Oefeningen Numerieke Wiskunde

Oefenzitting 2: Bewegende kommavoorstelling en foutenanalyse

1 Bewegende kommavoorstelling

Probleem 1. De IEEE standaard voorziet voor basis $b = 2$ in enkelvoudige-precisiegetallen en dubbele-precisiegetallen (zie boek H2 §5.5). Deze hebben, respectievelijk, een 32-bit en een 64-bit voorstelling. Hoeveel bits worden hiervan gebruikt voor de mantisse?

Probleem 2. Hoeveel dubbele-precisiegetallen kunnen er worden voorgesteld

- tussen de getallen 1 en 2?
- tussen de getallen 7 en 9?

Probleem 3. Hoe ziet de rij $x_1, x_2, \ldots$, met

$$x_1 = 1; \quad x_{n+1} = fl(x_n + 1),\ n = 1, 2, \ldots,$$

er voor grote waarden van n uit, op een machine met 3 decimale cijfers voor de mantisse?

Probleem 4. Analyseer de volgende algoritmen voor het bepalen van de basis, b, en het aantal cijfers in de mantisse, p:

bepaal_b <uit: b >
1. $A \leftarrow 1$
2. while $(A + 1) - A = 1$
   2.1. $A \leftarrow 2 * A$
3. $i \leftarrow 1$
4. while $(A + i) = A$
   4.1. $i \leftarrow i + 1$
5. $b \leftarrow (A + i) - A$

bepaal_p <in: b; uit: p >
1. $p \leftarrow 1$
2. $z \leftarrow b$
3. while $(z + 1) - z = 1$
   3.1. $p \leftarrow p + 1$
   3.2. $z \leftarrow z * b$

- Ga de werking van de algoritmen na voor $b = 10$ en $p = 3$.
- (extra opgave) Toon formeel de werking van de algoritmen aan voor om het even welke b en p.
  (Hint: op regel 2.1 van **bepaal_b** zou er ook $A \leftarrow A+1$ mogen staan. Waarom staat er wat er nu staat?)

2 Foutenanalyse

Het doel van een formele foutenanalyse is het vinden van een a priori bovengrens op het effect van de afrondingsfouten en hun voortplanting doorheen een berekening $y = f(x)$. We zullen dit effect meten met de relatieve fout op het eindresultaat.

Veronderstel dat de afrondingsfouten die gemaakt worden voldoen aan $\text{fl}(x) = x(1+\epsilon)$ met $|\epsilon| \leq \epsilon_{\text{mach}}$ en dat iedere elementaire bewerking exact wordt uitgevoerd en dan wordt afgerond. D.w.z.,

$$\begin{aligned} \text{fl}(a \circ b) &= (a \circ b)(1+\epsilon), \\ \text{fl}(\Box a) &= (\Box a)(1+\epsilon), \end{aligned}$$

waarbij $\circ$ een van de bewerkingen $+, -, *$ of $\div$ voorstelt en $\Box$ een elementaire functie zoals $\sqrt{}, \sin, \exp, \ldots$. Door het berekenen van een eerste orde benadering van de fout, bekom je uiteindelijk iets van de vorm

$$\left| \frac{\bar{y} - y}{y} \right| \leq E(x)\, \epsilon_{mach}\,,$$

waarbij $E(x)$ een functie is die enkel afhangt van x.

Bij het doen van een formele foutenanalyse kan je twee methodes volgen om een eerste orde benadering te bekomen van de fout. Eerst bekijken we een methode waarbij je de eerste orde benadering in één keer berekent via een Taylor benadering in meerdere veranderlijken. In Sectie 3 wordt een alternatieve methode voorgesteld die je zelf thuis kan bekijken.

2.1 Voorbeeld

Als voorbeeld voeren we een foutenanalyse uit voor de berekening van

$$y = \sqrt{1+x} - 1,$$

waarbij men de berekeningen uitvoert volgens de uitdrukking hierboven. Men veronderstelt dat x exact voorgesteld kan worden op de machine. Op die manier worden enkel de afrondingsfouten in rekening gebracht die gemaakt worden bij de berekeningen.

