

Numerieke Methoden voor Differentiaalvergelijkingen

Dr.ir. C. Vuik

wi2091

wi2092

2000



Technische Universiteit Delft
Faculteit Informatietechnologie en Systemen
Afdeling Toegepaste Wiskundige Analyse

Voorwoord

In dit college worden numerieke methoden besproken voor het oplossen van differentiaalvergelijkingen. Aan de orde komen aspecten, die in praktische problemen een grote rol spelen. In dit college beperken we ons tot 1 dimensie, uitgezonderd het hoofdstuk over de warmtevergelijking waarin een tijd en ruimte dimensie gebruikt worden. Een aantal technieken zoals: interpolatie, numerieke integratie en het oplossen van niet-lineaire vergelijkingen, kunnen ook buiten de context van differentiaalvergelijkingen gebruikt worden. Het college bestaat uit twee modules: beginwaarde en randwaarde problemen.

Achtergrondinformatie is te vinden in:

Numerical Analysis (6^e editie)

R.L. Burden, J.D. Faires

Brodes/Cole publishing company, Pacific Grove, 1997

Applied Numerical Analysis (6^e editie)

C.F. Gerald, P.O. Wheatly

Addison Wesley, Reading, 1999

Afternotes on Numerical Analysis

G.W. Stewart

SIAM, Philadelphia, 1996

Moduul I

Beginwaarde problemen

Moduul II

Randwaarde problemen

Inhoudsopgave

1	Inleiding	1
1.1	Enkele historische opmerkingen	1
1.2	Wat is numerieke wiskunde?	1
1.3	Waarom numerieke wiskunde?	2
1.4	Afrondfouten	3
1.5	Het O-symbool van Landau	5
1.6	Samenvatting	6
2	Interpolatie	7
2.1	Inleiding	7
2.2	Lineaire interpolatie	7
2.3	Lagrange interpolatie	9
2.4	Interpolatie met functiewaarden en afgeleiden	12
2.4.1	Taylorpolynoom	12
2.4.2	Interpolatie in het algemeen	13
2.4.3	Hermite interpolatie	14
2.5	Interpolatie met splines	16
2.6	Samenvatting	18
3	Numeriek differentiëren	20
3.1	Inleiding	20
3.2	Eenvoudige differentieformules voor de eerste afgeleide	21
3.3	Algemene formules voor de eerste afgeleide	25
3.4	Relatie tussen differentieformules en interpolatie	26
3.5	Differentieformules voor hogere afgeleiden	27
3.6	Richardson's extrapolatie	28
3.6.1	Inleiding	28
3.6.2	Schatting van de fout in de praktijk	28
3.6.3	Nauwkeuriger formules via Richardson's extrapolatie	30
3.7	Samenvatting	31
4	Numerieke tijdsintegratie voor beginwaarde problemen	32
4.1	Inleiding	32
4.2	Eenstaps methoden	33
4.3	Testvergelijking en versterkingsfactor	35
4.4	Afbreekfout en consistentie	35

4.5	Stabiliteit	37
4.6	Convergentie	40
4.7	Stelsels eerste orde beginwaarde problemen	42
4.8	Stabiliteit voor stelsels	44
4.9	Stijve stelsels differentiaalvergelijkingen	47
4.10	Runge-Kutta methoden	49
4.11	Meerstaps methoden	50
4.12	Samenvatting	53
5	De eindige differentie methode voor randwaarde problemen	54
5.1	Inleiding	54
5.2	De eindige differentie methode	55
5.3	Herhaling van enkele lineaire algebra begrippen	56
5.4	Consistentie, stabiliteit en convergentie	57
5.5	De conditie van de discretisatie matrix	60
5.6	Neumann randvoorwaarde	61
5.7	Het algemene probleem	62
5.8	De convectie-diffusie vergelijking	63
5.9	Samenvatting	64
6	De instationaire warmtevergelijking	66
6.1	Inleiding	66
6.2	Afleiding van de instationaire warmtevergelijking	66
6.3	De gediscretiseerde vergelijking	67
6.4	Samenvatting	70
7	Numerieke integratie	71
7.1	Introductie	71
7.2	Eenvoudige numerieke integratieformules	71
7.3	Newton Cotes formules	76
8	Niet-lineaire vergelijkingen	84
8.1	Introductie	84
8.2	Een eenvoudige nulpuntsmethode	84
8.3	Vaste punt iteratie	86
8.4	De Newton-Raphson methode	87
8.5	Stelsels niet-lineaire vergelijkingen	90
8.6	Toepassingen op niet-lineaire differentiaalvergelijkingen	91
	8.6.1 Niet-lineaire beginwaarde problemen	91
	8.6.2 Niet-lineaire randwaarde problemen	91
8.7	Samenvatting	92

Hoofdstuk 1

Inleiding

1.1 Enkele historische opmerkingen ¹

De moderne toegepaste wiskunde begint in de 17e en 18e eeuw met mensen als Stevin, Descartes, Newton en Euler. Numerieke aspecten maakten op natuurlijke wijze deel uit van de analyse; de term numerieke wiskunde was onbekend. Numerieke methoden bedacht door Newton, Euler en later Gauss spelen ook vandaag nog een belangrijke rol.

In de 18e en 19e eeuw worden voor allerlei deelgebieden van de natuurkunde, zoals de mechanica en de stromingsleer, fundamentele wetten geformuleerd in de vorm van eenvoudig ogende wiskundige vergelijkingen. Deze bleken echter, tot veler teleurstelling, alleen in heel speciale gevallen analytisch te kunnen worden opgelost. De opkomende techniek ontwikkelt zich dan ook goeddeels los van de wiskunde. De komst van de computer heeft hierin verandering gebracht. Met gebruik van de computer is het mogelijk met gedetailleerde en realistische wiskundige modellen en numerieke methoden nauwkeurige kwantitatieve informatie te verwerven, betreffende een veelheid van verschijnselen en processen in de natuur en in de techniek. De toepassing van computers en numerieke methoden is algemeen geworden. Steekproeven wijzen uit dat in ca. 70% van de artikelen in de vaktijdschriften van de ingenieurswetenschappen wiskundige modellen en methoden worden gebruikt, die niet triviaal zijn.

Berekeningen zijn meestal goedkoper dan experimenten; verder kunnen experimenten onmogelijk of gevaarlijk zijn. Vaak kunnen experimenten alleen op verkleinde schaal worden uitgevoerd, wat de resultaten minder betrouwbaar maakt.

1.2 Wat is numerieke wiskunde?

Numerieke wiskunde is de leer der methoden voor het getalmatig benaderen van oplossingen van wiskundige vergelijkingen met eindige rekenprocessen. Of, nog korter: de leer der constructieve methoden in de wiskunde.

In grote delen van de wiskunde zijn de belangrijkste concepten die van afbeelding en verzameling. Voor de numerieke wiskunde moet hieraan worden toegevoegd het concept van bereken-

¹voor meer informatie zie <http://ta.twi.tudelft.nl/nw/users/vuik/wi211/wi211.html>

baarheid. Berekenbaarheid houdt in dat het resultaat verkregen kan worden met een eindig aantal operaties (zodat de rekentijd eindig is) op een eindige deelverzameling van de rationale getallen (omdat het geheugen van de computer eindig is). In het algemeen zal het resultaat een benadering zijn, omdat de meeste wiskundige vergelijkingen operatoren bevatten, die op een oneindig voortlopend proces gebaseerd zijn, zoals integralen en afgeleiden, en omdat de oplossingen functies zijn, waarvan het domein en bereik ook irrationele getallen bevatten.

Omdat in het algemeen slechts benaderingen van oplossingen verkregen kunnen worden, kunnen in de numerieke wiskunde slechts die problemen zinvol behandeld worden, die bestand zijn tegen kleine verstoringen, oftewel stabiel zijn. De vraag naar de stabiliteit heeft de numerieke wiskunde gemeen met de klassieke wiskunde. Een belangrijk hulpmiddel bij de studie van stabiliteit is de *functionaalanalyse*. Deze discipline speelt ook een belangrijke rol bij de studie van de fout: het verschil tussen de numerieke benadering en de exacte oplossing.

De consequenties van het rekenen met een eindige deelverzameling van de rationale getallen, zijn velerlei. Zo kan een computer bijvoorbeeld geen onderscheid maken tussen polynomen van voldoende hoge graad. Zodoende kan men niet zonder meer vertrouwen op methoden die gebaseerd zijn op de hoofdstelling van de algebra (namelijk, dat een n^e graads polynoom precies n nulpunten heeft). De fouten die het gevolg zijn van het werken met een eindig aantal cijfers heten *afrondfouten*. Aan afrondfouten zal in het vervolg van dit hoofdstuk nog enige aandacht geschonken worden.

Een belangrijk aspect van de numerieke wiskunde is de nadruk op *efficiency*. Een vergroting van efficiency, dat wil zeggen vermindering van het aantal benodigde operaties en/of het benodigde geheugen, wordt elders in de wiskunde niet als een essentiële vooruitgang gezien, in de numerieke wiskunde echter wel. Deze vooruitgang is van groot praktisch belang. Het einde hiervan is nog lang niet in zicht. Hier liggen nog vele mogelijkheden voor creatieve geesten. Verder zal nog veel overhoop gehaald worden door veranderingen in computerarchitectuur.

1.3 Waarom numerieke wiskunde?

Een groot voordeel van numerieke wiskunde is dat een numeriek antwoord verkregen kan worden voor problemen, die geen "analytische" oplossing hebben. Neem bijvoorbeeld de integraal

$$\int_0^{\pi} \sqrt{1 + \cos^2(x)} dx ,$$

die de lengte geeft van één boog van de kromme gegeven door $y = \sin x$. Er bestaat voor deze integraal geen oplossing in gesloten vorm. Echter met een numerieke methode kan de integraal eenvoudig bepaald worden. Een bijkomend voordeel is dat een numerieke methode alleen gebruik maakt van functie-evaluaties en de operaties: optellen, afrekken, vermenigvuldigen en delen. Omdat deze operaties precies de functies zijn die een computer kan uitvoeren, vormen numerieke wiskunde en computers een perfecte combinatie.

Een voordeel van een analytische methode is dat deze een oplossing geeft in termen van

wiskundige functies. Hieruit kan men inzicht verkrijgen in het gedrag en de eigenschappen van de oplossing. Bij een numerieke oplossing is dit niet het geval. Aan de andere kant, er wordt vaak gebruik gemaakt van een visualizatie om inzicht in het gedrag van de oplossing te verkrijgen. Het maken van een grafiek met een numerieke methode is meestal efficiënter dan het evalueren van de analytische oplossing in een groot aantal punten.

1.4 Afrondfouten

Een computer werkt niet met oneindige precisie. Reële getallen worden in computers opgeslagen als

$$\pm 0.d_1d_2\dots d_n \cdot \beta^e$$

waarbij $d_1 > 0$ en $0 \leq d_i < \beta$. We noemen dit een floating point getal waarbij: de mantisse is $0.d_1d_2\dots d_n$, de basis is β en e (geheel getal) is de exponent. Vaak is $\beta = 2$, $n = 24$ (*enkele precisie*), zoals voor vele PC's. In dubbele precisie geldt $n = 56$. We zeggen dan dat de machine rekent met n cijfers.

Laat voor $x \in \mathbb{R}$ gelden

$$0.d_1\dots d_n \cdot \beta^e \leq x < 0.d_1d_2\dots(d_n + 1) \cdot \beta^e$$

waarbij we x gemakshalve positief veronderstellen. Afronding houdt in dat x vervangen wordt door het dichtstbijzijnde floating point getal, dat we $fl(x)$ zullen noemen. De hierdoor veroorzaakte fout wordt *afrondfout* genoemd. Laten we schrijven

$$fl(x) = x(1 + \epsilon). \quad (1.1)$$

We noemen ϵx de *absolute fout* en ϵ de *relatieve fout*. Het verschil tussen de getallen waar x tussen ligt is β^{e-n} , dus vanwege de wijze van afronding geldt $|fl(x) - x| \leq \frac{1}{2}\beta^{e-n}$, zodat voor de absolute fout geldt:

$$|\epsilon x| \leq \frac{1}{2}\beta^{e-n}.$$

Aangezien $|x| \geq \beta^{e-1}$ (omdat $d_1 > 0$) geldt voor de relatieve fout:

$$|\epsilon| \leq eps \quad (1.2)$$

met de relatieve precisie van de computer eps gedefinieerd door

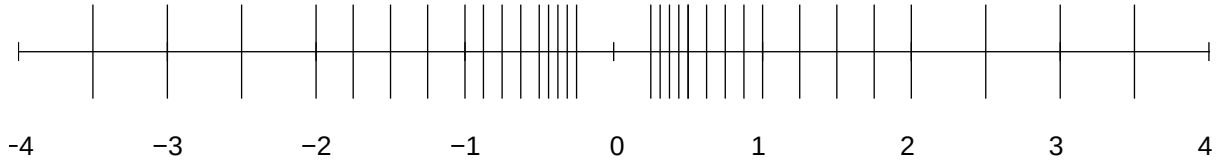
$$eps = \frac{1}{2}\beta^{1-n}. \quad (1.3)$$

Met $\beta = 2$ en $n = 24$ geldt dus $eps \cong 6 \cdot 10^{-8}$: er wordt met ongeveer 7 decimale cijfers gerekend.

Figuur 1.1 toont de ligging van de floating point getallen $0.1d_2d_3 \cdot \beta^e$; $e = -1, 0, 1, 2$ in het binaire stelsel ($\beta = 2$). Deze floating point getallen liggen ongelijk verdeeld en er is een gat

bij 0. Als een rekenresultaat in dit gat terecht komt noemen we dit *underflow*. De meeste machines geven een waarschuwing, vervangen het resultaat door 0 en gaan door. Als de absolute waarde van een rekenresultaat te groot is, is er sprake van *overflow*. De machine waarschuwt en stopt.

Hoe voeren computers rekenkundige operaties uit met floating point getallen? Rekenproces-



Figuur 1.1: De ligging van $\pm 0.1d_2d_3 \cdot \beta^e$, $\beta = 2$, $e = -1, 0, 1, 2$.

soren zijn ingewikkeld. Meestal wordt het volgende model gebruikt ter benadering van de werkelijkheid. Laat $\circ$ een rekenkundige operatie zijn ($+$, $-$, $*$ of $/$), en laten x en y twee floating point getallen zijn. Dan is de machine-uitkomst van de operatie:

$$z = fl(x \circ y) \quad (1.4)$$

met $x \circ y$ de exacte uitkomst. Deze is in het algemeen geen floating point getal, zodat er een fout gemaakt wordt. Volgens (1.1) geldt

$$z = \{x \circ y\}(1 + \epsilon). \quad (1.5)$$

voor één of andere ϵ die aan (1.2) voldoet.

Stel dat x en y geen floating point getallen zijn, maar benaderd worden met $fl(x)$ en $fl(y)$ dus $fl(x) = x(1 + \epsilon_1)$, $fl(y) = y(1 + \epsilon_2)$, dan geeft (1.5):

$$z = \{x(1 + \epsilon_1) \circ y(1 + \epsilon_2)\}(1 + \epsilon). \quad (1.6)$$

Stel $x > 0$, $y > 0$, en $\circ = -$ (aftrekking). Dan geldt voor $|\epsilon|, |\epsilon_1|, |\epsilon_2| \ll 1$:

$$\begin{aligned} z &\cong x - y + \epsilon(x - y) + x\epsilon_1 - y\epsilon_2 = \\ &= (x - y) \left(1 + \epsilon + \frac{x\epsilon_1 - y\epsilon_2}{x - y} \right) = (x - y)(1 + \epsilon_3). \end{aligned} \quad (1.7)$$

Als $|x - y|$ klein is, kan de relatieve fout in z groot worden:

$$\begin{aligned} |\epsilon_3| &\leq |\epsilon| + |x\epsilon_1 - y\epsilon_2|/|x - y| \\ &\leq |\epsilon| + (|x\epsilon_1| + |y\epsilon_2|)/|x - y| \\ &\leq eps\{1 + (|x| + |y|)/|x - y|\}. \end{aligned} \quad (1.8)$$

Dit is slechts een bovengrens; wat ϵ_3 wordt hangt van het toeval af. Maar het gelijktken kan aangenomen worden (namelijk als $\epsilon = -\epsilon_1 = \epsilon_2 = eps$). Als $|\epsilon_3| \gg eps$ (relatieve fout in resultaat veel groter dan de relatieve fout in de operanden) spreken we van *cijferverlies*.

Voorbeeld (cijferverlies)

$$\begin{array}{r} \text{Beschouw de som} \quad 472635.0000 \\ + \quad 27.5013 \\ - \quad 472630.0000 \\ \hline 32.5013 \end{array}$$

Als we met 6 cijfers rekenen dan krijgen we

$$\begin{array}{r} 472635.0000 \\ + \quad 27.5013 \\ \hline 472663.0000 \end{array}$$

en daarna

$$\begin{array}{r} 472663 \\ - \quad 472630 \\ \hline 33 \end{array}$$

Merk op dat er slechts 2 cijfers goed zijn van het berekende antwoord, terwijl we toch met 6 cijfers gerekend hebben. De fout wordt niet veroorzaakt door de aftrekking maar door het feit dat bij de optelling de laatste 4 cijfers van 27.5013 weggegooid zijn. Verder blijkt dat de berekening wel een goed resultaat geeft als de volgorde veranderd wordt. Als we eerst de grote getallen verwerken, dan krijgen we

$$\begin{array}{r} 472635 \\ - \quad 472630 \\ \hline 5 \end{array}$$

en daarna

$$\begin{array}{r} 5.0000 \\ + \quad 27.5013 \\ \hline 32.5013 \end{array}$$

hetwelk gelijk is aan het exacte antwoord.

1.5 Het O-symbool van Landau

Bij het analyseren van numerieke methoden is het onderzoeken van de fout belangrijk. Hierbij is een indicatie van de orde van grootte belangrijker dan een nauwkeurige uitdrukking voor de fout. Om schrijfwerk te besparen maken we gebruik van het (grote O) O-symbool van Landau.

Definitie

Gegeven de functies f en g dan geldt $f(x) = O(g(x))$ als $x \rightarrow 0$ als er een positieve r en M bestaat zodanig dat

$$|f(x)| \leq M|g(x)| \quad \text{voor alle } x \in [-r, r].$$

Bij het schatten van fouten maken we vaak gebruik van de volgende rekenregels:

Rekenregels

Als $f(x) = O(x^p)$ en $g(x) = O(x^q)$ als $x \rightarrow 0$, met $p \geq 0$ en $q \geq 0$ dan is

- a) $f(x) = O(x^s)$ voor alle s met $0 \leq s \leq p$.
- b) $\alpha f(x) + \beta g(x) = O(x^{\min\{p,q\}})$ voor alle $\alpha, \beta \in \mathbb{R}$.
- c) $f(x)g(x) = O(x^{p+q})$.
- d) $\frac{f(x)}{|x|^s} = O(x^{p-s})$ als $0 \leq s \leq p$.

1.6 Samenvatting

In dit hoofdstuk zijn de volgende begrippen behandeld:

- numerieke wiskunde
- afrondfouten
- O -symbool van Landau

Hoofdstuk 2

Interpolatie

2.1 Inleiding

De vraagstelling: het bepalen van tussenliggende waarden uit een beperkt aantal metingen, of het voorspellen van waarden in punten die buiten het meetbereik liggen (extrapolatie) komt in praktijk veel voor. Laten we als voorbeeld het aantal kippen beschouwen in de Nederlandse pluimveehouderijen. In Tabel 2.1 staan de aantallen in miljoenen tussen 1970 en 1995.

jaar	1970	1975	1980	1985	1990	1995
aantal	53	68	82	92	94	92

Tabel 2.1: Het aantal kippen (in miljoenen) in Nederland (bron: NRC 09-12-1998)

We kunnen ons afvragen hoe deze getallen gebruikt kunnen worden om het aantal kippen te schatten in de tussenliggende jaren bijvoorbeeld in 1992, of te voorspellen in het jaar 2000. Een aantal interpolatie/extrapolatie methoden om dit te doen worden in dit hoofdstuk besproken.

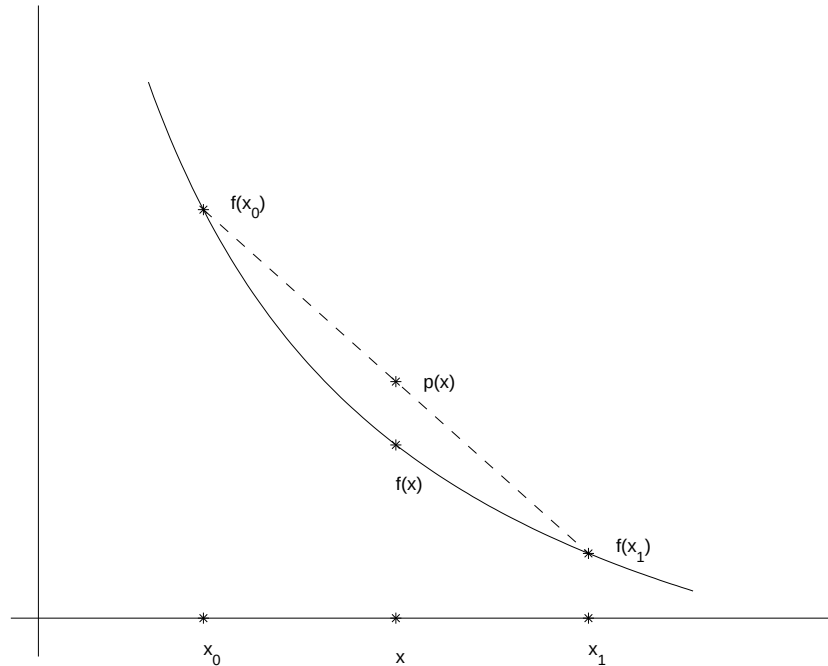
Ook bij visualizeren van beelden op een computerscherm kan veel geheugenruimte bespaard worden als niet alle punten opgeslagen worden, maar alleen een beperkt aantal punten. Door deze punten is dan een kromme gedefinieerd waarmee een min of meer realistisch beeld op het scherm weergegeven kan worden.

Als laatste toepassing het bepalen van goniometrische functies op een rekenmachine. Het berekenen van een functiewaarde kost veel rekentijd. Een oplossing hiervoor: sla een aantal functiewaarden op in het geheugen en bepaal op een goedkope manier de waarden in tussenliggende punten uit de bekende functiewaarden.

2.2 Lineaire interpolatie

De eenvoudigste manier van extrapoleren is de nulde graads extrapolatie. Stel de waarde in een bepaald punt is bekend. Dan kiezen we de waarden in de omgeving gelijk aan deze waarde. Een bekend voorbeeld hiervan is de voorspelling dat het weer van morgen hetzelfde is als vandaag. Deze voorspelling blijkt in 80% van de gevallen goed te zijn.

Een betere vorm van interpolatie is de rechte lijn tussen 2 punten (zie Figuur 2.1). Stel we



Figuur 2.1: Lineaire interpolatie.

weten de meetwaarden in de punten x_0 en x_1 : $f(x_0)$ en $f(x_1)$. Als we verder geen informatie hebben ligt het voor de hand in het punt x^* de waarde van de rechte lijn door de punten $(x_0, f(x_0))$, $(x_1, f(x_1))$ te nemen. Het is eenvoudig in te zien dat de rechte lijn gegeven wordt door:

$$p(x) = f(x_0) + \frac{x - x_0}{x_1 - x_0}(f(x_1) - f(x_0))$$

of

$$p(x) = \frac{x - x_1}{x_0 - x_1}f(x_0) + \frac{x - x_0}{x_1 - x_0}f(x_1).$$

De functie p is een lineair interpolatiepolynoom die gelijk is aan $f(x_0)$ in x_0 en $f(x_1)$ in x_1 . Een voor de hand liggende vraag is uiteraard: hoe groot is de fout bij lineaire interpolatie

Voorbeeld (lineaire interpolatie)

Stel de waarde van de sinusfunctie is bekend in 36° en 38° . De lineaire interpolatiebenadering in 37° is dan 0.601723. Het verschil met de exacte waarde is slechts 0.9×10^{-4} .

α	$\sin \alpha$
36°	0.58778525
37°	0.60181502
38°	0.61566148

Tabel 2.2: De waarde van $\sin \alpha$

en waar hangt die van af? Indien we weten dat de functie die we interpoleren tenminste 2 maal continu differentieer is dan kunnen we een uitspraak over de fout maken. Merk op met $[a, b, x_0]$ bedoelen we het gesloten interval opgespannen door de uitersten van a, b en x_0 .

Stelling 2.2.1 Laat x_0 en x_1 punten zijn in $[a, b]$, $x_0 \neq x_1$ en $f \in C[a, b] \cap C^2(a, b)$. Dan geldt voor het lineaire interpolatie polynoom p van f op de steunpunten x_0, x_1 en voor elke $x \in [a, b]$: er is een $\xi \in (x_0, x_1, x)$ zodat

$$f(x) - p(x) = \frac{1}{2}(x - x_0)(x - x_1)f''(\xi) . \quad (2.1)$$

Bewijs:

Als $x = x_0$ of $x = x_1$, dan is $f(x) - p(x) = 0$ en kan men ξ willekeurig kiezen. Neem aan $x \neq x_0$ en $x \neq x_1$. Voor elke waarde van x is er een getal q zodat

$$f(x) - p(x) = q(x - x_0)(x - x_1) .$$

Beschouw voor dit getal q de volgende functie

$$\varphi(t) = f(t) - p(t) - q(t - x_0)(t - x_1) .$$

Hiervoor geldt $\varphi(x_0) = \varphi(x_1) = \varphi(x) = 0$. Volgens de stelling van Rolle zijn er minstens twee verschillende punten y en z in (x_0, x_1, x) zodat $\varphi'(y) = \varphi'(z) = 0$. Opnieuw volgens Rolle is er ook een $\xi \in (y, z)$ en dus $\xi \in (x_0, x_1, x)$ zo dat $\varphi''(\xi) = 0$. Omdat $\varphi''(t) = f''(t) - 2q$ betekent dit dat $q = \frac{1}{2}f''(\xi)$. $\square$

Als $x \notin [x_0, x_1]$ dan gebruiken we het interpolatiepolynoom om te extrapoleren. De fout wordt dan nog steeds beschreven door (2.1).

Bij interpolatie volgt de bovengrens

$$|f(x) - p(x)| \leq \frac{1}{8}(x_1 - x_0)^2 \max_{\xi \in [x_0, x_1]} |f''(\xi)| .$$

In praktische toepassingen zijn de waarden $f(x_0)$ en $f(x_1)$ meestal door meting of berekening verkregen. Deze waarden bevatten dus een fout. Neem aan dat de absolute fout hoogstens ε is. Het verschil tussen het exacte polynoom p en verstoorde polynoom $\hat{p}$ wordt begrensd door:

$$|p(x) - \hat{p}(x)| \leq \frac{|x_1 - x| + |x - x_0|}{x_1 - x_0} \varepsilon .$$

Bij interpolatie is deze fout altijd begrensd door ε . Bij extrapolatie kan de fout groter worden dan ε . Stel $x \geq x_1$ dan wordt de additionele onnauwkeurigheid begrensd door:

$$|p(x) - \hat{p}(x)| \leq (1 + 2 \frac{x - x_1}{x_1 - x_0}) \varepsilon .$$

De totale fout is de som van de interpolatie/extrapolatie fout en de meetfout.

2.3 Lagrange interpolatie

Indien men meer gegevens dan 2 punten heeft, is het verstandig om ook van die extra punten gebruik te maken. Een voor de hand liggende methode is een hogere graads interpolatie te gebruiken. Omdat een polynoom van de graad n , $n + 1$ onafhankelijke parameters bevat zijn dus $n + 1$ gegevens nodig.

Als generalisatie van lineaire interpolatie kan men de approximatie beschouwen van een functie f door een hoogstens n -de graadspolynoom L_n , zodanig dat de waarden van f en L_n in een gegeven $n+1$ -tal verschillende punten $x_0, \dots, x_n$ aan elkaar gelijk zijn. Dit noemen we n -orde Lagrange interpolatie.

Een polynoom L_n dat aan deze eisen voldoet is gemakkelijk aan te geven in een vorm, die een duidelijke generalisatie van (2.1) is:

$$L_n(x) = \sum_{k=0}^n f(x_k) L_{kn}(x) ,$$

waarbij

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

Dat dit het goede polynoom is kunen we als volgt zien:

- elke L_{kn} is een polynoom van de graad n ,
- omdat $L_{kn}(x_j) = \delta_{kj}$ geldt $L_n(x_k) = f(x_k), k \in \{0, \dots, n\}$.

Het polynoom L_n wordt het Lagrange interpolatiepolynoom genoemd. De polynomen L_{kn} heten de Lagrange coëfficiënten. Zij kunnen ook geschreven worden als:

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

Een generalisatie van Stelling 2.2.1 is nu:

Stelling 2.3.1 *Laat $x_0, \dots, x_n$ verschillende punten zijn in $[a, b]$. Laat $f \in C^n[a, b]$ en $f^{(n+1)}$ bestaat op (a, b) . Dan geldt voor het bijbehorende polynoom L_n en voor elke $x \in [a, b]$: er is een $\xi \in (x_0, x_1, \dots, x_n, x)$ zodat*

$$f(x) - L_n(x) = (x-x_0)\dots(x-x_n) \frac{f^{(n+1)}(\xi)}{n+1!} .$$

Bewijs:

Het bewijs is geheel analoog aan dat van Stelling 2.2.1. □

Als we Lagrange interpolatie gebruiken op tabelwaarden, dan krijgt men het beste resultaat als de steunpunten zo gekozen worden dat x in het (of een) middelste interval ligt. Leg zelf uit waarom.

Naast de interpolatiefout willen we weten welke extra fouten er kunnen optreden bij hogere graads interpolatie als de functie (of tabel) waarden niet exact zijn. Neem aan dat de absolute fout in de waarden hoogstens ε is. De fout in het verstoorde interpolatiepolynoom is dan hoogstens

$$\sum_{k=0}^n |L_{kn}(x)| \varepsilon .$$

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

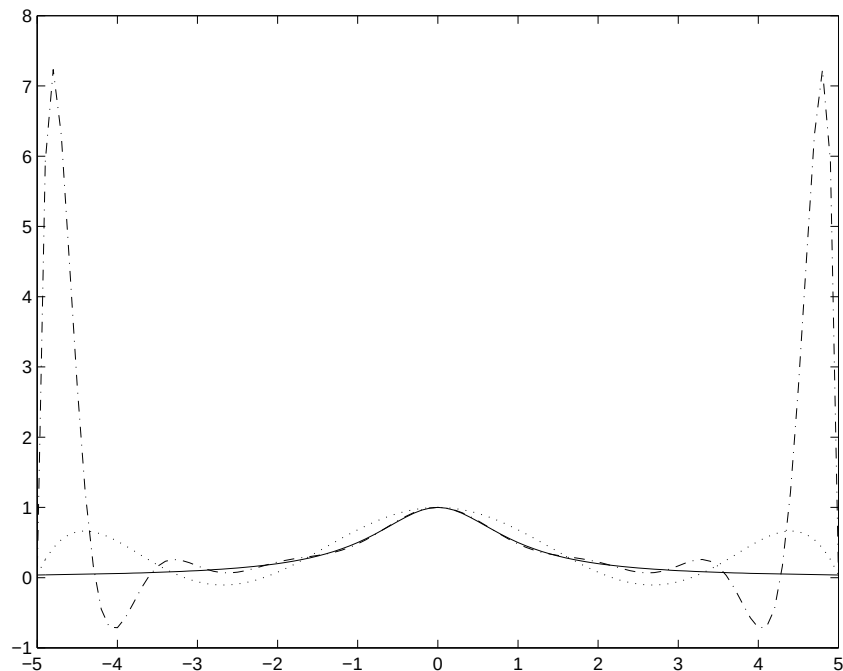
Tabel 2.3: Bovengrenzen voor $\sum_{k=0}^n |L_{kn}(x)|$.

