Mittsemesterprüfung

Numerische Mathematik I

FS 2014

Prof. R. Hiptmair

Name		Note
Vorname		
Studiengang		
Leginummer		
Datum	1.4.2014	

1	2	3	Total
(8)	(8)	(9)	(25)

- **Nur Stifte und Legi auf dem Tisch!**
- Mobiltelefone, Tablets, etc. **ausgeschaltet** in der Tasche
- Aufgabenblätter sind in den braunen Kuverts.
  Diese bitte mit Ihrem Namen beschriften.
- **Braune Kuverts dürfen erst auf Anweisung geöffnet werden!**
- Bitte nicht mit Bleistift / Rot / Grün schreiben!
- Zugelassene Hilfsmittel: Keine.
- Prüfungsdauer: 20 Minuten.
- Bitte füllen Sie zuerst das Deckblatt aus.
- **Schreiben Sie Lösungen in die dafür vorgesehenen Felder.**
- **Nach Prüfungsende die ausgefüllten Aufgabenblätter bitte in das mit Ihrem Namen beschriftete braune Kuvert stecken.**

Viel Erfolg!

Mittsemesterprüfung

Numerische Mathematik I

D-MATH

FS 2014

Prof. R. Hiptmair

Die Lösungswege müssen,   **nachvollziehbar dargestellt sein.**

Problem 1 Fehlerabschätzung für Polynominterpolation [8 points]

Es sei für $a < b$ die Abbildung $P^{(n)}_{[a,b]} : C^0([a, b]) \to \mathbb{P}_n, n \geq 1$, der Polynominterpolationsoperator zu äquidistanten Knoten $x_j = a + j(b - a)/n$, $j = 0, \ldots, n$. In der Vorlesung wurde folgende Fehlerschranke für die Maximumnorm des Interpolationsfehlers hergeleitet

$$\left\| f - P^{(n)}_{[a,b]} f \right\|_{\infty,[a,b]} \leq \frac{1}{(n+1)!} \left\| f^{(n+1)} \right\|_{\infty,[a,b]} \|\omega_n\|_{\infty,[a,b]} , \quad \omega_n(t) := \prod_{j=0}^{n} (t - x_j) . \tag{1.1}$$

(1a) [4 points] Leiten Sie aus (1.1) eine möglichst scharfe Schranke (in Abhängigkeit von n) für $\left\| f - P^{(n)}_{[1,2]} f \right\|_{\infty,[1,2]}$ und $f = \log$ (natürlicher Logarithmus) ab.

HINWEIS: Sie können die Stirlingsche Formel verwenden:

$$\sqrt{2\pi}\, n^{n+1/2} e^{-n} \leq n! \leq e\, n^{n+1/2} e^{-n} \quad \text{für alle} \quad n \geq 1 . \tag{1.2}$$

Solution: Let $f(t) = \log(t)$ for $t \in [1, 2]$. For $n \geq 1$, the $(n + 1)$-th derivative of f reads

$$f^{(n+1)}(t) = (-1)^n \frac{n!}{t^{n+1}}.$$

Hence

$$\left\| f^{n+1} \right\|_{\infty,[1,2]} := \max_{t \in [1,2]} |f^{(n+1)}(t)| = \max_{t \in [1,2]} \left| (-1)^n \frac{n!}{t^{n+1}} \right| = n! \ .$$

Moreover, for equally spaced nodes $x_{j+1} = x_j + jh$ with $h = (b - a)/n$ and $j = 0, \ldots, n$, it holds

$$|\omega_n(t)| = \prod_{j=0}^{n} |t - x_j| \leq \frac{h^{n+1}}{4} n! = \frac{n!}{4n^{n+1}}$$

Using Stirling's formula (1.2) one has

$$\left\| f - P^{(n)}_{[1,2]} f \right\|_{\infty,[1,2]} \leq \frac{\left\| f^{(n+1)} \right\|_{\infty,[1,2]}}{(n+1)!} \|\omega_n\|_{\infty,[1,2]} \leq \frac{n!}{4(n+1)n^{n+1}} \leq \frac{n^{n+1/2} \mathrm{e}^{1-n}}{4(n+1)n^{n+1}} = \frac{\mathrm{e}}{4(n+1)n^{1/2}\mathrm{e}^n}$$

