

**Universiteit Utrecht**



*Department  
of Mathematics*

## NUMERIEKE WISKUNDE, 1-ste deel

### Inleiding in de Numerieke Analyse

door

Numerieke Wiskunde Groep  
Departement Wiskunde  
Universiteit Utrecht

November 2011



## Voorwoord

**Numeriek Wiskunde.** Een numeriek wiskundige ontwerpt en analyseert *reken-schema's (algoritmen)* voor het *numeriek* oplossen van (gewoonlijk) rekenintensieve rekenproblemen. Hij let daarbij op

- *efficiëntie* (met het rekenschema produceert een computer snel de oplossing)
- *nauwkeurigheid* (de oplossing is nauwkeurig binnen van te voren opgegeven foutgrenzen)
- *betrouwbaarheid* (er is een zekere garantie dat de gewenste nauwkeurigheid ook daadwerkelijk gehaald wordt)
- *robuustheid* (de performance (efficiëntie, nauwkeurigheid, betrouwbaarheid) wordt ook gehaald voor naburige problemen)

De rekenschema's moeten geïmplementeerd worden voor verwerking op een computer (dit is het werkterrein van de informaticus) en ze kunnen voor zekere toepassingen geoptimaliseerd worden (het werkterrein van de computational scientist). De grenzen tussen Numerieke Wiskunde en Informatica en Computational Science is vaag: de efficiëntie van een rekenschema hangt af van de architectuur van de computer waarop het schema moet “draaien” en wat efficiënt is en nauwkeurig hangt af van de toepassing.

De rekenschema's zijn numeriek (en niet symbolisch), d.w.z., de geproduceerde oplossingen bestaan uit getallen. Voor analyse (t.b.v. inzicht, voorspellingen, beleid, ontwerp, etc.) worden de oplossingen gewoonlijk gevisualiseerd middels grafieken of animaties. In het weerpraatje in het 8 uur journaal zijn de beelden van het weer van afgelopen dag foto's (satellietbeelden), maar de beelden (met pijlen die de windrichting aangeven, lijnen van gelijke luchtdruk, bewegende wolkenpartijen) die het weer voorspellen zijn animaties verkregen met rekenschema's als boven bedoeld.

**De cursus.** Het doel van deze cursus is tweeledig. (i) Voor eenvoudige problemen—de functies in deze cursus zijn reëelwaardig en gedefiniëerd op een reëel interval—willen we laten zien wat efficiëntie, nauwkeurigheid, betrouwbaarheid en robuustheid is en welke overwegingen een rol spelen bij het afleiden van schema's met dergelijke eigenschappen. (ii) Verder zullen we benaderingstrategieën bekijken die vaak de basis vormen voor rekenschema's voor ingewikkeldere problemen.

Bij tal van mathematische problemen is de exacte oplossing niet in handzame vorm te vinden. Zo is  $\int_a^b f(x) dx$  maar voor relatief weinig functies  $f$  als getalwaarde te bepalen, zijn differentiaalvergelijkingen nog zeldzamer in eindige termen van elementaire functies oplosbaar en zijn wortels van vergelijkingen  $f(x) = 0$  slechts zelden te bepalen. En de oplossingen van deze problemen worden zelfs ongedefiniëerd als men van de optredende functies slechts in een aantal punten de waarden kent, terwijl dit bijv. bij de experimentele wetenschappen toch een veel voorkomende situatie is. In dit college, en in vervolg colleges, zullen wij laten zien hoe men de genoemde problemen benaderd kan oplossen door de optredende functies te benaderen door functies uit zinnige functieklassen waarvoor de gestelde problemen wèl oplosbaar zijn.

Belangrijke vragen die hierbij een rol zullen spelen zijn:

- Wat is (een bovengrens voor) de fout in de numerieke benadering;
- Wat zijn de kosten verbonden aan het berekenen van deze benadering;

- Wat is het effect op het eindresultaat van fouten in de ingangsgegevens en van fouten die gemaakt worden door het rekenen in eindige precisie.

**Het dictaat.** Dit dictaat is een bewerking van een collegedictaat geschreven door A. van der Sluis en M. de Gee. Verschillende leden van de numerieke wiskunde groep, i.h.b. G. Sleijpen en R. Stevenson, hebben daaraan bijgedragen.

## Inhoudsopgave

<b>Voorwoord</b>	<b>i</b>
<b>1 Interpolatie</b>	<b>2</b>
1.0 Inleiding . . . . .	2
1.1 Lineaire interpolatie . . . . .	2
1.2 Lagrange interpolatie . . . . .	5
1.3 Interpolatie met functiewaarden en afgeleiden . . . . .	11
1.4 Inverse (lineaire) interpolatie . . . . .	13
<b>2 Numerieke differentiatie</b>	<b>15</b>
2.1 Eenvoudige formules. Effect van afrondfouten . . . . .	15
2.2 Nauwkeuriger formules . . . . .	17
2.3 Experimenteel gevonden functiewaarden . . . . .	18
<b>3 Foutschattingen bij numerieke processen</b>	<b>20</b>
3.1 Fouten . . . . .	20
3.2 Afrondfouten . . . . .	22
3.3 Foutvoortplanting . . . . .	26
3.4 Representatiefouten . . . . .	27
3.5 Extrapolatieprocessen . . . . .	29
<b>4 Numerieke integratie</b>	<b>35</b>
4.0 Inleiding . . . . .	35
4.1 De trapeziumregel . . . . .	35
4.2 Automatische integratie en Romberg schema's . . . . .	39
4.3 Andere kwadratuurformules . . . . .	42
4.4 Convergentie van kwadratuurschema's . . . . .	49
<b>A Interpolatie met afgeleiden</b>	<b>51</b>
<b>B Rombergschema's</b>	<b>53</b>
<b>C De Euler Mac-Laurin “reeks”</b>	<b>55</b>
<b>Vraagstukken</b>	<b>57</b>
1 Interpolatie . . . . .	58
2 Numerieke differentiatie . . . . .	63
3 Foutschattingen bij numerieke processen . . . . .	65
4 Numerieke integratie . . . . .	69
<b>Index</b>	<b>74</b>

# 1 Interpolatie

## 1.0 Inleiding

Een basis element in de numerieke wiskunde is het benaderen van een functie  $f$  door een andere functie  $\tilde{f}$  uit een geschikte functieklassse. Wanneer bijvoorbeeld  $f$  slechts bekend is in een aantal punten, kan dan deze  $\tilde{f}$ , die ook in tussengelegen punten gedefiniëerd is, gebruikt worden om de waarde van  $f$  in zo'n tussenpunt te benaderen of om de extremen, de integraal of de afgeleide van  $f$  te benaderen. Ook wanneer  $f$  in principe in ieder punt beschikbaar is, maar moeilijk te evalueren valt, kan  $\tilde{f}$  hier voor gebruikt worden. Vaak zijn de functies  $f$  oplossingen van differentiaalvergelijkingen en kunnen de functiewaarden alleen (en ook nog maar alleen bij benadering) berekend worden in zekere rekenpunten. Bij bijvoorbeeld de weersvoorspellingen berekent men de temperatuur alleen om de 15 minuten. Of zijn functiewaarden alleen bekend in zekere meetpunten (de temperatuur van gisteren is alleen gemeten in een aantal meetstations).

In dit hoofdstuk benaderen we  $f$  door een interpolatiepolynoom, d.w.z. een polynoom dat met  $f$  overeenstemt in een aantal (steun)punten.

## 1.1 Lineaire interpolatie

**1.1A** Zij  $a < b$  twee punten in  $\mathbb{R}$  en  $f$  een functie die gedefiniëerd is op  $[a, b]$ . Laat verder  $c$  en  $d$  twee verschillende punten zijn in  $[a, b]$ . Het *lineaire interpolatiepolynoom* van  $f$  op de *steunpunten*  $c$  en  $d$  is dan gedefiniëerd als het eerste graads polynoom  $p$  dat in  $c$  en  $d$  dezelfde waarde aanneemt als  $f$ .

Men gaat snel na dat het bedoelde polynoom gegeven wordt door

$$p(x) = f(c) + \frac{x-c}{d-c}(f(d) - f(c)) \quad (1)$$

oftewel

$$p(x) = \frac{x-d}{c-d}f(c) + \frac{x-c}{d-c}f(d). \quad (2)$$

Grafisch komt dit er op neer dat men op  $[a, b]$  de grafiek van  $f$  benadert door de rechte lijn door de punten  $(c, f(c))$  en  $(d, f(d))$ .

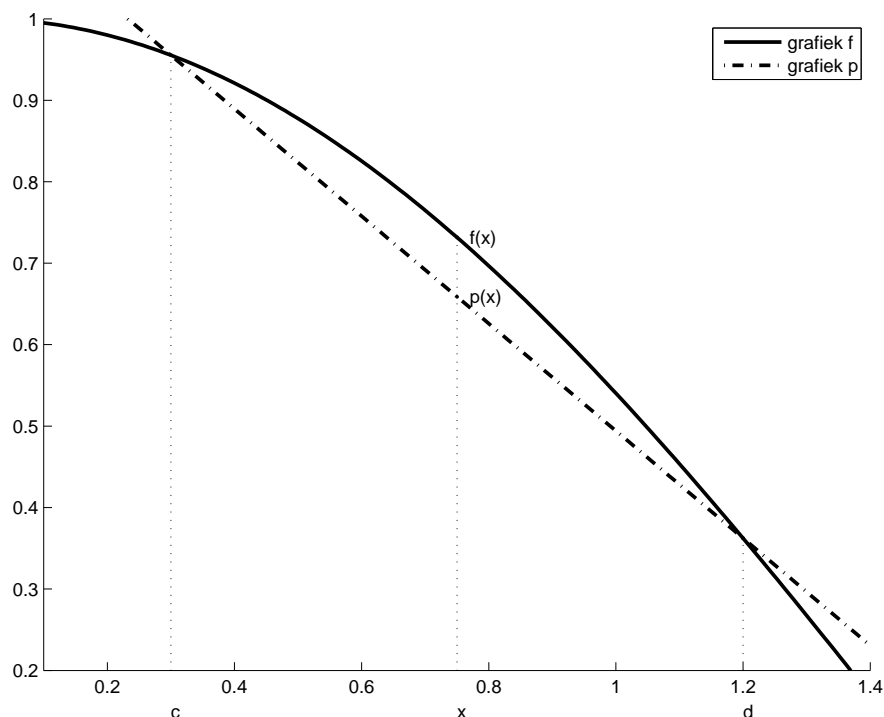
Lineaire interpolatie is een veel nauwkeuriger proces dan de meeste mensen denken. Beschouw bijvoorbeeld volgend tabelletje van de sinus.

$x$	$35^\circ$	$36^\circ$	$37^\circ$	$38^\circ$	$39^\circ$
$\sin(x)$	0.573576	0.587785	0.601815	0.615661	0.629320

Stel eens dat we  $\sin(37^\circ)$  niet gekend hadden en deze door lineaire interpolatie uit  $\sin(36^\circ)$  en  $\sin(38^\circ)$  hadden willen bepalen. Dan was er uitgekomen 0.601723, hetgeen slechts 9 eenheden van de vijfde decimaal verschilt van het ware antwoord.

Dit is eigenlijk wel een verrassend resultaat, en er rijzen dan natuurlijk meteen vragen, zoals

- zou je ook voor andere argumenten dan  $37^\circ$  zo'n goed resultaat krijgen;
- zou dat ook zo zijn op andere intervallen  $[c, d]$  met  $d - c = 2$ ;
- wat gebeurt er op langere of kortere intervallen;
- hoe zit dit bij andere functies.



FIGUUR 1: De grafiek van een functie  $f$  en van een lineair interpolatie polynoom  $p$  van  $f$ .

**1.1B** Kortom, als je een functie benadert met een andere wil je weten wat de fout ongeveer is en waar die van afhangt. Voor lineaire interpolatie doet de volgende stelling hierover een uitspraak (hierin duidt een notatie als  $[a, b, c, d]$  het gesloten interval aan opgespannen door de uitersten van het rijtje  $a, b, c, d$  en evenzo  $]a, b, c, d[$  het bijbehorende open interval;  $a, b, c, d$  hoeven daarbij geen stijgend rijtje voor te stellen; met name is dus bijv.  $[5, -3]$  toegestaan met dezelfde betekenis als  $[-3, 5]$ ).

**Stelling 1.1.1** *Laat  $c$  en  $d$  punten in  $[a, b]$  zijn met  $c \neq d$ . Zij  $f \in C[a, b]$ , en laat  $f'$  en  $f''$  bestaan op  $]a, b[$ . Dan geldt voor het lineaire interpolatiepolynoom  $p$  van  $f$  op de steunpunten  $c$  en  $d$ , en voor elke  $x \in [a, b]$  dat er een  $\xi \in ]c, d, x[$  is, zo dat*

$$f(x) - p(x) = (x - c)(x - d) \frac{f''(\xi)}{2}. \quad (3)$$

**Bewijs.** Als  $x = c$  of  $x = d$ , dan is  $f(x) - p(x) = 0$  en kan men  $\xi$  willekeurig kiezen. Zij dus nu  $x \neq c, x \neq d$ . Dan is er, voor deze vaste waarde van  $x$ , een getal  $q$  zo dat

$$f(x) - p(x) = q(x - c)(x - d).$$

Beschouw voor deze vaste  $q$  de functie

$$\varphi : t \mapsto f(t) - p(t) - q(t - c)(t - d).$$

Hiervoor geldt:  $\varphi(c) = \varphi(d) = \varphi(x) = 0$ . Volgens de stelling van Rolle zijn er minstens twee (verschillende) punten  $y$  en  $z$  in  $]c, d, x[$  zodat  $\varphi'(y) = \varphi'(z) = 0$ . Opnieuw volgens Rolle is er ook een  $\xi \in ]y, z[$ , dus  $\xi \in ]c, d, x[$ , zo dat  $\varphi''(\xi) = 0$ . Wegens  $\varphi''(t) = f''(t) - 2q$  volgt nu  $q = \frac{1}{2}f''(\xi)$ .  $\square$

**Opmerking 1.1.1** Hoewel we  $p$  een *interpolatiepolynoom* noemen kunnen we  $p$  natuurlijk ook gebruiken om te *extrapoleren* (dat wil zeggen, het benaderen van  $f(x)$  met  $p(x)$  voor  $x \notin [c, d]$ ). Stelling 1.1.1 doet dan nog steeds een uitspraak over de fout.

**1.1C** Als we ons beperken tot echte interpolatie (dus  $x \in [c, d]$ ), dan kunnen we met stelling 1.1.1 eenvoudig een bovengrens voor de absolute interpolatiefout vinden:

**Stelling 1.1.2** Zij  $f \in C[c, d]$ ,  $f \in C^2]c, d[$ . Definiëer

$$R_f(c, d) = \|f - p\|_{\infty, [c, d]} (= \sup\{|(f - p)(x)| : x \in [c, d]\}),$$

waarbij  $p$  het lineaire interpolatiepolynoom is van  $f$  op de steunpunten  $c$  en  $d$ . Dan geldt

$$\frac{1}{8}(d - c)^2 \inf_{]c, d[} |f''(x)| \leq R_f(c, d) \leq \frac{1}{8}(d - c)^2 \sup_{]c, d[} |f''(x)|.$$

I.h.b. geldt voor “vaste”  $m = \frac{d+c}{2}$  dat

$$\lim_{(d-c) \downarrow 0} \frac{R_f(c, d)}{\frac{1}{8}(d - c)^2} = |f''(m)|.$$

**Bewijs.** Dit volgt eenvoudig uit  $\max_{x \in [c, d]} |\frac{1}{2}(x - c)(x - d)| = |\frac{1}{2}(m - c)(m - d)| = \frac{1}{8}(d - c)^2$ ,  $R_f(c, d) \geq |f(m) - p(m)|$ , en, voor de laatste bewering, een toepassing van de insluitstelling voor limieten.  $\square$

In woorden zegt bovenstaande stelling dus dat, indien  $f''(m) \neq 0$ , de maximale fout bij lineaire interpolatie asymptotisch afneemt met het kwadraat van de afstand tussen de steunpunten.

**Voorbeeld 1.1.1** Neem  $f(x) = \sin(x)$  (voortaan  $x$  in radialen). In onderstaande tabel staan voor verschillende waarden van  $m = \frac{c+d}{2}$  en  $d - c$  de waarden van  $R_{\sin}(c, d)$  gegeven (op de regel met “R”) en de factor waarmee  $R_{\sin}(c, d)$  gereduceerd wordt als bij vaste waarde van  $m$  de waarde van  $d - c$  gehalveerd wordt (op de regels met “red”).

$m \setminus (d - c)$	.2	.1	.05	
1	.004204331	.001051650	.000262948	R
	3.998	3.9995		red
0.5	.002397355	.000599297	.000149821	R
	4.0003	4.00009		red
0.1	.000511833	.000125615	.0000312499	R
	4.075	4.0197		red
0	.0000641073	.00000801741	.00000100230	R
	7.996	7.9989		red

We zien, dat in de eerste drie rijen de reductiefactor het getal 4 heel aardig benadert. Dat in de derde rij de reductiefactoren wat verder van 4 afdiggen wordt verklaard doordat de tweede afgeleide van de sinus op de intervallen  $[0, 0.2]$ ,  $[0.05, 0.15]$  en  $[0.075, 0.125]$  in relatieve zin sterk variëert. In de vierde rij ziet men een reductiefactor 8 opdoemen. Men kan inderdaad bewijzen dat deze factor asymptotisch juist is (zie paragraaf 1.1D).

## 1.1D

**Opgave 1.1.1** Zij  $f$  een voldoende vaak differentiëerbare functie, en zij  $p$  het lineaire interpolatiepolynoom van  $f$  op de punten  $c \neq d$ . Laat zien dat voor alle  $x$  en  $a \notin ]x, c, d[$  er een  $\eta \in ]a, c, d, x[$  is met

$$f(x) - p(x) = \frac{1}{2}(x - c)(x - d)[f''(a) + (\frac{x+c+d}{3} - a)f'''(\eta)]. \quad (4)$$

**Aanwijzing:** Het bewijs loopt analoog aan dat van stelling 1.1.1; bedenk dat voor de overeenkomstig gedefiniëerde functie  $\varphi$  ook nog geldt:  $\varphi''(a) = 0$ .

**Opgave 1.1.2** In de situatie van opgave 1.1.1, en met  $\zeta := \frac{x+c+d}{3}$  het zwaartepunt van  $c, d$  en  $x$ , laat zien dat er een  $\theta \in ]a, \zeta[$  en een  $\phi \in ]\eta, \theta[$  zijn, zo dat

$$f(x) - p(x) = \frac{1}{2}(x - c)(x - d)[f''(\zeta) + (a - \zeta)(\theta - \eta)f'''(\phi)] \quad (5)$$

Merk op door  $a$  willekeurig dicht bij  $]x, c, d[$  te nemen dat dit betekent dat  $\xi$  uit (3) ongeveer gelijk is aan het zwaartepunt van  $x$  en de steunpunten  $c$  en  $d$ . Laat zien dat indien  $f$  een derde graads polynoom is er zelfs gelijkheid geldt.

**Opgave 1.1.3** In de situatie van opgave 1.1.2 en met  $m = \frac{d+c}{2}$ , bewijs dat indien  $f''(m) = 0$  er geldt

$$\lim_{(d-c) \downarrow 0} \frac{8 \cdot 27 \cdot R_f(c, d)}{\sqrt{3}(d - c)^3} = |f'''(m)|.$$

Verklaar hiermee de factor 8 in de laatste rij van de tabel in 1.1C. (Merk op dat het rekenwerk aanzienlijk vereenvoudigt als men  $m = 0$  veronderstelt. Ga na dat dit geen wezenlijke beperking is.)

## 1.2 Lagrange interpolatie

**1.2A** Als natuurlijke generalisatie van lineaire interpolatie kan men de approximatie beschouwen van een functie  $f$  door een  $n$ -de graads polynoom  $p$ , zodanig dat de waarden van  $f$  en  $p$  in een vooraf gegeven  $n + 1$ -tal *verschillende* punten  $x_0, \dots, x_n$  overeenstemmen. Men spreekt dan van  *$n$ -de orde Lagrange interpolatie*.

Een polynoom  $p$  dat aan deze eisen voldoet is gemakkelijk aan te geven, en wel in een vorm die een duidelijke generalisatie van (2) is:

$$p(x) = \sum_{k=0}^n f(x_k)L_{kn}(x) \quad (6)$$

waarin

$$L_{kn}(x) = \frac{(x - x_0) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n)}{(x_k - x_0) \cdots (x_k - x_{k-1})(x_k - x_{k+1}) \cdots (x_k - x_n)}. \quad (7)$$

Immers,  $L_{kn}$  heeft voor elke  $k$  de graad  $n$ , en  $L_{kn}(x_i) = 0$  voor  $i \neq k$ ,  $L_{kn}(x_k) = 1$ , zodat  $p(x_i) = f(x_i)$  voor  $i = 0, \dots, n$ .

Het polynoom  $p$  is het *Lagrange interpolatiepolynoom* van  $f$  op de steunpunten  $x_0, \dots, x_n$ .

**Opmerking 1.2.1** Het polynoom  $p$  is uniek. Immers zij  $\tilde{p}$  een ander  $n$ -de graads polynoom dat met  $f$  overeenstemt op de verschillende punten  $x_0, \dots, x_n$ . Dan is  $p - \tilde{p}$  een  $n$ -de graads polynoom met  $n + 1$  verschillende nulpunten, hetgeen betekent dat  $p - \tilde{p} = 0$ .

Het rechterlid van (6) is de *Lagrange representatie* van  $p$ . (Elk polynoom heeft talloze representaties; vergelijk bijvoorbeeld (1) en (2)). De polynomen  $L_{kn}$  uit (7) zijn de *Lagrange coëfficiënten*; zij hangen slechts af van  $x_0, \dots, x_n$ , en niet van  $f$ . In de literatuur worden ze vaak geschreven als

$$L_{kn}(x) = \frac{\omega(x)}{(x - x_k)\omega'(x_k)} \quad \text{met} \quad \omega(x) = \prod_{i=0}^n (x - x_i).$$

Stelling 1.1.1 laat zich eenvoudig generaliseren. Het bewijs laten we achterwegen: het loopt geheel analoog aan dat van stelling 1.1.1.

**Stelling 1.2.1** Laat  $x_0, \dots, x_n$  verschillende punten zijn in  $[a, b]$ . Laat  $f \in C[a, b]$ , en laat  $f', f'', \dots, f^{(n+1)}$  bestaan op  $]a, b[$ . Dan geldt voor het bijbehorende interpolatiepolynoom  $p$  en voor elke  $x \in [a, b]$  dat er een  $\xi \in ]x_0, x_1, \dots, x_n, x[$  is, zo dat

$$f(x) - p(x) = (x - x_0)(x - x_1) \cdots (x - x_n) \frac{f^{(n+1)}(\xi)}{(n+1)!}. \quad \square \quad (8)$$

**Opmerking 1.2.2** Wat betreft het punt  $\xi$  in stelling 1.2.1 kan men weer bewijzen dat

$$\xi \approx \frac{1}{n+2} \left( x + \sum_{i=0}^n x_i \right), \quad (9)$$

dus  $\xi$  is ongeveer het zwaartepunt van  $x, x_0, x_1, \dots, x_n$  (Vergelijk opgave 1.1.2)

**1.2B** Een belangrijk geval krijgt men als de afstand tussen twee opeenvolgende steunpunten steeds gelijk is, zeg gelijk aan  $h$ . Dan nummert men de steunpunten zo dat  $x_k = x_0 + kh$  voor  $k = 0, \dots, n$ . Men noemt dit *equidistante* Lagrange interpolatie. Schrijven we nu  $x = x_0 + qh$ , dan krijgt men:

$$L_{kn}(x) = \frac{q \cdots (q - (k-1))(q - (k+1)) \cdots (q - n)}{(-1)^{n-k} k! (n-k)!} \quad (10)$$

voor  $k = 0, \dots, n$ . Merk op dat het rechterlid van (10) onafhankelijk is van  $h$ , hetgeen betekent dat de coëfficiënten hergebruikt kunnen worden wanneer de stapgrootte  $h$  veranderd wordt.

**1.2C** Stel dat een functie  $f$  gegeven is voor  $x = 0, 0.1, 0.2, \dots$ . Als we dan met 4 punts interpolatie  $f(0.65)$  willen bepalen kunnen we dat bijv. doen met als steunpunten 0.4 t/m 0.7, 0.5 t/m 0.8 of 0.6 t/m 0.9. Zou dat verschil uitmaken? We proberen dit eens voor  $f(x) = e^x$ . Men krijgt dan resp. 1.9155477; 1.9155363; 1.9155489, zodat  $f - p = -6.9 \times 10^{-6}$ ;  $4.5 \times 10^{-6}$ ;  $-8.1 \times 10^{-6}$ . Dus als we de steunpunten zo kiezen dat het interpolatiepunt  $x$  in het middelste deelinterval ligt krijgen we het beste resultaat.

Is dit verschil stelselmatig? Bekijken we eens de restterm

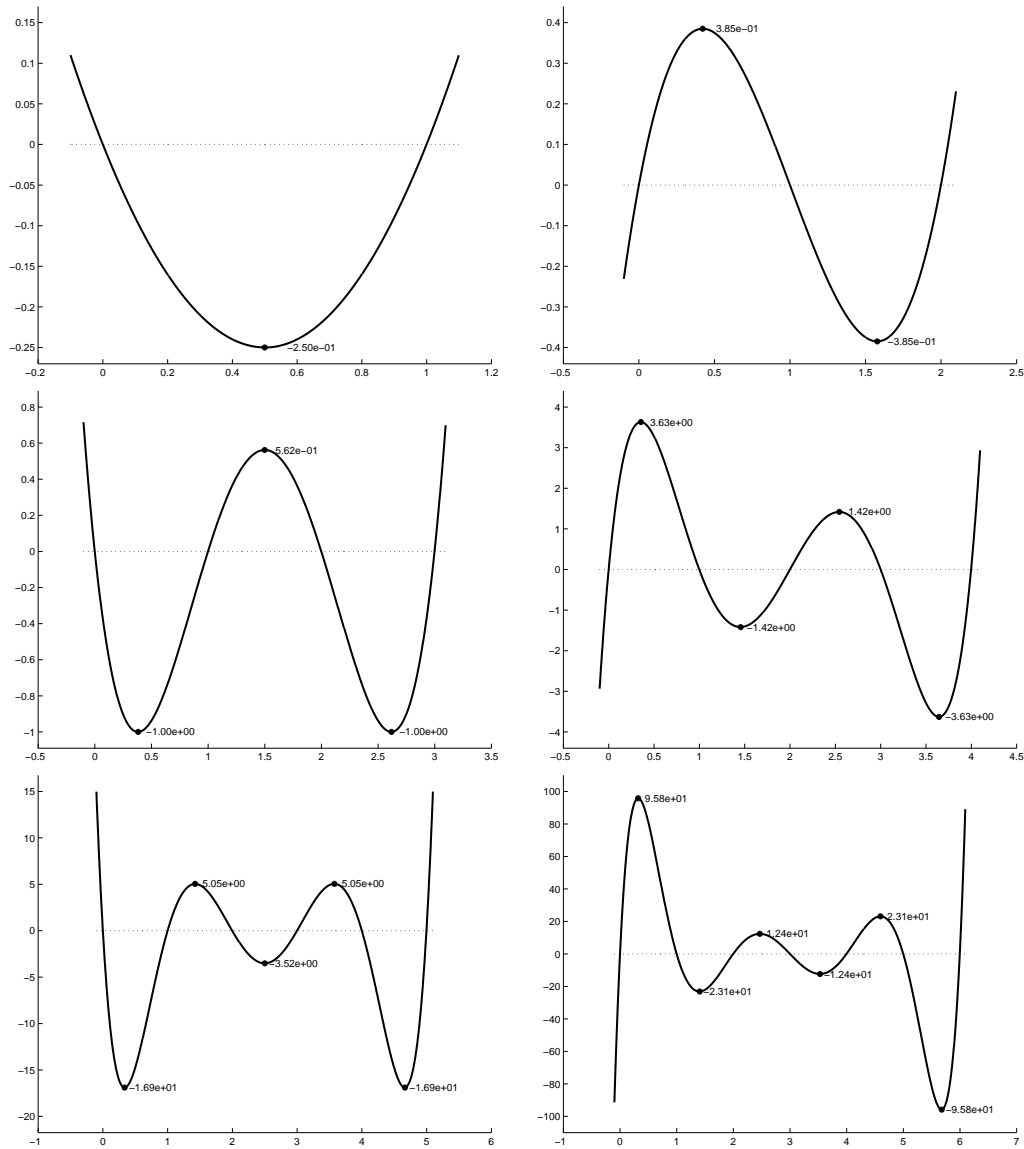
$$(x - x_0)(x - x_1) \cdots (x - x_n) \frac{f^{(n+1)}(\xi)}{(n+1)!} \quad (11)$$

Wanneer  $f^{(n+1)}$  op de gezamenlijke interpolatie intervallen slechts weinig variëert, hetgeen geen onredelijke aanname is aangezien men meestal interpoleert op kleine intervallen, dan wordt het gedrag van de fout in afhankelijkheid van  $x$  voornamelijk bepaald door het gedrag van het polynoom  $(x - x_0)(x - x_1) \cdots (x - x_n)$ .

In het equidistante geval dat we zojuist beschouwden hebben we  $x_i = x_0 + ih$ . Schrijven we weer  $x = x_0 + qh$  dan krijgen we

$$(x - x_0)(x - x_1) \cdots (x - x_n) = h^{n+1} q(q-1) \cdots (q-n). \quad (12)$$

Voor  $n = 1, 2, 3, 4, 5$  zijn de grafieken van  $q(q-1) \cdots (q-n)$  in figuur 2 geschetst. De



FIGUUR 2: De grafiek van de functies  $q \rightsquigarrow q(q-1) \cdots (q-n)$  voor  $n = 1, 2, 3, 4, 5, 6$  met daarin de waarden van lokale extrema (in drie decimalen).

waarden van de lokale extrema zijn in de tekening aangegeven. Het blijkt dat deze extrema inderdaad het kleinst zijn midden op het interpolatie interval zodat men bij langzaam variërende  $f^{(n+1)}$  dus inderdaad mag verwachten dat de interpolatiefouten het kleinst zijn in het (of een) middelste interpolatie interval. Hiermee moet men dan

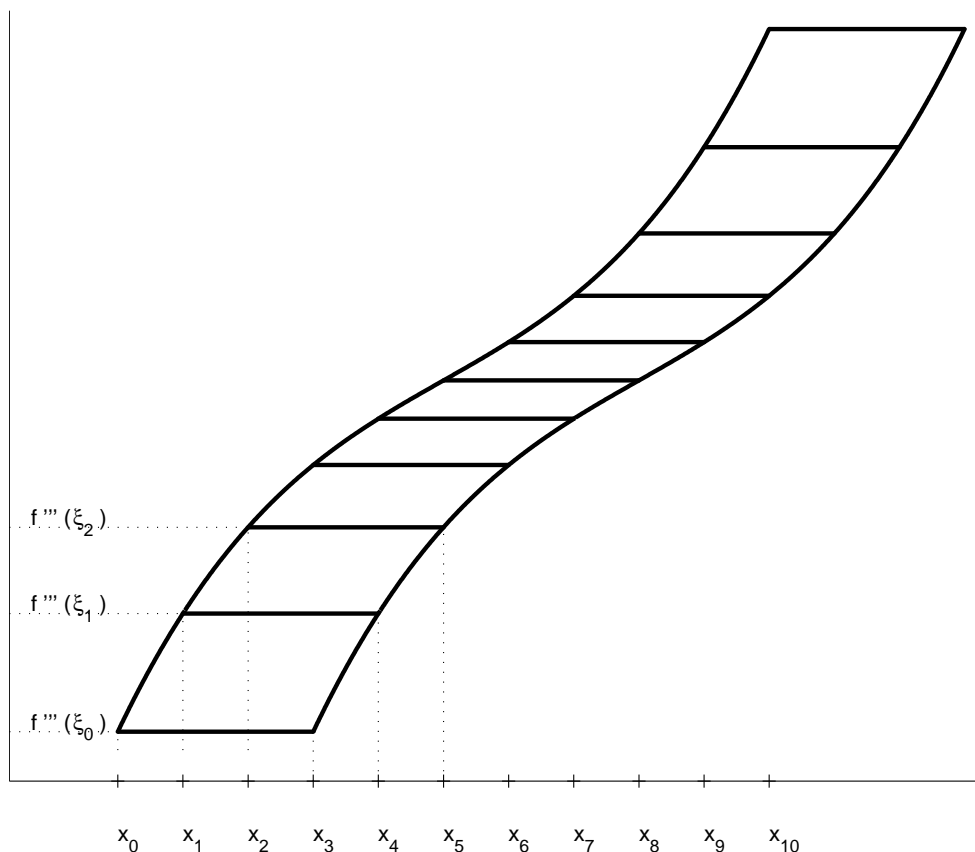
bij de keuze van de steunpunten rekening proberen te houden. Buiten het interval  $[0, n]$  groeit het rechterlid van (12) snel als functie van  $q$ . Dit verklaart de veel grotere fout bij *extrapolatie*.