Als de steunpunten equidistant zijn, $x_k = x_0 + kh$, dan neemt de waarde van $\sum_{k=0}^n |L_{kn}(x)|$ langzaam toe als n groter wordt. In Tabel 2.3 worden een aantal bovengrenzen gegeven.

In het algemeen verwacht je dat de benaderingsfout afneemt als de graad van het polynoom toeneemt. Dit is echter niet altijd het geval zoals het volgende voorbeeld laat zien.

Voorbeeld (interpolatie)

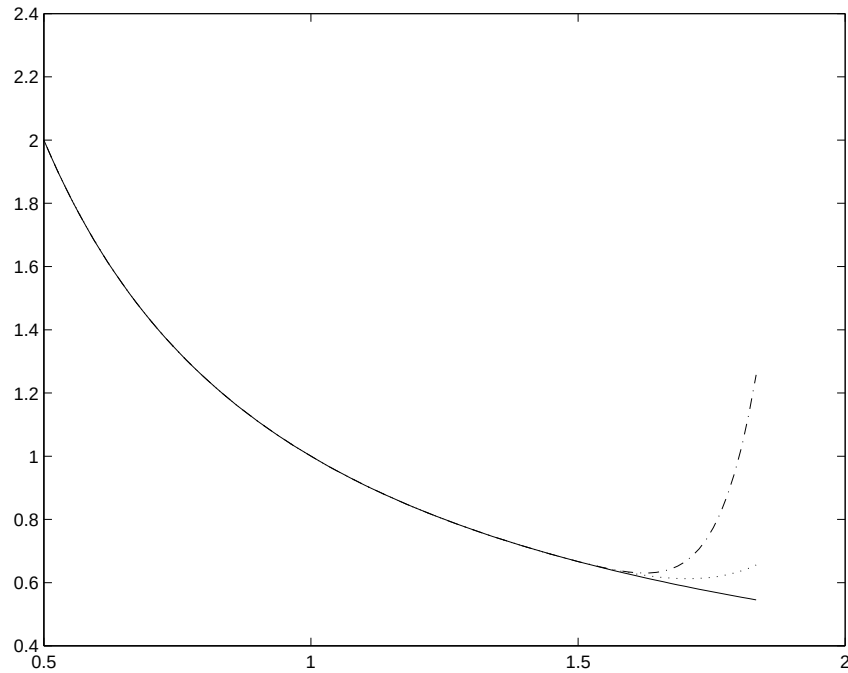
Beschouw de functie $\frac{1}{1+x^2}$ op het interval $[-5, 5]$. Bij de interpolatie gebruiken we de punten $x_k = -5 + \frac{10k}{n}$, $k = 0, \dots, n$. In Figuur 2.2 is de functie, het 6^e en het 14^e graads interpolatiepolynoom afgebeeld. Merk op dat op het interval $[-4, 4]$ de benadering van het 14^e graadspolynoom beter is dan die van het 6^e graadspolynoom. Echter in de buurt van de eindpunten vertoont het 14^e graadspolynoom grote afwijkingen.



Figuur 2.2: Interpolatie van $\frac{1}{1+x^2}$ met een 6^e graads Lagrange polynoom ($\cdots$) en een 14^e graad Lagrange polynoom ($-\cdot-\cdot-$).

Voorbeeld (extrapolatie)

Eenzelfde verschijnsel kan optreden bij extrapolatie. Beschouw de functie $\frac{1}{x}$. Het n -de graads interpolatiepolynoom wordt bepaald op de steunpunten $x_k = 0.5 + \frac{k}{n}$, $k = 0, \dots, n$. Als we de functie, het 6^e en het 10^e graads polynoom tekenen op het interval $[0.5, 1.8]$ (Figuur 2.3), dan zien we dat de polynomen niet te onderscheiden zijn van de functie op het interval $[0.5, 1.5]$. Echter bij extrapolatie ($x \geq 1.5$) treden grote fouten op. Hierbij is opnieuw de fout het grootst bij het 10-de graadspolynoom.



Figuur 2.3: Extrapolatie van $\frac{1}{x}$ met een 6^e graads Lagrange polynoom ($\cdots$) en een 10^e graads Lagrange polynoom ($-\cdot-\cdot-$).

2.4 Interpolatie met functiewaarden en afgeleiden

2.4.1 Taylorpolynoom

Een bekende methode om een functie te benaderen met een polynoom is het Taylorpolynoom. Als benaderingsmethode wordt deze techniek in de numerieke wiskunde niet zo vaak gebruikt. Het Taylorpolynoom wordt wel veel gebruikt in de analyse van numerieke processen. We vatten enkele resultaten samen.

Stelling 2.4.1 *Laat c een punt in het interval $[a, b]$ zijn en $f \in C^{n+1}[a, b]$. Laat het polynoom p_n gegeven zijn door:*

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

dan geldt voor de restterm $R_n(x) = f(x) - p_n(x)$: er bestaat een $\xi \in (c, x)$ zodat

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

Bewijs:

Voor het bewijs verwijzen we naar de eerstejaars analyse. □

In veel gevallen wordt de benadering door het Taylorpolynoom beter als de graad van het polynoom toeneemt. Maar dit is niet altijd het geval. Dit laten we zien in het volgende voorbeeld:

Voorbeeld (Taylorpolynoom)

De functie $f(x) = \frac{1}{x}$ willen we in $x = 3$ benaderen met een Taylorpolynoom van de graad n met het steunpunt $x = 1$. Voor de afgeleiden geldt: $f^{(k)}(x) = (-1)^k k! x^{-(k+1)}$. Het n^e graads Taylorpolynoom wordt dan gegeven door:

$$p_n(x) = \sum_{k=0}^n f^{(k)}(1)(x-1)^k/k! = \sum_{k=0}^n (-1)^k (x-1)^k .$$

De waarden van $p_n(3)$ als benadering van $f(3) = \frac{1}{3}$ staan in Tabel 2.4. Hieruit blijkt dat de

n	0	1	2	3	4	5	6	7
$p_n(3)$	1	-1	3	-5	11	-21	43	-85

Tabel 2.4: De waarde van het Taylorpolynoom in het punt $x = 3$ voor toenemende graad (n).
benadering onnauwkeuriger wordt als n toeneemt!

2.4.2 Interpolatie in het algemeen

In het algemene geval zoeken we bij een functie f een polynoom, zodanig dat in een aantal verschillende knooppunten $x_0, \dots, x_n$ niet alleen de waarden van f en p overeenstemmen, maar ook de waarden van hun afgeleiden tot en met de orde m_i .

In wiskundige termen: bij elke x_i is een niet negatief getal m_i gegeven. Veronderstel dat $f \in C^m[a, b]$ met $m = \max_{0 \leq i \leq n} m_i$. Dan is p het polynoom van de laagste graad zodanig dat

$$\frac{d^k p}{dx^k}(x_i) = \frac{d^k f}{dx^k}(x_i) \quad \text{voor elke } i = 0, 1, \dots, n \quad \text{en} \quad k = 0, 1, \dots, m_i .$$

Opmerkingen

1. De graad van dit polynoom is hoogstens $M = \sum_{i=0}^n m_i + n$.
2. Als $n = 0$ dan is p het Taylorpolynoom in x_0 van de graad m_0 .
3. Als $m_i = 0$ dan is p het n^e graads Lagrange polynoom van f op de punten $x_0, x_1, \dots, x_n$.

Als voorbeeld van algemene interpolatie beschouwen we in de volgende paragraaf de keuze $m_i = 1$. De resulterende polynomen worden Hermite polynomen genoemd.

2.4.3 Hermite interpolatie

Indien we van een functie niet alleen de functiewaarden in bepaalde punten kennen maar ook de afgeleiden, dan kunnen we die gegevens gebruiken om een hogere graads benaderingspolynoom te kiezen.

Stel bijvoorbeeld van een functie zijn in 2 punten zowel de functiewaarden als de afgeleiden bekend. Dan hebben we dus in feite 4 gegevens en kunnen een derde graads polynoom construeren met behulp van deze gegevens.

Dus gegeven:

$$\begin{array}{ll} (x_0, f(x_0)) & (x_1, f(x_1)) , \\ (x_0, f'(x_0)) & (x_1, f'(x_1)) . \end{array}$$

Het derde graads polynoom p_3 moet voldoen aan de volgende eisen:

$$\begin{array}{ll} p_3(x_i) = f(x_i) & i = 0, 1 , \\ p_3'(x_i) = f'(x_i) & i = 0, 1 . \end{array}$$

Naar analogie van de Lagrange interpolatiepolynomen is dit polynoom te schrijven als:

$$p_3(x) = \sum_{i=0}^1 f(x_i) H_{i1}(x) + f'(x_i) \hat{H}_{i1}(x) ,$$

met

$$\begin{array}{ll} H_{i1}(x_j) = \delta_{ij} , & \hat{H}_{i1}(x_j) = 0 , \\ H_{i1}(x_j) = 0 , & \hat{H}_{i1}'(x_j) = \delta_{ij} . \end{array}$$

Polynomen die behalve op functiewaarden ook gebaseerd zijn op gegeven afgeleiden staan bekend onder de naam Hermite polynoom.

De algemene uitdrukking voor Hermite polynomen die functiewaarden en eerste afgeleiden bevatten luidt:

$$H_{2n+1}(x) = \sum_{j=0}^n f(x_j) H_{jn}(x) + \sum_{j=0}^n f'(x_j) \hat{H}_{jn}(x) ,$$

waarbij

$$H_{jn}(x) = [(1 - 2(x - x_j) L_{jn}'(x_j)) L_{jn}^2(x)]$$

en

$$\hat{H}_{jn}(x) = (x - x_j) L_{jn}^2(x) .$$

Dat dit het goede polynoom is kunnen we als volgt zien: H_{jn} en $\hat{H}_{jn}$ zijn polynomen van de graad $2n + 1$. In steunpunt x_k geldt $H_{jn}(x_k) = 0$ als $k \neq j$. Als $k = j$ dan geldt:

$$H_{jn}(x_j) = (1 - 2(x_j - x_j) L_{jn}'(x_j)) L_{jn}^2(x_j) = 1 .$$

Omdat $\hat{H}_{jn}(x_k) = 0$ geldt dus $H_{2n+1}(x_j) = f(x_j)$.

Om te laten zien dat ook de afgeleiden overeenstemmen merken we op dat H'_{jn} gegeven wordt door:

$$H'_{jn}(x) = -2L'_{jn}(x_j)L_{jn}^2(x) + [1 - 2(x - x_j)L'_{jn}(x_j)]2L_{jn}(x)L'_{jn}(x) .$$

Door invullen is eenvoudig te controleren dat $H'_{jn}(x_k) = 0$ voor $k = 0, 1, \dots, n$. De afgeleide van $\hat{H}_{jn}$ wordt gegeven door

$$\hat{H}_{jn}(x) = L_{jn}^2(x) + 2(x - x_j)L_{jn}(x)L'_{jn}(x) .$$

Hieruit volgt $\hat{H}'_{jn}(x_k) = \delta_{jk}$ zodat $H'_{2n+1}(x_j) = f'(x_j)$.

Stelling 2.4.2 Als $f \in C^{2n+2}[a, b]$ dan geldt er is een $\xi \in (x_0, \dots, x_n, x)$ zodat

$$f(x) - H_{2n+1}(x) = \frac{(x - x_0)^2 \dots (x - x_n)^2}{(2n + 2)!} f^{(2n+2)}(\xi) .$$

Bewijs:

Het bewijs hiervan gaat analoog aan het bewijs van Stelling 2.2.1. De hulpfunctie wordt nu als volgt gekozen:

$$\varphi(t) = f(t) - H_{2n+1}(t) - \frac{(t - x_0)^2 \dots (t - x_n)^2}{(x - x_0)^2 \dots (x - x_n)^2} [f(x) - H_{2n+1}(x)] ,$$

waar voor bewezen kan worden dat $\varphi'(t)$ $2n + 2$ verschillende nulpunten heeft in $(x_0, \dots, x_n, x)$. $\square$

We zullen aan de hand van twee voorbeelden duidelijk maken waarom soms voor Hermite interpolatie gekozen wordt in plaats van Lagrange interpolatie.

Voorbeeld (seismiek)

Bij het opsporen van olie wordt vaak gebruik gemaakt van seismische golven. Een eenvoudig model voor golfvoortplanting wordt beschreven door het volgende stelsel differentiaalvergelijkingen:

$$\begin{aligned} \frac{dx}{dt} &= c \sin \theta , \\ \frac{dz}{dt} &= -c \cos \theta , \\ \frac{d\theta}{dt} &= -\frac{dc}{dz} \sin \theta . \end{aligned}$$

De positie wordt aangegeven met (x, z) terwijl θ opgevat kan worden als de hoek, die het golf-front maakt met de x -as. We veronderstellen dat de voortplantingssnelheid c alleen afhangt van de verticale positie en bekend is in een eindig aantal meetpunten. Bij het oplossen van dit stelsel hebben we een benadering nodig van $c(z)$ ook in de tussenliggende punten. Lineaire interpolatie in elk interval heeft tot gevolg dat $\frac{dc}{dz}$ niet bestaat in de steunpunten. Dit kan

grote fouten geven in de oplossing. Als zowel c als $\frac{dc}{dz}$ bekend is in de meetpunten, dan kan 3^e graads Hermite interpolatie gebruikt worden in elk interval. Hiervoor bestaat de eerste afgeleide wel in alle punten.

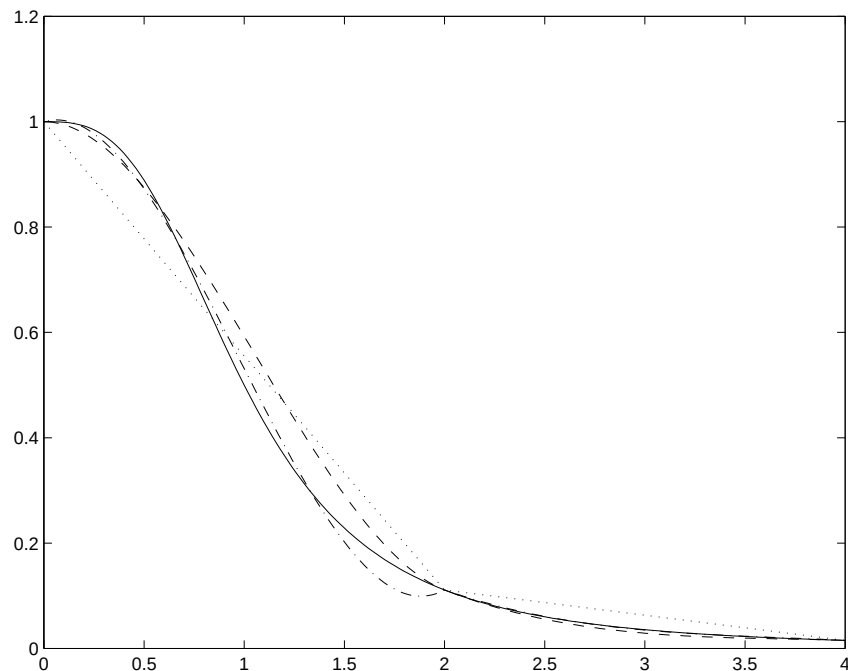
Voorbeeld (visualizatie)

Als we een figuur willen tekenen bepaald door een eindig aantal punten, ontstaan er bij lineaire interpolatie knikken in de grafiek bij de steunpunten. Dit geeft vaak een onrealistische weergave. Een beter resultaat wordt verkregen met Hermite interpolatie.

Stel de grafiek van de functie $f(x) = \frac{1}{1+x^3}$, $x \in [0, 4]$ wordt gebruikt om de helft van een symmetrische heuvel te visualiseren. Om geheugen en rekentijd te sparen werken visualisatie programma's vaak met eenvoudige bouwstenen. Stel dat 3^e graads polynomen daartoe behoren. In Figuur 2.4 is de grafiek benaderd met een aantal interpolatiepolynomen. De resulterende figuur bij lineaire interpolatie lijkt niet op een heuvel. Hermite interpolatie op dezelfde steunpunten geeft een veel beter resultaat. De vergelijking lijkt niet eerlijk omdat bij Hermite interpolatie 3^e graads polynomen zijn gebruikt. Daarom is de functie ook benaderd met 3^e orde Lagrange interpolatie. Ook dit resultaat is onbruikbaar.

2.5 Interpolatie met splines

In de voorgaande paragrafen hebben we gezien dat het benaderen van een functie op een interval tot problemen kan leiden. Het is vaak beter om het interval op te delen in deelintervallen en op elk deelinterval een interpolatie polynoom te construeren. Een probleem hierbij is dat de afgeleide vaak niet bestaat op de overgang van de deelintervallen. Hermite interpolatie



Figuur 2.4: Interpolatie van $\frac{1}{1+x^3}$. — functie, $\cdots$ lineaire interpolatie, — — — Hermite interpolatie, — · — · — 3^e orde Lagrange interpolatie.

verhelpt dit probleem, echter hierbij moet de afgeleide bekend zijn. Bij gemeten grootheden is de afgeleide in de knooppunten onbekend. Een oplossing hiervoor is het gebruik van zogenaamde splines (uit te spreken op zijn Engels).

Een spline is een stuksgewijs polynoom, die glad aansluit in de knooppunten. Wij zullen alleen splines van de eerste en derde orde beschouwen.

Definitie

Voor $f \in C[a, b]$ is de interpolatie spline van graad 1 : $s \in C[a, b]$ zodanig dat als $a = x_0 < x_1 < \dots < x_n = b$ een partitie is van $[a, b]$ dan is s lineair op elk deelinterval $[x_i, x_{i+1}]$ en $s(x_i) = f(x_i)$.

Merk op dat een interpolerende spline van graad 1 niets anders is dan lineaire interpolatie op elk deelinterval. Voor een spline van graad 3 komen wel nieuwe eigenschappen aan bod. Een spline van graad 3 bestaat uit een stuksgewijs derde graads polynoom. In de knooppunten is de waarde van de spline gelijk aan de gegeven functiewaarden en zijn de eerste en tweede afgeleide continu. Dit wordt ook wel een cubische spline genoemd.

Definitie

Voor $f \in C[a, b]$ heeft de interpolerende spline s van de graad 3 de volgende eigenschappen:

- a. s is een derde graads polynoom s_j op elk deelinterval $[x_j, x_{j+1}]$,
- b. $s(x_j) = f(x_j)$,
- c. $s_j(x_{j+1}) = s_{j+1}(x_{j+1})$,
 $s'_j(x_{j+1}) = s'_{j+1}(x_{j+1})$,
 $s''_j(x_{j+1}) = s''_{j+1}(x_{j+1})$,
- d. $s''_0(x_0) = s''_{n-1}(x_n) = 0$.

Merk op dat $s \in C^2[a, b]$. Verder kunnen de voorwaarden in d vervangen worden door andere (rand) voorwaarden.

We zullen nu laten zien hoe zo'n interpolerende spline bepaald kan worden. We schrijven s_j als:

$$s_j(x) = a_j(x - x_j)^3 + b_j(x - x_j)^2 + c_j(x - x_j) + d_j. \quad (2.2)$$

Verder definiëren we $h_j = x_{j+1} - x_j$, $f_j = f(x_j)$ en $G_j = s''_j(x_j)$. Uit b volgt $d_j = f_j$.

We gaan nu de verschillende voorwaarden uit c gebruiken.

$$s''_j(x_{j+1}) = s''_{j+1}(x_{j+1})$$

Uit (2.2) volgt $s''_j(x) = 6a_j(x - x_j) + 2b_j$. Invullen geeft :

$$\begin{array}{lll} G_j = 2b_j & \text{voor } x = x_j & \text{en} \\ G_{j+1} = 6a_j h_j + 2b_j & \text{voor } x = x_{j+1} & . \end{array}$$

We kunnen nu a_j en b_j uitdrukken in de G_j :

$$b_j = \frac{1}{2}G_j \quad \text{en} \quad a_j = \frac{1}{6h_j}(G_{j+1} - G_j).$$

$$s_j(x_{j+1}) = s_{j+1}(x_{j+1})$$

Hieruit volgt $a_j h_j^3 + b_j h_j^2 + c_j h_j + d_j = f_{j+1}$. Invullen van a_j, b_j en d_j geeft voor c_j de volgende uitdrukking:

$$c_j = \frac{f_{j+1} - f_j}{h_j} - h_j \frac{2G_j + G_{j+1}}{6}.$$

$$s'_j(x_{j+1}) = s'_{j+1}(x_{j+1})$$

Dit geeft $3a_j h_j^2 + 2b_j h_j + c_j = c_{j+1}$. Invullen geeft

$$\begin{aligned} & \frac{1}{2} h_j (G_{j+1} - G_j) + G_j h_j + \frac{f_{j+1} - f_j}{h_j} - h_j \frac{2G_j + G_{j+1}}{6} \\ &= \frac{f_{j+2} - f_{j+1}}{h_{j+1}} - h_{j+1} \frac{2G_{j+1} + G_{j+2}}{6}. \end{aligned}$$

Vereenvoudiging geeft:

$$h_j G_j + 2(h_j + h_{j+1})G_{j+1} + h_{j+1}G_{j+2} = 6\left(\frac{f_{j+2} - f_{j+1}}{h_{j+1}} - \frac{f_{j+1} - f_j}{h_j}\right).$$

Deze relatie is geldig voor $j = 0$ tot $j = n - 2$. Dit geeft $n - 1$ vergelijkingen voor de $n + 1$ onbekenden $G_0, \dots, G_n$. Uit d volgt $G_0 = G_n = 0$. Het resulterende stelsel wordt dan:

$$\begin{pmatrix} 2(h_0 + h_1) & h_1 & & & \\ h_1 & 2(h_1 + h_2) & h_2 & & \\ & \ddots & \ddots & \ddots & \\ & & h_{n-2} & 2(h_{n-2} + h_{n-1}) & \end{pmatrix} \begin{pmatrix} G_1 \\ G_2 \\ \vdots \\ G_{n-1} \end{pmatrix} = \begin{pmatrix} 6\left(\frac{f_2 - f_1}{h_1} - \frac{f_1 - f_0}{h_0}\right) \\ \vdots \\ \vdots \\ \vdots \end{pmatrix}$$

Uit dit stelsel kan de waarde van G_j berekend worden. Hiermee kunnen a_j, b_j en c_j bepaald worden.

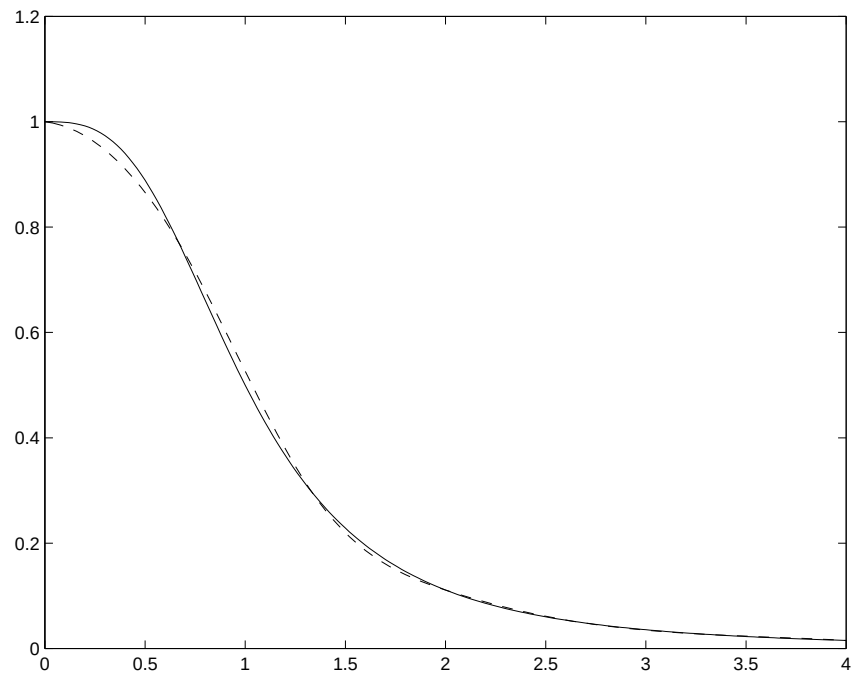
Voorbeeld (visualizatie)

We beschouwen hetzelfde voorbeeld als in de vorige paragraaf. In Figuur 2.5 is er een cubische spline gebruikt waarbij het interval opgedeeld is in 6 deelintervallen. Merk op dat de kwaliteit van de interpolatie beter is dan Hermite interpolatie. Een ander voordeel is dat de functiewaarden in de steunpunten voldoende zijn. Het is niet nodig om de afgeleide te weten in de steunpunten.

2.6 Samenvatting

In dit hoofdstuk zijn de volgende begrippen behandeld:

- lineaire interpolatie
- Lagrange interpolatie
- interpolatie met afgeleiden
- Taylor polynoom



Figuur 2.5: Interpolatie van $\frac{1}{1+x^3}$: — functie, - - spline van 3e orde

- interpolatie in het algemeen
- Hermite interpolatie
- spline interpolatie.

Hoofdstuk 3

Numeriek differentieren

3.1 Inleiding

Het volgende citaat ¹ bespreekt het verschijnsel "muisarm" of RSI (Repetitive Strain Injury):

RSI (Repetitive Strain Injury)

Door de opkomst van MS-Windows als besturingssysteem raken we steeds meer met de muis vergroeid. Na een korte periode van gewenning raakt men al maar enthousiaster. Handig hoor dat klikken, dubbelklikken en slepen. Na een aantal weken of tijdens een intensieve cursus beginnen de problemen. In schouder, arm en pols begint het te tintelen. In het begin gaat de pijn over na een nacht slapen, maar na een aantal maanden helpt dat niet meer. De muisarm dient zich aan. Net als bij een tennisarm ontstaan de problemen door overbelasting van spieren en pezen.

Waarom is de muis dan de boosdoener?

1. de muis ligt meestal niet goed op het bureau en gestrekt bedienen levert overbelasting op,
2. verkeerd knijpen in de muis kost te veel kracht,
3. dubbelklikken en positioneren vereisen te veel coordinatie en kracht (slepen!) en
4. de meeste muizen zijn niet ergonomisch ontworpen.

Om een muisarm te voorkomen wordt er onderzoek gedaan aan de krachten, die optreden in de gewrichten. Hiervoor wordt de positie van de muis gemeten als functie van de tijd. Om de krachten te kunnen bepalen moeten de optredende snelheden en versnellingen bepaald worden. Zoals bekend is de snelheid de eerste afgeleide van de verplaatsing naar de tijd en de versnelling de tweede afgeleide. De vraag is: hoe kunnen we de afgeleide van een functie benaderen als de functie slechts bekend is in een eindig aantal meetpunten. Bovendien bevatten de waarden een meetfout. Het effect van de meetfouten op de benaderde snelheden en versnellingen is van belang. In dit hoofdstuk zullen we aangeven hoe de afgeleiden numeriek bepaald kunnen worden en wat het effect is van meetfouten.

¹<http://www.acaserv.nl/ict/nieuws/rsi.htm>

3.2 Eenvoudige differentieformules voor de eerste afgeleide

Stel f is een continu differentieerbare functie. De voorwaartse differentie is gedefinieerd door:

$$\frac{f(x+h) - f(x)}{h} \quad , \quad h > 0 \quad ,$$

waarbij h de stapgrootte genoemd wordt. Per definitie geldt:

$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = f'(x) \quad ,$$

zodat de voorwaartse differentie naar de afgeleide convergeert. De afbreekfout wordt gedefinieerd door:

$$R_v(h) = f'(x) - \frac{f(x+h) - f(x)}{h} \quad .$$

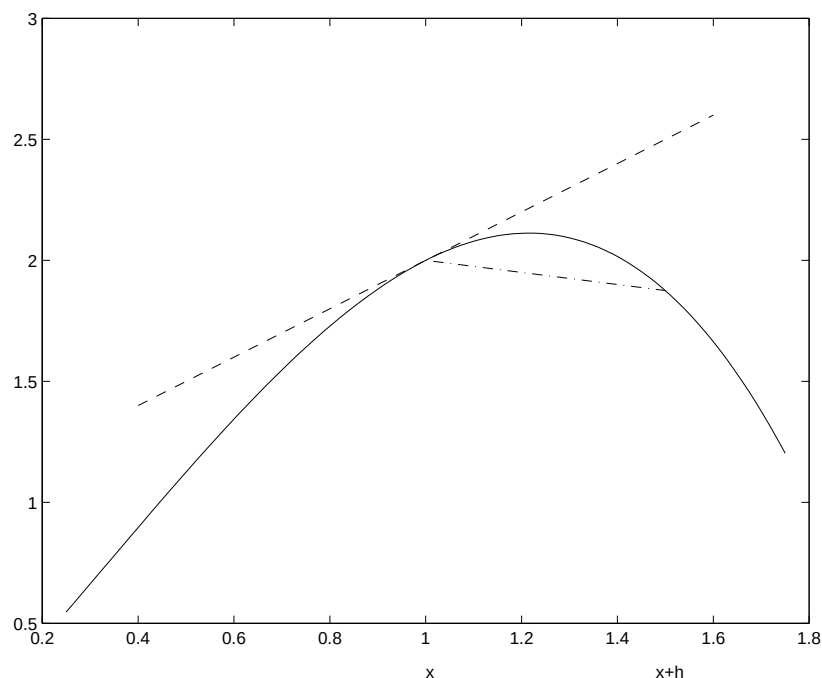
Stelling 3.2.1 *Als $f \in C^2[x, x+h]$ dan geldt $R_v(h) = -\frac{h}{2}f''(\xi)$, met ξ tussen x en $x+h$.*

Bewijs:

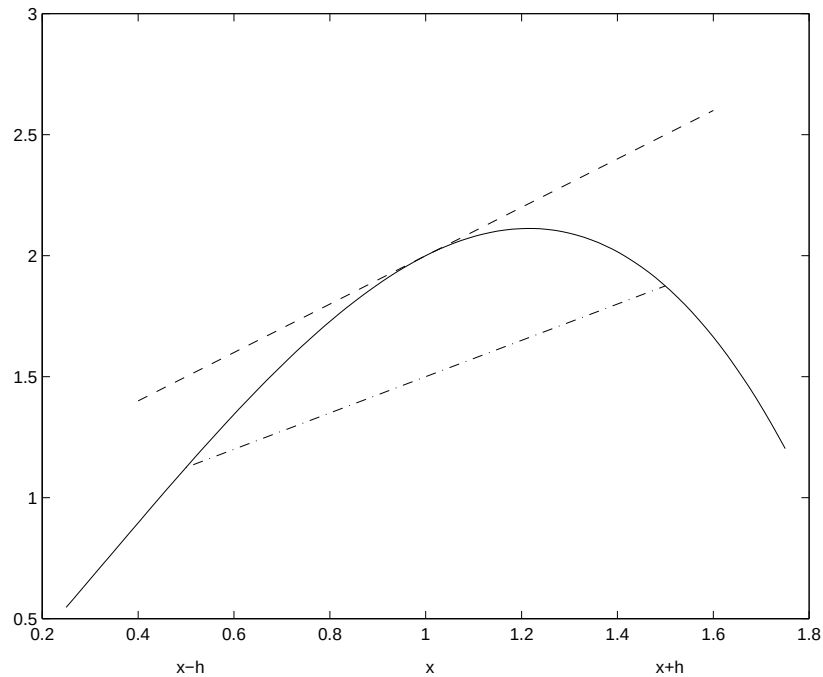
Het resultaat volgt eenvoudig uit een Taylorreeks van $f(x+h)$ rond het punt x . □

Voorbeeld (voorwaartse differentie)

We benaderen de afgeleide van de functie gegeven door $f(x) = -x^3 + x^2 + 2x$ in het punt $x = 1$ met stapgrootte $h = 0.5$. In Figuur 3.1 zien we dat de benadering van de afgeleide nog niet zo nauwkeurig is.



