Technische Universität Hamburg-Harburg
Institut für Numerische Simulation, E-10
Dr. Jens-Peter M. Zemke

Sommersemester 2008

Numerische Verfahren

Übungen und Lösungen, Blatt 1

Aufgabe 1: (Thema: relativer und absoluter Fehler)

a) Von ARCHIMEDES wurden die Zahlen $3\frac{1}{7}$ und $3\frac{10}{71}$ als obere und untere Schranken für π angegeben. Bestimmen Sie — nur unter Verwendung dieser Werte — Schranken für die *absoluten* und die *relativen* Fehler dieser Näherungen und ihres Mittelwertes.

Zur Rekapitulation:

Sei $\tilde{x} \in \mathbb{R}$ eine Näherung für $x \in \mathbb{R}$, dann ist mittels

$$|x - \tilde{x}|$$

der absolute und über

$$\frac{|x - \tilde{x}|}{|x|}$$

der relative Fehler definiert.

b) Sie wissen von einer Größe $x \in \mathbb{R}$ nach einer Messung, dass $x \in [1.005, 1.015]$ gilt. Nun benötigen Sie den Wert $y := x^2 - 1$, wobei ein maximaler relativer Fehler von 1% nicht überschritten werden darf. Zeigen Sie, dass es nicht garantiert werden kann, dass jedes $\tilde{x}$ aus $[1.005, 1.015]$ der geforderten Fehlergüte genügt. Welchen relativen Fehler muss ein $\tilde{x} \in [1.005, 1.015]$ erfüllen, so dass doch die Genauigkeitsanforderungen an die Auswertung von y erfüllbar sind?

Lösung zu Aufgabe 1:

zu a) Wir verwenden die eingängigen Bezeichungen π_u für die untere Schranke, π_o für die obere Schranke und $\pi_m = (\pi_u + \pi_o)/2$ für den Mittelwert der Schranken. Diese Werte sind im Einzelnen gegeben durch

$$\begin{aligned} \pi_u &:= 3\tfrac{10}{71} \approx \underline{3.14}0845, \\ \pi_o &:= 3\tfrac{1}{7} \approx \underline{3.14}2857, \\ \pi_m &:= 3\tfrac{141}{994} \approx \underline{3.141}851. \end{aligned}$$

Den *absoluten*, respektive *relativen* Fehler einer Näherung $\tilde{x}$ zu einem Wert x bezeichnen wir im folgenden mit $\epsilon_{\text{abs}}(\tilde{x})$, respektive $\epsilon_{\text{rel}}(\tilde{x})$:

$$\epsilon_{\text{abs}}(\tilde{x}) \equiv |\tilde{x} - x|, \qquad \epsilon_{\text{rel}}(\tilde{x}) \equiv \frac{|\tilde{x} - x|}{|x|} = \frac{\epsilon_{\text{abs}}(\tilde{x})}{|x|}.$$

Mit Hilfe dieser Notation schätzen wir zuerst den *absoluten* Fehler ab. Wir haben nur die Information, dass $\pi \in [\pi_u, \pi_o]$. Damit lässt sich leicht die Gültigkeit der folgenden Aussagen nachvollziehen:

$$\epsilon_{\text{abs}}(\pi_u) = |\pi_u - \pi| \leqslant |\pi_u - \pi_o| = \frac{1}{497} \approx 0.002012073,$$

$$\epsilon_{\text{abs}}(\pi_o) = |\pi_o - \pi| \leqslant |\pi_o - \pi_u| = \frac{1}{497} \approx 0.002012073.$$

Beide Abschätzungen sind in dem Sinne *scharf*, dass aufgrund mangelnder Information der „echte“ Wert von π ja ohne weiteres der jeweils *andere* Rand des Intervalles $[\pi_u, \pi_o]$ sein könnte. Die Abschätzung für den Mittelwert π_m sieht besser aus, es gilt

$$\epsilon_{\text{abs}}(\pi_m) \quad = \quad |\pi_m - \pi| \quad \leqslant \quad \frac{|\pi_o - \pi_u|}{2} \quad = \quad \frac{1}{994} \quad \approx \quad 0.001006037,$$

da der maximale Abstand von π_m zu irgendeinem anderen Punkt gerade durch den *Radius* des Intervalles $[\pi_u, \pi_o]$, also durch den halben Diameter (Durchmesser) $\pi_o - \pi_u$ gegeben ist. Auch diese Abschätzung ist *scharf*.