De berekende waarde voor y is dan

$$\bar{y} = fl\left(fl(\sqrt{fl(1+x)}) - 1\right) = \left((\sqrt{(1+x)(1+\epsilon_1)})(1+\epsilon_2) - 1\right)(1+\epsilon_3)$$

Hierbij stellen $\epsilon_i,\ i = 1, 2, 3$, de relatieve fouten voor die gemaakt worden bij afronding na iedere bewerking. We weten dat

$$|\epsilon_i| \leq \epsilon_{mach},\ i = 1, 2, 3.$$

We interpreteren, in de uitdrukking voor $\bar{y}$, x als een parameter en $\epsilon_i,\ i = 1, 2, 3$, als variabelen,

$$\bar{y} = F(\epsilon_1, \epsilon_2, \epsilon_3).$$

Vermits de ϵ_i zeer klein zijn, laat $F(\epsilon_1, \epsilon_2, \epsilon_3)$ zich goed benaderen door de Taylorontwikkeling rond $(0,0,0)$ afgebroken na de lineaire termen

$$\bar{y} = F(\epsilon_1, \epsilon_2, \epsilon_3) \approx F(0,0,0) + \epsilon_1 \frac{\partial F}{\partial \epsilon_1}(0,0,0) + \epsilon_2 \frac{\partial F}{\partial \epsilon_2}(0,0,0) + \epsilon_3 \frac{\partial F}{\partial \epsilon_3}(0,0,0).$$

De constante term $F(0,0,0)$ in de lineaire benadering is de exacte waarde y. Voor de berekening van de partiële afgeleiden gaan we als volgt te werk

$$\begin{aligned} F(\epsilon_1, 0, 0) &= \sqrt{(1+x)(1+\epsilon_1)} - 1 = \sqrt{1+x}(1+\epsilon_1)^{\frac{1}{2}} - 1 \\ \frac{\partial F}{\partial \epsilon_1}(\epsilon_1, 0, 0) &= \sqrt{1+x}\frac{1}{2}(1+\epsilon_1)^{-\frac{1}{2}} \\ \frac{\partial F}{\partial \epsilon_1}(0,0,0) &= \frac{1}{2}\sqrt{1+x} \end{aligned}$$

Na uitwerking van de andere partiële afgeleiden vinden we de lineaire benadering voor $\bar{y}$ en een benaderende formule voor de absolute en de relatieve fout

$$\begin{aligned} \bar{y} &\approx y + \frac{1}{2}\sqrt{1+x}\,\epsilon_1 + \sqrt{1+x}\,\epsilon_2 + y\epsilon_3 \\ \bar{y} - y &\approx \frac{1}{2}\sqrt{1+x}\,\epsilon_1 + \sqrt{1+x}\,\epsilon_2 + y\epsilon_3 \\ \frac{\bar{y}-y}{y} &\approx \frac{\sqrt{1+x}}{2(\sqrt{1+x}-1)}\epsilon_1 + \frac{\sqrt{1+x}}{\sqrt{1+x}-1}\epsilon_2 + \epsilon_3. \end{aligned}$$

In de limiet voor $x \to 0$ worden de coëfficiënten vóór ϵ_1 en ϵ_2 oneindig groot. Dit betekent dat voor kleine x de relatieve fout op de berekende y zeer groot kan zijn. Voor grote x daarentegen is het berekende resultaat nauwkeurig.

Een bovengrens voor de relatieve fout vinden we als volgt

$$\begin{aligned} \left|\frac{\bar{y}-y}{y}\right| &\approx \left|\frac{\sqrt{1+x}}{2(\sqrt{1+x}-1)}\epsilon_1 + \frac{\sqrt{1+x}}{\sqrt{1+x}-1}\epsilon_2 + \epsilon_3\right| \\ &\leq \left|\frac{\sqrt{1+x}}{2(\sqrt{1+x}-1)}\epsilon_1\right| + \left|\frac{\sqrt{1+x}}{\sqrt{1+x}-1}\epsilon_2\right| + |\epsilon_3| \\ &= \left|\frac{\sqrt{1+x}}{2(\sqrt{1+x}-1)}\right||\epsilon_1| + \left|\frac{\sqrt{1+x}}{\sqrt{1+x}-1}\right||\epsilon_2| + |\epsilon_3| \\ &\leq \epsilon_{mach}\left(\left|\frac{\sqrt{1+x}}{2(\sqrt{1+x}-1)}\right| + \left|\frac{\sqrt{1+x}}{\sqrt{1+x}-1}\right| + 1\right) \\ &= \epsilon_{mach}\left(\left|\frac{3\sqrt{1+x}}{2(\sqrt{1+x}-1)}\right| + 1\right) \end{aligned}$$

2.2 Oefeningen

Probleem 5. Voer de foutenanalyse uit voor de berekening van y van het voorbeeld indien de berekeningen verlopen volgens de wiskundig equivalente formule

$$y = \frac{x}{\sqrt{x+1}+1}.$$

Verklaar dat dit rekenschema nauwkeuriger is dan het vorige wanneer x klein is.