Figuur 3.1: De benadering van de afgeleide (- - -) in $x = 1$ met een voorwaartse differentie ($-\cdot-$).



Figuur 3.2: De benadering van de afgeleide (- - -) in $x = 1$ met een centrale differentie ($-\cdot-$).

Op dezelfde manier kan de achterwaartse differentie gedefinieerd worden:

$$\frac{f(x) - f(x - h)}{h} \quad , \quad h > 0 \quad .$$

Hiervoor geldt dat de afbreekfout gegeven wordt door $R_a(h) = \frac{h}{2}f''(\xi)$ met ξ tussen x en $x - h$. De nauwkeurigheid is vergelijkbaar met de voorwaartse differentie.

Merk op dat de fouten een verschillend teken hebben. Dit brengt ons op het idee om de formules te middelen zodat de fouten tegen elkaar wegvallen (Waarom?). Het resultaat van de middeling staat bekend als de centrale differentie:

$$\frac{f(x + h) - f(x - h)}{2h} \quad .$$

In Figuur 3.2 zien we dat dit inderdaad een betere benadering geeft. Als de derde afgeleide van f bestaat dan wordt de afbreekfout in de centrale differentie:

$$R_2(h) = f'(x) - \frac{f(x + h) - f(x - h)}{2h} \quad ,$$

gegeven door

$$R_2(h) = -\frac{h^2}{6}f'''(\xi) \quad ,$$

met ξ tussen $x - h$ en $x + h$. Om dit te bewijzen ontwikkelen we f in x in een Taylorpolynoom met restterm en gebruiken we de tussenwaardestelling. Merk op dat de fout in de centrale

differentie benadering veel sneller naar nul gaat als functie van h dan bij de voorwaartse formule.

Voorbeeld (afrondfouten)

We gaan de afgeleide van de functie e^x benaderen in het punt $x = 1$. De afgeleide is gelijk aan $e = 2.71828\dots$. We berekenen de afgeleide met de centrale differentie formule en rekenen met 6 cijfers. De gebruikte getallen staan in Tabel 3.1 en de benaderingen in Tabel 3.2. Voor

x	$\exp(x)$
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	3.00416
1.2000	3.32011

h	benadering	fout
0.2000	2.7364	-0.01814
0.1000	2.7228	-0.00451
0.0100	2.7185	-0.00021
0.0010	2.7200	-0.00171
0.0001	2.6999	0.01828

Tabel 3.2: De resultaten met de centrale differentie formule

Tabel 3.1: De gebruikte waarden

dalende h wordt de benadering eerst beter, maar daarna verslechtert hij. Deze verslechtering wordt veroorzaakt door afrondfouten. We werken dit uit voor:

$$f(1.001) - f(0.999) = 2.72100 - 2.71556 = 0.00564 .$$

De functiewaarden zijn belast met afrondfouten. De absolute waarde van deze fouten is hoogstens $5 \cdot 10^{-6}$. De relatieve fouten zijn dan hoogstens $5 \cdot 10^{-6}/2.7 \approx 2 \cdot 10^{-6}$. Een bovengrens voor de absolute fout in bovenstaand verschil is $10 \cdot 10^{-6}$, hetgeen een relatieve fout in het verschil veroorzaakt van

$$10^{-5}/0.00564 \approx 2 \cdot 10^{-3} .$$

Deze fout is 1000 maal zo groot dan de maximale relatieve fout in de functiewaarden. Dit effect wordt sterker naar mate h kleiner wordt.

Het bovenstaande voorbeeld laat zien dat de te bereiken nauwkeurigheid voor de bepaling van de afgeleide met een differentieformule begrensd is. We zullen dit voor de centrale differentie verder analyseren. Door het verkleinen van de stapgrootte h zal de fout $R_2(h)$ kleiner worden, echter de bovengrens $S_2(h)$ van de fout tengevolge van afrondfouten zal toenemen. Er is dus een optimale waarde van h waarbij de totale fout $|R_2(h)| + S_2(h)$ het kleinst is. Deze waarde van h is onbekend maar men kan wel een benadering hiervoor vinden.

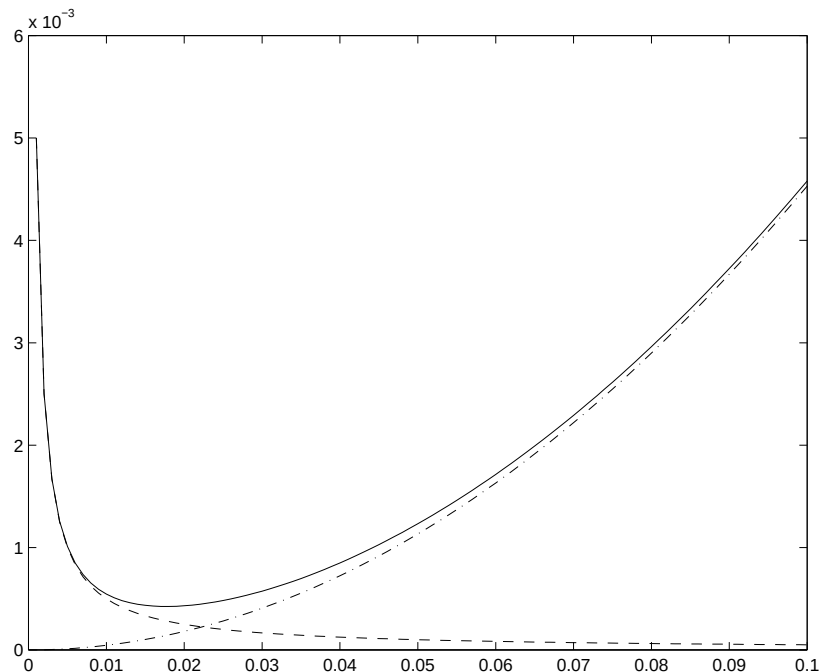
Stel dat de functiewaarden een afrondfout bevatten van hoogstens ε . Dit levert een bovengrens voor de fout $S_2(h) = \frac{\varepsilon}{h}$ (waarom?) in de numerieke benadering van $f'(1)$. De totale fout wordt dus gemajoreerd door de grootte $\varphi(h)$, die gegeven wordt door:

$$\varphi(h) = \frac{\varepsilon}{h} + \frac{h^2}{6}m ,$$

waarbij m een bovengrens is voor $|f'''(x)|$ op een voldoende grote omgeving van 1. De functie φ heeft een minimum $\sqrt[3]{\frac{9}{8}\varepsilon^2 m}$ voor $h_{\text{opt}} = \sqrt[3]{\frac{3\varepsilon}{m}}$ (leid dit zelf af). Voor ons voorbeeld staat $\varphi(h)$, $S_2(h)$ en $|R_2(h)|$ getekend in Figuur 3.3. Voor $m = e$ volgt $h_{\text{opt}} = 0.017$ en $\varphi(h_{\text{opt}}) = 0.00043$. Merk op dat de fout in de buurt van h_{opt} niet veel varieert, zodat een stapgrootte in de buurt van h_{opt} tot een vergelijkbare nauwkeurigheid al leiden.

Opmerkingen

1. Als de benadering van de afgeleide niet nauwkeurig genoeg is dan zijn er twee mogelijke aanpakken om dit te verbeteren:
 - de afrond- of meetfout in de functiewaarden verkleinen,
 - overgang op een hogere orde differentieformule. In dit geval is de afbreekfout $R(h)$ vaak kleiner dan bij een lage orde formule voor grotere waarden van h . Voor grote h neemt het effect van afrondfouten af.
2. Vaak worden differentieformules gebruikt voor het oplossen van differentiaalvergelijkingen. In dit geval wil men niet de afgeleide nauwkeurig bepalen maar de oplossing. Het gedrag ten opzichte van afrondfouten is dan meestal gunstiger.
3. Het effect van afrondfouten is het grootst als de fout in de functiewaarden groot is. Dit is vaak het geval bij metingen zoals aan de muis. In het geval van een door de computer gegenereerde tabel is de afrondfout meestal een stuk kleiner dan bij meetfouten. Desalniettemin voor kleine h is hij altijd merkbaar.



Figuur 3.3: De grafiek van de totale fout φ (—), de afbreekfout $|R_2(h)|$ (— · —) en de afrondfout $S_2(h)$ (- - -).

3.3 Algemene formules voor de eerste afgeleide

De differentieformules gegeven in de vorige paragraaf zijn op een ad-hoc manier geïntroduceerd. In een aantal toepassingen is het nodig om differentieformules te gebruiken met een hogere nauwkeurigheid (denk aan afrondfouten) of met roosterpunten, die niet equidistant verdeeld zijn. In deze paragraaf zullen we een methode presenteren om differentieformules af te leiden.

Stel er zijn $n+1$ verschillende roosterpunten $x_0, \dots, x_n$ gegeven. Gevraagd: bepaal de coëfficiënten α_i zodanig dat de orde van de fout:

$$|f'(x) - Q(h)| = O(h^k)$$

maximaal is. Hierbij wordt $Q(h)$ gegeven door:

$$Q(h) = \sum_{i=0}^n \alpha_i f_i .$$

De aanpak gaat als volgt: ontwikkel elke f_i in een Taylorpolynoom van voldoende hoge graad met x als steunpunt. Bepaal $\alpha_0, \dots, \alpha_n$ zodanig dat de orde van de fout maximaal is. We zullen twee voorbeelden uitwerken:

Centrale differentie

In dit voorbeeld kiezen we $x_{-1} = x - h$, $x_0 = x$ en $x_1 = x + h$. Taylorontwikkeling geeft:

$$\begin{aligned} f(x+h) &= f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{3!}f'''(x) + O(h^4) \\ f(x) &= f(x) \\ f(x-h) &= f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{3!}f'''(x) + O(h^4) \end{aligned}$$

De voorwaarden voor α_{-1} , α_0 en α_1 zijn nu:

$$\begin{aligned} f(x) : & \quad \alpha_{-1} + \alpha_0 + \alpha_1 = 0 , \\ f'(x) : & \quad -h\alpha_{-1} + h\alpha_1 = 1 , \\ f''(x) : & \quad h^2\alpha_{-1} + h^2\alpha_1 = 0 . \end{aligned}$$

Oplossing van dit stelsel geeft $\alpha_{-1} = \frac{-1}{2h}$, $\alpha_0 = 0$ en $\alpha_1 = \frac{1}{2h}$ waarmee we de reeds bekende formule:

$$Q(h) = \frac{f(x+h) - f(x-h)}{2h}$$

verkrijgen. Merk op dat de fout bij deze formule $O(h^2)$ is.

Eenzijdige differentieformule van de orde 2

Stel dat de afgeleide gewenst is aan de rand van je meetgebied. Bovendien is een orde 2 van de afbreekfout gewenst. Kies $x_0 = x$, $x_1 = x + h$ en $x_2 = x + 2h$ en bepaal de Taylorontwikkeling.

$$\begin{aligned} f(x) &= f(x) , \\ f(x+h) &= f(x) + hf'(x) + \frac{h^2}{2}f''(x) + O(h^3) , \\ f(x+2h) &= f(x) + 2hf'(x) + 2h^2f''(x) + O(h^3) . \end{aligned}$$

We verkrijgen nu de volgende voorwaarden:

$$\begin{aligned} f(x) : & \quad \alpha_0 + \alpha_1 + \alpha_2 = 0, \\ f'(x) : & \quad h\alpha_1 + h\alpha_2 = 1, \\ f''(x) : & \quad \frac{h^2}{2}\alpha_1 + 2h^2\alpha_2 = 0. \end{aligned}$$

Oplossing van dit stelsel geeft $\alpha_0 = \frac{-3}{2h}$, $\alpha_1 = \frac{2}{h}$ en $\alpha_2 = \frac{-1}{2h}$. De afbreekfout van de verkregen eenzijdige differentieformule

$$Q(h) = \frac{-3f(x) + 4f(x+h) - f(x+2h)}{2h} \quad (3.1)$$

is $O(h^2)$.

3.4 Relatie tussen differentieformules en interpolatie

In paragraaf 3.3 hebben we een algemene manier besproken om differentieformules af te leiden. Echter een duidelijk nadeel is dat we een stelsel vergelijkingen moeten oplossen waarbij het aantal onbekenden toeneemt naarmate de orde van nauwkeurigheid toeneemt.

In deze paragraaf zullen we een totaal andere aanpak volgen. In paragraaf functiewaarden met steunpunten behandeld. Dit interpolatiepolynoom kan natuurlijk ook gezien worden als een benadering van de oorspronkelijke functie. Het lijkt dus voor de hand te liggen, dat de afgeleide van het interpolatiepolynoom ook een benadering vormt voor de afgeleide van de functie. Vanuit deze redenering zullen we op een alternatieve wijze differentieformules afleiden.

We herhalen de formule voor Lagrange interpolatie. Veronderstel dat $x_0, x_1, \dots, x_n$ verschillend zijn en $f \in C^{n+1}[a, b]$. Dan geldt:

$$f(x) = \sum_{k=0}^n f(x_k) L_{kn}(x) + (x - x_0) \dots (x - x_n) \frac{f^{n+1}(\xi(x))}{(n+1)!},$$

waarbij $\xi(x) \in [a, b]$. Differentiëren van deze uitdrukking geeft:

$$\begin{aligned} f'(x) &= \sum_{k=0}^n f(x_k) L'_{kn}(x) + \frac{d}{dx} [(x - x_0) \dots (x - x_n)] \frac{f^{n+1}(\xi(x))}{(n+1)!} \\ &\quad + (x - x_0) \dots (x - x_n) \frac{d}{dx} \frac{f^{n+1}(\xi(x))}{(n+1)!} \end{aligned}$$

De foutschatting is nu een gecompliceerde uitdrukking tenzij x gelijk genomen wordt aan één van de steunpunten. Voor $x = x_j$ wordt de formule voor de afgeleide:

$$f'(x_j) = \sum_{k=0}^n f(x_k) L'_{kn}(x_j) + \prod_{\substack{k=0 \\ k \neq j}}^n (x_j - x_k) \frac{f^{n+1}(\xi(x))}{(n+1)!}.$$

Voorbeeld (voorwaartse differentie)

Voor $n = 1$, $x = x_0$ en $x_1 = x_0 + h$ geldt:

$$\begin{aligned} L_{01}(x) &= \frac{x - x_1}{x_0 - x_1}, & L'_{01}(x) &= \frac{-1}{h}, \\ L_{11}(x) &= \frac{x - x_0}{x_1 - x_0}, & L'_{11}(x) &= \frac{1}{h}. \end{aligned}$$

De formule is dan: $f'(x_0) = -\frac{1}{h}f(x_0) + \frac{1}{h}f(x_0 + h) + \frac{h}{2}f''(\xi)$. Dit is precies de formule voor de voorwaartse differentie.

3.5 Differentieformules voor hogere afgeleiden

In de inleiding hebben we een voorbeeld gezien waarbij het bepalen van de snelheid en de versnelling uit gemeten plaatswaarden nodig is. De versnelling is gelijk aan de tweede afgeleide van de plaats naar de tijd.

Hogere afgeleiden kunnen benaderd worden op dezelfde manier als de methoden waarmee de eerste afgeleide benaderd is. We beperken ons hier tot het bepalen van de tweede afgeleide door herhaald toepassen van de formules voor de eerste afgeleide. Voor de numerieke bepaling van de 2e afgeleide hebben we minimaal 3 punten nodig (waarom?). Stel we willen de 2e afgeleide benaderen met 3 punten, die op een equidistante afstand liggen:

$$\begin{array}{ccccccc} & & x - \frac{1}{2}h & & x + \frac{1}{2}h & & \\ & \times & \bullet & \times & \bullet & \times & \\ & x-h & & x & & x+h & \end{array}$$

Het ligt voor de hand om de 2e afgeleide te benaderen door het verschil te nemen van de eerste afgeleide in de punten $x+h$ en $x-h$. Echter bij toepassing van centrale differenties betekent dit, dat wij de functie ook in $x+2h$ en $x-2h$ nodig hebben. We kunnen dit ondervangen door de hulppunten $x \pm \frac{1}{2}h$ in te voeren. Centrale differentie geeft:

$$\frac{1}{h} \left(\frac{f'(x + \frac{1}{2}h) - f'(x - \frac{1}{2}h)}{h} \right)$$

als benadering van $f''(x)$. Op de beide eerste orde afgeleiden passen we opnieuw een centrale differentie toe:

$$\frac{1}{h} \left(\frac{f(x+h) - f(x)}{h} - \frac{f(x) - f(x-h)}{h} \right).$$

Vereenvoudiging van deze formule geeft:

$$Q(h) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}.$$

Met behulp van Taylorpolynomen volgt voor de afbreekfout:

$$f''(x) - Q(h) = \frac{-h^2}{12} f^{(4)}(\xi).$$

Merk op dat het effect van afrondfouten nu nog ernstiger is. Een bovengrens voor de fout ten gevolge van afrondfouten is:

$$S(h) = \frac{4\varepsilon}{h^2}.$$

Voor het oplossen van differentiaalvergelijkingen is het soms nodig om een benadering te

vinden voor $(pf')'$ waarbij p een gegeven functie is. Analoog aan bovenstaande kunnen we de volgende benadering van $(pf')'$ opschrijven:

$$\frac{(pf')'(x + \frac{1}{2}h) - (pf')'(x - \frac{1}{2}h)}{h}.$$

Opnieuw toepassen van centrale differenties geeft:

$$\frac{p(x + \frac{1}{2}h)(f(x + h) - f(x)) - p(x - \frac{1}{2}h)(f(x) - f(x - h))}{h^2}.$$

3.6 Richardson's extrapolatie

3.6.1 Inleiding

In paragraaf 3.2 is aangetoond dat de fout in de voorwaartse differentiebenadering afhangt van f'' . Meestal is f'' onbekend zodat de gegeven foutschatting onbruikbaar is in de praktijk. Richardson's extrapolatie kan gebruikt worden om een schatting te geven van de gemaakte fout. Daarnaast kan de extrapolatie gebruikt worden om met een lagere orde methode een hogere nauwkeurigheid te halen. Deze extrapolatie kan gebruikt worden als de vorm van de fout bekend is. Stel de formule $N(h)$ is bedoeld om een onbekende waarde M te benaderen. Neem aan dat de fout in deze benadering de vorm

$$M - N(h) = K_1 h^{\alpha_1} + K_2 h^{\alpha_2} + \dots \quad (3.2)$$

heeft, waarbij $K_i \in \mathbb{R}$ ongelijk aan nul is en voor $\alpha_i \in \mathbb{N}$ geldt: $1 \leq \alpha_1 < \alpha_2 < \dots$

In deze paragraaf zal Richardson's extrapolatie toegepast worden op differentieformules, echter de methode kan ook voor andere benaderingsmethoden (interpolatie, numerieke integratie etc) gebruikt worden, zolang de fout van de vorm (3.2) is.

Als voorbeeld van het gebruik van Richardson's extrapolatie zullen we in paragraaf 3.6.3 laten zien hoe we een hogere nauwkeurigheid kunnen verkrijgen als de α_i bekend zijn. Daarna zullen we in paragraaf 3.6.2 een foutschatting geven als de α_i onbekend zijn. Natuurlijk kan in beide toepassingen Richardson's extrapolatie gebruikt worden zowel als de α_i bekend, of als de α_i onbekend zijn.

3.6.2 Schatting van de fout in de praktijk

In de voorgaande paragrafen is er vaak een uitdrukking gegeven voor de fout in de differentieformules. Deze uitdrukking bevat vaak een hogere afgeleide van de functie. Echter als we de eerste afgeleide niet weten, hoe zouden we dan een hogere afgeleide kunnen bepalen? Kortom de gegeven schattingen zijn vaak theoretisch van belang, maar in de praktijk onbruikbaar. We zullen daarom een bruikbare methode geven met behulp van Richardson's extrapolatie.

Voor h klein genoeg kunnen we (3.2) benaderen door

$$M - N(h) = Kh^\alpha. \quad (3.3)$$

Voor een gegeven h kunnen we $N(h)$ uitrekenen zodat (3.3) nog 3 onbekenden bevat: M, K en α . Het idee is nu om $N(h), N(\frac{h}{2})$ en $N(\frac{h}{4})$ te bepalen. Dit geeft 3 vergelijkingen met 3 onbekenden:

$$\begin{aligned} M - N(h) &= Kh^\alpha, \\ M - N(\frac{h}{2}) &= K(\frac{h}{2})^\alpha, \\ M - N(\frac{h}{4}) &= K(\frac{h}{4})^\alpha \end{aligned} \quad (3.4)$$

Door de vergelijkingen van elkaar af te trekken houden we 2 onbekenden over:

$$\begin{aligned} N(\frac{h}{2}) - N(h) &= Kh^\alpha(1 - (\frac{1}{2})^\alpha), \\ N(\frac{h}{4}) - N(\frac{h}{2}) &= K(\frac{h}{2})^\alpha(1 - (\frac{1}{2})^\alpha). \end{aligned} \quad (3.5)$$

Het delen van deze uitdrukkingen geeft:

$$\frac{N(\frac{h}{2}) - N(h)}{N(\frac{h}{4}) - N(\frac{h}{2})} = 2^\alpha. \quad (3.6)$$

Hieruit kan α bepaald worden. Invullen van α in (3.5) geeft een benadering voor K . Beide constanten invullen in (3.4) geeft een schatting voor de fout $M - N(\frac{h}{4})$.

Voorbeeld (praktische foutschatting)

Stel we benaderen de afgeleide van e^x in het punt 1 met behulp van de voorwaartse differentie. Het exacte antwoord is: $M = e = 2.71828$. De resultaten voor $h = 0.1, \frac{h}{2}$ en $\frac{h}{4}$ staan in Tabel 3.3.

h	Q_h
0.1	2.8588
0.05	2.7874
0.025	2.7525

Tabel 3.3: De eerste orde afgeleide van e^x met de voorwaartse differentie

Als we deze resultaten invullen in (3.6) dan krijgen we

$$\frac{2.7874 - 2.8588}{2.7525 - 2.7874} = 2.0458.$$

Zoals verwacht is de waarde van α bijna gelijk aan 1. Hiermee kunnen we K bepalen uit (3.5):

$$K = \frac{2.7525 - 2.7874}{0.05(1 - \frac{1}{2})} = -1.3936.$$

De foutschatting is nu $K\frac{h}{4} = -0.0348$ terwijl de echte fout gelijk is aan:

$$e - N(\frac{h}{4}) = -0.0343.$$

In dit voorbeeld is de foutschatting heel betrouwbaar.

Om een betere benadering te vinden zou de foutschatting opgeteld kunnen worden bij de benadering:

$$N\left(\frac{h}{4}\right) + K\frac{h}{4} = 2.7177 .$$

Dit is inderdaad een hele goede benadering echter we kunnen hiervoor nu geen foutschatting meer geven.

De vraag zou kunnen zijn: waarom gaan we in bovenstaand voorbeeld niet uit van $\alpha = 1$; dit volgt immers uit de analyse? Inderdaad als de functie voldoende glad is dan geldt $\alpha = 1$ en zou α niet bepaald behoeven te worden. Echter in de praktijk hebben we vaak een ingewikkelder situatie. De volgende complicaties kunnen zich voordoen:

- Het is onbekend of hogere afgeleiden bestaan en/of begrensd zijn.
- Het uiteindelijke antwoord is een combinatie van verschillende benaderingsmethoden. De invloed van deze benaderingen op α is niet altijd duidelijk.
- Tijdens het implementeren van het algoritme in een computerprogramma zijn fouten gemaakt.

Ten gevolge van deze complicaties is het in de praktijk goed om te controleren of de berekende α overeenstemt met de α die volgt uit de theorie.

3.6.3 Nauwkeuriger formules via Richardson's extrapolatie

In een aantal toepassingen zijn de waarde van α_i in (3.2) bekend. We kunnen Richardson's extrapolatie dan gebruiken om nauwkeuriger formules te bepalen. Als voorbeeld zullen we de voorwaartse differentie beschouwen. De fout in de voorwaartse differentieformule blijkt door gebruik van hogere orde Taylorpolynomen geschreven te kunnen worden als:

$$f'(x) - Q_v(h) = K_1 h + K_2 h^2 + \dots . \quad (3.7)$$

Als we de differentie ook uitrekenen voor $\frac{h}{2}$ dan geldt:

$$f'(x) - Q_v\left(\frac{h}{2}\right) = K_1 \frac{h}{2} + K_2 \left(\frac{h}{2}\right)^2 + \dots . \quad (3.8)$$

Als we nu (3.8) met 2 vermenigvuldigen en (3.7) hier van aftrekken dan geldt:

$$f'(x) - (2Q_v\left(\frac{h}{2}\right) - Q_v(h)) = -\frac{1}{2}K_2 h^2 + \dots .$$

Hieruit volgt dat de differentieformule

$$\begin{aligned} 2Q_v\left(\frac{h}{2}\right) - Q_v(h) &= \frac{2f\left(x + \frac{h}{2}\right) - f(x)}{\frac{h}{2}} - \frac{f(x+h) - f(x)}{h} \\ &= \frac{-3f(x) + 4f\left(x + \frac{h}{2}\right) - f(x+h)}{h} , \end{aligned}$$

een afbreekfout $O(h^2)$ heeft. Deze formule is gelijk aan (3.1) met $h/2$ in plaats van h .

3.7 Samenvatting

In dit hoofdstuk zijn de volgende begrippen behandeld:

- differentie methode voor de eerste afgeleide
- het effect van meetfouten
- differentie methode voor hogere afgeleiden
- Richardson's extrapolatie
 - verhogen nauwkeurigheid
 - foutschatting in de praktijk.

Hoofdstuk 4

Numerieke tijdsintegratie voor beginwaarde problemen

4.1 Inleiding

Een beginwaarde probleem is meestal een wiskundige beschrijving (differentiaalvergelijking) van een tijdsafhankelijk probleem. De voorwaarden, die nodig zijn om een unieke oplossing te bepalen, worden gegeven op $t = 0$ en worden beginvoorwaarden genoemd.

Als voorbeeld beschouwen we de water afvoer uit een stuwmeer door een leiding (zie Figuur 4.1). Het water is in rust totdat op $t = 0$ de schuif plotseling geopend wordt. Door de traagheid van het water komt de stroming geleidelijk op gang. Als dit stromende water gebruikt wordt om electriciteit op te wekken, dan is het van belang om te weten hoe lang het duurt, voordat de turbines op vol vermogen werken.

Een beschrijving van dit proces wordt gegeven door de niet-lineaire differentiaalvergelijking

$$\frac{dQ}{dt} = p(t) - aQ^2,$$

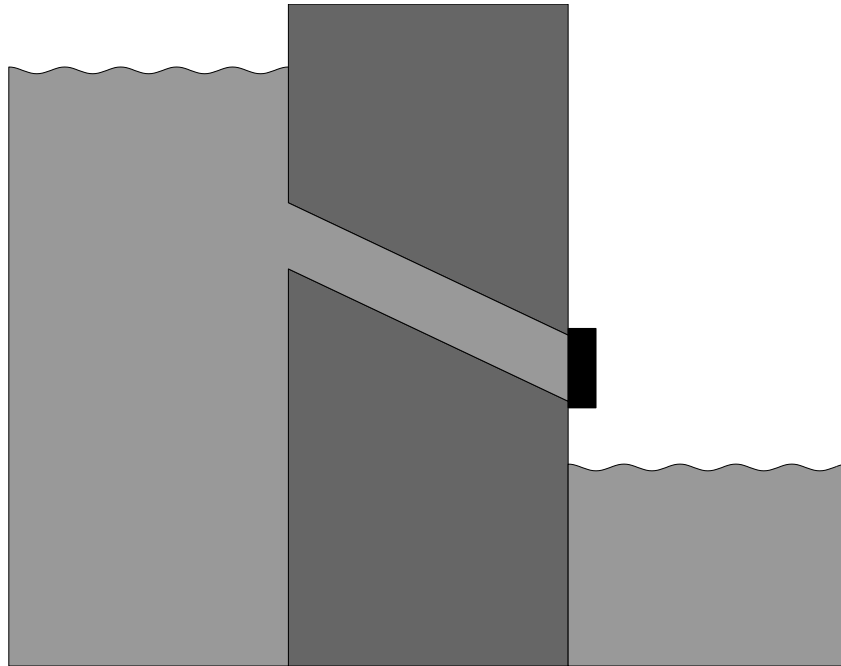
met beginvoorwaarde

$$Q(0) = 0.$$

Hierin is $Q(m^3/s)$ de afvoer, p is de drijvende kracht ($\frac{\text{kracht/lengte}}{\text{dichtheid}}$ ofwel $\frac{N/m}{kg/m^3} = m^3/s^2$) en aQ^2 is de wrijvingskracht. De drijvende kracht is onder andere afhankelijk van het waterpeil in het stuwmeer. Als we veronderstellen dat $p(t)$ constant is ($p(t) = p_0$) dan is een analytische oplossing bekend namelijk

$$Q = \sqrt{\frac{p_0}{a}} \tanh(t\sqrt{ap_0}).$$

De numerieke oplossing kan voor willekeurige functies p bepaald worden.



Figuur 4.1: Het stuwmeer, de afvoerbuiscap en de schuif in gesloten toestand

4.2 Eenstaps methoden

De vergelijking

De algemene vorm van een beginwaarde probleem van de eerste orde wordt gegeven door

$$y' = f(t, y), \quad t \in [0, T], \quad (4.1)$$

met beginvoorwaarde

$$y(0) = y_0.$$

Het rooster

De tijd s wordt onderverdeeld in intervallen ter lengte h . De knooppunten zijn $t_j = jh$ en u_j is een numerieke benadering van de exacte oplossing $y_j = y(t_j)$.

De voorwaartse methode van Euler

Als de term y' in (4.1) vervangen wordt door de voorwaartse differentie, dan geldt

$$\frac{u_{j+1} - u_j}{h} = f(t_j, u_j).$$

Dit kunnen we ook schrijven als

$$u_{j+1} = u_j + hf(t_j, u_j). \quad (4.2)$$

Als we starten met de beginwaarde $u_0 = y_0$ dan kan de numerieke benadering u_j verkregen worden door (4.2) herhaald toe te passen.