Jetzt wenden wir uns den relativen Fehlern zu. Der *relative* Fehler ist gegeben durch den *absoluten* Fehler geteilt durch den *Absolutbetrag* des exakten Wertes gegeben. Wir nutzen aus, dass wir schon Schranken für den absoluten Fehler berechnet haben, und verwenden folgende Beobachtung:

$$a \leqslant b, \quad c \geqslant d \qquad \Rightarrow \qquad \frac{a}{c} \leqslant \frac{b}{c} \leqslant \frac{b}{d}.$$

Wir benötigen also nur noch eine *untere* Schranke für den exakten Wert von π, die durch π_u explizit vorgegeben ist. Damit gilt für die relativen Fehler:

$$\begin{aligned}
\epsilon_{\text{rel}}(\pi_u) &\leqslant \frac{|\pi_u - \pi_o|}{|\pi_u|} \approx 0.000640615 \leqslant 0.65\% \\
\epsilon_{\text{rel}}(\pi_o) &\leqslant \frac{|\pi_u - \pi_o|}{|\pi_u|} \approx 0.000640615 \leqslant 0.65\% \\
\epsilon_{\text{rel}}(\pi_m) &\leqslant \frac{|\pi_u - \pi_o|}{2|\pi_u|} \approx 0.000320307 \leqslant 0.33\%
\end{aligned}$$

Beim Übergang zum Mittelwert gewinnt man also Genauigkeit in der Abschätzung. Dieser Gewinn ist allerdings erkauft durch den Verlust der Information, *auf welcher Seite* der exakte Wert liegt.

zu b) Sei $\tilde{x} \in [1.005, 1.015]$ beliebig, dann gilt mit $\tilde{y} := \tilde{x}^2 - 1$:

$$\frac{|y - \tilde{y}|}{|y|} \leqslant \frac{3.0225 \cdot 10^{-2} - 1.0025 \cdot 10^{-2}}{1.0025 \cdot 10^{-2}} \approx 200\%.$$

Die Abschätzung ist insofern scharf, als dass für den möglichen exakten Wert $x = 1.005$ und der Approximation $\tilde{x} = 1.015$ das Ungleichheitszeichen durch ein Gleichheitszeichen ersetzt werden kann. Quintessenz: Die geforderte Genauigkeit kann also bei weitem nicht garantiert werden.

Wir kommen jetzt zu der Frage, *wie genau* man die Messung durchführen müsste, damit die erforderliche Genauigkeit erzielt werden kann. Es gilt unter Beachtung von x, $\tilde{x} \in [1.005, 1.015]$:

$$\begin{aligned}
\epsilon_{\text{rel}}(\tilde{y}) &= \frac{|(\tilde{x}^2 - 1) - (x^2 - 1)|}{|x^2 - 1|} = \frac{|(\tilde{x} + x) \cdot (\tilde{x} - x)|}{|(x + 1) \cdot (x - 1)|} \\
&= \frac{(\tilde{x} + x)}{(x + 1) \cdot (x - 1)} \cdot |\tilde{x} - x| = \frac{(\tilde{x} + x) \cdot x}{(x + 1) \cdot (x - 1)} \cdot \epsilon_{\text{rel}}(\tilde{x}) \\
&\leqslant \frac{(1.015 + 1.015) \cdot 1.015}{(1.005 + 1) \cdot (1.005 - 1)} \cdot \epsilon_{\text{rel}}(\tilde{x}) \\
&\leqslant 205.5312 \cdot \epsilon_{\text{rel}}(\tilde{x})
\end{aligned}$$

Die Forderung, dass der letzte Ausdruck kleiner gleich 0.01 sein soll, führt auf die Schranke

$$\epsilon_{\text{rel}}(\tilde{x}) \leqslant \frac{0.01}{205.5312} \approx 4.865 \cdot 10^{-5}$$

für den relativen Fehler der Approximation $\tilde{x}$.

In dieser Aufgabe kann man sehr schön sehen, in welcher Weise die gegebene Genauigkeit für ein Funktionsargument die Güte der Genauigkeit des Funktionswertes beeinflussen kann.

Aufgabe 2: (Thema: Rundungsfehler)

Gegeben Sei das Polynom

$$p(x) \;=\; (x-2)^{11}\,.$$

Berechnen Sie mit Hilfe der *Intervallhalbierung* die Nullstelle 2. Gehen Sie dabei folgendermaßen vor: Programmieren Sie die Funktion zunächst in der *ausmultiplizierten Form*

```
x.^11-22*x.^10+220*x.^9-1320*x.^8+5280*x.^7-...
14784*x.^6+29568*x.^5-42240*x.^4+42240*x.^3-...
28160*x.^2+11264*x-2048
```

mit Hilfe des HORNER-Schemas. Wählen Sie als linken Startwert $\pi/2$ und als rechten Startwert π. Führen Sie die Intervallhalbierung durch, bis linker und rechter Intervallrand bis auf 10^{-5} zusammengerückt sind. (Hinweis: Über Sprachkonstrukte von Matlab können Sie sich allgemein mit `help lang` informieren. Es gibt so z.B. `while`-Schleifen.) Notieren Sie das Ergebnis. Probieren Sie nun verschiedene andere rechte Startwerte rechts von $x = 2$ aus, z.B. auch 2.2. Was beobachten Sie?