(1b) [4 points] Zeigen Sie, dass die Schranke aus (1.1) *skalierungsinvariant* ist, d.h., sie ändert sich nicht, wenn $[a, b]$ durch $[0, 1]$, die Knotenmenge $\mathcal{N}$ durch $T^{-1}(\mathcal{N})$, und f durch $f \circ T$ ersetzt wird, wobei T die affine Transformation $T(\xi) := a + (b - a)\xi$ bezeichne, $\xi \in [0, 1]$.

Solution: Let $x = T(t) \in [a, b]$ for $t \in [0, 1]$ and $t_j \in T^{-1}(\mathcal{N})$ for $j = 0, \ldots, n$. We first consider the following equalities

$$\begin{aligned}
\left\| (f \circ T)^{(n+1)} \right\|_{\infty,[0,1]} &= \left\| (b - a)^{n+1} \big(f^{(n+1)} \circ T\big) \right\|_{\infty,[0,1]} = \\
&= (b - a)^{n+1} \max_{t \in [0,1]} \left| \big(f^{(n+1)} \circ T\big)(t) \right| = \\
&= (b - a)^{n+1} \max_{x \in [a,b]} \left| f^{(n+1)}(x) \right| = (b - a)^{n+1} \left\| f^{(n+1)} \right\|_{\infty,[a,b]} ;
\end{aligned}$$

and

$$\begin{aligned}\|\omega_n\|_{\infty,[0,1]} &= \max_{t\in[0,1]}\left|\prod_{j=0}^{n}(t-t_j)\right| = \max_{t\in[0,1]}\left|\prod_{j=0}^{n}\left(t-\frac{j}{n}\right)\right| = \max_{x\in[a,b]}\left|\prod_{j=0}^{n}\left(\frac{x-a}{b-a}-\frac{j}{n}\right)\right| = \\ &= \max_{x\in[a,b]}\left|\prod_{j=0}^{n}\frac{1}{b-a}\left(x-a-\frac{j}{n}(b-a)\right)\right| = \\ &= \max_{x\in[a,b]}\left|\frac{1}{(b-a)^{n+1}}\prod_{j=0}^{n}(x-x_j)\right| = \frac{\|\omega_n\|_{\infty,[a,b]}}{(b-a)^{n+1}}.\end{aligned}$$

Therefore we have

$$\left\|(f\circ T)-P_{[0,1]}^{(n)}(f\circ T)\right\|_{\infty,[0,1]} \le \frac{1}{(n+1)!}(b-a)^{n+1}\left\|f^{(n+1)}\right\|_{\infty,[a,b]}\frac{\|\omega_n\|_{\infty,[a,b]}}{(b-a)^{n+1}}$$

which concludes the proof.

Problem 2 Lineares Gleichungssystem [8 points]

Für reguläres $\mathbf{A}\in\mathbb{R}^{n\times n}$ und $\mathbf{u},\mathbf{v}\in\mathbb{R}^n$ gilt

$$\left(\mathbf{A}+\mathbf{u}\mathbf{v}^\top\right)^{-1} = \mathbf{A}^{-1} - \frac{(\mathbf{A}^{-1}\mathbf{u})(\mathbf{v}^\top\mathbf{A}^{-1})}{1+\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u}}\ , \tag{2.1}$$

wenn $1+\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u}\neq 0$.

(2a) [2 points] Zeigen Sie, dass $\mathbf{A}+\mathbf{u}\mathbf{v}^\top$ singulär ist, wenn $\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u}=-1$.

HINWEIS: Bestimmen Sie einen nichttrivialen Vektor im Nullraum von $\mathbf{A}+\mathbf{u}\mathbf{v}^\top$.