De achterwaartse methode van Euler

Als de term y' in (4.1) vervangen wordt door de achterwaartse differentie, dan krijgen we

$$u_{j+1} = u_j + hf(t_{j+1}, u_{j+1}) . \quad (4.3)$$

Merk op dat er een belangrijk verschil is tussen beide methoden. In (4.2) is u_{j+1} gegeven als functie van t_j en u_j . Zo'n methode noemen we een expliciete methode. Voordeel van deze methoden is dat de numerieke benaderingen eenvoudig bepaald kunnen worden. Bij Euler achterwaarts is u_{j+1} impliciet gegeven. Voor het verkrijgen van u_{j+1} moet de (eventueel niet-lineaire) vergelijking

$$u_{j+1} - hf(t_{j+1}, u_{j+1}) = u_j$$

opgelost worden, hetgeen veel werk kan kosten. Methoden waarbij u_{j+1} ook in f voorkomt noemen we impliciet. In hoofdstuk 8 komen we terug op het numeriek oplossen van niet-lineaire vergelijkingen.

Trapeziumregel

Middeling van de beide methoden van Euler geeft

$$u_{j+1} = u_j + \frac{h}{2}[f(t_j, u_j) + f(t_{j+1}, u_{j+1})] . \quad (4.4)$$

De Trapeziumregel is impliciet. Een alternatief om deze methode af te leiden gaat als volgt. Integreer de differentiaalvergelijking (4.1) van t_j tot t_{j+1}

$$y_{j+1} = y_j + \int_{t_j}^{t_{j+1}} f(\xi, y(\xi)) d\xi .$$

Als de integraal benaderd wordt door het oppervlak van het trapezium met hoekpunten $(t_j, 0)$, $(t_j, f(t_j, y_j))$, $(t_{j+1}, 0)$ en $(t_{j+1}, f(t_{j+1}, y_{j+1}))$ dan komen we opnieuw uit op formule (4.4).

De methode van Heun

We kunnen de Trapeziumregel (4.4) expliciet maken door in het rechterlid een voorspelling u_{j+1}^* te gebruiken van u_{j+1} met behulp van de voorwaartse methode van Euler. Dit geeft de methode van Heun:

$$\begin{aligned} \text{(predictor)} \quad u_{j+1}^* &= u_j + hf(t_j, u_j) , \\ \text{(corrector)} \quad u_{j+1} &= u_j + \frac{h}{2}[f(t_j, u_j) + f(t_{j+1}, u_{j+1}^*)] . \end{aligned} \quad (4.5)$$

Dit is een voorbeeld van een predictor-corrector formule.

In de volgende paragrafen zullen we de verschillende methoden onderzoeken op hun eigenschappen, zodat een afgewogen keuze gemaakt kan worden wanneer men een praktisch probleem numeriek op wil lossen.

4.3 Testvergelijking en versterkingsfactor

De differentiaalvergelijking (4.1) is te algemeen geformuleerd om bruikbaar te zijn voor analyse. Daarvoor beperken we ons tot de homogene testvergelijking.

Testvergelijking

$$y' = \lambda y \quad \text{en} \quad y(0) = y_0 . \quad (4.6)$$

De exacte oplossing van deze vergelijking is $y(t) = y_0 e^{\lambda t}$.

Versterkingsfactor

De numerieke methoden uit de vorige paragraaf kunnen we toepassen op (4.6). Alle methoden blijken voor de testvergelijking geschreven te kunnen worden als

$$u_{j+1} = k(h\lambda)u_j . \quad (4.7)$$

De functie k heet de versterkingsfactor. Eenvoudig is na te gaan dat:

$$\text{Euler voorwaarts} \quad k(h\lambda) = 1 + h\lambda , \quad (4.8)$$

$$\text{Euler achterwaarts} \quad k(h\lambda) = \frac{1}{1 - h\lambda} , \quad (4.9)$$

$$\text{Trapeziumregel} \quad k(h\lambda) = \frac{1 + \frac{1}{2}h\lambda}{1 - \frac{1}{2}h\lambda} , \quad (4.10)$$

$$\text{Heun} \quad k(h\lambda) = 1 + h\lambda + \frac{1}{2}(h\lambda)^2 . \quad (4.11)$$

De versterkingsfactor speelt een grote rol bij het analyseren van de verschillende numerieke methoden.

4.4 Afbreukfout en consistentie

Om het verschil tussen de exacte en de numerieke oplossing te schatten definiëren we opnieuw de begrippen: consistentie, stabiliteit en convergentie. Elke eenstaps methode is te schrijven als

$$u_{j+1} = u_j + h\phi(u_{j+1}, u_j, t_{j+1}, t_j, h) .$$

Als ϕ niet afhangt van u_{j+1} dan noemen we de methode expliciet, anders impliciet. In paragraaf 4.11 worden meerstapsmethoden behandeld. Deze notatie zullen we gebruiken om een aantal begrippen te definiëren.

Afbreukfout

De afbreukfout e_{j+1} is gedefinieerd als

$$e_{j+1} = \frac{y_{j+1} - y_j}{h} - \phi(y_{j+1}, y_j, t_{j+1}, t_j, h) . \quad (4.12)$$

Merk op dat de exacte oplossing ingevuld is in de numerieke methode.

Consistentie

Een methode heet consistent als

$$\lim_{h \rightarrow 0} e_j = 0 ,$$

waarbij in de limiet j zo gekozen wordt, dat jh constant is. We noemen een methode van orde p als $e_j = O(h^p)$.

Afbreekfout bij de testvergelijking

Als we een eenstaps methode toepassen op de testvergelijking dan volgt uit (4.7)

$$e_{j+1} = \frac{y_{j+1} - k(h\lambda)y_j}{h} . \quad (4.13)$$

De exacte oplossing van de testvergelijking wordt gegeven door

$$y_{j+1} = e^{h\lambda} y_j . \quad (4.14)$$

Substitutie van (4.14) in (4.13) geeft

$$e_{j+1} = \frac{e^{h\lambda} - k(h\lambda)}{h} y_j .$$

De grootte van de afbreekfout wordt dus bepaald door het verschil tussen de versterkingsfaktor en de exponentiële functie. Om dit verschil te bepalen maken we gebruik van de Taylorontwikkeling van $e^{h\lambda}$:

$$e^{h\lambda} = \sum_{i=0}^{r-1} \frac{1}{i!} (h\lambda)^i + O(h^r) . \quad (4.15)$$

Als we (4.15) vergelijken met (4.8) dan volgt dat Euler voorwaarts een afbreekfout van de orde 1 heeft. De methode van Heun heeft volgens (4.11) een afbreekfout van de orde 2 en is dus nauwkeuriger.

Voor de achterwaartse methode van Euler volgt uit (4.9) dat

$$k(h\lambda) = 1 + h\lambda + (h\lambda)^2 + O(h^3) ,$$

zodat de afbreekfout bij Euler achterwaarts ook van de orde 1 is. Bij de Trapeziumregel is de orde van de afbreekfout gelijk aan 2.

Afbreekfout voor de algemene vergelijking

Voor een expliciete methode geldt

$$e_{j+1} = \frac{y_{j+1} - y_j}{h} - \phi(y_j, t_j, h) . \quad (4.16)$$

Taylorontwikkeling van y_{j+1} geeft

$$y_{j+1} = y_j + hy'_j + \frac{h^2}{2} y''_j + O(h^3) . \quad (4.17)$$

Omdat y voldoet aan de differentiaalvergelijking geldt:

$$y' = f(t, y) \quad (4.18)$$

en

$$y'' = f_t + f_y f. \quad (4.19)$$

Combinatie van (4.16), (4.17), (4.18) en (4.19) geeft

$$e_{j+1} = f_j + \frac{h}{2}(f_t + f_y f)_j - \phi(y_j, t_j, h) + O(h^2).$$

Bij de voorwaartse methode van Euler is $\phi(y_j, t_j, h) = f(t_j, y_j)$ en is de afbreekfout van de orde 1. Dit kunnen we als volgt bewijzen. Uit (4.16) en (4.17) volgt

$$e_{j+1} = y'_j + O(h) - \phi(y_j, t_j, h)$$

Omdat voor Euler voorwaarts $\phi(y_j, t_j, h) = f(t_j, y_j)$ volgt uit (4.18) $e_{j+1} = O(h)$. Bij de methode van Heun komen we uit op

$$e_{j+1} = f_j + \frac{h}{2}(f_t + f_y f)_j - \frac{1}{2}[f_j + f(t_{j+1}, y_j + hf_j)] + O(h^2). \quad (4.20)$$

Uit de stelling van Taylor voor een functie van twee veranderlijken volgt

$$f(t_{j+1}, y_j + hf_j) = f_j + h(f_t)_j + hf_j(f_y)_j + O(h^2). \quad (4.21)$$

Als (4.21) ingevuld wordt in (4.20) dan volgt

$$e_{j+1} = O(h^2).$$

Voor Euler achterwaarts gebruiken we de ontwikkeling

$$y_j = y_{j+1} - hy'_{j+1} + O(h^2). \quad (4.22)$$

Substitutie van (4.22) in (4.12) geeft

$$e_{j+1} = y'_{j+1} - f_{j+1} + O(h).$$

Omdat $y'_{j+1} = f_{j+1}$ volgt dat de afbreekfout van Euler achterwaarts van de orde 1 is. De orde van de afbreekfout van de Trapeziumregel is 2.

4.5 Stabiliteit

Een fysisch verschijnsel noemen we stabiel als een kleine verstoring van de parameters een klein verschil in de oplossing geeft. Voorbeelden van instabiele verschijnselen zijn: knik van een op druk belaste staaf of resonantie van een brug. In dit hoofdstuk gaan we er van uit dat het beginwaarde probleem een beschrijving is van een stabiele toepassing. We zullen nu beschouwen onder welke voorwaarden het beginwaarde probleem stabiel is. Als laatste onderzoeken we de stabiliteit van numerieke methoden.

Stabiliteit van een beginwaarde probleem

Stel dat de beginvoorwaarde een fout ε_0 bevat. De verstoorde oplossing $\tilde{y}$ voldoet aan

$$\tilde{y}' = f(t, \tilde{y}) \quad \text{en} \quad \tilde{y}(0) = y_0 + \varepsilon_0 .$$

De verschilfunctie noemen we ε : $\varepsilon(t) = \tilde{y}(t) - y(t)$. Een beginwaarde probleem heet stabiel als

$$\lim_{t \rightarrow \infty} |\varepsilon(t)| < \infty .$$

Als de limiet onbegrensd is spreken we van een instabiel beginwaarde probleem. Een beginwaarde probleem is absoluut stabiel als

$$\lim_{t \rightarrow \infty} |\varepsilon(t)| = 0 .$$

We beperken de stabiliteitsanalyse eerst tot de testvergelijking.

Stabiliteit voor de testvergelijking

Voor de testvergelijking hebben we

$$y' = \lambda y + g(t) \quad \text{en} \quad y(0) = y_0 \tag{4.23}$$

en

$$\tilde{y}' = \lambda \tilde{y} + g(t) \quad \text{en} \quad \tilde{y}(0) = y_0 + \varepsilon_0 .$$

Combinatie geeft dat ε voldoet aan

$$\varepsilon' = \lambda \varepsilon \quad \text{en} \quad \varepsilon(0) = \varepsilon_0 . \tag{4.24}$$

De oplossing van (4.24) is

$$\varepsilon(t) = \varepsilon_0 e^{\lambda t} .$$

De testvergelijking is stabiel als $\lambda \leq 0$ en absoluut stabiel als $\lambda < 0$.

Numerieke stabiliteit voor de testvergelijking

We beschouwen twee numerieke oplossingen van de testvergelijking (4.23): u_i met beginvoorwaarde $u_0 = y_0$ en v_i met beginvoorwaarde $v_0 = y_0 + \varepsilon_0$. Het blijkt dat voor alle van belang zijnde numerieke methoden het verschil van deze oplossingen $\varepsilon_i = v_i - u_i$ precies de numerieke oplossing is van (4.24). Het schema heet stabiel als

$$\lim_{j \rightarrow \infty} |\varepsilon_j| < \infty$$

en absoluut stabiel als

$$\lim_{j \rightarrow \infty} |\varepsilon_j| = 0 .$$

Uit (4.7) volgt $\varepsilon_{j+1} = k(h\lambda)\varepsilon_j$. Uit volledige inductie blijkt $\varepsilon_j = [k(h\lambda)]^j \varepsilon_0$. Een numeriek schema is dus stabiel dan en slechts dan als

$$|k(h\lambda)| \leq 1 . \tag{4.25}$$

In (4.25) is λ gegeven en de vraag is dus voor welke h is er aan ongelijkheid (4.25) voldaan.

Voor de voorwaartse methode van Euler geldt $k(h\lambda) = 1 + h\lambda$. Neem aan dat $\lambda \leq 0$ zodat het beginwaarde probleem stabiel is. Om aan (4.25) te volen moet h zo gekozen worden dat

$$-1 \leq 1 + h\lambda \leq 1 ,$$

ofwel

$$-2 \leq h\lambda \leq 0 .$$

Omdat $h > 0$ en $\lambda \leq 0$ volgt hieruit $h \leq \frac{2}{|\lambda|}$.

Voor Euler achterwaarts moet gelden

$$-1 \leq \frac{1}{1 - h\lambda} \leq 1 .$$

Eenvoudig is in te zien dat hier voor elke $h \geq 0$ aan voldaan is omdat $\lambda \leq 0$, zodat de achterwaartse methode van Euler onvoorwaardelijk stabiel is als $\lambda \leq 0$.

Eenzelfde analyse geeft dat voor $\lambda \leq 0$ de methode van Heun stabiel is als $h < \frac{2}{|\lambda|}$. De Trapeziumregel is ook onvoorwaardelijk stabiel als $\lambda \leq 0$.

Voorbeeld (numerieke stabiliteit)

Als voorbeeld nemen we het beginwaarde probleem:

$$\begin{aligned} y' &= -10y , & t \in [0, 1] , \\ y(0) &= 1 . \end{aligned}$$

De exacte oplossing wordt gegeven door $y(t) = e^{-10t}$. Uit de theorie volgt dat Euler voorwaarts stabiel is als $h \leq 0.2$. In Figuur 4.2 staat de numerieke oplossing voor de stapgroottes $h = \frac{1}{3}, \frac{1}{6}$ en $\frac{1}{12}$. De methode is inderdaad instabiel als $h = \frac{1}{3}$ immers $\varepsilon_j = \varepsilon_0 \cdot y_j$ zodat $|\varepsilon_j| > |\varepsilon_{j-1}|$.

Voor $h = \frac{1}{6}$ en $\frac{1}{12}$ blijkt dat de numerieke oplossing naar de exacte oplossing convergeert. De achterwaartse methode van Euler is stabiel voor alle stapgroottes (zie Figuur 4.3).

Numerieke stabiliteit voor een algemene vergelijking

Voor een algemeen beginwaarde probleem

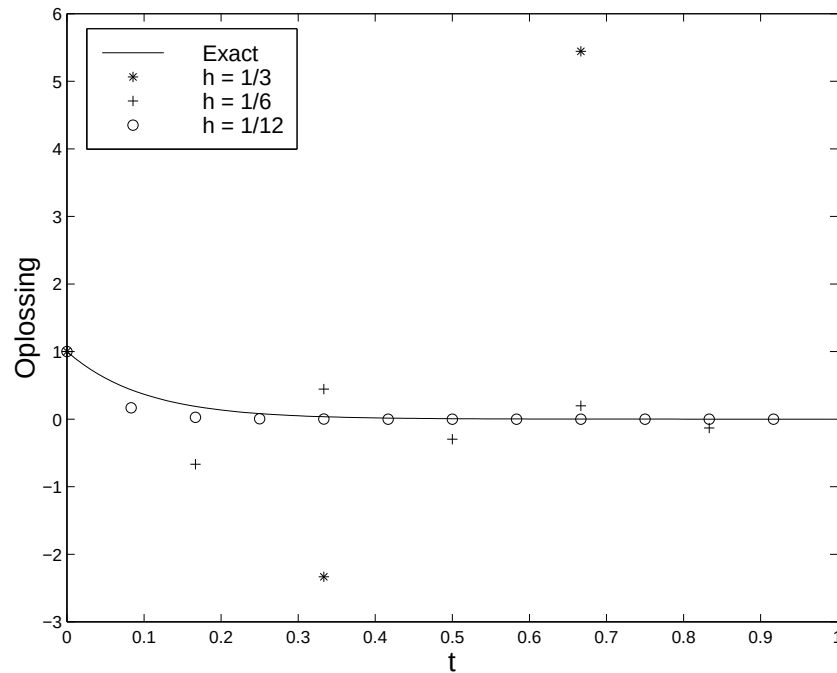
$$y' = f(t, y) \quad \text{en} \quad y(0) = y_0 ,$$

kunnen we een indruk krijgen van de stabiliteitseigenschappen door f te lineariseren rondom een gegeven punt, bijvoorbeeld (t_0, y_0) . Vervang $f(t, y)$ door de lineaire benadering

$$f(t_0, y_0) + (y - y_0) \frac{\partial f}{\partial y}(t_0, y_0) + (t - t_0) \frac{\partial f}{\partial t}(t_0, y_0) .$$

De differentiaalvergelijking gedraagt zich in de buurt van (t_0, y_0) als

$$y' = f(t_0, y_0) + (y - y_0) \frac{\partial f}{\partial y}(t_0, y_0) + (t - t_0) \frac{\partial f}{\partial t}(t_0, y_0) .$$



Figuur 4.2: Euler voorwaarts toegepast op $y' = -10y$

Als we dit vergelijken met de testvergelijking (4.23), dan blijkt dat we dezelfde analyse kunnen volgen, waarbij λ gelijk genomen wordt aan $\frac{\partial f}{\partial y}(t_0, y_0)$. Merk op dat de stabiliteitsvoorwaarde nu afhangt van y_0 en t_0 .

Voorbeeld (stabiliteit stuwmeer)

Het debiet Q dat uit het stuwmeer stroomt voldoet aan

$$\frac{dQ}{dt} = p(t) - aQ^2,$$

zodat $f(t, Q) = p(t) - aQ^2$. In de stabiliteitsanalyse gebruiken we

$$\frac{\partial f}{\partial Q}(t_0, Q_0) = -2aQ_0.$$

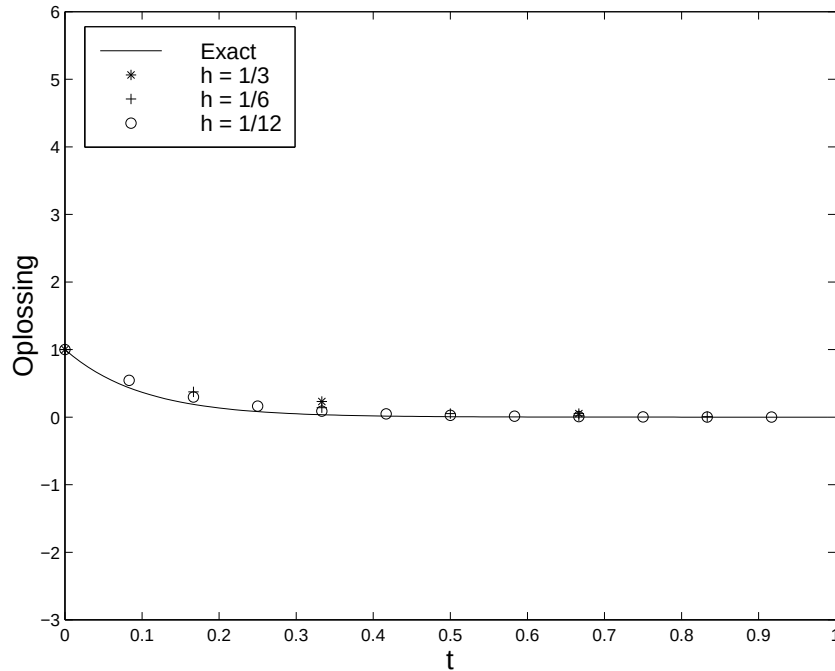
Merk op dat voor alle positieve a en Q de differentiaalvergelijking stabiel is. Euler voorwaarts is stabiel als

$$h \leq \frac{1}{aQ_0},$$

zodat de bovengrens voor h kleiner wordt als Q_0 toeneemt.

4.6 Convergentie

Bij het bepalen van een numerieke benadering is het van belang om te weten hoe groot de globale fout $E_j = y_j - u_j$ is.



Figuur 4.3: Euler achterwaarts toegepast op $y' = -10y$

Convergentie

Een schema heet convergent als voor de globale fout E_j geldt:

$$\lim_{h \rightarrow 0} E_j = 0, \quad \text{waarbij } jh \text{ constant is.}$$

We laten nu zien dat stabiliteit en consistentie opnieuw convergentie impliceren.

Stelling 4.6.1 *Als het numerieke schema stabiel en consistent is, dan convergeert de numerieke oplossing naar de exacte oplossing. Bovendien is de orde van de globale fout gelijk aan die van de afbreekfout.*

Bewijs:

We zullen het bewijs alleen geven voor de homogene testvergelijking $y' = \lambda y$. Voor de globale fout kan een recurrente betrekking worden afgeleid, namelijk

$$\begin{aligned} E_j &= y_j - u_j = y_j - k(h\lambda)u_{j-1} \\ &= y_j - k(h\lambda)y_{j-1} + k(h\lambda)(y_{j-1} - u_{j-1}) \\ &= he_j + k(h\lambda)E_{j-1}. \end{aligned} \tag{4.26}$$

Met inductie volgt uit (4.26) dat we E_j voor $e_0 = 0$ kunnen schrijven als

$$E_j = \sum_{i=0}^{j-1} [k(h\lambda)]^i he_{j-i}.$$

Uit de stabiliteit volgt nu

$$E_j \leq \sum_{i=0}^{j-1} |k(h\lambda)|^i h |e_{j-i}| \leq \sum_{i=0}^{j-1} h |e_{j-i}| \leq \max_{1 \leq i \leq j} |e_i|.$$

Hieruit blijkt dat de orde van de globale fout gelijk is aan die van de lokale fout. Merk op dat uit de consistentie volgt dat de methode convergent is. $\square$

Waarom hogere orde methoden?

Het gebruik van een hogere orde methode is vaak zinvol. We laten dit zien aan de hand van de methode van Euler en Heun. Neem aan dat de globale fout begrensd is door:

$$\begin{aligned} |E_j| &\leq h && \text{Euler voorwaarts ,} \\ |E_j| &\leq h^2 && \text{Heun .} \end{aligned}$$

Om de hoeveelheid werk te bepalen, tellen we het aantal functie evaluaties. Bij Euler voorwaarts is één functie evaluatie per tijdstap nodig en bij Heun twee. In de volgende tabel geven we de stapgrootte en aantal functie evaluaties, dat nodig is om een voorgeschreven precisie te halen. Uit de tabel blijkt dat een hogere orde methode de voorkeur heeft, als er een hoge

E_j	Euler voorwaarts		Heun	
	h	functie evaluaties	h	functie evaluaties
10^{-1}	10^{-1}	10	$1/3$	6
10^{-2}	10^{-2}	100	10^{-1}	20
10^{-4}	10^{-4}	10000	10^{-2}	200

Tabel 4.1: Kosten voor Euler voorwaarts en Heun voor een verschillende nauwkeurigheid

nauwkeurigheid vereist wordt. Deze observatie is algemeen geldig. Belangrijk is wel dat de exacte oplossing voldoende glad is.

4.7 Stelsels eerste orde beginwaarde problemen

Sommige toepassingen worden beschreven door een stelsel van gekoppelde eerste orde differentiaalvergelijkingen. Als voorbeeld nemen we de interactie van prooi- en roofdieren. Stel het aantal prooidieren geven we aan met $p(t)$ en het aantal roofdieren met $r(t)$. Een eenvoudig stelsel dat de groei of afname van de aantallen beschrijft is:

$$\begin{aligned} \frac{dp}{dt} &= (1 - p)p - pr , \\ \frac{dr}{dt} &= (p - 1)r , \end{aligned}$$

waarbij $p(0) = p_0$ en $r(0) = r_0$ gegeven zijn.

De algemene gedaante van zo'n stelsel noteren we als: $\mathbf{x}' = \mathbf{f}(t, \mathbf{x})$ en $\mathbf{x}(0) = \mathbf{x}_0$. De variable t is de onafhankelijk veranderlijke en x_i zijn de afhankelijk veranderlijken.

Numerieke methoden voor stelsels

Het is heel eenvoudig om een numerieke methode voor de vergelijking $y' = f(t, y)$, $y(0) = y_0$ uit te breiden naar stelsels. Als voorbeeld nemen we de methode van Heun

$$\begin{aligned} \mathbf{x}_{j+1}^* &= \mathbf{x}_j + h\mathbf{f}(t_j, \mathbf{x}_j) , \\ \mathbf{x}_{j+1} &= \mathbf{x}_j + \frac{h}{2}[\mathbf{f}(t_j, \mathbf{x}_j) + \mathbf{f}(t_{j+1}, \mathbf{x}_{j+1}^*)] . \end{aligned}$$

De resultaten ten aanzien van de afbreekfout blijken ook voor stelsels geldig.

Hogere orde beginwaarde problemen

Een hogere orde beginwaarde probleem wordt beschreven door de differentiaalvergelijking

$$y^{(n)} = f(t, y, y^{(1)}, \dots, y^{(n-1)}) , \quad (4.27)$$

met beginvoorwaarden

$$y(0) = y_0, \quad y'(0) = y'_0, \quad \dots, \quad y^{(n-1)}(0) = y_0^{(n-1)} ,$$

waarin $y^{(k)}$ een korte schrijfwijze is voor $\frac{d^k y}{dt^k}$.

Een manier om dit probleem op te lossen is de volgende transformatie:

$$\begin{aligned} x_1 &= y , \\ x_2 &= y^{(1)} , \\ &\vdots \\ x_n &= y^{(n-1)} . \end{aligned}$$

Hiermee gaat (4.27) over in het volgende eerste orde stelsel

$$\begin{aligned} x_1' &= x_2 , \\ &\vdots \\ x_{n-1}' &= x_n , \\ x_n' &= f(t, x_1, \dots, x_n) , \end{aligned}$$

met beginvoorwaarden $x_1(0) = y_0, \dots, x_n(0) = y_0^{(n-1)}$.

Voorbeeld (mathematische slinger)

De vergelijking van de mathematische slinger wordt gegeven door

$$\psi'' + \sin \psi = 0 ,$$

met beginvoorwaarden $\psi(0) = \psi_0$ en $\psi'(0) = 0$. Het bijbehorende eerste orde stelsel volgt uit $x_1 = \psi$ en $x_2 = \psi'$:

$$\begin{aligned} x_1' &= x_2 , \\ x_2' &= -\sin x_1 , \end{aligned}$$

met beginvoorwaarden $x_1(0) = \psi_0$ en $x_2(0) = 0$. Als we de voorwaartse methode van Euler gebruiken dan geldt:

$$\begin{aligned} x_1^{j+1} &= x_1^j + h x_2^j , \\ x_2^{j+1} &= x_2^j - h \sin x_1^j . \end{aligned}$$

4.8 Stabiliteit voor stelsels

We zullen eerst de stabiliteit beschouwen voor stelsels differentiaal vergelijkingen en daarna de stabiliteit van de gebruikte numerieke methoden.

Stabiliteit voor een stelsel met constante coëfficiënten

We beperken ons eerst tot het stelsel

$$\mathbf{x}' = A\mathbf{x} + \mathbf{g}(t) \quad \text{met} \quad \mathbf{x}(0) = \mathbf{x}_0, \quad (4.28)$$

waarbij A een reële, $n \times n$ matrix is met constante coëfficiënten. We beschouwen weer twee oplossingen: $\mathbf{x}$ met beginvoorwaarde $\mathbf{x}(0) = \mathbf{x}_0$ en $\tilde{\mathbf{x}}$ met beginvoorwaarde $\tilde{\mathbf{x}}(0) = \mathbf{x}_0 + \boldsymbol{\varepsilon}_0$. Voor het verschil geldt:

$$\boldsymbol{\varepsilon}' = A\boldsymbol{\varepsilon} \quad \text{met} \quad \boldsymbol{\varepsilon}(0) = \boldsymbol{\varepsilon}_0. \quad (4.29)$$

Dit is een gekoppeld stelsel lineaire differentiaalvergelijkingen. We proberen eerst een transformatie te vinden, die de grootheden ontkoppelt. Daarvoor nemen we aan dat A diagonaliseerbaar is, dat wil zeggen er is een niet-singuliere matrix S zodanig dat

$$AS = S\Lambda, \quad (4.30)$$

waarin $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ de eigenwaarden van A bevat.

Als A een complex eigenpaar $\lambda, \mathbf{v}$ heeft, dan geldt dat $\bar{\lambda}, \bar{\mathbf{v}}$ ook een eigenpaar is. Stel nu $\boldsymbol{\varepsilon} = S\boldsymbol{\eta}$ dan kunnen we (4.29) schrijven als:

$$\begin{aligned} \boldsymbol{\eta}' &= \Lambda\boldsymbol{\eta}, \\ \boldsymbol{\eta}(0) &= \boldsymbol{\eta}_0 = S^{-1}\boldsymbol{\varepsilon}_0. \end{aligned} \quad (4.31)$$

Het stelsel (4.31) is ontkoppeld en eenvoudig op te lossen

$$\boldsymbol{\eta} = e^{\Lambda t} \boldsymbol{\eta}_0.$$

De matrix $e^{\Lambda t}$ is een diagonaalmatrix met als diagonaalelementen $e^{\lambda_1 t}, \dots, e^{\lambda_n t}$. Hieruit volgt

$$\boldsymbol{\varepsilon} = Se^{\Lambda t} S^{-1} \boldsymbol{\varepsilon}_0. \quad (4.32)$$

Als alle eigenwaarden reëel zijn dan is (4.28) een stabiel stelsel als $\lambda_i \leq 0$. Als λ_i complex is dan kunnen we de oplossing schrijven als een oscillerende functie van t , waarbij de amplitude gegeven wordt door $e^{\text{Re}\lambda_i t}$. Dus in het algemeen geldt dat het stelsel (4.28) stabiel is als $\text{Re}\lambda_i \leq 0$ voor $1 \leq i \leq n$.

Stabiliteit van numerieke methoden

We beschouwen de numerieke oplossing van het stelsel differentiaalvergelijkingen

$$\mathbf{x}' = A\mathbf{x} \quad \text{met} \quad \mathbf{x}(0) = \mathbf{x}_0.$$

De numerieke oplossing $\mathbf{u}$ wordt gekarakteriseerd door

$$\mathbf{u}_{j+1} = G(hA)\mathbf{u}_j \quad \text{met} \quad \mathbf{u}_0 = \mathbf{x}_0.$$

De $n \times n$ matrix $G(hA)$ noemen we de versterkingsmatrix. De volgende uitdrukkingen zijn eenvoudig te verifiëren

Euler voorwaarts	$G(hA) = I + hA$,
Euler achterwaarts	$G(hA) = (I - hA)^{-1}$,
Trapeziumregel	$G(hA) = (I - \frac{1}{2}hA)^{-1}(I + \frac{1}{2}hA)$,
Heun	$G(hA) = I + hA + \frac{1}{2}h^2 A^2$.

Stelling 4.8.1 *Zij A een diagonaliseerbare $n \times n$ matrix en laat S de niet-singuliere matrix zijn zodanig dat*

$$S^{-1}AS = \text{diag}(\lambda_1, \dots, \lambda_n) .$$

Zij $G(hA)$ de versterkingsmatrix van een numerieke methode, dan geldt

$$G(hA) \text{ is diagonaliseerbaar en } S^{-1}G(hA)S = M ,$$

waarbij $M = \text{diag}(k(h\lambda_1), \dots, k(h\lambda_n))$, waarin $k(h\lambda)$ de versterkingsfactor is van de numerieke methode.