**1.2D** Bij functies die men in de praktijk tegenkomt zijn de hogere afgeleiden vaak bijzonder ingewikkelde uitdrukkingen, waarbij het vrijwel ondoenlijk is te schatten hoe groot de waarden wel kunnen worden. Toch kan men zich wel een indruk verschaffen van de waarden van  $f^{(n+1)}$ . Een van de methoden daarvoor werkt als volgt:

Stel: We kennen de functiewaarden in een (groot) aantal punten  $x_0, \dots, x_n, x_{n+1}, \dots$ , (die we eenvoudigheidshalve in volgorde van grootte genummerd denken:  $x_0 < x_1 < \dots$ ). Laat  $p_0$  het  $n$ -de orde interpolatiepolynoom zijn op  $x_0, \dots, x_n$ , dan kunnen we de waarde van het polynoom en die van de functie vergelijken in het punt  $x_{n+1}$ . We vinden zodoende de waarde van  $f^{(n+1)}$  in 'n punt  $\xi_0 \in [x_0, x_{n+1}]$ :

$$f^{(n+1)}(\xi_0) = \frac{(n+1)!}{(x_{n+1} - x_0)(x_{n+1} - x_1) \cdots (x_{n+1} - x_n)} [f(x_{n+1}) - p_0(x_{n+1})]. \quad (13)$$

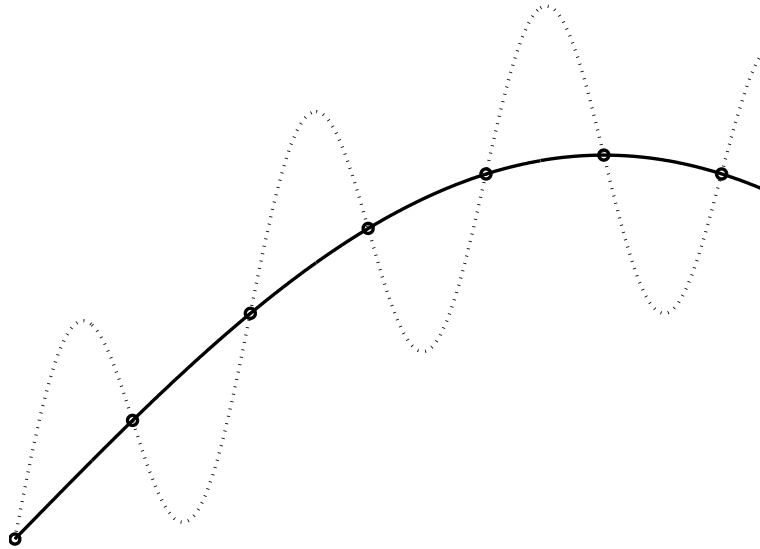
Evenzo vindt men uitgaande van het interpolatiepolynoom  $p_1$  op  $x_1, \dots, x_{n+1}$  en de functiewaarde in  $x_{n+2}$  de waarde van  $f^{(n+1)}$  in 'n  $\xi_1 \in [x_1, x_{n+2}]$  etc.



FIGUUR 3: De grafiek van de functie  $f'''$  loopt door de punten  $(\xi_i, f'''(\xi_i))$ . Van deze punten weten we dat ze op de horizontale lijnstukjes liggen (het  $i$ -de punt ligt ergens op het  $i$ -de lijnstukje).

Stel nu eens dat deze aldus verkregen waarden  $f^{(n+1)}(\xi_i)$  een gladjes verlopende rij vormen en laten we deze, althans in gedachten, eens grafisch uitzetten d.m.v. horizontale lijnstukjes op de hoogte  $f^{(n+1)}(\xi_i)$  boven het interval  $[x_i, x_{i+n+1}]$  (zie figuur 3 voor  $n = 2$ ). Dan snijdt de grafiek van  $f^{(n+1)}$  al deze horizontale lijnstukjes en men *hoopt* dat

hij binnen de aangegeven strook loopt. Op die manier krijgt men dan een indruk van de waarden van  $f^{(n+1)}$ . Men kan dit beschouwen als een bedenkelijke gang van zaken, maar men doet hetzelfde wanneer men, om enig zicht op het verloop van een lastige functie te krijgen, een aantal functiewaarden bepaalt (zie figuur 4) en dan aanneemt dat het verloop door de getrokken kromme wordt weergegeven en niet door de gestippelde.



FIGUUR 4: Als we van een functie een aantal waarden kennen (gemarkeerd met 'o'), dan nemen we aan dat de grafiek van de van de functie meer lijkt op de getrokken kromme dan op de gestippelde.

**1.2E** Tot dusver hebben we ons bezig gehouden met de fout die je bij interpolatie maakt als je gebruik maakt van exacte functiewaarden. In werkelijkheid heb je die vrijwel nooit. Zo moet men er in het voorbeeld in 1.1A mee rekening houden dat er (af rond)fouten ter grootte van  $\frac{1}{2}10^{-6}$  op de functiewaarden kunnen zitten. Het is een essentiële numeriek wiskundige vraagstelling bij allerhande numerieke processen hoe dergelijke “ingangsfouten” in het resultaat doorwerken. We gaan dat nu na voor interpolatie.

We merken eerst op dat als we  $n+1$  punts interpolatie toepassen op een polynoom  $f$  van graad hoogstens  $n$ , we wegens opmerking 1.2.1  $f$  zelf terug krijgen. In het bijzonder geldt dit voor  $f(x) = 1$ . Bijgevolg geldt de interessante relatie:  $\sum_{k=0}^n L_{k,n}(x) = 1$  voor alle  $x$ . I.h.a. zijn echter niet alle Lagrange coëfficiënten *positief*, zodat  $\sum_{k=0}^n |L_{k,n}(x)| \geq 1$ . Stel nu dat de functiewaarden  $f(x_k)$  met fouten  $\epsilon_k$  belast zijn,  $|\epsilon_k| \leq \bar{\epsilon}$ . Deze werken in het interpolatieresultaat door als  $\sum L_{k,n}(x)\epsilon_k$ , en dat is hoogstens  $\bar{\epsilon} \sum |L_{k,n}(x)|$ . Deze grens kan bereikt worden; namelijk, als  $\epsilon_k$  net telkens  $\pm\bar{\epsilon}$  is met het teken van  $L_{k,n}(x)$ . Als nu  $\max_x \sum |L_{k,n}(x)|$  veel groter dan 1 zou zijn zouden fouten in de  $f(x_k)$  dus zeer versterkt kunnen doorwerken. Bij equidistante interpolatie (d.w.z.  $x_k = x_0 + kh$ ) met niet al te veel punten valt dit echter nog al mee. Onderstaand geven we voor dit geval bovengrenzen voor  $\sum_{k=0}^n |L_{k,n}(x)|$  in afhankelijkheid van  $n$  en van de ligging van  $x$  ten

opzichte van  $x_0, \dots, x_n$ .

	$x \in [x_0, x_1]$	$x \in [x_1, x_2]$	$x \in [x_2, x_3]$	$x \in [x_3, x_4]$	$x \in [x_4, x_5]$
$n = 1$	<b>1</b>				
$n = 2$	<b>1.25</b>	<b>1.25</b>			
$n = 3$	1.63	<b>1.25</b>	1.63		
$n = 4$	2.3	<b>1.4</b>	<b>1.4</b>	2.3	
$n = 5$	3.1	1.6	<b>1.4</b>	1.6	3.1

Men ziet hieruit nog dat deze waarden bij stijgende  $n$  langzaam oplopen. Men kan bewijzen (zie bijvoorbeeld Natanson, “*Constructive Function Theory*, vol. III”, Theorem 1) dat voor willekeurige  $x_0 < \dots < x_n$ ,

$$\max_{x \in [x_0, x_n]} \sum_{k=0}^n |L_{kn}(x)| > \frac{\ln(n+1)}{8\sqrt{\pi}} \quad (14)$$

**1.2F** Veronderstel dat we geïnteresseerd zijn in de waarde van een functie  $f$  in een punt  $x$ , en dat de functie slechts bekend is in de punten  $\dots < x_{-1} < x_0 < x_1 < \dots$  waarbij  $x \in (x_0, x_1)$ . Indien we niet tevreden zijn met het resultaat van een zeker Lagrange interpolatie polynoom op een aaneengesloten rijtje steunpunten waaronder  $x_0$  en  $x_1$ , dan kunnen we slechts hopen een betere benadering te krijgen door meer steunpunten toe te voegen die evenwel verder verwijderd zijn van  $x$ .

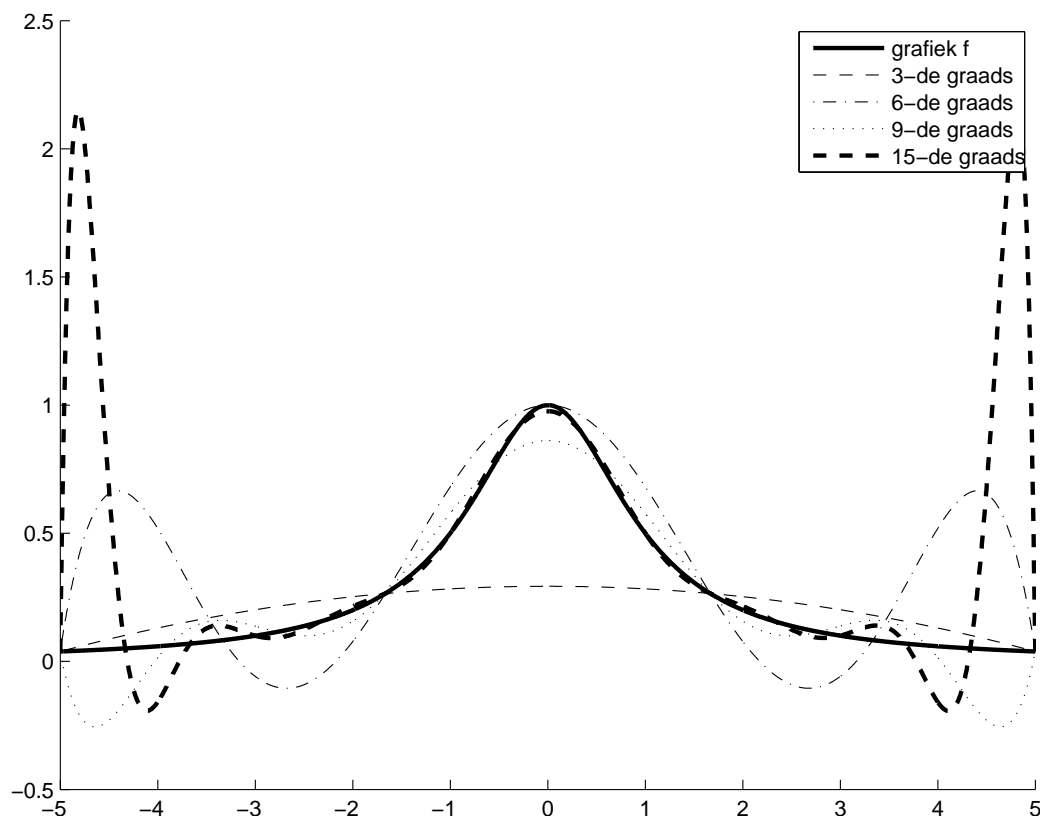
Bij equidistante steunpunten, efficiënte berekeningswijzen van het Lagrange interpolatiepolynoom waarbij de steunpunten in de volgorde  $x_0, x_1, x_2, \dots$  danwel  $x_0, x_1, x_{-1}, x_2, \dots$  worden toegevoegd staan bekend als *Newton's voorwaartse formule* respectievelijk *Gauß' voorwaartse formule* (zie vraagstuk 9). Op grond van inzichten verkregen in 1.2C kunnen we voor vaste graad  $n$ , aannemende dat  $f^{(n+1)}$  langzaam varieert, van de Gauß formule betere resultaten verwachten.

De vraag is of er voor  $n \rightarrow \infty$  convergentie optreedt van het resultaat van interpolatie naar  $f(x)$ . Dit blijkt slechts onder bijzondere voorwaarden het geval te zijn. Zo is bij Newton's voorwaartse formule een noodzakelijke (niet eens voldoende) voorwaarde dat  $f$  regulier (d.w.z. differentiëerbaar) is in het complexe halfvlak  $\operatorname{Re} z \geq x_0$ . En bij de formule van Gauß moet  $f$  zelfs in het hele complexe vlak regulier zijn. Voor een functie als  $\frac{1}{1+x^2}$ , die op de reële as  $C^\infty$  is, kan dus eenvoudig geen convergentie optreden bij Gauß, en ook niet bij Newton als  $x_0 < 0$ , wegens de singulariteiten voor  $x = \pm i$ .

Uit het hieronder gegeven voorbeeld blijkt zelfs dat wanneer men met het verhogen van de orde tegelijkertijd de afstand tussen de steunpunten verkleint convergentie geen uitgemaakte zaak is:

**Voorbeeld 1.2.1** [Runge] Zij  $p_n$  het Lagrange interpolatie polynoom van de functie  $\frac{1}{1+x^2}$  op de punten  $x_k^{(n)} = -5 + k \frac{10}{n}$ ,  $k = 0, 1, \dots, n$ . Er geldt  $\lim_{n \rightarrow \infty} p_n(x) = f(x)$  dan en slechts dan als  $\int_{-5}^5 \log |x - \xi| d\xi < \int_{-5}^5 \log |x - i| d\xi \Leftrightarrow |x| < 3.6334 \dots$ , zie figuur 5.

Problemen als in bovenstaand voorbeeld worden veroorzaakt door (14). Voor een overzicht van *positieve* resultaten aangaande convergentie van interpolatieschema's bij het verhogen van de orde, zie P.J. Davis, “*Interpolation and Approximation*”, Ch. IV. Een voldoende voorwaarde voor uniforme convergentie is ruwweg dat de afstand tussen het betreffende gebied en de singulariteiten groot genoeg is. Deze afstand bepaalt immers de grootte van de afgeleiden.



FIGUUR 5: De getrokken kromme is de grafiek van de functie  $f : x \rightsquigarrow 1/(1+x^2)$ . De andere krommen zijn de grafieken van Lagrange interpolatie polynomen op steunpunten die equidistant verdeeld liggen tussen  $-5$  en  $+5$ . De benaderingskwaliteit verbetert in het midden bij toenemend aantal steunpunten, terwijl die aan de rand verslechtert.

**Opgave 1.2.1** Zij  $f \in C[a, b]$ ,  $x_k^{(n)} = a + k \frac{b-a}{n}$ ,  $k = 0, \dots, n$ . Benader  $f$  door  $p_n$  gedefinieerd als het stuksgewijs lineair interpolatie polynoom op de intervallen  $[x_k^{(n)}, x_{k+1}^{(n)}]$ . Bewijs dat  $\lim_{n \rightarrow \infty} \|f - p_n\|_{\infty, [a, b]} = 0$ .

### 1.3 Interpolatie met functiewaarden en afgeleiden

#### 1.3A De Taylor polynomen.

Een heel bekende methode om een functie te benaderen met een polynoom is het gebruiken van de *Taylor ontwikkeling*. Als eigenlijke benaderingsmethode wordt deze techniek in de numerieke wiskunde niet zo vaak gebruikt; veel belangrijker echter is zijn rol in de analyse van andere numerieke processen. We vatten hier kort enige resultaten uit de analyse samen.

**Stelling 1.3.1** Laat  $c$  een punt in het interval  $[a, b]$  zijn, en  $f \in C^n[a, b]$ . Het  $n$ -de orde Taylor polynoom  $p_n$  van  $f$  in  $c$  is gegeven door

$$p_n(x) = \sum_{k=0}^n (x-c)^k \frac{f^{(k)}(c)}{k!}.$$

Voor  $x \in [a, b]$  geldt voor de restterm  $R_n(x) = f(x) - p_n(x)$  dat

$$\text{a) } R_n(x) = \int_c^x \frac{(x-t)^{n-1}}{(n-1)!} (f^{(n)}(t) - f^{(n)}(c)) dt \quad (= o((x-c)^n) \quad (x \rightarrow c)).$$

Als zelfs  $f \in C^{n+1}[c, x]$ , dan

$$\text{b) } R_n(x) = \int_c^x \frac{(x-t)^n}{n!} f^{(n+1)}(t) dt \quad (= \mathcal{O}((x-c)^{n+1}) \quad (x \rightarrow c)).$$

en er bestaat dan een,  $x$  afhankelijke,  $\xi \in ]c, x[$  zodat

$$\text{c) } R_n(x) = \frac{(x-c)^{n+1}}{(n+1)!} f^{(n+1)}(\xi)$$

### Opmerking 1.3.1

- Er geldt  $p_n^{(k)}(c) = f^{(k)}(c)$  voor  $k = 0, \dots, n$ .
- Bovenstaande stelling doet geen uitspraak over convergentie van  $R_n(x)$  naar nul voor  $n \rightarrow \infty$ .

**1.3B Interpolatie in het algemeen.** In het algemene geval stelt men zich bij interpolatie de opgave om bij een functie  $f$  een polynoom  $p$  te construeren, zodanig dat in een aantal gegeven steunpunten niet alleen de waarden van  $f$  en  $p$  overeenstemmen, maar ook de waarden van een aantal van hun afgeleiden. Het aantal afgeleiden mag per steunpunt verschillen. Het zal duidelijk zijn, dat zowel Lagrange interpolatie als approximatie met Taylor polynomen bijzondere gevallen zijn.

We bekijken hiertoe een rij  $(x_0, x_1, \dots, x_n)$  getallen  $x_i$  die niet allemaal verschillend hoeven te zijn. Zo komt de waarde 0 drie maal in de rij  $(0, 0, \frac{1}{2}, 1, 1, 0)$  voor en 1 twee maal.

**Notatie 1.3.1** Laat  $(x_0, x_1, \dots, x_n)$  een rij getallen zijn. We zeggen dat

$$p = f \text{ op } (x_0, x_1, \dots, x_n)$$

als  $p(x_i) = f(x_i)$  voor alle  $i = 0, \dots, n$  en bovendien  $p^{(j)}(x_i) = f^{(j)}(x_i)$  als de waarde  $x_i$   $j + 1$  of meer keer voorkomt in de rij  $(x_0, x_1, \dots, x_n)$ . We nemen hierbij aan dat  $f$  in de  $x_i$  gedefiniëerd is en in dat punt ook voldoende vaak differentiëerbaar is.

Dus als  $p(0) = f(0)$ ,  $p'(0) = f'(0)$ ,  $p''(0) = f''(0)$ ,  $p(\frac{1}{2}) = f(\frac{1}{2})$ ,  $p(1) = f(1)$  en  $p'(1) = f'(1)$ , dan is  $p = f$  op  $(0, 0, \frac{1}{2}, 1, 1, 0)$ .

**Stelling 1.3.2** Laat  $(x_0, x_1, \dots, x_n)$  een rij getallen zijn.

Dan is er een uniek polynoom  $p$  van graad hoogstens  $n$  waarvoor  $p = f$  op  $(x_0, x_1, \dots, x_n)$ :  $p$  is het interpolatie polynoom voor  $f$  op  $(x_0, x_1, \dots, x_n)$ .  $\square$

Anders dan bij approximatie met Lagrange en Taylor polynomen is in dit algemene geval een expliciete gedaante van het interpolatiepolynoom niet zo eenvoudig aan te geven. We volstaan daarom met het geven van bovenstaand resultaat over existentie, uniciteit en onderstaand resultaat over de restterm. De bewijzen van deze stellingen kan men vinden in bijlage A. In deze bijlage (in het bewijs van Stelling A.3) en in Vraagstuk 9 wordt overigens een recursieve definitie gegeven voor het interpolatiepolynoom, waarin het interpolatiepolynoom uitgedrukt wordt in lager graads interpolatiepolynomen (dat zijn polynomen die  $f$  in één punt minder interpoleren). Met behulp van deze uitdrukking (een generalisatie van de voorwaarts Newton formule als vermeld in 1.2F) kan je polynoomwaarden gemakkelijk en efficiënt berekenen.

**Opmerking 1.3.2**

• Een essentieel element bij algemene interpolatie is, dat men in ieder steunpunt slechts een aaneengesloten stel afgeleiden, beginnend vanaf de nulde (de functiewaarde zelf) mag nemen. Als men dat niet doet, zijn existentie en uniciteit niet altijd gegarandeerd.

**Voorbeeld 1.3.1** Er bestaat geen eerste graads polynoom met

$$p(0) = 1, \quad p''(0) = 1$$

en er bestaan er oneindig veel van de tweede graad die aan deze eisen voldoen.

• Naast Lagrange interpolatie ( $x_i \neq x_j$  alle  $i \neq j$ ) en Taylor polynomen ( $x_i = x_0$  alle  $i$ ) heeft ook het geval  $x_{2i} = x_{2i+1}$  (alle  $i$ ,  $2i + 1 \leq n$ ) en  $x_{2i} \neq x_{2j}$  alle  $i \neq j$  een afzonderlijke naam; men spreekt dan van *Hermite interpolatie*.

**Stelling 1.3.3** Laat  $(x_0, x_1, \dots, x_n)$  een rij getallen zijn. Zij  $p$  het polynoom van graad hoogstens  $n$  dat gelijk is aan  $f$  op  $(x_0, x_1, \dots, x_n)$ . Als  $f \in C[a, b]$  en  $f^{(n+1)}$  bestaat op  $]a, b[$ , dan bestaat er voor elke  $x \in [a, b]$  een  $\xi \in ]x_0, \dots, x_n[$  zodanig dat

$$f(x) - p(x) = (x - x_0)(x - x_1) \cdots (x - x_n) \frac{f^{(n+1)}(\xi)}{(n+1)!} \quad \square \quad (15)$$

Als voor de  $\xi$  in stelling 1.2.1 geldt ook hier weer (9).

Er is bijvoorbeeld precies een polynoom  $p$  van graad hoogstens 5 zodat  $p = f$  op  $(0, 0, 0, \frac{1}{2}, 1, 1)$ . Voor dit polynoom is er voor iedere  $x \in \mathbb{R}$  een  $\xi$  tussen 0, 1 en  $x$  zodat

$$f(x) - p(x) = x^3(x - \frac{1}{2})(x - 1)^2 \frac{f^{(6)}(\xi)}{6!}.$$

**1.4 Inverse (lineaire) interpolatie**

Bij het tabelletje in 1.1A van de sinus zou men zich de vraag kunnen stellen: voor welke waarde van  $x$  heeft  $\sin(x)$  de waarde 0.60000? De vraag kan benaderd beantwoord worden door de tabel van rechts naar links te lezen en lineair te interpoleren in 0.600000. Men vat dus de waarden van de sinus op als argumenten, de waarden van  $x$  als functiewaarden, en interpoleert in feite op de inverse functie, in dit geval de boogsinus. Dit heet daarom *inverse interpolatie*.

Met (2), en m.b.v. de waarden van  $\sin(36^\circ)$  en  $\sin(37^\circ)$  vinden we in dit geval  $x = 36.8706^\circ$ . Hoe nauwkeurig is dit antwoord? Met stelling 1.1.1 vinden we voor de fout:

$$-0.001815 \times 0.012215 \times \frac{f''(\xi)}{2} = -0.000011 \times f''(\xi),$$

waarin  $f$  echter niet de sinus, maar de inverse functie is. Van de sinus kennen we de inverse functie, maar in het algemeen is de inverse niet expliciet bekend.

Een mogelijkheid biedt dan “invers differentiëren”. Voor een inverteerbare en voldoende gladde functie  $g$ , zij  $y = g(x)$  en dus  $x = g^{-1}(y)$ . Door de relatie  $g^{-1} \circ g = Id$  te differentiëren m.b.v. de kettingregel vindt men  $(g^{-1})'(y) = \frac{1}{g'(x)}$ . Hiermee verkrijgt men

$$(g^{-1})''(y) = \frac{d}{dy} \left( \frac{1}{g'(g^{-1}(y))} \right) = -\frac{g''(g^{-1}(y)) \frac{d}{dy}(g^{-1}(y))}{(g'(g^{-1}(y)))^2} = -\frac{g''(x)}{(g'(x))^3} \quad (16)$$

Met  $g(x) = \sin(x)$  kunnen wij nu de fout schatten, waarbij we dan ook  $g'(x) = \cos(x)$  nodig hebben. Wegens  $\cos 36^\circ = 0.809$ ,  $\cos 37^\circ = 0.799$  vinden we als bovengrens voor de fout  $\approx 0.000011 \times 0.6 / (0.8)^3 \approx 1.3 \cdot 10^{-5}$  radialen  $\approx 7.5 \cdot 10^{-4}$  graden. Het juiste antwoord is  $36.8699$  graden, en de fout  $\approx 0.0007^\circ$ .

## 2 Numerieke differentiatie

### 2.1 Eenvoudige formules. Effect van afrondfouten

**2.1A** Vanwege de definitie  $f'(x_0) = \lim_{h \rightarrow 0} \frac{f(x_0+h) - f(x_0)}{h}$  zou men  $f'(x_0)$  kunnen benaderen met het eindige quotiënt, een zogenaamde *differentie quotiënt*. Als we schrijven

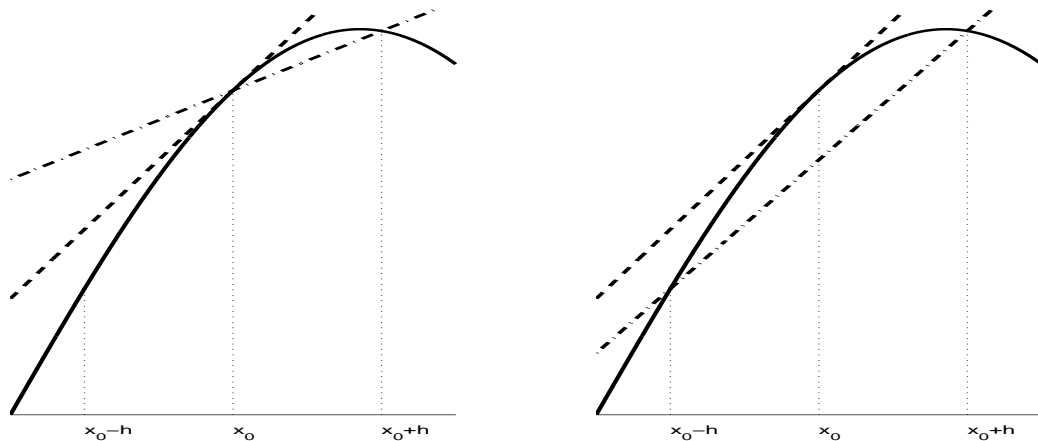
$$f'(x_0) = \frac{f(x_0+h) - f(x_0)}{h} + R_1(h), \quad (17)$$

dan is  $R_1(h)$  de fout bij de *stapgrootte*  $h$ .

Aannemend dat  $f \in C^2[x_0, x_0+h]$ , vinden we door  $f(x_0+h)$  in een Taylor reeks te ontwikkelen een uitdrukking voor  $R_1(h)$ :

$$R_1(h) = -\frac{h}{2}f''(\xi), \quad \text{met } \xi \text{ tussen } x_0 \text{ en } x_0+h. \quad (18)$$

Voor  $h$  klein zal  $f''(\xi)$  niet veel verschillen van  $f''(x_0)$ , en we zien dat, indien  $f''(x_0) \neq 0$ , asymptotisch de fout linear met  $h$  afneemt.



FIGUUR 6: De richting van de raaklijn (---) aan de grafiek (—) van de functie  $f$  is de afgeleide  $f'(x)$  die we willen benaderen. De richting van de koorde (- · - ·) in het rechter figuur doet dat veel beter dan die in het linker figuur.

**2.1B** Uit het linker plaatje in figuur 6 zien we dat het quotiënt in (17) niet zo'n goede keuze is. Veel betere resultaten kunnen we verwachten van de in het rechter plaatje gesuggereerde benadering. We schrijven

$$f'(x_0) = \frac{f(x_0+h) - f(x_0-h)}{2h} + R_2(h). \quad (19)$$

Als de derde afgeleide van  $f$  bestaat en continu is vinden we door  $f(x_0+h)$  en  $f(x_0-h)$  volgens Taylor te ontwikkelen in  $x_0$  de restterm

$$R_2(h) = -\frac{h^2}{6}f'''(\xi), \quad x_0-h < \xi < x_0+h. \quad (20)$$

De fout gaat nu inderdaad veel sneller naar nul voor  $h \rightarrow 0$ , nl. indien  $f'''(x_0) \neq 0$ , asymptotisch met het kwadraat van de stapgrootte  $h$ .

**2.1C** Laten we voor verschillende stapgrootten  $h$ , rekenend met zes cijfers, eens de afgeleiden van de exponentiële functie  $e^x$  in  $x_0 = 1$  bepalen met behulp van (19).

$x$	$e^x$
0.8000	2.22554
0.9000	2.45960
0.9900	2.69123
0.9990	2.71556
0.9999	2.71801
1.0000	2.71828
1.0001	2.71855
1.0010	2.72100
1.0100	2.74560
1.1000	3.00417
1.2000	3.32012

$h$	benadering	fout
0.2000	2.73644	0.01816
0.1000	2.72285	0.00457
0.0100	2.71850	0.00022
0.0010	2.72000	0.00172
0.0001	2.70000	0.01828

We zien dat voor dalende  $h$  de benadering aanvankelijk beter wordt, maar daarna verslechtert. De reden voor dit verschijnsel is niet moeilijk aan te geven. Voor bijv.  $h = 0.0010$  hebben we de volgende situatie:

$$\begin{aligned} f(x_0 + h) &= 2.72100 \\ f(x_0 - h) &= \frac{2.71556}{0.00544} \end{aligned}$$

Bij het aftrekken vallen de eerste drie cijfers tegen elkaar weg, en we houden er slechts drie over. Dit heeft de volgende consequenties:

De oorspronkelijke waarden van  $f$  zijn belast met afrondfouten. De absolute grootte van deze fouten is hoogstens  $5 \cdot 10^{-6}$ , en de relatieve fouten zijn dus hoogstens  $5 \cdot 10^{-6} / 2.7 \approx 2 \cdot 10^{-6}$ .

De absolute afrondfout in het verschil kan dus oplopen tot  $10 \cdot 10^{-6}$  (namelijk bij tegengestelde tekens van de afrondfouten), hetgeen een relatieve fout in het verschil veroorzaakt van

$$10 \cdot 10^{-6} / 0.0054 \approx 2 \cdot 10^{-3},$$

dat is dus 1000 maal zo groot als de maximale relatieve fout in de  $f$ -waarden. Dit effect wordt natuurlijk sterker naarmate  $h$  kleiner wordt.