Was beobachten Sie dagegen, wenn Sie dasselbe mit dem Polynom in der Darstellung `(x-2).^11` machen? Versuchen Sie eine Erklärung.

Ein Hinweis zur Visualisierung: Plotten Sie beide Funktionen zunächst mittels `fplot` auf dem Intervall $[1.8, 2.2]$, dann mittels `plot` über dem Vektor x gegeben durch

```
x = linspace(1.8,2.2,8000);
```

Wählen Sie mit Hilfe der Funktion `axis` geeignete Ausschnitte und erklären Sie, was Sie dabei beobachten. Veranschaulichen Sie sich Ihre Beobachtung bei der Intervallhalbierung im Lichte dieses Resultats.

Lösung zu Aufgabe 2:

Das M-File `aufg01f01.m` enthält das Polynom in der ausmultiplizierten Form. Es wird mit einem gegebenen Vektor x wie folgt aufgerufen:

```
y = aufg01f01(x);
```

Nach dem Aufruf (mit den verschiedenen Startwerten) wird eine teilweise drastische Miskonvergenz sichtbar. Die Intervallhalbierung „sieht“ Nullstellen, wo keine sein dürfen, während bei dem Polynom in der Form `(x-2).^11` so etwas nicht passiert. Der Grund dafür sind Auslöschungsphänomene bei der Auswertung des Polynoms nach dem HORNER-Schema in der Nähe der Nullstelle, also Rundungsfehler. Interessant dabei ist, dass für *verschiedene* rechte Ränder die Intervallhalbierung *verschiedene* Rundungsfehler und damit auch *verschiedene* „Nullstellen“ produziert.

Man kann aber zeigen, dass jede dieser Pseudo-Nullstellen die Nullstelle eines (anderen) Polynoms ist, das man erhält, indem jeder Koeffizient leicht gestört wird. Wir lernen aus dieser Aufgabe etwas über die Kondition mehrfacher Nullstellen von Polynomen und auch über die sinnhafte Implementation algebraischer Ausdrücke. So verlieren wir in diesem Fall an Genauigkeit, wenn wir das Polynom erst auf Standardform bringen.

Das M-File `aufg01f02.m` führt die Intervallhalbierung durch und plottet anschließend die Konvergenzhistorie der unteren und oberen Intervallränder. (Man kann entsprechende Zeilen im Code „scharfmachen“ und bekommt dann zusätzlich eine Historie der y-Werte.) Das M-File wird im Fall der Ränder $\pi/2$ und π folgendermaßen aufgerufen:

```
[nst,val,xa,ya] = aufg01f02(pi/2,pi,1e-5);
```

Die Aufrufparameter im Einzelnen aufgeschlüsselt sind gegeben als:

Erster Parameter $\approx$ linker Rand,

Zweiter Parameter $\approx$ rechter Rand,

Dritter Parameter $\approx$ Toleranz.

Die Rückgabeparameter im Einzelnen aufgeschlüsselt sind gegeben als:

Erster Parameter $\approx$ Nullstellenschätzung,

Zweiter Parameter $\approx$ y-Wert dazu,

Dritter Parameter $\approx$ x-Historie,

Vierter Parameter $\approx$ y-Historie.

Man kann das M-File auch ohne Rückgabeparameter aufrufen und erhält nur den Plot:

```
aufg01f02(pi/2,pi,1e-5);
```

Wenn man die ausmultiplizierte Funktion mit `fplot` (nur in Matlab, nicht in Octave) zeichnet,

```
fplot('aufg01f01(x)',[1.8 2.2]);
```

erkennt man bei einiger Vergrößerung, etwa

```
axis([1.8 2.2 -1e-9 1e-9]);
```

ein erratisches Zickzackmuster. Gibt man den x-Vektor noch feiner explizit vor, nämlich etwa wie vorgeschlagen

```
x = linspace(1.8,2.2,8000);
```

und zeichnet dann mit `plot`, i.e.,

```
plot(x,aufg01f01(x));
```

so erkennt man bei obiger `axis`-Wahl einen „ausgeschmierten“ Bereich. Es werden scheinbar alle Werte in einem bestimmten Streifen um die Null in einer gewissen Nähe von $x = 2$ angenommen. Mit

```
axis([1.8 2.2 -1e-11 1e-11]);
```

wird das noch deutlicher. Das Intervallhalbierungsverfahren erwischt also quasi „zufällig“ positive und negative Funktionswerte und steuert damit *eine von unabsehbar vielen* „falschen“ Nullstellen in der Nähe der 2 an.