Bewijs:

Het bewijs volgt door uitschrijven. □

Voor de stabiliteit van een numerieke methode toegepast op (4.28) beschouwen we het gedrag van de numerieke oplossing van

$$\epsilon' = A\epsilon , \quad \epsilon(0) = \epsilon_0 . \tag{4.33}$$

Uit bovenstaande volgt $\epsilon_{j+1} = G(hA)\epsilon_j$. Voor stabiliteit moet ϵ_j begrensd blijven. Met de substitutie $\epsilon_j = S\eta_j$ volgt

$$\eta_{j+1} = M\eta_j .$$

Voor de stabiliteit moet gelden $|k(h\lambda_i)| \leq 1$ voor $1 \leq i \leq n$.

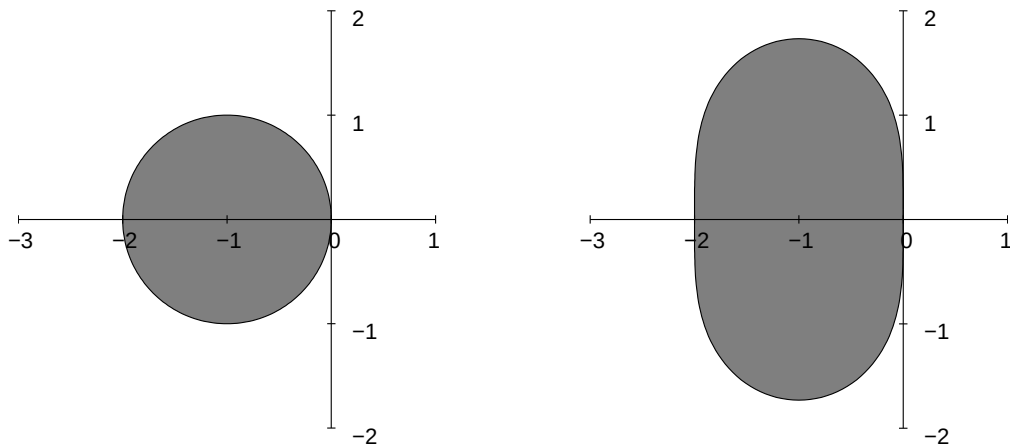
Stabiliteitsgebied

Voor complexe eigenwaarden van A kan er handig gebruik gemaakt worden van stabiliteitsgebieden. Het stabiliteitsgebied is de verzameling van punten z in het complexe vlak waarvoor geldt $|k(z)| \leq 1$.

De stabiliteitsgebieden van de expliciete methoden staan in Figuur 4.4 en die van de impliciete methoden in Figuur 4.5. Een stabiliteitsgebied kan als volgt gebruikt worden: stel λ_i is een complexe eigenwaarde, dan moet h zo klein genomen worden dat $h\lambda_i$ binnen het stabiliteitsgebied valt. Merk op dat de expliciete methoden altijd instabiel zijn als $\text{Re}\lambda_i = 0$ (zuiver imaginaire eigenwaarden).

Euler voorwaarts

Heun



Figuur 4.4: Stabiliteitsgebieden van de expliciete methoden

Stabiliteit van een algemeen stelsel

Voor stelsels van de vorm

$$\mathbf{x}' = \mathbf{f}(t, \mathbf{x}), \quad \mathbf{x}(0) = \mathbf{x}_0,$$

gelden lokaal dezelfde eigenschappen als voor stelsels van het type (4.28). De rol van A wordt dan ingenomen door de matrix van partiële afgeleiden van $\mathbf{f}$ gegeven door

$$\begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}.$$

Voorbeeld (stabiliteit mathematische slinger)

We beschouwen de stabiliteit van de numerieke oplossing van de vergelijking van de mathematische slinger:

$$\psi'' + \sin \psi = 0, \quad \psi(0) = \psi_0 \quad \text{en} \quad \psi'(0) = 0.$$

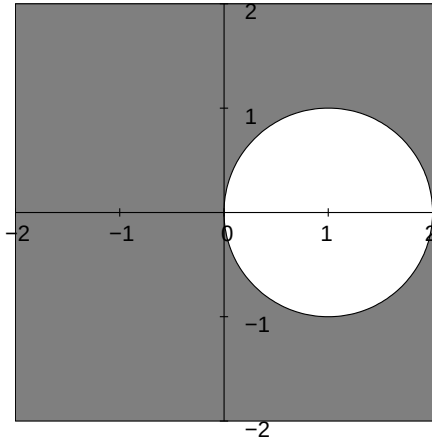
We kunnen dit schrijven als het stelsel

$$\begin{aligned} x_1' &= x_2 = f_1(t, x_1, x_2), \\ x_2' &= -\sin x_1 = f_2(t, x_1, x_2). \end{aligned}$$

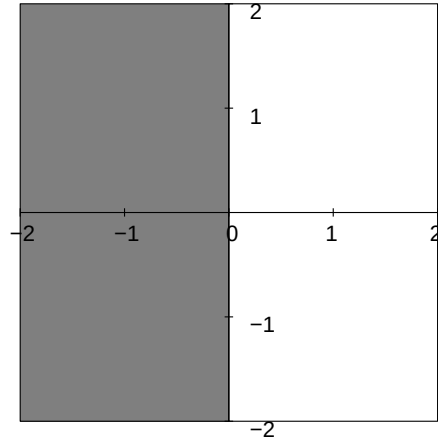
De matrix van afgeleiden is nu

$$\begin{bmatrix} 0 & 1 \\ -\cos x_1 & 0 \end{bmatrix}.$$

Euler achterwaarts



Trapeziumregel



Figuur 4.5: Stabiliteitsgebied van de impliciete methoden

Als we aannemen dat $-\frac{\pi}{2} < \psi_0 < \frac{\pi}{2}$ dan zijn de eigenwaarden

$$\lambda_{1,2} = \pm i\sqrt{\cos x_1}.$$

Merk op dat de voorwaartse methode van Euler en de methode van Heun instabiel zijn voor elke h . De Trapeziumregel is stabiel en de achterwaartse methode van Euler leidt tot demping van de slingerbeweging. In een volgende paragraaf zullen we een expliciete methode behandelen, die wel stabiel is voor stelsels met imaginaire eigenwaarden.

4.9 Stijve stelsels differentiaalvergelijkingen

We beperken ons in deze paragraaf tot stelsels van de vorm

$$\mathbf{y}' = A\mathbf{y}, \quad (4.34)$$

waarbij de eigenwaarden van A reëel zijn en genoteerd worden als

$$\lambda_n \leq \lambda_{n-1} \dots \leq \lambda_1 < 0.$$

Stijve stelsels differentiaalvergelijkingen

Het stelsel (4.34) wordt stijf genoemd als

$$\frac{|\lambda_n|}{|\lambda_1|} \gg 1.$$

De oplossing van (4.34) wordt gegeven door $\mathbf{y} = \sum_{k=1}^n c_k e^{\lambda_k t} \mathbf{v}_k$, waarbij $\mathbf{v}_k$ de eigenvector is behorende bij λ_k . Merk op dat componenten, die corresponderen met sterke negatieve λ 's snel

uitdempen en slechts bij het inschakelen (voor kleine t) van belang zijn. Deze componenten worden snelle componenten genoemd. Meestal is men niet geïnteresseerd in inschakelverschijnselen en dus ook niet in de snelle componenten.

Expliciete methoden

Expliciete methoden zijn voorwaardelijk stabiel, dat wil zeggen alleen stabiel als h voldoende klein gekozen wordt. Als voorbeeld nemen we Euler voorwaarts. De stabiliteitsvoorwaarde wordt gegeven door

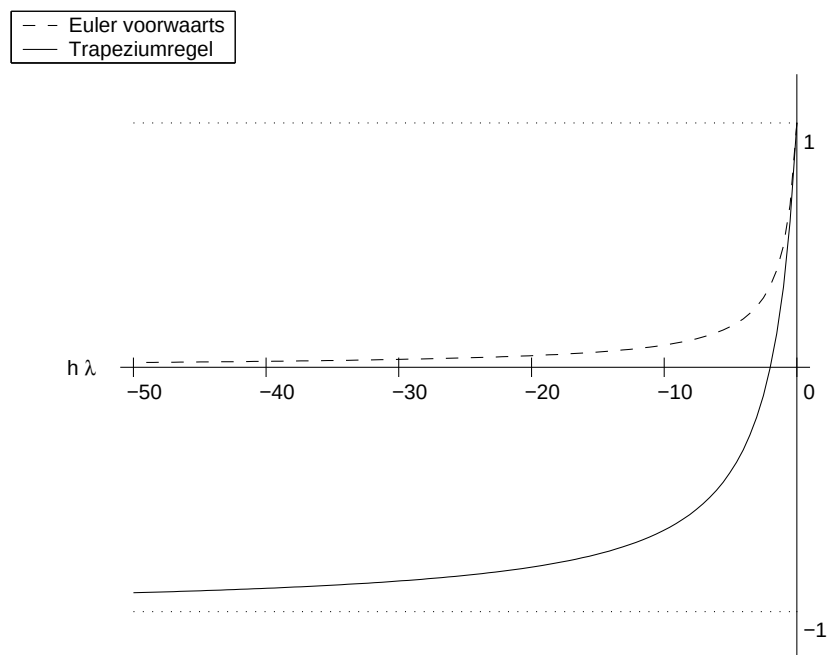
$$h \leq \frac{2}{|\lambda_n|}.$$

De nauwkeurigheid van de oplossing wordt voornamelijk bepaald door de langzame componenten. Op grond hiervan zou met een grotere stapgrootte volstaan kunnen worden. Conclusie bij stijve stelsels differentiaalvergelijkingen resulteert de stabiliteitseis voor expliciete methoden in waarden van h , die onpraktisch klein zijn.

Impliciete methoden

Impliciete methoden kunnen onvoorwaardelijk stabiel zijn. Men kan de stapgrootte dan afstemmen op de gewenste nauwkeurigheid van de langzame componenten. De nauwkeurigheid ten opzichte van de snelle componenten is dan wel gering, maar omdat de integratie stabiel is heeft dit geen grote gevolgen.

We vergelijken het gedrag van Euler achterwaarts en de Trapeziumregel ten opzichte van de snelle componenten. De versterkingsfactoren van beide methoden zijn gegeven in Figuur 4.6. Er blijkt een belangrijk verschil te zijn voor het gedrag voor $h\lambda \rightarrow -\infty$.



Figuur 4.6: Versterkingsfactoren van Euler voorwaarts en de Trapeziumregel

Superstabil

Een numerieke methode heet superstabil als

$$\lim_{h\lambda \rightarrow -\infty} |k(h\lambda)| < 1 .$$

Uit Figuur 4.6 blijkt dat Euler achterwaarts superstabil is, terwijl voor de Trapeziumregel geldt:

$$\lim_{h\lambda \rightarrow -\infty} |k(h\lambda)| = 1 .$$

Dit betekent dat beginverstoringen in de snelle componenten niet of zeer traag uitdempen bij het gebruik van de Trapeziumregel.

Discussie

In deze paragraaf hebben we ons beperkt tot een eenvoudig stelsel differentiaalvergelijkingen (4.34). Het zal duidelijk zijn dat ook bij ingewikkelder stelsels sprake van stijfheid kan zijn. In dergelijke gevallen zijn impliciete methoden aan te raden. Bij toepassing van impliciete methoden moet in elke stap een stelsel vergelijkingen opgelost worden. Meestal is dit stelsel niet-lineair. Hoewel er zeer goede numerieke methoden beschikbaar zijn voor het oplossen van stelsels vergelijkingen, kunnen de kosten toch sterk oplopen. Het is daarom vaak een kwestie van wikken en wegen, waarbij expliciete methoden niet op voorhand zijn uit te sluiten.

4.10 Runge-Kutta methoden

Een belangrijke klasse van eenstaps methoden zijn de Runge-Kutta methoden. Een eenvoudig voorbeeld van de Runge-Kutta methode is de methode van Heun. Deze is beschreven in (4.5). In een nieuwe notatie luidt de methode

$$\begin{cases} k_1 = hf(t_j, u_j) , \\ k_2 = hf(t_{j+1}, u_j + k_1) , \\ u_{j+1} = u_j + \frac{1}{2}(k_1 + k_2) . \end{cases} \quad (4.35)$$

De methode RK₄

We volstaan met het geven van een veel gebruikte Runge-Kutta methode, te weten de methode RK₄. Deze luidt

$$\begin{cases} k_1 = hf(t_j, u_j) , \\ k_2 = hf(t_{j+\frac{1}{2}}, u_j + \frac{1}{2}k_1) , \\ k_3 = hf(t_{j+\frac{1}{2}}, u_j + \frac{1}{2}k_2) , \\ k_4 = hf(t_{j+1}, u_j + k_3) , \\ u_{j+1} = u_j + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) . \end{cases} \quad (4.36)$$

De methode is expliciet. De versterkingsfactor volgt na substitutie van $f(t, y) = \lambda y$ en luidt

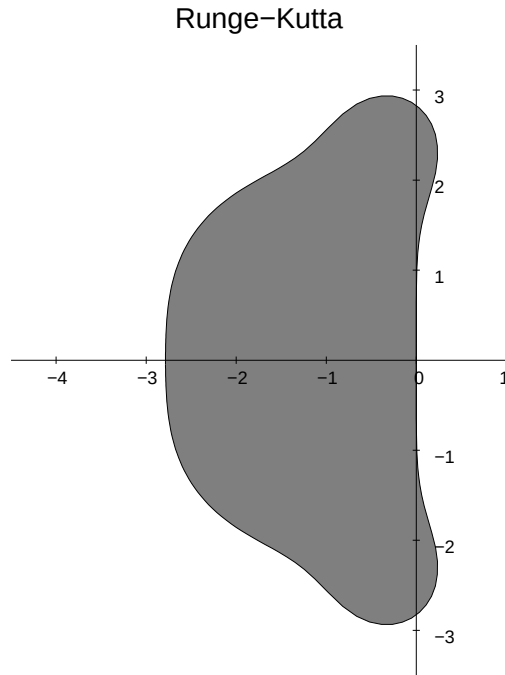
$$k(h\lambda) = 1 + h\lambda + \frac{(h\lambda)^2}{2} + \frac{(h\lambda)^3}{3!} + \frac{(h\lambda)^4}{4!} . \quad (4.37)$$

Eenvoudig is na te gaan dat

$$\frac{e^{h\lambda} - k(h\lambda)}{h} = O(h^4) ,$$

en daarmee is de orde van de lokale fout gelijk aan 4. Ook voor willekeurige functies f is dit het geval, maar het bewijs hiervan vergt veel rekenen en wordt hier achterwege gelaten.

Het stabiliteitsgebied van de RK₄ methode staat in Figuur 4.7. We zien dat de methode



Figuur 4.7: Stabiliteitsgebied van de Runge-Kutta methode RK₄

voorwaardelijk stabiel is. Tevens merken we op dat de methode ook stabiel kan zijn in het geval van $\mathbf{y}' = A\mathbf{y}$ met een matrix A , die zuiver imaginaire eigenwaarden heeft. De eerder door ons behandelde expliciete methoden zijn altijd instabiel voor een dergelijk stelsel.

4.11 Meerstaps methoden

De methoden, die we tot nu toe behandeld hebben, worden eenstaps methoden genoemd, omdat de benadering in t_{j+1} alleen afhangt van informatie van het vorige tijdstip t_j . Hoewel sommige methoden gebruik maken van functie evaluaties in tussengelegen punten, wordt deze informatie niet bewaard om te gebruiken in toekomstige benaderingen. Alle informatie gebruikt door deze methoden is verkregen binnen het interval waarover de oplossing benaderd wordt.

Omdat de benaderende oplossing beschikbaar is in de roosterpunten $t_0, t_1, \dots, t_j$ voordat de benadering in t_{j+1} bepaald wordt, lijkt het redelijk om methoden te ontwikkelen, die deze informatie gebruiken. Deze methoden worden meerstaps methoden genoemd.

Adams-Bashforth methode

Voor de afleiding van de Adams-Bashforth methode gaan we uit van de Trapeziumregel

$$u_{j+1} = u_j + \frac{h}{2}[f(t_j, u_j) + f(t_{j+1}, u_{j+1})] \quad (4.38)$$

We kunnen deze methode expliciet maken door $f(t_{j+1}, u_{j+1})$ te extrapoleren met behulp van $f(t_{j-1}, u_{j-1})$ en $f(t_j, u_j)$. Het lineaire interpolatiepolynoom wordt gegeven door

$$l(t) = f(t_j, u_j) + (t - t_j) \frac{f(t_j, u_j) - f(t_{j-1}, u_{j-1})}{h} .$$

Als we nu $f(t_{j+1}, u_{j+1})$ in (4.38) vervangen door $l(t_{j+1})$, dan krijgen we de Adams-Bashforth methode

$$u_{j+1} = u_j + \frac{3}{2}hf(t_j, u_j) - \frac{1}{2}hf(t_{j-1}, u_{j-1}) . \quad (4.39)$$

Merk op dat per tijdstap slechts één functie evaluatie nodig is.

Stabiliteit

We maken gebruik van de testvergelijking $y' = \lambda y$. Voor de verstoring ε geldt de recursie

$$\varepsilon_{j+1} = (1 + \frac{3}{2}h\lambda)\varepsilon_j - \frac{1}{2}h\lambda\varepsilon_{j-1} . \quad (4.40)$$

Dit is een differentievergelijking met constante coëfficiënten. Deze kunnen we oplossen door te substitueren $\varepsilon_j = cz^j$. De karakteristieke vergelijking wordt

$$z^2 - (1 + \frac{3}{2}h\lambda)z + \frac{1}{2}h\lambda = 0 ,$$

met als oplossingen

$$z_1 = \frac{1 + \frac{3}{2}h\lambda + \sqrt{D}}{2} , \quad (4.41)$$

$$z_2 = \frac{1 + \frac{3}{2}h\lambda - \sqrt{D}}{2} , \quad (4.42)$$

waarbij $D = 1 + h\lambda + \frac{9}{4}(h\lambda)^2$. De algemene oplossing van (4.40) is nu

$$\varepsilon_j = c_1 z_1^j + c_2 z_2^j .$$

De constanten c_1 en c_2 volgen uit de beginvoorwaarden. Merk op dat de Adams-Bashforth methode stabiel is wanneer

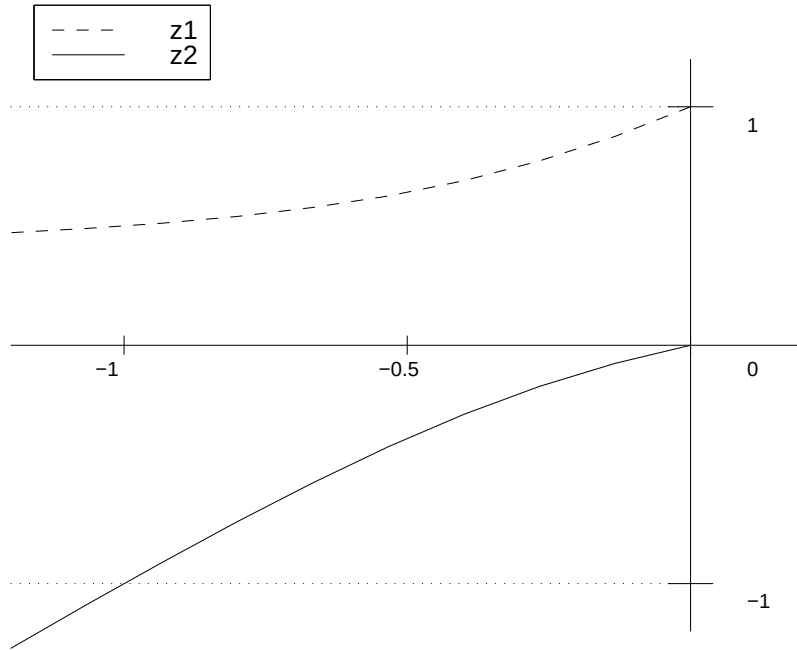
$$|z_{1,2}(h\lambda)| \leq 1 . \quad (4.43)$$

De wortels z_1 en z_2 zijn gegeven in Figuur 4.8. Merk op dat aan (4.43) voldaan is voor $h \leq \frac{1}{|\lambda|}$. De methode is dus voorwaardelijk stabiel.

Afbreukfout

Met behulp van Taylorontwikkeling volgt uit (4.41) en (4.43):

$$\begin{aligned} z_1(h\lambda) &= 1 + h\lambda + \frac{1}{2}(h\lambda)^2 - \frac{1}{4}(h\lambda)^3 + O(h^4) , \\ z_2(h\lambda) &= \frac{1}{2}h\lambda + O(h^2) . \end{aligned}$$



Figuur 4.8: De versterkingsfactoren z_1 en z_2 van de Adams-Bashforth methode

De exacte oplossing van de testvergelijking is

$$e^{h\lambda} = 1 + h\lambda + \frac{1}{2}(h\lambda)^2 + \frac{1}{6}(h\lambda)^3 + O(h^4) .$$

Merk op dat $\frac{|z_1(h\lambda) - e^{h\lambda}|}{h} \approx \frac{5}{12}h^2|\lambda|^3$.

Dus $z_1(h\lambda)$ is een goede benadering van $e^{h\lambda}$ en wordt daarom de 'principal root' genoemd. Daarentegen heet $z_2(h\lambda)$ de 'spurious root'. Deze heeft niets met de differentiaalvergelijking te maken, maar is een gevolg van de gekozen numerieke methode.

Vergelijking met de methode van Heun

Wat betreft stabiliteit moet de stapgrootte bij Adams-Bashforth 2 maal zo klein genomen worden dan bij de methode van Heun. Omdat Heun per tijdstap 2 maal zoveel werk kost zijn de methoden op dit punt vergelijkbaar.

Wat betreft nauwkeurigheid kiezen we de stapgrootte zo dat de afbreekfout kleiner is dan ϵ . Voor Adams-Bashforth geldt dan

$$h = \sqrt{\frac{12\epsilon}{5|\lambda|^3}}$$

en voor Heun

$$h = \sqrt{\frac{6\epsilon}{|\lambda|^3}} .$$

We kunnen met dezelfde hoeveelheid werk 2 stappen Adams-Bashforth doen ten opzichte van 1 stap Heun. Hiermee wordt het interval waarover we de oplossing benaderen bij Adams-Bashforth

$$\sqrt{\frac{96\epsilon ps}{10|\lambda|^3}},$$

en bij Heun

$$\sqrt{\frac{60\epsilon ps}{10|\lambda|^3}},$$

zodat Adams-Bashforth minder werk kost dan Heun.

Discussie

Meerstaps methoden zijn minder populair dan Runge-Kutta methoden. Bij meestaps methoden treden opstartproblemen op, omdat de oplossing op meerdere tijdsniveaus bekend dient te zijn. Verder is de controle van de 'spurious roots' een moeilijke zaak.

4.12 Samenvatting

In dit hoofdstuk zijn de volgende begrippen behandeld:

- eenstapsmethoden (Euler, Heun, Trapezium)
- impliciete, expliciete methoden
- versterkingsfactor
- afbreekfout, consistentie
- stabiliteit, convergentie
- stelsels eerste orde beginwaarde problemen
- stabiliteit stelsels, stabiliteitsgebied
- stijve stelsels, superstabil
- Runge-Kutta methoden
- meestapsmethode (Adams-Bashforth)

Hoofdstuk 5

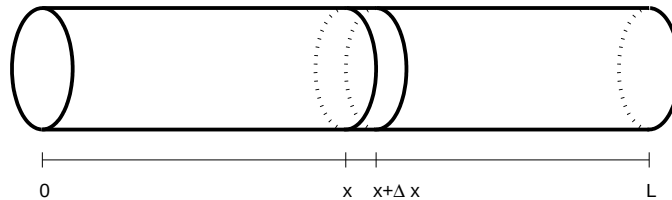
De eindige differentie methode voor randwaarde problemen

5.1 Inleiding

Bij het simuleren van praktische problemen wordt vaak gebruik gemaakt van randwaarde problemen. Een eendimensionaal randwaarde probleem is een differentiaalvergelijking op een begreind gebied, waarbij de functie en/of zijn afgeleiden gegeven zijn in beide randpunten.

Stationair warmte probleem

Als voorbeeld nemen we de temperatuurverdeling in een staaf (Figuur 5.1). De temperatuur



Figuur 5.1: De vorm van de staaf

van de staaf geven we aan met $T(x)$ ($^{\circ}\text{K}$). We nemen aan dat de temperatuur gegeven is in beide uiteinden: $T(0) = T_0$ en $T(L) = T_L$. Verder wordt er in de staaf warmte geproduceerd. Deze warmteproductie geven we aan met $Q(x)$ (J/m^3s). Bij dit voorbeeld zijn we geïnteresseerd in de temperatuur na lange tijd. Dit betekent dat de temperatuur stationair geworden is en dus alleen een functie is van x . Voor het afleiden van de differentiaalvergelijking maken we gebruik van de wet van behoud van energie toegepast op het controle volume tussen x en $x + \Delta x$ (zie Figuur 5.1).

Door geleiding is er warmtetransport door de zijvlakken op x en $x + \Delta x$. Volgens de wet van Fourier wordt dit warmtetransport per oppervlak en per seconde gegeven door:

$$-\lambda \frac{dT}{dx},$$

waarbij λ (J/msK) de warmtegeleidingscoëfficiënt genoemd wordt. De warmtebalans is nu:

$$-\lambda A \frac{dT}{dx}(x) + \lambda A \frac{dT}{dx}(x + \Delta x) + A Q(x) \Delta x = 0 .$$

Delen door $A \Delta x$ geeft

$$-\lambda \frac{\frac{dT}{dx}(x + \Delta x) - \frac{dT}{dx}(x)}{\Delta x} + Q(x) = 0 .$$

Na het nemen van de limiet $\Delta x \rightarrow 0$ krijgen we het volgende randwaarde probleem:

$$\begin{aligned} -\lambda \frac{d^2 T}{dx^2}(x) &= Q(x) \quad 0 < x < L , \\ T(0) &= T_0 \quad \text{en} \quad T(L) = T_L . \end{aligned}$$

Dit voorbeeld zal vaak gebruikt worden ter illustratie van de eindige differentie methode.

In een aantal toepassingen is de warmtestroom aan één van de uiteinden gegeven in plaats van de temperatuur. Voor $x = L$ leidt dit via de wet van Fourier tot de randvoorwaarde

$$\lambda A \frac{dT}{dx}(L) = A q_L ,$$

waarbij de warmtestroom q_L ($J/m^2 s$) gegeven is.

5.2 De eindige differentie methode

De algemene vorm van een randwaarde probleem van de tweede orde in één dimensie wordt gegeven door

$$-(p(x)y')' + r(x)y' + q(x)y = f(x) , \quad 0 < x < L ,$$

met de randvoorwaarden

$$a_0 y(0) + b_0 y'(0) = c_0 \quad \text{en} \quad a_L y(L) + b_L y'(L) = c_L .$$

Hierbij is aangenomen dat $p(x) > 0$ en $q(x) \geq 0$ voor alle $x \in [0, L]$. Deze problemen worden Sturm-Liouville problemen genoemd. Merk op dat het gegeven randwaarde probleem meerdere oplossingen heeft als $a_0 = a_L = 0$. Voor de randvoorwaarden worden de volgende aanduidingen gebruikt (we beperken ons tot $x = 0$):

$$\begin{aligned} \text{Dirichlet randvoorwaarde:} \quad & a_0 y(0) = c_0 \quad \text{dus} \quad b_0 = 0 , \\ \text{Neumann randvoorwaarde:} \quad & b_0 y'(0) = c_0 \quad \text{dus} \quad a_0 = 0 , \\ \text{Robbins randvoorwaarde:} \quad & a_0 \neq 0 \quad \text{en} \quad b_0 \neq 0 . \end{aligned}$$

De vergelijking

We beschrijven de eindige differentie methode voor het probleem:

$$\begin{aligned} -y'' + q(x)y &= f(x) , \quad 0 < x < 1 , \\ y(0) &= 0 \quad \text{en} \quad y(1) = 0 . \end{aligned}$$

Discretisatie

We delen het interval $[0, 1]$ op in n gelijke delen met lengte h . De knooppunten zijn $x_j = jh$ voor $j = 0, \dots, n$. de numerieke benadering van $y_j = y(x_j)$ geven we aan met u_j . Om de numerieke oplossing te bepalen wordt gebruik gemaakt van de differentiaalvergelijking in het punt x_j :

$$-y_j'' + q_j y_j = f_j \quad 1 \leq j \leq n-1. \quad (5.1)$$

De tweede afgeleide in (5.1) benaderen we met een centrale differentie zodat geldt:

$$-\frac{u_{j-1} - 2u_j + u_{j+1}}{h^2} + q_j u_j = f_j \quad 1 \leq j \leq n-1. \quad (5.2)$$

De waarde u_0 en u_n in de randpunten volgen uit de randvoorwaarden

$$u_0 = 0 \quad \text{en} \quad u_n = 0. \quad (5.3)$$

Het stelsel vergelijkingen

Het differentieschema (5.2) en (5.3) geeft $n+1$ vergelijkingen voor de onbekenden $u_0, \dots, u_n$. Als we u_0 en u_n elimineren uit (5.2) dan krijgen we het stelsel

$$A\mathbf{u} = \mathbf{f}, \quad (5.4)$$

met $A = K + M$, waarbij K en M gegeven worden door:

$$K = \frac{1}{h^2} \begin{bmatrix} 2 & -1 & & \emptyset \\ -1 & 2 & -1 & \\ & -1 & 2 & -1 \\ \emptyset & & \ddots & \ddots & \ddots \end{bmatrix} \quad \text{en} \quad M = \begin{bmatrix} q_1 & & \emptyset \\ & \ddots & \\ & \emptyset & \ddots \\ & & & q_{n-1} \end{bmatrix}.$$

$$\text{De vectoren } \mathbf{u} \text{ en } \mathbf{f} \text{ zijn: } \mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_{n-1} \end{pmatrix} \text{ en } \mathbf{f} = \begin{pmatrix} f_1 \\ \vdots \\ f_{n-1} \end{pmatrix}.$$

5.3 Herhaling van enkele lineaire algebra begrippen

In deze paragraaf worden een aantal begrippen uit de lineaire algebra opgesomd, die gebruikt worden bij het bestuderen van het verschil tussen de numerieke u_j en de exakte oplossing y_j van het randwaarde probleem.

Conditie

De geschaalde Euclidische norm van een vector is gedefinieerd als:

$$\|\mathbf{x}\| = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2}.$$

We maken veel gebruik van de ongelijkheid $\|A\mathbf{x}\| \leq \|A\| \|\mathbf{x}\|$.