*We zijn hier aangelopen tegen bekend probleem in de numerieke wiskunde: Het aftrekken van twee bijna even grote getallen.*

Voor de benaderingsformule (19) betekent dit dat de te bereiken nauwkeurigheid gelimiteerd is. Enerzijds wil je  $h$  zo klein mogelijk hebben om de rest klein te maken; anderzijds mag  $h$  niet klein zijn om bovengenoemde reden. Er zal dus een optimale waarde van  $h$  zijn waarbij de uitkomst zo nauwkeurig mogelijk is. Aangezien de afrondfout onbekend is kunnen we deze waarde niet bepalen. Wel kan men een majorant, d.w.z. bovengrens, van de fout minimaliseren, en wel als volgt.

Laat de waarden van  $f$  belast zijn met afrondfouten hoogstens  $\epsilon$ . Dit levert dan een bijdrage van hoogstens  $S_2(h) = 2\epsilon/2h$  tot de absolute fout in de numerieke benadering voor  $f'(x_0)$ . De tweede bijdrage is de restterm  $R_2(h)$  als in (20). De totale fout wordt dus gemajoreerd door de grootte  $\varphi(h)$  die gegeven wordt door

$$\varphi(h) = \frac{\epsilon}{h} + \frac{h^2}{6}m \quad (21)$$

waarin  $m$  een majorant is voor  $|f'''(x)|$  op een voldoende grote maar ook weer niet al te grote omgeving  $[x_0 - h_0, x_0 + h_0]$  van  $x_0$ . Deze functie heeft een minimum  $\sqrt[3]{\frac{9}{8}\epsilon^2 m}$  voor  $h = \sqrt[3]{3\epsilon/m}$ . Wanneer voor deze  $h$  geldt  $h \leq h_0$  dan beschouwt men hem als de optimale.

In het voorbeeld met  $e^x$  vinden we zo met  $x_0 = 1$ ,  $\epsilon = 5 \cdot 10^{-6}$ ,  $h_0 = 0.2$ ,  $m = 3.32$  dat  $h = 0.0165$  en  $\phi(h) = 0.00045$ .

**2.1D** De tweede afgeleide kunnen we benaderen door met (19) eerst  $f'(x_0 + \frac{1}{2}h)$  en  $f'(x_0 - \frac{1}{2}h)$  te benaderen, en hiermede vervolgens de benadering

$$\frac{f'(x_0 + \frac{1}{2}h) - f'(x_0 - \frac{1}{2}h)}{h} \quad (22)$$

voor  $f''(x_0)$  te construeren. Met enig rekenen en Taylor vinden we

$$f''(x_0) = \frac{f(x_0 + h) - 2f(x_0) + f(x_0 - h)}{h^2} + R_3(h), \quad R_3(h) = -\frac{h^2}{12}f^{(4)}(\xi). \quad (23)$$

**2.1E** De effecten van afrondfouten zijn nu nog ernstiger. Analoog aan 2.1C vinden we nu  $S_3(h) = \frac{4\epsilon}{h^2}$  als bovengrens voor het effect van de afrondfouten in de  $f$ -waarde op de numerieke benadering voor  $f''(x_0)$  en dat loopt snel op bij dalende  $h$ .

Formule (23) levert met  $h = 0.01$  en de tabel van 2.1C voor  $f''(1)$  een waarde 2.70 en voor  $h = 0.1$  een waarde 2.72.

Men kan een analogon van (21) opstellen, en vindt dan een optimale foutschatting 0.0042 voor  $h \approx 0.1$ .

**2.1F** Formules (19) en (23) zijn equidistante formules. Een generalisatie van (23) wordt gegeven door

$$f''(x_0) = \frac{h_1 f(x_0 + h_2) - (h_1 + h_2)f(x_0) + h_2 f(x_0 - h_1)}{h_1 h_2 (h_1 + h_2)/2} + R_4(h_1, h_2) \quad (24)$$

waarin

$$R_4(h_1, h_2) = -\frac{h_2^2 f'''(\xi_2) - h_1^2 f'''(\xi_1)}{3(h_1 + h_2)}, \quad \xi_2 \in (x_0, x_0 + h_2) \text{ en } \xi_1 \in (x_0 - h_1, x_0). \quad (25)$$

Merk op dat i.h.a. slechts  $R_4(h_1, h_2) = \mathcal{O}(\max\{h_1, h_2\})$ .

## 2.2 Nauwkeuriger formules

De formules in sectie 2.1 zijn enigzins adhoc afgeleid en hebben geen erge hoge nauwkeurigheidssorde.

Een systematischer manier, waarmee ook hogere nauwkeurigheidssorden bereikt kunnen worden (als we afzien van het effect van afrondfouten), is de te differentiëren functie te benaderen met een Lagrange interpolatiepolynoom  $p_n$  op steunpunten  $\{x_k : 0 \leq k \leq n\}$  en dit polynoom te differentiëren.

M.b.v. de Lagrange representatie uit (6) krijgt men zo

$$f^{(i)}(a) \approx p_n^{(i)}(a) = \sum_{k=0}^n L_{kn}^{(i)}(a) f(x_k) \quad (26)$$

waarbij de coëfficiënten  $L_{kn}^{(i)}(a)$  dus onafhankelijk zijn van  $f$ . Duidelijk is dat de benadering (26) slechts zinnig is voor  $i \leq n$ . Aangezien (26) exact is voor  $f \in P_n$ , kunnen

we voor vaste  $\{x_k : 0 \leq k \leq n\}$ ,  $i$  en  $a$ , de factoren  $L_{kn}^{(i)}(a)$  behalve door differentiatie van de Lagrange coëfficiënten ook bepalen door een stelsel van  $n + 1$  vergelijkingen op te lossen dat verkregen wordt door voor  $f(x)$  achtereenvolgens  $1, x, \dots, x^n$  te nemen. Men kan eenvoudig aantonen dat dit stelsel een unieke oplossing heeft.

Voor equidistante puntenrijen  $\{x_k\}$ ,  $i = n$  en met  $h$  de afstand tussen opvolgende steunpunten krijgt men de volgende formules:

$$f'(x_0 + \frac{1}{2}h) \approx [f(x_1) - f(x_0)]/h \quad (27)$$

$$f''(x_1) \approx [f(x_2) - 2f(x_1) + f(x_0)]/h^2 \quad (28)$$

$$f'''(x_1 + \frac{1}{2}h) \approx [f(x_3) - 3f(x_2) + 3f(x_1) - f(x_0)]/h^3 \quad (29)$$

etc., welke voor voldoende gladde  $f$  een restterm  $\mathcal{O}(h^2)$  hebben.

Formules met een restterm van hogere orde kan men verkrijgen door de techniek uit deze sectie met  $n > i$  toe te passen. Een voorbeeld van een dergelijke formule met  $n = 4$  en  $i = 1$  is

$$f'(a) \approx \frac{1}{12h}[f(a - 2h) - 8f(a - h) + 8f(a + h) - f(a + 2h)] \quad (30)$$

met restterm  $\mathcal{O}(h^4)$ . De coëfficiënt voor  $f(a)$  blijkt hier nul te zijn.

### 2.3 Experimenteel gevonden functiewaarden

Wanneer men een functie  $f$  wil differentiëren waarvan men slechts experimenteel gevonden functiewaarden kent dan kunnen de tot dusverre ontwikkelde differentiatieformules wel heel slecht uit de bus komen.

**Voorbeeld 2.3.1** Stel dat men van de functie  $f(x) = 1 + x$  door meting functiewaarden bepaalt voor  $x_k = \frac{1}{2} + \frac{k}{20}$  ( $k = 0, \dots, 4$ ), en dat de resultaten van onderstaande tabel worden verkregen. Men ziet dat de meetfout circa 0.005 bedraagt. Benadert men nu  $f'(0.60)$  met behulp van de formule (30) dan vindt men  $f'(0.60) \approx 1.10$ , hetgeen 10% fout is.

$x$	$f(x)$
0.50	1.503
0.55	1.548
0.60	1.596
0.65	1.655
0.70	1.697

Een andere aanpak is nu dat we  $f$  niet benaderen met een 4-de graads polynoom en dat differentiëren (zoals in (30)), maar dat we een lineaire functie zo goed mogelijk aan de gevonden functiewaarden aanpassen, en dan die lineaire functie differentiëren. Kiezen we  $\|g\| = \sqrt{\sum_{k=0}^4 |g(x_k)|^2}$  als norm op functies gedefiniëerd op de punten  $x_k$ , dan levert elementair rekenwerk dat de beste benadering van  $f$  door een lineaire functie gegeven wordt door

$$x \mapsto \frac{\alpha_0\beta_1 - \alpha_1\beta_0}{\alpha_0\alpha_2 - \alpha_1^2}x + \frac{\alpha_2\beta_0 - \alpha_1\beta_1}{\alpha_0\alpha_2 - \alpha_1^2}$$

waarbij  $\alpha_0 = n + 1$ ,  $\alpha_1 = \sum x_k$ ,  $\alpha_2 = \sum x_k^2$ ,  $\beta_0 = \sum f(x_k)$  en  $\beta_1 = \sum f(x_k)x_k$ . Het minimaliseren van de afstand in bovenstaande norm is bekend als de *kleinste kwadraten*

*methode.* Het differentiëren van het verkregen lineair polynoom levert nu de benadering  $f'(0.6) \approx \frac{\alpha_0\beta_1 - \alpha_1\beta_0}{\alpha_0\alpha_2 - \alpha_1^2} = 0.99$  hetgeen een heel wat beter resultaat is dan wat eerder werd verkregen.

Op soortgelijke wijze kan men, als men wat meer punten heeft,  $f$  aanpassen aan een polynoom van hogere graad, en dat dan differentiëren.

### 3 Foutschattingen bij numerieke processen

#### 3.1 Fouten

**3.1A** Vrijwel altijd zal het resultaat  $a$  van een numerieke benadering van een grootheid  $\alpha$  niet exact zijn. De mate waarin  $a$  van  $\alpha$  afwijkt, kunnen we uitdrukken in

- de *fout*:  $\alpha - a$
- de *absolute fout*:  $|\alpha - a|$
- de (*absolute*) *relatieve fout*:  $|\frac{\alpha-a}{\alpha}|$

De fout zelf is doorgaans onbekend (want anders zou men het ware antwoord  $\alpha$  ook kunnen vinden). Men is vaak al blij als men een majorant (d.w.z. bovengrens) voor de absolute fout heeft. Zo weet men dat de absolute fouten in de sinustabel uit 1.1A hoogstens  $\frac{1}{2}10^{-6}$  zijn, maar de (gewone, absolute, relatieve) fout zelf kent men niet. Het is dan ook onjuist te zeggen dat deze tabel de sinuswaarden oplevert met een *fout* van  $\frac{1}{2}10^{-6}$ : de waarden worden opgeleverd met een *onbetrouwbaarheid* van  $\frac{1}{2}10^{-6}$ . Wij zullen dit onderscheid verder steeds aanhouden.

Onder *onbetrouwbaarheid* (respectievelijk *relatieve onbetrouwbaarheid*) verstaan we dus bovengrenzen (majoranten) voor de absolute (respectievelijk relatieve) fout.

**Eigenschap.** (bewijs dit zelf):

- $a$  heeft onbetrouwbaarheid  $\epsilon \Leftrightarrow a = \alpha + \epsilon', |\epsilon'| \leq \epsilon,$
- $a$  heeft relatieve onbetrouwbaarheid  $\epsilon \Leftrightarrow a = \alpha(1 + \epsilon'), |\epsilon'| \leq \epsilon.$

#### 3.1B Foutfilosofie.

Het is doorgaans niet interessant om een onbetrouwbaarheid erg nauwkeurig op te geven: de mededeling, dat een relatieve onbetrouwbaarheid 1.1% is, is nauwelijks inhoudsrijker dan dat hij 1% is. Het is natuurlijk wel van belang, te weten of hij 1% dan wel 5% is, want nu is er onderhand sprake van een verschil in orde van grootte. Hetzelfde geldt natuurlijk voor de (absolute) onbetrouwbaarheid.

**Voorbeeld 3.1.1** We bekijken weer dezelfde sinus-tabel uit 1.1A, maar nu met wat meer decimalen om in het vervolg afrondfouten te kunnen verwaarlozen.

$x$	$\sin(x)$
$35^\circ$	0.57357644
$36^\circ$	0.58778525
$37^\circ$	0.60181502
$38^\circ$	0.61566148
$39^\circ$	0.62932039

Passen wij nu lineaire interpolatie toe met steunpunten  $36^\circ$  en  $38^\circ$  om  $\sin(37^\circ)$  te benaderen, dan vinden we

$$\sin(37^\circ) \approx 0.60172337,$$

waarin dus een fout zit van 0.00009165. Als bovengrens voor de fout vinden we met stelling 1.1.2:  $|\text{fout}| \leq \frac{1}{8}(4\pi/360)^2 \sin(38^\circ) = 0.00009377$ . Als majorant voor de fout is deze informatie veel te nauwkeurig. Immers, we zien hieruit dat de 7 in de vierde decimaal van het berekende antwoord misschien wel een 6 of een 8 had moeten zijn, en dat alle volgende cijfers betekenisloos zijn. Dezelfde conclusie had men kunnen trekken als er een onbetrouwbaarheid van 0.0001 was gegeven.

Een ander aspect dat door het bovenstaande voorbeeld goed geïllustreerd wordt, is dat van de overbodige decimalen. Dikwijls geven rekenapparaten (computers, pocket-calculators, etc.) de uitkomst standaard in de -voor hen- grootst mogelijke nauwkeurigheid, dat wil zeggen, in het maximaal beschikbare aantal cijfers. Echter doordat een approximatieproces is toegepast, gebeurt het maar zelden dat al deze cijfers inderdaad ook juist zijn. Als men, zoals in het bovenstaande voorbeeld, een antwoord heeft met een onbetrouwbaarheid van ongeveer 0.0001, dan is het aan te bevelen om in een verslag niet meer dan 4 of 5 cijfers achter de komma te vermelden, dus in bovenstaand geval 0.6017 of 0.60172; men handhaaft zo de onbetrouwbaarheid op ongeveer hetzelfde niveau (deze wordt 0.00012 resp. 0.0001), en valt de lezer niet lastig met overbodige cijfers.

Het is *geen* goede gewoonte om (zoals nogal eens aanbevolen wordt) een uitkomst af te ronden op een zodanig aantal cijfers dat het resultaat gelijk is aan het ware antwoord afgerond op dit aantal cijfers. Dat zou bijvoorbeeld betekenen dat een uitkomst 0.4345 met een onbetrouwbaarheid 0.0001, moet worden afgerond op 0.43, met een onbetrouwbaarheid 0.0046. De onbetrouwbaarheid kan zo dus aanzienlijk toenemen.

Het is ook *geen* goede gewoonte het aantal correcte cijfers te hanteren als maat voor de nauwkeurigheid van een uitkomst. Immers, een uitkomst  $a = 1.9999$  voor een grootte  $\alpha = 2$  heeft geen enkel correct cijfer, maar is wel redelijk nauwkeurig. Omgekeerd geldt natuurlijk wel, dat een groot aantal correcte cijfers (afgezien van eventuele nullen vooraan) een kleine relatieve onbetrouwbaarheid garandeert.

Zelfs indien men over een uitkomst met een kleine relatieve onbetrouwbaarheid beschikt (zeg  $10^{-10}$ ) dan is het nog niet altijd verstandig om de eerste tien relevante cijfers (dat zijn de cijfers die overblijven na weglating van eventuele nullen vooraan) ook alle te geven. Dikwijls is de lezer niet geïnteresseerd in een dergelijke nauwkeurigheid.

### Samenvattend:

- Geef van het resultaat niet meer cijfers dan er nodig zijn om de onbetrouwbaarheid ongeveer op hetzelfde niveau te handhaven.
- Geef van het resultaat niet meer cijfers dan er gewenst worden.
- Geef van de relatieve of absolute onbetrouwbaarheid één, of hoogstens twee, relevante cijfers.

### 3.1C De bronnen van fouten kunnen globaal worden ingedeeld in drie groepen:

- modelfouten
- representatiefouten
- afrondfouten

**Modelfouten** ontstaan als men tracht de fysische (of economische, biologische, technische) realiteit te beschrijven door middel van een wiskundige vergelijking; vaak worden hierbij kleine (en soms ook niet zo kleine) effecten verwaarloosd. In het algemeen zijn deze fouten vrij klein bij problemen uit de fysica, maar aanzienlijk groter bij problemen uit de economie of de biologie. Deze fouten vallen eigenlijk buiten het kader van de numerieke wiskunde.

**Representatiefouten** ontstaan doordat men het wiskundige probleem, dat als model van het fysische (economische, enz.) probleem dient, en waarvan men de oplossing niet kan bepalen, vervangt door een naburig probleem dat men wel kan oplossen. Voorbeeld: men vervangt een afgeleide door een differentiequotient, zie hoofdstuk 2. Heel

vaak hebben representatiefouten een bepaalde structuur. Een belangrijke tak van de numerieke wiskunde houdt zich bezig met het opsporen en uitbuiten van deze structuur. Zie verder sectie 3.4 en sectie 3.5.

**Afrondfouten** ontstaan doordat men rekenapparaten slechts in eindige precisie rekenen. Zie verder sectie 3.2.

### 3.2 Afrondfouten

**3.2A** Afrondfouten ontstaan doordat rekenapparaten niet met de verzameling  $\mathbb{R}$  van reële getallen werken, maar met een eindige deelverzameling  $Q$  van  $\mathbb{Q}$ . De elementen van  $Q$  worden *machinegetallen* genoemd. Gewoonlijk heeft deze verzameling de volgende vorm

$$Q = \{y = \pm.d_1 \dots d_s \times \beta^e : d_i \in \{0, \dots, \beta - 1\}, d_1 > 0 \text{ tenzij } y = 0, e \in \mathbb{Z}, m \leq e \leq M\}.$$

Hierbij heet  $\pm.d_1 \dots d_s$  de *mantisse*, en  $e$  de *exponent* (t.o.v. het grondtal  $\beta$ ).

Computers rekenen meestal *binair*, d.w.z.  $\beta = 2$ , hoewel er ook wel *hexadecimale arithmetiek* toegepast wordt, d.w.z.  $\beta = 16$ . In het binaire geval zijn typische waarden (*IEEE standaard*)  $s = 24$ ,  $m = -125$  en  $M = 128$  (enkelvoudige precisie), danwel  $s = 53$ ,  $m = -1021$  en  $M = 1024$  (dubbele precisie). Rekenmachines rekenen meestal *decimaal* ( $\beta = 10$ ). Typische waarden zijn  $s = 10$ ,  $m = -98$ ,  $M = 100$ .

Onder het *bereik van een machine* verstaat men de verzameling

$$\{a \in \mathbb{R} \mid q_1 \leq |a| \leq q_2\} \cup \{0\}$$

waarin  $q_1$  en  $q_2$  het kleinste resp. grootste positieve getal in  $Q$  is. Getallen binnen dit bereik kunnen door de machine in goede relatieve precisie worden weergegeven. Als bij een berekening een uitkomst ontstaat met absolute waarde groter dan  $q_2$  dan spreekt men van *overflow*. Vaak wordt er dan een foutmelding gegenereerd, echter sommige rekenmachines leveren zonder verdere waarschuwing  $\pm q_2$  af. Bij een uitkomst ongelijk aan 0 in  $] -q_1, q_1 [$  (*underflow*) wordt doorgaans op 0 afgerond zonder verdere waarschuwing.

**Eigenschap.** (Bewijs dit zelf) Zij  $\alpha \neq 0$  een getal dat binnen het bereik van de machine ligt. Zij  $a$  het getal in  $Q$  (of een van de twee getallen in  $Q$ ) het dichtst bij  $\alpha$ . Als de machine  $\alpha$  afrondt tot  $a$ , spreken we van een correcte afronding. Dan heeft  $a$  een relatieve fout van hoogstens

$$\frac{1}{2} \beta^{1-s}.$$

In het algemeen zal na iedere bewerking  $+$ ,  $-$ ,  $\times$ ,  $/$  en  $\sin$ ,  $\cos$ ,  $\exp$ , enzovoort, van getallen in  $Q$  het exacte resultaat niet meer in  $Q$  zitten. Er zal dus (opnieuw) afgerond moeten worden. Op deze wijze ontstaat in grotere programma's een hele keten van afrondfouten, die bovendien op elkaar inwerken. Hierin valt meestal geen structuur te ontdekken. Het beste dat men kan bereiken is het geven van een bovengrens voor de uiteindelijke afrondfout (dat is het totale effect van alle voorafgaande afrondingen) in het eindresultaat.

**3.2B** In praktijk komt het relatief zelden voor dat de begrenzingen van het machinereik een rol spelen; bovendien, als dit wel het geval is, kan men zich daartegen wapenen door het inbouwen van controles in het programma.

We nemen daarom aan dat het volgende geldt: Er bestaat een (klein) getal  $\bar{\xi}$ , de *relatieve machine precisie*, en een functie  $\text{rd}$  (round) :  $\mathbb{R} \rightarrow Q$  zo dat voor elk reëel getal  $\alpha$  geldt

$$\text{rd}(\alpha) = \alpha(1 + \xi) \quad \text{met } |\xi| \leq \bar{\xi} \quad (31)$$

en bovendien voor elke  $a \in Q$

$$\text{rd}(a) = a, \quad (32)$$

met de afspraak, dat getallen die men van buiten in het rekenapparaat invoert, door  $\text{rd}$  worden afgebeeld in  $Q$ . We nemen verder aan dat de opteloperatie ‘+’ die de machine uitvoert op getallen  $a$  en  $b$  uit  $Q$  als resultaat levert

$$a \text{ ‘+’ } b = \text{rd}(a + b) = (a + b)(1 + \xi), \quad |\xi| \leq \bar{\xi}$$

en analoog voor de *machineoperaties* ‘-’, ‘×’, ‘/’. D.w.z., we veronderstellen een optimale arithmetiek. Om dit te realiseren wordt er intern met één of meerdere extra cijfers gerekend.

**3.2C** Om te laten zien hoe afrondfoutenanalyse werkt, bekijken we nu het voorbeeld van het optellen van een aantal machinegetallen  $a_0, \dots, a_n$ , in deze volgorde. Allereerst tellen we  $a_0$  en  $a_1$  op:

$$a_0 \text{ ‘+’ } a_1 = (a_0 + a_1)(1 + \xi_1), \quad |\xi_1| \leq \bar{\xi}.$$

Hierbij wordt  $a_2$  opgeteld, enzovoort:

$$\begin{aligned} (a_0 \text{ ‘+’ } a_1) \text{ ‘+’ } a_2 &= ((a_0 + a_1)(1 + \xi_1) + a_2)(1 + \xi_2) = S_2^*, \quad |\xi_i| \leq \bar{\xi}, \quad i = 1, 2; \\ ((a_0 \text{ ‘+’ } a_1) \text{ ‘+’ } a_2) \text{ ‘+’ } a_3 &= (((a_0 + a_1)(1 + \xi_1) + a_2)(1 + \xi_2) + a_3)(1 + \xi_3) = S_3^* \\ & \quad |\xi_i| \leq \bar{\xi}, \quad i = 1, 2, 3. \end{aligned}$$

(Om de uitdrukking nog enigszins overzichtelijk te houden, nemen we  $n = 3$ .) We krijgen zo:

$$\begin{aligned} S_3^* &= a_0(1 + \xi_1)(1 + \xi_2)(1 + \xi_3) + a_1(1 + \xi_1)(1 + \xi_2)(1 + \xi_3) + \\ & \quad + a_2(1 + \xi_2)(1 + \xi_3) + a_3(1 + \xi_3). \end{aligned}$$

We merken nu op dat er zeker een  $\xi_4$  is met  $|\xi_4| \leq \bar{\xi}$  zo dat

$$(1 + \xi_1)(1 + \xi_2)(1 + \xi_3) = (1 + \xi_4)^3$$

(ga na), en evenzo is er een  $\xi_5$  met  $|\xi_5| \leq \bar{\xi}$ , zo dat

$$S_3^* = a_0(1 + \xi_4)^3 + a_1(1 + \xi_4)^3 + a_2(1 + \xi_5)^2 + a_3(1 + \xi_3). \quad (33)$$

Bijgevolg, als  $S_3$  de ware som  $a_0 + a_1 + a_2 + a_3$  voorstelt, geldt

$$S_3^* - S_3 = a_0[(1 + \xi_4)^3 - 1] + a_1[(1 + \xi_4)^3 - 1] + a_2[(1 + \xi_5)^2 - 1] + a_3\xi_3$$

Nu geldt  $[(1 + \xi_4)^3 - 1] = 3\xi_4 + 3\xi_4^2 + \xi_4^3 \approx 3\xi_4$  en  $[(1 + \xi_5)^2 - 1] = 2\xi_5 + \xi_5^2 \approx 2\xi_5$  zodat, indachtig de foutfilosofie uit 3.1B,

$$|S_3^* - S_3| \lesssim (3|a_0| + 3|a_1| + 2|a_2| + |a_3|)\bar{\xi}.$$

Het teken “ $\lesssim$ ” wordt uitgesproken als “kleiner dan of ongeveer gelijk aan”.

**3.2D** We constateren dat het ons weinig gebaat heeft al de optredende  $\xi$ 's door middel van indices van elkaar te onderscheiden; dit blijkt vaak het geval te zijn bij afrondfoutenanalyse. We laten daarom deze indices meestal weg. Dit betekent dan wel dat een en hetzelfde symbool  $\xi$  een aantal verschillende waarden kan representeren, zelfs binnen één formule. We maken daarom de volgende notatieafspraken:

Bij afrondfoutenanalyse zal  $\xi$  steeds een getal voorstellen met absolute waarde kleiner dan of gelijk aan  $\bar{\xi}$ , maar iedere keer dat  $\xi$  voorkomt, kan hij een andere waarde hebben.

Met deze afspraak krijgen we voor de berekende som  $S_3^*$

$$\begin{aligned} S_3^* &= (((a_0 + a_1)(1 + \xi) + a_2)(1 + \xi) + a_3)(1 + \xi) = \\ &= a_0(1 + \xi)^3 + a_1(1 + \xi)^3 + a_2(1 + \xi)^2 + a_3(1 + \xi) \end{aligned} \quad (34)$$

(vergelijk (33)). We hebben hier al stilzwijgend gebruik gemaakt van een volgende notatieafpraak:

Een relatie als  $f(\xi) = g(\xi)$ , met  $f$  en  $g$  functies van  $\mathbb{R}$  naar  $\mathbb{R}$ , moet in de afrondfouten analyse gelezen worden als:  
bij elke  $\xi_1$ ,  $|\xi_1| \leq \bar{\xi}$ , is er een  $\xi_2$ ,  $|\xi_2| \leq \bar{\xi}$ , zodat  $f(\xi_1) = g(\xi_2)$ .

Merk op, dat volgens deze afspraak het “=” teken geen symmetrische relatie weergeeft. Het symbool  $\xi$  is dus vergelijkbaar met de ordesymbolen  $o$  en  $\mathcal{O}$  van Landau. Zo geldt er (ga na!)

$$\xi = 2\xi$$

maar niet

$$2\xi = \xi.$$

Dit is vergelijkbaar met

$$\mathcal{O}(x^2) = \mathcal{O}(x) \quad (x \rightarrow 0)$$

maar niet

$$\mathcal{O}(x) = \mathcal{O}(x^2) \quad (x \rightarrow 0).$$

Ga ook na, dat de volgende uitdrukkingen juist zijn:

$$\begin{aligned} \xi - \xi &= 2\xi \\ (1 + \xi)(1 + \xi) &= (1 + \xi)^2. \end{aligned}$$

**3.2E** Bij afrondfoutenanalyse komen we blijkbaar allerlei machten van  $(1 + \xi)$  tegen, ook negatieve en niet gehele machten, en we willen weten, hoever die van 1 af kunnen liggen.

In eerste benadering (dat wil zeggen voor  $\xi \rightarrow 0$ ,  $k$  vast) geldt

$$(1 + \xi)^k - 1 \simeq k\xi \quad (35)$$

(eerste orde Taylor benadering). De vraag is voor welke waarden van  $k$  dit bij een vaste waarde van  $\xi$  nog bruikbaar is. Hierover doet het volgende lemma een uitspraak.

**Lemma 3.2.1** *Zij  $k \in \mathbb{R}$ . Dan geldt*

$$1 + kx \leq (1 + x)^k \leq 1 + kx(1 + kx) \quad \text{voor } k \geq 1 \quad \text{en} \quad |kx| \leq 1 \quad (36)$$

$$1 - kx \leq \frac{1}{(1 + x)^k} \leq 1 - kx(1 - 2kx) \quad \text{voor } k \geq 1 \quad \text{en} \quad |kx| \leq \frac{1}{2}. \quad (37)$$

**Bewijs.** Noem  $f(x) = (1 + x)^k - (1 + kx)$ . Dan geldt  $f(0) = 0$ ,  $f'(x) \geq 0$  voor  $x \geq 0$  en  $k \geq 1$  en  $f'(x) \leq 0$  voor  $-1 \leq x \leq 0$  en  $k \geq 1$ . Hiermee volgt de linker ongelijkheid in (36).

Aangaande de rechter ongelijkheid in (36) merken we op dat  $1 + x \leq e^x$  voor alle  $x \in \mathbb{R}$ , zodat voor  $k \geq 0$ ,  $|kx| \leq 1$  geldt

$$\begin{aligned} (1 + x)^k &\leq e^{kx} = 1 + kx + \frac{k^2 x^2}{2} \left( 1 + \frac{kx}{3} + \frac{k^2 x^2}{3 \cdot 4} + \frac{k^3 x^3}{3 \cdot 4 \cdot 5} + \dots \right) \\ &\leq 1 + kx + \frac{k^2 x^2}{2} \left( 1 + \frac{1}{3} + \frac{1}{3^2} + \frac{1}{3^3} + \dots \right) = 1 + kx + \frac{3}{4} k^2 x^2, \end{aligned}$$

en dus is de rechter ongelijkheid in (36) bewezen.

De linker ongelijkheid in (37) volgt uit  $(1 + x)^k (1 - kx) \leq e^{kx} e^{-kx} = 1$ . De rechter ongelijkheid volgt m.b.v. (36) uit

$$\frac{1}{(1 + x)^k} \leq \frac{1}{1 + kx} = 1 - kx + \frac{k^2 x^2}{1 + kx}.$$

□

We zien hieruit dat als  $k\xi \ll 1$  (en dit is heel vaak het geval) de relatie (35) in de zin van de foutfilosofie bevredigend vervuld is, en evenzo de relatie

$$\frac{1}{(1 + \xi)^k} - 1 \approx k\xi \quad (38)$$

(merk op dat de  $\xi$ 's links en rechts tegengesteld teken hebben).

### 3.2F Toepassingen.

Ga na dat in het algemeen geldt:

**Eigenschap.** Zij  $S_n^*$  de door het rekenapparaat geleverde benadering voor de som  $S_n = \sum_{i=0}^n a_i$ , gesommeerd vanaf  $a_0$ . Dan

$$|S_n^* - S_n| \lesssim (n|a_0| + \sum_{i=1}^n (n+1-i)|a_i|)\bar{\xi}.$$

**Opgave 3.2.1** Zij  $a_0, \dots, a_n$  een stel machinegetallen en  $S_n = \sum_{i=0}^n a_i$ . Men kan deze getallen uiteraard in allerlei volgordes optellen, en krijgt dan steeds verschillende berekende waarden  $S_n^*$ . Toon aan dat de onbetrouwbaarheid in  $S_n^*$  zoals deze gegeven wordt door eigenschap 3.2F minimaal is als men de getallen optelt in volgorde van opklimmende absolute waarde.