Probleem 6. Voer de foutenanalyse uit voor de berekening van

$$y = \frac{1-\cos(x)}{x^2}.$$

Naar welk getal convergeert y voor heel kleine x? Is de formule dan nog nauwkeurig?

Probleem 7. Beschouw volgend algoritme voor het berekenen van de som van n getallen:

som <in: $a_1, \ldots, a_n$; uit: $S = \sum_{i=1}^{n} a_i$ >
1. $S \leftarrow a_1$
2. for $i = 2 : 1 : n$
   2.1. $S \leftarrow S + a_i$

Toon aan dat de absolute fout van dit algoritme gelijk is aan

$$\bar{S} - S \approx \epsilon_2(a_1 + a_2) + \epsilon_3(a_1 + a_2 + a_3) + \ldots + \epsilon_n(a_1 + a_2 + \ldots + a_n)$$

of

$$\bar{S} - S \approx \sum_{k=2}^{n} \epsilon_k \, b_k, \quad \text{met} \quad b_k = \sum_{i=1}^{k} a_i.$$

(Zie ook slides en handboek.)

3 Alternatieve methode en extra oefeningen

3.1 Alternatieve methode

Bij deze methode ga je waar nodig de stelling van Taylor in één veranderlijke toepassen om zo in een aantal stappen een eerste orde benadering te bekomen.

We illustreren de methode op hetzelfde voorbeeld: de berekening van

$$y = \sqrt{1+x} - 1.$$

De berekende waarde voor y is

$$\bar{y} = fl\left(fl(\sqrt{fl(1+x)}) - 1\right) = \left((\sqrt{(1+x)(1+\epsilon_1)})(1+\epsilon_2) - 1\right)(1+\epsilon_3) \qquad (1)$$

met $|\epsilon_i| \leq \epsilon_{mach},\ i = 1, 2, 3.$

Ditmaal herschrijven we $\bar{y}$ als $\bar{y} \approx y(1+\delta)$ door in (1) gebruik te maken van de Taylor reeks rond nul

$$f(\epsilon) = f(0) + f'(0)\epsilon + f''(0)\frac{\epsilon^2}{2} + \dots$$

en hogere orde termen te verwaarlozen. Bv.

$$\sqrt{1+\epsilon} \approx 1 + \frac{1}{2}\epsilon, \quad \text{en} \quad (1+\epsilon_1)(1+\epsilon_2) \approx 1 + \epsilon_1 + \epsilon_2,$$

waarbij we veronderstellen dat ϵ, ϵ_1 en ϵ_2 klein genoeg zijn. We krijgen dus

$$\begin{aligned}
\bar{y} &= \left((\sqrt{(1+x)(1+\epsilon_1)})(1+\epsilon_2) - 1 \right)(1+\epsilon_3) \\
&= \left(\sqrt{1+x}\,\sqrt{1+\epsilon_1}\,(1+\epsilon_2) - 1 \right)(1+\epsilon_3) \\
&\approx \left(\sqrt{1+x}\,(1+\frac{1}{2}\epsilon_1)\,(1+\epsilon_2) - 1 \right)(1+\epsilon_3) \\
&\approx \left((\sqrt{1+x})(1+\frac{1}{2}\epsilon_1 + \epsilon_2) - 1 \right)(1+\epsilon_3) \\
&= \left(y + (\sqrt{1+x})(\frac{1}{2}\epsilon_1 + \epsilon_2) \right)(1+\epsilon_3) \\
\implies \bar{y} - y &\approx \frac{1}{2}\sqrt{1+x}\,\epsilon_1 + \sqrt{1+x}\,\epsilon_2 + y\epsilon_3 \\
\frac{\bar{y}-y}{y} &\approx \frac{\sqrt{1+x}}{2(\sqrt{1+x}-1)}\epsilon_1 + \frac{\sqrt{1+x}}{\sqrt{1+x}-1}\epsilon_2 + \epsilon_3.
\end{aligned}$$

waarbij we steeds de tweede orde termen verwaarloosd hebben. Dit is uiteraard hetzelfde resultaat als voor de eerste methode.

3.2 Extra oefeningen

Probleem 8. Voer een foutenanalyse uit voor de volgende berekeningen. Waar verwacht je numerieke moeilijkheden?