Stel dat we geïnteresseerd zijn in de vector $\mathbf{x}$, die voldoet aan $A\mathbf{x} = \mathbf{b}$. Als het rechterlid $\mathbf{b}$

verstoord wordt met een fout $\Delta \mathbf{b}$, dan zal tengevolge daarvan in de oplossing $\mathbf{x}$ een fout $\Delta \mathbf{x}$ optreden, zodat we in feite oplossen

$$A(\mathbf{x} + \Delta \mathbf{x}) = \mathbf{b} + \Delta \mathbf{b} .$$

Voor de relatieve fout geldt dan

$$\frac{\|\Delta \mathbf{x}\|}{\|\mathbf{x}\|} \leq \|A\| \|A^{-1}\| \frac{\|\Delta \mathbf{b}\|}{\|\mathbf{b}\|} .$$

De grootheid $K(A) = \|A\| \|A^{-1}\|$ wordt het conditiegetal genoemd. Voor een symmetrische matrix met eigenwaarden $\lambda_1, \dots, \lambda_n$ geldt

$$\|A\| = \lambda_{\max} = \max_{1 \leq j \leq n} |\lambda_j| \quad \text{en} \quad \|A^{-1}\| = \frac{1}{\lambda_{\min}} = \frac{1}{\min_{1 \leq j \leq n} |\lambda_j|} ,$$

zodat $K(A) = \frac{\lambda_{\max}}{\lambda_{\min}}$. Voor een schatting van het conditiegetal van een symmetrische matrix, is het van belang om de grootste en de kleinste eigenwaarden te kennen. Een handige stelling hierbij is:

De stelling van Gershgorin

De eigenwaarden van de matrix A liggen in de vereniging van de cirkels

$$|z - a_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| \quad \text{waarbij} \quad z \in \mathbb{C} .$$

5.4 Consistentie, stabiliteit en convergentie

We willen laten zien dat het verschil tussen de numerieke en de exakte oplossing naar nul gaat als de stapgrootte h naar nul gaat. Om dit aan te tonen zullen we een aantal begrippen definiëren.

Afbreekfout

De afbreekfout $\boldsymbol{\varepsilon}$ van het schema $A\mathbf{u} = \mathbf{f}$ is gedefinieerd als

$$\varepsilon_j = (A\mathbf{y} - \mathbf{f})_j , \quad j = 1, \dots, n-1 ,$$

waarbij de componenten van $\mathbf{y}$ de exacte oplossing, gegeven worden door $y_j = y(x_j)$.

Voor het stelsel gegeven door (5.4) $\frac{-y_{j-1} + 2y_j - y_{j+1}}{h^2} + q_j y_j = f_j$ geldt met de foutschatting voor de centrale differentie

$$\varepsilon_j = -y_j'' + q_j y_j - f_j + O(h^2) .$$

Gecombineerd met (5.1) volgt dan:

$$\varepsilon_j = O(h^2) . \tag{5.5}$$

De orde van de afbreekfout is gelijk aan 2 voor dit schema.

Consistentie

Een eindige differentie schema heet consistent als

$$\lim_{h \rightarrow 0} \|\epsilon\| = 0 .$$

Uit (5.5) volgt $\|\epsilon\| = O(h^2)$ zodat stelsel (5.4) consistent is.

Stabiliteit

Een eindige differentie schema heet stabiel, als er een constante M bestaat, die niet afhangt van h , zodanig dat

$$\|A^{-1}\| \leq M , \quad \text{voor } h \rightarrow 0 .$$

Merk op dat stabiliteit impliceert dat het stelsel een eenduidige oplossing heeft. De matrix A uit (5.4) is symmetrisch. Dit betekent dat de eigenwaarden reëel zijn en

$$\|A^{-1}\| = \frac{1}{\lambda_{\min}} .$$

Als $0 < q_{\min} \leq q(x) \leq q_{\max}$ dan volgt uit de stelling van Gershgorin

$$q_{\min} \leq \lambda_j \leq q_{\max} + \frac{4}{h^2} \quad \text{voor } j = 1, \dots, n-1 .$$

Hieruit volgt $\|A^{-1}\| \leq \frac{1}{q_{\min}}$ zodat het schema stabiel is.

Als $q(x) = 0$ voor alle $x \in [0, 1]$, dan hebben we niets aan de stelling van Gershgorin voor het schatten van de kleinste eigenwaarde (waarom niet?). Zonder bewijs merken we op dat de eigenwaarden van $A = K$ gegeven worden door

$$\lambda_j = (2 - 2 \cos(n-j)h\pi)/h^2 , \quad j = 1, \dots, n-1 .$$

Dit geeft

$$\lambda_{\min} = (2 - 2 \cos h\pi)/h^2 = (4 \sin^2 \frac{h\pi}{2})/h^2 \approx \pi^2 .$$

Dus ook voor $q \equiv 0$ geldt dat het schema stabiel is. Het conditiegetal van A is $\frac{4}{(\pi h)^2}$.

Convergentie

Een schema heet convergent als voor de globale fout $\mathbf{y} - \mathbf{u}$ geldt

$$\lim_{h \rightarrow 0} \|\mathbf{y} - \mathbf{u}\| = 0 .$$

In de volgende stelling wordt er een verband aangegeven tussen de begrippen: consistentie, stabiliteit en convergentie.

Stelling 5.4.1 *Als een schema stabiel en consistent is, dan is het schema convergent.*

Bewijs:

De globale fout $\mathbf{y} - \mathbf{u}$ voldoet aan het stelsel

$$A(\mathbf{y} - \mathbf{u}) = A\mathbf{y} - A\mathbf{u} = \mathbf{f} + \boldsymbol{\varepsilon} - \mathbf{f} = \boldsymbol{\varepsilon} ,$$

waar uit volgt: $\mathbf{y} - \mathbf{u} = A^{-1}\boldsymbol{\varepsilon}$. Als we hiervan de norm nemen dan geldt:

$$\|\mathbf{y} - \mathbf{u}\| \leq \|A^{-1}\| \|\boldsymbol{\varepsilon}\| .$$

Uit de stabiliteit en de consistentie volgt nu direct

$$\lim_{h \rightarrow 0} \|\mathbf{y} - \mathbf{u}\| = 0 .$$

□

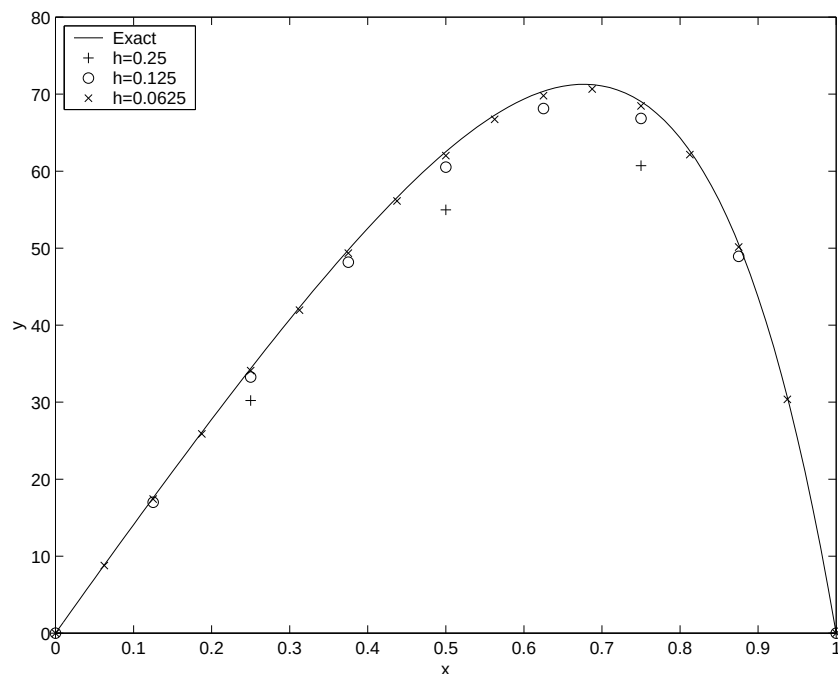
Merk op dat consistentie alleen, niet voldoende is voor convergentie.

Voorbeeld (convergentie)

Stel het warmtetransport in een staaf wordt beschreven door het randwaarde probleem:

$$\begin{aligned} -y'' &= 25e^{5x} , & 0 < x < 1 , \\ y(0) &= y(1) = 0 . \end{aligned}$$

Omdat er een warmtebron ($25e^{5x}$) is, verwachten we dat de temperatuur y positief is. In Figuur 5.2 staat de grafiek van y en van de numerieke oplossingen bepaald met de stapgrootten: $h = \frac{1}{4}, \frac{1}{8}$ en $\frac{1}{16}$. Er is een snelle convergentie van de numerieke oplossing naar de exacte



Figuur 5.2: De exacte en numerieke oplossingen van het warmteprobleem

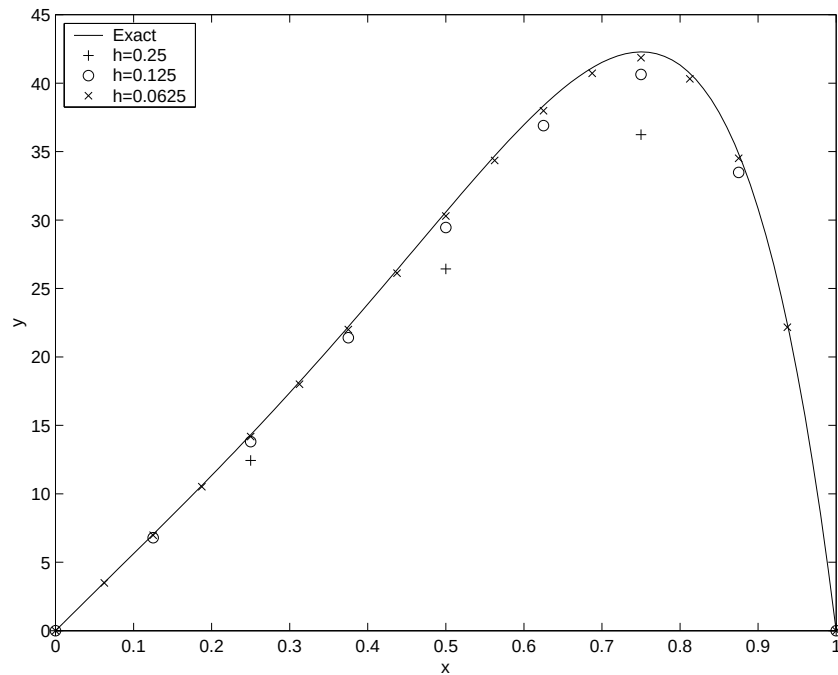
oplossing. De behaalde nauwkeurigheid bij stapgrootte $h = \frac{1}{16}$ is in een praktische situatie

waarschijnlijk voldoende.

Als een variatie op deze toepassing nemen we

$$\begin{aligned} -y'' + 9y &= 25e^{5x}, & 0 < x < 1, \\ y(0) &= y(1) = 0. \end{aligned}$$

De term $9y$ beschrijft afgifte van warmte aan de omgeving, die evenredig is met de temperatuur. De exacte en numerieke oplossingen staan in Figuur 5.3. Merk op dat de maximale



Figuur 5.3: De exacte en numerieke oplossingen van het warmteprobleem met afkoeling

temperatuur inderdaad lager is dan in het eerste voorbeeld (Figuur 5.2). Verder is er weinig verschil tussen het convergentiegedrag van beide problemen.

5.5 De conditie van de discretisatie matrix

Voor $q(x) = 0$, $0 \leq x \leq 1$ hebben we gezien dat

$$\lambda_{\min} \simeq \pi^2 \quad \text{en} \quad \lambda_{\max} \simeq \frac{4}{h^2}.$$

Dit betekent dat $K(A) = \frac{4}{\pi^2 h^2}$. Als h naar nul gaat, dan gaat de conditie van A naar oneindig. Dit ziet er niet zo goed uit. In de praktijk valt het gedrag van de fout vaak mee. Bij de afleiding van de relatieve fout is uitgegaan van een pessimistische situatie. Voor een realistischer schatting kunnen we de volgende analyse gebruiken. De numerieke oplossing voldoet aan

$$Au = f.$$

Bij een verstoring $\Delta \mathbf{f}$ geldt

$$A(\mathbf{u} + \Delta \mathbf{u}) = \mathbf{f} + \Delta \mathbf{f} ,$$

zodat $\Delta \mathbf{u}$ voldoet aan $A\Delta \mathbf{u} = \Delta \mathbf{f}$. Hiermee leiden we af

$$\|\Delta \mathbf{u}\| = \|A^{-1}\Delta \mathbf{f}\| \leq \frac{1}{\lambda_{\min}} \|\Delta \mathbf{f}\| .$$

Voor de relatieve fout volgt dan

$$\frac{\|\Delta \mathbf{u}\|}{\|\mathbf{u}\|} \leq \frac{1}{\lambda_{\min}} \frac{\|\mathbf{f}\|}{\|\mathbf{u}\|} \cdot \frac{\|\Delta \mathbf{f}\|}{\|\mathbf{f}\|} .$$

De rol van $K(A)$ is hierbij overgenomen door het "effectieve" conditiegetal $\frac{1}{\lambda_{\min}} \frac{\|\mathbf{f}\|}{\|\mathbf{u}\|}$. We weten dat $\lambda_{\min} \simeq \pi^2$ en in veel toepassingen is $\frac{\|\mathbf{f}\|}{\|\mathbf{u}\|}$ begrensd als h naar nul gaat, zodat het "effectieve" conditiegetal wel begrensd is.

5.6 Neumann randvoorwaarde

Tot nu toe hebben we randwaarde problemen bekeken met twee Dirichlet randvoorwaarden. Hoe moeten we een Neumann randvoorwaarde verwerken? We zullen deze vraag beantwoorden voor de voorwaarde

$$\frac{dy}{dx}(1) = 0 . \quad (5.6)$$

Discretisatie

We introduceren een virtueel punt $x_{n+1} = (n+1)h = 1+h$. Voor $j = n$ geldt

$$\frac{-u_{n-1} + 2u_n - u_{n+1}}{h^2} + q_n u_n = f_n . \quad (5.7)$$

Discretisatie van (5.6) geeft

$$\frac{u_{n+1} - u_{n-1}}{2h} = 0 .$$

Hieruit volgt $u_{n+1} = u_{n-1}$. Dit substitueren we in (5.7) en delen door 2:

$$\frac{-u_{n-1} + u_n}{h^2} + \frac{1}{2}q_n u_n = \frac{1}{2}f_n . \quad (5.8)$$

Deling door 2 is uitgevoerd om een symmetrische matrix A te verkrijgen. Merk op dat de vector $\mathbf{u}$ nu lengte n heeft.

Convergentie

Met Taylorontwikkeling volgt

$$u_{n+1} = u_{n-1} + O(h^3) ,$$

zodat de afbreekfout in (5.8) gelijk is aan $O(h)$ en bij alle andere vergelijkingen is een afbreekfout $O(h^2)$. Bij Dirichlet randvoorwaarden is de lokale afbreekfout in elke vergelijking

$O(h^2)$. Betekent dit dat de globale fout $O(h)$ is? Dit blijkt mee te vallen, zoals we hieronder zullen zien.

We nemen $q(x) = 0$, $0 \leq x \leq 1$. Het is bekend dat het resulterende stelsel stabiel is. We splitsen de afbreekfout in twee vectoren:

$$A\mathbf{y} = \mathbf{f} + h^2\mathbf{p} + h\mathbf{q}, \quad \text{met} \quad \mathbf{q} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ q_n \end{pmatrix}.$$

De globale fout $\mathbf{e} = \mathbf{y} - \mathbf{u}$ splitsen we ook $\mathbf{e} = \mathbf{e}^{(1)} + \mathbf{e}^{(2)}$, waarbij

$$A\mathbf{e}^{(1)} = h^2\mathbf{p} \quad \text{en} \quad A\mathbf{e}^{(2)} = h\mathbf{q}.$$

Omdat A stabiel is geldt $\|\mathbf{e}^{(1)}\| = O(h^2)$. Voor $\mathbf{e}^{(2)}$ merken we op dat

$$e_i^{(2)} = h^3 q_n i, \quad i = 1, \dots, n.$$

Omdat $hi \leq 1$ volgt $\|\mathbf{e}^{(2)}\| = O(h^2)$. Dus voor de globale fout geldt $\|\mathbf{e}\| = O(h^2)$.

5.7 Het algemene probleem

Het algemene randwaarde probleem wordt gegeven door

$$-(py')' + ry' + qy = f, \quad 0 < x < 1,$$

met randvoorwaarden

$$y(0) = \alpha \quad \text{en} \quad y'(1) = \beta.$$

Discretisatie in het punt x_j geeft:

$$\frac{-p_{j+\frac{1}{2}}(u_{j+1} - u_j) + p_{j-\frac{1}{2}}(u_j - u_{j-1}))}{h^2} + r_j \frac{u_{j+1} - u_{j-1}}{2h} + q_j u_j = f_j, \quad j = 1, \dots, n. \quad (5.9)$$

Merk op als $r_j \neq 0$ dan is de discretisatie matrix niet symmetrisch. De randvoorwaarde voor $x = 0$ geeft $u_0 = \alpha$ hetgeen gesubstitueerd kan worden in (5.9) voor $j = 1$. De randvoorwaarde voor $x = 1$ discretiseren we als volgt:

$$\frac{1}{2}(p_{n+\frac{1}{2}} \frac{u_{n+1} - u_n}{h} + p_{n-\frac{1}{2}} \frac{u_n - u_{n-1}}{h}) = p(1)\beta.$$

Merk op dat links het gemiddelde genomen is van de benaderingen voor

$$p_{n+\frac{1}{2}} \frac{dy}{dx}(1 + \frac{1}{2}h) \quad \text{en} \quad p_{n-\frac{1}{2}} \frac{dy}{dx}(1 - \frac{1}{2}h).$$

Dit gemiddelde is een benadering voor $p_n \frac{dy}{dx}(1) = p(1)\beta$. Voor deze formulering is gekozen omdat nu de term $p_{n+\frac{1}{2}}(u_{n+1} - u_n)$ in (5.9) voor $j = n$ vervangen kan worden door $-p_{n-\frac{1}{2}}(u_n - u_{n-1}) + 2hp(1)\beta$.

5.8 De convectie-diffusie vergelijking

In de vorige paragraaf hebben we de discretisatie afgeleid voor het algemene randwaarde probleem. We hebben hierbij centrale differenties toegepast. In toepassingen blijkt dit niet altijd tot (fysisch) acceptabele oplossingen te leiden. We illustreren dit aan de hand van het randwaarde probleem:

$$\begin{aligned} -y'' + vy' &= 0, & 0 < x < 1, \\ y(0) &= 1, & y(1) = \alpha, \end{aligned}$$

met $v > 0$. We kunnen dit interpreteren als een warmte transport probleem. Hierbij is er niet alleen transport door geleiding (diffusie) $-y''$ maar ook door stroming (convectie) vy' . Bovenstaande vergelijking staat bekend als de convectie-diffusie vergelijking.

Centrale differenties

Toepassen van centrale differenties geeft als stelsel

$$\frac{1}{h^2} \begin{pmatrix} 2 & -1 + \frac{vh}{2} & & & \\ -1 - \frac{vh}{2} & 2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ \emptyset & & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ \vdots \\ u_{n-1} \end{pmatrix} = \frac{1}{h^2} \begin{pmatrix} 1 + \frac{vh}{2} \\ 0 \\ \vdots \\ 0 \\ (1 - \frac{vh}{2})\alpha \end{pmatrix}.$$

Als h zo gekozen wordt dat $vh = 2$ dan wordt de matrix een onderdriehoeksmatrix. De oplossing is dan $u_i = 1$, $i = 1, \dots, n$ onafhankelijk van de waarde van α . Dit kan niet goed zijn. Het blijkt dat de oplossing acceptabel is zolang h zo klein is dat $vh < 2$.

Upwind differenties

Als de voorwaarde $vh < 2$ tot hele kleine stapgroottes leidt, is het beter om upwind discretisatie te gebruiken. Hierbij wordt vy'_j benaderd door

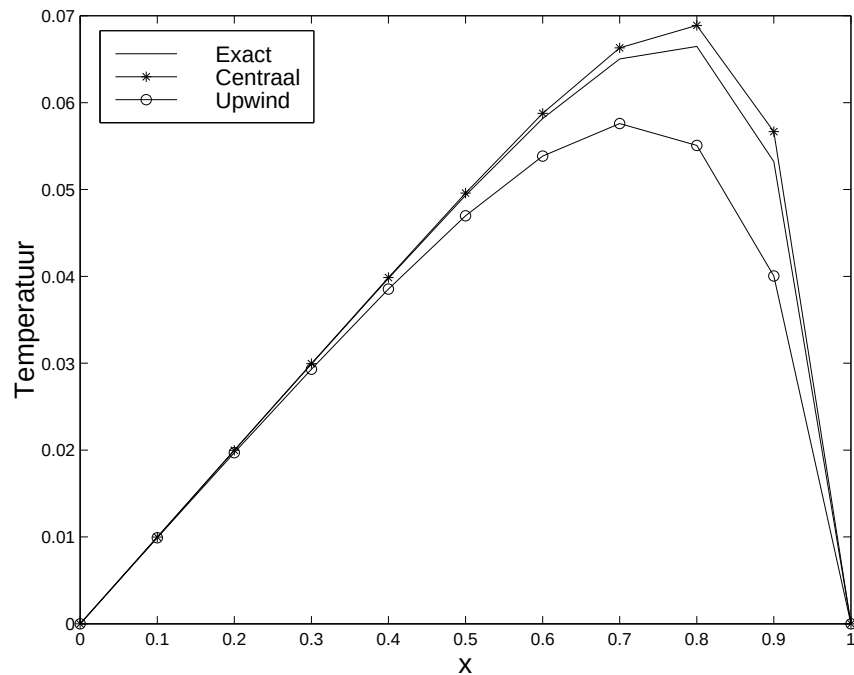
$$\begin{aligned} v \frac{u_j - u_{j-1}}{h} & \quad \text{als } v \geq 0 \quad \text{en} \\ v \frac{u_{j+1} - u_j}{h} & \quad \text{als } v < 0. \end{aligned}$$

Voorbeeld (convectie-diffusie)

Om het verschil van centrale en upwind discretisatie te zien, beschouwen we het randwaardeprobleem

$$\begin{aligned} -y'' + vy' &= 1, & 0 < x < 1, \\ y(0) &= y(1) = 0. \end{aligned}$$

Voor $h = 0.1$ zijn de oplossingen bepaald voor $v = 10, 20$ en 100 (zie Figuur 5.4, 5.5 en 5.6). Voor $v = 10$ geldt $vh < 2$ en geeft centrale discretisatie een beter resultaat dan een upwind discretisatie. Voor $v = 20$ geldt $vh = 2$. De numerieke oplossing voor de centrale differentie hangt nu niet af van de randvoorwaarde in $x = 1$. Toevallig is in dit probleem de numerieke oplossing nog redelijk. Voor $v = 100$ zien we dat de oplossing bepaald via upwind discretisatie nog steeds een goede oplossing geeft terwijl centrale discretisatie oscillaties geeft in de oplossing. Bovendien treden er grote fouten op.

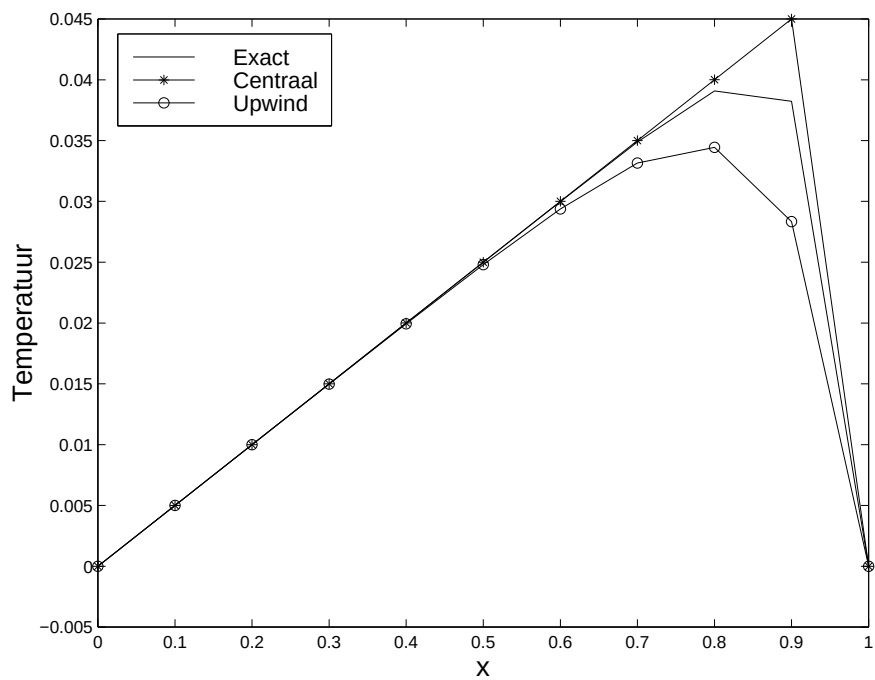


Figuur 5.4: De oplossing van het convectie-diffusie probleem voor $v = 10$

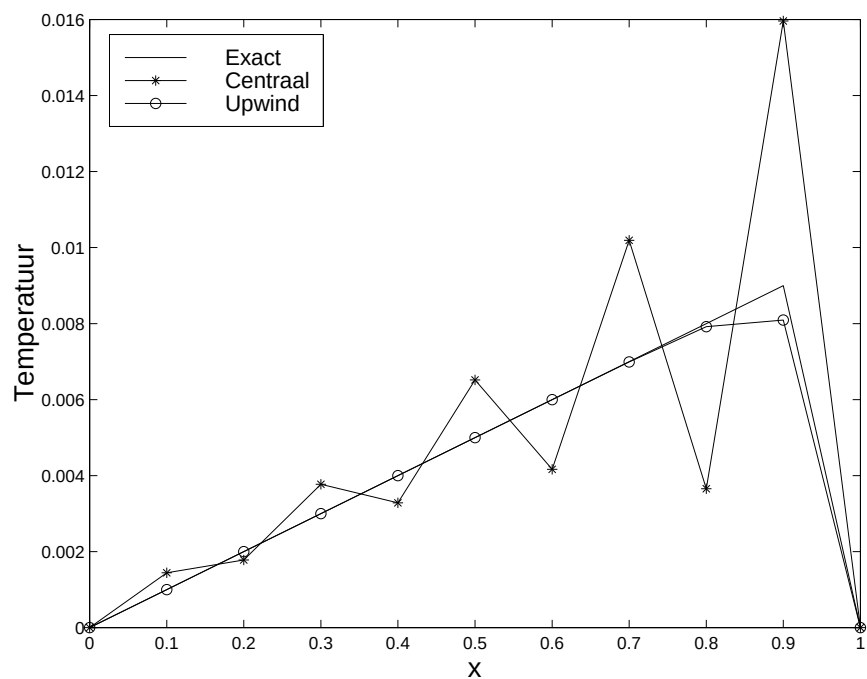
5.9 Samenvatting

In dit hoofdstuk zijn de volgende begrippen behandeld:

- model warmtetransport
- randwaarde probleem
- Dirichlet, Neumann, Robbins randvoorwaarde
- eindige differentie
- conditiegetal
- consistentie, afbreekfout
- stabiliteit
- convergentie, globale fout
- convectie-diffusie probleem
- upwind discretisatie



Figuur 5.5: De oplossing van het convectie-diffusie probleem voor $v = 20$



Figuur 5.6: De oplossing van het convectie-diffusie probleem voor $v = 100$

Hoofdstuk 6

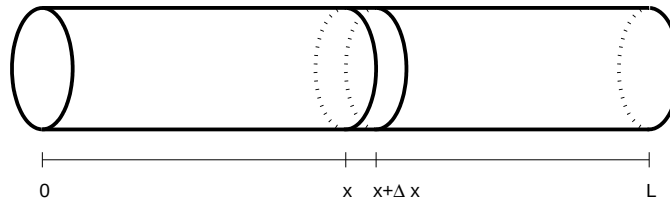
De instationaire warmtevergelijking

6.1 Inleiding

De tijdsafhankelijke temperatuurverdeling in een staaf kan beschreven worden met een parabolische partiële differentiaalvergelijking. Om zo'n begin-, randwaarde probleem op te lossen maken we gebruik van de methoden uit hoofdstuk 5 en 4. Omdat praktische voorbeelden aan de orde zullen komen in het practicum, zullen we in dit hoofdstuk geen voorbeelden geven.

6.2 Afleiding van de instationaire warmtevergelijking

Als voorbeeld nemen we de temperatuurverdeling in een staaf (Figuur 6.1). De temperatuur geven we aan met $T(x, t)$ (K). We nemen aan dat de temperatuur gegeven is in beide



Figuur 6.1: De vorm van de staaf

uiteinden $T(0, t) = T_l$ en $T(L, t) = T_r$. De warmteproductie in de staaf geven we aan met $Q(x, t)$ (J/m^3s). Voor het afleiden van de warmtevergelijking maken we gebruik van de wet van behoud van energie toegepast op het controle volume tussen x en $x + \Delta x$ en van het tijdstip t tot $t + \Delta t$. Volgens de wet van Fourier wordt de warmtestroomdichtheid gegeven door $-\lambda \frac{\partial T}{\partial x}$.

De energie balans is nu

$$\begin{aligned} \rho c T(x, t + \Delta t) A \Delta x &= \rho c T(x, t) A \Delta x \\ -\lambda \frac{\partial T}{\partial x}(x) A \Delta t + \lambda \frac{\partial T}{\partial x}(x + \Delta x) A \Delta t + Q(x) A \Delta x \Delta t, \end{aligned}$$

waarbij geldt: ρ (kg/m^3) is de soortelijke massa, C (J/kgK) is de soortelijke warmte en λ (J/msK) is de warmtegeleidingscoëfficiënt. Delen door $A \Delta x \Delta t$ en herschikken van de ter-

men geeft

$$\rho c \frac{T(x, t + \Delta t) - T(x, t)}{\Delta t} = \lambda \frac{\frac{\partial T}{\partial x}(x + \Delta x, t) - \frac{\partial T}{\partial x}(x, t)}{\Delta x} + Q .$$

Na het nemen van de limieten $\Delta x \rightarrow 0$ en $\Delta t \rightarrow 0$ krijgen we het volgende begin-, randwaarde probleem:

$$\begin{aligned} \rho c \frac{\partial T}{\partial t} &= \lambda \frac{\partial^2 T}{\partial x^2} + Q, \quad 0 < x < L, \quad 0 < t, \\ T(0, t) &= T_l(t) \quad \text{en} \quad T(L, t) = T_r(t), \quad 0 \leq t, \\ T(x, 0) &= T_0(x), \quad 0 \leq x \leq L. \end{aligned}$$

6.3 De gediscretiseerde vergelijking

Voor een eenvoudig warmteprobleem laten we zien hoe plaats- en tijdsdiscretisatie eruit komen te zien. Er is hierbij een duidelijk verschil tussen een expliciete (Euler voorwaarts) en een impliciete (Euler achterwaarts) tijdsintegratie methode.