**Opgave 3.2.2** Men kan de getallen  $a_0, \dots, a_n$  ook als volgt optellen:

- bereken  $a'_0 = a_0 + a_1$ ,  $a'_1 = a_2 + a_3$ ,  $a'_2 = a_4 + a_5$ , enzovoort (als  $n$  even is, denk dan nog een term  $a_{n+1} = 0$  toegevoegd),
- bereken  $a''_0 = a'_0 + a'_1$ ,  $a''_1 = a'_2 + a'_3$ ,  $a''_2 = a'_4 + a'_5$ , enzovoort,
- enzovoort.

Toon aan dat

$$|S_n^* - S_n| \lesssim \left( \sum_{i=0}^n |a_i| \right) \lceil 2 \log(n+1) \rceil \bar{\xi},$$

waarbij hier  $\lceil a \rceil = \min\{j \in \mathbb{Z} \mid j \geq a\}$ .

Toon aan, dat deze onbetrouwbaarheid in elk geval veel kleiner is dan die in eigenschap 3.2F, wanneer men daarin de ongunstigste sommatievolgorde kiest. Ga na, dat deze onbetrouwbaarheid ook veel kleiner zal zijn dan bij de in opgave 3.2.1 genoemde gunstigste sommatie volgorde, tenzij de getallen  $a_i$  zeer uiteenlopend van grootte zijn.

**Opgave 3.2.3** Zij  $I_n^*$  de door het rekenapparaat geleverde benadering voor het inproduct  $I_n = \sum_{i=1}^n a_i b_i$ , gesommeerd vanaf  $a_1 b_1$ . Bewijs dat

$$|I_n^* - I_n| \lesssim (n|a_1 b_1| + \sum_{i=2}^n (n+2-i)|a_i b_i|) \bar{\xi}.$$

### 3.3 Foutvoortplanting

**3.3A** Zowel door onnauwkeurigheid in de ingangsgegevens van een berekening als door onnauwkeurigheid in tussenresultaten ten gevolge van afrondfouten ontstaat er een fout in het eindresultaat. Bij het bepalen van de onbetrouwbaarheid in het eindresultaat is het heel gebruikelijk om slechts eerste orde effecten in rekening te brengen, net als we al in 3.2D deden. Bewijs zelf dat hiervoor de onderstaande rekenregels gelden:

Laat  $a$  en  $b$  benaderingen zijn voor  $\alpha$  en  $\beta$  met relatieve onbetrouwbaarheden  $\epsilon_1$  en  $\epsilon_2$  respectievelijk,  $\epsilon_1 \ll 1$  en  $\epsilon_2 \ll 1$ . Dan geldt voor de exacte bewerkingen (dus zonder afrondfout) dat het resultaat een relatieve onbetrouwbaarheid heeft volgens de volgende tabel.

Operatie	Relatieve onbetrouwbaarheid in het resultaat
$ab$	$\approx \epsilon_1 + \epsilon_2$
$1/b$	$\approx \epsilon_2$
$a/b$	$\approx \epsilon_1 + \epsilon_2$
$a + b$	$\max(\epsilon_1, \epsilon_2)$ als $\alpha$ en $\beta$ hetzelfde teken hebben
$a + b$	$\frac{\epsilon_1 + \epsilon_2}{\delta}$ als $\beta = (-1 + \delta)\alpha$ , $\delta \in ]0, 1]$
$a^k$	$\approx k\epsilon_1$ als $k\epsilon_1 \ll 1$
$\sqrt{a}$	$\approx \frac{1}{2}\epsilon_1$
$a^b$	$\approx  \beta \epsilon_1 +  \beta \log \alpha \epsilon_2$ als $ \beta \epsilon_1 +  \beta \log \alpha \epsilon_2 \ll 1$
$\sin(a)$	$\epsilon_1$ als $ \alpha  \leq \frac{\pi}{2}$

N.B. Als de bewerkingen op het rekenapparaat worden uitgevoerd, dan komen de afrondfouten er nog bij.

**Voorbeeld 3.3.1** Laat men als resultaat van een computerberekening getallen  $a$  en  $b$  verkregen hebben zodat  $a = \alpha(1 + k\xi)$ ,  $b = \beta(1 + m\xi)$ . Dan is  $\frac{a}{b} \approx \frac{\alpha}{\beta}(1 + (k + m)\xi)$ . Wordt echter het quotiënt  $\frac{a}{b}$  op de computer bepaald, dan krijgt men

$$a \text{ ' / ' } b = \frac{a}{b}(1 + \xi) \approx \frac{\alpha}{\beta}(1 + (k + m)\xi)(1 + \xi) \approx \frac{\alpha}{\beta}(1 + (k + m + 1)\xi).$$

**3.3B** Uit het bovenstaande schema blijkt, dat bij de meeste elementaire bewerkingen en standaardfuncties de relatieve fouten in de operand(en) vermenigvuldigd met een beperkte factor in het resultaat terecht komen. Een duidelijke uitzondering op deze regel is het aftrekken van twee bijna gelijke getallen; hierbij is de factor waarmee de relatieve fouten in de operanden worden vermenigvuldigd, omgekeerd evenredig met de relatieve afstand tussen de operanden. (Vergelijk ook 2.1C en 2.1E).

### 3.4 Representatiefouten

**3.4A** Zowel bij interpolatie als bij numerieke differentiatie hebben we gezien dat de representatiefout de volgende vorm had: bekende factor vermenigvuldigd met een hogere afgeleide in een onbekend punt. Deze situatie is essentieel anders dan die bij afrondfouten: deze hebben geen structuur, en zelfs hun teken is niet bekend; ze hebben slechts een vaste bovengrens (zie 3.2A). Representatiefouten daarentegen hebben geen “vaste” bovengrens (bij interpolatie en differentiatie dalen ze als de stapgrootte daalt), hoewel de som van representatiefout en afrondfout vaak wel een minimum heeft (zie bijvoorbeeld 2.1C). In het vervolg van dit hoofdstuk zullen we de structuur van de representatiefout gebruiken om al rekenend de fout te schatten, en m.b.v. deze schatting het resultaat te verbeteren. We ontwikkelen deze algemeen bruikbare techniek aan de hand van het voorbeeld betreffende numerieke differentiatie.

**3.4B Schatting van de fout op grond van bekend foutgedrag: Numerieke differentiatie.** Met  $R(h) = f'(x_0) - \frac{f(x_0+h) - f(x_0-h)}{2h}$  weten we dat  $R(h)$  ongeveer evenredig is met  $h^2$  als we  $h$  naar 0 laten gaan (zie (20)) mits  $f'''(x_0) \neq 0$ . Dus met de notaties  $I = f'(x_0)$  en  $T(h) = \frac{f(x_0+h) - f(x_0-h)}{2h}$  geldt

$$I - T\left(\frac{1}{2}h\right) \approx \frac{1}{4}[I - T(h)]. \quad (39)$$

Door  $I - T(h)$  te schrijven als  $I - T\left(\frac{h}{2}\right) + T\left(\frac{h}{2}\right) - T(h)$ , volgt hier eenvoudig uit dat

$$I - T\left(\frac{1}{2}h\right) \approx \frac{1}{3}\left[T\left(\frac{1}{2}h\right) - T(h)\right]. \quad (40)$$

Dus als men voor zekere waarden van  $h$  de grootheden  $T(h)$  en  $T\left(\frac{1}{2}h\right)$  berekent dan kan men met (40) de fout in  $T\left(\frac{1}{2}h\right)$  schatten.

**3.4C** Wanneer men dit wil toepassen rijst natuurlijk wel de vraag naar de betrouwbaarheid van deze foutschatting, dus de vraag of bij de beschouwde waarde van  $h$  er wel in redelijke mate evenredigheid is van  $R(h)$  met  $h^2$ . Enig vertrouwen hierin kan men verwerven door te kijken naar de berekenbare waarde

$$q(h) = \frac{T(2h) - T(h)}{T(h) - T\left(\frac{1}{2}h\right)}. \quad (41)$$

Deze grootheid is immers gelijk aan

$$\frac{R(2h) - R(h)}{R(h) - R(\frac{1}{2}h)} \quad (42)$$

en dit is 4 als  $R(h)$  evenredig is met  $h^2$ .

**Voorbeeld 3.4.1**  $f(x) = e^x$ ,  $x_0 = 1$

$h$	$q(h)$	$R(h)$	$\frac{1}{3}(T(h) - T(2h))$
.2	4.03	-.018158	-.018304
.1	4.008	-.004533	-.004542
.05	4.002	-.001133	-.001133

Men ziet dat  $q(h) \approx 4$  en dat  $\frac{1}{3}(T(h) - T(2h))$  nauwkeurig overeenstemt met de in dit geval bekende ware fout  $R(h)$ .

Uit het feit dat  $q(h) \approx 4$  volgt niet noodzakelijk dat  $R(h) \approx h^2$  en dus dat (40) geldt. Men kan zich natuurlijk meer vertrouwen verwerven door  $q(h)$  voor verschillende waarden van  $h$  te berekenen en te kijken of er steeds ongeveer 4 uitkomt. Men kan zelfs zeggen dat (40) met zekerheid geldt als  $q(\eta) \approx 4$  voor alle  $\eta \in ]0, h_0]$ , zoals men ziet uit stelling 3.4.1 In de praktijk is deze voorwaarde natuurlijk niet te verifiëren.

**Stelling 3.4.1** *Zij  $\lim_{h \rightarrow 0} T(h) = I$ . Laten er getallen  $h_0 > 0$ ,  $a, b > 1$  bestaan, zodat voor alle  $h \in ]0, h_0]$  geldt  $a \leq q(h) \leq b$ . Dan geldt voor alle  $h \in ]0, 2h_0]$*

$$\frac{1}{b-1} \leq \frac{I - T(\frac{1}{2}h)}{T(\frac{1}{2}h) - T(h)} \leq \frac{1}{a-1} \quad (43)$$

en zo voor  $\lambda \geq 0$ ,

$$1 + \lambda(1-b) \leq \frac{I - T(\frac{h}{2}) - \lambda(T(\frac{h}{2}) - T(h))}{I - T(\frac{h}{2})} \leq 1 + (1-a)\lambda,$$

hetgeen een bruikbaar resultaat oplevert als  $\frac{1}{b-1} \approx \lambda \approx \frac{1}{a-1}$ , vgl. (40).

**Bewijs.** f Voor  $h \in ]0, 2h_0]$  geldt:

$$\begin{aligned} \frac{1}{b} &\leq \frac{T(\frac{1}{2}h) - T(\frac{1}{4}h)}{T(h) - T(\frac{1}{2}h)} \leq \frac{1}{a}, \\ \frac{1}{b^2} &\leq \frac{T(\frac{1}{4}h) - T(\frac{1}{8}h)}{T(h) - T(\frac{1}{2}h)} \leq \frac{1}{a^2} \quad (\text{wegens } \frac{T(\frac{1}{4}h) - T(\frac{1}{8}h)}{T(h) - T(\frac{1}{2}h)} = \\ &= \frac{T(\frac{1}{4}h) - T(\frac{1}{8}h)}{T(\frac{1}{2}h) - T(\frac{1}{4}h)} \frac{T(\frac{1}{2}h) - T(\frac{1}{4}h)}{T(h) - T(\frac{1}{2}h)}), \\ \frac{1}{b^3} &\leq \frac{T(\frac{1}{8}h) - T(\frac{1}{16}h)}{T(h) - T(\frac{1}{2}h)} \leq \frac{1}{a^3}, \quad \text{etc.} \end{aligned}$$

Optellen levert

$$\sum_{i=1}^{n-1} \frac{1}{b^i} \leq \frac{T(\frac{1}{2}h) - T(2^{-n}h)}{T(h) - T(\frac{1}{2}h)} \leq \sum_{i=1}^{n-1} \frac{1}{a^i}$$

Door de limiet voor  $n \rightarrow \infty$  te nemen volgt het gestelde.  $\square$

In de numerieke praktijk moet men voor kleine waarden van  $h$  wel voorzichtig zijn met de hierboven geschetste methode om vertrouwen te verwerven. Immers, door het rekenen op de computer werkt men eigenlijk niet met de grootheid  $R$ , maar met  $R+$  afrondfout. Zoals we zagen in hoofdstuk 2 zal in het bijzonder in het voorbeeld van numerieke differentiatie voor kleine waarden van  $h$  de afrondfout gaan overheersen, zodat van de gewenste asymptotiek niets overblijft.

### 3.5 Extrapolatieprocessen

**3.5A** Vaak gaat men bij toepassing van het voorgaande nog een stapje verder. Men merkt dan op dat (40) een *schatting* geeft voor de fout (dit wat anders dan een *boven-grens*). Door deze schatting aan te brengen als correctie krijgt men

$$T(\frac{1}{2}h) + \frac{1}{3}[T(\frac{1}{2}h) - T(h)] \quad (44)$$

als benadering voor  $I$ . Merk op dat de aanpak om een benadering te verbeteren door te corrigeren met een schatting van de fout ook al werd gevolgd in de vraagstukken 2 en 3 behorend bij hoofdstuk 1. Verwacht kan worden dat (44) een veel betere benadering zal geven dan  $T(\frac{1}{2}h)$  zelf. I.h.a. heeft men echter geen informatie meer over de fout in de nieuwe benadering.

**3.5B** Men noemt het toevoegen van de correctieterm aan  $T(\frac{1}{2}h)$  in (44) wel *Richardson extrapolatie*. Men kan deze benaming begrijpen vanuit de volgende gedachtengang:  $T(h)$  gedraagt zich blijkbaar ongeveer als  $I + ch^2$ . Als men nu  $T(h_1)$  en  $T(h_2)$  kent voor gegeven  $h_1$  en  $h_2$  en men wil  $T(h)$  voor een andere waarde van  $h$  kennen, dan zou men een functie  $\phi$  kunnen bepalen,  $\phi(h) = p + qh^2$ , zodat  $\phi(h_1) = T(h_1)$ ,  $\phi(h_2) = T(h_2)$ , en dan  $\phi(h)$  nemen als benadering voor  $T(h)$ . Dit is dus een soort interpolatie opgave. Wegens  $\lim_{h \rightarrow 0} T(h) = I$  zal dan  $\phi(0)$  een benadering voor  $I$  zijn. Kiest men  $h_1$  en  $h_2 = \frac{1}{2}h_1$ , dan vindt men voor  $\phi(0)$  juist de uitdrukking in (44). Maar omdat 0 niet tussen  $h_1$  en  $h_2$  ligt is er wel sprake van extrapolatie.

**3.5C Romberg schema's.** Richardson extrapolatie is de eerste stap van een veel verdergaand proces. We illustreren dit proces weer aan de hand van ons numerieke differentiatie voorbeeld. Als in 3.4B  $f$   $2m + 1$  keer continu differentiëerbaar is op een omgeving van  $x_0$ , dan zijn er coëfficiënten  $c_1, \dots, c_{m-1}$  zo dat

$$I = T(h) + c_1 h^2 + c_2 h^4 + \dots + c_{m-1} h^{2m-2} + \mathcal{O}(h^{2m}) \quad (45)$$

voor  $h \rightarrow 0$ . Merk op, dat de coëfficiënten  $c_i$  met  $i < m$  niet van  $m$  afhangen. Als zelfs  $f \in C^\infty[x_0 - h_0, x_0 + h_0]$  en  $h < h_0$  dan kunnen we schrijven

$$I \sim T(h) + c_1 h^2 + c_2 h^4 + c_3 h^6 + \dots \quad (46)$$

Het  $\sim$  wijst erop dat we ons niet bezig houden met convergentie van de "som" in het rechterlid; men noemt (46) wel een asymptotische expansie. We kunnen hierin de  $c_i$  zelfs expliciet aangeven, maar dat blijkt niet nodig te zijn.

Schrijven we (45) op met  $\frac{1}{2}h$  in plaats van  $h$ , dan vinden we uiteraard

$$I = T\left(\frac{1}{2}h\right) + \frac{1}{4}c_1h^2 + \frac{1}{16}c_2h^4 + \dots \quad (47)$$

Vermenigvuldig dit laatste met 4 en trek (45) er vanaf. Er komt dan

$$3I = 4T\left(\frac{1}{2}h\right) - T(h) + d_2h^4 + d_3h^6 + \dots$$

zodat

$$I = \frac{1}{3}\left[4T\left(\frac{1}{2}h\right) - T(h)\right] + d'_2h^4 + d'_3h^6 + \dots \quad (48)$$

Men ziet dus dat, terwijl  $T(h)$  een benadering voor  $I$  geeft met een fout  $\approx c_1h^2$  voor  $h \rightarrow 0$  (als  $c_1 \neq 0$ ),  $\frac{1}{3}[4T(\frac{1}{2}h) - T(h)]$  een benadering voor  $I$  geeft met een fout  $\approx d'_2h^4$  voor  $h \rightarrow 0$  (als  $d'_2 \neq 0$ ), en aangezien  $h^4$  veel sneller naar 0 gaat dan  $h^2$  verwacht men dat voor voldoende kleine  $h$ ,  $\frac{1}{3}[4T(\frac{1}{2}h) - T(h)]$  een veel betere benadering voor  $I$  zal geven dan  $T(\frac{1}{2}h)$ . Merk op dat we deze benadering ook al gevonden hadden in (44), echter daar zonder foutschatting.

**3.5D** Uitgaande van de expansie (48) i.p.v. (45) kunnen we bovenstaande methode herhalen. Zo voortgaand komen we tot het volgende schema: Kies een waarde voor  $h$ . Definieer

$$T_{i0} = T(2^{-i}h), \quad T_{ij} = \frac{4^j T_{i,j-1} - T_{i-1,j-1}}{4^j - 1} \quad (49)$$

We geven deze grootheden in onderstaand schema weer, met daarnaast een schematische aanduiding hoe ze samenhangen.

$h$	$T_{00}(= T(h))$			$-T_{i-1,j-1}$
$\frac{1}{2}h$	$T_{10}(= T(\frac{1}{2}h))$	$T_{11}$		+
$\frac{1}{4}h$	$T_{20}(= T(\frac{1}{4}h))$	$T_{21}$	$T_{22}$	
$\frac{1}{8}h$	$T_{30}(= T(\frac{1}{8}h))$	$T_{31}$	$T_{32}$	
$\frac{1}{16}h$	$T_{40}(= T(\frac{1}{16}h))$	$T_{41}$	$T_{42}$	
				$4^j T_{i,j-1} \xrightarrow{\div(4^j - 1)} T_{ij}$

(men noemt dit wel een *Romberg schema*). Zolang men nog beschikt over een ontwikkeling van de vorm (45), dat wil zeggen, zolang  $j < m$ , zal iedere kolom sneller convergeren dan zijn voorganger.

**3.5E Andere dalingsfactoren.** In het voorgaande baseerden we het Romberg schema op de stapgrootte rij

$$h, \frac{1}{2}h, \frac{1}{4}h, \dots,$$

de zgn. *Romberg rij*. Deze rij gaat nogal snel naar nul, en voor het geval dat  $T$  een numerieke differentiatieresultaat voorstelt worden de waarden zo nogal onnauwkeurig (zie 2.1C).

Er is echter niets op tegen te werken met een rij  $h, \alpha h, \alpha^2 h, \dots$  met  $\alpha > \frac{1}{2}$ , bijvoorbeeld  $\alpha = 0.7$  of  $0.8$ . I.v.m. numerieke integratie (zie hoofdstuk 4) gebruikt men nogal eens de zogenaamde *Bulirsch rij*:

$$h, \frac{2}{3}h, \frac{1}{2}h, \frac{1}{3}h, \frac{1}{4}h, \frac{1}{6}h, \frac{1}{8}h, \frac{1}{12}h, \dots$$

Men moet voor deze rijen uiteraard wel de formules aanpassen.

In het algemeen, als  $h_0, h_1, h_2, h_3, \dots$  een rij van stapgrootten is met  $h_0 > h_1 > \dots > 0$  dan berekent men de waarden  $T_{ij}$  van het Romberg schema dat hierop gebaseerd is recursief als volgt

$$\begin{aligned} T_{i0} &= T(h_i) && \text{voor } i = 0, 1, 2, \dots \\ T_{ij} &= \frac{h_{i-j}^2 T_{i,j-1} - h_i^2 T_{i-1,j-1}}{h_{i-j}^2 - h_i^2} && \text{voor iedere } i, j \in \mathbf{N}, i \geq j. \end{aligned} \quad (50)$$

(Ga na dat dit, voor het geval  $h_i = 2^{-i}h$ , overeenstemt met de formule in (49).) Aannemende een expansie als in (46) kan bewezen worden dat

$$I - T_{ij} = h_{i-j}^2 h_{i-j+1}^2 \cdots h_i^2 (\tilde{c}_j + \mathcal{O}(h_{i-j}^2)) \quad (51)$$

voor zekere constanten  $\tilde{c}_j$ . Overigens verloopt de afleiding van dit algemene schema en het bewijs van (51) via een generalisatie van het idee uit 3.5B van 2-punts extrapolatie tot meerpunts extrapolatie, en niet via een generalisatie van de aanpak uit 3.5C/3.5D.

**3.5F Superlineaire convergentie.** De essentiële overeenkomst tussen de schema's uit 3.5D en 3.5E is dat het quotiënt  $h_{i+1}/h_i$  tussen twee opeenvolgende stapgroottes van 1 weg begrensd is. Voor dergelijke schema's geldt, dat de convergentie langs de diagonaal  $T_{00}, T_{11}, T_{22}, \dots$  (en ook langs de codiagonalen) nog sneller is dan die van de kolommen. Hiervoor geldt de volgende stelling (bewijs in Bijlage B).

### Stelling 3.5.1

Veronderstel dat  $I$  een ontwikkeling heeft als in (46), en dat  $h_{i+1}/h_i \leq \alpha < 1$ . Dan zijn er getallen  $y_n$  zo dat  $|I - T_{nn}| \leq y_n$  en  $\lim_{n \rightarrow \infty} y_{n+1}/y_n = 0$ .  $\square$

Het convergentiegedrag dat in stelling 3.5.1 wordt beschreven noemt men wel *superlineair*: De bovengrens  $y_n$  daalt met een factor die naar 0 gaat. Dit is in tegenstelling met wat men mag verwachten van de convergentie in de kolommen; voor de Rombergrij  $h_i = h_0/2^i$  geldt bijvoorbeeld:  $|I - T_{ij}|/|I - T_{i-1,j}| \approx \frac{1}{4^{j+1}}$  voor  $i \rightarrow \infty$ . Een rij benaderingen die een dergelijk gedrag vertoont, noemt men wel *lineair convergent*.

**3.5G Foutschatting.** Het resultaat van stelling 3.5.1 wordt soms gebruikt om uit het gedrag van de rij  $T_{11}, T_{22}, \dots, T_{nn}$  conclusies te trekken aangaande de fout in  $T_{nn}$  (bijvoorbeeld: "Als  $T_{n-1,n-1}$  en  $T_{n,n}$  in  $k$  cijfers overeenstemmen, dan zijn die  $k$  cijfers vast wel goed"). De wiskundige fundering van dit soort strategieën is nogal wankel.

Veronderstel dat men een Romberg schema als in 3.5D gemaakt heeft, met  $0 \leq i, j \leq n$ . We beschouwen dan het punt rechtsonder in het schema (immers, we verwachten dat daar de "beste" getallen staan):

$$\begin{array}{ccc} & & T_{n-2,n-2} \\ & & T_{n-1,n-2} \quad T_{n-1,n-1} \\ T_{n,n-2} & T_{n,n-1} & T_{n,n} \end{array}$$

Het opbouwen van de  $(n-1)$ ste kolom berust op de gedachte dat voor de  $(n-2)$ de kolom geldt  $(I - T_{i,n-2})/(I - T_{i+1,n-2}) \approx 4^{n-1}$ . Door nu het quotiënt  $(T_{n-2,n-2} - T_{n-1,n-2})/(T_{n-1,n-2} - T_{n,n-2})$  te berekenen kunnen we ons hierin een zeker vertrouwen verschaffen; namelijk, dit quotiënt moet dan ook ongeveer de waarde  $4^{n-1}$  hebben. We

zitten nu eigenlijk in dezelfde situatie als in sectie 3.4, alleen met hogere ordes. Analoog aan 3.4C zullen we er nu vertrouwen in hebben dat

$$I - T_{n,n-2} \approx \frac{T_{n-1,n-2} - T_{n,n-2}}{4^{n-1} - 1} \quad (52)$$

en dus dat  $|I - T_{n,n-1}| \ll |I - T_{n,n-2}|$ .

Vervolgens bekijken we de  $(n - 1)$ ste kolom. Hiervan vermoeden we op grond van (45) dat er geldt  $(I - T_{i,n-1})/(I - T_{i+1,n-1}) \approx 4^n$ . Echter, omdat we slechts over twee elementen van deze kolom beschikken, bestaat er geen methode om dit vermoeden ook te toetsen. Daarom kunnen we in het equivalent van (52),

$$I - T_{n,n-1} \approx \frac{T_{n-1,n-1} - T_{n,n-1}}{4^n - 1} \quad (53)$$

en dus  $|I - T_{n,n}| \ll |I - T_{n,n-1}|$  aanzienlijk minder vertrouwen hebben dan in (52). Tenslotte is m.b.v. de Romberg tabel een foutschatting voor  $T_{nn}$  helemaal niet meer te geven.

**3.5H Voorbeelden.** We geven nu twee voorbeelden van een Romberg schema voor de eerste afgeleide van  $f(x) = -1/\tan(x)$  voor  $x = 0.04$  (een niet zo gunstig punt voor deze functie wegens de nabijheid van de singulariteit in  $x = 0$ ). De functiewaarden komen uit een tabel met 8 cijfers achter de komma.

**Voorbeeld 3.5.1** Gebruikt zijn de formules (49) met  $h = 0.0128$  (de kolom  $h_i$  geeft de gebruikte argumenten bij  $T_{i,0}$ ; elk element is bepaald uit zijn linksboven- en linksonder buur):

$h_i$	$T_{i0}$	$T_{i1}$	$T_{i2}$	$T_{i3}$
0.0128	696.6346914			
		623.4601726		
0.0064	641.7538023		625.3455055	
		625.2276722		625.3334226
0.0032	629.3592047		625.3336114	
		625.3269902		625.3334448
0.0016	626.3350438		625.3334744	
		625.3330438		625.3334257
0.0008	625.5835438		625.3334260	
		625.3334021		
0.0004	625.3959375			

Het proces is niet voortgezet na de kolom  $T_{i3}$  omdat deze kolom niet monotoon is (wat men wel zou verwachten omdat  $I - T_{i3} \approx ch_i^8$ ). We wijten dit aan de afrondfouten in de functiewaarden. Niettemin is het resultaat al erg bevredigend, gezien het ware antwoord  $I = 625.33344002 \dots$  (in 8 cijfers achter de komma).

Hieronder geven we  $R_{ij} = I - T_{ij}$  aan:

$R_{i0}$	$R_{i1}$	$R_{i2}$	$R_{i3}$
-71.301			
	1.873267		
-16.420		-0.0120645	
	0.105778		0.0000174
-4.026		-0.0001714	
	0.006450		-0.0000048
-1.002		-0.0000074	
	0.000396		0.0000143
-0.250		0.0000140	
	0.000038		
-0.061			

Men ziet dat de getallen  $R_{i0}$  mooi met een factor 4 dalen, de getallen  $R_{i1}$  met een factor 16 (op de laatste na), en dat de eerste twee getallen  $R_{i2}$  een verhouding 70 (verwachting 64) hebben. In de onderste getallen  $R_{i1}$  en  $R_{i2}$  manifesteren zich kennelijk ook al afrondfouten.

**Voorbeeld 3.5.2** Gebruikt zijn de formules (50) voor de Bulirsch rij met  $h = 0.0256$

$h$	$T_{i0}$	$T_{i1}$	$T_{i2}$	$T_{i3}$	$T_{i4}$	$T_{i5}$	$T_{i6}$
0.0256	1058.9377906						
		495.5225783					
0.0192	812.4436352		640.1425224				
		603.9875364		624.4282997			
0.0128	696.6346914		626.6381123		625.3572216		
		620.9754683		625.2991640		625.3330932	
0.0096	663.5337813		625.4479360		625.3339415		625.3334398
		624.3298191		625.3317679		625.3334344	
0.0064	641.7538023		625.3481040		625.3335585		
		625.0935328		625.3333435			
0.0048	634.4649344		625.3349836				
		625.2746209					
0.0032	629.3592047						

Men ziet, dat nu veel minder kleine waarden van  $h$  nodig zijn. Maar vooral ziet men, dat het meest rechtse getal een zeer nauwkeurige benadering is van het ware antwoord

$I = 625.33344002$ . Schrijven we weer  $R_{ij} = I - T_{ij}$ , dan krijgen we

$R_{i0}$	$R_{i1}$	$R_{i2}$	$R_{i3}$	$R_{i4}$	$R_{i5}$	$R_{i6}$
-493.60						
	129.810					
-187.11		-15.2090				
	22.347		0.905140			
-71.30		-1.3046		-0.023782		
	4.368		0.034276		0.0003468	
-38.20		-0.1145		-0.000501		0.0000002
	1.004		0.001672		0.0000056	
-16.42		-0.0146		-0.000008		
	0.240		0.000096			
-9.13		-0.0015				
	0.059					
-4.03						

Uit (51) leiden we af dat  $Q_{ij} := \frac{R_{ij}}{R_{i+2j}} = \frac{h_{i-j}^2 h_{i-j+1}^2 (\tilde{c}_j + \mathcal{O}(h_{i-j}^2))}{h_{i+1}^2 h_{i+2}^2 (\tilde{c}_j + \mathcal{O}(h_{i+2-j}^2))} \rightarrow 4^{j+1}$  ( $i \rightarrow \infty$ ) (indien  $\tilde{c}_j \neq 0$ ). De hieronder gegeven waarden voor  $Q_{ij}$  voor  $j = 0, \dots, 4$

$Q_{i0}$	$Q_{i1}$	$Q_{i2}$	$Q_{i3}$	$Q_{i4}$
6.08				
	29.7			
4.90		132.8		
	22.3		541	
4.34		89.0		2798
	18.2		355	
4.18		74.2		
	17.1			
4.08				

lijken nog niet erg op de getallen 4, 16, 64, 256 en 1024 die men mag verwachten als de asymptotiek in bevredigende mate bereikt is. Kennelijk is hiervoor de waarde  $h$  te groot. Wel ziet men fraai de superlineaire convergentie langs de diagonaal:  $|R_{i+1,i+1}/R_{ii}|$  neemt achtereenvolgens de waarden 0.299, 0.117, 0.0595, 0.0263, 0.0146 en 0.0006 aan.

**3.5I** Tenslotte merken we nog op, dat men ook Romberg schema's kan vormen als de ontwikkeling (45) of (46) ook andere dan alleen de even machten van  $h$  bevat. Wel moet men natuurlijk van tevoren weten wat voor soort ontwikkeling men heeft en de factoren (dat zijn de  $4^j$  in (49)) daaraan aanpassen. Bevat de ontwikkeling bijvoorbeeld alleen positieve gehele machten van  $h$  dan moet men in (50) de factoren  $h_l^2$  vervangen door  $h_l$ .

**Literatuur** over Romberg schema's en andere extrapolatieprocessen:

C. Brezinski, "Acceleration de la convergence", Springer 1977.

## 4 Numerieke integratie

### 4.0 Inleiding

Bij de berekening van integralen  $\int_a^b f(t) dt$  zullen we vaak op functies stuiten waarvan moeilijk of in het geheel niet expliciet een primitieve aan te geven is. Om dan toch een benadering voor de integraal te vinden kunnen we bijv. de integrand door een interpolatiepolynoom vervangen en dit polynoom vervolgens integreren. Hoe goed zo'n benadering is, zullen we uitgebreid onderzoeken. Tevens zal blijken dat we hiervoor het interpolatiepolynoom niet expliciet behoeven op te stellen.