Aufgabe 3: (Thema: Polynominterpolation)

a) Es sei $M = M_1, \ldots, M_n$ Ihre Matrikelnummer. Nehmen Sie von dieser die letzten fünf Ziffern und bilden Sie mit diesen die Zahl $m = m_1, \ldots, m_5$ mit $m_i = M_{n-5+i}$, $i = 1, \ldots, 5$. Bestimmen Sie mit Hilfe von Matlab, aber ohne den Gebrauch der Funktionen `polyfit` oder `polyval` ein Polynom p, welches die Zahlen $i \in \{1, 2, 3, 4, 5\}$ auf die $i$te Ziffer m_i der Zahl m abbildet. Plotten Sie p auf dem Intervall $[1, 5]$.

b) Es sei nun

$$\tilde{m}_3 \;:=\; m_3 + \frac{1}{100}$$

Bestimmen Sie ein Polynom $\tilde{p}$, welches $i \in \{1, 2, 4, 5\}$ auf die $i$te Ziffer von m und die 3 auf $\tilde{m}_3$ abbildet. Plotten Sie die Fehlerfunktion $f := p - \tilde{p}$ auf dem Intervall $[0, 50]$.

c) Zeigen Sie, dass die Fehlerfunktion f nicht von Ihrer Matrikelnummer abhängt, dass f genau vier Nullstellen hat und berechnen Sie den Funktionsterm explizit.

Lösung zu Aufgabe 3:

zu a) In der Materialsammlung finden Sie die Programme `aufg01f03.m` und `aufg01f04.m`. Das Programm `aufg01f03.m` berechnet das Polynom mit Hilfe einer VANDERMONDE-Matrix. Später werden wir sehen, dass dieses Vorgehen numerisch nicht besonders günstig ist. In `aufg01f04.m` wurde die Berechnung nach Algorithmus 2.14 und 2.15 aus dem Skript (Dividierte Differenzen und HORNER-Schema) realisiert.

zu b) In diesem Aufgabenteil sollte Ihnen vor allem auffallen, dass eine einzige sehr kleine Störung der Anfangsdaten großen Einfluss auf das resultierende Interpolationspolynom hat. Die Fehlerfunktion f hängt aber nicht von den *Daten*, sondern *nur* von der *Störung* ab. Eine mögliche Realisierung finden Sie in der Materialsammlung unter `aufg01f05.m`.

zu c) Wir stellen sowohl p wie $\tilde{p}$ mit der LAGRANGEschen Interpolationsformel (Definition 2.5 im Skript) dar. Dann ist

$$f(x) = p(x) - \tilde{p}(x) = \sum_{j=1}^{5} m_j l_j(x) - \sum_{j=1}^{5} \tilde{m}_j \tilde{l}_j(x).$$

Da die Polynome $l_j(x)$ nur von den Knoten x_i abhängen, gilt $l_j = \tilde{l}_j$. Wegen $m_i = \tilde{m}_i$ für $i \in \{1, 2, 4, 5\}$ bzw. $m_3 = \tilde{m}_3 - 1/100$ gilt

$$f(x) = \sum_{1,2,4,5} m_j l_j(x) - \sum_{1,2,4,5} m_j l_j(x) + m_3 l_3(x) - \tilde{m}_3 l_3(x) = \frac{1}{100} l_3(x).$$

Die Fehlerfunktion hängt also nicht von dem gewählten Satz m_i (die fünf letzten Ziffern Ihrer Matrikelnummer) ab. Da p und $\tilde{p}$ in den Knoten $1, 2, 4$ und 5 übereinstimmen, hat f mindestens vier Nullstellen. Wurden p und $\tilde{p}$ mit minimalen Grad $d \leqslant 4$ bestimmt, dann ist auch f ein Polynom vom Grade höchstens gleich vier. Aus dem Fundamentalsatz der Algebra folgt, dass keine weiter Nullstellen existieren können[1].

Wir kennen also die vier einzigen Nullstellen und können daher explizit schreiben

$$f(x) = \alpha(x-1)(x-2)(x-4)(x-5).$$

Aus $f(3) = \frac{1}{100}$ folgt dann $\alpha = \frac{1}{400}$.

[1] Alternativ könnte man auch mit Satz 2.4 aus dem Skript argumentieren. Gäbe es eine weitere Nullstelle, dann hätten p und $\tilde{p}$ fünf gemeinsame Punkte und müssten daher nach Satz 2.4 identisch sein. Das widerspräche aber $p(3) - \tilde{p}(3) = 1/100$.