De vergelijking

Als warmteprobleem nemen we

$$\begin{aligned} \frac{\partial y}{\partial t} &= \frac{\partial^2 y}{\partial x^2}, \quad 0 < x < 1, \quad 0 < t \leq T \\ y(0, t) &= y_l(t) \quad \text{en} \quad y(1, t) = y_r(t), \quad 0 \leq t \leq T \\ y(x, 0) &= y_0(x), \quad 0 \leq x \leq 1. \end{aligned} \tag{6.1}$$

Semi-discretisatie

We discretiseren (6.1) in de x -richting met behulp van de eindige differentie methode. Deel het interval $[0, 1]$ op in n gelijke delen met lengte h . De knooppunten zijn $x_i = ih$ voor $i = 0, \dots, n$. De numerieke benadering van $y(x_i, t)$ geven we aan met $u_i(t)$. De vector $\mathbf{u}(t)$ wordt gegeven door $\mathbf{u}(t) = (u_1(t), \dots, u_{n-1}(t))^T$. De waarden in de randpunten zijn weggelaten, omdat deze bekend zijn uit de randvoorwaarden. Gebruikmakend van de technieken uit hoofdstuk 5 krijgen we het volgende stelsel eerste orde differentiaalvergelijkingen

$$\begin{aligned} \frac{d\mathbf{u}}{dt} &= K\mathbf{u} + \mathbf{r}, \quad 0 < t \leq T. \\ \mathbf{u}(0) &= \mathbf{y}_0. \end{aligned} \tag{6.2}$$

De matrix K en vector $\mathbf{r}$ worden gegeven door

$$K = \frac{1}{h^2} \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ \emptyset & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix} \quad \text{en} \quad \mathbf{r} = \frac{1}{h^2} \begin{pmatrix} y_l(t) \\ 0 \\ \vdots \\ 0 \\ y_r(t) \end{pmatrix}.$$

Men spreekt hierbij van semi-discretisatie (of methode der lijnen), omdat er (vooralsnog) wel in x maar niet in t wordt gediscretiseerd.

Tijdsintegratie (Euler voorwaarts)

Het stelsel (6.2) kan in de tijd geïntegreerd worden met één van de methoden uit hoofdstuk 4. We kiezen eerst voor de voorwaartse methode van Euler. Stel we nemen m stappen in de tijd ter lengte $k = \frac{T}{m}$ en $t_j = jk$. De numerieke oplossing geven we aan met w_i^j , hetwelk een benadering is van $u_i(t_j)$ en dus van $y(x_i, t_j)$. In inwendige roosterpunten geldt dan de volgende vergelijking

$$\frac{\omega_i^{j+1} - \omega_i^j}{k} = \frac{\omega_{i-1}^j - 2\omega_i^j + \omega_{i+1}^j}{h^2}. \quad (6.3)$$

De afbreekfout van deze differentievergelijking is

$$\tau_{i,j} = \frac{k}{2} \frac{\partial^2 u}{\partial t^2}(x_i, \mu_j) - \frac{h^2}{12} \frac{\partial^4 u}{\partial x^4}(\xi_i, t_j), \quad (6.4)$$

waarbij $\mu_j \in (t_j, t_{j+1})$ en $\xi_i \in (x_{i-1}, x_{i+1})$. De methode is dus $O(k + h^2)$. Het is verstandig om k en h zo te kiezen dat beide componenten aan de rechterkant van (6.4) ongeveer even groot zijn. Dit betekent dat als je de plaatsstap halveert, het voor de hand ligt om de tijdstap 4 maal zo klein te nemen.

Merk op dat we ω_i^{j+1} eenvoudig kunnen bepalen via (6.3) immers

$$\omega_i^{j+1} = (1 - \frac{2k}{h^2})\omega_i^j + \frac{k}{h^2}(\omega_{i-1}^j + \omega_{i+1}^j), \quad 1 \leq i \leq n-1, \quad (6.5)$$

waarbij ω_0^{j+1} en ω_n^{j+1} volgen uit de randvoorwaarden. Merk op dat de waarden in het roosterpunt (x_i, t_{j+1}) afhangt van de waarden in de roosterpunten (x_{i-1}, t_j) , (x_i, t_j) en (x_{i+1}, t_j) . Deze punten zijn met een $\times$ aangegeven in Figuur 6.2. De waarden, die al bekend zijn uit de begin- en randvoorwaarden zijn aangegeven met een $\circ$. In matrix vectornotatie geldt $u^{j+1} = (I + kK)u^j + kr^j$.

Stabiliteit

Omdat Euler voorwaarts voorwaardelijk stabiel is, is het van belang om te bepalen hoe groot de tijdstap genomen kan worden zonder dat er instabiliteiten optreden. De matrix K uit (6.2) is symmetrisch, dus de eigenwaarden zijn reëel. Toepassing van de stelling van Gershgorin geeft

$$-\frac{4}{h^2} \leq \lambda_i \leq 0, \quad \text{voor } 1 \leq i \leq n-1.$$

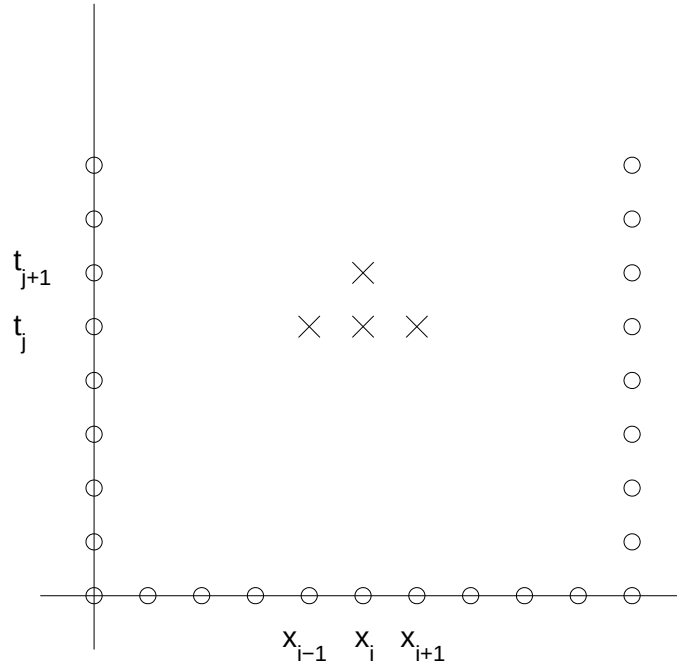
De matrix heeft geen positieve eigenwaarden zodat het stelsel differentiaalvergelijkingen (6.2) stabiel is. In paragraaf 5.5 hebben we een expliciete uitdrukking voor de eigenwaarden van K gegeven. Hieruit blijkt

$$\frac{\max |\lambda_i|}{\min |\lambda_i|} \approx \frac{4}{\pi^2 h^2}.$$

Voor kleine waarden van h is (6.2) een stijf stelsel.

We weten dat Euler voorwaarts stabiel is als de tijdstap k voldoet aan

$$k \leq \frac{2}{\max |\lambda_i|} \leq \frac{h^2}{2}.$$



Figuur 6.2: De afhankelijkheid bij Euler voorwaarts

Merk op dat als de plaatsstap gehalveerd wordt, de tijdstap 4 maal zo klein genomen moet worden om te voldoen aan de stabiliteitsvoorwaarde. Voor kleine h kan het verstandiger zijn om over te stappen op de achterwaartse methode van Euler, die onvoorwaardelijk stabiel is.

Tijdsintegratie (Euler achterwaarts)

Als we Euler achterwaarts toepassen dan krijgen we de formule

$$\frac{\omega_i^{j+1} - \omega_i^j}{k} = \frac{\omega_{i-1}^{j+1} - 2\omega_i^{j+1} + \omega_{i+1}^{j+1}}{h^2},$$

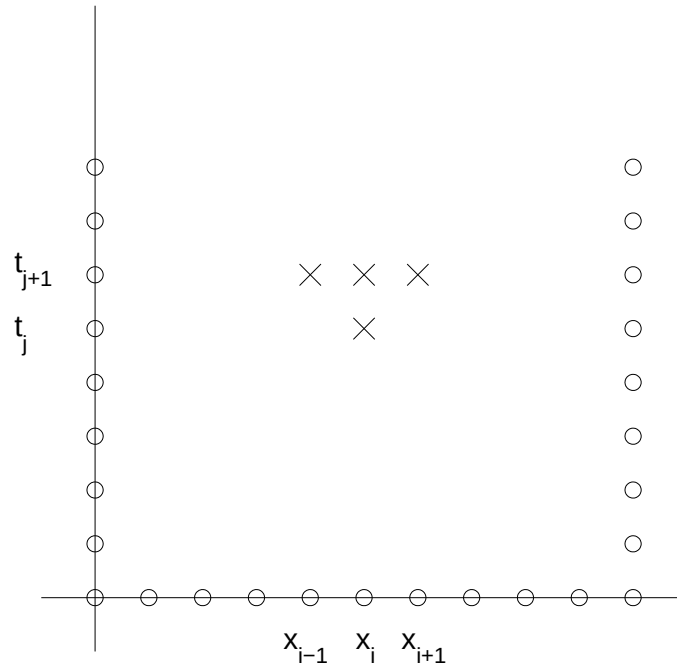
in inwendige roosterpunten. Ook hiervan is de afbreekfout $O(k + h^2)$. Voor het bepalen van ω_i^{j+1} moeten we nu het volgende lineaire stelsel oplossen

$$A\mathbf{x} = \mathbf{b}, \quad (6.6)$$

$$\text{met } A = (I - kK), \quad \mathbf{x} = \begin{pmatrix} u_1^{j+1} \\ \vdots \\ u_{n-1}^{j+1} \end{pmatrix} \text{ en } \mathbf{b} = \begin{pmatrix} u_1^j \\ \vdots \\ u_{n-1}^j \end{pmatrix} + \frac{k}{h^2} \begin{pmatrix} y_l(t_j) \\ 0 \\ \vdots \\ 0 \\ y_r(t_j) \end{pmatrix}.$$

Merk op dat A een matrix is met veel nullen. Toch kost het oplossen van (6.6) meer werk dan het bepalen van $\omega_{i,j+1}$ via (6.5). Een voordeel is wel, dat Euler achterwaarts onvoorwaardelijk stabiel is, zodat de tijdstap afgesteld kan worden op de gewenste nauwkeurigheid. Ook hierbij is een afweging nodig: goedkoop oplossen met veel tijdstappen (Euler voorwaarts) versus duur oplossen met weinig tijdstappen (Euler achterwaarts).

Nu hangt de waarde in het roosterpunt (x_i, t_{j+1}) af van de waarden in de roosterpunten (x_i, t_j) , (x_{i-1}, t_{j+1}) en (x_{i+1}, t_{j+1}) , zie Figuur 6.3.



Figuur 6.3: De afhankelijkheid bij Euler achterwaarts

Tijdsintegratie (Trapezium)

We kunnen ook de Trapezium regel gebruiken als tijdsintegratie methode. De uitwerking hiervan gaat analoog. In de literatuur wordt de Trapeziumregel ook wel aangeduid als de Crank-Nicolson methode.

6.4 Samenvatting

In dit hoofdstuk zijn de volgende begrippen behandeld:

- instationaire warmtevergelijking
- semi-discretisatie
- afbreekfout
- stabiliteit
- Crank-Nicolson

Hoofdstuk 7

Numerieke integratie

7.1 Introductie

Voor het bepalen van volume, massa, lengte enz. is het nodig om de integraal van een functie uit te rekenen. Vaak is het onmogelijk om van een integrand de primitieve te vinden. Om dan toch een antwoord te bepalen wordt er veelal gebruik gemaakt van numerieke integratie.

Als voorbeeld nemen we de productie van een spoiler, die geplaatst wordt op de cabine van een vrachtwagen (Figuur 7.1). De vorm van de spoiler wordt beschreven door een sinusfunctie met een periode van 2π meter. De spoiler wordt door walsen verkregen uit een vlakke plaat. De fabrikant wil weten hoe breed de vlakke plaat moet zijn opdat de horizontale afmeting van de spoiler 80 cm is. Het antwoord hierop is de booglengte van de kromme gegeven door

$$\begin{aligned}x(t) &= t & 0 \leq t \leq 0.8 \\y(t) &= \sin t\end{aligned}$$

Voor het bepalen van de booglengte kunnen we gebruik maken van de formule

$$l = \int_0^{0.8} \sqrt{1 + \left(\frac{dy}{dt}\right)^2} dt = \int_0^{0.8} \sqrt{1 + (\cos t)^2} dt.$$

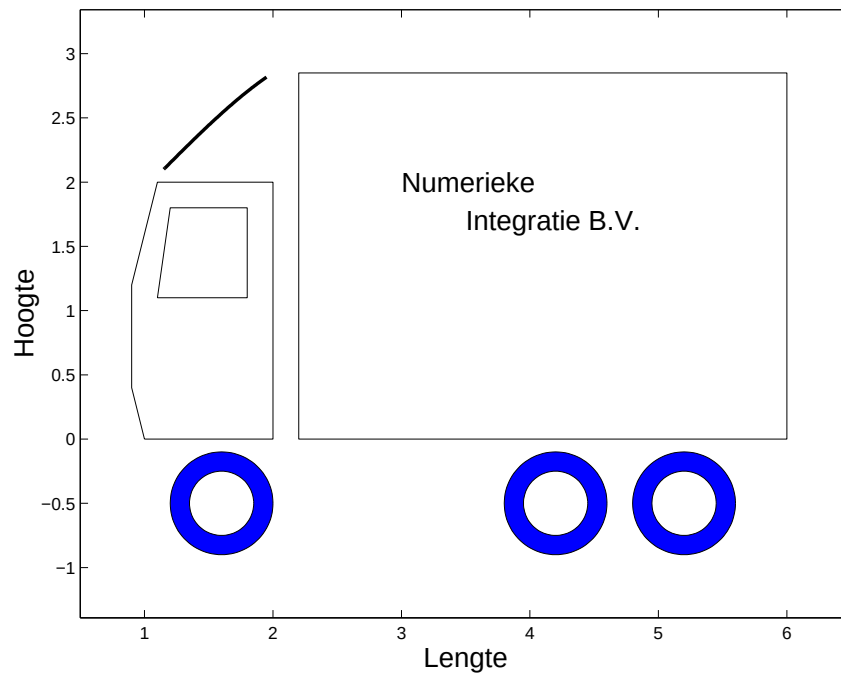
Deze integraal is niet eenvoudig te bepalen. We zullen in dit hoofdstuk laten zien hoe de lengte bepaald kan worden met behulp van numerieke integratie.

7.2 Eenvoudige numerieke integratieformules

Na een herhaling van de definitie van een integraal zullen we een aantal eenvoudige integratieregels geven. Hiervan zal de benaderingsfout en het effect van afrondfouten onderzocht worden.

Definitie

Een partitie P van $[a, b]$ is een eindig aantal punten x_k waarbij $a = x_0 < x_1 < \dots < x_n = b$. Een bijbehorende strooiing T is een verzameling tussenpunten t_k zodat $x_{k-1} \leq t_k \leq x_k$. De lengte van een interval geven we aan met $h_k = x_k - x_{k-1}$ en de maaswijdte $m(P) = \max_{1 \leq k \leq n} \{h_k\}$. De



Figuur 7.1: Vrachtwagen met spoiler op de cabine

Riemansom voor een functie f die continu is op $[a, b]$ is nu gedefinieerd als:

$$R(f, P, T) = \sum_{k=1}^n f(t_k) h_k .$$

Kies een rij verdelingen $P_1, P_2, \dots$ en bijbehorende strooiingen T_n zodanig dat $m(P_n) \rightarrow 0$ dan convergeert

$$R(f, P_n, T_n) \text{ naar een limiet } I = \int_a^b f(x) dx .$$

Numerieke integratieregels lijken veel op de Riemansom. Het belangrijkste verschil is de gewenste efficiëntie bij numerieke integratie.

De Rechthoekregel

Neem een equidistante verdeling zodat de knooppunten gegeven worden door: $x_k = a + kh$, $k = 0, 1, \dots, n$ met $h = \frac{b-a}{n}$. De strooipunten nemen we gelijk aan de linkerknooppunten: $t_k = x_{k-1}$. De Riemansom wordt dan gegeven door

$$I_R = h[f(a) + f(a+h) + \dots + f(b-h)] .$$

Stelling 7.2.1 *Zij f een differentieerbare functie op $[a, b]$. Stel M , is het maximum van $|f'|$ op $[a, b]$ dan geldt:*

$$\left| \int_a^b f(x) dx - I_R \right| \leq \frac{1}{2} M_1 (b-a) h$$

Bewijs:

We beschouwen eerst het interval $[x_{k-1}, x_k]$. Uit een Taylor ontwikkeling volgt:

$$f(x) = f(x_{k-1}) + (x - x_{k-1})f'(\xi(x)) \quad \text{met} \quad x_{k-1} \leq \xi(x) \leq x .$$

Hieruit volgt:

$$\left| \int_{x_{k-1}}^{x_k} [f(x) - f(x_{k-1})] dx \right| \leq M_1 \int_{x_{k-1}}^{x_k} (x - x_{k-1}) dx ,$$

zodat

$$\left| \int_{x_{k-1}}^{x_k} f(x) dx - hf(x_{k-1}) \right| \leq \frac{1}{2} M_1 h^2 .$$

Voor de totale fout geldt dus:

$$\left| \int_a^b f(x) dx - I_R \right| \leq \sum_{k=1}^n \left| \int_{x_{k-1}}^{x_k} f(x) dx - hf(x_{k-1}) \right| \leq \frac{1}{2} M_1 h^2 n = \frac{1}{2} M_1 (b-a) h .$$

□

Voorbeeld(spoiler)

Voor het voorbeeld genoemd in de introductie willen we de lengte van de vlakke plaat benaderen met een fout van 1 cm. Als we de Rechthoekregel toepassen betekent dit dat

$$\frac{1}{2} M_1 0.8 h \leq 0.01 .$$

De afgeleide van de integrand $f(t) = \sqrt{1 + (\cos t)^2}$ wordt gegeven door:

$$f'(t) = \frac{-\cos t \sin t}{\sqrt{1 + (\cos t)^2}} = \frac{-\frac{1}{2} \sin 2t}{\sqrt{1 + (\cos t)^2}} .$$

Hieruit volgt $|f'| \leq M_1 = \frac{1}{2}$ zodat een stapgrootte $h = 0.05$ voldoende zou moeten zijn. De integraal is 1.0759 m in 5 cijfers nauwkeurig. Voor $n = 16$ geeft de Rechthoekregel 1.0807 zodat de fout inderdaad kleiner is dan 1 cm.

De nauwkeurigheid van de Rechthoekregel is $O(h)$. Hierna wordt een methode gegeven, die met dezelfde hoeveelheid werk, een nauwkeuriger antwoord geeft.

Midpuntregel

Neem een equidistante verdeling zodat de knooppunten gegeven worden door: $x_k = kh$, $k = 0, 1, \dots, n$ en $h = \frac{b-a}{n}$. De middenpunten $\frac{x_k + x_{k-1}}{2}$ worden als strooipunten gekozen. De Riemannsom wordt dan gegeven door:

$$I_m = h[f(a + \frac{1}{2}h) + f(a + \frac{3}{2}h) + \dots + f(b - \frac{1}{2}h)] .$$

Stelling 7.2.2 Zij f een tweemaal differentieerbare functie op $[a, b]$. Stel M_2 is het maximum van $|f''|$ op $[a, b]$ dan geldt:

$$\left| \int_a^b f(x) dx - I_m \right| \leq \frac{1}{24} M_2 (b-a) h^2 .$$

Bewijs deze stelling zelf.

(Hint: ontwikkel f rond $\frac{x_k + x_{k-1}}{2}$ in een Taylorpolynoom van de graad 1).

□

Voorbeeld (spoiler)

In Tabel 7.1 zijn de fouten gegeven voor verschillende waarden van de stapgrootte h . Op grond van de theorie verwachten we dat de fout bij de rechthoeksregel afneemt met een factor

h	rechthoeksregel	Midpuntregel
0.8	− 0.055	− 0.0117
0.4	− 0.0336	− 0.0028
0.2	− 0.0182	− 0.00068
0.1	− 0.0094	− 0.00017
0.05	− 0.0048	− 0.000043

Tabel 7.1: De fout voor verschillende waarden van h .

2 en die bij de Midpuntregel met een factor 4 als h gehalveerd wordt. De resultaten zijn in overeenstemming met onze verwachtingen.

Ook bij numerieke integratie kunnen meet- en afrondfouten een belangrijke rol spelen. Neem aan dat de functiewaarden verstoord zijn met een fout ε :

$$\hat{f}(x) = f(x) + \varepsilon(x) .$$

Als we de integraal benaderen met de Rechthoekregel dan geldt:

$$\left| \int_a^b f(x) dx - h \sum_{i=0}^{n-1} \hat{f}(x_i) \right| \leq \frac{1}{2} M_1 (b-a) h + h \sum_{k=0}^{n-1} \varepsilon(x_k) .$$

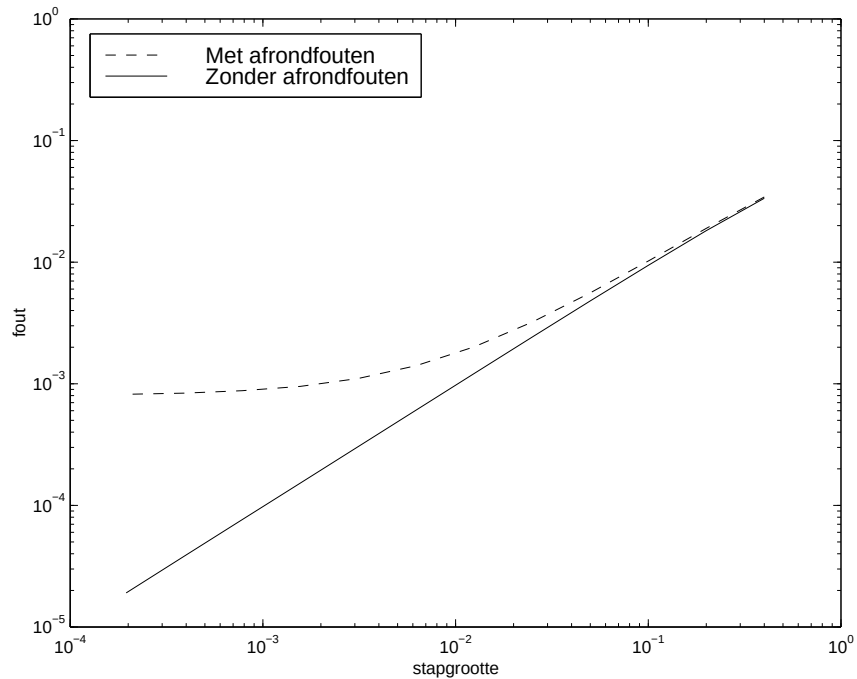
Onder de aanname $|\varepsilon(x)| \leq \varepsilon_{\max}$ volgt:

$$\left| \int_a^b f(x) dx - \hat{I}_R \right| \leq \left(\frac{1}{2} M_1 h + \varepsilon_{\max} \right) (b-a) .$$

Merk op dat het geen zin heeft om h veel kleiner te kiezen dan $\frac{2\varepsilon_{\max}}{M_1}$.

Voorbeeld (spoiler)

We hebben de berekening van de lengte van de vlakke plaat opnieuw gedaan met de veronderstelling dat er een kleine fabricagefout opgetreden is. Als gevolg daarvan bevat de integrand een afrondfout $\varepsilon(x) = 10^{-3}$. In Figuur 7.2 zien we het effect van de afrondfouten: de totale fout blijft groter dan 0.8×10^{-3} . Verder heeft het geen zin om de stapgrootte kleiner te kiezen



Figuur 7.2: De fout in de bepaling van de lengte van de vlakke plaat.

dan $4\varepsilon_{\max} = 4 \cdot 10^{-3}$.

Als laatste kan voor de numerieke bepaling van een integraal I onderscheid gemaakt worden tussen een goed en een slecht gesteld probleem. Neem aan dat de fout begrensd wordt door de ongelijkheid

$$|f(x) - \hat{f}(x)| \leq |f(x)|\varepsilon.$$

Een bovengrens voor de relatieve fout in het antwoord is nu

$$\frac{\left| \int_a^b f(x)dx - \int_a^b \hat{f}(x)dx \right|}{\int_a^b f(x)dx} \leq \frac{\int_a^b |f(x)|dx}{\int_a^b f(x)dx} \cdot \varepsilon.$$

Definieer $K_I = \frac{\int_a^b |f(x)|dx}{\int_a^b f(x)dx}$ als het conditiegetal van de integraal I . Als $K_I \gg 1$ dan is het numeriek bepalen van I een slecht gesteld probleem.

Voorbeeld (winst/verlies)

De winst of het verlies per dag van een autofabrikant is afhankelijk van het seizoen. In de winter worden minder auto's verkocht dan in de zomer. In de winter zijn de kosten hoger dan de inkomsten en wordt er dus verlies geleden. We veronderstellen de volgende winstformule (in miljard \$) in het eerste kwartaal:

$$w_{\text{lente}}(t) = 0.01 + \sin\left(\pi t - \frac{\pi}{2}\right), \quad t \in [0, 1]$$

en in het derde kwartaal:

$$w_{\text{herfst}}(t) = 0.01 + \sin(\pi t + \frac{\pi}{2}) , \quad t \in [0, 1] .$$

De totale winst in het eerste kwartaal (W_{lente}) is gelijk aan:

$$W_{\text{lente}} = \int_0^1 w_{\text{lente}}(t) dt .$$

In beide kwartalen bedraagt de winst 0.01 (is 10 miljoen \$). Omdat

$$\int_0^1 |w_{\text{lente}}(t)| dt \simeq 0.63$$

is $K_I = 63$ en is het bepalen van de integraal een slecht gesteld probleem. Als we beide integralen bepalen met de Rechthoekregel krijgen we voor $n = 50$ de antwoorden:

$$\begin{aligned} W_{\text{lente}} &= -0.01 , \\ W_{\text{herfst}} &= 0.03 . \end{aligned}$$

De kleine stapgrootte $h = 0.02$ is niet voldoende klein om de winst correct te bepalen.

7.3 Newton Cotes formules

In deze paragraaf zullen we algemene integratieregels beschrijven. We beginnen met de Trapeziumregel en we geven het verband tussen numerieke integratie en interpolatie. Daarna behandelen we integratieregels die gebaseerd zijn op hogere orde interpolatie. Als een bijzonder geval daarvan beschouwen we de Newton Cotes formules.

De Trapeziumregel

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 paragraaf 2.2):

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

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

$$\int_a^b p(x)dx = \frac{b-a}{2}(f(a) + f(b)) . \quad (7.1)$$

Deze benadering heet de Trapeziumregel, omdat (7.1) gelijk is aan het oppervlak van het trapezium met hoekpunten $(a, 0)$, $(b, 0)$, $(b, f(b))$ en $(a, f(a))$. Merk op dat de Trapeziumregel ook aan de orde geweest is in paragraaf 4.2 bij het oplossen van beginwaarde problemen.

De restterm

Bij lineaire interpolatie op a en b hebben we de volgende restterm

$$f(x) - p(x) = \frac{1}{2}(x-a)(x-b)f''(\xi(x)) , \quad (7.2)$$

waarbij we $\xi(x)$ schrijven in plaats van ξ om de afhankelijkheid van ξ ten aanzien van x te benadrukken. Als we de linker- en rechterterm van (7.2) integreren dan krijgen we:

$$\int_a^b f(x)dx - \frac{b-a}{2}(f(b) + f(a)) = \frac{1}{2} \int_a^b (x-a)(x-b)f''(\xi(x))dx . \quad (7.3)$$

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

Stelling 7.3.1 *Zij $f \in C^2[a, b]$. Dan is er een $\eta \in [a, b]$ zo dat*

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

Bewijs:

Zioj $m = \min f''(x)$ en $M = \max f''(x)$. Dan geldt, wegens $(x-a)(b-x) \geq 0$ op $[a, b]$:

$$\begin{aligned} m \int_a^b (x-a)(b-x)dx &\leq \int_a^b (x-a)(b-x)f''(\xi(x))dx \\ &\leq M \int_a^b (x-a)(b-x)dx , \end{aligned}$$

zodat er een $\mu \in (m, M)$ is waarvoor geldt

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

Omdat f'' continu is, is er een $\eta \in [a, b]$ zodat $\mu = f''(\eta)$. Invullen in (7.3) voltooit het bewijs. $\square$

Als we de Trapeziumregel gebruiken om een integraal te benaderen dan is de fout meestal groter dan de gewenste nauwkeurigheid. Een nauwkeuriger antwoord kan verkregen worden met de gerepeteerde Trapeziumregel. In dit geval verdeelt men het interval in n delen ter grootte $h = \frac{b-a}{n}$. Op elk deelinterval $[x_{k-1}, x_k]$ met $x_k = a + kh$ wordt dan de Trapeziumregel toegepast:

$$\int_{x_{k-1}}^{x_k} f(x)dx = \frac{h}{2}[f(x_k) + f(x_{k-1})] - \frac{h^3}{12}f''(\xi_k) ,$$

met $x_{k-1} \leq \xi_k \leq x_k$. Dus

$$\begin{aligned} \int_a^b f(x) dx &= h \left[\frac{1}{2} f(x_0) + f(x_1) + \dots + f(x_{n-1}) + \frac{1}{2} f(x_n) \right] \\ &\quad - \frac{h^3}{12} \sum_{k=1}^n f''(\xi_k) . \end{aligned} \quad (7.4)$$

Omdat voor een continue functie f'' geldt:

$$\min_k f''(\xi_k) \leq \frac{1}{n} \sum_{k=1}^n f''(\xi_k) \leq \max_k f''(\xi_k)$$

bestaat er een $\xi \in [a, b]$ zodat

$$\sum_{k=1}^n f''(\xi_k) = n f''(\xi) .$$

Dit invullen in (7.4) geeft

$$\int_a^b f(x) dx = I_T(h) - \frac{b-a}{12} h^2 f''(\xi) ,$$

waarbij $I_T(h) = h \left[\frac{1}{2} f(x_0) + f(x_1) + \dots + \frac{1}{2} f(x_n) \right] .$

Opmerkingen

1. De Restterm bij de gerepeteerde Trapeziumregel is $O(h^2)$, terwijl de Rechthoekregel een restterm $O(h)$ heeft. Bij beiden methoden wordt dezelfde hoeveelheid werk gebruikt. De Trapeziumregel heeft dus duidelijk de voorkeur.
2. De gerepeteerde Trapeziumregel kan ook als volgt geïnterpreteerd worden. Stel s is de lineaire spline benadering van f op $a = x_0, \dots, x_n = b$. Dan geldt

$$I_T(h) = \int_a^b s(x) dx .$$

Voorbeeld (winst/verlies)

We hebben de winst uit het voorbeeld gegeven in paragraaf 7.2 ook benaderd met de Trapeziumregel. Voor $n = 50$ krijgen we de antwoorden:

$$W_{\text{lente}} = 0.01 \quad \text{en} \quad W_{\text{herfst}} = 0.01 .$$

Merk op dat de resultaten nu exact zijn.

Algemene kwadratuurformules

We hebben gezien dat lineaire interpolatie aanleiding geeft tot de Trapeziumregel. Als we

andere interpolatieformules gebruiken krijgen we verschillende integratieregels.

Laat $x_0, \dots, x_m \in [a, b]$ gegeven zijn. Zij p het Lagrange interpolatiepolynoom van f op de punten $x_0, \dots, x_m$. We nemen dan $\int_a^b p(x)dx$ als benadering voor $\int_a^b f(x)dx$. Omdat $p(x) = \sum_{k=0}^m f(x_k)L_{km}(x)$ geldt

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

met als coëfficiënten

$$w_k = \int_a^b L_{km}(x)dx ,$$

die blijkbaar onafhankelijk van f zijn. We noemen $\sum_{k=0}^m w_k f(x_k)$ de bij $x_0, \dots, x_m$ behorende interpolatoire kwadratuurformule voor $\int_a^b f(x)dx$. De x_k heten de steunpunten en w_k de gewichten van de kwadratuurformule.

