-1,$
- $(DFT(b))_k = \dfrac{1}{n} \displaystyle\sum_{j=0}^{n-1} b_j \overline{\omega}^{jk} = \frac{i}{n} \sum_{j=0}^{n-1} \overline{\omega}^{jk} \overset{c)}{=} \begin{cases} i, & \text{für } k = 0 \\ 0, & \text{für } k = 1, \ldots, n-1 \end{cases},$
- $(DFT(c))_k = \dfrac{1}{n} \displaystyle\sum_{j=0}^{n-1} c_j \overline{\omega}^{jk} = \frac{1}{n} \sum_{j=0}^{n-1} (-1)^j \overline{\omega}^{jk} = \frac{1}{n} \sum_{j=0}^{n-1} (e^{i\pi})^j \overline{\omega}^{jk}$
  $= \dfrac{1}{n} \displaystyle\sum_{j=0}^{n-1} (e^{i\pi})^j e^{-i\frac{2\pi}{n}jk} = \frac{1}{n} \sum_{j=0}^{n-1} e^{-i\frac{2\pi}{n}j(k-n/2)} \overset{c)}{=} \begin{cases} 1, & \text{für } k = \frac{n}{2} \\ 0, & \text{sonst} \end{cases}.$