In hoofdstuk 1 gebruikten we lineaire interpolatie om een aantal wezenlijke aspecten van interpolatie te demonstreren. Analoog zullen we in dit hoofdstuk de trapeziumregel als voorbeeld te gebruiken voor de i.h.a. efficiëntere methoden die men in de praktijk gebruikt.

Formules om integralen benaderd te berekenen worden vaak *kwadratuurformules* genoemd.

### 4.1 De trapeziumregel

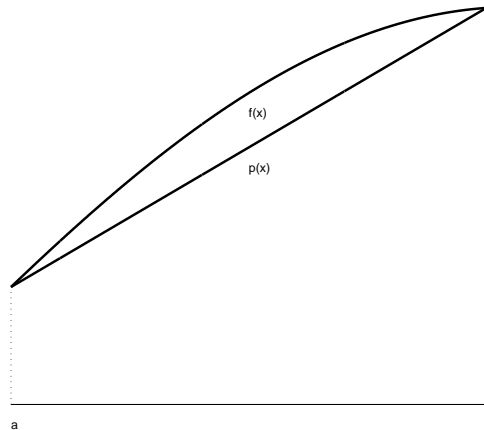
**4.1A** Laat  $a$  en  $b$  twee punten in  $\mathbb{R}$  zijn,  $a < b$ , en  $f$  een continue functie op  $[a, b]$ . Laat  $p$  het lineaire interpolatiepolynoom van  $f$  op  $a$  en  $b$  zijn (zie 2):

$$p(x) = \frac{x-b}{a-b}f(a) + \frac{x-a}{b-a}f(b)$$

We gebruiken nu  $\int_a^b p(x) dx$  als benadering voor  $\int_a^b f(x) dx$ :

$$\int_a^b p(x) dx = \frac{b-a}{2}(f(a) + f(b)) \quad (54)$$

(ga na). De benadering  $\frac{b-a}{2}(f(a) + f(b))$  voor  $\int_a^b f(x) dx$  heet de *trapeziumregel*: Het oppervlak onder de grafiek van  $f$  tussen  $a$  en  $b$  wordt benaderd met het oppervlak van het trapezium met als hoekpunten  $(a, 0)$ ,  $(b, 0)$ ,  $(b, f(b))$  en  $(a, f(a))$  (zie figuur 7).



FIGUUR 7: De trapeziumregel: het oppervlak onder de grafiek van de functie  $f$  wordt benaderd door het oppervlak van het trapezium dat gevormd wordt door de grafiek van het lineaire interpolatiepolynoom  $p$  van  $f$ .

**4.1B De restterm.** In 1.1B vonden we als restterm bij lineaire interpolatie op  $a$  en  $b$

$$f(x) - p(x) = \frac{1}{2}(x-a)(x-b)f''(\xi(x)) \quad (55)$$

waarbij we  $\xi(x)$  schrijven in plaats van  $\xi$  om de afhankelijkheid van  $\xi$  ten aanzien van  $x$  nog eens te benadrukken. Hoewel we geen resultaten hebben over de continuïteit van  $x \mapsto \xi(x)$ , noch hoeven aan te nemen dat  $f''$  continu is, is het duidelijk dat  $x \mapsto \frac{1}{2}(x-a)(x-b)f''(\xi(x))$  een continue functie is: immers, het is het verschil van twee continue functies. Het rechterlid van (55) is dus zeker integreerbaar.

Sterker nog,  $x \mapsto f''(\xi(x))$  is continu op  $[a, b]$ . Immers,  $\phi \equiv f - p$  is continue differentieerbaar, en  $\phi(a) = \phi(b) = 0$ . Omdat  $f''(\xi(x)) = 2\phi(x)/((x-a)(x-b))$  hebben we continuïteit van  $x \mapsto f''(\xi(x))$  op  $(a, b)$  en omdat

$$\lim_{x \rightarrow a} \frac{2\phi(x)}{(x-a)(x-b)} = \frac{2}{a-b} \lim_{x \rightarrow a} \frac{\phi(x) - \phi(a)}{x-a} = \frac{2}{a-b} \phi'(a)$$

volgt continuïteit in  $a$ . Evenzo volgt continuïteit in  $b$ .

Integreren van (55) levert

$$\int_a^b f(x) dx - \frac{b-a}{2}(f(a) + f(b)) = \frac{1}{2} \int_a^b (x-a)(x-b)f''(\xi(x)) dx. \quad (56)$$

Deze restterm kan men schatten met behulp van de volgende stelling:

**Stelling 4.1.1** *Zij  $f \in C^2[a, b]$ . Dan is er een  $\eta \in [a, b]$  zo dat*

$$\int_a^b f(x) dx - \frac{b-a}{2}(f(a) + f(b)) = -\frac{(b-a)^3}{12} f''(\eta).$$

Het bewijs van deze stelling maakt gebruik van de volgende stelling.

**4.1C** Stelling 4.1.2 wordt de *tussenwaardenstelling van de integraalrekening* genoemd. De stelling maakt gebruik van tekenvastе functies: een functie  $g$  is *tekenvast* op  $[a, b]$  als of  $g(t) \geq 0$  voor alle  $t \in [a, b]$  of  $g(t) \leq 0$  voor alle  $t \in [a, b]$

**Stelling 4.1.2** *Zij  $f$  en  $g$  reëel-waardige integreerbare functies op  $[a, b]$ ,  $f$  is continu op  $[a, b]$ ,  $g$  is begrensd en tekenvast. Dan is er een  $\eta \in (a, b)$  zodat*

$$\int_a^b g(x)f(x) dx = f(\eta) \int_a^b g(x) dx$$

**Bewijs.** We mogen aannemen dat  $g \geq 0$  is op  $[a, b]$ . Anders vervangen we  $g$  door  $-g$ . Zij  $m = \min f(x)$ ,  $M = \max f(x)$ . Dan geldt,

$$m \int_a^b g(x) dx \leq \int_a^b g(x)f(x) dx \leq M \int_a^b g(x) dx.$$

Dus ligt  $d \equiv \int_a^b g(x)f(x) dx / \int_a^b g(x) dx$  tussen  $m$  en  $M$  en is er volgens de tussenwaardenstelling een  $\eta \in [a, b]$  zodat  $f(\eta) = d$ .

Merk op dat  $d = m$  of  $d = M$  alleen kan als  $f$  constant is op  $J \equiv \{t \in [a, b] \mid g(t) > 0\}$ . Als  $f$  niet constant is op  $J$ , is  $m < d < M$  en is er een  $\eta \in (a, b)$  met de gewenste eigenschap. Als  $f$  constant is op  $J$ , voldoet iedere  $\eta \in J$ , dus ook een  $\eta \in (a, b)$ .  $\square$

**Bewijs van Stelling 4.1.1.** Er geldt  $(x-a)(b-x) \geq 0$  op  $[a, b]$ . Verder hebben we al gezien dat  $x \mapsto f''(\xi(x))$  continu is op  $(a, b)$ . Dus volgt uit voorgaande stelling dat er een  $\eta \in [a, b]$  is waarvoor

$$\int_a^b (x-a)(b-x)f''(\xi(x)) dx = f''(\eta) \int_a^b (x-a)(b-x) dx = f''(\eta) \frac{(b-a)^3}{6}. \quad \square$$

**4.1D De gerepeteerde trapeziumregel.** Met stelling 4.1.1 vinden we eenvoudig een bovengrens voor de fout van de trapeziumregel:

$$\left| \int_a^b f(x) dx - \frac{b-a}{2}(f(a) + f(b)) \right| \leq \frac{(b-a)^3}{12} \max_{[a,b]} |f''(x)|. \quad (57)$$

Dikwijls zal deze bovengrens de maximaal toelaatbare fout (de zogenaamde tolerantie) overtreffen. Men verdeelt dan het interval in een aantal delen en past de trapeziumregel toe op elk deelinterval. Een op deze wijze gevormde kwadratuurformule noemt men een *gerepeteerde kwadratuurformule*.

**Voorbeeld 4.1.1** Verdeel het interval  $[a, b]$  in  $n$  stukken ter grootte  $h$ . Definiër  $x_i = a + ih$ . Voor de trapeziumregel op een deelinterval geldt

$$\int_{x_{i-1}}^{x_i} f(x) dx = \frac{h}{2}[f(x_i) + f(x_{i-1})] - \frac{1}{12}h^3 f''(\xi_i)$$

met  $x_{i-1} \leq \xi_i \leq x_i$ . Dus:

$$\int_a^b f(x) dx = h \left[ \frac{1}{2}f(x_0) + f(x_1) + \cdots + f(x_{n-1}) + \frac{1}{2}f(x_n) \right] - \frac{1}{12}h^3 \sum_{i=1}^n f''(\xi_i). \quad (58)$$

Wegens  $\min_i f''(\xi_i) \leq \frac{1}{n} \sum_{i=1}^n f''(\xi_i) \leq \max f''(\xi_i)$  is er (voor een continue functie  $f''$ ) een  $\xi \in [a, b]$  zodat

$$\sum_{i=1}^n f''(\xi_i) = n f''(\xi).$$

Met  $nh = b - a$  vinden we zo:

$$\int_a^b f(x) dx = T(h) - \frac{b-a}{12}h^2 f''(\xi) \quad (59)$$

waarin

$$T(h) = h \left[ \frac{1}{2}f(x_0) + f(x_1) + \cdots + f(x_{n-1}) + \frac{1}{2}f(x_n) \right]. \quad (60)$$

De aldus verkregen kwadratuurformule  $T(h)$  zullen we aanduiden als de  $n \times$  *gerepeteerde trapeziumregel*, waarbij dus, enigszins ongelukkig, de  $1 \times$  gerepeteerde trapeziumregel gelijk is aan de ongerepeteerde trapeziumregel.

Uit (59) volgt onmiddellijk dat het resultaat van de gerepeteerde trapeziumregel convergeert naar de integraal als  $f \in C^2[a, b]$ ; uit (59) krijgen we namelijk als bovengrens voor de fout

$$\left| \int_a^b f(x) dx - T(h) \right| \leq \frac{b-a}{12}h^2 \max_{[a,b]} |f''(x)|. \quad (61)$$

Maar ook voor wat minder nette functies treedt nog convergentie op. Zij  $f \in C^1[a, b]$ ; dan

$$\begin{aligned} \int_a^b f(x) dx &= \int_a^b \left(x - \frac{b+a}{2}\right)' f(x) dx = \left[ \left(x - \frac{b+a}{2}\right) f(x) \right]_{x=a}^{x=b} - \int_a^b \left(x - \frac{b+a}{2}\right) f'(x) dx \\ &= \frac{b-a}{2}(f(a) + f(b)) - \int_a^b \left(x - \frac{b+a}{2}\right) f'(x) dx, \end{aligned}$$

zo dat

$$\left| \int_a^b f(x) dx - \frac{b-a}{2}(f(a) + f(b)) \right| \leq \max |f'(x)| \int_a^b \left| x - \frac{b+a}{2} \right| dx.$$

Dit geeft een andere bovengrens voor de restterm bij de ongerepeteerde trapeziumregel dan (57):

$$\left| \int_a^b f(x) dx - \frac{b-a}{2}(f(a) + f(b)) \right| \leq \frac{(b-a)^2}{4} \max_{[a,b]} |f'(x)|. \quad (62)$$

Met deze grens vinden we in plaats van (61):

$$\left| \int_a^b f(x) dx - T(h) \right| \leq \frac{b-a}{4} h \max_{[a,b]} |f'(x)|. \quad (63)$$

We zien nu dus nog convergentie, maar mogelijkijkerwijze wel langzamer; immers, het rechterlid in (61) daalt evenredig met  $h^2$ , terwijl het rechterlid van (63) slechts evenredig met  $h$  is.

Tenslotte, veronderstel dat  $f$  slechts continu is op  $[a, b]$ . Dan voeren we de getallen  $t_i$ ,  $i = 0, \dots, n+1$  in als volgt:  $t_0 = x_0 = a$ ,

$$t_i = x_0 + (i - \frac{1}{2})h, \quad i = 1, \dots, n,$$

$$t_{n+1} = x_n = b.$$

Dan geldt:

$$h \left( \frac{1}{2} f(x_0) + f(x_1) + \dots + f(x_{n-1}) + \frac{1}{2} f(x_n) \right) = \sum_{i=0}^n (t_{i+1} - t_i) f(x_i).$$

Voor dit soort sommen geldt de volgende stelling:

**Stelling 4.1.3** *Zij  $f : [a, b] \rightarrow \mathbb{R}$  een Riemann-integreerbare functie. Beschouw uitdrukkingen van de vorm*

$$\sum_{i=0}^n (t_{i+1} - t_i) f(\tau_i)$$

waarin  $a = t_0 \leq \tau_0 \leq t_1 \leq \tau_1 \leq \dots \leq t_n \leq \tau_n \leq t_{n+1} = b$ . Dan geldt: voor elke  $\epsilon > 0$  is er een  $\delta > 0$  zo dat als  $t_{i+1} - t_i \leq \delta$ , voor alle  $i \leq n$ , dan

$$\left| \int_a^b f(x) dx - \sum_{i=0}^n (t_{i+1} - t_i) f(\tau_i) \right| \leq \epsilon.$$

Als de  $\{t_i\}$  en de  $\{\tau_i\}$  voldoen aan de voorwaarden van stelling 4.1.3, dan noemt men  $\sum_{i=0}^n (t_{i+1} - t_i) f(\tau_i)$  een *Riemann-som*.  $T(h)$  is dus zo'n Riemann-som.)

We vatten het bovenstaande samen:

**Stelling 4.1.4** *Zij  $f \in C[a, b]$ ; dan geldt*

$$(i) \int_a^b f(x) dx - T(h) \rightarrow 0 \text{ als } h \rightarrow 0,$$

Als  $f \in C^1[a, b]$ , dan

$$(ii) \left| \int_a^b f(x) dx - T(h) \right| \leq \frac{b-a}{4} h \max_{[a,b]} |f'(x)|,$$

Als  $f \in C^2[a, b]$ , dan

$$(iii) \left| \int_a^b f(x) dx - T(h) \right| \leq \frac{b-a}{12} h^2 \max_{[a,b]} |f''(x)|,$$

en bovendien is er een  $\xi \in [a, b]$  zo dat

$$(iv) \int_a^b f(x) dx - T(h) = -\frac{b-a}{12} h^2 f''(\xi).$$

□

## 4.2 Automatische integratie en Romberg schema's

**4.2A** Wanneer men de gerepeteerde trapeziumregel in praktische situaties wil gebruiken om een integraal met een zekere nauwkeurigheid te benaderen, dan moet men (bij voorkeur) over een majorant van  $f''$  beschikken, en dat kan men in de praktijk meestal niet.

Men zou daarom graag een proces willen hebben dat de gerepeteerde trapeziumregel voor een dalende rij waarden van  $h$  toepast en daarbij gaandeweg een indruk van de fout oplevert, om te stoppen als de gewenste nauwkeurigheid is bereikt. Om aan een dergelijk proces te komen grijpen we terug naar de restterm zoals die in (58) gegeven is en noteren deze als  $R(h)$ :

$$R(h) = -\frac{1}{12} h^3 \sum_1^n f''(\xi_i). \quad (64)$$

Wanneer we hierin  $h \sum_1^n f''(\xi_i)$  als Riemann som voor  $\int_a^b f''(t) dt = f'(b) - f'(a)$  beschouwen geldt dus

$$R(h) = -\frac{1}{12} h^2 [f'(b) - f'(a)] + o(h^2). \quad (65)$$

Blijkbaar is voor  $f'(a) \neq f'(b)$  en  $h$  klein de fout ongeveer evenredig met  $h^2$ . Bijgevolg kunnen we dezelfde foutschattingstechniek als in sectie 3.4 hanteren. Als we met  $I$  en  $T(h)$  de ware waarde van de integraal aanduiden, respectievelijk de benadering ervoor verkregen met de gerepeteerde trapeziumregel met stapgrootte  $h$ , dan geldt  $I - T(h) \approx 4[I - T(\frac{1}{2}h)]$ , en dus

$$I - T(\frac{1}{2}h) \approx \frac{1}{3} [T(\frac{1}{2}h) - T(h)] \quad (66)$$

en het rechterlid kan worden berekend. Dus als men een fout  $< \epsilon$  wenst gaat men door totdat  $|T(\frac{1}{2}h) - T(h)| < 3\epsilon$  en neemt dan  $T(\frac{1}{2}h)$  als de gewenste benaderingswaarde. Zodoende krijgt men dus een *automatisch integratieproces*. De vraag bij dit proces is natuurlijk of  $h$  al zo klein is dat  $R(h)$  al behoorlijk evenredig is met  $h^2$ , en dat vereist weer kennis van het verloop van  $f''$  die men doorgaans niet heeft. Als bijv.  $f''$  tekenvast is zullen Riemann sommen al voor betrekkelijk grote waarden van  $h$  een (in relatieve zin) goede benadering voor  $\int f''$  geven (een fout van 10% hierin zal ons niet deren; zie de foutschattingfilosofie, 3.1B); als echter  $|\int f''| \ll \int |f''|$  dan zal dit pas voor heel

kleine waarden van  $h$  het geval zijn; en voor  $\int f'' = 0$  (d.w.z.  $f'(a) = f'(b)$ , hetgeen in het bijzonder gebeurt bij periodieke functies  $f$ ) is het voor geen enkele waarde van  $h$  het geval. Overeenkomstig sectie 3.4 kan men enig *vertrouwen* in de evenredigheid van  $R(h)$  met  $h^2$  verwerven door er op te letten of

$$[T(2h) - T(h)]/[T(h) - T(\frac{1}{2}h)] \quad (67)$$

ongeveer de waarde 4 heeft. Als dit tijdens het halveringsproces enkele keren achtereen het geval is zal men hopen dat  $R(h)$  behoorlijk evenredig is met  $h^2$ , en daarmee dat (66) een goede foutschatting geeft.

#### Opmerking 4.2.1

- Het kan aanbevelenswaardig zijn het interval in 2 of meer delen op te delen als de integrand steile pieken heeft op een deel van het interval en zeer gladjes verloopt op een ander deel, omdat men bij glad verloop met veel grotere  $h$  (en dus minder werk) kan volstaan dan bij pieken.
- Een automatische manier van opdelen van het integratie interval  $[a, b]$  krijgt men als volgt: Zij  $T_0(a, h) = \frac{h}{2}[f(a) + f(a + h)]$  het trapezium resultaat voor het interval  $[a, a + h]$  en  $T_1(a, h) = \frac{h}{4}[f(a) + 2f(a + \frac{1}{2}h) + f(a + h)]$  het resultaat van de tweemaal gerepeteerde trapeziumregel. Verlaag dan  $h$  zo ver dat  $\frac{1}{3}|T_0(a, h) - T_1(a, h)| \leq \epsilon h/(b - a)$ , en accepteer  $T_1(a, h)$  als voldoende goede benadering voor  $\int_a^{a+h} f(x) dx$  (vergelijk (66)); beschouw nu  $a + h$  als nieuw beginpunt. Verschillende methoden voor automatische integratie berusten op deze gedachte, zij het dat ze vaak andere methoden dan de trapeziumregel als basis hebben (zie sectie 4.3); ze verschillen onderling ook in de strategie om de waarde van  $h$  voor het aan de beurt zijnde deelinterval te bepalen. Dergelijke methoden noemt men *variabele stap methoden*, ook wel *adaptieve methoden*.
- Uit (65) ziet men nog dat als  $f'(a) = f'(b)$  men kan verwachten dat  $R(h)$  sneller naar 0 gaat dan  $\sim h^2$ , hetgeen betekent dat  $\frac{1}{3}(T(h) - T(2h))$  geen zinnige schatting van  $R(h)$  geeft.

#### Voorbeeld 4.2.1

$$f(x) = 100((e^{x-1} - 1) \sin x)^2, \quad a = 0, \quad b = 1.$$

$n$	$T(h)$	$R(h)$	$\frac{T(h)-T(2h)}{3}$	$\frac{T(4h)-T(2h)}{T(2h)-T(h)}$	$\frac{T(h)-T(2h)}{15}$
1	0	1.89080656			
2	1.77923834	0.11156822	.59307945		.11861589
4	1.88397718	0.00682938	.03491295	16.987	.00698259
8	1.89038207	0.00042449	.00213496	16.353	.00042699
16	1.89078005	0.00002651	.00013266	16.093	.00002653
32	1.89080489	0.00000167	.00000828	16.022	.00000166
64	1.89080644	0.00000012	.00000052	16.026	.00000010

De vierde kolom suggereert dat de fout evenredig is met  $h^4$ . In 4.2B zullen we zien dat dit inderdaad het geval is. We hebben dit gebruikt om de foutschatting weergegeven in de laatste kolom op te stellen.

**4.2B De Euler-Maclaurin reeks.** Nu we met succes de techniek uit sectie 3.4 hebben toegepast op numerieke integratie met de gerepeteerde trapeziumregel, rijst de vraag of we misschien ook een extrapolatieproces als in sectie 3.5 kunnen maken. Essentieel hierbij is de vraag of we de fout  $R(h)$  van de gerepeteerde trapeziumregel kunnen ontwikkelen in machten van  $h$ :

$$R(h) = c_2 h^2 + c_3 h^3 + c_4 h^4 + \cdots + c_m h^m + \mathcal{O}(h^{m+1}).$$

Dat dit inderdaad kan blijkt uit stelling 4.2.1:

**Stelling 4.2.1**

Er bestaan getallen  $B_2, B_4, B_6, \dots$ , zodanig dat voor  $f \in C^{2m}[a, b]$ ,  $h = \frac{1}{n}(b-a)$ ,

$$\int_a^b f(x) dx - T(h) = - \sum_{i=1}^m h^{2i} B_{2i} [f^{(2i-1)}(b) - f^{(2i-1)}(a)] / (2i)! + o(h^{2m}). \quad (68)$$

**Bewijs.** Zie Bijlage C. □

**Opmerking 4.2.2** bij stelling 4.2.1.

(i) Voor  $f \in C^\infty[a, b]$  kan men  $m$  willekeurig groot maken; men krijgt dan de formule

$$\int_a^b f(x) dx - T(h) \sim - \sum_{i=1}^{\infty} h^{2i} B_{2i} [f^{(2i-1)}(b) - f^{(2i-1)}(a)] / (2i)!. \quad (69)$$

Net als in (46) gebruiken we hier het symbool  $\sim$  om aan te geven dat we niet weten of de reeks  $\sum_{i=1}^{\infty} h^{2i} B_{2i} [f^{(2i-1)}(b) - f^{(2i-1)}(a)] / (2i)!$  wel convergeert, en dat ons dat ook eigenlijk niet interesseert; (69) betekent dan ook niet anders dan dat (68) geldt voor elke  $m$ . We spreken in een dergelijk geval liever van een asymptotische ontwikkeling dan van een reeks. Desalniettemin wordt het rechterlid van (69) in de literatuur vaak de *Euler-Maclaurin reeks* genoemd.

(ii) De coëfficiënten  $B_{2i}$  zijn de zogenaamde *Bernoulli getallen*. Vele relaties zijn hiervoor bekend (zie bijvoorbeeld Abramowitz-Stegun, “*Handbook of mathematical functions*”, p. 803-810). Enige waarden:

$$B_2 = \frac{1}{6}, \quad B_4 = -\frac{1}{30}, \quad B_6 = \frac{1}{42}, \quad B_8 = -\frac{1}{30}, \\ B_{10} = \frac{5}{66}, \quad B_{12} = -\frac{691}{2730}, \quad B_{14} = \frac{7}{6}, \quad B_{16} = -\frac{3617}{510}.$$

Asymptotisch geldt:  $B_{2i} \simeq (-1)^{i+1} 2(2i)! / (2\pi)^{2i}$ .

(iii) Er bestaat ook een variant van (68) zonder  $o(h^{2m})$ -term:

$$\int_a^b f(x) dx - T(h) = - \sum_{i=1}^{m-1} h^{2i} B_{2i} [f^{(2i-1)}(b) - f^{(2i-1)}(a)] / (2i)! - \\ -(b-a) h^{2m} B_{2m} f^{(2m)}(\xi) / (2m)! \quad (70)$$

**4.2C Romberg integratie.** De theorie uit sectie 3.5 kan nu rechtstreeks worden toegepast op het benaderen van  $\int_a^b f(x) dx$  door middel van de gerepeteerde trapeziumregel. Ook stelling 3.5.1 blijft onverkort van toepassing.

Bij numerieke integratie wordt vaak aanbevolen om de Bulirsch rij  $h_0, \frac{1}{2}h_0, \frac{1}{3}h_0, \frac{1}{4}h_0, \frac{1}{6}h_0, \frac{1}{8}h_0, \frac{1}{12}h_0, \dots$  te gebruiken met  $h_0 = b - a$  ( $\frac{1}{2}h_0, \frac{1}{3}h_0, \frac{1}{4}h_0, \dots$  is de Bulirsch rij in 3.5E met  $h = \frac{1}{2}h_0$ ). Het oogmerk hierbij is het beperken van het benodigde aantal evaluaties van  $f$ . In feite is de Bulirsch rij hiervoor ontworpen.

**Opgave 4.2.1** Ga na dat het opstellen van de eerste 7 rijen in het Romberg schema corresponderend met de Romberg danwel Bulirsch rij respectievelijk 65 en 17 functie evaluaties vergt.

**Opmerking 4.2.3** (i) Men ziet nu het gebeuren van de tabel in 4.2A verklaard: de eerste term uit de Euler-Maclaurin reeks is hier nul, zodat de restterm voor  $h \rightarrow 0$  evenredig is met  $ch^4$ . Vandaar de factoren dichtbij 16 in de vierde kolom, en de goede benadering van  $R(h)$  door de waarden uit de laatste kolom.

(ii) Als we een periodieke functie integreren over een periodiciteitsinterval zijn alle termen van de Euler-Maclaurin reeks nul. De restterm van de gerepeteerde trapeziumregel gaat dan sneller naar nul dan elke macht van  $h$ .

**Voorbeeld 4.2.2**

$$\int_0^\pi \frac{1}{1 + \cos^2 x} dx = \pi/\sqrt{2} = 2.221441470$$

$n = \frac{\pi}{h}$	benadering	fout	factor
1	1.570796327	0.650645143	4.82
2	2.356194491	-0.134753021	34.97
4	2.225294797	-0.003853327	1155.07
8	2.221444806	-0.000003336	?
16	2.221441470	0	

Het zal duidelijk zijn dat onder deze omstandigheden het opstellen van een Romberg schema volstrekt zinloos is.

### 4.3 Andere kwadratuurformules

**4.3A** Zoals lineaire interpolatie aanleiding geeft tot de trapeziumregel, zo geven ook andere interpolatieformules aanleiding tot kwadratuurformules.

Laat  $x_0, \dots, x_n$  verschillende punten in  $[a, b]$  zijn. Zij  $p$  het Lagrange interpolatiepolynoom van  $f$  op de punten  $x_0, \dots, x_n$ . Dan zullen we dus  $\int_a^b p(x) dx$  als benadering voor  $\int_a^b f(x) dx$  nemen. Nu weten we dat  $p(x) = \sum_{k=0}^n f(x_k)L_{kn}(x)$  (zie (6)), zodat

$$\int_a^b p(x) dx = \sum_{k=0}^n w_k f(x_k) \quad (71)$$

met als coëfficiënten

$$w_k = \int_a^b L_{kn}(x) dx \quad (72)$$

die blijkbaar onafhankelijk van  $f$  zijn.

We noemen  $\sum_{k=0}^n w_k f(x_k)$  de bij  $x_0, \dots, x_n$  behorende *interpolatoire kwadratuurformule* voor  $\int_a^b f(x) dx$ . De  $x_k$  heten de *steunpunten* (net als bij Lagrange interpolatie, zie sectie 1.2), de  $w_k$  de *gewichten* van de kwadratuurformule.

**Stelling 4.3.1** Als  $f$  een polynoom van graad hoogstens  $n$  is, dan is elke  $n + 1$ -punts interpolatoire kwadratuurformule exact:

$$\int_a^b f(x) dx = \sum_{k=0}^n w_k f(x_k).$$

**Bewijs.** Zoals blijkt uit opmerking 1.2.1 valt  $f$  samen met zijn Lagrange interpolatiepolynoom op  $x_0, \dots, x_n$ , en dus geldt  $\int f(x) dx = \int p(x) dx$   $\square$

Omgekeerd geldt:

**Stelling 4.3.2** Wanneer  $x_0, \dots, x_n$  en  $w_0, \dots, w_n$  gegeven getallen-rijtjes zijn en er geldt

$$\int_a^b p(x) dx = \sum_{k=0}^n w_k p(x_k)$$

voor alle polynomen  $p$  van graad  $\leq n$ , dan is  $\sum_{k=0}^n w_k f(x_k)$  de unieke bij  $x_0, \dots, x_n$  behorende interpolatoire kwadratuurformule voor  $\int_a^b f(x) dx$ .

**Bewijs.** Op grond van stelling 4.3.1 kunnen we volstaan met het bewijzen dat de gewichten  $w_k$  uniek zijn. Zij  $0 \leq k \leq n$  gegeven. Definiëer de unieke  $q \in P_n$  door  $q(x_k) = 1$ ,  $q(x_\ell) = 0$  voor  $\ell \neq k$ . Dan geldt  $\int_a^b q(x) dx = w_k$ .  $\square$

**4.3B** We kunnen stelling 4.3.2 ook gebruiken om de gewichten  $w_k$  uit te rekenen als de punten  $x_0, \dots, x_n$  gegeven zijn. Neem namelijk maar voor  $p(x)$  achtereenvolgens  $1, x, \dots, x^n$  of een willekeurige andere basis van  $P_n$ . Dit levert een stelsel van  $n + 1$  vergelijkingen in de  $n + 1$  onbekenden  $w_k$  op, welk stelsel dus uniek oplosbaar is. (In verband met optredende afrondfouten is het voor grote waarden van  $n$  niet aanbevelenswaardig de basis  $1, x, \dots, x^n$  te gebruiken).

**Voorbeeld 4.3.1** Neem  $a = x_0 = 0$ ,  $b = x_1 = 1$ . Dan is de kwadratuurformule  $w_0 f(0) + w_1 f(1)$ . We weten dat moet gelden:

$$1 = \int_0^1 1 dx = w_0 + w_1$$

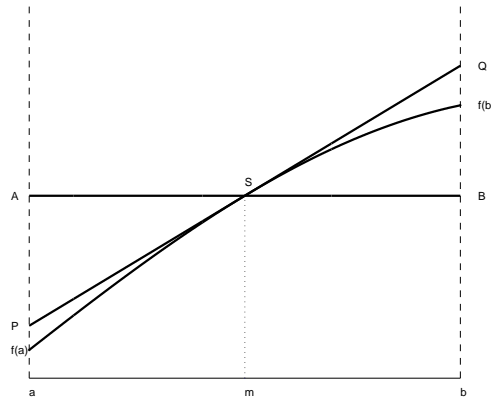
$$\frac{1}{2} = \int_0^1 x dx = w_0 \cdot 0 + w_1 \cdot 1 = w_1$$

zodat  $w_0 = w_1 = \frac{1}{2}$ ; we vinden zo de trapeziumregel terug.

**Voorbeeld 4.3.2** Neem  $a = x_0 = -1$ ,  $x_1 = 0$ ,  $b = x_2 = 1$ ,  $n = 2$ . Op dezelfde wijze als boven vinden we

$$\begin{aligned} w_0 + w_1 + w_2 &= 2 \\ -w_0 + w_2 &= 0 \\ w_0 + w_2 &= \frac{2}{3} \end{aligned}$$

Hieruit volgt  $w_0 = w_2 = \frac{1}{3}$ ,  $w_1 = \frac{4}{3}$ .