(a) $y = x\sin(x)$

(b) $y = \dfrac{1-\cos(x)}{\sin(x)}$

(c) $y = \dfrac{1-e^{-2x}}{x}$

(d) $y = (1+x)^{\frac{1}{x}}$

(e) $y = \sqrt{e^x - 1}$

(f) $y = \sin(\frac{1}{x})$

(g) $y = \left(1 + x^2\right)^{x^2}$

(h) $y = \dfrac{e^{x^2} - e^{-x^2}}{2x^2}$

Probleem 9. Beschouw volgende algoritmen voor het berekenen van product en scalair product:

- **product** <in: $a_1, \ldots, a_n$; uit: $P = \Pi_{i=1}^n a_i >$
  1. $P \leftarrow a_1$
  2. voor $i = 2 : 1 : n$
     2.1 $P \leftarrow P * a_i$

- **scalair_product** <in: $a_1, \ldots, a_n, b_1, \ldots, b_n$; uit: $SP = \sum_{i=1}^n a_i b_i >$
  1. $SP \leftarrow a_1 b_1$
  2. voor $i = 2 : 1 : n$
  2.1 $SP \leftarrow SP + a_i * b_i$

Maak een afschatting van de absolute fout m.b.v. een foutenanalyse.

Probleem 10. Analyseer het volgende recursieve algoritme:
SOM <in: $a_1, \ldots, a_n$; uit: $S = \sum_{i=1}^n a_i >$
1. als $n = 1$
   1.1 $S \leftarrow a_1$
   1.2 anders
   $S \leftarrow S(a_1, \ldots, a_{n\text{div}2}) + S(a_{n\text{div}2+1}, \ldots, a_n)$

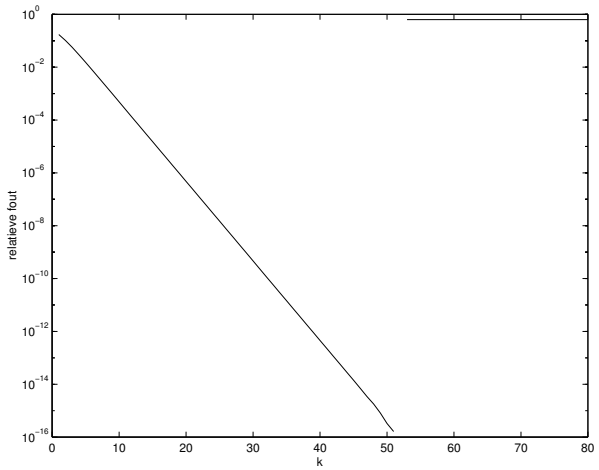
De bewerking 'div' levert het quotient van de gehele deling, dus $(2n+1)\text{div}2 = (2n)\text{div}2 = n,\ n \in \mathbb{N}$.
Toon aan dat in eindige precisie, de fout $S - \bar{S}$ voldoet aan:

$$n \leq 2^k,\ k \in \mathbb{N} \rightarrow |S - \bar{S}| \leq k\epsilon_{\text{mach}} \left(|a_1| + |a_2| + \cdots + |a_n|\right).$$

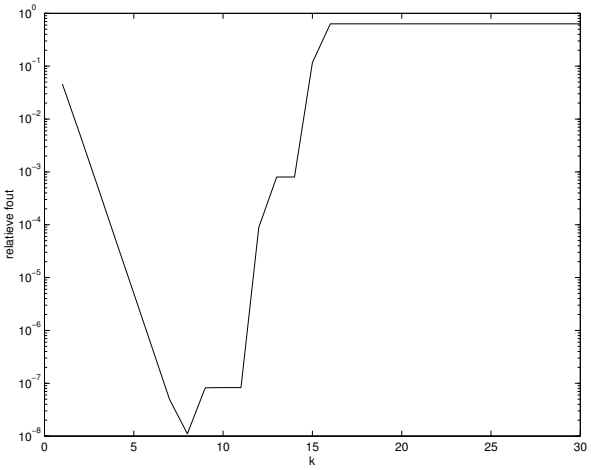
Een recursief algoritme leent zich tot een bewijs door ...

Probleem 11. (Examenvraag)
We weten dat $\lim_{x\to\infty}(1 + 1/x)^x = e^1$. We gaan de waarde e benaderen door $\tilde{e}_k = (1 + 1/x_k)^{x_k}$ uit te rekenen voor

- $x_k = 2^k$ en
- $x_k = 10^k$.



(a) Relatieve fout



(b) Relatieve fout

In de figuren hieronder geven we de relatieve fout weer, d.w.z. $\left|\frac{\tilde{e}_k - e}{e}\right|$. De eerste figuur geeft de relatieve fout voor de machten van 2 terwijl de tweede figuur de fouten voor machten van 10 weergeeft.

Waarom zijn de twee grafieken zo verschillend?
Kan men hieruit afleiden met welke basis en met hoeveel beduidende cijfers de computer werkt?
Verklaar in detail je antwoord.