Solution: We show that there exists $\mathbf{x}\neq 0$ such that $\mathbf{x}\in\ker\left(\mathbf{A}+\mathbf{u}\mathbf{v}^\top\right)$. Let $\mathbf{x}:=\mathbf{A}^{-1}\mathbf{u}$, since $\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u}=-1$, therefore $\mathbf{x}\neq 0$ and

$$\begin{aligned}\left(\mathbf{A}+\mathbf{u}\mathbf{v}^\top\right)\mathbf{x} &= \left(\mathbf{A}+\mathbf{u}\mathbf{v}^\top\right)\mathbf{A}^{-1}\mathbf{u} = \\ &= \mathbf{u}+\mathbf{u}\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u} = \\ &= \mathbf{u}-\mathbf{u} = 0.\end{aligned}$$

(2b) [4 points] Ihnen wird der MATLAB-Code aus Listing 2.1 vorgelegt und es wird Ihnen gesagt, dass er auf der Formel (2.1) basiert. Ferner, dass die MATLAB-Funktion `smvslv(d,b)` zwei Spaltenvektoren als Eingaben erwartet und ein lineare Gleichungssystem mit rechter Seite `b` löst.

Geben Sie die Koeffizientenmatrix dieses linearen Gleichungssystems an.

HINWEIS: `c=b./d` $\Longleftrightarrow$ `c(i)=b(i)/d(i)` for `i=1:length(d)`

Listing 2.1: Lösungsalgorithmus für spezielles lineares Gleichungsystem auf der Grundlage von (2.1)

```
function x = smvslv(d,b)
n = length(d);
```

```
if (length(b) ~= n), error('size mismatch'); end
w = 1./d;
alpha = 1 + sum(w);
if (abs(alpha) < eps*sum(abs(d)))
  warning('System close to singular');
end
x = b./d - w*(w'*b)/alpha;
```

Solution: Let $\mathbf{A} := \operatorname{diag}(d_1, \dots, d_n) \in \mathbb{R}^{n \times n}$ with $d_i \neq 0$ for all $i = 1, \dots, n$. Let $w_i = 1/d_i$ for all $i = 1, \dots, n$, it holds

$$\mathbf{w}^\top \mathbf{b} = \sum_{i=1}^{n} \frac{b_i}{d_i}.$$

Moreover, if $\mathbf{u} = (1, \dots, 1)^\top \in \mathbb{R}^n$, then $\mathbf{w} = \mathbf{A}^{-1}\mathbf{u}$ and the routine in Listing 2.1 computes

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} - \frac{\mathbf{w}(\mathbf{w}^\top \mathbf{b})}{1 + \mathbf{u}^\top \mathbf{w}} = \left(\mathbf{A}^{-1} - \frac{(\mathbf{A}^{-1}\mathbf{u})(\mathbf{A}^{-1}\mathbf{u})^\top}{1 + \mathbf{u}^\top \mathbf{A}^{-1}\mathbf{u}}\right)\mathbf{b}.$$

By (2.1), since $1 + \mathbf{u}^\top \mathbf{A}^{-1}\mathbf{u} = 1 + \sum_{i=1}^{n} 1/d_i \neq 0$, it can be concluded that $\mathbf{x} \in \mathbb{R}^n$ is the solution of the system

$$(\mathbf{A} + \mathbf{u}\mathbf{u}^\top)\mathbf{x} = \mathbf{b}.$$

(2c) [2 points] Was ist der asymptotische Rechenaufwand der Funktion `smvslv` aus Listing 2.1 bezüglich der Vektorlänge $n \to \infty$?

Solution: The computational effort is $\mathcal{O}(n)$. Indeed the routine requires only a fixed number of vector-vector operations.

Problem 3 Auslöschung [9 points]

Betrachten Sie die folgende MATLAB-Funktion

Listing 3.1: MATLAB-Implementierung der rationalen Funktion y

```
function y=ratfn(x)
if ((x<=-1) || (x>=1))
   error ('out of range');
end
y=1/(1+x)-1/(1-x);
```

(3a) [2 points] Was versteht man unter dem Phänomen der "Auslöschung" im Kontext numerischer Berechnungen?