FIGUUR 8: De *midpuntregel*. Het figuur maakt duidelijk waarom benaderend integreren middels de *midpuntregel* zoveel beter is dan men op grond van een 0-de orde benadering zou verwachten: het oppervlak onder de grafiek van de 0-de orde benadering (de rechte  $AB$ ) is gelijk aan het oppervlak onder de grafiek van de eerste orde benadering (de rechte  $PQ$ ).

**4.3C Speciaal geval: de *midpuntregel*.** Door nulde orde interpolatie van  $f$  in het punt  $m = \frac{1}{2}(a + b)$  op het segment  $[a, b]$  vinden we de volgende kwadratuurformule voor  $\int_a^b f(x) dx$ :

$$(b - a)f(m). \quad (73)$$

Dit noemen we de *midpuntregel* (zie figuur 8). We benaderen dus blijkbaar de grafiek van  $f$  door de rechte  $AB$  en dat lijkt nogal grof. Evenwel is  $\text{opp}(a, b, B, A) = \text{opp}(a, b, Q, P)$  waarbij  $PQ$  een willekeurige lijn door  $S$  is, en als we  $PQ$  de raaklijn in  $S$  nemen zien we dat de *midpuntregel* een veel beter resultaat geeft dan aanvankelijk werd verwacht.

Op grond van stelling 4.3.1 weten we dat de *midpuntregel* exact is voor constante functies. Aangezien met  $\phi(x) = x - m$ ,  $\int_a^b \phi(x) dx = 0 = \phi(m)$  is de *midpuntregel* ook exact voor  $\phi$ . Daar iedere  $p \in P_1$  te schrijven is als  $p = c_1\phi + c_2$  voor zekere constanten  $c_1$  en  $c_2$ , is de *midpuntregel* blijkbaar zelfs exact op  $P_1$ .

Integratie van de standaard restterm voor het 0-de orde interpolatiepolynoom in  $m$  levert

$$\int_a^b f(x) dx = (b - a)f(m) + \int_a^b (x - m)f'(\xi(x)) dx$$

Hierin is  $x \mapsto (x - m)$  niet tekenvast op  $[a, b]$ , en is op grond van het voorafgaande  $f'$  een factor welke we liever niet tegenkomen.

Aannemende dat  $f \in C^2[a, b]$ , definiëren we daarom  $q \in P_1$  door  $q(m) = f(m)$  en  $q'(m) = f'(m)$ . Er geldt dan  $\int_a^b q(x) dx = (b - a)q(m) = (b - a)f(m)$ , en zo m.b.v. de representatie van  $f - q$  uit stelling 1.3.3,

$$\int_a^b f(x) dx = (b - a)f(m) + \int_a^b (x - m)^2 \frac{f''(\xi(x))}{2} dx$$

Aangezien nu  $(x - m)^2$  wel tekenvast is, vinden we analoog aan stelling 4.1.1

$$\int_a^b f(x) dx - (b - a)f(m) = \frac{(b - a)^3}{24} f''(\xi) \quad (74)$$

voor zekere  $\xi \in [a, b]$ .

Voor de  $n$  keer gerepeteerde midpuntregel vindt men nu, met  $h = (b - a)/n$ ,  $m_i = a + (i + \frac{1}{2})h$ :

$$\int_a^b f(x) dx = h[f(m_0) + \cdots + f(m_{n-1})] + R(h) \quad (75)$$

waarbij  $R(h) = \frac{b-a}{24}h^2 f''(\xi)$  voor een  $\xi \in [a, b]$ , en ook  $R(h) = \frac{h^2}{24}(f'(b) - f'(a)) + o(h^2)$ .

**4.3D Speciaal geval: de Simpson regel.** We kiezen nu als steunpunten  $x_0 = a$ ,  $x_1 = \frac{1}{2}(a+b) = m$ ,  $x_2 = b$ , en benaderen  $f$  met het tweedegraads interpolatiepolynoom op  $x_0, x_1, x_2$ . Voor het geval  $a = -1, b = 1$  hebben we de gewichten al uitgerekend in voorbeeld 4.3.2. Analoog vinden we, bijv. door exactheid te eisen voor  $f(x) = 1$ ,  $f(x) = x - x_1$ ,  $f(x) = (x - x_1)^2$  waarbij  $x_1 = (a + b)/2$ , als kwadratuurformule voor  $\int_a^b f(x) dx$  met willekeurige  $a$  en  $b$ :

$$\frac{b-a}{6}[f(x_0) + 4f(x_1) + f(x_2)]. \quad (76)$$

Dit is de zogenaamde *regel van Simpson*.

Stelling 4.3.1 laat zien dat Simpson exact is  $P_2$ . Nu geldt echter met  $\phi(x) = (x - x_1)^3$ :

$$\int_a^b \phi(x) dx = 0 \quad \text{en} \quad \frac{b-a}{6}[\phi(a) + 4\phi(x_1) + \phi(b)] = 0$$

zodat Simpson ook nog exact is voor deze speciale derdegraads functie, en dus voor iedere derdegraads functie.

Zij nu  $f \in C^4[a, b]$ . Analoog aan 4.3C, definiëren we  $q \in P_3$  door  $q(x_i) = f(x_i)$ ,  $i = 0, 1, 2$  en  $q'(x_1) = f'(x_1)$ . Omdat  $q$  een derdegraads polynoom is, geldt

$$\int_a^b q(x) dx = \sum_{i=0}^2 w_i q(x_i) = \sum_{i=0}^2 w_i f(x_i).$$

Stelling 1.3.3 levert dat

$$\int_a^b f(x) dx = \int_a^b q(x) dx + \int_a^b (x - x_0)(x - x_1)^2(x - x_2) \frac{f^{(4)}(\xi(x))}{4!} dx.$$

Aangezien  $(x - x_0)(x - x_1)^2(x - x_2)$  tekenvast is concluderen we dat

$$\int_a^b f(x) dx - \sum_{i=0}^2 w_i f(x_i) = -\frac{1}{90} \left(\frac{b-a}{2}\right)^5 f^{(4)}(\xi)$$

voor zekere  $\xi \in [a, b]$ .

Voor de  $n$  keer gerepeteerde Simpson regel vindt men nu met  $h = (b - a)/2n$ ,  $x_i = a + ih$

$$\int_a^b f(x) dx = \frac{h}{3}[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \cdots + 4f(x_{2n-1}) + f(x_{2n})] + R(h) \quad (77)$$

waarbij  $R(h) = -\frac{b-a}{180}h^4 f^{(4)}(\xi)$  voor een  $\xi \in [a, b]$ , en ook  $R(h) = -\frac{h^4}{180}(f^{(3)}(b) - f^{(3)}(a)) + o(h^4)$ .

**Opgave 4.3.1** Zij  $(T_{ij})_{i=0,1,2,\dots, j=0,1,2,\dots, i}$  het Romberg schema (zie (49)) ter benadering van  $\int_a^b f(x)dx$ , waarbij  $T_{i0}$  het resultaat is van de  $2^i \times$  gerepeteerde trapeziumregel bij een equidistante opdeling van  $[a, b]$  (Romberg rij).

- Ga na dat  $T_{i1}$  het resultaat is van de  $2^{i-1} \times$  gerepeteerde Simpson regel.
- Zie in dat  $T_{jj}$  van de vorm  $\sum_{k=0}^n w_k f(x_k)$  is met  $x_k$  verschillend in  $[a, b]$ . Wat is  $n$  als functie van  $j$ ?
- Tot op welke graad is  $T_{jj}$  exact? (gebruik (70) en realiseer dat voor  $f \in P_{2i-1}$  geldt  $f^{(2i-1)}(b) - f^{(2i-1)}(a) = 0$ ).
- Aannemende dat de  $w_k$ 's in b) ongelijk nul zijn, laat m.b.v. b), c) en stelling 4.3.2 zien dat voor  $j \geq 3$ ,  $T_{jj}$  geen interpolatoire kwadratuurformule voorstelt.

**4.3E Newton-Cotes formules.** Laat  $x_0, \dots, x_n$  equidistant in  $[a, b]$  liggen met  $x_0 = a$ ,  $x_n = b$ . De kwadratuurformules die ontstaan door integratie van het  $n$ -de orde interpolatiepolynoom  $p_n$  van  $f$  noemt men  $(n+1)$ -punts *Newton-Cotes formules*. Speciale gevallen: trapeziumregel ( $n=1$ ) en Simpson regel ( $n=2$ ). Voor de rest-termen van deze formules geldt de volgende stelling (voor een bewijs zie bijvoorbeeld: E. Isaacson en H.B. Keller, Analysis of numerical methods, Wiley 1966):

**Stelling 4.3.3** Voor de restterm  $R_n$  van de  $(n+1)$ -punts *Newton-Cotes kwadratuurformule* voor het benaderen van  $\int_a^b f(x)dx$  geldt met  $h = \frac{b-a}{n}$ : als  $n$  even is en  $f \in C^{n+2}[a, b]$ , dan

$$R_n = C_n h^{n+3} f^{(n+2)}(\xi), \quad (78)$$

$$C_n = \frac{1}{(n+2)!} \int_0^n t^2(t-1)(t-2) \cdots (t-n) dt; \quad (79)$$

als  $n$  oneven is en  $f \in C^{n+1}[a, b]$  dan

$$R_n = D_n h^{n+2} f^{(n+1)}(\xi) \quad (80)$$

$$D_n = \frac{1}{(n+1)!} \int_0^n t(t-1) \cdots (t-n) dt. \quad (81)$$

□

**4.3F Transformatie van het integratie interval, repeteren en Romberg.** Wanneer men een kwadratuurformule heeft voor een zekere integratie interval dan kan men er daaruit een afleiden voor een ander integratie interval:

**Stelling 4.3.4** Zij  $\sum_0^n w_i f(x_i)$  een kwadratuurformule voor  $\int_a^b f(x) dx$ , en laat gelden

$$\int_a^b f(x) dx = \sum_0^n w_i f(x_i) + \alpha(b-a)^{k+1} f^{(k)}(\xi), \quad \xi \in [a, b] \quad (82)$$

voor alle  $k \times$  continu differentieerbare functies  $f$ . Dan geldt

$$\int_c^d g(y) dy = \frac{d-c}{b-a} \sum_0^n w_i g(y_i) + \alpha(d-c)^{k+1} g^{(k)}(\xi), \quad \xi \in [c, d] \quad (83)$$

met  $y_i = \frac{d-c}{b-a} x_i + \frac{bc-ad}{b-a}$ , voor alle  $k \times$  continu differentieerbare functies  $g$ .

**Bewijs.** Substitueer  $y = \frac{d-c}{b-a}x + \frac{bc-ad}{b-a}$ . Dan geldt

$$\begin{aligned} \int_c^d g(y) dy &= \frac{d-c}{b-a} \int_a^b g\left[\frac{d-c}{b-a}x + \frac{bc-ad}{b-a}\right] dx = \\ &= \frac{d-c}{b-a} \sum_0^n w_i g(y_i) + \frac{d-c}{b-a} \alpha (b-a)^{k+1} \frac{d^k}{dx^k} g\left[\frac{d-c}{b-a}x + \frac{bc-ad}{b-a}\right] \Big|_{x=\xi} \end{aligned}$$

Hieruit volgt het gestelde.  $\square$

**Opmerking 4.3.1** (i) De stelling ziet er op het eerste gezicht wat afschrikwekkend uit. Er staat echter niets anders dan dat de punten  $y_i$  in  $[c, d]$  “gelijkvormig” liggen met de punten  $x_i$  in  $[a, b]$ , en dat de gewichten evenredig zijn met de lengte van het interval.

(ii) Neemt men in de stelling  $f = g$  dan ziet men dat bij verkorting van het integratie interval de fout bijzonder snel afneemt bij grote waarden van  $k$ .

Een belangrijke toepassing van stelling 4.3.4 is het ontwerpen van gerepeteerde kwadratuurformules. We zien hieruit dat als we met een kwadratuurformule als in (82) een  $n \times$  gerepeteerde formule maken, we daarvoor met  $h = (b-a)/n$  de volgende restterm krijgen

$$\alpha h^{k+1} \sum_{i=1}^n g^{(k)}(\xi_i) = \begin{cases} \alpha (b-a) h^k g^{(k)}(\xi), \\ \alpha h^k (g^{(k-1)}(b) - g^{(k-1)}(a)) + o(h^k). \end{cases}$$

Met dergelijke gerepeteerde formules geldt een theorie analoog aan die van de gerepeteerde trapeziumregel: men kan automatisch integreren, er bestaat een asymptotische ontwikkeling analoog aan de Euler-Maclaurin reeks, en men kan een Romberg schema vormen.

**4.3G Gauss formules.** We zagen dat de midpunt- en Simpson regel, en algemeen de  $(n+1)$ -punts Newton-Cotes formules met  $n$  even exact zijn op een graad welke één hoger is dan men op voorhand zou verwachten. In deze paragraaf onderzoeken wij dit verschijnsel in zijn algemeenheid.

**Stelling 4.3.5** *Zij  $x_0, \dots, x_n$  verschillende punten in  $[a, b]$ . Definieer*

$$q(x) = \prod_{i=0}^n (x - x_i) \in P_{n+1}.$$

*Zij  $m > n$ . Dan is de interpolatoire kwadratuurformule behorend bij  $x_0, \dots, x_n$  exact op  $P_m$  dan en slechts dan als  $\int_a^b q(x)p(x) dx = 0$  voor alle  $p \in P_{m-n-1}$  (d.w.z. als  $q$  loodrecht staat op  $P_{m-n-1}$  t.o.v. het inproduct  $(f, g) = \int_a^b f(x)g(x) dx$  op  $C[a, b]$ ).*

**Bewijs.** Zij  $p \in P_{m-n-1}$ . Dan is  $qp$  in  $P_m$ . Indien de kwadratuurformule exact is op  $P_m$ , dan volgt uit  $\sum_{i=0}^n w_i q(x_i)p(x_i) = 0$  dat  $\int_a^b q(x)p(x) dx = 0$ .

Omgekeerd zij  $f \in P_m$ . Schrijf  $f = pq + r$  voor zekere  $p \in P_{m-n-1}$  en  $r \in P_n$ . De existentie van dergelijke  $p$  en  $r$  wordt gegarandeerd door de zgn. Chinese reststelling, hetgeen niets anders is dan staartdeling met rest. Uit  $\int f = \int pq + \int r$ ,  $\int r = \sum w_i r(x_i)$

(immers  $r \in P_n$ ),  $r(x_i) = f(x_i)$  ( $\forall i$ ) en de veronderstelling  $\int pq = 0$  volgt nu  $\int f = \sum w_i f(x_i)$ .  $\square$

De polynomen corresponderend met de midpunt- en Simpson regel worden gegeven door respectievelijk  $q(x) = x - m$  en  $q(x) = (x - a)(x - m)(x - b)$ . Inderdaad geldt  $\int_a^b q(x) dx = 0$  terwijl  $\int_a^b q(x)x dx \neq 0$ , d.w.z. er wordt precies één extra orde van exactheid verkregen.

Een interessante vraag is hoe de steunpunten gekozen moeten worden opdat een zo hoog mogelijke graad van exactheid bereikt wordt. Uiteraard is er geen  $q \in P_{n+1}$  met  $\int q P_{n+1} = 0$  (immers  $\int q^2 > 0$ ), d.w.z. in stelling 4.3.5 is  $m - n - 1 < n + 1$  oftewel  $m < 2n + 2$ .

Definieer nu de rij  $(q_n)_{n \geq 0}$  van zgn. *Gauss-Legendre polynomen* recursief door

$$q_n = x^n - \sum_{k=0}^{n-1} \frac{(x^n, q_k)}{(q_k, q_k)} q_k. \quad (84)$$

(De rij  $1, x, x^2, \dots$  mag vervangen worden door een willekeurige rij  $p_0, p_1, p_2, \dots$  z.d.d. voor alle  $n$ ,  $\{p_0, \dots, p_n\}$  een basis is van  $P_n$ )

**Opgave 4.3.2** a) Met  $P_{-1} := 0$ , bewijs met inductie dat

$$q_n \in P_n, \quad q_n - x^n \in P_{n-1}, \quad q_n \perp P_{n-1}, \quad (85)$$

en  $\text{span}\{q_0, \dots, q_n\} = P_n$ .

b) Laat zien dat er slechts één  $q_n$  is welke aan (85) voldoet.

c) Bewijs dat  $q_n$  precies  $n$  verschillende nulpunten heeft, alle in  $(a, b)$ . (Hint: Stel  $q_n$  heeft slechts  $k < n$  verschillende nulpunten  $x_1, \dots, x_k$ . Definieer  $p(x) = \prod_{i=1}^k (x - x_i)$  en leidt een tegenspraak af met (85)).

d) Het zgn. *Gram-Schmidt orthogonalisatie proces* (84) is voor grote  $n$  gevoelig voor afrondfouten. Bewijs daarom met  $q_{-1} := 0$ , dat

$$q_{n+1} = \left( x - \frac{(xq_n, q_n)}{(q_n, q_n)} \right) q_n - \frac{(xq_n, q_{n-1})}{(q_{n-1}, q_{n-1})} q_{n-1},$$

hetgeen een efficiënte en stabiele berekeningswijze van deze orthogonale polynomen geeft. (Hint: Laat zien dat het rechterlid element is van  $P_{n+1}$ , leidende coëfficiënt gelijk aan 1 heeft, en loodrecht staat op  $P_{n-2}$ ,  $q_{n-1}$  en  $q_n$ ).

De interpolatoire kwadratuurformule behorend bij de nulpunten  $x_0, \dots, x_n$  van  $q_n$ , de zgn.  $(n + 1)$ -punts *Gauss* of *Gauss-Legendre formule*, is dus de unieke  $(n + 1)$ -punts formule welke exact is op de maximale graad  $2n + 1$ . Voor de restterm kan bewezen worden dat

$$R = \frac{(b - a)^{2n+3} ((n + 1)!)^4}{(2n + 3) ((2n + 2)!)^3} f^{(2n+2)}(\xi). \quad (86)$$

**Voorbeeld 4.3.3** In onderstaande tabel geven we de waarde van de restterm weer als men ter berekening van  $\int_0^\pi \sin(x) dx$  een zeker aantal punten investeert in gerepeteerde trapezium, gerepeteerde Simpson, diagonaal van het Romberg schema gebaseerd op de trapeziumregel met Romberg rij ( $h_0 = \pi$ ,  $\frac{1}{2}h_0$ ,  $\frac{1}{4}h_0$ ,  $\frac{1}{8}h_0$ ,  $\frac{1}{16}h_0, \dots$ ), diagonaal van

het Romberg schema gebaseerd op de trapeziumregel met Bulirsch rij en gerepeteerde 5-punts Gauss.

methode	aantal malen gerepeteerd	aantal punten	fout
trap.	4	5	$1.04 \cdot 10^{-1}$
Simpson	2	5	$4.56 \cdot 10^{-3}$
diag. R. schema, Romberg rij		5	$1.43 \cdot 10^{-3}$
diag. R. schema, Bulirsch rij		4	$2.57 \cdot 10^{-3}$
5 p. Gauss	1	5	$1.11 \cdot 10^{-7}$
trap.	8	9	$2.58 \cdot 10^{-2}$
Simpson	4	9	$2.69 \cdot 10^{-4}$
Romberg schema, Romberg rij		9	$-5.55 \cdot 10^{-6}$
R. schema, Bulirsch rij		9	$2.83 \cdot 10^{-7}$
5 p. Gauss	2	10	$1.1 \cdot 10^{-10}$
trap.	16	17	$6.43 \cdot 10^{-3}$
Simpson	8	17	$1.66 \cdot 10^{-5}$
Romberg schema, Romberg rij		17	$5.42 \cdot 10^{-9}$
R. schema, Bulirsch rij		17	$1.92 \cdot 10^{-12}$
5 p. Gauss	4	20	$1.1 \cdot 10^{-13}$

Opvallend is dat 5-punts Gauss enkelvoudig toegepast reeds zo'n goed resultaat geeft, alsmede dat halvering van het interval bij 5-punts Gauss de nauwkeurigheid zo snel doet toenemen, overigens geheel in overeenstemming met opmerking 4.3.1(ii). I.h.b. daar de sinus analytisch is op  $\mathbb{C}$  kan men van de 9- respectievelijk 17-punts Gauss formules nog veel betere resultaten verwachten.

Met het oog op de, voor gladde  $f$ , bijzonder gunstige restterm (86) zou men kunnen denken dat er in ieder geval voor dergelijke functies er geen redenen zijn om andere dan ongerepeteerde Gauss formules te gebruiken. Bedenk echter dat de techniek van repeteren/Romberg de mogelijkheid biedt al rekenend fout te schatten, en zo nodig goedkoop het resultaat te verbeteren.

#### 4.4 Convergentie van kwadratuurschema's

**4.4A** Onder een *kwadratuurschema* verstaat men een rij kwadratuurformules van de vorm

$$(f \mapsto \sum_{i=0}^{m(n)} w_i^{(n)} f(x_i^{(n)}))_{n \in \mathbb{N}} \quad (87)$$

ter benadering van  $f \mapsto \int_a^b f(x) dx$ . Voorbeelden hiervan zijn:

- een  $n \times$  gerepeteerde interpolatoire kwadratuur formule met stapgrootte  $h = (b - a)/n$ .
- de diagonaal van het bijbehorend Romberg schema
- de  $(n + 1)$ -punts Newton-Cotes formule op  $[a, b]$ .
- de  $(n + 1)$ -punts Gauss formule op  $[a, b]$ .

Een voor de hand liggende vraag is nu, of er wellicht geldt dat  $R_n := \int_a^b f(x)dx - \sum_{i=0}^{m(n)} w_i^{(n)} f(x_i^{(n)})$  naar nul gaat voor  $n \rightarrow \infty$ .

**Stelling 4.4.1** *Noodzakelijk en voldoende opdat  $\lim_{n \rightarrow \infty} R_n(f) = 0$  voor elke  $f \in C[a, b]$  is, dat*

- a)  $\lim_{n \rightarrow \infty} R_n(p) = 0$  voor elk polynoom  $p$ , en
- b)  $\sum_{i=0}^{m(n)} |w_i^{(n)}| \leq M$  voor zekere  $M$  onafhankelijk van  $n$ .

Voor de geïnteresseerde lezer geven we het bewijs:

**Bewijs.** De zgn. *Stelling van Weierstrass* uit de approximatie theorie zegt dat er voor iedere  $f \in C[a, b]$  en  $\tilde{\epsilon} > 0$  er een polynoom  $p$  is met  $\max_{x \in [a, b]} |(f - p)(x)| \leq \tilde{\epsilon}$ . Zij nu  $f \in C[a, b]$  en  $\epsilon > 0$  gegeven. Zij  $p$  een polynoom met  $\max_{x \in [a, b]} |(f - p)(x)| \leq \epsilon/[2((b - a) + M)]$ . Zij  $n$  groot genoeg opdat  $|R_n(p)| \leq \epsilon/2$ . Dan volgt uit

$$R_n(f) = \int (f - p) + \sum w_i^{(n)} (p(x_i^{(n)}) - f(x_i^{(n)})) + R_n(p),$$

dat  $|R_n(f)| \leq ((b - a) + M) \max_{x \in [a, b]} |(f - p)(x)| + |R_n(p)| \leq \epsilon$

De noodzaak van conditie a) is duidelijk. Stel nu dat  $\lim_{n \rightarrow \infty} R_n(f) = 0$  voor elke  $f \in C[a, b]$ . Dan geldt i.h.b. dat voor elke  $f \in C[a, b]$ , de rij  $(\sum w_i^{(n)} f(x_i^{(n)}))_{n \in \mathbb{N}}$  begrensd is. De zgn. *Stelling van Banach-Steinhaus* uit de functionaalanalyse zegt dan dat er een  $M$  is z.d.d.

$$|\sum w_i^{(n)} f(x_i^{(n)})| \leq M \max_{x \in [a, b]} |f(x)| \quad (88)$$

voor alle  $n$  en voor alle  $f \in C[a, b]$ . Eenvoudig construeert men voor iedere  $n$  een  $f_n \in C[a, b]$  z.d.d. in (88) het linkerlid gelijk is aan  $\sum |w_i^{(n)}|$  en het rechterlid gelijk is aan 1, waarmee ook de noodzaak van conditie b) aangetoond is.  $\square$

**Opmerking 4.4.1** Stel  $\tilde{f}(x_i^{(n)}) = f(x_i^{(n)}) + \epsilon(x_i)$  met  $|\epsilon(x_i)| \leq \epsilon$ . Dan geldt

$$|\sum w_i^{(n)} (\tilde{f}(x_i^{(n)}) - f(x_i^{(n)}))| \leq \epsilon \sum |w_i^{(n)}|,$$

en deze ongelijkheid is i.h.a. scherp. Los van de convergentievraag toont dit dus de relevantie aan van conditie b) voor de numerieke praktijk.

Het is eenvoudig in te zien dat een rij van  $n \times$  gerepeteerde interpolatoire kwadratuurformules aan de voorwaarden van stelling 4.4.1 voldoet. Uit stelling B.1 van Bijlage B valt af te leiden dat, in ieder geval voor de trapeziumregel en de Romberg rij, de diagonaal van het Romberg schema ook aan deze voorwaarden voldoet.

**Opgave 4.4.1** Bewijs dat een  $(n + 1)$ -punts kwadratuurformule welke exact is op  $P_{2n}$  slechts positieve gewichten heeft. Laat hiermee zien dat de rij van  $(n + 1)$ -punts Gauss formules aan de voorwaarden van stelling 4.4.1 voldoet.

Op grond van (14) zou men de vrees kunnen krijgen dat de rij van  $(n + 1)$ -punts Newton-Cotes formules niet aan conditie b) van stelling 4.4.1 voldoet, het geen inderdaad het geval blijkt te zijn (zie Natanson, “*Constructive Function Theory*, vol. III”).

**Literatuur** bij hoofdstuk 4:

- H. Brass, “*Quadraturverfahren*”, Vandenhoeck & Ruprecht 1977.
- P.J. Davis & P. Rabinowitz, “*Methods of numerical integration*”, Academic Press 1975.

## A Interpolatie met afgeleiden

Een expliciete oplossing van het interpolatieprobleem zoals gesteld in definitie 1.3.1 is in het algemeen moeilijk aan te geven. We moeten daarom de existentie van een oplossing apart bewijzen; verder moeten we de uniciteit aantonen, en de restterm aangeven. We doen dit in een wat eigenaardige volgorde: eerst geven we de restterm (van een oplossing waarvan we nog niet hebben aangetoond dat hij bestaat!), vervolgens tonen we aan dat een eventuele oplossing uniek is, en tenslotte bewijzen we de existentie van de oplossing.

In het volgende is  $(x_0, \dots, x_n)$  steeds een stijgende rij (niet noodzakelijk verschillende) getallen. Dus  $x_{i-1} \leq x_i$  ( $i = 1, \dots, n$ ).

**Stelling A.1** *Laat  $f \in C[a, b]$  zodanig zijn dat  $f^{(n+1)}$  bestaat op  $]a, b[$ . Veronderstel verder, als  $x_i = a$  of  $x_i = b$  voor zekere  $0 \leq i \leq n$ , dat  $f$  voldoende vaak rechts-, respectievelijk linksdifferentiëerbaar is. Laat  $p$  een polynoom zijn van graad hoogstens  $n$  waarvoor  $p = f$  op  $(x_0, \dots, x_n)$ . Dan bestaat er voor elke  $x \in [a, b]$  een  $\xi \in ]x_0, \dots, x_n, x[$  zodanig dat*

$$f(x) - p(x) = (x - x_0)(x - x_1) \cdots (x - x_n) \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

**Bewijs.** Kies een  $x \in [a, b]$ . In dit bewijs is  $x$  verder vast en is  $t$  variabel. Als  $x$  samenvalt met een van de punten  $x_0, \dots, x_n$  kan men  $\xi$  willekeurig kiezen. Zij dus  $x \neq x_i$  voor alle  $i$ . Dan is er een getal  $\alpha$  zo dat

$$f(x) - p(x) = \alpha r(x) \quad \text{met} \quad r(t) := \prod_{i=0}^n (t - x_i)$$

(neem  $\alpha = (f(x) - p(x))/r(x)$ ). Beschouw voor deze  $\alpha$  de functie

$$\phi(t) := f(t) - p(t) - \alpha r(t)$$

Hernummer de rij  $(x_0, \dots, x_n, x)$  tot een stijgende rij  $(\tilde{x}_0, \dots, \tilde{x}_{n+1})$ . Merk op dat  $r = \mathbf{0}$  op  $(x_0, \dots, x_n)$ . Hierbij is  $\mathbf{0}$  de 0-functie. Dit impliceert dat  $\phi = \mathbf{0}$  op de rij  $(\tilde{x}_0, \dots, \tilde{x}_{n+1})$ . Als  $\tilde{x}_i < \tilde{x}_{i+1}$  dan is er volgens Rolle een  $\xi_i \in ]\tilde{x}_i, \tilde{x}_{i+1}[$  waarvoor  $\phi'(\xi_i) = 0$ . Als  $\tilde{x}_i = \tilde{x}_{i+1}$  dan impliceert  $\phi = \mathbf{0}$  op  $(\tilde{x}_i, \tilde{x}_{i+1})$  dat  $\phi'(\tilde{x}_i) = 0$ . Kortom, er is een rij  $(\xi_0, \dots, \xi_n)$  waarop  $\phi' = \mathbf{0}$ , met  $\xi_i$  tussen  $\tilde{x}_i$  en  $\tilde{x}_{i+1}$  ( $i = 0, \dots, n$ ) en  $\xi_0 < \xi_n$ . Evenzo volgt dat  $\phi'' = \mathbf{0}$  op een rij  $(\xi'_0, \dots, \xi'_{n-1})$  met  $\xi'_i$  tussen  $\xi_i$  en  $\xi_{i+1}$  ( $i = 0, \dots, n-1$ ) en  $\xi'_0 < \xi'_{n-1}$ . Zo doorgaand kan men inzien dat  $\phi^{(n+1)}$  minstens één nulpunt  $\xi$  heeft op  $(a, b)$ . Deze  $\xi$  is de gezochte; immers, omdat de  $(n+1)$ -ste afgeleide van  $p$  nul en die van  $r$  gelijk aan  $(n+1)!$  is hebben we:

$$0 = \phi^{(n+1)}(\xi) = f^{(n+1)}(\xi) - \alpha(n+1)!, \quad \text{zodat} \quad \alpha = f^{(n+1)}(\xi)/(n+1)! \quad \square$$

**Stelling A.2** *Als  $p$  en  $q$  polynomen zijn van graad hoogstens  $n$  zo dat  $p = q$  op  $(x_0, \dots, x_n)$  dan is  $p = q$ .*

**Bewijs.** Als  $p = q$  op  $(x_0, \dots, x_n)$  dan zegt stelling A.1 met  $f = q$  dat

$$q(x) - p(x) = (x - x_0) \cdots (x - x_n) q^{(n+1)}(\xi)/(n+1)!$$

Aangezien  $q^{(n+1)} = 0$  volgt hieruit  $p = q$ . □

**Stelling A.3** *Er bestaat precies één polynoom  $p$  van graad hoogstens  $n$  zo dat  $p = f$  op  $(x_0, \dots, x_n)$ .*

**Bewijs.** De uniciteit volgt uit stelling A.2.

We passen inductie toe naar  $n$  om de existentie te bewijzen. Het polynoom  $p$  van hoogstens graad  $n$  waarvoor  $p = f$  op  $(x_0, \dots, x_n)$  noteren we met  $p_{(x_0, \dots, x_n)}$ . We moeten de existentie van  $p_{(x_0, \dots, x_n)}$  aantonen. Het is duidelijk dat  $p_{(x_i)}$  bestaat:  $p_{(x_i)}$  is de constant  $f(x_i)$ .

Als  $x_0 = x_n$  (en dus  $x_i = x_0$  alle  $i = 1, \dots, n$ ) dan is  $p_{(x_0, \dots, x_n)}$  het Taylor polynoom

$$p_{(x_0, \dots, x_n)} = \sum_{j=0}^n (x - x_0)^j \frac{f^{(j)}(x_0)}{j!} :$$

deze formule toont in deze situatie aan dat  $p_{(x_0, \dots, x_n)}$  bestaat.