Stelling 7.3.2 *Als f een polynoom van hoogstens graad m is, dan is elke $m+1$ -punts interpolatoire kwadratuurformule exact:*

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

Bewijs:

Uit stelling 2.3.1 volgt dat f samenvalt met zijn Lagrange interpolatiepolynoom op $x_0, \dots, x_m$ en dus geldt

$$\int_a^b f(x)dx = \int_a^b p(x)dx .$$

□

Omgekeerd geldt

Stelling 7.3.3 *Wanneer $x_0, \dots, x_m$ en $w_0, \dots, w_m$ gegeven getallen zijn en er geldt*

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

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

Bewijs:

Zij p het interpolatiepolynoom op $x_0, \dots, x_m$ van f . De vraag is dan of geldt

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

Dit is zo omdat

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

□

We kunnen deze stelling ook gebruiken om de gewichten w_k uit te rekenen als de punten $x_0, \dots, x_m$ gegeven zijn. Neem voor $p(x)$ achtereenvolgens $1, x, \dots, x^m$. Dit levert een stelsel van $m + 1$ vergelijkingen in de $m + 1$ onbekenden w_k op. Men kan aantonen dat dit stelsel eenduidig oplosbaar is.

Midpuntregel

Als we de nulde orde interpolatie van f nemen in het punt $m = \frac{1}{2}(a + b)$ dan vinden we als

kwadratuurformule voor $\int_a^b f(x) dx$:

$$(b - a)f(m) .$$

Dit noemen we de Midpuntregel. Deze regel is nauwkeuriger dan men op het eerste gezicht zou verwachten. Als $f \in C^2[a, b]$ dan geldt

$$f(x) = f(m) + (x - m)f'(m) + \frac{1}{2}(x - m)^2 f''(\xi(x))$$

en

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

Omdat $(x - m)^2 \geq 0$ is kunnen we afleiden

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

Voorbeeld (winst/verlies)

Als we de winst of het verlies uitrekenen met de gerepeteerde Midpuntregel

$$h[f(a + \frac{1}{2}h) + \dots + f(b - \frac{1}{2}h)] ,$$

dan krijgen we voor $n = 50$

$$W_{\text{lente}} = 0.01 \quad \text{en} \quad W_{\text{herfst}} = 0.01 .$$

Newton-Cotes formules

Laat $x_0, \dots, x_m$ equidistant in $[a, b]$ liggen met $x_0 = a$ en $x_m = b$. De kwadratuurformules, die ontstaan door integratie van het m -de orde interpolatiepolynoom p_m van f noemt men Newton-Cotes formules.

Voor de resttermen van deze formules geldt de volgende stelling:

Stelling 7.3.4 Voor de restterm R_m van de Newton Cotes kwadratuurformules voor het benaderen van $\int_a^b f(x)dx$ geldt:

als m even is en $f \in C^{m+2}[a, b]$ dan

$$\begin{aligned} R_m &= C_m \left(\frac{b-a}{m} \right)^{m+3} f^{(m+2)}(\xi) \quad \text{met} \\ C_m &= \frac{1}{(m+2)!} \int_0^m t^2(t-1)\dots(t-m)dt; \end{aligned}$$

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

$$\begin{aligned} R_m &= D_m \left(\frac{b-a}{m} \right)^{m+2} f^{(m+1)}(\xi) \quad \text{met} \\ D_m &= \frac{1}{(m+1)!} \int_0^m t(t-1)\dots(t-m)dt. \end{aligned}$$

Simpsonregel

We kiezen nu als steunpunten $x_0 = a$, $x_1 = \frac{1}{2}(a+b)$ en $x_2 = b$. De integratieregels is dan

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

Het is niet moeilijk om in te zien dat de Simpsonregel exact is voor de polynomen $f(x) = 1$, $f(x) = x - x_1$ en $f(x) = (x - x_1)^2$.

Gauss formules

De integratieregels $\sum_{k=0}^m w_k f(x_k)$ hebben we afgeleid door de steunpunten x_k van te voren te kiezen en vervolgens de w_k te bepalen zodat de regel exact is voor polynomen van de graad $\leq m$. Men zou echter ook de vraag kunnen stellen bepaal alle w_k en x_k zo dat polynomen van een zo hoog mogelijke graad nog exact worden geïntegreerd, in de hoop dat dit ook voor willekeurige functies tot betere resultaten zal leiden. Door te eisen dat $1, x, \dots, x^{2m+1}$ exact worden geïntegreerd ontstaat er een stelsel van $2m+2$ niet-lineaire vergelijkingen in de $2m+2$ onbekenden w_k en x_k . Men kan aantonen dat dit stelsel oplosbaar is. De resulterende kwadratuurformules heten Gauss formules. Voor de $m+1$ punts Gauss formule geldt de volgende restterm:

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

Voorbeeld (2 punts Gauss formule)

Stel dat we c_0, c_1, x_0 en x_1 willen bepalen zo dat de integratieformule

$$\int_{-1}^1 f(x) dx \approx c_0 f(x_0) + c_1 f(x_1)$$

het exacte resultaat geeft als f een polynoom is van de graad 3. We kiezen voor f de functies $1, x, x^2$ en x^3 zodat c_0, c_1, x_0 en x_1 moeten voldoen aan

$$\begin{aligned} c_0 + c_1 &= \int_{-1}^1 1 dx = 2, \\ c_0 x_0 + c_1 x_1 &= \int_{-1}^1 x dx = 0, \\ c_0 x_0^2 + c_1 x_1^2 &= \int_{-1}^1 x^2 dx = \frac{2}{3}, \\ c_0 x_0^3 + c_1 x_1^3 &= \int_{-1}^1 x^3 dx = 0. \end{aligned}$$

Het is eenvoudig om in te zien dat de oplossing van dit stelsel gegeven wordt door

$$c_0 = 1, \quad c_1 = 1, \quad x_0 = \frac{-\sqrt{3}}{3} \text{ en } x_1 = \frac{\sqrt{3}}{3},$$

zodat de integratieformule gegeven wordt door

$$\int_{-1}^1 f(x) dx \approx f\left(\frac{-\sqrt{3}}{3}\right) + f\left(\frac{\sqrt{3}}{3}\right).$$

Voorbeeld (vergelijken methoden)

In Tabel 7.2 staan de resttermen als men ter berekening van $\int_0^\pi \sin x dx$ een zeker aantal punten investeert in Trapezium, Simpson en 5-punts Gauss. 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.

methode	aantal malen gerepeteerd	aantal punten	fout
Trapezium	4	5	$1.04 \cdot 10^{-1}$
Simpson	2	5	$4.56 \cdot 10^{-3}$
5p. Gauss	1	5	$1.11 \cdot 10^{-7}$
Trapezium	8	9	$2.58 \cdot 10^{-2}$
Simpson	4	9	$2.69 \cdot 10^{-4}$
5 p. Gauss	2	10	$1.1 \cdot 10^{-10}$
Trapezium	16	17	$6.43 \cdot 10^{-3}$
Simpson	8	17	$1.66 \cdot 10^{-5}$
5 p. Gauss	4	20	$1.1 \cdot 10^{-13}$

Tabel 7.2: Vergelijking verschillende integratiemethoden.

Hoofdstuk 8

Niet-lineaire vergelijkingen

8.1 Introductie

In dit hoofdstuk geven we numerieke methoden voor het bepalen van x , die voldoet aan de niet-lineaire vergelijking $f(x) = 0$.

Als voorbeeld nemen we de drukval in een stromende vloeistof. Bij lage stroomsnelheden is de stroming laminair terwijl bij hoge snelheden de stroming turbulent genoemd wordt. Het Reynoldsgetal kan gebruikt worden om te zien of een stroming turbulent is. Voor een stroming in een ronde pijp met diameter $D(m)$ wordt het Reynoldsgetal gegeven door

$$Re = \frac{Dv}{\nu},$$

waarbij v (m/s) de gemiddelde vloeistofsnelheid is en ν (m^2/s) de viscositeit is van de vloeistof. Als het Reynoldsgetal kleiner is dan 2100 dan is de stroming laminair, terwijl als $Re \geq 3000$ dan is er een turbulente stroming.

Voor een turbulente stroming wordt het drukverschil tussen de uit- en inlaat gegeven door

$$P_{\text{uit}} - P_{\text{in}} = \frac{\rho w L v^2}{2gD},$$

hierbij is w de wrijvingsfactor, ρ (kg/m^3) is de soortelijke dichtheid, L (m) is de lengte en g (m/s^2) is de valversnelling. Als de vloeistof deeltjes bevat (zand, papiergezels) dan voldoet de wrijvingsfactor w aan de vergelijking:

$$\frac{1}{\sqrt{w}} = \frac{\ln(Re\sqrt{w}) + 14 - \frac{5.6}{k}}{k},$$

waarbij k een parameter is, die bekend is uit experimenten.

In dit hoofdstuk zullen methoden behandeld worden om w te bepalen uit deze vergelijking, als de waarden van Re en k gegeven zijn.

8.2 Een eenvoudige nulpuntsmethode

Bisectie

De eerste methode, de Bisectie methode, is gebaseerd op de tussenwaarde stelling. Stel f is

een continue functie gedefinieerd op een interval $[a, b]$ waarbij $f(a)$ en $f(b)$ een tegengesteld teken hebben. Volgens de tussenwaarde stelling bestaat er een getal p in (a, b) waar $f(p) = 0$. We nemen aan dat er slechts één zo'n p is. In de methode wordt het interval steeds gehalveerd waarbij in elke stap het interval gekozen wordt waar p in ligt.

We starten de methode met $a_1 = a$ en $b_1 = b$ en nemen voor p_1 het gemiddelde van a_1 en b_1 :

$$p_1 = \frac{1}{2}(a_1 + b_1) .$$

Als $f(p_1) = 0$ dan zijn we klaar, anders heeft $f(p_1)$ hetzelfde teken als $f(a_1)$ of $f(b_1)$. Als $f(p_1)f(a_1) > 0$ dan nemen we $a_2 = p_1$ en $b_2 = b_1$ anders $a_2 = a_1$ en $b_2 = p_1$. Daarna herhalen we de procedure met het interval $[a_2, b_2]$.

Stopcriterium

De Bisectie methode is een iteratieve methode. Dat betekent dat een gebruiker op moet geven wanneer de methode moet stoppen. De volgende stopcriteria kunnen gebruikt worden

$$\frac{|p_n - p_{n-1}|}{|p_n|} \leq \varepsilon ,$$

of

$$|f(p_n)| < \varepsilon .$$

Convergentie

De Bisectie methode kan traag convergeren. Het kan ook gebeuren dat $|p_{i-1} - p| \ll |p_i - p|$. Een groot voordeel is dat de methode altijd convergeert naar een oplossing. De Bisectie methode wordt vaak gebruikt om een goede startoplossing te genereren voor efficiëntere methoden, die later in dit hoofdstuk behandeld zullen worden.

Stelling 8.2.1 *Stel $f \in C[a, b]$ en $f(a) \cdot f(b) < 0$, dan genereert de Bisectie methode een rij $\{p_n\}$ die convergeert naar een nulpunt p van f waarbij*

$$|p_n - p| \leq \frac{b - a}{2^n} , \quad n \geq 1 .$$

Bewijs:

Voor elke $n \geq 1$ hebben we $b_n - a_n = \frac{1}{2^{n-1}}(b - a)$ en $p \in (a_n, b_n)$. Omdat $p_n = \frac{1}{2}(a_n + b_n)$ volgt

$$|p_n - p| \leq \frac{1}{2}(b_n - a_n) = \frac{b - a}{2^n} .$$

□

Afrondfouten

Als we van een functie f de berekende waarden een afrond fout $\bar{\varepsilon}$ hebben dan kan men niet het echte nulpunt p bepalen, immers elk punt van de verzameling

$$I = \{x \in [a, b] \mid |f(x)| < \bar{\varepsilon}\}$$

zou het (of een) nulpunt van f kunnen zijn. Een dergelijk interval noemt men een onbetrouwbaarheidsinterval. Het heeft bijvoorbeeld geen zin om als stopcriterium

$$|f(p_n)| < \varepsilon$$

te nemen als $\varepsilon < \bar{\varepsilon}$ gekozen is. Grote kans dat het algoritme dan altijd blijft itereren. Met behulp van het Taylorpolynoom blijkt dat als p een enkelvoudig nulpunt is van f dan is $f'(p) \neq 0$ en is I ongeveer gelijk aan

$$I \approx [p - \frac{\bar{\varepsilon}}{|f'(p)|}, p + \frac{\bar{\varepsilon}}{|f'(p)|}] .$$

Merk op dat als $|f'(p)|$ dicht bij 0 ligt het bepalen van p een slecht gesteld probleem is.

8.3 Vaste punt iteratie

Een vast punt van een gegeven functie g is een getal p zodanig dat $g(p) = p$. In deze paragraaf beschouwen we het vinden van oplossing voor een vast punt probleem en het verband tussen deze problemen en het vinden van nulpunten.

Het vinden van een nulpunt en vast punt problemen zijn als volgt gerelateerd: als we het nulpunt p zoeken zodat $f(p) = 0$, dan kunnen we een niet unieke functie g definiëren als $g(x) = x - f(x)$ of $g(x) = x + \sqrt{f(x)}$. Aan de andere kant als g een vast punt heeft in p , dan heeft de functie gedefinieerd door $f(x) = x - g(x)$ een nulpunt in p .

In de volgende stelling geven we voldoende voorwaarden voor de existentie en eenduidigheid van een vast punt.

Stelling 8.3.1 1. Als $g \in C[a, b]$ en $g(x) \in [a, b]$ voor alle $x \in [a, b]$ dan heeft g een vast punt in $[a, b]$.

2. Als bovendien $g'(x)$ bestaat voor $x \in [a, b]$ en er is een positieve constante $k < 1$ zodanig dat

$$|g'(x)| \leq k \quad \text{voor alle } x \in [a, b] ,$$

dan is het vaste punt in $[a, b]$ uniek.

Bewijs:

1. Als $g(a) = a$ of $g(b) = b$ dan heeft g een vast punt in een eindpunt. Anders geldt $g(a) > a$ en $g(b) < b$. De functie $h(x) = g(x) - x$ is continu op $[a, b]$ en $h(a) > 0$ en $h(b) < 0$. Uit de tussenwaarde stelling volgt er is een p zodat $h(p) = 0$, dus p is een vast punt van g .

2. Stel $|g'(x)| \leq k < 1$ en er zijn twee vaste punten p en q waarbij $p < q$. Uit de middelwaardestelling volgt er is een $\xi \in [p, q]$ zodat

$$\frac{g(p) - g(q)}{p - q} = g'(\xi) .$$

Dan volgt

$$|p - q| = |g(p) - g(q)| = |g'(\xi)| |p - q| < |p - q| ,$$

dit is een tegespraak, dus $p = q$ en het vaste punt is uniek. □

Om een vast punt van een functie g te benaderen, kiezen we een startwaarde p_0 en bepalen p_n door middel van $p_n = g(p_{n-1})$ voor $n \geq 1$. Als de rij convergeert naar p en g is een continue functie dan geldt

$$p = \lim_{n \rightarrow \infty} p_n = \lim_{n \rightarrow \infty} g(p_{n-1}) = g\left(\lim_{n \rightarrow \infty} p_{n-1}\right) = g(p) ,$$

zodat p een vast punt is van g . Dit noemen we de vaste punt (of Picard) iteratie.

Convergentie van de methode volgt uit de Vaste Punt Stelling.

Stelling 8.3.2 *Stel $g \in C[a, b]$, $g(x) \in [a, b]$, $x \in [a, b]$ en $|g'(x)| \leq k < 1$ voor $x \in [a, b]$. Dan geldt dat de vaste punt iteratie convergeert naar p voor elke waarde $p_0 \in [a, b]$.*

Bewijs:

Onder de gegeven voorwaarden heeft g een uniek vast punt p . Uit de middelwaarde stelling volgt:

$$|p_n - p| = |g(p_{n-1}) - g(p)| = |g'(\xi)| |p_{n-1} - p| \leq k |p_{n-1} - p| .$$

Met inductie volgt

$$\lim_{n \rightarrow \infty} |p_n - p| \leq \lim_{n \rightarrow \infty} k^n |p_0 - p| = 0 ,$$

dus p_n convergeert naar p . □

De bovenstaande stelling kan ook gebruik worden om een goed stopcriterium te geven.

Merk op dat voor $m > n \geq 1$ geldt

$$\begin{aligned} |p_m - p_n| &= |p_m - p_{m-1} + p_{m-1} + \dots + p_{n+1} - p_n| \\ &\leq |p_m - p_{m-1}| + |p_{m-1} - p_{m-2}| + \dots + |p_{n+1} - p_n| \\ &\leq k^{m-n} |p_n - p_{n-1}| + \dots + k |p_n - p_{n-1}| \\ &= (k + \dots + k^{m-n}) |p_n - p_{n-1}| . \end{aligned}$$

Omdat $\lim_{m \rightarrow \infty} p_m = p$ geldt

$$|p - p_n| = \lim_{m \rightarrow \infty} |p_m - p_n| \leq k \sum_{i=0}^{\infty} k^i |p_n - p_{n-1}| = \frac{k}{1-k} |p_n - p_{n-1}| .$$

Dit betekend dat als we stoppen wanneer

$$|p_n - p_{n-1}| \leq \frac{1-k}{k} \varepsilon ,$$

dan geldt $|p - p_n| \leq \varepsilon$.

8.4 De Newton-Raphson methode

De Newton-Raphson methode is één van de krachtigste en bekendste numerieke methoden voor het oplossen van een niet-lineaire vergelijking $f(x) = 0$. We zullen de methode uitleggen

aan de hand van een Taylorpolynoom.

Veronderstel $f \in C^2[a, b]$. Laat $\bar{x} \in [a, b]$ een benadering zijn van de oplossing p zodanig dat $f'(\bar{x}) \neq 0$ en veronderstel dat $|\bar{x} - p|$ klein is. Beschouw het eerste graads Taylorpolynoom:

$$f(x) = f(\bar{x}) + (x - \bar{x})f'(\bar{x}) + \frac{(x - \bar{x})^2}{2}f''(\xi(x)) ,$$

waarbij $\xi(x) \in (x, \bar{x})$. Omdat $f(p) = 0$ geldt

$$0 = f(\bar{x}) + (p - \bar{x})f'(\bar{x}) + \frac{(p - \bar{x})^2}{2}f''(\xi(x)) .$$

Omdat we aangenomen hebben dat $|p - \bar{x}|$ klein is geldt:

$$0 \approx f(\bar{x}) + (p - \bar{x})f'(\bar{x}) .$$

Als we hieruit p oplossen krijgen we

$$p \approx \bar{x} - \frac{f(\bar{x})}{f'(\bar{x})} .$$

Dit motiveert de Newton-Raphson methode, die start met een beginbenadering p_0 en een rij $\{p_n\}$ genereert volgens

$$p_n = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})} , \quad \text{voor } n \geq 1 .$$

We kunnen dezelfde stopcriteria gebruiken als bij de bisectie methode:

$$\frac{|p_n - p_{n-1}|}{|p_n|} < \varepsilon \quad \text{of} \quad |f(p_n)| < \varepsilon .$$

In de stelling hierna geven we het verband tussen de Newton-Raphson iteratie en een vaste punt methode.

Stelling 8.4.1 *Laat $f \in C^2[a, b]$. Als $p \in [a, b]$ zodat $f(p) = 0$ en $f'(p) \neq 0$ dan is er een $\delta > 0$ zodanig dat Newton's methode een rij $\{p_n\}$ genereert, die naar p convergeert voor elke $p_0 \in [p - \delta, p + \delta]$.*

Bewijs:

Beschouw Newton's methode als de vaste punt iteratie $p_n = g(p_{n-1})$ met

$$g(x) = x - \frac{f(x)}{f'(x)} .$$

We proberen eerst een interval $[p - \delta, p + \delta]$ te vinden zodat $|g'(x)| \leq k < 1$ voor $(p - \delta, p + \delta)$. Omdat $f'(p) \neq 0$ en f' continu is bestaat er een $\delta_1 > 0$ zodat $f'(x) \neq 0$ voor $x \in [p - \delta_1, p + \delta_1]$. Dus g is gedefinieerd en continu op $[p - \delta_1, p + \delta_1]$. Bovendien geldt

$$g'(x) = 1 - \frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2} = \frac{f(x)f''(x)}{(f'(x))^2} .$$

Omdat $f \in C^2[a, b]$ is $g'(x) \in C^1[p - \delta_1, p + \delta_1]$. Merk op dat $g'(p) = 0$ omdat $f(p) = 0$. Omdat g' continu is bestaat er een $\delta < \delta_1$ zodat

$$|g'(x)| \leq k < 1 \quad \text{voor alle } x \in [p - \delta, p + \delta] .$$

Als laatste moeten we laten zien dat $g : [p - \delta, p + \delta] \rightarrow [p - \delta, p + \delta]$. Uit de middelwaardestelling volgt

$$|g(x) - g(p)| = |g'(\xi)| |x - p| \quad \text{voor } \xi \text{ tussen } x \text{ en } p .$$

Omdat $x \in [p - \delta, p + \delta]$ geldt $|x - p| < \delta$ en dus

$$|g(x) - g(p)| = |g(x) - p| < |x - p| < \delta .$$

Hiermee is de stelling bewezen. □

Om een goede vergelijking van convergentie mogelijk te maken voeren we de volgende definitie in.

Definitie

Stel $\{p_n\}_{n=0}^\infty$ is een rij die convergeert naar p , met $p_n \neq p$ voor alle n .

Als er positieve constanten λ en α bestaan met

$$\lim_{n \rightarrow \infty} \frac{|p_{n+1} - p|}{|p_n - p|^\alpha} = \lambda , \tag{8.1}$$

dan convergeert $\{p_n\}$ naar p met orde α en asymptotische constante λ .

In het algemeen convergeert een hogere orde methode sneller dan een lagere orde methode. De waarde van de asymptotische constante is minder belangrijk. Twee gevallen zijn belangrijk

- als $\alpha = 1$ dan noemen we het proces lineair convergent,
- als $\alpha = 2$ dan noemen we het proces kwadratisch convergent.

Het is eenvoudig om in te zien dat elke convergerende vaste punt methode tenminste lineair convergent is.

Voor Newton Raphson merken we op

$$0 = f(p) = f(p_n) + (p - p_n)f'(p_n) + \frac{(p - p_n)^2}{2} f''(\xi_n) , \xi_n \in (p_n, p) .$$

Volgens de definitie geldt

$$0 = f(p_n) + (p_{n+1} - p_n)f'(p_n) .$$

Aftrekken geeft

$$p_{n+1} - p_n = (p_n - p)^2 \frac{f''(\xi)}{2f'(p_n)} .$$

Dit is inderdaad van de gedaante (8.1) met $\alpha = 2$ en $\lambda = \frac{f''(p)}{2f'(p)}$, zodat Newton Raphson kwadratisch convergent is.

Er zijn allerlei varianten op de Newton Raphson methode bekend. We zullen deze alleen noemen.

Secant methode (Koorden Newton)

Vervang $f'(p_{n-1})$ door $\frac{f(p_{n-1}) - f(p_{n-2})}{p_{n-1} - p_{n-2}}$. Voordeel hiervan is dat f' niet bepaald hoeft te worden.

Regula Falsi methode

Stel er zijn twee benaderingen gegeven p_0 en p_1 zodanig dat $f(p_0) \cdot f(p_1) < 0$. Het schema om p te benaderen is nu:

- stap 1 $q_0 = f(p_0)$
 $q_1 = f(p_1)$
 $n = 2$
- stap 2 doe zolang $n \leq N_0$
- stap 3 $p = p_1 - q_1(p_1 - p_0)/(q_1 - q_0)$
 $q = f(p)$
 $n = n + 1$
- stap 4 als $q \cdot q_1 < 0$ dan $p_0 = p_1$; $q_0 = q$
- stap 5 $p_1 = p$
 $q_1 = q$

8.5 Stelsels niet-lineaire vergelijkingen

Een stelsel van niet-lineaire vergelijkingen heeft de vorm

$$\begin{aligned} f_1(x_1, \dots, x_n) &= 0, \\ &\vdots \\ f_n(x_1, \dots, x_n) &= 0. \end{aligned}$$

We noteren dit ook als

$$\mathbf{F}(\mathbf{x}) = \mathbf{0}. \quad (8.2)$$

De Newton-Raphson methode voor (8.2) wordt gegeven door

$$\mathbf{x}^{(p)} = \mathbf{x}^{(p-1)} - J(\mathbf{x}^{(p-1)})^{-1} \mathbf{F}(\mathbf{x}^{(p-1)}).$$

Hierbij wordt $J(\mathbf{x})$ de Jacobiaan matrix genoemd, die gedefinieerd is als:

$$J(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_n(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial f_n(\mathbf{x})}{\partial x_n} \end{pmatrix}.$$

Als het exact uitrekenen van $\frac{\partial f_j}{\partial x_k}(\mathbf{x})$ onmogelijk is, dan kunnen we de partiële afgeleiden vervangen door eindige differentie benaderingen. Bijvoorbeeld

$$\frac{\partial f_j}{\partial x_k}(\mathbf{x}) \approx \frac{f_j(\mathbf{x} + \mathbf{e}_k h) - f_j(\mathbf{x})}{h},$$

waarbij e_k de k^e eenheidsvector is. Dit noemen we een quasi-Newton methode.

8.6 Toepassingen op niet-lineaire differentiaalvergelijkingen

In deze paragraaf zullen we laten zien, hoe bovenstaande methoden gebruikt kunnen worden voor het oplossen van niet-lineaire differentiaalvergelijkingen.

8.6.1 Niet-lineaire beginwaarde problemen

Als we een impliciete tijdsintegratie methode gebruiken voor het oplossen van een beginwaarde probleem,

$$\begin{aligned} y' &= f(t, y), \quad t \in [0, T], \\ y(0) &= y_0 \end{aligned}$$

dan moet er in elke tijdstip een niet-lineaire vergelijking opgelost worden. Als voorbeeld nemen we Euler achterwaarts, bepaal u_{j+1} uit

$$u_{j+1} - hf(t_{j+1}, u_{j+1}) = u_j.$$

Als startoplossing kiest men vaak $u_{j+1}^{(0)} = u_j$ terwijl de iteranden $u_{j+1}^{(n)}$ bepaald worden uit $u_{j+1}^{(n-1)}$ met behulp van één van de methoden genoemd in paragraaf 8.2, 8.3 of 8.4. Omdat u_j vaak een goede beginschatting is zijn er vaak maar enkele iteraties nodig.

Voor het oplossen van een eerste orde beginwaarde probleem

$$x' = f(t, x) \quad \text{en} \quad x(0) = x_0$$

met een impliciete methode maakt men gebruik van een methode uit paragraaf 8.5.

8.6.2 Niet-lineaire randwaarde problemen

Een niet-lineair randwaarde probleem kunnen we schrijven als

$$\begin{aligned} -(py')' + g(y', y, x) &= 0, \quad 0 < x < 1, \\ y(0) &= 0 \quad \text{en} \quad y(1) = 0. \end{aligned} \tag{8.3}$$

De algemene aanpak is dat (8.3) wordt opgelost met een iteratief proces waarin alleen lineaire vergelijkingen opgelost worden. De lineaire randwaarde problemen lossen we op met methoden uit Hoofdstuk 5. De iteratieve methode genereert een rij iteranden $y^{(0)}, y^{(1)}, y^{(2)}, \dots$ zodanig dat $y^{(n)} \rightarrow y$, $n \rightarrow \infty$ met y de oplossing van (8.3).

Picard

De iteratiemethode van Picard werkt als volgt:

```

Kies een startfunctie  $y^{(0)}(x)$ 
for  $n = 1$  step 1 until nit do
begin
    bepaal  $y^{(n)}$  uit
         $-(py')' = -g(y^{(n-1)'}, y^{(n-1)}, x)$ ,
         $y(0) = 0$  en  $y(1) = 0$ 
end
```

Hierbij is *nit* het aantal iteraties dat wordt uitgevoerd.

Newton-Raphson

Met behulp van Taylor ontwikkeling volgt:

$$g(y', y, x) \simeq g(y^{(n-1)'}, y^{(n-1)}, x) + (y' - y^{(n-1)'})r + (y - y^{(n-1)})q$$

waarbij

$$\begin{aligned} r &= \frac{\partial g}{\partial y'}(y^{(n-1)'}, y^{(n-1)}, x) \quad \text{en} \\ q &= \frac{\partial g}{\partial y}(y^{(n-1)'}, y^{(n-1)}, x) . \end{aligned}$$

De Newton-Raphson methode is nu:

```

Kies een startfunctie  $y^{(0)}(x)$ 
for  $n = 1$  step 1 until nit do
begin
  bepaal  $y^{(n)}$  uit
     $-(py')' + ry' + qy = y^{(n-1)'}r + y^{(n-1)}q - g(y^{(n-1)'}, y^{(n-1)}, x)$ 
     $y(0) = 0$  en  $y(1) = 0$ 
end

```

8.7 Samenvatting

In dit hoofdstuk zijn de volgende begrippen behandeld:

- Bisectie methode
- stopcriteria
- convergentie
- vaste punt iteratie (Picard)
- Newton Raphson
- Secant methode, Regula Falsi methode
- stelsels niet-lineaire vergelijkingen
- niet-lineaire beginwaarde problemen
- niet-lineaire randwaarde problemen

Index

- absoluut stabiel, 38
- achterwaartse differentie, 22
- achterwaartse methode van Euler, 34
- Adams-Bashforth, 50
- afbreekfout, 35, 57
- afrondfout, 3

- beginwaarde probleem, 33
- Bisectie, 84

- centrale differentie, 22, 25, 27
- cijferverlies, 4
- conditiegetal, 57
- consistent, 36, 58
- convectie-diffusie vergelijking, 63
- convergent, 41
- convergentie, 58
- Crank-Nicolson, 70
- cubische spline, 17

- differentievergelijking, 51
- Dirichlet, 55

- eindige differentie, 55
- Euclidische norm, 56
- expliciet, 34
- extrapolatie, 12
- extrapoleren, 9

- floating point getal, 3
- functie evaluatie, 42

- Gershgorin, 57
- globale fout, 41

- Hermite interpolatie, 14
- hogere orde beginwaarde, 43

- impliciet, 34
- interpolatie, 7

- Jacobiaan, 90

- kwadratisch convergent, 89

- Lagrange interpolatie, 9, 26
- lineair convergent, 89
- lineaire interpolatie, 7

- meerstaps methoden, 50
- methode van Heun, 34

- Neumann, 55, 61
- Newton-Raphson, 87
- niet-lineair stelsel, 90
- niet-lineaire differentiaalvergelijkingen, 91
- niet-lineaire vergelijking, 84

- O-symbool, 5
- onbetrouwbaarheidsinterval, 86

- principal root, 52

- quasi-Newton, 91

- randwaarde probleem, 54
- Regula Falsi, 90
- Richardson's extrapolatie, 28
- Robbins, 55
- Runge-Kutta methoden, 49

- Secant, 90
- semi-discretisatie, 67
- splines, 16
- spurious root, 52
- stabiel, 38
- stabiliteit, 37, 44, 58
- stabiliteitsgebied, 45
- stelsel, 42
- stijve stelsels, 47
- stopcriterium, 85
- superstabiel, 49

- Taylorpolynoom, 12

testvergelijking, 35
Trapeziumregel, 34

upwind discretisatie, 63

vast punt probleem, 86
versterkingsfactor, 35
versterkingsmatrix, 45
voorwaartse differentie, 21, 26
voorwaartse methode van Euler, 33

warmtevergelijking, 66