Solution: Cancellation designates the strong *amplification of relative errors* when subtracting two numbers of about the same modulus and opposite sign.

Supplement: Cancellation itself does not introduce big relative errors.

Supplement: Inevitably roundoff will cause small relative errors and, thus, cancellation can lead to a severe impact of roundoff on the result of a numerical computation.

(3b) **[4 points]** Sei $\widetilde{y}(x)$ das Ergebnis von `ratfn` berechnet in MATLAB mit IEEE Gleitpunktarithmetik für ein zulässigen Argument x aus der entsprechenden Menge der Gleitpunktzahlen.

Sei $\delta(x)$ der relative Fehler von $\widetilde{y}$, also $\widetilde{y}(x) = y(x)(1+\delta(x))$. Geben Sie für $\delta(x)$ eine hinreichend scharfe Schranke an, in Abhängigkeit von der Rundungseinheit (Maschinengenauigkeit) $u \approx 10^{-16}$.

HINWEIS: Verwenden Sie folgende Abschätzung: Wenn $0 \leq u < 1$, $n < \frac{1}{u}$, und $|\delta_i| \leq u$, dann gilt

$$\prod_{i=1}^{n}(1+\delta_i)^{\pm 1} = 1 + \theta_n \quad \text{wobei} \quad |\theta_n| \leq \frac{nu}{1-nu} = \gamma_n(u)\ .$$

Solution: We analyze the propagation of round-off errors in the IEEE floating-point arithmetic by performing a round-off analysis on $\widetilde{y}$. Consider the error generated by operations in $\widetilde{y}$

$$\widetilde{y}(x) = \left(\frac{1+\delta_2}{(1+x)(1+\delta_1)} - \frac{1+\delta_4}{(1-x)(1+\delta_3)}\right)(1+\delta_5) \tag{3.1}$$

where $|\delta_i| \leq u$ for all i. Using the hint we have

$$\begin{aligned}
\widetilde{y}(x) &= \left(\frac{(1+\delta_2)(1+\delta_1)^{-1}(1+\delta_5)}{1+x} - \frac{(1+\delta_4)(1+\delta_3)^{-1}(1+\delta_5)}{1-x}\right) \\
&= \left(\frac{(1+\theta_3)}{1+x} - \frac{(1+\theta_3')}{1-x}\right) \\
&= y(x)\left(1 + \frac{-\theta_3 + x\theta_3 + x\theta_3' + \theta_3'}{2x}\right)
\end{aligned}$$

Therefore, we have

$$|\delta(x)| = \left|-\frac{\theta_3}{2x} + \frac{\theta_3}{2} + \frac{\theta_3'}{2} + \frac{\theta_3'}{2x}\right| \leq \frac{3u}{|x|(1-3u)} + \frac{3u}{(1-3u)},$$

where the last inequality is obtained by

$$|\theta_3| \leq \frac{3u}{1-3u} \qquad \text{and} \qquad |\theta_3'| \leq \frac{3u}{1-3u}\ .$$

(3c) **[3 points]** In Listing 3.2 sind drei verschiedene bei rundungsfehlerfreier Rechnung äquivalente MATLAB-Implementierungen der rationalen Funktion `ratfn` aus Listing 3.1 angegeben.

Figure 3.1 stellt die sich unter dem Einfluss von Rundungsfehlern für `f1`, `f2` und `f3` ergebenden relativen Fehler bei der Rechnung in IEEE-Gleitpunktarithmetik doppelter Genauigkeit für $x \in \{3^{-k}\}_{k=1,\ldots,30} \bigcup \{1-3^{-k}\}_{k=1,\ldots,30}$ dar.