Neem nu aan dat  $x_0 < x_n$  en dat de interpolatie polynomen  $p_0 := p_{(x_0, \dots, x_{n-1})}$  en  $p_1 := p_{(x_1, \dots, x_n)}$  (beide van graad  $n - 1$ ) bestaan (inductie hypothese).

Beschouw

$$q(x) = \frac{x - x_n}{x_0 - x_n} p_0(x) + \frac{x - x_0}{x_n - x_0} p_1(x).$$

Het is duidelijk dat  $q$  een polynoom is van graad ten hoogste  $n$ . Invullen laat zien dat  $q(x_0) = p_0(x_0) = f(x_0)$ ,  $q(x_i) = \frac{x_i - x_n}{x_0 - x_n} f(x_i) + \frac{x_i - x_0}{x_n - x_0} f(x_i) = f(x_i)$ , en  $q(x_n) = f(x_n)$ . Verder is

$$q'(x) = \frac{1}{x_0 - x_n} (p_0(x) - p_1(x)) + \frac{x - x_n}{x_0 - x_n} p_0'(x) + \frac{x - x_0}{x_n - x_0} p_1'(x)$$

en zien we dat  $q'(x_0) = f'(x_0)$  als  $x_0 = x_1$ , etc.. Algemener geldt

$$q^{(j)}(x) = \frac{j}{x_0 - x_n} (p_0^{(j-1)}(x) - p_1^{(j-1)}(x)) + \frac{x - x_n}{x_0 - x_n} p_0^{(j)}(x) + \frac{x - x_0}{x_n - x_0} p_1^{(j)}(x)$$

en er volgt dat  $q = f$  op  $(x_0, \dots, x_n)$ . □

**Opmerking A.2** Merk op dat het bewijs van stelling A.3 een inductieve constructie geeft van het interpolatie polynoom. In feite volgen we met het bewijs de voorwaartse formule van Newton (zie Vraagstuk 9).

De volgende opgave laat zien dat het bewijs van stelling A.3 ook eenvoudig met behulp van een bekend resultaat uit de Lineaire Algebra volgt uit stelling A.2.

**Opgave A.2** Voor de rij  $\mathbf{x} := (x_0, \dots, x_n)$  en de functie  $f$  is de vector  $V_{\mathbf{x}}(f) := (f_0, \dots, f_n)^T$  van functiewaarden en afgeleiden gedefinieerd door

$$f_i := f^{(j)}(x_i) \quad (i = 0, \dots, n) \quad \text{met } j \leq i \quad \text{zodat } x_{i-j-1} < x_i, \quad x_{i-j} = x_i.$$

Hierbij is  $f^{(0)} = f$  en  $x_i := x_{-\infty}$  als  $i < 0$ .

a) Laat zien dat  $V_{(0,0,0,\frac{1}{2},1,1)}(f) = (f(0), f'(0), f''(0), f(\frac{1}{2}), f(1), f'(1))^T$ .

Beschouw nu de afbeelding  $A$  die aan een vector  $(\alpha_0, \alpha_1, \dots, \alpha_n)^T$  in  $\mathbb{R}^n$  de vector  $V_{\mathbf{x}}(p)$  toevoegt met  $p$  het polynoom  $p(x) = \alpha_0 + \alpha_1 x + \dots + \alpha_n x^n$ .

b) Laat zien dat  $A$  een lineaire afbeelding is van  $\mathbb{R}^n$  naar  $\mathbb{R}^n$ .

c) Geef de matrix van  $A$  in geval  $\mathbf{x} = (0, 0, 0, \frac{1}{2}, 1, 1)$ ,

d) Laat zien dat:  $A$  is injectief  $\Leftrightarrow A$  is surjectief  $\Leftrightarrow A$  is bijectief.

e) Formuleer injectiviteit en surjectiviteit van  $A$  in termen van uniciteit en existentie van een interpolatie polynoom.

## B Rombergschema's: stabiliteit en superlineaire convergentie langs de diagonaal

**B1** We veronderstellen dat de benadering  $T(h)$  voor de grootheid  $I$  een asymptotische ontwikkeling heeft van de vorm

$$I - T(h) = c_1 h^2 + c_2 h^4 + c_3 h^6 + \dots \quad (89)$$

(waarmee bedoeld wordt

$$I - T(h) = \sum_{i=1}^m c_i h^{2i} + o(h^{2m}) \quad (90)$$

voor alle  $m \geq 1$ ).

We berekenen voor een zekere rij stapgroottes  $\{h_i\}$  de waarden  $T(h_i)$ , en stellen op basis hiervan een Rombergschema op. Terwille van de eenvoud nemen we  $h_i = h_0(\frac{1}{2})^i$ , maar dat is niet essentieel.

De benaderingen  $T_{ij}$  zijn dan gedefiniëerd in (49):

$$T_{ij} = \frac{4^j T_{i,j-1} - T_{i-1,j-1}}{4^j - 1}. \quad (91)$$

Laat  $\epsilon_0$  een bovengrens zijn voor de totale fout  $R_{i0} = I - T(h_i)$ , dus  $R_{i0}$  is afrondfout plus representatiefout. Als we de afrondfouten die optreden in relatie (91) buiten beschouwing laten, dan voldoet ook  $R_{ij} = I - T_{ij}$  aan (91):

$$R_{ij} = \frac{4^j R_{i,j-1} - R_{i-1,j-1}}{4^j - 1} \quad (92)$$

en dus

$$|R_{ij}| \leq \frac{4^j + 1}{4^j - 1} \max(|R_{i,j-1}|, |R_{i-1,j-1}|). \quad (93)$$

Deze formule herhaald toepassen levert

$$|R_{ij}| \leq \frac{4^j + 1}{4^j - 1} \frac{4^{j-1} + 1}{4^{j-1} - 1} \dots \frac{4^{n+1} + 1}{4^{n+1} - 1} \max_{0 \leq k \leq j-n} |R_{i-k,n}| \leq K \max_{0 \leq k \leq j-n} |R_{i-k,n}| \quad (94)$$

waarin

$$K = \lim_{m \rightarrow \infty} \prod_{k=1}^m \frac{4^k + 1}{4^k - 1} = 1.969 \dots \quad (95)$$

In het bijzonder geldt met  $n = 0$ :

**Stelling B.1 (Stabiliteit van het Rombergschema)** *Bij verwaarlozing van de afrondfouten in het Rombergschema zelf geldt*

$$|R_{ij}| \leq K \epsilon_0. \quad \square$$

**B2** Omdat in elke volgende kolom van het Rombergschema een term van de ontwikkeling (89) geëlimineerd wordt, geldt (als we afzien van afrondfouten)

$$R_{ij} = d_{j,1} h_i^{2j+2} + d_{j,2} h_i^{2j+4} + d_{j,3} h_i^{2j+6} + \dots$$

Bijgevolg is er voor gegeven  $h_0$  voor elke  $j$  een  $a_j$  zo dat

$$|R_{ij}| \leq a_j 4^{-i(j+1)}. \quad (96)$$

Toepassen van (92) geeft achtereenvolgens

$$\begin{aligned} |R_{i,j+1}| &\leq a_j \frac{4^{j+1} 4^{-i(j+1)} + 4^{-(i-1)(j+1)}}{4^{j+1} - 1} = a_j 4^{-i(j+1)} \frac{2}{1 - 4^{-j+1}} \\ |R_{i,j+2}| &\leq a_j \frac{4^{j+2} 4^{-i(j+1)} + 4^{-(i-1)(j+1)}}{4^{j+2} - 1} \frac{2}{1 - 4^{-(j+1)}} = \\ &= a_j 4^{-i(j+1)} \frac{2}{1 - 4^{-(j+1)}} \frac{5/4}{1 - 4^{-(j+2)}} \\ |R_{i,j+3}| &\leq a_j \frac{4^{j+3} 4^{-i(j+1)} + 4^{-(i-1)(j+1)}}{4^{j+3} - 1} \frac{2}{1 - 4^{-(j+1)}} \frac{5/4}{1 - 4^{-(j+2)}} = \\ &= a_j 4^{-i(j+1)} \frac{1 + 4^0}{1 - 4^{-(j+1)}} \frac{1 + 4^{-1}}{1 - 4^{-(j+2)}} \frac{1 + 4^{-2}}{1 - 4^{-(j+3)}} \end{aligned}$$

Algemeen geldt (te bewijzen met inductie)

$$|R_{i,j+k}| \leq a_j 4^{-i(j+1)} \prod_{m=0}^{k-1} \frac{1 - 4^{-m}}{1 - 4^{-(j+m+1)}}$$

zodat

$$|R_{i,j+k}| \leq 2K a_j 4^{-i(j+1)}, \quad (97)$$

met  $K$  als in (95). Met  $j + k = i$ :

$$|R_{ii}| \leq 2K \min_{j \leq i} a_j 4^{-i(j+1)}.$$

We vinden zo

$$\begin{aligned} |R_{00}| &\leq 2K a_0 &&= y_0 \\ |R_{11}| &\leq 2K \min(a_0 4^{-1}, a_1 4^{-2}) &&= y_1 \\ |R_{22}| &\leq 2K \min(a_0 4^{-2}, a_1 4^{-4}, a_2 4^{-6}) &&= y_2 \\ |R_{33}| &\leq 2K \min(a_0 4^{-3}, a_1 4^{-6}, a_2 4^{-9}, a_3 4^{-12}) &&= y_3 \\ |R_{44}| &\leq 2K \min(a_0 4^{-4}, a_1 4^{-8}, a_2 4^{-12}, a_3 4^{-16}, a_4 4^{-20}) &&= y_4 \end{aligned}$$

Zolang in deze minima de term met  $a_0$  de kleinste is, geldt  $y_{i+1}/y_i = 1/4$ . Omdat de term met  $a_1$  steeds per slag een factor 16 zakt en die met  $a_0$  slechts met een factor 4, geldt na een eindig aantal slagen dat de term met  $a_1$  kleiner wordt dan die met  $a_0$ . Vanaf dat moment wint men per slag zeker een factor 16. Weer na een eindig aantal slagen wordt de term met  $a_2$  kleiner dan die met  $a_1$  en  $a_0$ , en vanaf dat moment geldt  $y_{i+1}/y_i \leq 1/64$ , enzovoort. Hiermee is stelling 3.5.1 bewezen voor het geval  $h_i = h_0(\frac{1}{2})^i$ .

## C De Euler Mac-Laurin “reeks”

We bewijzen de stelling:

**Stelling C.1** *Er bestaan getallen  $e_2, e_4, e_6, \dots$  zodanig dat voor  $f \in C^{2m}[a, b]$  en  $n \in \mathbb{N}$*

$$\int_a^b f(x)dx - T(h) = \sum_{i=1}^m h^{2i} e_{2i} [f^{(2i-1)}(b) - f^{(2i-1)}(a)] + o(h^{2m}), \quad (98)$$

waarbij  $T(h)$  het resultaat is van de  $n \times$  gerepeteerde trapeziumregel met stapgrootte  $h = \frac{b-a}{n}$ .

**Bewijs.** We definiëren de rij functies  $\{\varphi_j\}$  en getallen  $\{c_{2j}\}$ :

$$\begin{aligned} \varphi_1(t) &= t \\ \left. \begin{aligned} \varphi_2(t) &= \frac{t^2}{2!} + c_2 \\ \varphi_3(t) &= \frac{t^3}{3!} + c_2 t \end{aligned} \right\} c_2 \text{ zo dat } \varphi_3\left(\frac{1}{2}\right) = \varphi_3\left(-\frac{1}{2}\right) = 0 \\ \left. \begin{aligned} \varphi_4(t) &= \frac{t^4}{4!} + c_2 \frac{t^2}{2!} + c_5 \\ \varphi_5(t) &= \frac{t^5}{5!} + c_2 \frac{t^3}{3!} + c_4 t \end{aligned} \right\} c_4 \text{ zo dat } \varphi_5\left(\frac{1}{2}\right) = \varphi_5\left(-\frac{1}{2}\right) = 0 \\ &\text{etc.} \end{aligned}$$

zodat  $\varphi_j = \varphi'_{j+1}$ . Definiëer nog  $e_j = \varphi_j\left(\frac{1}{2}\right)$ , dus  $e_{2j+1} = 0$  voor  $j > 0$ . Merk op dat  $\varphi_{2j}\left(\frac{1}{2}\right) = \varphi_{2j}\left(-\frac{1}{2}\right)$ . Dan geldt voor  $g \in C^{(2m)}\left[-\frac{1}{2}, \frac{1}{2}\right]$ :

$$\begin{aligned} \int_{-\frac{1}{2}}^{\frac{1}{2}} g(t) dt &= \int_{-\frac{1}{2}}^{\frac{1}{2}} g(t) d\varphi_1(t) = g(t)\varphi_1(t) \Big|_{-\frac{1}{2}}^{\frac{1}{2}} - \int_{-\frac{1}{2}}^{\frac{1}{2}} g'(t)\varphi_1(t) dt = \\ &= g(t)\varphi_1(t) \Big|_{-\frac{1}{2}}^{\frac{1}{2}} - \int_{-\frac{1}{2}}^{\frac{1}{2}} g'(t) d\varphi_2(t) = g(t)\varphi_1(t) \Big|_{-\frac{1}{2}}^{\frac{1}{2}} - g'(t)\varphi_2(t) \Big|_{-\frac{1}{2}}^{\frac{1}{2}} + \\ &+ \int_{-\frac{1}{2}}^{\frac{1}{2}} g''(t)\varphi_2(t) dt \quad \text{etc.} \end{aligned} \quad (99)$$

zodat men met volledige inductie krijgt

$$\begin{aligned} \int_{-\frac{1}{2}}^{\frac{1}{2}} g(t) dt &= \sum_{j=0}^{2m-1} (-1)^j g^{(j)}(t)\varphi_{j+1}(t) \Big|_{-\frac{1}{2}}^{\frac{1}{2}} + \int_{-\frac{1}{2}}^{\frac{1}{2}} g^{(2m)}(t)\varphi_{2m}(t) dt = \\ &= \frac{1}{2} [g\left(\frac{1}{2}\right) + g\left(-\frac{1}{2}\right)] + \sum_{j=1}^m e_{2j} [g^{(2j-1)}\left(\frac{1}{2}\right) - g^{(2j-1)}\left(-\frac{1}{2}\right)] + \\ &+ \int_{-\frac{1}{2}}^{\frac{1}{2}} g^{(2m)}(t)\varphi_{2m}(t) dt. \end{aligned} \quad (100)$$

Beschouw nu

$$\int_a^b f(x) dx = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x) dx = h \sum \int_{-\frac{1}{2}}^{\frac{1}{2}} f(m_i + ht) dt \quad (101)$$

met  $x_i = a + ih$ ,  $h = (b - a)/n$ ,  $m_i = (x_i + x_{i+1})/2$ . Op ieder van de termen van de tweede som passen we het voorgaande toe met  $g(t) = f(m_i + ht)$ . Dan is  $g^{(2j-1)}(t) = h^{2j-1} f^{(2j-1)}(m_i + ht)$ . Dus

$$\begin{aligned} \int_a^b f(x) dx &= T(h) + \sum_{j=1}^m h^{2j} e_{2j} [f^{(2j-1)}(b) - f^{(2j-1)}(a)] + \\ &+ h \sum_{i=0}^{n-1} h^{2m} \int_{-\frac{1}{2}}^{\frac{1}{2}} f^{(2m)}(m_i + ht) \varphi_{2m}(t) dt \end{aligned} \quad (102)$$

Wegens de continuïteit van  $f^{(2m)}$  is er bij elke  $\epsilon > 0$  een  $h_0 > 0$  zodat  $|f^{(2m)}(m_i + ht) - f^{(2m)}(m_i)| < \epsilon$  voor  $h < h_0$  en  $|t| \leq \frac{1}{2}$ , zodat, met  $\beta_m = \int_{-\frac{1}{2}}^{\frac{1}{2}} |\varphi_{2m}(t)| dt$

$$\left| \int_{-\frac{1}{2}}^{\frac{1}{2}} [f^{(2m)}(m_i + ht) - f^{(2m)}(m_i)] \varphi_{2m}(t) dt \right| < \beta_m \epsilon. \quad (103)$$

Echter,  $\int_{-\frac{1}{2}}^{\frac{1}{2}} \varphi_{2m}(t) dt = \varphi_{2m+1}(\frac{1}{2}) - \varphi_{2m+1}(-\frac{1}{2}) = 0$ .

Dus  $\left| \int_{-\frac{1}{2}}^{\frac{1}{2}} f^{(2m)}(m_i + ht) \varphi_{2m}(t) dt \right| < \beta_m \epsilon$ . Hiermee is bewezen dat de term  $h \sum$  in (102) te schrijven is als  $o(h^{2m})$ .  $\square$

De getallen  $B_{2i} = -(2i)!e_{2i}$  noemt men Bernoulli getallen. Men vindt ze bijv. in Abramowitz-Stegun. Men kan ze ook zelf uitrekenen door (98) bijv. toe te passen met  $b = 1$ ,  $a = 0$ ,  $h = 1$ ,  $f(x) = x^{2m}$  met achtereenvolgens  $m = 1, 2, \dots$  en te bedenken dat dan de term  $o(h^{2m})$  ontbreekt omdat  $f^{(2m)}$  dan steeds constant is (zie in de afleiding waar de o-termen vandaan komen). Dat geeft

$$\frac{1}{2m+1} = \frac{1}{2} + \sum_{i=1}^m e_{2i} (2m)(2m-1) \cdots (2m-2i+2)$$

en men vindt  $e_2 = \frac{-1}{12}$ ,  $e_4 = \frac{1}{720}$ ,  $e_6 = \frac{-1}{30240}, \dots$



## 1 Interpolatie

### Vraagstuk 1

- a) Toon aan dat de absolute fout bij equidistante 2e orde Lagrange interpolatie op een interval  $[a, b]$  van een 3-maal continu differentieerbare functie  $f$  niet groter is dan

$$\frac{1}{36}\sqrt{3}(b-a)^3 \max_{x \in [a,b]} \frac{|f'''(x)|}{3!}.$$

- b) Leidt een dergelijke uitdrukking af voor equidistante 3e orde Lagrange interpolatie.

**Vraagstuk 2** Stel iemand wil een tabel van goniometrische functies opstellen om daarin te kunnen interpoleren.

- a) Met welke stapgrootte  $h$  moet men  $\sin(x)$  tabuleren opdat de absolute fout bij lineair interpoleren gegarandeerd kleiner is dan  $\frac{1}{2} \cdot 10^{-4}$ ?

Zij  $p(x)$  het lineaire interpolatiepolynoom van  $\sin(x)$  op de steunpunten  $x_0$  en  $x_0 + h$ . Uit Stelling 1.1.1 volgt dat

$$\sin(x) - p(x) = -\frac{\sin(\xi)}{2}(x - x_0)(x - x_0 - h), \quad \xi \in (x_0, x_0 + h). \quad (1)$$

Vanwege  $\sin(\xi) \approx \sin(x) \approx p(x)$ , wordt de hierboven gegeven fout bij lineaire interpolatie benaderd door

$$-\frac{1}{2}p(x)(x - x_0)(x - x_0 - h).$$

Deze benadering gaan we nu gebruiken om het resultaat te corrigeren. D.w.z. als nieuwe benadering van  $\sin(x)$  beschouwen we

$$p(x)\left(1 - \frac{1}{2}(x - x_0)(x - x_0 - h)\right). \quad (2)$$

- b) Laat zien dat  $\forall x \in [x_0, x_0 + h]$ ,  $\exists \eta \in [x_0, x_0 + h]$  z.d.d.  $p(x) = \sin(\eta)$ . Bewijs hiermee dat de benadering (2) een absolute fout van ten hoogste  $\frac{h^3}{8}$  heeft. Hoe groot mag  $h$  zijn opdat (2) een absolute fout kleiner dan  $\frac{1}{2}10^{-4}$  oplevert?

- c) Bereken voor  $x = 0.41$  de benaderingen  $p(x)$  en (2) gebruikmakend van

$$\sin(0.4) = 0.3894183423$$

$$\sin(0.5) = 0.4794255386$$

- d) Volgens Opgave 1.1.2 is  $\zeta = \frac{x_0 + x_0 + h + x}{3}$  een goede benadering van  $\xi$  uit (1). Bepaal de lineaire interpolatie benadering van  $\sin(\zeta)$ . Benader hiermee de grootte

$$p(x) - \frac{1}{2}\sin(\zeta)(x - x_0)(x - x_0 - h).$$

Vergelijk de drie gevonden antwoorden met de waarde

$$\sin(0.41) = 0.3986093279.$$

**Vraagstuk 3** a) Ga na dat  $\tan'(x) = 1 + \tan^2(x)$  en dus  $\tan''(x) = 2\tan(x)(1 + \tan^2(x))$

- b) Geef een uitdrukking voor de fout bij lineaire *interpolatie* op de tangens op een interval ter lengte  $h$  welke geen punt van de vorm  $k\pi + \frac{\pi}{2}$  ( $k \in \mathbb{Z}$ ) bevat. Gebruik deze uitdrukking om een gecorrigeerde benadering op te stellen waarvoor de maximale absolute fout  $\mathcal{O}(h^3)$  is.
- c) Benader m.b.v. onderstaande tabel aldus  $\tan(0.7855)$ . Hoe groot is de fout?

$x$	$\tan(x)$
0.784	0.99720758
0.785	0.99920399
0.786	1.00120440
0.787	1.00320882

- d) Hoe fijn moet een tabulering op  $[\pi/6, \pi/3]$  zijn om op de boven geschetste wijze te kunnen interpoleren z.d.d. de relatieve fout gegarandeerd  $\leq \frac{1}{2}10^{-6}$  is?

#### Vraagstuk 4

$x$	$\tan(x)$	$\cot(x)$
1.566	208.5	0.004796
1.567	263.4	0.003796
1.568	357.6	0.002796

- a) Bij lineaire interpolatie in de buurt van  $x = 1.567$  verwacht men moeilijkheden. Waarom? (N.B.  $\pi/2 = 1.5708$ .) Geef een ondergrens voor de maximale absolute fout die gemaakt wordt bij lineaire interpolatie tussen 1.566 en 1.568. Wat verwacht u van toepassing van de benadering uit vraagstuk 3b)? Waarom?
- b) Benader  $\tan(1.567)$  via lineaire interpolatie op de cotangens. Geef een bovengrens voor de fout, als 1.566 en 1.568 als steunpunten gebruikt worden.
- c) Voor welke waarden van  $x$  geldt:  $\tan x = 300$ ? Benader deze  $x$  met inverse lineaire interpolatie en geef een bovengrens voor de fout. Merk op dat deze interpolatie weinig extra informatie geeft.
- d) Benader  $\tan(x) = 300$  met inverse lineaire interpolatie op de cot. Geef een bovengrens voor de fout. Vergelijk de antwoorden in c. en d. met  $\arctan(300) = 1.5674630$ .

**Vraagstuk 5** Gegeven zijn in onderstaande tabel de waarden van  $\sin(x)$  voor  $x = 0.0(0.2)1.0$ . Gevraagd wordt  $\sin(0.1)$ .

- a) Hoeveelste orde interpolatie moet men gebruiken om met Newton's voorwaartse formule een absolute nauwkeurigheid van  $\frac{1}{2}10^{-5}$  te halen als  $x_0 = 0.0$ ?
- b) En hoeveelste orde Gauß interpolatie met  $x_0 = 0.2$ ?

- c) Waar moet de  $x_0$  voor Newton interpolatie gekozen worden om hetzelfde resultaat te krijgen als bij Gauß interpolatie?
- d) Maak plausibel dat bij voldoende hoge orde, de absolute fout bij Gauß interpolatie kleiner is dan die bij Newton interpolatie, als bij beiden de orde en  $x_0$  hetzelfde genomen worden.

$x$	$\sin(x)$
0.0	0.000000
0.2	0.198669
0.4	0.389418
0.6	0.564642
0.8	0.717356
1.0	0.841471

**Vraagstuk 6** In deze opgave willen we nagaan hoe het staat met de nauwkeurigheid waarmee men een functie kan benaderen als men in 4 steunpunten over de functiewaarden en over de waarden van de afgeleiden beschikt. Zij  $f$  4 keer continu differentieerbaar. We vergelijken nu de uitkomsten op het interval  $[0, h]$  van:

- (1) 4-punts Lagrange interpolatie op de steunpunten  $-h, 0, h$  en  $2h$ .
  - (2) 2-punts Hermite interpolatie op de steunpunten  $0, h$  (d.w.z. gebruik  $f(0), f'(0), f(h)$  en  $f'(h)$ ).
  - (3) Een Taylor reeks rond  $x = 0$ , afgekapt na de 3e graadsterm.
  - (4) Een Taylor reeks rond  $x = \frac{1}{2}h$ , afgekapt na de 3e graadsterm.
- a) Ga na dat in elk der 4 genoemde methoden de fout gegeven wordt door  $R_i(x) = \frac{f^{(4)}(\xi)}{4!} \pi_i(x)$  met  $\pi_i(x)$  een 4e graads polynoom behorend bij methode (i).
- b) Laat zien waar  $\pi_i$  op  $[0, h]$  extremen aanneemt (inwendig- of randextremen). Bereken de extremen.
- c) Bewijs: voor alle  $x \in ]0, h[$ :  $\pi_1(x) > \pi_2(x) > 0$ . Schets nu  $\pi_1, \pi_2, \pi_3$  en  $\pi_4$  op het segment  $[0, h]$ .

**Vraagstuk 7** Gegeven is een equidistante tabel van de functie  $f \in C^4$ , stapgrootte  $h$ , waarbij  $f(x_0)$  en  $f(x_0 + h)$  verschillend teken hebben.

- a) We bepalen het tussenliggende nulpunt  $\alpha$  met behulp van inverse lineaire interpolatie. Bewijs dat het resultaat  $\tilde{\alpha}$  een absolute fout van hoogstens

$$\frac{h^2}{8} \left| \frac{f''(\alpha)}{f'(\alpha)} \right| + \mathcal{O}(h^3) \quad (h \downarrow 0)$$

heeft (mits  $f'(\alpha) \neq 0$ ). (Ter herinnering: als  $f^{-1}$  de inverse functie van  $f$  is, dan  $(f^{-1})''(y) = -\frac{f''(x)}{(f'(x))^3}$  waarbij  $y = f(x)$ ).

- b) Veronderstel dat de tabelwaarden van  $f$  een relatieve fout  $\epsilon$  hebben, met  $|\epsilon| \leq \bar{\epsilon} \ll 1$ . Bewijs dat dit een extra absolute fout van hoogstens  $\approx 2\bar{\epsilon}h$  in  $\tilde{\alpha}$  veroorzaakt. (Hint: schrijf  $\tilde{\alpha} = x_0 - \frac{hf(x_0)}{f(x_0+h)-f(x_0)}$ ).

**Vraagstuk 8** We bezien in deze opgave het polynoom

$$\omega_n(x) = x(x-1)(x-2)\cdots(x-n)$$

en we nemen aan dat  $n$  oneven is.

a) Laat zien dat  $\omega_n$  een lokaal extreem heeft in  $x = \frac{n}{2}$ , en dat

$$\frac{n!}{2^{n+1}} \leq |\omega_n(\frac{n}{2})| \leq (n+1)\frac{n!}{2^{n+1}}.$$

b) Laat zien dat  $\omega_n$  ook op  $[0, 1]$  een lokaal extreem heeft. Laat dit extreem aangenomen worden voor  $x = y$ . Laat zien dat

$$y(1-y) \leq \left| \frac{\omega_n(y)}{(n-1)!} \right| \leq ny(1-y).$$

c) Laat zien dat  $y$  voldoet aan

$$\frac{1}{y} = \frac{1}{1-y} + \frac{1}{2-y} + \cdots + \frac{1}{n-y}.$$

(Hint:

$$\omega_n'(x) = \sum_{j=0}^n \frac{\omega_n(x)}{x-j}.)$$

Toon hiermee aan:  $y < \frac{1}{2}$  voor  $n \geq 1$ .

d) Bewijs dat

$$\begin{aligned} \log n &\leq 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n} \leq \frac{1}{y} \leq \frac{1}{2} + \frac{1}{1\frac{1}{2}} + \frac{1}{2\frac{1}{2}} + \cdots + \frac{1}{n - \frac{1}{2}} \\ &\leq 2 + \log n. \end{aligned}$$

e) Laat zien:

$$\frac{1}{n(n+1)} \frac{2^n}{(2 + \log n)} \leq \left| \frac{\omega_n(y)}{\omega_n(\frac{n}{2})} \right| \leq \frac{2^{n+1}}{\log n}.$$

Ga na dat voor grote waarden van  $n$  deze grenzen ongeveer een factor  $2n^2$  verschillen.

f) Becommentarieer dit resultaat in verband met Lagrange interpolatie.

**Vraagstuk 9** Zij  $x_0, x_1, x_2, \dots$  een willekeurige rij van verschillende reële getallen, en zij  $f$  een voldoende gladde functie  $\mathbb{R} \rightarrow \mathbb{R}$ . Zij  $p_n$  het Lagrange interpolatiepolynoom van  $f$  op  $x_0, \dots, x_n$ .

a) Voor  $n \geq 1$ , bewijs dat er een constante  $a_n$  is z.d.d.

$$p_n(x) = p_{n-1}(x) + a_n(x-x_0)\cdots(x-x_{n-1}). \quad (3)$$

De constante  $a_n$  is dus de leidende coëfficiënt van  $p_n$ . We schrijven in het vervolg  $a_n$  als  $f[x_0, \dots, x_n]$ . Duidelijk is dat  $f[x_0, \dots, x_n]$  onafhankelijk is van de ordening van de steunpunten, en dat de formule (3) gebruikt kan worden voor het toevoegen van een willekeurig nieuw steunpunt aan een willekeurig rijtje van bestaande steunpunten. We definiëren nog

$$f[x_i] = f(x_i), \quad (4)$$

zodat het nuldegraads interpolatie polynoom van  $f$  met steunpunt  $x_i$  gegeven wordt door  $x \mapsto f[x_i]$ .

Formule (3) suggereert een efficiënte berekeningswijze voor het verhogen van de orde. Hiervoor dienen we dan wel een manier te vinden om de coëfficiënten  $f[\dots]$  te berekenen:

- b) Veronderstel dat het interpolatiepolynoom  $q$  van  $f$  met steunpunten  $x_1, \dots, x_{n-1}$  beschikbaar is. Geef m.b.v. (3) nu twee formules voor  $p_n(x)$ . De eerste formule door eerst  $x_0$  toe te voegen en vervolgens  $x_n$ , de tweede formule door deze steunpunten in omgekeerde volgorde toe te voegen. Laat nu zien dat

$$f[x_0, \dots, x_n] = \frac{f[x_1, \dots, x_n] - f[x_0, \dots, x_{n-1}]}{x_n - x_0} \quad (5)$$

- c) Zij  $x_i = i$ ,  $f(x_0) = 1$ ,  $f(x_1) = 3$  en  $f(x_2) = 2$ . Laat zien hoe je m.b.v. (3), (4) en (5) de interpolatiepolynomen  $p_0$ ,  $p_1$  en  $p_2$  opstelt.

(Newton's en Gauß' voorwaartse formules welke genoemd worden in 1.2F zijn rechtstreekse toepassingen van de hier aangegeven berekeningswijze).

- d) Geef een uitdrukking voor  $f(x_n) - p_{n-1}(x_n)$ . Merk op dat  $f(x_n) = p_n(x_n)$ , en laat m.b.v. (3) zien dat er een  $\xi \in ]x_0, \dots, x_n[$  is z.d.d.

$$f[x_0, \dots, x_n] = \frac{f^{(n)}(\xi)}{n!}$$

- e) Vergelijk nu  $p_n(x) - p_{n-1}(x)$  met de fout  $f(x) - p_{n-1}(x)$ , en beschrijf een methode waarmee de interpolatie orde telkens verhoogd wordt totdat een schatting van de fout beneden een opgegeven tolerantie is.

## 2 Numerieke differentiatie

**Vraagstuk 10** Zij gegeven een equidistante getabelleerde voldoende gladde functie  $f$ . We willen de afgeleide  $f'(t)$  bepalen in een willekeurig punt  $t$  zodat  $-\frac{1}{2}h \leq t \leq \frac{1}{2}h$ . Hiervoor staan verschillende methoden ter beschikking.

a) Toon met behulp van Taylor reeksen aan:

$$f'(t) = \frac{(2t+h)f(h) - 4tf(0) + (2t-h)f(-h)}{2h^2} + R,$$

$$R = \frac{1}{6}f'''(\xi)(3t^2 - h^2)$$

$$(\text{u mag ook aantonen dat } R = \frac{1}{6}f'''(0)(3t^2 - h^2) + \mathcal{O}(h^3)).$$

Ga na, dat men dezelfde formule (en dus ook dezelfde restterm) verkrijgt, als men

- b)  $f'(-h/2)$  en  $f'(h/2)$  benadert met  $\frac{f(0)-f(-h)}{h}$  resp.  $\frac{f(h)-f(0)}{h}$  en daarop lineaire interpolatie toepast, of
- c) op  $f(-h)$ ,  $f(0)$  en  $f(h)$  kwadratische interpolatie toepast, en het verkregen polynoom differentieert.

**Vraagstuk 11** Stel er is een functie  $f : \mathbb{R} \rightarrow \mathbb{R}$  gegeven, die 4-maal continu differentieerbaar is. Als we de tweede afgeleide van  $f$  in  $x_0$  numeriek willen bepalen kunnen we gebruik maken van formule (23). Stel  $M_4$  is een majorant van  $|f^{(4)}(x)|$  in een omgeving van  $x_0$  en de onbetrouwbaarheid in de functiewaarden is  $\epsilon$ . Bepaal de waarde  $h$ , uitgedrukt in  $M_4$  en  $\epsilon$ , waarvoor de majorant van de totale fout minimaal is.

**Vraagstuk 12** Men doet een experiment waarbij een functie  $f : [0, 1] \rightarrow \mathbb{R}$ , met  $f \in C^\infty[0, 1]$ , gemeten kan worden. Men is geïnteresseerd in een nauwkeurige numerieke benadering van  $f'$ . Men doet daartoe eerst een experiment, waarbij men  $f$  bepaalt in de equidistant verdeelde tijdstippen  $t_i = ih$  met  $i \geq 0$ , met de grove stapgrootte  $h = 0.1$ . M.b.v. deze gegevens bepaalt men een schatting voor de optimale stapgrootte  $h_{\text{opt}}$  zodat de totale fout in de benadering van  $f'$  minimaal is. Daarna herhaalt men het experiment en meet  $f$  in de tijdstippen  $t_i = ih_{\text{opt}}$ .

**Tabel** (eerste experiment)

$t_i$	$f(t_i)$
0.0	-0.69314718
0.1	-0.51082562
0.2	-0.35667494
0.3	-0.22314355
0.4	-0.10536051
0.5	0.0
0.6	0.09531017
0.7	0.18232155
0.8	0.26236426
0.9	0.33647223
1.0	0.40546510

- a) We benaderen  $f'(x_0)$  met formule (19). Bewijs (20). De onbetrouwbaarheid in de functiewaarden, d.w.z., een bovenschatting voor de absolute waarde van de fout, is  $10^{-7}$ . Geef een schatting van de optimale stapgrootte  $h_1$  en de majorant van de totale fout bij deze  $h_1$ .
- b) Voor de benadering van  $f'(x_0)$  kunnen we ook gebruik maken van (30). Zij  $R_4(h)$  als volgt gedefinieerd:

$$f'(x_0) = \frac{1}{12h} \{f(x_0 - 2h) - 8f(x_0 - h) + 8f(x_0 + h) - f(x_0 + 2h)\} + R_4(h).$$

Het is niet eenvoudig om m.b.v. Taylor reeksen te bewijzen dat  $R_4(h) = C_1 h^4 f^{(5)}(\xi)$ ,  $\xi \in [x_0 - 2h, x_0 + 2h]$ . Waarom niet? Bewijs de schatting

$$|R_4(h)| \leq c_2 h^4 \max_{\xi \in [0,1]} |f^{(5)}(\xi)|$$

en bepaal  $c_2$ . Toon ook aan dat  $R_4(h) = c_3 h^4 f^{(5)}(x_0) + \mathcal{O}(h^6)$  en bepaal  $c_3$ .

- c) Geef een uitdrukking voor de onbetrouwbaarheid in deze benadering (formule (30)) voor  $f'(x_0)$  ten gevolge van een onbetrouwbaarheid  $\epsilon$  in de functiewaarden. Geef voor  $\epsilon = 10^{-7}$  een schatting van de stapgrootte  $h_2$  zodat de majorant van de totale fout minimaal is. Hoe groot is deze majorant dan?

Zie in dat men bij de eerste methode meer werk moet doen om de tabel voor de stapgrootte  $h_1$  op te stellen en dat de totale fout  $\alpha$  groter is dan bij de tweede methode met stapgrootte  $h_2$ .

De vergelijking van de hoeveelheid werk is nog gunstiger voor de tweede methode als we in plaats van  $h_2$  een zo groot mogelijke stapgrootte  $\hat{h}$  nemen zodanig dat de totale fout gelijk is aan  $\alpha$ . Dit leidt echter tot de 5<sup>e</sup> graadsvergelijking

$$\frac{18\epsilon}{12h} + \frac{48}{12 \cdot 120} h^4 114 = \alpha.$$

### 3 Foutschattingen bij numerieke processen

**Vraagstuk 13** Hieronder is een tabel aangegeven met functiewaarden van  $f(x) = x^3 + x + 1$ .

- Bepaal m.b.v. inverse lineaire interpolatie het nulpunt van deze functie. (Toon eerst aan, dat  $f$  lokaal een inverse heeft.)
- Noteer de (locale) inverse van  $f$  met  $g$ , en het gebruikte interpolatiepolynoom van  $g$  met  $q$ . Zij  $R(y) = g(y) - q(y)$ . Geef een afschatting voor  $\max |R(y)|$ , waarbij het maximum genomen is over alle  $y$  in  $[-0.043, 0.184]$
- Geef ook een schatting voor  $R(0)$ .

$x$	$f(x)$
-1.0	-1.000
-0.9	-0.629
-0.8	-0.312
-0.7	-0.043
-0.6	+0.184
-0.5	+0.375
-0.4	+0.536
-0.3	+0.673
-0.2	+0.792
-0.1	+0.899
0.0	+1.000

**Vraagstuk 14** Stel, we willen met een rekenapparaat (dat aritmetische bewerkingen uitvoert met een relatieve nauwkeurigheid  $\xi$ ,  $|\xi| \leq \bar{\xi}$ ) voor twee getallen  $a$  en  $b$ :  $a^2 - b^2$  berekenen. Dit kan op twee manieren:

$$\begin{array}{l} \underline{\text{A}} \quad a^2 - b^2 = (a + b) \cdot (a - b); \\ \underline{\text{B}} \quad a^2 - b^2 = a \cdot a - b \cdot b. \end{array}$$

- Laat  $a$  en  $b$  machinegetallen zijn. Toon aan, dat de relatieve fout in het resultaat berekend volgens A hoogstens  $3\bar{\xi}$  is, en volgens B  $(1 + (a^2 + b^2)/|a^2 - b^2|)\bar{\xi}$ . Welke methode zou u prefereren en waarom?
- Stel nu, dat  $a$  en  $b$  geen machinegetallen zijn. Toon aan, dat de relatieve fouten volgens A en B nu worden gemajoreerd door resp.

$$(3 + 2(a^2 + b^2)/|a^2 - b^2|)\bar{\xi} \quad \text{en} \quad (1 + 3(a^2 + b^2)/|a^2 - b^2|)\bar{\xi}.$$

### Vraagstuk 15

- a) Gegeven is:  $\sin(0.47380) = 0.45627$ , en  $\sin(0.47370) = 0.45618$ . Bereken hiermee  $y = \sin(0.47380) - \sin(0.47370)$ . Bereken  $y$  ook eens m.b.v. de formule:

$$\sin(a) - \sin(b) = 2 \sin((a-b)/2) \cos((a+b)/2).$$

(Gebruik, dat  $\cos(0.47375) = 0.88986$  en  $\sin(0.00005) = 0.50000 \cdot 10^{-4}$ .) Vergelijk de uitkomsten met het juiste antwoord (in 5 decimalen):

$$8.8986 \cdot 10^{-5}.$$

- b) Onder de aanname dat de elementaire operaties  $+$ ,  $-$ ,  $\times$  en  $/$  exact uitgevoerd worden, geef voor twee machinegetallen  $a$  en  $b$  de relatieve fouten in de resultaten van  $\sin(a) - \sin(b)$  volgens de ‘directe’ methode en volgens de formule.
- c) Hoe zou u de volgende grootheden berekenen? Beargumenteer uw berekeningswijze middels een afrondfoutenanalyse onder de aanname als in b).

$$x^{\frac{1}{2}} - (x-1)^{\frac{1}{2}} \quad \text{voor } x \gg 1;$$

$$\log(x) - \log(x-1) \quad \text{voor } x \gg 1;$$

$$(e^x - 1)/x \quad \text{voor } |x| \ll 1.$$

**Vraagstuk 16** Gegeven is een polynoom  $ax^2 + bx + c$ , met  $a, b, c \neq 0$  en  $b^2 - 4ac > 0$ . De wortels van dit polynoom duiden we aan met  $\lambda_1, \lambda_2$  waarbij  $|\lambda_1| > |\lambda_2|$ . We willen deze wortels berekenen. Neem aan dat  $b > 0$  (geen beperking) en laat  $dq := \frac{b^2}{4ac}$ . Neem verder aan dat  $a, b, c$  niet persé machinegetallen zijn. De elementaire rekenoperaties hebben relatieve onbetrouwbaarheid  $\bar{\xi}$  en  $(\sqrt{x})^* = \sqrt{x}(1 + \xi)$  met  $|\xi| \leq \bar{\xi}$ . We bekijken de situatie waarbij  $|dq| \geq 2$ .

- a) Ga na dat  $(\sqrt{b^2 - 4ac})^*$  een relatieve onbetrouwbaarheid  $(4\frac{1}{2} + \frac{4}{|dq|})\bar{\xi}$  heeft.
- b) Ga na dat de bekende  $a, b, c$ -formule:  $\lambda_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$  goed werkt voor de absoluut grootste wortel  $\lambda_1$ . Ga na dat de formule slecht werkt voor de absolute kleinste wortel  $\lambda_2$  als  $|dq| \gg 1$ .
- c) Een andere formule voor  $\lambda_2$  is:

$$\lambda_2 = \frac{-2c}{b + \sqrt{b^2 - 4ac}}$$

(in te zien met  $\lambda_1 \lambda_2 = \frac{c}{a}$ ) Ga na dat deze formule wel goed werkt (als  $|dq| \geq 2$ ).

**Vraagstuk 17** Gevraagd te berekenen de integralen  $I_n = \int_0^1 e^{-t} t^n dt$ ,  $n = 0, \dots, 100$ . Met partiële integratie ziet men in, dat  $I_n = nI_{n-1} - 1/e$ . Aangezien  $I_0 = 1 - 1/e$ , kan men de  $I_n$  dus recursief berekenen.

- a) Stel  $I_0$  heeft een fout  $\delta$ ,  $|\delta| \leq 10^{-16}$ , veroorzaakt door afronding. Ga na, dat zelfs al zou exact gerekend worden, de fout  $\delta$  kan bewerkstelligen, dat  $I_n$  vanaf  $n \approx 17$  geen enkel significant cijfer bevat. Ga dus na, dat de relatieve fouten vanaf  $n \approx 17$  groter dan 1 worden. Toon hiertoe o.a. aan, dat

$$\frac{1}{(n+1)e} \leq I_n \leq \frac{1}{n+1}.$$

- b) Stel een recursie op voor  $J_n = \int_1^\infty e^{-t^n} dt$ ,  $n = 0, \dots, 100$ . Ga na, dat de afrondfout  $\delta$  in  $J_0$ , onder aanname van verder exact rekenen, in dit geval niet leidt tot snel groeiende relatieve fouten in  $J_n$ .
- c) Wat denkt u van het berekenen van  $I_n$  met behulp van

$$I_n = \int_0^\infty e^{-t^n} dt - J_n = n! - J_n?$$

- d) Men zou  $I_n$  voor  $n = 1, \dots, 100$  kunnen berekenen door de integrand in een Taylor reeks te ontwikkelen en term voor term te integreren. Stel, dat elke term nu behept is met een relatieve fout  $\delta_k$ , met  $|\delta_k| \leq \delta$ . Ga na, dat de bijdrage van deze  $\delta_k$ 's tot de relatieve fout in het eindantwoord begrensd wordt door  $e^2 \delta$ . (Bij de evaluatie van de aldus ontstane reeks behoeven we dus geen instabiliteit te vrezen.)

**Vraagstuk 18** Stel, we wensen een niet exact bekende grootheid, zeg  $I$ , te benaderen en hiervoor hebben we een benaderingsmethode  $T_1(h)$ . We gaan ervan uit, dat

$$I - T_1(h) = ch^\alpha + \mathcal{O}(h^\beta), \quad \text{voor } h \rightarrow 0, \text{ met } 0 < \alpha < \beta.$$

- a) Bewijs dat

$$Q_1(h) := \frac{T_1(h) - T_1(\frac{1}{2}h)}{T_1(\frac{1}{2}h) - T_1(\frac{1}{4}h)} = 2^\alpha + \mathcal{O}(h^\epsilon) \quad \text{voor } h \rightarrow 0$$

met  $\epsilon = \beta - \alpha$ .

- b) Laat zien dat er precies één paar  $(\lambda, \mu) \in \mathbb{R}^2$  is, zodat voor de nieuwe benaderingsformule

$$T_2(h) := \lambda T_1(h) + \mu T_1(\frac{1}{2}h)$$

geldt:

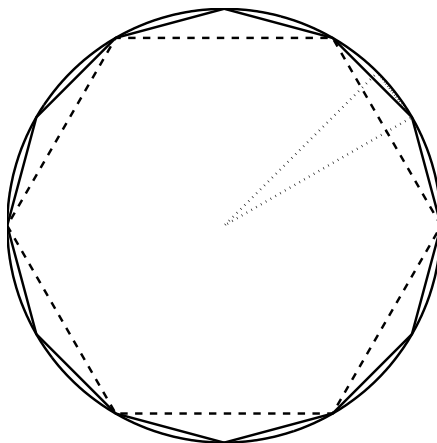
$$I - T_2(h) = \mathcal{O}(h^\beta) \quad \text{voor } h \rightarrow 0,$$

en geef een uitdrukking voor  $\lambda$  en  $\mu$  in termen van  $\alpha$ .

**Vraagstuk 19** Archimedes (250 v. Chr.) verkreeg onder- en bovengrenzen voor  $\pi$  door het opmeten van de omtrek van regelmatige ingeschreven danwel omgeschreven veelhoeken van een cirkel met diameter 1. In deze opgave beperken we ons tot de metingen van de ingeschreven veelhoeken.

- a) Zij  $T_0(h)$  de omtrek van de  $n$ -zijdige regelmatige ingeschreven veelhoek waarbij  $nh = 1$ . Bewijs dat  $T_0(h) = h^{-1} \sin(\pi h)$ , zie figuur 9.
- b) Laat zien dat er constanten  $(c_i)$  zijn zo dat  $\forall m \in \mathbb{N}$

$$\pi - T_0(h) = \sum_{i=1}^m c_i h^{2i} + \mathcal{O}(h^{2m+2}) \quad (h \rightarrow 0).$$



FIGUUR 9: *Ingeschreven regelmatige veelhoek. Merk op dat de hoek tussen de gestippelde lijnen gelijk is aan  $\frac{1}{2}\frac{2\pi}{n}$ . In het figuur is  $n = 12$ .*

- c) Bepaal  $\alpha_1, \beta_1$  z.d.d.  $T_1(h/2) := \alpha_1 T_0(h/2) + \beta_1 T_0(h)$  voldoet aan

$$\pi - T_1(h/2) = \mathcal{O}(h^4) \quad (h \rightarrow 0).$$

Huygens gebruikte dit idee al in 1654. Archimedes' metingen liepen tot  $n = 96$ . Aannemende dat Huygens de metingen van Archimedes gebruikte, en aannemende dat die metingen exact waren, wat waren de fouten in de beste benaderingen die beiden verkregen?

- d) Verbeter Huygens, d.w.z. bepaal  $\alpha_2, \beta_2$  z.d.d.  $T_2(h/4) := \alpha_2 T_1(h/4) + \beta_2 T_1(h/2)$  voldoet aan

$$\pi - T_2(h/4) = \mathcal{O}(h^6) \quad (h \rightarrow 0).$$

Wat is de fout in  $T_2(1/96)$ ?

## 4 Numerieke integratie

**Vraagstuk 20** We bezien de volgende kwadratuurformule:

$$\int_0^h f(x) dx = w_1 f(0) + w_2 f(h) + w_3 f'(0) + w_4 f'(h) + R(h).$$

- a) Bepaal de gewichten  $w_i$  zó dat polynomen van zo hoog mogelijke graad nog exact geïntegreerd worden.
- b) Bepaal  $c$  en  $k$  in de formule voor  $R(h)$ , aannemende dat deze de gedaante

$$ch^{k+1} f^{(k)}(\xi) \quad (0 < \xi < h)$$

heeft.

- c) Welk voordeel biedt deze formule als we deze  $n$ -maal gerepeteerd toepassen om  $\int_a^b f(t) dx$  nauwkeurig te berekenen? Kunt u iets zeggen over het gedrag van de fout bij het gerepeteerd toepassen van de kwadratuurformule? Vergelijk deze gerepeteerde kwadratuurformule met de gerepeteerde trapeziumregel. Vergelijk ook de bijbehorende fouten.
- d) Toon door integratie van een geschikt interpolatiepolynoom aan dat de aanname in b. correct is als  $f$   $k$  maal continu differentieerbaar is.

**Vraagstuk 21** In deze opgave willen we via enkele eenvoudige methoden de integraal  $\int_0^1 e^x dx$  benaderen.

- a) Benader de integraal met de trapeziumregel en vervolgens met de midpuntregel. Bereken de fout t.o.v. het werkelijke antwoord. Wat valt u op bij vergelijking van het teken van de fout bij trapezium- resp. midpuntregel? Was dit te verwachten?
- b) Pas ook eens de  $2 \times$  gerepeteerde trapeziumregel en midpuntregel toe. Vergelijk de verschillende fouten uit onderdeel a. en b. met elkaar en ga na of deze aan het door de theorie voorspelde gedrag voldoen.
- c) Benader ook m.b.v. de Simpson regel. Vergelijk de fout met de antwoorden uit a.

$x$	$e^x$
0.25	1.28403
0.5	1.64872
0.75	2.11700
1	2.71828

**Vraagstuk 22**

- a) Toon aan dat voor  $f \in C^4[a, b]$  ( $a < b$ )

$$\int_a^b f(t) dt = (b-a)f(m) + \frac{(b-a)^3}{24} f''(m) + R_1, \quad R_1 = \frac{(b-a)^5}{80 \times 24} f^{(4)}(\xi)$$

waarin  $m = a + \frac{b-a}{2}$ ,  $\xi \in [a, b]$ .

- b) We vervangen nu  $f''(m)$  door een differentiequotient op  $f'$ . Ga na dat

$$\int_a^b f(t) dt = (b-a)f(m) + \frac{(b-a)^2}{24}(f'(b) - f'(a)) + R_2, \quad R_2 = \mathcal{O}((b-a)^5).$$

Vind een expliciete uitdrukking voor  $R_2$ . Voor welke graads polynomen is deze kwadratuurformule exact? Hoe luidt de  $n \times$  gerepeteerde versie van de hierdoor gedefinieerde kwadratuurformule?

- c) Zij  $q(t)$  het hoogstens tweedegraads interpolatie polynoom van  $f'(t)$  op de steunpunten  $a$ ,  $b$  en  $m$ , en zij  $p(t)$  gedefinieerd door  $p'(t) = q(t)$ ,  $p(m) = f(m)$ . Bewijs, zonder  $p$  en  $q$  expliciet te berekenen, dat

$$f(t) - p(t) = \frac{f^{(4)}(\eta)}{6} \int_m^t (x-a)(x-m)(x-b) dx,$$

waarbij  $\eta$  van  $t$  afhangt.

- d) Toon aan dat de in b. gedefinieerde kwadratuurformule toegepast op  $p(t)$  hetzelfde resultaat oplevert als die formule toegepast op  $f(t)$ . Bewijs hiermee dat

$$R_2 = \frac{-7}{24 \times 240} (b-a)^5 f^{(4)}(\theta).$$

- e) Vergelijk voor kleine stapgrootten de nauwkeurigheid van de  $2n$  gerepeteerde versie van de kwadratuurformule uit b. met die van de  $n$  maal gerepeteerde versie van de Simpson regel, als u voor beide ongeveer evenveel functie-evaluaties gebruikt.

### Vraagstuk 23

- a) Stel we willen  $\int_0^1 t^{7/2} dt$  benaderen met de gerepeteerde Simpson regel. Zie in, dat de restterm voor dit geval geen bruikbare informatie geeft over het gedrag van de fout. Men zij echter op zijn hoede: er volgt niet, dat er geen convergentie zou zijn! Toon convergentie aan voor een continue integrand, door de gerepeteerde Simpson regel te interpreteren als Riemann-som. (Geef een schets van de situatie.)
- b) Laat zien, dat men de regel van Simpson voor  $\int_a^b f(t) dt$  kan krijgen, door een geschikte lineaire combinatie te nemen van de midpunt- en trapeziumregel. Leid hiermede af voor de restterm  $R$  bij Simpson:

$$|R| \leq \frac{(b-a)^3}{18} \max |f''|, \quad |R| \leq \frac{(b-a)^4}{36} \max |f'''|.$$

Welk convergentie gedrag mag u dus minstens verwachten voor de integraal uit a.?

- c) U kunt de Simpson regel ook krijgen door een geschikte lineaire combinatie te nemen van de midpunt- en de 2-maal gerepeteerde trapeziumregel. Leiden deze combinaties tot een gunstiger foutschatting dan die in b.?
- d) Geef een schatting voor de restterm van de midpuntregel en de trapeziumregel voor het geval dat  $f \in C^1[a, b]$ . (Voor beide regels voldoet  $\frac{(b-a)^2}{4} \max |f'|$ .)

- e) Hoeveel steunpunten zijn nodig om  $\int_0^1 x^{\frac{5}{2}} dx$  met de gerepeteerde Simpson regel te benaderen met een fout van hoogstens  $10^{-4}$ ? Vergelijk uw resultaat met onderstaande tabel en geef een verklaring voor hetgeen u constateert.

Resultaten verkregen bij werkelijke berekening van  $\int_0^1 x^{\frac{5}{2}} dx$  met gerepeteerde Simpson regel ( $h = 2^{-n}$ ).

$n$	fout	verhouding
1	$1.19610^{-3}$	-
2	$1.21710^{-4}$	9.8
3	$1.18010^{-5}$	10.3
4	$1.10810^{-6}$	10.7
5	$1.02110^{-7}$	10.8
6	$0.92810^{-8}$	11.01
7	$0.83610^{-9}$	11.1

Verhouding is dus de factor waarmee de fout per verhoging van  $n$  afneemt.

### Vraagstuk 24

- a) Laat  $Q_n(f) = \sum_{i=0}^n w_i^{(n)} f(x_i)$  een ongerepeteerde kwadratuurformule zijn voor  $\int_a^b f(x) dx$ . Zij  $f^*(x) = f(x) + \epsilon(x)$ , met  $|\epsilon(x)| \leq \bar{\epsilon}$ . Toon aan, dat zowel voor de ongerepeteerde als de  $N \times$  gerepeteerde kwadratuurformule de extra fout t.g.v. de perturbatie  $\epsilon(x)$  in absolute waarde kleiner dan of gelijk is aan

$$\bar{\epsilon} \sum_{i=0}^n |w_i^{(n)}|.$$

- b) Toon aan, dat voor zowel de ongerepeteerde als voor de  $N \times$  gerepeteerde versies van de trapezium-, midpunt- en Simpson regels de in a. bedoelde extra fout hoogstens  $(b-a)\bar{\epsilon}$  bedraagt.
- c) Zie in, dat dit het beste is wat te verwachten was, gezien de verandering van de ware integraal als men de functie  $f$  'met  $\epsilon$  verstoort'.

**Vraagstuk 25** We bezien kwadratuurformules voor integralen van het type:

$$\int_0^h \sqrt{x} f(x) dx.$$

Zij  $f \in C^2[0, h]$ .

- a) Ga aan de hand van de restterm na waarom de trapeziumregel het in het algemeen slecht zal doen voor dit type integralen.
- b) Nieuwe formule: Zij  $\phi_1(x)$  het lineair interpolatiepolynoom van  $f$  met steunpunten 0 en  $h$ . Integreer nu:  $\int_0^h \sqrt{x} \phi_1(x) dx$ .
- c) Laat zien dat de zo verkregen kwadratuurformule een fout heeft van:

$$-\frac{2}{35} h^{7/2} f''(\xi_1), \quad \xi_1 \in (0, h).$$

- d) Nog een formule: Zij  $\phi_2^{(\alpha)}(x)$  de som van de 0<sup>e</sup> en 1<sup>e</sup> graadsterm van de Taylor reeks van  $f$  rond  $\alpha h$  ( $0 \leq \alpha \leq 1$ ). Bewijs dat de fout

$$R = \int_0^h \sqrt{x} f(x) dx - \int_0^h \sqrt{x} \phi_2^{(\alpha)}(x) dx$$

gelijk is aan:

$$R = f''(\xi) h^{\frac{7}{2}} \left[ \frac{1}{7} - \frac{2}{5} \alpha + \frac{1}{3} \alpha^2 \right] \quad \text{met } \xi \in [0, h].$$

(hiervoor behoeft  $\int_0^h \sqrt{x} \phi_2^{(\alpha)}(x) dx$  niet expliciet berekend te worden). De uitdrukking in  $\alpha$  is minimaal  $\frac{4}{175}$ , voor  $\alpha = \frac{3}{5}$ .

- e) Noem  $\phi_2 = \phi_2^{(\frac{3}{5})}$ . Integreer  $\int_0^h \sqrt{x} \phi_2(x) dx$ .

**Vraagstuk 26** Een Romberg schema kan men opstellen, uitgaande van een gegeven rij benaderingswaarden  $T(h)$ , behorende bij een rij van (dalende) stapgrootten  $h$ . Deze rij (van  $h$ -waarden) kan bijv. een halverings-rij zijn (iedere waarde van  $h$  is de helft van de voorgaande) of een zgn. Bulirsch-rij (zie 3.5E van het dictaat).

- Hoe zou u uit een gegeven rij van benaderingswaarden kunnen afleiden of deze behoort bij een Bulirsch rij dan wel een rij waarin met een constante factor gereduceerd wordt?
- Hieronder is een Romberg schema gegeven. Is dit schema waarschijnlijk gebaseerd op een Bulirsch rij of niet?
- Welke zijn de lineaire combinaties bij de eerste kolom geweest, waarmee de 2e kolom verkregen is?

+1.598722				
+1.346600	+1.094479			
+1.246763	+1.146926	+1.154418		
+1.203650	+1.160536	+1.162480	+1.163018	
+1.183280	+1.162910	+1.163249	+1.163300	+1.163309
+1.173269	+1.163258	+1.163308	+1.163312	+1.163312
+1.168287	+1.163305	+1.163312	+1.163312	+1.163312

**Vraagstuk 27** Vaak kan men een functie zeer efficiënt benaderen door de functie als integraal te representeren en deze integraal vervolgens met Gaußkwadratuur te benaderen. Het volgende geeft hiervan een voorbeeld.

- a) Laat zien:

$$\ln(1+x) = x/2 \int_{-1}^1 \frac{1}{1+x(1+t)/2} dt.$$

- b) Hoeveel punts Gauß(-Legendre) kwadratuur is nodig om deze integraal te benaderen met een relatieve fout van hooguit  $10^{-15}$  als  $x \in (0, 1]$ ?

- c) Men kan het interval, waarop de  $\ln$  benaderd moet worden, reduceren tot  $x \in (0, \sqrt{2} - 1]$ . Men berekent dan  $\ln(1+x)$  voor  $x > \sqrt{2} - 1$  met

$$\ln(1+x) = \ln\left(\frac{1+x}{\sqrt{2}}\right) + \frac{1}{2} \ln 2.$$

Ga na. Hoeveel punts Gauß kwadratuur is nu nog maar nodig om de  $\ln$  met eenzelfde relatieve nauwkeurigheid te benaderen voor  $x \in (0, \sqrt{2} - 1]$ ?

- d) Kunt u dit trucje wederom toepassen? Is dat voordelig voor de hoeveelheid rekenwerk?
- e) Bedenk nu zelf een methode om de arctangens efficiënt te benaderen.

## Index

- (absolute) relatieve fout, 20
- absolute fout, 20
- adaptieve methoden, 40
- afrondfout, 22
- automatisch integratieproces, 39
  
- bereik van een machine, 22
- Bernoulli getallen, 41
- binair, 22
- Bulirsch rij, 30
  
- decimaal, 22
- differentie quotiënt, 15
  
- equidistante, 6
- Euler-Maclaurin reeks, 41
- exponent, 22
- extrapoleren, 4
  
- fout, 20
  
- Gauss-Legendre formule, 48
- Gauss-Legendre polynomen, 48
- Gauß' voorwaartse formule, 10
- gerepeteerde kwadratuurformule, 37
- gerepeteerde trapeziumregel, 37
- gewichten, 42
- Gram-Schmidt orthogonalisatie proces, 48
- grondtal, 22
  
- Hermite interpolatie, 13
- hexadecimale arithmetiek, 22
  
- IEEE standaard, 22
- interpolatie polynoom, 12
- interpolatoire kwadratuurformule, 42
- inverse interpolatie, 13
  
- kleinste kwadraten methode, 19
- kwadratuurformules, 35
- kwadratuurschema, 49
  
- Lagrange coëfficiënten, 6
- Lagrange interpolatie, 5
- Lagrange interpolatiepolynoom, 5
- Lagrange representatie, 6
- lineair convergent, 31
- lineaire interpolatiepolynoom, 2
  
- machinegetallen, 22
- machineoperaties, 23
- majorant van de fout, 20
- mantisse, 22
- midpuntregel, 44
- modelfout, 21
  
- Newton's voorwaartse formule, 10
- Newton-Cotes formules, 46
- numeriek differentiëren, 15
- numerieke integratie, 35
  
- onbetrouwbaarheid, 20
- overflow, 22
  
- regel van Simpson, 45
- relatieve machine precisie, 23
- relatieve onbetrouwbaarheid, 20
- representatiefout, 21
- Richardson extrapolatie, 29
- Riemann-som, 38
- Romberg integratie, 42
- Romberg rij, 30
- Romberg schema, 30
  
- schatting voor de fout, 29
- stapgrootte, 15
- Stelling van Banach-Steinhaus, 50
- Stelling van Weierstrass, 50
- steunpunten, 2, 42
- superlineaire convergentie, 31
  
- Taylor ontwikkeling, 11
- Taylor polynoom, 11
- tekenvast, 36
- trapeziumregel, 35
- tussenwaardenstelling van de integraalrekening, 36
  
- underflow, 22
  
- variabele stap methoden, 40
- vertrouwensgetal, 40