Listing 3.2: Verschiedene algebraisch äquivalente Implementierungen von `ratfn`

```
f1 = @(x) ((1/(1.0+x))-(1/(1.0-x)));
f2 = @(x) (-2*x/(1.0-x^2));
f3 = @(x) (-2*x/((1.0-x)*(1.0+x)));
```

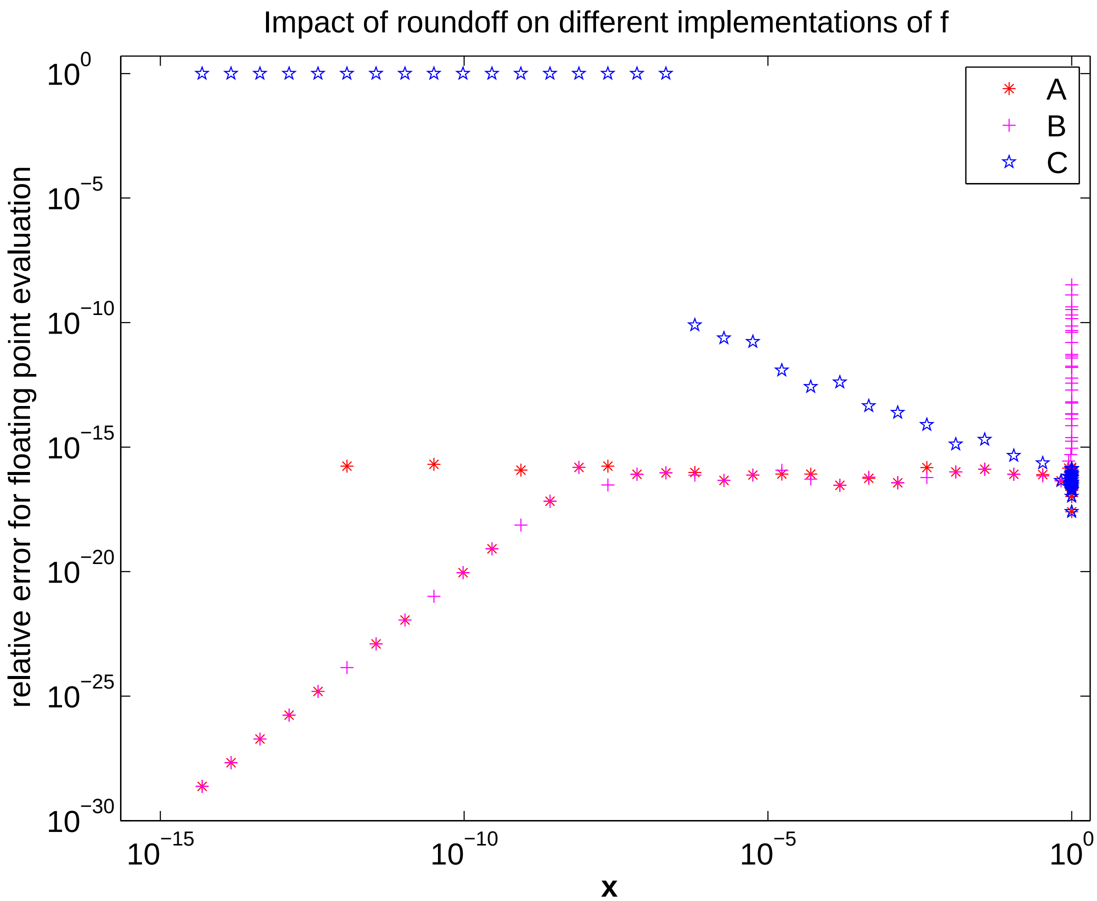

Abbildung 3.1: Relativer Fehler für verschiedene Implementierungen `f1`, `f2`, `f3` der rationalen Funktion y

Welcher Fehlergraph entspricht welcher Implemenentierung? Begründen Sie ihre Antwort kurz.

Solution: The blue one ($\star$) is $f1$ because of the cancellation for $x \approx 0$.
The red one ($*$) is $f3$ because no cancellation amplifies roundoff error.
The pink one ($+$) is $f2$ because of cancellation for $|x| \approx 1$.